



UNIVERSITY OF GENOA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

From Perception to Action: Data-Efficient and Open-Set Learning for Adaptive Human-Robot Interaction

by

Stefano Berti

Thesis submitted for the degree of *Doctor of Philosophy* (38° cycle)

March 2026

Giulia Pasquale
Lorenzo Natale
Paolo Massobrio

Supervisor
Supervisor
Head of the PhD program

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

To the people who stood by me throughout this journey.
Your support made the difficult parts lighter and the meaningful ones deeper.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Stefano Berti

July 2026

Acknowledgements

I would like to thank my supervisors, Giulia Pasquale and Lorenzo Natale, for their guidance, trust, and support throughout this Ph.D. Their advice helped me shape not only this thesis, but also the way I approach research problems and scientific work more broadly.

I am grateful to all the collaborators, co-authors, and colleagues who contributed to this journey. Working within the Humanoid Sensing and Perception environment gave me the opportunity to discuss ideas openly, learn continuously, and grow through shared projects, deadlines, and experiments. Many of the results presented in this dissertation were made possible by that collective effort.

Finally, I would like to thank my family and the people close to me for their patience, encouragement, and constant support over these years. Their presence gave me perspective beyond research and helped me carry this work through to the end.

Abstract

Robots operating alongside people must recognize actions and concepts under limited supervision, handle novelty reliably, and ultimately translate perception into appropriate behavior. This thesis is guided by the following question:

“How can a robot learn to interpret human actions and task-relevant concepts in a data-efficient and open-world setting, while preserving the reliability and semantic flexibility required for safe interaction and future embodied decision-making?”

These objectives are addressed through three connected research directions.

A first direction focuses on action recognition under scarce supervision and unknown classes. To the best of our knowledge, it introduces for the first time the problem of *few-shot open-set action recognition* (FSOSAR), where a model is given only a small support set of labeled examples for the classes of the current recognition problem and must both classify known query actions and reject actions outside that support set. To tackle this setting, we first proposed *Feature-Residual Discrimination* (FR-Disc) and validated it in a simplified one-shot setting using 3D skeleton sequences (Berti et al., 2022). FR-Disc operates by letting a few-shot backbone select the most likely support class and then using a lightweight discriminator on the resulting support–query feature residual to decide acceptance versus rejection. We later extended and evaluated FR-Disc more extensively in the RGB video domain by benchmarking against representative open-set strategies across standard action recognition datasets and backbones (Berti et al., 2026).

A second direction investigates how action recognition and visual understanding can be expanded with semantic flexibility, so that the robot is not limited to a fixed vocabulary of predefined labels. This direction studies how large frozen pretrained vision-language backbones can be adapted to new actions, compositions, or concepts by learning a small set of prompt or token parameters while keeping the backbone fixed. In the video domain, the first contribution presents *DualVClip* (Calabrese et al., 2024), a prompt-learning approach that extends a CLIP-based open-vocabulary video–text model to zero-shot multi-label recognition

of *verb-object* actions, and analyzes how verb-object compositionality and the design of verb-object class splits affect generalization. The second contribution studies vision-language model *personalization*: expanding the model vocabulary with a new token that represents a novel concept learned from a few images, and evaluating how the learned concept composes with known context words in both discriminative (personalized image retrieval) and generative (personalized image generation) settings via the *ConCon-Chi* benchmark (Rosasco et al., 2024).

A third direction connects perception with action and analyzes how these perceptual components may support safe embodied interaction. The thesis reports an embodied human-robot interaction study on *ergoCub*, a humanoid robot, where the same interaction skills are implemented either with a task orchestrator that is manually designed using behavior trees, or learned end-to-end via behavior cloning. To this end, in the final phase of the PhD, we developed *metaCub*, a software system for teleoperation and data collection, and used it to record interaction demonstrations. These data is used to train an end-to-end behavior-cloning approach using vision-language-action (VLA) policy models, i.e., policies conditioned on visual inputs and language. The analysis highlights practical limitations encountered in this setting, including limited perceptual generalization across people and limited temporal modeling in current VLA policies, and motivates future work on integrating stronger action-recognition representations into end-to-end policies.

Keywords: Human-Robot Interaction, Action Recognition, Open-Set Recognition, Vision-Language Models, Personalized Concept Learning.

Table of contents

List of Publications	xi
List of Figures	xii
List of Tables	xiv
List of Acronyms	xvi
I Contributions	1
1 Introduction	2
1.1 Motivation	2
1.2 Research Objectives	4
1.3 Thesis Outline	5
2 Few-Shot Open-Set Action Recognition	7
2.1 Introduction	7
2.2 Background	8
2.2.1 Problem Formulation	9
2.2.2 Few-Shot Learning Paradigms	12
2.2.3 Adopted Few-Shot Models	16
2.3 Methods	18
2.3.1 Feature-Residual Discriminator	18
2.3.2 Open-Set Extension of TRX with FR-Disc	19
2.3.3 Open-Set Extension of SAFSAR with FR-Disc	21
2.4 Experiments	22
2.4.1 First Evaluation on Skeleton Input	22

2.4.2	Extended RGB Evaluation	23
2.5	Summary	24
3	Vision-Language Models for Semantic Generalization and Personalization	26
3.1	Introduction	26
3.2	DualVCLIP: Zero-Shot Multi-Label Action Recognition	29
3.2.1	Background	29
3.2.2	Method	31
3.2.3	Experimental Validation	33
3.3	ConCon-Chi: Personalized Retrieval and Generation	34
3.3.1	Background	34
3.3.2	Benchmark Design	35
3.3.3	Dataset Acquisition	36
3.3.4	Experimental Validation	38
3.4	Summary	39
4	Behavior Cloning and Toward End-to-End Policies	41
4.1	Introduction	41
4.2	Background	42
4.2.1	A Modular Interaction Demonstrator on ergoCub	43
4.2.2	Behavior Cloning from Demonstrations	46
4.2.3	Recent Policy Learning for Humanoids	48
4.3	Methods	48
4.3.1	The metaCub Teleoperation and Data Collection Interface	49
4.3.2	An End-to-End Interaction Demo on ergoCub	50
4.4	Summary	54
II	Research Chapters	55
5	One-Shot Open-Set Skeleton-Based Action Recognition	56
5.1	Introduction	57
5.2	Related Work	60
5.2.1	Few-Shot Action Recognition (Videos)	60
5.2.2	Few-Shot Action Recognition (Skeleton)	60
5.2.3	Few-Shot Open-Set Recognition (Images)	61
5.3	Background	61

5.4	Problem Formulation	63
5.5	Proposed Solution	64
5.5.1	Model	64
5.5.2	Training	65
5.6	Experimental Validation	66
5.6.1	Comparison with Baseline (Quantitative Validation)	67
5.6.2	Discriminator's Confusion Matrix (Quantitative Validation)	69
5.6.3	Qualitative Results	69
5.7	Conclusion	75
6	A Baseline Study and Benchmark for Few-Shot Open-Set Action Recognition with Feature Residual Discrimination	76
6.1	Introduction	77
6.2	Related Work	79
6.2.1	Few-Shot Action Recognition	79
6.2.2	Open-Set Recognition	79
6.2.3	Few-Shot Open-Set Recognition	80
6.3	Problem Definition and Metrics	80
6.3.1	Task Definition	80
6.3.2	Performance Metrics	81
6.4	Methods	81
6.4.1	Few-Shot Action Recognition Models	81
6.4.2	Open-Set Techniques	82
6.5	Experimental Setup	84
6.5.1	FSOS-AR Benchmark Suite	84
6.5.2	Implementation Details	85
6.6	Results	85
6.6.1	Evaluating the Softmax Baseline	87
6.6.2	Evaluating Open-Set Techniques	88
6.7	Conclusion	90
7	The Impact of Compositionality in Zero-Shot Multi-Label Action Recognition for Object-Based Tasks	92
7.1	Introduction	93
7.2	Related Work	96
7.2.1	Multi-Label Action Recognition	96

7.2.2	Multi-Label Zero-Shot Action Recognition	96
7.2.3	Compositionality in Action Recognition	97
7.3	Background	97
7.4	Problem Definition	99
7.5	Method	100
7.5.1	Experiments	101
7.6	Results	102
7.6.1	Comparison with SOTA	102
7.6.2	Compositionality	103
7.6.3	Spatial vs Temporal Generalization	104
7.7	Conclusions	105
8	ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language Tasks	107
8.1	Introduction	108
8.2	Related Work	110
8.3	Concept-Context Chimera Benchmark	112
8.3.1	Dataset Overview	112
8.3.2	Acquisition and Annotation	115
8.4	Personalized Text-to-Image Retrieval	116
8.4.1	Benchmark Task	117
8.4.2	Benchmark Results	118
8.4.3	Analysis of Results	119
8.5	Personalized Text-to-Image Generation	122
8.5.1	Benchmark Task and Results	122
8.5.2	Analysis of Results	123
8.6	Conclusion	125
III	Conclusion and Outlook	126
9	Discussion	127
9.1	From Recognition to Interaction	127
9.2	Cross-Cutting Lessons	128
9.3	Limits and Open Problems	129
	References	130

Appendix A	Supplementary Material for Chapter 6	143
A.1	Implementation Details	143
A.1.1	MLS vs MSS	145
A.2	Metrics Definition	147
A.2.1	AUROC	147
A.2.2	AUPR	147
A.2.3	FS ACC	147
A.2.4	OS ACC	148
A.2.5	OSCR	148
A.3	Discriminator Architecture	148
A.3.1	SAFSAR	148
A.3.2	STRM	149
A.4	Splits	150
A.4.1	Diving48	150
A.4.2	NTURGBD120	150
Appendix B	Supplementary Material for Chapter 8	153
B.1	Concept-Context Chimera Benchmark	153
B.1.1	Concepts: Pictures and Descriptions	154
B.1.2	Contexts: Kinds and Environments	154
B.1.3	Acquisition and Annotation: Procedure Details	155
B.1.4	Benchmark Splits	155
B.1.5	Dataset Examples	157
B.2	Personalized Text-to-Image Retrieval	158
B.2.1	Metrics Definition	158
B.2.2	Concept and Context Bias: Implementation Details	161
B.2.3	Methods: Implementation Details	162
B.2.4	Additional Benchmark Metrics	163
B.2.5	Additional Analysis of Results	163
B.2.6	TIR Failure Cases	164
B.3	Personalized Text-to-Image Generation	164
B.3.1	Metrics: Implementation Details	164
B.3.2	Methods: Implementation Details	164
B.3.3	Benchmark Results: FID Numbers	166
B.3.4	TIG Failure Cases	166

List of Publications

Included Publications

- Stefano Berti, Andrea Rosasco, Michele Colledanchise, and Lorenzo Natale. *One-Shot Open-Set Skeleton-Based Action Recognition*. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022.
- Stefano Berti, Giulia Pasquale, and Lorenzo Natale. *A Baseline Study and Benchmark for Few-Shot Open-Set Action Recognition with Feature Residual Discrimination*. Accepted at the *International Conference on Pattern Recognition (ICPR)*, 2026. Also available as *arXiv preprint arXiv:2603.04125*.
- Carmela Calabrese, Stefano Berti, Giulia Pasquale, and Lorenzo Natale. *The Impact of Compositionality in Zero-Shot Multi-Label Action Recognition for Object-Based Tasks*. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2024.
- Andrea Rosasco, Stefano Berti, Giulia Pasquale, Damiano Malafronte, Shogo Sato, Hiroyuki Segawa, Tetsugo Inada, and Lorenzo Natale. *ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language Tasks*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Excluded Publications

The following co-authored publication belongs to the author's broader research activity but is not reproduced in this dissertation because it falls outside the main thematic arc of the thesis.

- Andrea Rosasco, Stefano Berti, Fabrizio Bottarel, Michele Colledanchise, and Lorenzo Natale. *Towards Confidence-guided Shape Completion for Robotic Applications*. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 580–586, 2022.

List of Figures

1.1	The ergoCub humanoid platform used in this thesis	2
2.1	Comparison of few-shot, open-set, and few-shot open-set recognition	10
2.2	Backbone-specific insertion of Feature-Residual Discrimination	22
3.1	Two research directions built on the same vision-language foundation . . .	28
3.2	DualVCLIP architecture	32
3.3	Concept-context structure of ConCon-Chi	37
3.4	Concept and context bias on ConCon-Chi	38
4.1	State machine of the interaction policy	44
4.2	Behavior tree used in the ergoCub demonstration	45
4.3	Representative snapshots from the modular ergoCub demo	46
4.4	Hardware and command interface for behavior cloning on ergoCub	49
4.5	metaCub teleoperation interface and data collection	50
4.6	Representative snapshots from the behavior-cloning demo	52
4.7	Comparison between modular and behavior-cloning pipelines	52
5.1	An overview of our approach for the One-Shot Three-Way problem.	59
5.2	Comparison between the proposed model (<i>TRX-OS</i>) and the baseline (<i>EXP</i>)	68
5.3	First qualitative example visualization.	71
5.4	Second qualitative example visualization.	71
5.5	Third qualitative example visualization.	72
5.6	Support set as a reference for the two unknown queries.	73
5.7	First unknown query (Cross arms)	74
5.8	Second unknown query (Relaxing)	74
6.1	Comparison between the closed-set Softmax baseline (SAFSAR) and its open-set extension FR-Disc on an unknown query.	78

6.2	Overview of closed- and open-set prediction flows	82
6.3	Analysis of open-set scoring and performance correlation.	88
6.4	Qualitative comparison between the SAFSAR Softmax baseline and FR-Disc on the SSv2 test set.	89
7.1	Overview of Dual-VCLIP.	94
7.2	Dual-VCLIP per-frame inference examples.	98
7.3	Confusion matrix for unseen classes.	105
8.1	Concept-context structure of ConCon-Chi.	107
8.2	Example concepts in ConCon-Chi.	113
8.3	Fine-grained context queries.	114
8.4	Concept-context matrices.	114
8.5	Concept and context biases.	116
8.6	New-meaning learning analysis.	119
8.7	Concept-concept compositionality.	120
8.8	Personalized TIG benchmark.	121
8.9	Qualitative results for personalized TIG.	124
A.1	Qualitative comparison between the baseline SAFSAR and its adaptation to the open-set scenario with FR-Disc on the NTURGBD dataset	144
A.2	Qualitative comparison between the baseline SAFSAR and its adaptation to the open-set scenario with FR-Disc on the SSv2 dataset	145
A.3	Open-set performances comparison between MLS and MSS for the SAFSAR model.	146
A.4	Performances of SAFSAR and STRM on easy spatial dataset (UCF101) and fine-grained temporal datasets (SSv2, Diving48)	152
B.1	Concepts in ConCon-Chi.	153
B.2	Context kinds and environments.	156
B.3	Ground-truth images per query.	157
B.4	Examples involving the chimeric concepts.	159
B.5	Examples involving the common concepts.	160
B.6	Performance per query kind.	161
B.7	Retrieval failure cases.	162
B.8	Generation failure cases.	165

List of Tables

2.1	Comparison of few-shot paradigms most relevant to this thesis	15
2.2	TRX and SAFSAR as representative base models	17
2.3	Summary of the two FSOSAR studies	24
3.1	High-level summary of the two studies in this research direction	40
4.1	Modular interaction pipeline versus behavior cloning	53
5.1	FSOS-ACC metric for our model (TRX-OS) and the baseline (EXP).	68
5.2	Confusion matrix of our model for the OS task with respect to the test classes.	70
5.3	First qualitative example results.	70
5.4	Second qualitative example results.	71
5.5	Third qualitative example results.	72
6.1	Combined results for SAFSAR in 5-way 1-shot and 5-shot settings.	85
6.2	Results for STRM in 5-way 5-shot settings.	86
7.1	Charades split details.	102
7.2	Comparison of zero-shot action recognition methods on Charades.	103
7.3	mAP on different subsets of the test set.	104
7.4	Verb-split vs object-split results.	105
8.1	Comparison of related datasets.	110
8.2	Personalized TIR benchmark.	118
8.3	Concept-context compositionality.	119
8.4	Comparing generated and GT images.	121
9.1	Thesis-level view of the three research directions	128
A.1	Training iterations and learning rates across datasets for SAFSAR 5-way 5-shot	143

A.2	Training iterations and learning rates across datasets for SAFSAR 5-way 1-shot	144
A.3	Training iterations and learning rates across datasets for STRM 5-way 5-shot	144
A.4	Gain from using MLS instead of MSS on SAFSAR.	146
B.1	Concept descriptions.	167
B.2	Broken FashionIQ URLs by split.	168
B.3	Additional metrics for the personalized TIR benchmark.	168
B.4	Personalized TIG benchmark.	168

List of Acronyms

AR Action Recognition

CLIP Contrastive Language-Image Pretraining

FSAR Few-Shot Action Recognition

FSOS Few-Shot Open-Set

FSOS-AR Few-Shot Open-Set Action Recognition

HRI Human-Robot Interaction

MLS Maximum Logit Score

MSS Maximum Softmax Score

OSCR Open-Set Classification Rate

TIG Text-to-Image Generation

TIR Text-to-Image Retrieval

VLM Vision-Language Model

VLA Vision-Language-Action

ZSL Zero-Shot Learning

Part I

Contributions

Chapter 1

Introduction

1.1 Motivation



Figure 1.1 The ergoCub humanoid platform used throughout this thesis, shown in illustrative household interaction scenarios. The image is a conceptual rendering rather than an experimental result.

Robots are progressively moving beyond isolated industrial workcells and into environments in which they must operate alongside people, adapt to changing situations, and respond safely to partially specified tasks. In these settings, perception is not merely a support module: it is a prerequisite for effective Human-Robot Interaction (HRI), collaboration, and long-term autonomy ([Akkaladevi and Heindl, 2015](#); [Rodomagoulakis et al., 2016](#); [Roitberg et al., 2014](#)). A robot that acts in a human-centered environment must be able to interpret what peo-

ple are doing, reason about task-relevant objects and contexts, and recognize when its current knowledge is insufficient. A central component of this problem is action understanding. In practical deployments, however, standard action recognition pipelines are often mismatched with the constraints of robotics. They typically assume a closed world in which all relevant categories are known in advance and a large amount of labeled data is available for each of them. This assumption is difficult to sustain when a robot is expected to learn user-specific behaviors, adapt to new tasks on demand, or operate in environments where novel actions and object configurations are routinely encountered (Ayub and Wagner, 2020; Chiatti et al., 2020; Wang et al., 2021c). For this reason, two requirements become especially important: *data efficiency* and *open-world reliability*.

The first requirement motivates the study of few-shot action recognition, namely the ability to acquire new action categories from a very small number of demonstrations (Perrett et al., 2021; Tang et al., 2024; Thatipelli et al., 2022). The second requires models to move beyond the traditional closed-set assumption and to explicitly reject actions that do not belong to the robot’s current vocabulary, rather than forcing every observation into a known class. In an embodied setting, failure to detect novelty is not only a recognition error: it can become a safety and interaction problem. This thesis therefore begins from the premise that useful perception for HRI must be uncertainty-aware (open-set aware) as well as accurate.

Anyway, even a robust few-shot open-set recognizer remains limited if the robot can only reason over pre-defined labels. Human activities are often compositional, object-dependent, and described more naturally through language than through a rigid class taxonomy. Recent vision-language models (VLMs) have made it possible to leverage broad semantic priors and open-vocabulary representations, enabling a shift from closed label prediction toward richer forms of generalization (Li et al., 2022; Radford et al., 2021; Rombach et al., 2022). This opens two complementary directions: zero-shot reasoning over unseen action compositions, and personalized learning of novel concepts from only a few examples.

From a robotics perspective, these developments suggest a broader trajectory. Few-shot open-set recognition addresses the problem of *when to trust a prediction*; vision-language modeling addresses the problem of *how to generalize beyond a fixed vocabulary*. Together, these capabilities are necessary steps toward robots that can use perception and semantic understanding to support embodied decision-making and action. The focus of this thesis is mainly on the first two aspects, but it is guided throughout by the longer-term objective of enabling robots to operate in open-ended environments where perception, semantic understanding, and control cannot be designed in isolation.

This dissertation studies that progression explicitly through three connected research directions. The first concerns action recognition under limited supervision and open-set conditions. The second concerns semantic and compositional generalization with vision-language models. The third discusses novel approaches that directly link robot perception and motor actuation. The common thread is the search for representations that are data-efficient, uncertainty-aware, and flexible enough to remain useful when the world does not conform to a closed training taxonomy.

1.2 Research Objectives

The overarching goal of this thesis is to study how a robot can acquire perceptual representations that remain useful under realistic open-world conditions, where data are scarce, task specifications evolve over time, and not every relevant action or concept can be enumerated in advance. More specifically, the thesis investigates how to bridge the gap between classical recognition pipelines, which are accurate but rigid, and more adaptive systems that can learn from few examples, reject novelty, and exploit semantic priors to extend their competence.

Within this context, the central research question of the dissertation is how to design perceptual representations and learning strategies that remain data-efficient and reliable under open-world conditions, while still supporting semantic flexibility for safe interaction and future embodied decision-making.

This question is pursued through four main research objectives.

1. **Develop data-efficient methods for action understanding in robotic settings.**

Robots operating in human environments cannot rely on collecting and annotating large datasets for every user, task, or interaction pattern. A first objective is therefore to study learning formulations in which new action classes can be acquired from very limited supervision. This includes few-shot protocols that are closer to the conditions of deployment, where the cost of adaptation must remain low and personalization must be possible from only a handful of examples.

2. **Equip action recognition systems with explicit open-set awareness.**

In realistic HRI scenarios, not every observed behavior belongs to the robot’s current action vocabulary. A second objective is thus to investigate how few-shot recognition can be coupled with mechanisms for detecting and rejecting unknown actions. The

aim is not only to improve benchmark robustness, but to make uncertainty operationally meaningful: the system should know when to abstain instead of returning an overconfident but unsafe prediction.

3. Move beyond fixed action vocabularies through semantic and compositional representations.

Even when uncertainty is handled correctly, closed label sets remain a severe limitation. A third objective is to explore whether vision-language models can support more flexible forms of action and concept understanding by leveraging textual priors, compositional structure, and open-vocabulary representations. In this thesis, this direction is studied both through zero-shot multi-label action recognition for *verb-object* tasks and through personalized learning of novel concepts from a few images.

4. Position these perceptual advances within a broader embodied perspective.

The final objective is to connect the previous contributions to the longer-term goal of end-to-end robotic intelligence. Robust recognition, novelty awareness, and semantic flexibility are valuable in their own right, but they are especially relevant when considered as ingredients for embodied decision-making. The thesis therefore also asks how the representations developed in the earlier chapters may inform, condition, or motivate future work on integrated sensory-motor policies.

This thesis develops and connects a set of contributions that progressively address perceptual and representational problems toward embodied intelligence. The first part of this program focuses on reliable action recognition under novelty, the second on semantic generalization beyond closed vocabularies, and the final part investigate novel approaches that directly link perception and action execution..

1.3 Thesis Outline

This dissertation is organized into three parts.

Part I (Contributions) introduces the research directions and provides the unifying narrative. Chapter 1 motivates the thesis from the perspective of open-world HRI and states the guiding research question and objectives. Chapter 2 presents few-shot open-set action recognition, introducing FSOSAR and the Feature-Residual Discrimination approach, from an initial one-shot skeleton setting to an extensive RGB video evaluation. Chapter 3 develops the vision-language direction through two contributions: DualVCLIP for zero-shot

multi-label recognition of verb–object actions, and ConCon-Chi for evaluating personalized concept learning and compositional generalization. Chapter 4 describes the application of these perceptual themes to embodiment by explaining a deployed ergoCub interaction demo and the comparison between modular pipelines to teleoperation and behavior cloning. The background sections of Part I are not intended as an exhaustive literature review, nor do they aim to cover every existing paradigm of few-shot, open-set, or vision-language action recognition. Their goal is to provide the *minimal pointers* needed to follow the unifying narrative of the thesis and to position the contributions, while the detailed and self-contained discussion of the related work for each contribution is given in the corresponding research chapter of Part II.

Part II (Research Chapters) collects the full papers in full. Chapter 5 presents the one-shot skeleton work (Berti et al., 2022). Chapter 6 presents the RGB video extension and benchmarking of FR-Disc (Berti et al., 2026). Chapter 7 presents DualVCLIP (Calabrese et al., 2024). Chapter 8 presents ConCon-Chi (Rosasco et al., 2024). These chapters preserve the paper material substantially unchanged and serve as the documentary core of the dissertation.

Part III (Conclusion and Outlook) closes the thesis. Chapter 9 discusses cross-cutting lessons that emerge across the three directions and outlines how uncertainty handling, semantic transfer, and embodiment constraints jointly shape future work toward adaptive and reliable human-centered robot intelligence.

Chapter 2

Few-Shot Open-Set Action Recognition

2.1 Introduction

This chapter studies how a robot can recognize actions when supervision is scarce and the action vocabulary is incomplete. This line of work treats *few-shot open-set action recognition* (FSOSAR) as a distinct setting with its own protocols and failure modes, and proposes Feature-Residual Discrimination as a solution. The corresponding research is documented in the research chapters reproduced later as Chapter 5 and Chapter 6. Our first study validates residual-based rejection in a deliberately simplified one-shot skeleton setting, where the modality and protocol make the core mechanism easier to isolate and analyze (Berti et al., 2022); our second study extends the analysis to RGB video, stronger backbones, and multiple datasets (Berti et al., 2026). This chapter introduces the problem setting, situates it within the few-shot literature, and clarifies why metric- and prototype-based models provide a natural foundation for open-set action recognition.

Action recognition is a core capability for HRI, because robots operating around people must interpret gestures, task-oriented behaviors, and interaction cues in order to respond appropriately (Akkaladevi and Heindl, 2015; Rodomagoulakis et al., 2016; Roitberg et al., 2014). In a laboratory benchmark, this problem is often cast as supervised closed-set classification over a fixed set of action labels. In deployment, however, that assumption is fragile. The robot may need to learn a new action from only a handful of demonstrations, and it may also observe behaviors that do not belong to its current action vocabulary.

These two difficulties point to two distinct recognition requirements. The first is *few-shot action recognition*: learning to classify novel actions from a small support set rather than from a large labeled training corpus (Perrett et al., 2021; Tang et al., 2024; Thatipelli et al., 2022). The second is *open-set recognition*: detecting when a query does not belong to the

known label space and should therefore be rejected instead of being forced into an incorrect class (Dhamija et al., 2018; Jeong et al., 2021; Liu et al., 2020; Vaze et al., 2021). Although they address different aspects of recognition, robotics applications often require both. A robot that confidently assigns an unseen action to the “closest” known class is not merely making a classification mistake; it is failing to represent that its knowledge is currently insufficient.

This makes few-shot open-set action recognition a particularly meaningful problem for HRI. Few-shot learning addresses data efficiency and rapid adaptation, while open-set recognition addresses reliability under novelty. Their combination is especially relevant in action recognition, where the set of possible human behaviors is open-ended and where the semantics of an action may depend on temporal structure rather than on a single frame. For this reason, this line of work does not treat novelty detection as an auxiliary post-processing step. Instead, it treats open-set awareness as part of the few-shot recognition problem itself.

The main idea developed across Chapters 5 and 6 is that support-query mismatch can be exploited explicitly for open-set rejection. Our first study explores this mechanism in a simplified one-shot skeleton setting, where 3D skeletons make the temporal structure explicit and reduce some of the complexity introduced by appearance variation (Berti et al., 2022). Our second study then tests whether the residual-based principle remains effective on the RGB video domain, stronger few-shot backbones, and a broader benchmark suite (Berti et al., 2026). Before discussing those models, it is useful to formalize the problem carefully and locate it within the main few-shot learning paradigms.

2.2 Background

This section first formalizes the recognition problem, then reviews the main few-shot learning paradigms, and finally summarizes the few-shot backbones on which the research chapters build. The overview of the few-shot learning paradigms below is deliberately kept concise and is meant to provide the minimal pointers needed to position our contributions; a more detailed and self-contained treatment of the related work, including stronger transfer-learning baselines and video-specific few-shot action recognition methods, is given in Chapters 5 and 6.

2.2.1 Problem Formulation

In all settings considered in this chapter, an action instance is a temporal observation

$$x = (o_1, \dots, o_T) \in \mathcal{X},$$

where each observation o_t may be an RGB frame, a depth frame, or a 3D skeleton representation. In the research chapters of this thesis, the input is either a sequence of 3D skeletons (Chapter 5) or an RGB video clip (Chapter 6). To keep the notation consistent throughout this chapter, θ always denotes the set of learnable parameters of the model, and a subscript θ marks a function that is learned from data. In particular, h_θ is the closed-set classifier, f_θ the few-shot (support-conditioned) classifier, ϕ_θ the embedding (feature-extraction) function, a_θ the binary acceptance function used for open-set rejection, and g_θ the resulting open-set predictor; $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ the label space, with superscripts specifying the relevant subset (e.g. $\mathcal{Y}^{\text{train}}$ for training classes and $\mathcal{Y}^S$ for the classes present in the current support set S). Figure 2.1 provides a visual comparison of the three recognition settings most directly relevant to this chapter: few-shot learning, open-set recognition, and their combination in few-shot open-set recognition. Standard closed-set classification remains the conceptual baseline discussed below and it is not represented in the figure.

Closed-Set Classification

In standard supervised classification, we assume a labeled dataset

$$D^{\text{train}} = \{(x_i, y_i)\}_{i=1}^n, \quad x_i \in \mathcal{X}, \quad y_i \in \mathcal{Y}^{\text{train}},$$

and we learn a classifier

$$h_\theta : \mathcal{X} \rightarrow \mathcal{Y}^{\text{train}}.$$

At test time, every query x_q is assumed to belong to one of the classes seen during training, so the prediction problem is

$$\hat{y}_q = h_\theta(x_q), \quad y_q \in \mathcal{Y}^{\text{train}}.$$

This is the classical closed-world assumption used by most large-scale action recognition benchmarks. It works well when the label space is fixed and well annotated, but it is mismatched with HRI scenarios in which new or user-specific actions may appear after deployment.

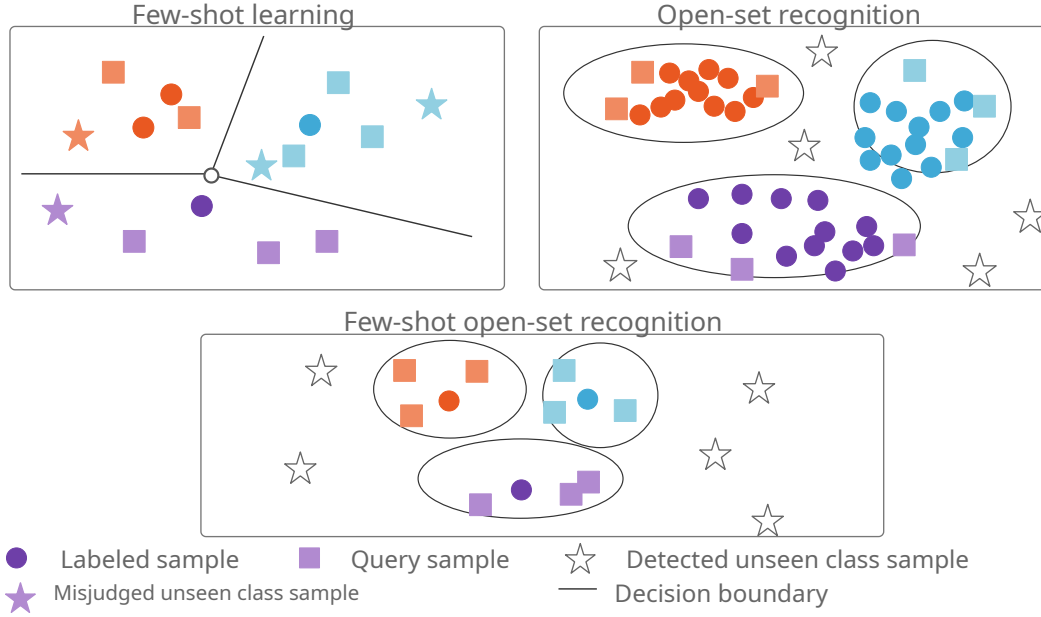


Figure 2.1 Comparison of the three recognition settings most directly relevant to this chapter. Few-shot open-set action recognition combines episodic support-based classification with explicit rejection of queries that do not belong to the current support label set.

Few-Shot Classification

Few-shot learning relaxes the assumption that every test class is represented by many labeled training samples. Instead, at test time the model receives a small *support set*

$$S = \bigcup_{c \in \mathcal{Y}^S} S_c, \quad S_c = \{(x_i^c, c)\}_{i=1}^N, \quad |\mathcal{Y}^S| = K,$$

where K is the number of classes in the episode and N is the number of labeled examples, or *shots*, per class. A query x_q must then be assigned to one of the support labels:

$$\hat{y}_q = f_\theta(x_q, S), \quad y_q \in \mathcal{Y}^S.$$

Many few-shot methods operate by mapping support and query instances to a common embedding space. In the prototype-based case, a class representative is computed as

$$p_c = \frac{1}{|S_c|} \sum_{(x,y) \in S_c} \phi_\theta(x),$$

and the query is classified by nearest prototype,

$$\hat{y}_q = \arg \min_{c \in \mathcal{Y}^S} d(\phi_\theta(x_q), p_c),$$

where ϕ_θ is a learned feature extractor and $d(\cdot, \cdot)$ is a distance function. We note that in action recognition, ϕ_θ must encode both visual appearance and temporal structure.

Open-Set Classification

Open-set recognition removes a different assumption: not every test query is guaranteed to belong to the known label space. It is therefore useful to separate the problem into two coupled tasks. The first is the usual closed-set prediction over the known classes,

$$\hat{y}_q = h_\theta(x_q), \quad y_q \in \mathcal{Y}^{\text{train}}.$$

The second is a binary acceptance decision,

$$a_\theta(x_q) \in \{0, 1\},$$

where $a_\theta(x_q) = 1$ means that the query is accepted as belonging to the known label space and $a_\theta(x_q) = 0$ means that it is rejected as unknown. The final open-set predictor combines the two,

$$g_\theta(x_q) = \begin{cases} h_\theta(x_q), & \text{if } a_\theta(x_q) = 1, \\ \emptyset, & \text{if } a_\theta(x_q) = 0. \end{cases}$$

This formulation is intentionally independent of any particular open-set mechanism. In some models the acceptance decision is derived from class scores or distances, while in others it is produced by an auxiliary module or discriminator. Since those choices belong to the method design rather than to the task definition, they are discussed later in the chapter.

Few-Shot Open-Set Action Recognition

Few-shot open-set action recognition combines the two requirements above. The model is given an episodic support set S with K classes and N shots per class, but the query may either belong to one of those support classes or fall outside them. Let

$$\mathcal{Q}^{\text{known}} = \{x_q : y_q \in \mathcal{Y}^S\}, \quad \mathcal{Q}^{\text{unknown}} = \{x_q : y_q \in \mathcal{Y}^U\},$$

with $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. The desired predictor is

$$g_\theta(x_q, S) = \begin{cases} f_\theta(x_q, S), & \text{if } a_\theta(x_q, S) = 1, \\ \emptyset, & \text{if } a_\theta(x_q, S) = 0, \end{cases} \quad (2.1)$$

where, following the notation introduced above, f_θ produces the closed-set support label and a_θ decides whether the query is accepted or rejected as unknown. This makes rejection harder than in a fixed-label open-set classifier, because the boundary between known and unknown changes with the support set of each episode. The classifier must identify the correct support label when the query is known, and it must reject the query when the action is not represented in the current support set. In other words, novelty is defined relative to the episode, not to the global training distribution.

In image-based few-shot open-set recognition, a single frame may already provide enough evidence for classification or rejection (Jeong et al., 2021; Liu et al., 2020). In action recognition, however, the relevant evidence is often distributed across time. The difference between two actions may depend on motion direction, temporal ordering, or interaction dynamics. For this reason, few-shot open-set action recognition should be treated as a temporal recognition problem in its own right rather than as a straightforward extension of image-based FSOSR.

2.2.2 Few-Shot Learning Paradigms

Few-shot learning is commonly organized into three broad families: *memory-based* methods, *metric-based* methods, and *optimization-based* methods (Chen et al., 2021b). Memory-based methods store and retrieve task-specific information; metric-based methods compare support and query examples in a learned representation space; optimization-based methods learn how to adapt model parameters from a small support set. Many of these approaches are trained episodically in a *meta-learning* framework, but few-shot learning also includes strong transfer-based baselines that learn a representation on base classes and reuse it for novel classes (Chen et al., 2019; 2021b).

In this thesis, the most relevant family is the metric-based one, because the action-recognition backbones used later make predictions by comparing the query with the support set. Within the metric-based family, *prototype-based* methods form an important subclass: each class is summarized by a prototype or centroid, and queries are classified by their similarity or distance to these representatives. This subclass is especially relevant here because prototype-query residuals provide a natural starting point for open-set rejection. For

completeness, we also briefly review optimization-based methods, which offer a different view of adaptation but are less central to the models used in this thesis.

Metric-Based Methods

Metric-based methods learn a feature space or a comparison function such that classification can be performed directly from support-query similarity. Matching Networks compare a query to the support set through attention (Vinyals et al., 2016); Relation Networks learn an explicit nonlinear relation module over support-query pairs (Sung et al., 2018). In generic form, these methods compute a class score

$$s_c(x_q, S) = \mathcal{R}_\theta(\phi_\theta(x_q), \{\phi_\theta(x) : (x, y) \in S_c\}),$$

where $\mathcal{R}_\theta$ denotes a generic relation or scoring function. Its role is to take the query embedding together with the *set* of support embeddings belonging to class c and return a single scalar compatibility score for that class. Depending on the method, this may be done by comparing the query to each support example and then aggregating the results, or by first building a class-level summary from S_c and then comparing the query to that summary. The prediction is then

$$\hat{y}_q = \arg \max_{c \in \mathcal{Y}^S} s_c(x_q, S).$$

Their main strength is that adaptation to a new task happens through comparison rather than through retraining. This is attractive in video action recognition because, from one episode to the next, the action classes and support examples change, but the model still has to solve the same core problem: compare each query video with the available support examples and decide which class it matches best. TRX and STRM are both examples of this broader design philosophy in the action-recognition setting (Perrett et al., 2021; Thatipelli et al., 2022).

Prototype-Based Methods

Prototype-based methods are a particularly simple and influential subclass of metric-based methods. Prototypical Networks represent each class by the mean of its support embeddings and classify queries by distance to class prototypes (Snell et al., 2017). Meta-Baseline shows that nearest-centroid reasoning remains surprisingly competitive when paired with a strong

representation (Chen et al., 2021b). In this setting,

$$p_c = \frac{1}{|S_c|} \sum_{(x,y) \in S_c} \phi_\theta(x),$$

and

$$\hat{y}_q = \arg \min_{c \in \mathcal{Y}^S} d(\phi_\theta(x_q), p_c).$$

Prototype-based methods are especially relevant for this thesis because they expose a natural residual signal,

$$r_q^c = \phi_\theta(x_q) - p_c, \quad (2.2)$$

which can be inspected to determine not only *which* class is closest, but also *how* the query mismatches the closest class representation. This residual (Eq. 2.2) is the quantity that the Feature-Residual Discrimination mechanism inspects for open-set rejection. The open-set rejection mechanism developed in Chapters 5 and 6 builds on this observation.

Optimization-Based Methods

Optimization-based methods take a different view. Instead of relying primarily on support-query similarity, they learn an initialization or an update rule that can adapt the model quickly to a new task. Model-Agnostic Meta-Learning (MAML) is the canonical example (Finn et al., 2017). Given an episodic support set S_i , an inner-loop update produces task-specific parameters

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{S_i}(f_{\theta}),$$

which are then evaluated on the query set:

$$\hat{y}_q = f_{\theta'_i}(x_q).$$

This family is attractive when rapid task-specific adaptation of parameters is important. In video action recognition, however, it can be computationally heavy because spatio-temporal backbones are large and episode-level inner-loop optimization is costly. More importantly for this thesis, optimization-based methods do not expose support-query residuals as directly as metric- and prototype-based models do, making them a less natural starting point for residual-based open-set scoring. Table 2.1 summarizes the few-shot paradigms discussed above and highlights why prototype- and metric-based approaches are the most natural fit for the residual-based open-set methods developed in this thesis.

Table 2.1 Comparison of the few-shot paradigms most relevant to this thesis. Prototype-based methods are listed separately because they play a central role in the action-recognition line of the dissertation, even though they belong to the broader metric-based family.

Paradigm	Core mechanism	Test-time adaptation	Why it matters for FSOSAR
Metric-based	Learn embeddings or relation functions so that support-query similarity directly induces the prediction.	No gradient update is required; the query is compared to support examples or class summaries.	A natural fit for episodic action recognition, where temporal matching between support and query is more feasible than retraining a classifier.
Prototype-based	Represent each class with a prototype or centroid in feature space and classify by distance or cosine similarity.	No gradient update is required; the query is compared to class prototypes.	Particularly compatible with open-set rejection because query-prototype residuals can be used as explicit novelty signals.
Optimization-based	Learn an initialization or update rule so that a few gradient steps on the support set adapt the model to a new task.	Yes; each episode includes an inner-loop parameter update.	Powerful in principle, but heavier for large video backbones and less suited for rejection mechanisms.

2.2.3 Adopted Few-Shot Models

In this thesis, we start from closed-set few-shot action recognition models and extend them to deal with open-set conditions. TRX (Perrett et al., 2021) provides the relational, metric-based backbone model used in our skeleton-based study (Berti et al., 2022) to establish the feasibility of residual-based rejection. SAFSAR is then adopted as a strong semantic-aware RGB backbone (Tang et al., 2024) to test the same principle under broader benchmark conditions in Chapter 6 (Berti et al., 2026). In the following, we review these models and identify the open-set limitations that motivate the extensions proposed in this thesis.

Temporal-Relational CrossTransformers (TRX)

Temporal-Relational CrossTransformers (TRX) is a metric-based few-shot action recognition model designed to compare support and query videos through temporal relations rather than through globally pooled features (Perrett et al., 2021). Instead of representing a whole video with a single embedding, TRX breaks the video into small ordered tuples of frames and compares support and query videos at that level. For each temporal tuple $\pi \in \Pi$, the model computes a query representation z_q^π and a class-specific reference representation $\tilde{p}_c^\pi$ built from the support examples of class c through cross-attention. The final class score is then obtained by averaging the tuple-level distances:

$$s_c(x_q, S) = -\frac{1}{|\Pi|} \sum_{\pi \in \Pi} d(z_q^\pi, \tilde{p}_c^\pi).$$

Intuitively, the model asks the following question for many short temporal snippets: does this part of the query video look like the corresponding temporal relation observed in the support examples of class c ? A class receives a higher score when this agreement is consistently strong across the temporal tuples. The key intuition is therefore that a query action should match a support class not only at the level of appearance, but at the level of temporal relations between sub-events.

TRX is particularly relevant to Chapter 5, where it is adapted to 3D skeleton sequences by replacing the RGB frame encoder with a skeleton embedding network (Berti et al., 2022). The open-set extension is then introduced by designing a discriminator on the residual between the query and the selected class representation. The resulting model remains aligned with the TRX philosophy: the support set is not merely a label dictionary, but a structured reference against which the query is compared.

Semantic-Aware Few-Shot Action Recognition (SAFSAR)

Semantic-Aware Few-Shot Action Recognition (SAFSAR) is a more recent RGB-video model that combines a 3D spatio-temporal backbone with textual semantics (Tang et al., 2024). Its motivation is different from TRX: instead of relying on explicit tuple-level relation modeling, SAFSAR argues that strong 3D video features and an effective text-video fusion mechanism can already provide highly discriminative few-shot representations. Concretely, SAFSAR uses a 3D video encoder based on VideoMAE (Tong et al., 2022) and injects textual information so that the video representation becomes more semantically consistent with the action label space. The downstream classification can then rely on cosine similarity, leveraging the discriminative structure of the learned video-text representation.

SAFSAR computes a video representation $z_q = E_v(x_q)$, a textual semantic representation $e_c = E_t(t_c)$ for class description t_c , and a fused feature $\tilde{z}_q^c = \Gamma(z_q, e_c)$ used to score each class. We refer to the original publication for additional details (Tang et al., 2024). SAFSAR was adopted because, at the time of writing the contribution reported in Chapter 6, it represented a strong RGB few-shot action recognition baseline, especially in low-shot settings. This made it a suitable backbone for evaluating whether residual discrimination could improve open-set reliability without weakening closed-set recognition.

Table 2.2 TRX and SAFSAR as representative closed-set few-shot action recognizers used in this thesis.

Model	Input representation	Main few-shot mechanism	Role in this thesis
TRX	Embeddings extracted from RGB frame tuples; adapted in Chapter 5 to 3D skeleton sequences	Metric-based relational matching with query-conditioned class representations built from temporal tuples	Provides an explicit support-query comparison structure and naturally exposes residuals between the query and the closest class representation.
SAFSAR	RGB video with a 3D video backbone and textual semantics	Semantic-aware video representation with simple cosine-based scoring over few-shot classes	Serves as a strong RGB baseline for testing whether residual discrimination can improve open-set rejection without sacrificing closed-set few-shot accuracy.

2.3 Methods

The main methodological contribution of this line of work is the introduction of *few-shot open-set action recognition* as a task in which a model must recognize novel known actions from only a few examples while also rejecting queries that belong to unseen classes. This setting exposes a limitation of standard few-shot action recognizers: they always assign each query to one of the support classes, even when none of them is appropriate.

Feature-Residual Discrimination (FR-Disc) is then proposed to address this limitation. Once a few-shot backbone has selected the most likely support class, the prediction is not accepted only because it is the best among the available options. Instead, FR-Disc inspects the *residual* between the query representation and the representation associated with the selected class. If that residual is compatible with the variability of known actions, the prediction is accepted; otherwise, it is rejected as unknown. Chapter 5 first validates this principle in a one-shot skeleton setting, while Chapter 6 evaluates whether the principle remains effective on RGB videos across stronger backbones and multiple datasets.

2.3.1 Feature-Residual Discriminator

The method can be described in a backbone-agnostic form before considering the specific TRX and SAFSAR instantiations used in the two research papers. Consider an episodic few-shot recognizer that, given a query x_q and support set S , produces class compatibility scores

$$s_c(x_q, S), \quad c \in \mathcal{Y}^S.$$

Higher scores correspond to better matches (for distance-based models this can be written by taking the negative distance). The closed-set predicted class is then

$$\hat{c} = \arg \max_{c \in \mathcal{Y}^S} s_c(x_q, S),$$

where $\hat{c}$ denotes the support class selected by the few-shot backbone. FR-Disc augments this prediction with a residual branch. If z_q denotes the query representation used by the backbone and $p_{\hat{c}}$ denotes the representation associated with the selected class, the residual is

$$r_q = z_q - p_{\hat{c}}.$$

This choice is intentionally simple. In metric- and prototype-based backbones, the query and class representations already live in a common embedding space, so their difference provides

a direct description of the mismatch between the two. Using the residual vector, rather than reducing the comparison immediately to a single Euclidean or cosine score, preserves both magnitude and direction of that mismatch and lets the discriminator learn which components are most informative for accepting or rejecting the prediction. This residual is then processed by a lightweight discriminator

$$u_q = \text{Disc}_\omega(r_q) \in [0, 1],$$

which acts as a knownness score. The final decision becomes

$$g(x_q, S) = \begin{cases} \hat{c}, & \text{if } u_q > \tau, \\ \emptyset, & \text{otherwise,} \end{cases}$$

with threshold τ .

At training time, the discriminator is optimized jointly with the few-shot backbone. The key design choice, introduced in our first study (Berti et al., 2022) and retained in the extended evaluation (Berti et al., 2026), is that the residual branch is trained on *hard comparisons*: positive samples are known queries that the few-shot model already classifies correctly, while negative samples are unknown queries compared against their closest known class. The general loss formulation can therefore be written as

$$\mathcal{L}_{\text{Disc}} = \mathbb{1}[\hat{c} = y_q, y_q \in \mathcal{Y}^S] \cdot \text{BCE}(u_q, 1) + \mathbb{1}[y_q \in \mathcal{Y}^U] \cdot \text{BCE}(u_q, 0),$$

and is combined with the closed-set few-shot loss:

$$\mathcal{L} = \mathcal{L}_{\text{FS}} + \alpha \mathcal{L}_{\text{Disc}}.$$

Here $\mathcal{L}_{\text{FS}}$ denotes the standard closed-set few-shot loss of the underlying backbone. Therefore, the discriminator does not operate on logits alone; it is trained on the residual pattern of a *correct known match* and on the residual pattern produced by an *unknown query forced toward its nearest known class*.

2.3.2 Open-Set Extension of TRX with FR-Disc

In the study reproduced in Chapter 5 (Berti et al., 2022), the few-shot backbone is a TRX-style relational matcher (Perrett et al., 2021) operating on temporal tuples in the skeleton-based setting of that work. In the original TRX formulation, Q_p denotes the query

representation associated with tuple $p \in \Pi$, $Q = \{Q_p\}_{p \in \Pi}$ is the resulting collection of query tuples, and S^c is the support representation of class c . The model then applies a learned linear projection Λ to each query tuple representation and compares it with a query-conditioned class representation t_p^c built from the support examples of class c . In other words, for each tuple p , TRX asks how close the projected query tuple ΛQ_p is to the class-specific reference tuple t_p^c . The overall class distance is obtained by averaging these tuple-level discrepancies:

$$T(Q, S^c) = \frac{1}{|\Pi|} \sum_{p \in \Pi} \|t_p^c - \Lambda Q_p\|_2.$$

The Euclidean norm is used here to remain consistent with the TRX matching rule: the closed-set classifier already ranks classes by the accumulated distance between transformed query tuples and query-conditioned class representations. Other distances or learned comparison functions could be used in different backbones, but changing the metric in this instantiation would also change the underlying few-shot recognizer being extended. The few-shot prediction is therefore

$$\hat{c}_{\text{TRX}} = \arg \min_{c \in \mathcal{Y}^S} T(Q, S^c).$$

To attach FR-Disc, the tuple-wise query representations and class representations are stacked:

$$\tilde{q} = \text{stack}(\{\Lambda Q_p : p \in \Pi\}), \quad t^c = \text{stack}(\{t_p^c : p \in \Pi\}),$$

and the residual passed to the discriminator is

$$r_q^{\text{TRX}} = \tilde{q} - t^{\hat{c}_{\text{TRX}}}.$$

The resulting knownness score is

$$u_q^{\text{TRX}} = \text{Disc}_{\omega}^{\text{TRX}}(r_q^{\text{TRX}}).$$

We care to note that in TRX the residual is not a single global vector, but a structured tensor built from temporal relations. The discriminator therefore sees not only whether the query is close to the selected class, but *how* that closeness is distributed across relational fragments of the sequence. This is precisely what makes the approach suitable for distinguishing semantically similar actions that differ mainly in temporal evolution.

2.3.3 Open-Set Extension of SAFSAR with FR-Disc

In Chapter 6, the same principle is retained but instantiated on a very different backbone. SAFSAR computes a compact semantic-aware video representation with a 3D video encoder and textual guidance (Tang et al., 2024). The closed-set prediction can be described as

$$\hat{c}_{\text{SAFSAR}} = \arg \max_{c \in \mathcal{Y}^S} s_c^{\text{SAFSAR}}(x_q, S),$$

where the score is derived from cosine-style comparison in the backbone feature space.

Let z_q^{SAFSAR} denote the query feature produced by the SAFSAR backbone, and let p_c^{SAFSAR} denote the support-derived representation of class c . The residual branch can then be written as

$$r_q^{\text{SAFSAR}} = z_q^{\text{SAFSAR}} - p_{\hat{c}_{\text{SAFSAR}}}^{\text{SAFSAR}}.$$

This follows the same construction used in our extended RGB study (Berti et al., 2026), where the residual is computed from the query feature and the support representation selected by the closed-set arg max. The discriminator then produces

$$u_q^{\text{SAFSAR}} = \text{Disc}_{\omega}^{\text{SAFSAR}}(r_q^{\text{SAFSAR}}).$$

Compared with the TRX case, the residual here is more compact and lives in a high-level video-semantic embedding space rather than in a stack of tuple-wise relational features. The conceptual role, however, is unchanged: the residual branch asks whether the selected class is merely the *least wrong* option in the support set, or whether the query is actually compatible with the geometry of known classes in feature space.

Although TRX and SAFSAR are different closed-set models, FR-Disc can be inserted into both because it relies on the same underlying idea: if few-shot classification compares a query with support-derived class representations, then the mismatch between the query and the selected class is itself informative. Chapter 5 uses TRX to isolate this mechanism in a simplified one-shot setting, while Chapter 6 tests whether it remains effective with RGB video, stronger backbones, and multiple datasets.

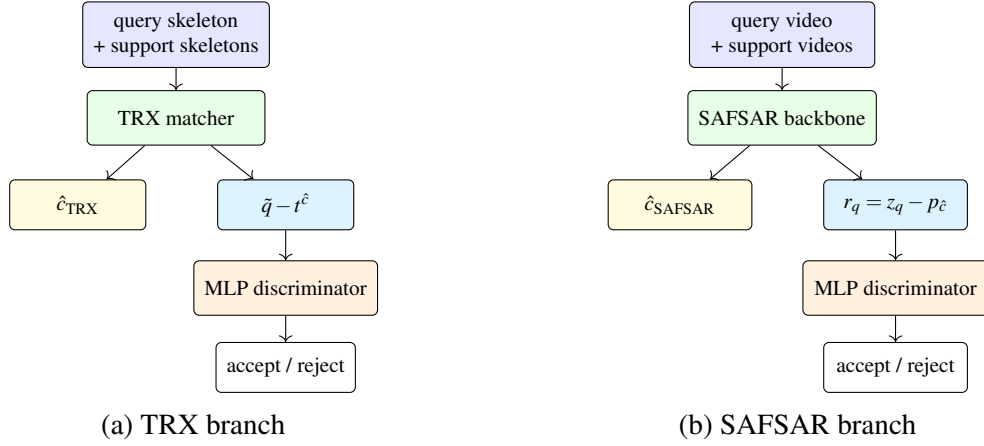


Figure 2.2 Backbone-specific insertion of Feature-Residual Discrimination. In both cases the closed-set few-shot model first selects the best known class, then a discriminator inspects the residual between the query representation and the representation associated with that class. The main difference is the feature space in which the residual is formed: a structured temporal-relational tensor for TRX and a compact semantic-aware video embedding for SAFSAR.

2.4 Experiments

The two research chapters later reproduced in Part II study the same idea under different experimental conditions. The first study validates residual-based rejection in a simplified one-shot skeleton setting (Berti et al., 2022); the second extends the analysis to RGB video, stronger backbones, and multiple datasets (Berti et al., 2026). The first setting isolates the core mechanism, while the second evaluates whether the same principle remains effective under broader and more realistic conditions.

2.4.1 First Evaluation on Skeleton Input

The first study, later reproduced as Chapter 5 (Berti et al., 2022), focuses on a robotics-oriented scenario. The input modality is a sequence of 3D skeletons, the few-shot regime is one-shot, and the evaluation asks whether a TRX-style recognizer extended with a residual discriminator can both classify known actions and reject unknown ones (Berti et al., 2022). The experiments are conducted on NTU RGB+D 120 (Liu et al., 2019) after restricting the setting to single-person actions and re-extracting skeletons. This makes the input lower-dimensional and more structured than raw RGB video, so the study can focus more directly on pose and temporal evolution, while abstracting away appearance cues such as texture, color, clothing, and background.

The episodic evaluation is designed to make the open-set component explicit. In each episode, K support classes are selected as known classes, while the remaining classes are treated as unknown. The main quantitative metric is FSOS-ACC, which summarizes in a single score whether known queries are classified correctly and unknown queries are rejected correctly. In addition, the study includes qualitative case analyses to inspect where the discriminator succeeds or fails when actions are semantically close.

The main experimental question of the first study (Berti et al., 2022) is whether residual-based rejection can already be useful in a first evaluation based on one-shot skeleton input. The answer is positive: across the tested episodic configurations, the proposed model improves over the threshold-based baseline in the joint handling of known and unknown queries. At the same time, the scope of the evaluation remains intentionally limited. The input is restricted to skeletons, the setting is one-shot, and the qualitative analysis highlights cases where full visual appearance would likely be necessary. For the detailed quantitative comparison, the reader is referred to Chapter 5 (Berti et al., 2022). At thesis level, the main point is that this study provides an initial validation of the idea and motivates its later extension to the RGB video domain.

2.4.2 Extended RGB Evaluation

The second study, later reproduced as Chapter 6, expands the evaluation along three axes: modality, backbone strength, and experimental scale (Berti et al., 2026). Rather than a single skeleton benchmark, it considers RGB video and a benchmark suite spanning five datasets: Diving48 (Li et al., 2018), SSv2 (Goyal et al., 2017; Materzynska et al., 2020), NTU RGB+D (Liu et al., 2019; Shahroudy et al., 2016), HMDB51 (Kuehne et al., 2011), and UCF101 (Soomro et al., 2012). The evaluation covers two backbones, SAFSAR and STRM, and compares FR-Disc with several representative open-set strategies, including score-based baselines (Dhamija et al., 2018), Entropic Open-Set (Dhamija et al., 2018), and Garbage Class (Jeong et al., 2021; Liu et al., 2020).

The main question of the study is whether the effectiveness of the residual-based solution holds once the evaluation becomes broader, the representations richer, and the baselines stronger. The answer is positive. The evaluation covers both 5-way 1-shot and 5-way 5-shot protocols and reports closed-set, open-set, and joint metrics: FS ACC and OS ACC for accuracy on known and unknown handling, AUROC and AUPR for threshold-independent open-set detection, and OSCR for the combined trade-off between correct classification and unknown rejection. In this context, *explicit open-set strategies* refers to methods that add

a dedicated mechanism for reject decisions, such as Entropic Open-Set, Garbage Class, or FR-Disc, rather than relying only on the scores produced by the original closed-set classifier. The experiments show, first, that strong closed-set few-shot models are already competitive open-set baselines; second, that explicit open-set strategies can behave quite differently across datasets and backbones; and third, that FR-Disc is the most consistent of the evaluated methods (Berti et al., 2026). The detailed numerical comparisons are given in Chapter 6.

2.5 Summary

The chapters later collected in Part II develop the FSOSAR work across two complementary studies: an initial validation in Chapter 5 (Berti et al., 2022) and a broader evaluation in Chapter 6 (Berti et al., 2026). Table 2.3 summarizes the two studies and highlights their main experimental differences.

Table 2.3 Summary of the two experimental studies that support the few-shot open-set action recognition work.

Study	Input	Backbone	Datasets	Protocol	Metrics	Main takeaway
Humanoids 2022	3D skeletons	TRX + residual discriminator	NTU RGB+D 120 single-person subset	One-shot episodes with $K \in \{5, 10, 15\}$ known classes and the remaining classes treated as unknown	FSOS-ACC + qualitative analysis	Validates residual-based rejection in a robotics-oriented temporal setting, while also showing the limitations of skeleton-only perception.
ICPR 2026	RGB video	SAFSAR and STRM + FR-Disc	Diving48, SSv2, NTU RGB+D, HMDB51, UCF101	5-way 1-shot and 5-shot episodes with comparison against multiple open-set techniques	FS ACC, OS ACC, AUROC, AUPR, OSCR	Shows that the same residual-based principle remains effective under broader and stronger evaluation conditions, and that FR-Disc is the most consistent explicit open-set method among the compared baselines.

At the core of both studies lies the same methodological idea: actions are represented through support-query comparisons, and the mismatch between the query and the selected class representation is used to decide whether the prediction should be accepted. The first study serves as an initial validation, while the second evaluates the same idea under broader experimental conditions.

Even so, few-shot open-set action recognition still operates within a closed (episodic) label space. Once the model can reject unknown actions, the next question is whether it can also reason beyond a fixed vocabulary of predefined labels. This motivates the next part of the thesis, which turns from novelty-aware recognition to semantic generalization with vision-language models.

Chapter 3

Vision-Language Models for Semantic Generalization and Personalization

3.1 Introduction

This chapter studies how a computer vision system can generalize beyond a fixed vocabulary of labels. The corresponding work is documented in the chapters reproduced later as Chapter 7 and Chapter 8. Chapter 7 presents *DualVCLIP* (Calabrese et al., 2024), a prompt-learning method for zero-shot multi-label action recognition built by combining OpenVCLIP (Weng et al., 2023) and DualCoOp (Sun et al., 2022), and evaluated in the paper on verb–object action labels. Chapter 8 presents *ConCon-Chi* (Rosasco et al., 2024), a benchmark for personalized vision-language retrieval and generation from a few example images. Together, these two studies examine how language-aligned visual representations can support semantic generalization when the target meanings are only partially available during training.

Both contributions assume that a large pretrained vision-language model already contains useful semantics that should be preserved. In *DualVCLIP*, adaptation takes the form of learned prompt contexts concatenated to class labels: following *DualCoOp*, the model learns a positive and a negative domain-specific prompt, and applies them to the OpenVCLIP video backbone so as to support multi-label zero-shot action recognition. In personalization, the same prompt-learning approach is used differently: rather than adapting the context around an existing label, the model learns the embedding of a new textual token that must represent a novel visual concept and remain compositional with the contexts used at retrieval or generation time.

Although the two problems are different, they share the same high-level strategy: keep the pretrained vision-language backbone fixed and adapt only a small number of language-side

parameters. This choice is important because it preserves the broad semantics acquired during pretraining, limits catastrophic forgetting, and keeps the learned representation compositional.

Vision-language models provide a natural starting point for this question. A central example is CLIP (Radford et al., 2021) (Contrastive Language-Image Pretraining), which learns aligned visual and textual representations from large-scale image-text pairs. In CLIP-style models, an image or a video observation x is encoded by a visual encoder E_v , a text prompt t is encoded by a text encoder E_t , and the similarity between the two is measured in a shared embedding space:

$$z_x = E_v(x), \quad z_t = E_t(t), \quad s(x, t) = \cos(z_x, z_t).$$

Given a set of textual class descriptions $\mathcal{T}$, zero-shot prediction can then be written as

$$\hat{t} = \arg \max_{t \in \mathcal{T}} s(x, t).$$

This simple mechanism is powerful because it turns language into an interface for semantic transfer. Instead of learning a dedicated classifier for every downstream label set, the model can reuse the large-scale alignment learned during pretraining and compare an observation directly with textual descriptions.

For robotics, this is attractive for at least two reasons. First, the semantic categories that matter in Human-Robot Interaction are often open-ended. A robot may need to understand a new verb-object action, a task-specific verbal description, or a user-specific concept that was not part of its original supervised taxonomy. Here verb-object actions are actions described by the combination of a verb and an object, such as *holding a box*, *opening a box*, or *passing a box*. Second, language often exposes internal structure that a flat label space hides. These actions share object semantics but differ in temporal evolution; personalized concepts may need to be recognized or generated across many known contexts rather than treated as isolated labels.

At the same time, transferring CLIP-like models to these settings is not immediate. The first difficulty is temporal and compositional. CLIP was originally designed for image-text alignment, whereas action recognition requires reasoning over time and often over multiple simultaneously valid labels. This is especially visible in verb-object actions, where the object often constrains the semantic field of the action while the temporal pattern determines which verb is actually being performed. The second difficulty is semantic novelty in a stronger sense: even an open-vocabulary model may fail when the concept itself is uncommon, user-specific, or absent from the language statistics of pretraining. This chapter focuses

on these two issues: adapting image-language alignment to video action recognition, and evaluating what happens when the model must learn genuinely new meanings from only a few example images. As in the rest of Part I, the discussion of CLIP and of zero-shot vision-language methods here is intended only to provide the minimal pointers required to follow our contributions; a broader account of the related work, including CLIP adaptations to video, prompt tuning, compositional zero-shot learning, and personalized vision-language learning, is given together with each contribution in Chapters 7 and 8.

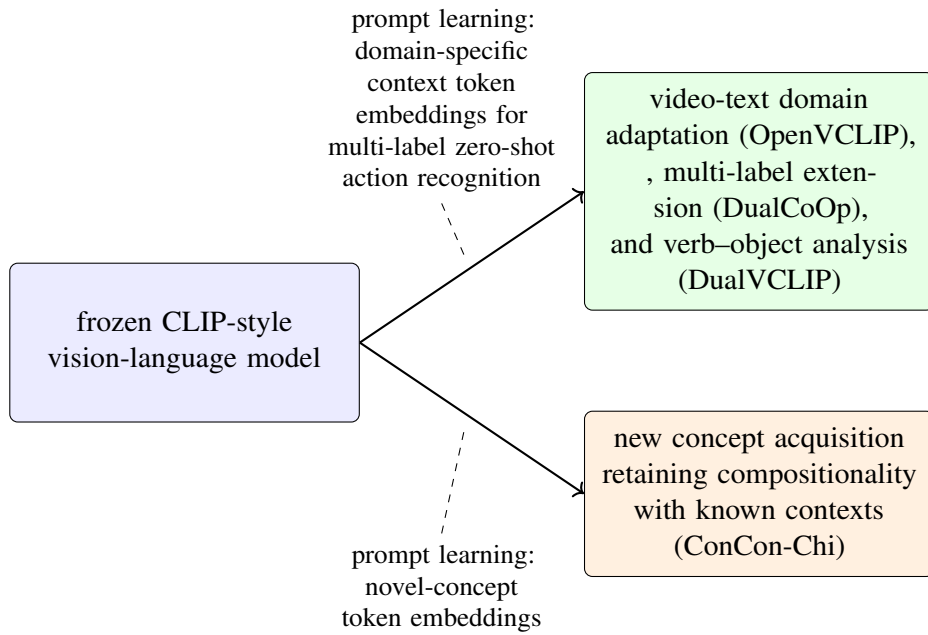


Figure 3.1 Two research directions built on the same frozen CLIP-style foundation. DualVCLIP adapts CLIP to video, imports the positive/negative prompt-learning mechanism of DualCoOp, and studies zero-shot multi-label action recognition under verb-object compositional structure. ConCon-Chi instead focuses on learning a novel concept token from a few examples and on evaluating whether that concept remains compositional with known textual contexts.

The two studies therefore share the same broad adaptation principle, but apply it to different problems. DualVCLIP learns prompt contexts for known labels in order to adapt a CLIP-based video recognizer to the multi-label zero-shot setting. ConCon-Chi addresses a different question: how to evaluate methods that learn the embedding of a new concept token from a few examples, and then use it compositionally in retrieval or generation. The rest of the chapter presents these two studies in turn, so that each paper can be followed as a coherent unit rather than through alternating background, method, and experiment sections.

3.2 DualVCLIP: Zero-Shot Multi-Label Action Recognition

This section presents our DualVCLIP study, from its technical starting points to the proposed method and its empirical validation. The main contribution is DualVCLIP (Calabrese et al., 2024), a CLIP-based video model for zero-shot multi-label action recognition obtained by combining OpenVCLIP (Weng et al., 2023) with DualCoOp (Sun et al., 2022).

3.2.1 Background

CLIP makes it possible to compare visual observations directly against textual class descriptions, but it was originally designed for image-text alignment rather than for video. OpenVCLIP (Weng et al., 2023) proposes an adaptation of CLIP to videos. Since naive frame averaging tends to discard the temporal relations that distinguish many actions, OpenVCLIP extends the pretrained visual encoder with temporal attention while trying to preserve the open-vocabulary properties inherited from CLIP.

Let $x = (f_1, \dots, f_T)$ denote a video clip and let q_t , K_t , and V_t be the query, key, and value tensors associated with frame t inside a transformer block. Standard self-attention processes each frame independently, whereas OpenVCLIP expands the attention neighborhood so that the representation of frame t can interact with nearby frames:

$$y_t = \text{Softmax} \left(\frac{q_t K_{\mathcal{N}(t)}^\top}{\sqrt{d}} \right) V_{\mathcal{N}(t)},$$

where $\mathcal{N}(t)$ denotes a temporal neighborhood around frame t . By stacking such layers, the model accumulates temporal context without abandoning the original CLIP embedding space.

This design offers a clean way to bring CLIP into video recognition while keeping the text encoder intact. The model can still compare videos to textual prompts in a zero-shot fashion, but it does so after exposing the visual representation to temporal structure. For action labels that share the same object, this is essential: the object often determines the semantic field of the action, while the temporal pattern determines which alternative action involving that object is actually taking place. In this sense, OpenVCLIP provides the temporally aware visual backbone on top of which DualVCLIP is built.

Once a CLIP-based backbone is available, the next issue is not only recognizing classes from textual descriptions, but doing so when labels are not mutually exclusive. In many

realistic settings, several labels may be simultaneously valid, so a single global prediction is no longer sufficient. DualCoOp (Sun et al., 2022) addresses this case by proposing a prompt-learning mechanism for zero-shot classification with non-exclusive labels. This is also the point where the connection with the previous chapter becomes clearer: when many linguistically valid labels are candidates but only some are actually present, the model must become more robust to absent labels rather than forcing every query into a single choice.

Instead of using a single textual template for each class name y_c , DualCoOp learns a positive prompt context P^+ and a negative prompt context P^- . These prompt contexts are shared across classes, but are learned for the target domain rather than taken from a hand-written template. For each class, the corresponding textual embeddings are

$$t_c^+ = E_t([P^+, y_c]), \quad t_c^- = E_t([P^-, y_c]).$$

Given a visual representation z_x , the class is treated as present when the image is more compatible with the positive than with the negative prompt. In its simplest form, this can be written as

$$\ell_c(x) = s(x, t_c^+) - s(x, t_c^-),$$

with independent decisions for different classes. The appeal of this formulation is that the model preserves the pretrained visual-textual alignment while adapting only a small number of prompt parameters. It can also be viewed as increasing robustness to absent candidate classes, because the model is encouraged to represent not only evidence in favor of a label but also evidence against it. In an open-vocabulary setting, where many linguistically valid labels may simply not be present in the observation, this positive/negative comparison is especially useful.

DualCoOp also makes an architectural observation that matters here. In multi-label recognition, a single pooled visual embedding may mix evidence for several classes. For this reason, the method promotes class-specific matching before the final aggregation, so that the model can preserve which visual evidence is relevant for each label. This is the second direct ingredient of DualVCLIP. The method combines OpenVCLIP, which provides temporally aware video features, with the positive/negative prompt mechanism of DualCoOp, which provides a natural way to handle non-mutually-exclusive labels.

3.2.2 Method

DualVCLIP (Calabrese et al., 2024) addresses zero-shot multi-label action recognition in video by combining OpenVCLIP with DualCoOp. The paper evaluates the method on action labels that are often written as verb–object descriptions and may co-occur within the same clip. The goal is therefore twofold: preserve the open-vocabulary transfer properties of CLIP, and improve discriminability in a multi-label temporal setting.

Let $x = (f_1, \dots, f_T)$ be a video clip and let $\mathcal{Y} = \{y_c\}_{c=1}^C$ be the set of action labels. The OpenVCLIP visual encoder produces a sequence of temporally contextualized frame embeddings

$$V = \{v_t\}_{t=1}^T, \quad v_t \in \mathbb{R}^d.$$

For each class y_c , DualVCLIP constructs positive and negative prompts

$$\hat{y}_c^+ = [P^+, y_c], \quad \hat{y}_c^- = [P^-, y_c],$$

and encodes them with the CLIP text encoder:

$$t_c^+ = E_t(\hat{y}_c^+), \quad t_c^- = E_t(\hat{y}_c^-).$$

Similarity is then computed at the frame level,

$$s_{c,t}^+ = \cos(v_t, t_c^+), \quad s_{c,t}^- = \cos(v_t, t_c^-),$$

so that the model retains class-specific temporal evidence instead of collapsing the video to a single representation too early.

The key modification introduced by DualVCLIP is *class-specific temporal aggregation*. For each class, the positive similarities define temporal weights

$$\alpha_{c,t} = \frac{\exp(s_{c,t}^+)}{\sum_{t'=1}^T \exp(s_{c,t'}^+)},$$

which are then used to aggregate both the positive and negative branches. The softmax is applied across the frames of the video because the goal is to turn frame-level positive similarities into a normalized temporal weighting for class c : the model should decide not only *whether* a class is present, but also *where in time* the main evidence for that class is concentrated. This normalization also makes the aggregation comparable across videos of different length. The weights are computed only from the positive branch because they are

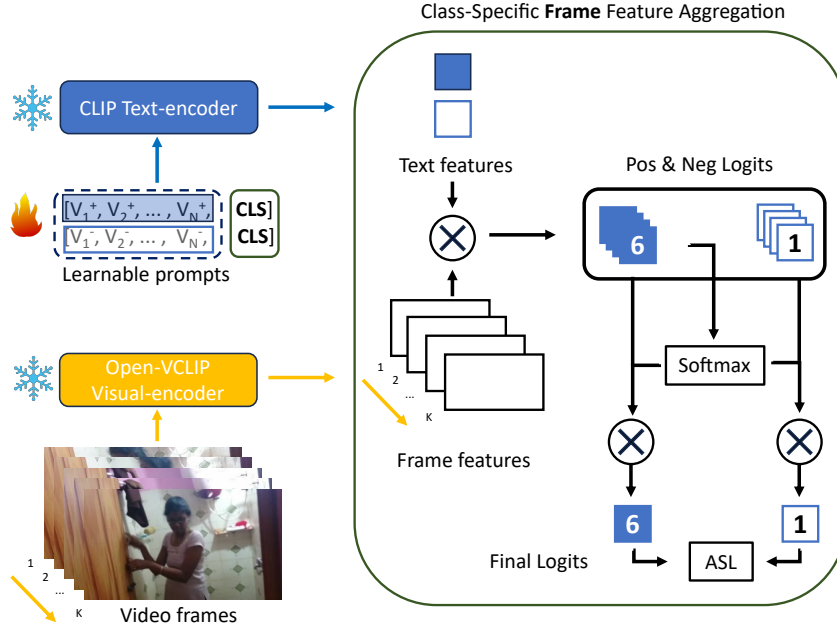


Figure 3.2 Overview of DualVCLIP. A video backbone adapted from CLIP provides temporally contextualized frame features, while two learned prompts generate positive and negative textual embeddings for each class. Temporal logits are then aggregated in a class-specific way. In the figure, the example values 6 and 1 denote the aggregated positive and negative logits, respectively, for the illustrated query class. During training, these outputs are used within the Asymmetric Loss (ASL), while only the prompt parameters are optimized, enabling multi-label zero-shot action recognition in video.

meant to identify the frames that best support the presence of class c . Once those frames are selected, the same temporal support is used to pool both $s_{c,t}^+$ and $s_{c,t}^-$, so the final decision compares positive and negative evidence on the very frames that are most indicative of the candidate action, instead of allowing the negative branch to redirect the attention toward frames that are irrelevant to that class:

$$S_c^+ = \sum_{t=1}^T \alpha_{c,t} s_{c,t}^+, \quad S_c^- = \sum_{t=1}^T \alpha_{c,t} s_{c,t}^-.$$

The final multi-label decision is made independently for each class:

$$\hat{m}_c = \mathbb{1}[S_c^+ > S_c^-].$$

During training, the learnable parameters are the prompt contexts P^+ and P^- , while the weights of the OpenVCLIP/CLIP backbone are kept frozen and are not updated.

Methodologically, DualVCLIP shows that the prompt-learning mechanism introduced by DualCoOp can be transferred to video by coupling it with an OpenVCLIP backbone, without abandoning the zero-shot properties of the underlying vision-language model.

3.2.3 Experimental Validation

The study evaluates DualVCLIP on Charades, a dataset of everyday home activities where each video may contain several simultaneously valid actions (Sigurdsson et al., 2016). This choice is motivated by the fact that many Charades labels can be decomposed into verb-object pairs, and many labels share the same object while differing in the verb. The paper therefore introduces class-level zero-shot splits that hide roughly half of the action labels during training. Since most Charades labels can be decomposed into verb-object pairs, the zero-shot classes can also be controlled according to which component is unseen. Besides a random split, the paper defines a *verb split*, where the unseen classes are chosen so that many test labels contain verbs absent during training while almost all objects remain familiar, and an *object split*, where the opposite happens and many unseen classes contain objects not observed during training. Concretely, the verb split leaves 18 of the 38 verbs unseen but only 2 of the 37 objects, whereas the object split leaves 18 objects unseen and 13 verbs unseen (Calabrese et al., 2024). This experimental setup is important because it makes the role of compositional semantics directly testable: instead of evaluating only random unseen classes, it probes whether the transferred semantics are more robust for object semantics or for the temporal/verb component of the action.

The quantitative evaluation uses mean Average Precision (mAP) in both the zero-shot setting and the generalized zero-shot setting. The method is compared with strong zero-shot baselines such as CLIP-Hitchhiker (Bain et al., 2022) and VideoCoCa (Yan et al., 2022), with an OpenVCLIP variant adapted to multi-label prediction, and with MSQNet (Mondal et al., 2023) where a homogeneous comparison is possible. The most informative comparison is against the OpenVCLIP adaptation, because it isolates the effect of prompt learning and class-specific temporal aggregation. On the full Charades test set, DualVCLIP reaches an mAP of 22.98, compared with 18.82 for the OpenVCLIP adaptation (Calabrese et al., 2024). On the zero-shot setting built from the proposed class-level random, verb, and object splits, it reaches an mAP of 16.65, showing that a lightweight prompt-based modification can retain nontrivial recognition ability even when the target actions are unseen during training.

Equally important is the diagnostic outcome of the split analysis. The experiments indicate that leaving objects unseen and leaving verbs unseen do not affect the model in

the same way. Performance benefits more from having seen the object during training than from having seen the verb, which suggests that the transferred representation is more reliable for object semantics than for the temporal component that distinguishes actions sharing the same object. The paper also shows that object augmentation during training helps reduce this imbalance, improving robustness when compositional generalization depends more strongly on the verb-object structure (Calabrese et al., 2024). Overall, these results clarify what kind of semantic transfer is currently easier to obtain with CLIP-style models on video: object-centered generalization is more stable, while verb-level temporal compositionality remains more difficult.

3.3 ConCon-Chi: Personalized Retrieval and Generation

This section presents our ConCon-Chi study as a second coherent block. The central contribution is ConCon-Chi (Rosasco et al., 2024), a dataset and benchmark designed to evaluate whether a personalized vision-language method learns a genuinely new concept and whether that concept remains compositional with known contexts.

3.3.1 Background

We have seen how, in DualVCLIP, prompt learning targets textual *context* tokens around known labels, so as to improve their discriminability in the target domain. Personalized retrieval and generation use prompt learning as well. Here the goal is to learn the embedding of a new textual token that represents a novel concept from only a handful of example images. This is the setting of personalized Text-to-Image Retrieval (TIR) (Cohen et al., 2022) and, in one important family of methods, personalized Text-to-Image Generation (TIG) (Gal et al., 2023a). In both cases, the central question is whether the learned concept token can later be combined correctly with known textual contexts.

In personalized TIR, the model receives a small concept set

$$I_c = \{i_1, \dots, i_k\},$$

containing example images of a target concept c , together with a textual context u . The task is to rank an image pool $\mathcal{P}$ and retrieve images that depict the learned concept in the requested context. The retrieval problem can be written as

$$\pi = \mathcal{R}(I_c, u, \mathcal{P}),$$

where $\mathcal{R}$ denotes a retrieval model that induces a ranking π over the pool $\mathcal{P}$ from the concept examples and the textual context. Representative methods in this family learn a token or pseudo-word embedding that can later be inserted into the text query, effectively expanding the vocabulary of the pretrained vision-language model (Cohen et al., 2022). Related work in zero-shot composed image retrieval, such as Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023), follows a similar intuition through learned mappings from visual examples to textual embeddings.

In personalized TIG, given the same few concept images I_c and a textual context u , the model must generate images that preserve the identity of the concept while satisfying the context:

$$\tilde{x} \sim p_{\psi}(\cdot \mid I_c, u).$$

Textual Inversion (Gal et al., 2023a) learns a new token embedding in the vocabulary of a frozen text-to-image model, while DreamBooth (Ruiz et al., 2023) instead fine-tunes the model around a rare textual identifier tied to the concept. These methods differ in optimization strategy, but they face the same evaluation problem: a method may appear to succeed even if it relies mostly on familiar category priors or memorized contexts rather than on a genuinely new concept representation.

This is the main limitation of much of the related validation setting adopted at the time, including studies on multi-concept personalization and compositional generation (Kumari et al., 2023a; b). The concepts are often instances of common categories, the contexts are relatively sparse, and the generation benchmarks usually do not provide real images for the target prompts. As a consequence, it is hard to determine whether a method has learned a genuinely new meaning or has instead benefited from concept priors or concept-context bias. This is precisely the gap addressed by ConCon-Chi.

3.3.2 Benchmark Design

ConCon-Chi is a dataset and benchmark that we proposed in (Rosasco et al., 2024). It addresses a simple question: once a vision-language model has learned a new concept from a few examples, how can one tell whether it has actually learned that concept, rather than overfitting a narrow context or falling back to a familiar category prior? The benchmark answers this question through an explicit organization by concepts, contexts, and their compositions.

The defining idea of ConCon-Chi is that concepts and contexts are controlled independently. A concept may be, for example, a personalized object or creature learned from a few

images; a context may specify where that concept appears, what it is doing, or with which other entities it is combined. Queries can involve one, two, or three concepts together with a textual context. This structure matters because it prevents the benchmark from collapsing into a collection of near-duplicate personalization examples.

This organization also makes two shortcut failures explicit. If each context appeared with only one concept, a retrieval or generation model could succeed mostly by recognizing the context. Conversely, if each concept appeared in only a few highly repetitive contexts, the model could memorize the concept without learning how to recombine it. ConCon-Chi is designed precisely to avoid these shortcuts by increasing both the number of contexts per concept and the number of concepts that can appear under related contexts. For this reason, it is useful to think of the benchmark as a concept-context matrix.

To build the dataset, we manually created the chimeric concepts and then acquired images by recreating the concept–context “grid” structure in a variety of everyday settings, including household, office, and outdoor environments. Overall, we collected more than 4000 images using an iPhone (Rosasco et al., 2024).

3.3.3 Dataset Acquisition

Chimeric concepts in ConCon-Chi are manually crafted physical objects, built by combining parts that evoke semantically unrelated categories into a single artifact. To populate the concept–context grid, we staged acquisitions across three environment types (house, office, and outdoor), and we recreated the intended context conditions systematically according to the grid structure. The resulting image pool contains more than 4000 photographs, acquired with an iPhone. (Rosasco et al., 2024)

ConCon-Chi also proposes a way to evaluate the semantic novelty of the concept to be learned, going beyond the validation strategy commonly used in earlier personalized retrieval and generation work. Rather than relying only on instances of common categories, it introduces *chimeric* concepts inspired by the Chimera setting for new-meaning learning (Lazaridou et al., 2017). These objects are crafted by combining elements of semantically unrelated known categories into a single concept. As a result, the benchmark evaluates whether a method can represent a concept that is not readily reducible to an already familiar category.

The dataset reinforces this effect through hard negatives. Concepts drawn from the same broad source categories as the chimera components are included among the common concepts, and multiple concepts can appear under similar or identical contexts. As a result, a

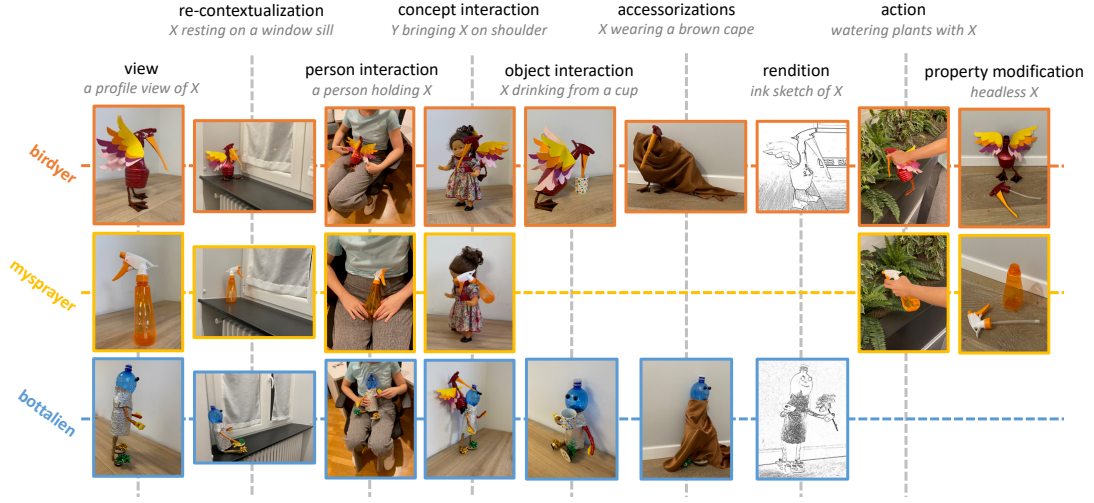


Figure 3.3 Concept-context structure of ConCon-Chi. Each row corresponds to a concept and each column to a context. This matrix-style organization is a defining property of the dataset and makes it possible to evaluate whether a personalized representation captures both the learned concept and its composition with known contexts.

method cannot obtain strong performance merely by recognizing that a prompt refers to *a bag-like thing* or *a toy-like creature*. It must distinguish the intended concept and preserve its relation to the requested context.

The same concept-context structure supports both personalized retrieval and personalized generation. In the retrieval benchmark, each query q is associated with a ground-truth image set

$$G(q) \subseteq \mathcal{P},$$

where $\mathcal{P}$ is the pool of candidate images. A method must rank the pool according to how well each image matches the learned concept and the requested context. In the generation benchmark, the same query is used as a prompt for image synthesis. The presence of real image realizations for the prompts makes it possible to compare generated samples not only to the training images of the concept, but also to an empirical distribution of target outputs.

This structure makes it possible to evaluate two aspects separately. The first is whether a method actually learns a new concept, rather than mapping it back to a familiar category already present in the pretrained model. The second is whether the learned concept can still be combined correctly with multiple known contexts, including contexts shared across different concepts. In this sense, ConCon-Chi is useful because it evaluates not only concept recognition, but also concept-context compositionality.

3.3.4 Experimental Validation

The study evaluates personalized vision-language methods on ConCon-Chi. The benchmark contains 20 concepts, 101 contexts, and a test split with 1084 queries over an image pool of 4008 items, with on average about six real images per query (Rosasco et al., 2024). This scale matters because it supports several complementary analyses: concept-context bias, personalized retrieval, personalized generation, and performance breakdowns on common versus chimeric concepts.

One of the first experiments studies concept-context bias. Following prior work on composed retrieval, especially Pic2Word (Saito et al., 2023), the analysis varies the relative contribution of concept and context information in the query. On several existing datasets, context-only retrieval remains close to peak performance, indicating that the benchmark can often be solved through contextual shortcuts. ConCon-Chi behaves differently: performance collapses when only the context or only the concept is retained, showing that both components are genuinely needed.

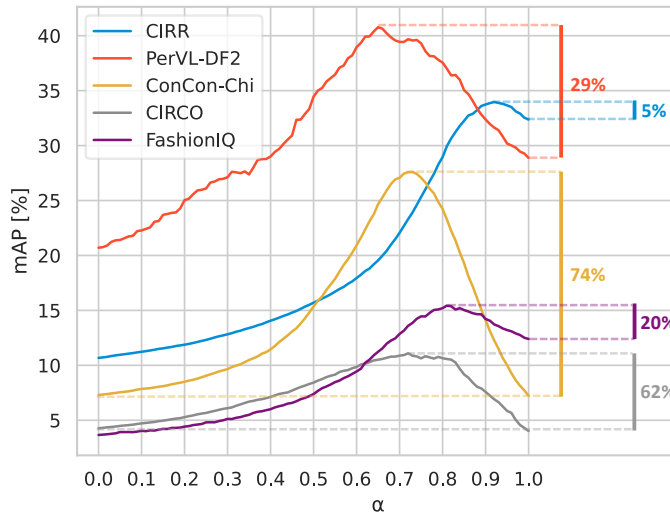


Figure 3.4 Concept and context bias on ConCon-Chi. Retrieval quality drops sharply when the query emphasizes only the concept or only the context, indicating that the benchmark discourages shortcuts based on either component alone.

The personalized retrieval benchmark then compares PALAVRA (Cohen et al., 2022) with two methods originally proposed for zero-shot composed image retrieval, Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023). With five example images per concept, SEARLE reaches the best retrieval performance on the benchmark, with an mAP of 30.74 (Rosasco et al., 2024). The analysis shows that methods can fail for opposite

reasons: PALAVRA recognizes the concept relatively well but struggles to preserve context composition, while Pic2Word is stronger on context but weaker on concept discrimination. SEARLE performs best because it reaches the best balance between these two requirements.

The generation benchmark arrives at a similar conclusion from a different angle. By evaluating methods such as Textual Inversion (Gal et al., 2023a) and DreamBooth (Ruiz et al., 2023) against real image realizations, the paper shows that personalized generation still faces a difficult trade-off between preserving the concept and satisfying the requested context. Moreover, common concepts remain consistently easier than chimeric ones, indicating that current methods still rely heavily on priors from the pretrained model. In this sense, ConCon-Chi is useful not only as a new benchmark, but also as a way to examine how dataset structure and evaluation protocol affect the assessment of personalized vision-language methods.

3.4 Summary

The research chapters later collected in Part II provide the full technical development of DualVCLIP and ConCon-Chi. Chapter 7 presents the action-recognition study in full (Calabrese et al., 2024), and Chapter 8 presents the ConCon-Chi benchmark together with its retrieval and generation analyses (Rosasco et al., 2024). Table 3.1 summarizes the two studies at a high level.

Across both chapters, language is the interface through which visual representations are queried or adapted. DualVCLIP shows that prompt learning can extend a CLIP-based video model to zero-shot multi-label action recognition. ConCon-Chi shows that, once prompt learning is used to represent personalized concepts, evaluating new-meaning learning and concept-context compositionality becomes a central problem in its own right.

This progression moves the dissertation from novelty-aware recognition to semantically extensible perception. Even so, the output of the system is still a recognition, retrieval, or generation decision. The next research direction asks how such perceptual capabilities matter once the robot must act rather than only interpret. This motivates Chapter 4, which turns to behavior cloning and the broader question of end-to-end policies.

Table 3.1 High-level summary of the two studies that support the vision-language research direction of this thesis.

Study	Problem	Experimental scope	Main takeaway
RO-MAN 2024	Zero-shot multi-label action recognition for verb-object actions	Charades benchmark with class-level random, verb, and object splits; comparison against CLIP-based and video-language baselines with mAP evaluation.	Prompt tuning and class-specific temporal aggregation make CLIP-style video features usable for zero-shot multi-label action recognition, but generalization is stronger for object semantics than for verb-level temporal compositionality.
CVPR 2024	Personalized text-to-image retrieval and generation	ConCon-Chi benchmark with 20 concepts, 101 contexts, 1084 retrieval/generation queries, chimeric concepts, bias analysis, and comparison of retrieval and generation methods.	Current personalized vision-language methods remain far from robust new-meaning learning; benchmark structure is crucial for measuring whether a method truly composes learned concepts with known contexts.

Chapter 4

Behavior Cloning and Toward End-to-End Policies

4.1 Introduction

This chapter reports the final research direction explored during the last phase of the PhD. Unlike Chapters 2 and 3, it does not correspond to research chapters reproduced later in the dissertation, because this line of work has so far taken the form of system development, deployment, teleoperation infrastructure, and ongoing policy-learning experiments. For this reason, the goal of the chapter is not to present a finalized method, but to explain the shift from perception modules studied in isolation toward embodied interaction on a humanoid robot, and to analyze the limitations of this shift and possible extensions.

The first outcome of this shift was the deployment of an interaction demonstrator on the ergoCub robot. At a high level, the demo lets the robot greet a user, establish mutual-gaze interaction, receive a box, hold or transport it, and hand it back, while also supporting social actions such as handshaking. The first method developed in this thesis, namely few-shot open-set action recognition, was one of the perceptual components integrated into that system together with mutual gaze estimation, 3D truncation-resistant pose estimation, face and box localization, behavior-tree-based decision logic, and motion primitives executed through robot controllers. The resulting system was repeatedly executed in the laboratory and in public demonstrations, so its value was not only functional but operational: it had to remain understandable, recoverable, and robust under repeated live use.

This deployed demo made explicit both the strengths and the limitations of a modular perception–decision–action loop. For a compact set of interaction skills and strong safety constraints, a behavior tree remains highly effective because it is interpretable, debuggable,

and operationally robust. At the same time, each additional behavior tends to require new perception outputs, new state variables, and new branching conditions. This observation motivated a second line of work during the final year of the PhD: an end-to-end behavior-cloning alternative to the modular perception–behavior-tree–action loop, implemented on the same robot and for the same family of interaction skills.

This dual implementation allows us to compare an end-to-end learning from demonstration framework with the previously manually designed behavior orchestration, and to find limitations in novel approaches. In particular, the policy must be trained on demonstrations that span all targeted behaviors, and in our setting the visual part of the policy showed limited robustness across different people and interaction styles. This is precisely where the earlier research directions become relevant again. A central motivation that emerges from this chapter is that end-to-end policies may benefit from stronger perceptual representations, especially temporally informed representations from action-recognition models, in order to improve robustness and reduce sample complexity.

The chapter is therefore organized around this broader argument. It first presents the modular ergoCub demo and the system-level choices behind it, then introduces behavior cloning and recent humanoid policy learning, and finally discusses the implemented teleoperation system (*metaCub*) and the behavior-cloning implementation developed on ergoCub, including the practical limitations that motivate future integration of action-recognition representations inside the perceptual front-end of the policy.

Indeed, the ergoCub project targets safe and collaborative humanoid operation in shared human workspaces. For this reason, it provides a natural setting to study how learned robot policies should be supported by perception of human actions, task-relevant objects, and unexpected situations.

4.2 Background

As part of the research reported in this thesis, I first contributed to implementing and deploying a modular human-robot collaboration pipeline that has been used on ergoCub over the last years for live demonstrations. Then, during the last year of the PhD, I promoted and helped implement a behavior-cloning pipeline on the same robot. This makes it possible to compare, for the same dynamic HRI setting, a manually designed orchestration based on perception modules and behavior trees with an end-to-end learning-from-demonstration alternative.

4.2.1 A Modular Interaction Demonstrator on ergoCub

We first introduce the ergoCub demo from a high-level view. The robot observes the user and the surrounding scene, extracts a small set of task-relevant perceptual variables, compresses them into a compact interaction state, and passes that state to a behavior tree that selects gaze targets, response actions, and emotion outputs. These commands are then executed by dedicated controllers and motion primitives. In the specific interaction showcased in the demo, the relevant tasks are greeting, handshaking, receiving a box from the user, and handing the box back.

Many HRI systems are built by decomposing the problem into perception, decision, and control layers (Bonasso et al., 1997; Simmons et al., 1997). Let o_t denote the sensory observation at time t , including camera data and any task-relevant robot state, and let z_t^{act} denote the observation stream used by the action recognition module, which may consist of 3D skeletons or image observations depending on the recognizer. In the ergoCub demo described in this chapter, the behavior tree consumes a compact symbolic state with five inputs:

$$\begin{aligned} \hat{y}_t^{\text{act}} &= \mathcal{A}(z_{t-L:t}^{\text{act}}), & m_t &= \mathcal{G}(o_t) \in \{0, 1\}, & \mathbf{p}_t^{\text{face}} &= \mathcal{F}(o_t) \in \mathbb{R}^3 \cup \{\emptyset\}, \\ \mathbf{p}_t^{\text{box}} &= \mathcal{D}(o_t) \in \mathbb{R}^3 \cup \{\emptyset\}, & b_t &\in \{0, 1\}. \end{aligned}$$

where $\mathcal{A}$ is the action recognition module, $\mathcal{G}$ produces a binary mutual-gaze variable, $\mathcal{F}$ returns the face position in cartesian coordinates expressed in the robot root frame when the face is detected, $\mathcal{D}$ returns the position of the manipulated object (in this demo, a box) in the same frame when it is detected, and b_t is an internal task-state variable indicating whether the robot is currently holding that box. The role of these geometric quantities is different: $\mathbf{p}_t^{\text{face}}$ supports social gaze and face-oriented interaction, $\mathbf{p}_t^{\text{box}}$ tells the robot where to reach, and b_t is updated by the execution logic after grasp or release events.

The behavior tree then maps these five quantities to three outputs:

$$(\mathbf{g}_t, r_t, e_t) = \pi_{\text{BT}}(\hat{y}_t^{\text{act}}, m_t, \mathbf{p}_t^{\text{face}}, \mathbf{p}_t^{\text{box}}, b_t),$$

where $\mathbf{g}_t \in \mathbb{R}^3$ is the cartesian point that the robot should look at, r_t is the response action selected by the tree, and e_t is the robot emotion to express. In practice, the gaze target gives priority to the detected box position whenever a box is present and the robot is not already holding it, and otherwise falls back to the detected face position. These three outputs are executed separately: $\mathbf{g}_t$ is sent to the head controller, e_t is sent to the facial-expression

controller that drives the head display, and r_t selects a prerecorded joint-space sequence that is replayed by the robot.

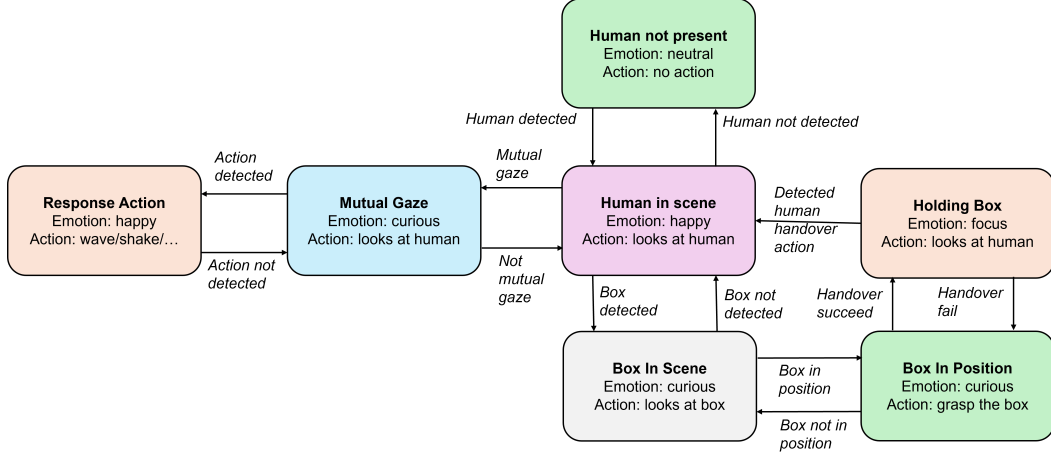


Figure 4.1 Approximate state machine of the interaction policy used in the modular ergoCub demo (some transitions and states has been dropped for clarity). The robot switches between idle, social interaction, box observation, box manipulation, and handover states according to human presence, recognized actions, mutual gaze, and box position.

In the deployed version of this modular demo (Fig. 4.1), the response motions were largely executed as fixed trajectory replays, with limited online geometric adaptation. As a result, behaviors such as box grasping and handshaking were reliable only within a constrained set of spatial conditions (for example, the box being presented in a particular pose and location, or the handover being performed in a repeatable configuration). Extending the same demo to handle broader variations would require additional components to estimate where the relevant interaction targets are (for example, the box pose or the human hand pose) and to adapt the motion primitives accordingly.

The box-held state is not estimated independently from raw perception at every time step. Instead, it is updated from the outcome of the manipulation primitive:

$$b_{t+1} = \begin{cases} 1, & \text{if a grasp primitive succeeds at time } t, \\ 0, & \text{if a release or handover primitive succeeds at time } t, \\ b_t, & \text{otherwise.} \end{cases}$$

This matches the practical logic of the demo: when the box is detected at position $\mathbf{p}_t^{\text{box}}$, the robot may decide to approach and grasp it; once that grasp succeeds, the controller assumes that the box is in hand until a later release resets the state.

This modularity is attractive for several reasons. Each module has a clear role, failure modes are easier to localize, and safety logic can be inserted explicitly. In the context of this thesis, the action-recognition module can also expose an *unknown* outcome, which is operationally meaningful because the behavior tree can fall back to a safe waiting state instead of committing to a wrong response.

The weakness of this architecture is that the interfaces are designed manually. The behavior tree only sees the few variables exposed upstream, and each new interaction requires engineering effort at multiple levels: perceptual predicates, transition conditions, and motion primitives. As the interaction vocabulary becomes richer, the pipeline becomes harder to maintain, because potentially informative details of motion and scene context remain in the raw sensor stream and never reach the symbolic decision layer. Figure 4.2 shows the behavior tree used in the deployed demo, while Figure 4.3 provides representative snapshots of the resulting interaction.

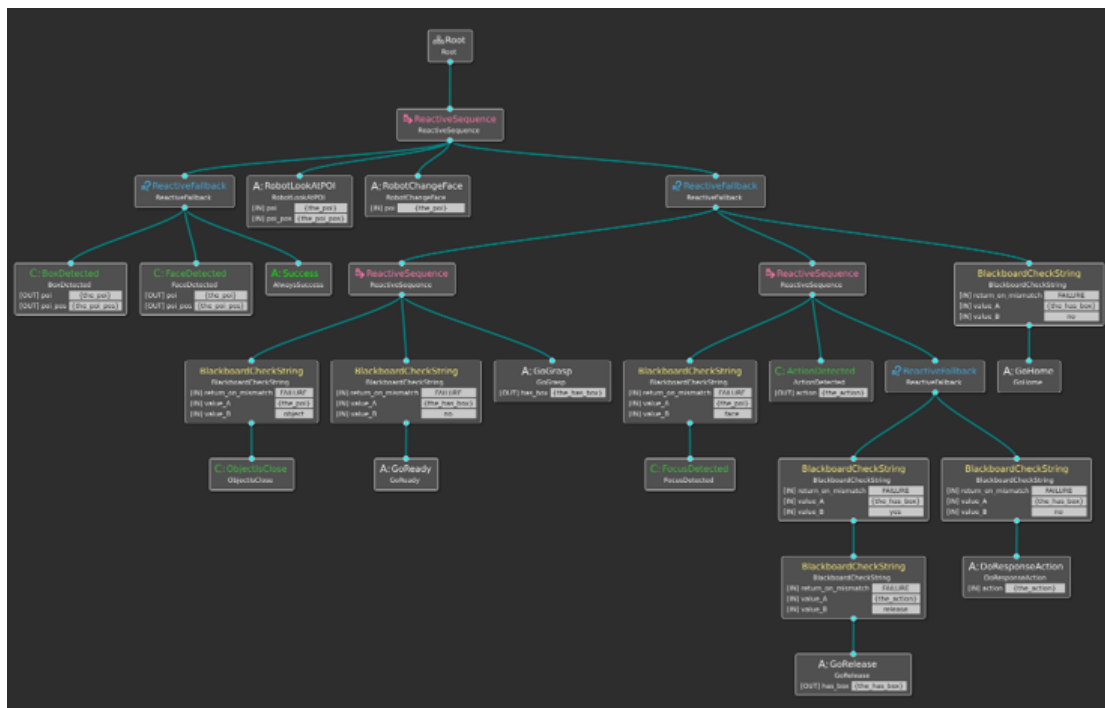


Figure 4.2 Behavior tree orchestrating gaze, response actions, and box-related branches during live interaction in the deployed ergoCub demo.

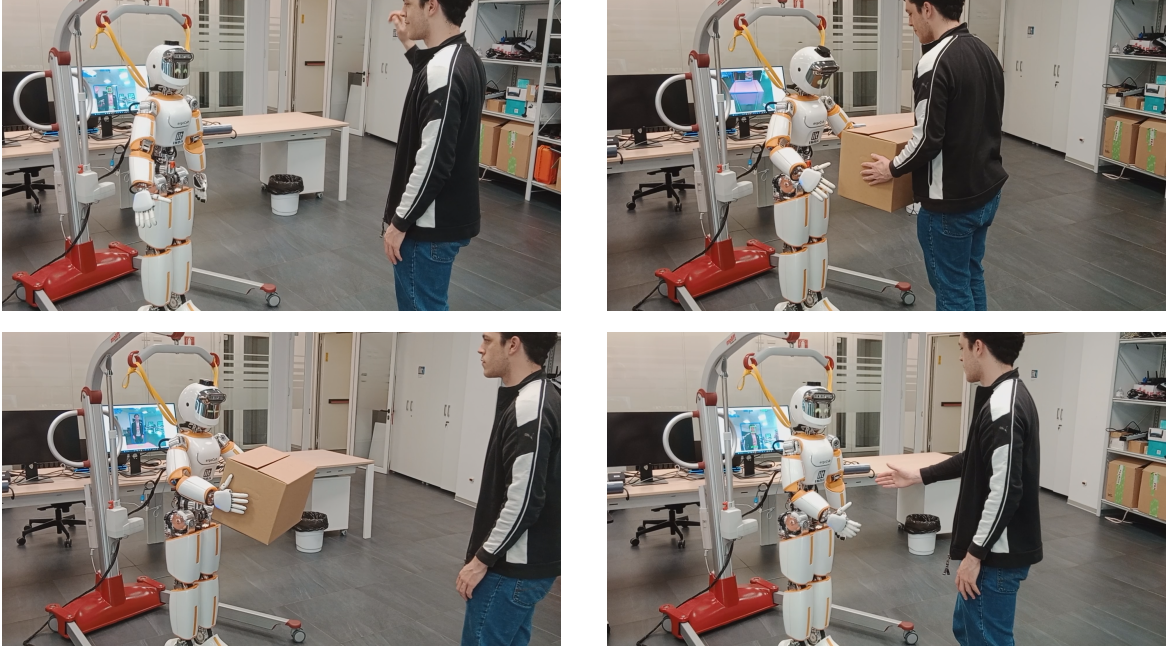


Figure 4.3 Selected moments from the deployed modular ergoCub demo: greeting, box reception, box transport, and handshake.

4.2.2 Behavior Cloning from Demonstrations

Behavior cloning can represent a different approach: instead of factorizing the interaction loop into separate modules, example demonstrations of the desired behavior are collected and a policy is trained directly from these observation-action trajectories (Chi et al., 2025; Levine et al., 2016). Let

$$\mathcal{D} = \left\{ \left(\mathbf{o}_{1:T_i}^{(i)}, \mathbf{a}_{1:T_i}^{(i)} \right) \right\}_{i=1}^M$$

denote a dataset of M demonstrations, where $\mathbf{o}_{1:T_i}^{(i)}$ is the observation sequence and $\mathbf{a}_{1:T_i}^{(i)}$ the corresponding action sequence. A policy π_θ is trained to predict the robot action from the current observation or a recent history:

$$\hat{\mathbf{a}}_{t:t+K-1} = \pi_\theta(o_{t-H:t}),$$

where the policy may predict either the next control command or a short horizon of future commands. This is the form commonly used by modern visuomotor behavior-cloning policies, including sequence-prediction methods such as Diffusion Policy (Chi et al., 2025).

For the ergoCub setting considered here, the observation is decomposed as

$$o_t = (I_t, p_t^{\text{robot}}),$$

where I_t denotes the camera images and p_t^{robot} the robot proprioception. A single control command can then be written as

$$a_t = (\xi_t^L, \xi_t^R, \mathbf{g}_t, \eta_t),$$

where ξ_t^L and ξ_t^R denote continuous control references for the left and right arm, respectively, $\mathbf{g}_t \in \mathbb{R}^3$ is the gaze target expressed in the robot root frame, and $\eta_t \in [0, 1]$ is a scalar emotion output. Depending on the controller, ξ_t^L and ξ_t^R may represent desired joint values or desired end-effector poses. In the ergoCub setup, the most natural choice is a cartesian action interface for the two hands, coupled with a separate head-gaze target. The emotion scalar is converted at execution time into one of a small number of display expressions by partitioning the interval $[0, 1]$ into fixed bins. In the broader control stack, the torso was intentionally restricted to yaw only, without pitch or roll, to keep the upper-body motion stable while still allowing the robot to turn toward the human.

For an explicit policy, the learning objective can be written as

$$\mathcal{L}_{\text{BC}}(\theta) = \frac{1}{\sum_i T_i} \sum_{i=1}^M \sum_{t=1}^{T_i} \ell\left(\pi_{\theta}(o_{t-H:t}^{(i)}), \mathbf{a}_{t:t+K-1}^{(i)}\right),$$

where ℓ is typically a regression loss over the predicted command sequence, for example

$$\begin{aligned} \ell_t &= \ell(\hat{\mathbf{a}}_{t:t+K-1}, \mathbf{a}_{t:t+K-1}) \\ &= \sum_{\tau=t}^{t+K-1} \left(\left\| \hat{\xi}_{\tau}^L - \xi_{\tau}^L \right\|_2^2 \right. \\ &\quad \left. + \left\| \hat{\xi}_{\tau}^R - \xi_{\tau}^R \right\|_2^2 \right. \\ &\quad \left. + \left\| \hat{\mathbf{g}}_{\tau} - \mathbf{g}_{\tau} \right\|_2^2 \right. \\ &\quad \left. + \left| \hat{\eta}_{\tau} - \eta_{\tau} \right|^2 \right), \end{aligned}$$

In this formulation the learned policy predicts continuous control references directly, rather than selecting among a finite set of prerecorded motions.

What changes conceptually is where the engineering effort is placed. In the modular pipeline, each piece of decision logic requires dedicated upstream signals and explicit

conditions: seeing the box requires a detector that returns a stable estimate of its position; preventing social motions while the robot is holding the box requires explicit state variables and branches in the behavior tree; and each new interaction variant typically introduces new predicates and transitions. In a behavior-cloning formulation, the policy is trained from demonstrations to map the sensory stream directly to control commands, reducing the amount of hand-designed interfaces between perception and action. While this can be seen as a simplification, it introduces a different requirement: the demonstrations must jointly cover the full multi-task interaction repertoire to be reproduced by the policy, and the policy must learn to disambiguate which behavior is appropriate from visual context and temporal evolution alone.

4.2.3 Recent Policy Learning for Humanoids

Recent robot-learning systems use demonstration data for training visuomotor policies. Diffusion Policies ([Chi et al., 2025](#)) is one representative example: it formulates behavior cloning as the generation of action sequences through a conditional denoising process and shows strong results on a wide range of manipulation tasks, including bimanual setting.

Considering humanoid platforms, NVIDIA Isaac GR00T ([Zhu et al., 2025](#)) scales the approach to a huge amount of demonstration data. GR00T N1 is presented as an open foundation model for humanoid robots, trained from egocentric human videos, real and simulated robot trajectories, and synthetic data; NVIDIA reports language-conditioned bimanual manipulation on humanoid platforms and positions the model as a reusable starting point for post-training on new robots and tasks. These systems are relevant here not because they solve our specific HRI problem directly, but because they make visible a broader shift in the literature: increasingly large amounts of demonstration data are used to train reusable visuomotor policies for humanoids, rather than engineering every interaction as branching logic over hand-designed perceptual modules.

4.3 Methods

This section describes how the same interaction scenario was reformulated as a behavior-cloning problem on ergoCub. The goal is make the comparison concrete on the robot: the same skill set is executed both through a perception–behavior-tree–action loop and through a policy learned from demonstrations.

4.3.1 The metaCub Teleoperation and Data Collection Interface

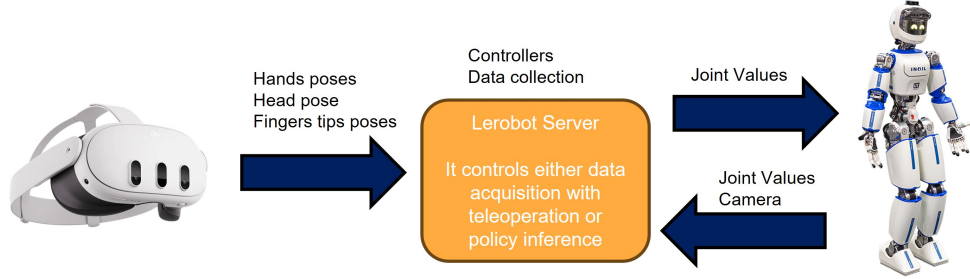


Figure 4.4 Teleoperation and command interface used to collect demonstrations and execute a learned policy on ergoCub.

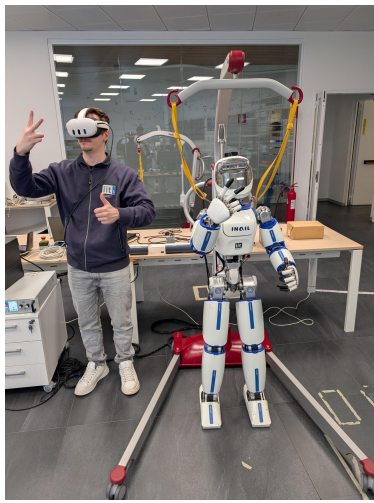
To collect demonstration data for behavior cloning, we implemented *metaCub*, a bimanual teleoperation system that streams the operator motion and the robot’s egocentric view and converts them into robot command trajectories (Fig. 4.4). The operator wears a Meta Quest 3 headset running an OpenXR module that estimates the pose of the head and both hands with respect to the Oculus world frame (Fig. 4.5a). These hand poses are then transformed into desired end-effector poses for ergoCub, expressed with respect to the robot root frame. A dedicated bimanual controller running at 100 Hz solves the dual-arm manipulation problem to satisfy both end-effector targets and generates smooth command trajectories at the same rate. In parallel, a head controller aligns the robot head orientation with the operator head orientation while compensating for torso motion, yielding a more stable and fluid teleoperation experience.

Finger motion is controlled by solving an inverse-kinematics problem that matches fingertip targets with respect to the palm frame. On the visual side, the headset displays the robot egocentric camera stream. Since ergoCub uses an Intel RealSense sensor that provides two infrared images separated by approximately 7 cm, the left and right infrared images are streamed to the corresponding displays in the headset, providing a stereo view with depth cues that supports geometrically demanding tasks and fine-grained manipulation (Fig. 4.5b). The images are compressed and streamed over Wi-Fi.

The data collection protocol involved two people. One user teleoperated the robot through the headset and generated the bimanual motion commands, while another user interacted physically and socially with the robot in front of it, for example by approaching, offering a handshake, or presenting the box. Overall, we collected approximately 40 minutes of teleoperation data with different users interacting with the robot and with different visual backgrounds. The goal of this variation was to expose the learned policy to multiple human

appearances, interaction styles, and environments, so that the policy would not be tied to a single person or a single laboratory configuration. A complementary setting in which the policy is learned from only one interacting human and one environment remains unexplored in this work and is an important direction for understanding the limits of data-efficient behavior cloning.

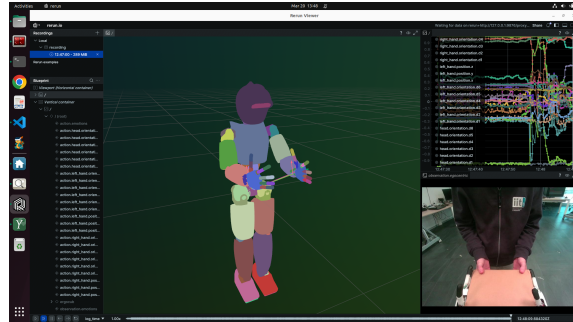
Finally, the recorded dataset stores the commanded bimanual end-effector poses (rather than robot joint angles), which facilitates transferring the learned policy across different humanoid platforms. A view of such synchronized data is depicted in Fig. 4.5c.



(a) Teleoperation setup with Meta Quest 3.



(b) Egocentric stereo IR stream.



(c) Synchronized data logging.

Figure 4.5 metaCub teleoperation interface and data-collection pipeline on ergoCub. (a) The operator wears a Meta Quest 3 headset to teleoperate the robot. (b) The two infrared views from the RealSense sensor are streamed to the left/right displays to provide stereo depth cues. (c) Overview of the recorded signals used to build the demonstration dataset.

4.3.2 An End-to-End Interaction Demo on ergoCub

The same interaction skills included in the deployed modular demo were also implemented through behavior cloning, by collecting demonstrations and training a policy to output continuous command variables. In this setup, demonstrations are recorded through a bimanual

teleoperation interface developed within *metaCub*. The learned policy then generates the action commands at inference time (for example, desired end-effector poses for the two hands, a gaze target, and an emotion value), and these commands are tracked by the same low-level controllers used for teleoperation and execution.

The demonstrations must cover the targeted interaction repertoire, including the multi-task nature of the demo, and they must do so under enough variation for the policy to learn not only the arm motion but also when each behavior should be expressed. This was one of the main motivations for developing the teleoperation and collection tools around *metaCub*: without a stable interface for recording synchronized robot observations and commands, the move from a scripted demo to a learnable policy wouldn't be possible.

In practice, this approach can support increased reactivity and robustness across demonstrations and across executions. The modular version mainly replays fixed responses and therefore relies on the environment matching the assumptions implicit in those trajectories. In the behavior-cloning version, the policy conditions its continuous end-effector commands on the current RGB observations and proprioception, which makes it possible to reach for the box at different positions and to execute a handover or handshake with spatially consistent motion. Importantly, this does not require adding explicit box or hand pose estimation modules and new behavior-tree conditions for each spatial variant of the same interaction.

At the same time, the development experience of the last months also exposed the limits of a purely end-to-end approach. The most evident issue was not low-level control, but perceptual generalization: compared with object-centered interaction cues, the policy was less robust to variation across people, action styles, and social context. This matters because, in a dynamic HRI setting, deciding whether the human is greeting, offering a handshake, or initiating a handover is itself a temporal perception problem. These observations motivate a more nuanced direction than a simple “modular versus end-to-end” dichotomy: one may instead seek policies whose perceptual front-end is strengthened with representations learned by action-recognition models, so as to improve robustness and reduce sample complexity while retaining the advantages of demonstration-based control.

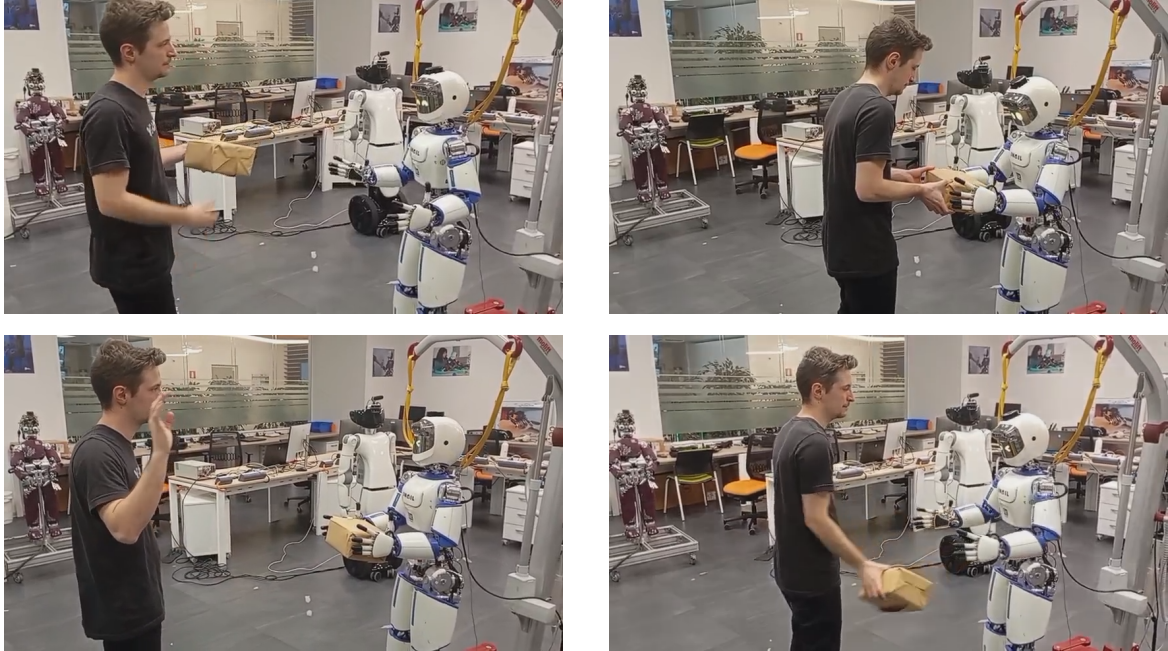


Figure 4.6 Selected moments from the behavior-cloning-oriented demo, centered on box exchange and the continuous control behavior that a learned policy should reproduce.

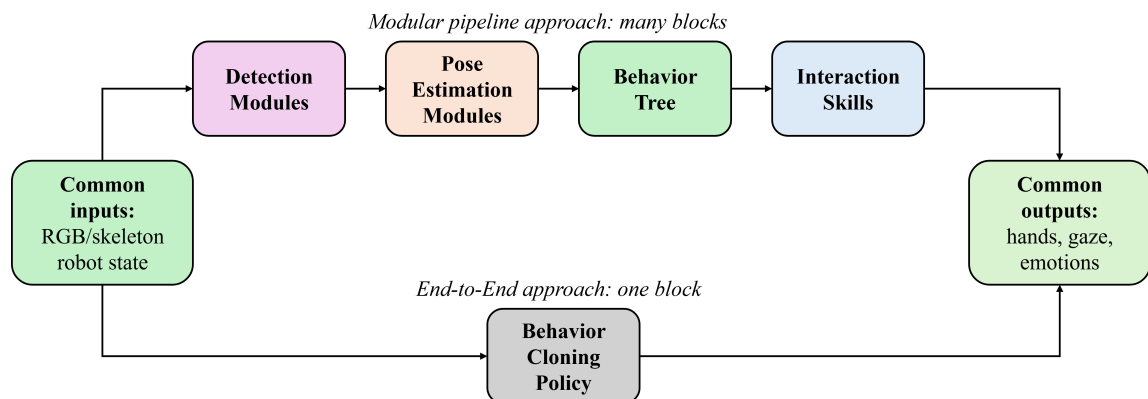


Figure 4.7 Comparison between the modular architecture and the behavior-cloning pipeline. The modular architecture is composed of several explicit blocks, while behavior cloning replaces the perception–state–decision–control chain with one end-to-end learned policy.

Table 4.1 Comparison between the deployed modular HRI pipeline and a behavior-cloning reformulation for the same interaction skills.

Aspect	Modular pipeline	Behavior cloning / end-to-end policy
Inputs to the decision layer	Recognized action, binary mutual gaze, human face position in the robot root frame, box position in the robot root frame, and internal box-held state.	Camera images and robot proprioception.
Outputs of the decision layer	Gaze target in cartesian coordinates, response action, and robot emotion.	Continuous control references, such as desired end-effector poses or joint values, together with the head-gaze target and a scalar emotion output.
Supervision	Requires designed predicates, transition conditions, and motion primitives.	Requires demonstrations aligned with the continuous command interface used at execution time.
Adding a new skill	Typically requires changes in perception, logic, and control.	In principle requires new demonstrations and policy adaptation or post-training.
Interpretability	High: failures can often be traced to a module or branch.	Lower: action decisions emerge from the learned policy.
Uncertainty handling	Can explicitly abstain or wait if upstream modules expose rejection.	Usually implicit; rejection and fallback behavior must be designed separately.
Deployment profile	Strong for repeated demos with a compact interaction vocabulary.	Promising when behavior is richer, more continuous, or harder to script.

4.4 Summary

The modular ergoCub demo and the behavior-cloning variant emphasize different design trade-offs. The modular system showed that a carefully engineered stack can produce reliable interaction in public settings, especially when the interaction vocabulary is compact, and safety and interpretability matter more than expressive richness. In that regime, explicit rejection, clear behavior-tree branches, and inspectable controllers remain substantial advantages.

The behavior-cloning implementation is useful for a different reason. It shows that some interaction skills, especially those involving continuous spatial adaptation such as handover and handshake, can be reformulated more naturally as learned visuomotor behavior rather than as fixed motion primitives selected by symbolic logic. This is the main motivation for transitioning from a modular perception–action pipeline toward end-to-end perception–action learning.

At the same time, the last months of development also made clear that behavior cloning is not a complete solution by itself. The policy must cover a multi-task interaction repertoire through demonstrations, its uncertainty handling is less explicit, and its perceptual front-end showed limited generalization across people and social interaction styles. In our setting, this suggests that the main bottleneck is not only control but also perception: the policy needs temporally meaningful cues about what the human is doing, and learning those cues from scratch inside a behavior-cloning pipeline may be data-inefficient.

For this reason, the natural continuation is not simply to replace the handcrafted stack with a monolithic policy. A more promising direction is to inject stronger perceptual representations into the policy, especially representations learned by action-recognition models, so as to improve robustness and sample complexity while preserving the benefits of demonstration-based control. In that sense, the dissertation moves from perception to action, but also returns from action to perception by asking how the perceptual advances of the earlier chapters can support more robust embodied policies. In the following chapters the research directions explored in the thesis are presented in full.

Part II

Research Chapters

Chapter 5

One-Shot Open-Set Skeleton-Based Action Recognition

Abstract

Action recognition is a fundamental capability for humanoid robots to interact and cooperate with humans. This application requires the action recognition system to be designed so that new actions can be easily added, while “unknown” actions are identified and ignored. In recent years, deep-learning approaches represented the principal solution to the Action Recognition problem. However, most models often require a large dataset of manually-labeled samples. In this work we target One-Shot deep-learning models, because they can deal with just a single instance for class. Unfortunately, One-Shot models assume that, at inference time, the action to recognize falls into the support set and they fail when the action lies outside the support set. Few-Shot Open-Set Recognition (FSOSR) solutions attempt to address that flaw, but current solutions consider only static images and not sequences of images. Static images remain insufficient to discriminate actions such as sitting-down and standing-up.

In this paper we propose a novel model that addresses the FSOSR problem with a One-Shot model that is augmented with a discriminator that rejects “unknown” actions. This model is useful for applications in humanoid robotics, because it allows to easily add new classes and determine whether an input sequence is among the ones that are known to the system. We show how to train the whole model in an end-to-end fashion and we perform quantitative and qualitative analyses. Finally, we provide real-world examples.

5.1 Introduction

Action Recognition (AR) represents a fundamental task for Human-Robot Interaction ([Akkaladevi and Heindl, 2015](#); [Rodomagoulakis et al., 2016](#); [Roitberg et al., 2014](#)), video surveillance ([Han et al., 2018](#); [Khan et al., 2020](#)), and sport-analysis ([Soomro and Zamir, 2014](#); [Zhu et al., 2006](#)), to mention just a few. Different AR solutions require different data, such as still images ([Liu et al., 2020](#)) ([Jeong et al., 2021](#)), sequences of images ([Bo et al., 2020](#)) ([Careaga et al., 2019](#)), and sequences of 2D/3D skeleton data ([Ren](#)

et al., 2020) (Wang et al., 2021a). The AR-based on sequences of 2D/3D skeleton data has the advantage to use less computation power and ignores the background, focusing only on the human pose (Perrett et al., 2021). To train an AR pipeline, we often need big labeled datasets, especially for Deep-Learning models. These datasets require manual collection and annotation, and the resulting model requires a fine-tuning, or worse, full retraining to simply add or change a class. These models are badly suited for those applications in robotics in which the robot is required to learn new actions from a few demonstrations and, perhaps more importantly, should be able to recognize when a given input belong to one of the “known” classes. A class of models called Few-Shot Learning (FSL) aims at solving the problem above by adapting to recognize new classes given few samples. The FSL models for AR are called Few-Shot Action-Recognition (FSAR) models (Guo and Cheung, 2020; Wu et al., 2019). Usually, in a few-shot problem, the model receives as input a set of few examples for each class called *support set* and a query; then, the model outputs the probabilities for the query to belong to a class in the support set. However, in practical applications, it is not possible to foresee in advance all the actions humans will perform while interacting with the robot. In this “Open Set” (OS) scenario, we cannot add all the possible classes of actions (which are infinite) to the support set and reliably classify each sample. The OS problem combined with the FSL problem yields the Few-Shot Open-Set Recognition (FSOSR) (Liu et al., 2020) problem and its relative application to AR is called Few-Shot Open-Set Action Recognition (FSOSAR). Typically this problem is solved by creating a “reject” class, whose instances can be sampled from the training set and can be inferred from the distances between the query and the elements of the support set (Liu et al., 2020) (Jeong et al., 2021). While the FSOSAR problem has been investigated for images (Liu et al., 2020) (Jeong et al., 2021), to the best of our knowledge, there is no work that addresses the FSOSAR problem for sequences.

In this work, we propose a solution to the FSOSAR problem (that can easily be extended to the FSOSR problem) by defining a discriminator that assigns a confidence score to the difference between the features of the query and the features of the predicted class. With episodic end-to-end learning, we force this difference to have a high norm value when the difference is high and low otherwise. Moreover, the discriminator forces this difference to contain information about similarities and differences between the query prototype and the query-class prototype of the predicted class. Finally, we propose to use features obtained from a 3D skeleton of the human, resulting in a simpler model that is less computationally expensive and easier to train. An overview of our approach is depicted in Figure 5.1. Our contributions can be summarized as follows:

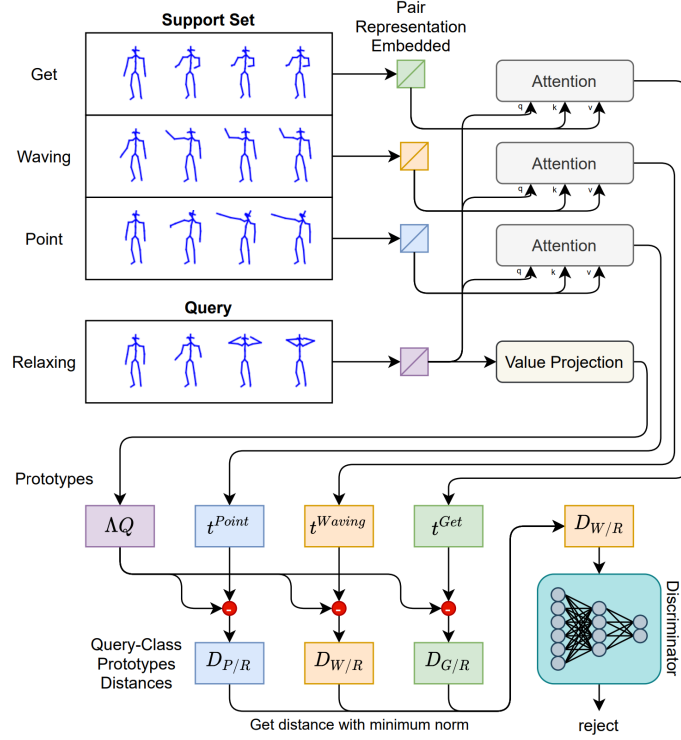


Figure 5.1 The query class "relaxing" is not present in the support set, but the FS part of the model chooses the "waving" class as the solution. Our discriminator *Disc* discards the FS prediction because it provides a low confidence score (0.11). Thus, our approach selects "reject" (i.e. "unknown") as the final class.

- We created a model that deals with Few-Shot Open-Set learning for sequences of 3D skeleton.
- We proposed an episodic end-to-end learning technique that balances the outcome of the discriminator and allows to jointly train the FSL part and the OS part of the model.
- We demonstrated qualitatively that the model is capable of differentiating similar but semantically different actions.

We released the code¹ and the dataset² in order to make all our experiments easily reproducible.

¹Available at <https://github.com/stefanoberti/ISBFSAR>

²Available at <https://bit.ly/30cD9yb>

5.2 Related Work

In this section we give a brief review of various FSL methods, analyzing the advantages and disadvantages of different types of input data and the main differences with respect to our approach.

5.2.1 Few-Shot Action Recognition (Videos)

A recent work ([Careaga et al., 2019](#)) proposed to aggregate frame-level features of both the RGB image and the optical flow across time. For the time-aggregation of the features, they tested Pooling, Long-Short Term Memory (LSTM), ConvLSTM, and 3D-Convolutions. As a FSL algorithm, they tested Matching Networks, Prototypical Networks, and Learned Distance Metrics. Another work ([Bo et al., 2020](#)) proposed to use multiple Temporal Attention Vectors to encode video-wide temporal information that adapts to the various lengths of frame sequences. The importance scores are then used to select important temporal information.

The works above use sequences of images as input data that allow taking into consideration the whole context, like the objects that the human is interacting with and the surrounding elements of the environment. In contrast, our work focus on 3D human skeleton data as we do not need to train a Convolutional Neural Network to extract features from the important parts of the RGB images. As a result, our model requires much less computational power to train and fewer samples when the action is mainly discriminated through the human pose.

5.2.2 Few-Shot Action Recognition (Skeleton)

A recent work ([Wang et al., 2021a](#)) tackles the Skeleton-Based FSAR problem with a variant of Dynamic Time Warping. It tries to find the best alignment in the temporal and simulated camera viewpoint spaces between the query and the support frames. A temporal block encoder expressed the sequences. In short, Dynamic Time Warping (DTW) is a sequence-alignment technique that finds an optimal monotonic correspondence between the frames of two sequences that may differ in speed or temporal offset, so that frame-by-frame distances are minimized along the matched path. In ([Wang et al., 2021a](#)) this alignment is performed jointly over two axes: the temporal axis and a set of simulated camera viewpoints, i.e. the warping is also used to compensate for the orientation of the skeletons.

Our work also aims at the best temporal alignment between the query and the support set, but in contrast we do not explicitly simulate different camera viewpoints. It is important

to note that, in our method, temporal alignment and orientation alignment are decoupled: the temporal matching is handled by the cross-transformer of the base model (TRX), while robustness to the skeleton orientation is delegated to the embedding layer rather than to the warping procedure. For this reason, the camera-viewpoint component of DTW is not required in our setting: the embedding layer, trained with a dataset where actions are recorded from different orientations, is designed to project poses observed under different rotations close to each other in feature space.

5.2.3 Few-Shot Open-Set Recognition (Images)

The FSOSAR problem addresses two problems at the same time: the FSAR problem and the OS problem. FSOSAR has been studied in the literature by (Liu et al., 2020) and (Jeong et al., 2021), but mainly using still images as input data. Some approaches (Liu et al., 2020) propose to solve the FSOSR problem with a combination of Meta-Learning, a random selection of novel classes per episode, a loss that maximizes the posterior entropy for examples of those classes, and a new metric learning formulation based on the Mahalanobis distance (Mahalanobis, 1930). Another approach (Jeong et al., 2021) extended the one above, taking inspiration from the concept of reconstruction-based OS recognition methods, where an algorithm is trained to reconstruct the training data and it fails at reconstructing unseen samples. In this work the authors trained a transformation function based on the fact that the distance between a query feature and its class prototype is closer than the distance between an unknown class feature and that prototype. Then they compute the distance between the transformed query and the transformed elements of the support set that are not chosen as the class and they check that the difference is under a certain threshold. In this way, they avoid dependency on pseudo-unseen examples.

In contrast to these works, our approach considers sequences as input data and can therefore discriminate ambiguous frames based on previous ones. For example, by using sequences we can understand if a person is sitting down or standing up, but that is tricky to understand from a single image.

5.3 Background

The Temporal-Relational Cross-Transformer (TRX)(Perrett et al., 2021) is a few-shot method for action recognition that considers the similarity between all the possible sub-sequences of length 2 of the query frames and all the possible sub-sequences of length 2 of all the

elements in the support set (the model was also extended to sub-sequences of length > 2 , but in this paper we focus only on pairs of frames). We are interested in the K -way N -shot problem with $N = 1$ (also known as One-Shot). With an attention mechanism, the TRX model constructs “query pair”-specific class prototypes and selects the class relative to the element in the support set which has a minimum query-class difference. Given a query $Q \triangleq \{q_1, \dots, q_F\}$, the query representation for the pair p is defined as:

$$Q_p \triangleq [\phi(q_{p_1}) + PE(p_1), \phi(q_{p_2}) + PE(p_2)] \in R^{2 \times D}$$

where p_1 and p_2 are ordered indices of frames such that $1 \leq p_1 < p_2 \leq F$, F is the total number of frames, $\phi : R^{H \times W \times 3} \rightarrow R^D$ is a convolutional network that extracts a D -dimensional embedding from each input frame and PE is the positional encoding (Gehring et al., 2017). We define the set of all possible pairs as:

$$\Pi \triangleq \{(n_1, n_2) \in \mathbb{N}^2 : 1 \leq n_1 < n_2 \leq F\}$$

The support set S contains a support $S^c = \{s_1^c, \dots, s_F^c\}$ for each class c . A single frame-pair representation of the video with class c in the support set with respect to the ordered pair of indices $m = (m_1, m_2) \in \Pi$ is:

$$S_m^c \triangleq [\phi(s_{m_1}^c) + PE(m_1), \phi(s_{m_2}^c) + PE(m_2)] \in R^{2 \times D}$$

and the set of all pair representations in the support set for class c is:

$$S^c \triangleq \{S_m^c : (m \in \Pi)\}$$

By using three different linear projections called queries

$$\Upsilon : R^{D \times 2} \rightarrow R^{D_\Upsilon},$$

with key

$$\Gamma : R^{D \times 2} \rightarrow R^{D_\Gamma},$$

and value

$$\Lambda : R^{D \times 2} \rightarrow R^{D_\Lambda},$$

we can compute the query-specific class prototype for the query Q_p with respect to the class c :

$$t_p^c \triangleq \text{Softmax}(L(\Gamma \cdot S_m^c) \cdot L(\Upsilon \cdot Q_p)) \cdot (\Lambda \cdot S_m^c)$$

where $t_p^c \in R^{D_v}$ and L is the standard layer normalization (Ba et al., 2016). Then we compute the distance between the query Q_p mapped with Λ and each query-specific class prototype of the support set S^c :

$$T(Q, S^c) \triangleq \frac{1}{|\Pi|} \sum_{p \in \Pi} \|t_p^c - \Lambda \cdot Q_p\|$$

In this way, we can compute a distance between the query and each class of the support set. The model is then trained with a cross-entropy loss over the query-class negative distances $T(Q, S^c)$.

5.4 Problem Formulation

In this section, we formally formulate the problem of FSL and we extend it to the FSOSR. We pose the problem as a single-instance classification problem, where each class encodes a different action. Let $D^S \triangleq \{x_i^S, y_i^S\}_{i=1}^{NK}$ be the support set, where x_i^S and y_i^S represent an instance and its label respectively, K the number of classes, and N the number of labeled samples for each class. Given a known query set $D^{Kn} \triangleq \{x_i^{Kn} \in X^{Kn}, y_i^{Kn} \in Y^S\}_{i=1}^Q$, where Q is the number of query samples for each class, the aim of FSL is to infer for each query sample x_i^{Kn} the corresponding class y_i^{Kn} that is assumed to always be contained inside the support labels Y^S .

In an FSOSR problem, the query set contains also instances of classes that are not part of the support set. The query set $D^Q \triangleq D^{Kn} \cup D^{Un}$ equals the union of both the known query set D^{Kn} and the unknown query set $D^{Un} \triangleq \{x_i^{Un} \in X^{Un}, y_i^{Un} \in Y^{Un}\}_{i=1}^{N^{Un}}$, where $Y^S \cap Y^U = \emptyset$ holds and N^{Un} is the number of unknown instances. Thus, a FSOSR problem requires to create a classifier f such that given a support set S , for each query sample x_i and true class y_i , the following holds:

$$f(x_i, S) = \begin{cases} y_i, & \text{if } y_i \in Y^S \\ False, & \text{otherwise} \end{cases} \quad (5.1)$$

$$f(x_i, S) = y_i, y_i \in Y^S \quad (5.2)$$

Since previous work focuses on the FSOSR problem for still images (Liu et al., 2020) (Jeong et al., 2021) and fail to correctly discriminate some actions as for “standing-up”

and “sitting-down”, we address the problem using sequences of 3D skeletons. We focus on sequences of 3D skeletons because we are mostly interested in the human pose, we want to be able to differentiate temporal actions, and because working with skeleton data requires much less computational power (at least for the training).

5.5 Proposed Solution

In this section, we present a novel model for FSOSR with its relative end-to-end training technique that allows to increasingly train the discriminator with a balanced number of positive and negative samples. An overview of our proposed solution is depicted in Figure 5.1.

5.5.1 Model

Our proposed model is a variation of the TRX (Perrett et al., 2021) model. To adapt it for sequences of skeleton, we replaced the convolutional neural network $\phi : R^{H \times W \times 3} \rightarrow R^D$ with an *embedding* network $\Psi : R^{J \times 3} \rightarrow R^D$ where J is the number of joints of the skeleton, that are represented as points in three dimension. We used a Multi-Layer Perceptron (Kubat, 1999) as an embedding network, with a hidden layer of size $J \times 3 \times 2$ and an output layer of size D . We also added a ReLU (Agarap, 2018) activation function after each layers. Given different skeleton rotations during training, the embedding function Ψ is encouraged to project similar poses observed under different rotations into nearby D -dimensional vectors. This is a design choice rather than a formally quantified invariance property, and it is supported here only qualitatively (Section 5.6.3).

To extend to the OS this FSAR model, we need to add a “reject” class that indicates that the query class y_i is not contained inside the support set classes Y^S . We define $t^c \triangleq \text{stack}(\{t_p^c : p \in \Pi\})$ and $\tilde{q} \triangleq \text{stack}(\{\Lambda \cdot Q_p : p \in \Pi\})$, where *stack* is a function that concatenates the elements a sequence along a new dimension, such that $t^c \in \mathbb{R}^{|\Pi| \times D^\Lambda}$ represent the concatenation of the pairs of query-class prototypes and $\tilde{q} \in \mathbb{R}^{|\Pi| \times D^\Lambda}$ represents the concatenation of the pairs of query prototypes. We could extend to the OS by comparing the query-class prototypes t^c of each class c with the query prototypes $\tilde{q}$, but this would lead to many negative samples and an unbalanced training of the discriminator. Instead, we propose an alternative approach in which an action is associated to a set of known actions and a discriminator is trained to accept or reject the best association. We add this behavior to our model in the following way: first we get the class c_{FS} that is chosen by the FSL part of

the model

$$c_{FS} \triangleq \underset{c}{\operatorname{argmin}} \{T(Q, S^c)\}$$

, then we give the matrix $\tilde{q} - t^{c_{FS}}$ to a discriminator *Disc*. The discriminator *Disc* assigns a score $[0, 1]$ to its input that represents the confidence of associating the query-class features difference with two semantically close skeleton sequences. We implemented *Disc* with a Multi-Layer Perceptron as follows: the first layer $l^1 : \mathbb{R}^{D^\Lambda} \rightarrow D^r$ reduces the dimension of each element of the sequence, obtaining a matrix with size $\mathbb{R}^{|\Pi| \times D^r}$. Then it flattens the last two-dimension and uses three layers to progressively reduce the dimension of the array to one neuron, with a ReLU (Agarap, 2018) function between each pair of layers. Finally, we apply a Sigmoid function to obtain a score in $[0, 1]$ that we compare with a threshold value to either "accept" or "reject" the class predicted by the FSL part of the model. Thus, the output of our model holds the following:

$$f(x_i, S) = \begin{cases} c_{FS}, & \text{if } \text{Disc}(\tilde{q} - t^{c_{FS}}) > \tau \\ \text{False}, & \text{otherwise} \end{cases} \quad (5.3)$$

where τ is a threshold that depends on the similarity of the actions we are considering. Basically, if the actions in the support set are very similar, a higher threshold value τ may be needed to correctly reject unknown queries.

5.5.2 Training

The model is trained end-to-end with episodic learning. To train the FSAR part of the model, we follow the approach of TRX (Perrett et al., 2021): given a query $Q_i = \{q_{i1}, \dots, q_{iF}\}$ with true class $y_i \in Y^S$ and a support set $S_i = \{S_i^1, \dots, S_i^K\}$ with K examples (one for each class c), we compute the FS loss with a cross-entropy between the query-class negative distances and a one-hot target vector for the task i is:

$$\ell_{FS} \triangleq \frac{1}{|B|} \sum_{i=0}^B CE(\text{stack}(\{-T(Q_i, S_i^c) : c \in Y^S\}), \tilde{y}_i)$$

where B is the number of training samples, CE stands for Cross-Entropy and $\tilde{y}_i$ is an one-hot vector with the entry corresponding to y_i set to 1.

To jointly train the discriminator, we need to feed it with positive and negative samples. To train the model end-to-end we need to consider only true positive samples, that are samples Q_i that were correctly classified by the FSAR part of the model, and so that $c_{FS,i} = y_i$. We

define z as the number of samples correctly classified in a mini-batch by the FSAR part of the model. Now that we have z positive samples, we can obtain the negative samples by just sampling queries Q_i with label y_i from the unknown set D^U . By sampling z negative samples we can keep the training of the discriminator balanced, where the number of samples that it receives at each training step is $z \times 2$ and it increases during training accordingly to the accuracy of the FSAR part of the model. This also allows us to back-propagate the OS loss to the FSAR part of the model only where the query and the most similar element in the support set are similar, providing more information to the training. We define this loss as:

$$\ell_{OS} \triangleq \frac{1}{|B|} \sum_{i=0}^B \tilde{\ell}_{OS}(Q_i, S_i^c)$$

where, once defined $\tilde{o}_i \triangleq \text{Disc}(\tilde{q}_i - t_i^c)$:

$$\tilde{\ell}_{OS}(Q_i, S_i^c) = \begin{cases} \text{BCE}(\tilde{o}_i, 1), & \text{if } c_{FS,i} = y_i \\ \text{BCE}(\tilde{o}_i, 0), & \text{if } y_i \notin Y^S \\ 0, & \text{if } y_i \in Y_i^S \wedge c_{FS,i} \neq y_i \end{cases}$$

where BCE stands for Binary Cross-Entropy. By combining the FS loss and the OS loss we obtain the total loss:

$$\ell_{FSOS} \triangleq \ell_{FS} + \sigma \ell_{OS}$$

where σ is a regularizer value.

5.6 Experimental Validation

In this section, we provide the experimental validation of our model. In particular, we investigate the performance of our model in recognizing new classes from a small set of examples and the ability of the model to detect and discard “unknown” actions. We compare our model with the baseline, providing quantitative and qualitative results. We then show that, for the classes in the test set, our approach solves the problem defined in Section 5.4. The experimental validation characterizes in:

- Quantitative validation: we compared the accuracy of the FSOS task of our model with respect to a baseline, where the discriminator is not present and a query is “rejected” if the exponential of the negative query-class distance is lower than a certain threshold.

- Qualitative validation: we discussed some predictions of our model over the test set, demonstrating its strength to avoid recognizing wrong actions.

Unfortunately, to the best of our knowledge, there are no other works in the literature that studied the FSOSAR problem for sequences of 3D skeleton and therefore we cannot make a fair comparisons with other methods on this task. Thus, instead, we created a baseline where the discriminator is removed and the OS loss is computed directly on the negative query-class prototypes distance without further processing. Since the baseline does not use a discriminator explicitly trained on the OS task, we expect it to have lower accuracy.

5.6.1 Comparison with Baseline (Quantitative Validation)

We declare a simple baseline called *EXP* where the discriminator is removed and the confidence is computed with an exponential function over the maximum negative distance $\max_c \{-T(Q, S^c) : c \in Y^S\}$. Indeed, since the query-class negative distances are bounded in $(-\infty, 0]$, its exponential has value inside $[0, 1]$ that can be used as a confidence score to either confirm or “reject” the class predicted by the FSL part of the model.

As dataset, we consider the NTURGBD-120 dataset (Liu et al., 2019). This dataset contains 120 classes involving daily actions, medical actions, and two people interactions. Each class has about a thousand samples with relative RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos. However, some samples do not have skeleton data and, as pointed out also in the literature (Trivedi et al., 2021), most of the given skeletons have low quality and are noisy. Because of this, we decided to re-extract the skeleton from the RGB videos. Thus, we took 16 equidistant frames from each video, we extracted the 3D skeleton with MetrABS (Sáránci et al., 2020) and we centered the skeleton with respect to the pelvis coordinate. The output of the MetrABS heads is already bounded inside $[-1, 1]$, and therefore it did not need normalization. On the other hand, we decided not to normalize the rotation because we want the embedding layer to be invariant to the rotation of the skeletons. We removed the 26 classes representing two people interactions since our problem focuses on a single skeleton at a time. To evaluate the performance of our method, we use an approach similar to the One-Shot Action Recognition described in the literature (Liu et al., 2019), but adapted for the OS task: we randomly select k classes and for each of them we add in the support set the corresponding exemplar (Liu et al., 2019) (one fixed action instance for class), then we compute the metric on all the elements of the test set. In this way, we have a known set D^K that contains all the samples of the k selected class and an unknown set D^U that contains all the samples from the remaining classes. We compute a

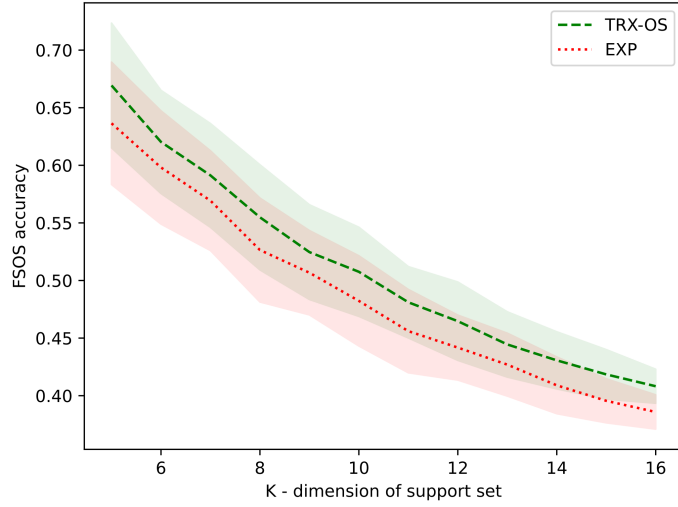


Figure 5.2 by the FSOS-ACC metric on our dataset. The *TRX-OS* model outperforms the baseline for every K .

	K=5	K=10	K=15
TRX-OS	0.67 ± 0.05	0.52 ± 0.04	0.43 ± 0.02
EXP	0.63 ± 0.05	0.51 ± 0.04	0.41 ± 0.02
RANDOM	0.17	0.09	0.06

Table 5.1 We also show random results for reference, where each query has $1/(K+1)$ (where the +1 comes from the reject class) probability to be classified as the true class.

metric that we call *FSOS-ACC*, which is the accuracy between the predicted classes and the target classes, considering also the “reject” class. We expect our model to classify as “reject” all the instances in the unknown set D^U_n and to classify with the correct classes in the support set all the instances of the known set D^K_n . We performed the comparisons for growing values of K ; for each K we train the model 100 times and we report the mean accuracy and its standard deviation.

Figure 5.2 depicts the results of our comparison, while Table 5.1 shows the numerical values. For higher values of k , there is an increasing number of instances that should be rejected by the discriminator. Moreover, the more classes there are in the support set, the harder is to match the query with the true class because it is more probable that some classes are similar. With $k = 5$ we can reject or select the true class the 67% of the times, that means that our model can be reliably used in real Human-Robot scenarios when the number of classes is limited. Our model outperforms the baseline for all the values of k . This means that the discriminator learns useful concepts that are also back-propagated to the FSL part of the model thanks to the Open-Set loss ℓ_{OS} .

5.6.2 Discriminator’s Confusion Matrix (Quantitative Validation)

Another quantitative analysis refers to the Open-Set Accuracy of our model. In fact, if we set our support set size to one we can build a confusion matrix on the output of the discriminator to get insights on which actions are more difficult to classify. We use the exemplar of one class from the test set as the support set, then we attempt to classify all the elements of the test set and we computed the number of instances that are classified as *True* by our model for each class. We show a subset of the results of this analysis in Table 5.2. We expect the elements on the diagonal to have a value close to one and all the other elements of the matrix to have a value close to zero. Our model performs well in this task for some classes like *Take off glasses* and *Falling*, but other classes like *Tear up paper* and *Apply cream on face* have low performances. This happens mainly for two reasons, as discussed also in the following Subsection 5.6.3:

- The skeleton model that we used for training and testing does not track the hands of the human. As a consequence, it is impossible to differentiate between actions like *Apply cream on face* and *Hush (quite gesture)* because the focus of these actions is on the hand movement and the fingers position.
- By using skeleton data only, we ignore which object the human is interacting with. Due to this fact, actions like *Stample book* and *Tear up paper* are very similar.

We emphasize that these confusions stem from a limitation of the input modality and of the dataset (untracked hands and absence of object information) rather than from the proposed open-set mechanism itself: such classes are intrinsically indistinguishable from the 3D skeleton alone. Consequently, the aggregate results reported above would appear cleaner had these intrinsically ambiguous classes been excluded from the evaluation; we deliberately keep them to report performance under the full, unfiltered benchmark. An important aspect of this analysis is that we set the threshold value τ of Equation 5.3 to 0.5, but different values could be more suitable for different applications.

5.6.3 Qualitative Results

In our application when the robot detects an action performed by the human it triggers his motors and assists the human in the completion of the action. Since these actions involve also dangerous Human-Robot Interactions (e.g. the robot passes a box to the human), it is needed that the robot does not misclassify critical human actions. In the following, we show and discuss some qualitative examples that justify the existence of our model. We show three

	Take off glasses	Tear up paper	Pointing something	Falling	Hush (quite)	Stample book	Apply cream on face	Arm circles
Take	0.91	0.01	0.03	0.	0.01	0.	0.05	0.
Tear	0.	0.42	0.02	0.	0.	0.49	0.06	0.
Pointing	0.	0.03	0.82	0.	0.	0.	0.	0.
Falling	0.	0.	0.	0.87	0.	0.	0.	0.
Hush	0.09	0.	0.02	0.	0.74	0.	0.48	0.
Stample	0.	0.06	0.	0.	0.	0.52	0.	0.
Apply	0.12	0.	0.02	0.	0.48	0.	0.6	0.
Arm	0.	0.03	0.02	0.	0.	0.	0.	0.71

Table 5.2 On the rows, we have the classes that were added in the support set, and on the columns, we have the classes that were tested against the exemplar for the class in the support set.

	Take (know)	Put (unknown)
Point	0.	0.24
Reach	0.	0.04
Take	0.99	0.72
OS-Score	0.87	0.09

Table 5.3 The first three rows shows the FSL scores for the known/unknown query and all the elements in the support set, the last row shows the discriminator's confidence between the known/unknown query and the element of the support set with highest FSL score.

examples in the 3-Way 1-Shot scenario for ease of representation. We define the OS-Score as the output of the Discriminator, where a value close to one stands for "Confirm" and a value close to zero stands for "Reject".

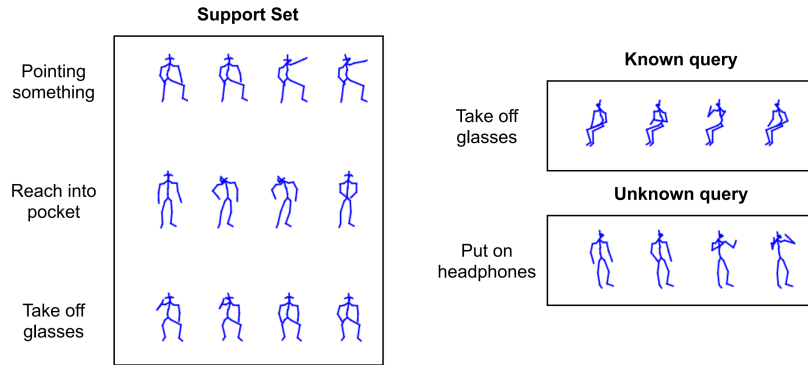


Figure 5.3 On the left the support set and on the right the known query and the unknown query.

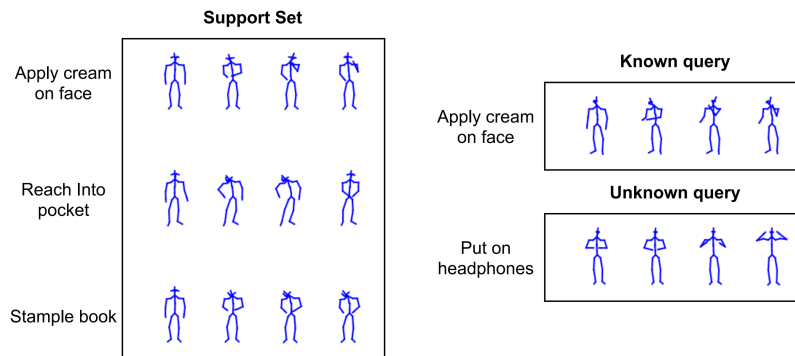


Figure 5.4 On the left the support set and on the right the known query and the unknown query.

	Apply (known)	Put (unknown)
Apply	0.99	0.98
Reach	0.0	0.1
Stample	0.	0.1
OS-Score	0.98	0.31

Table 5.4 The first three rows shows the FSL scores for the known/unknown query and all the elements in the support set, the last row shows the discriminator's confidence between the known/unknown query and the element of the support set with highest FSL score.

In the first example, depicted in Figure 5.3 with results shown in Table 5.3, both the FS and OS parts of the model correctly classify the known query *Take off glasses*, providing qualitative evidence that the model can cope with different rotations of the skeleton in this example. The unknown query *Put on headphones* is similar to the exemplar of the class *Take off glasses* in the support set, but in the first action, the object is brought closer to the

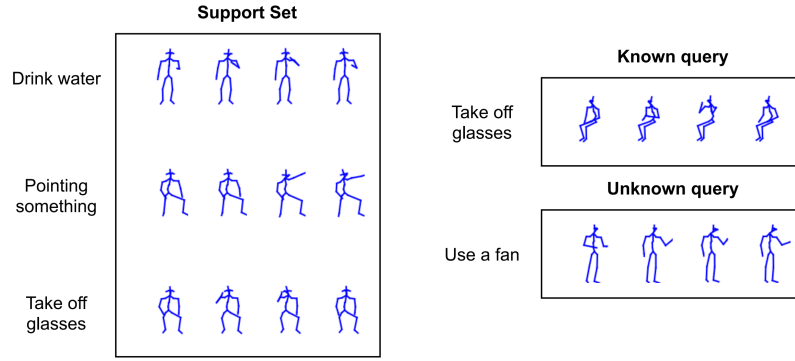


Figure 5.5 On the left the support set and on the right the known query and the unknown query.

	Take (known)	Use (unknown)
Drink	0.	0.02
Point	0.	0.97
Take	0.99	0.01
OS-Score	0.87	0.80

Table 5.5 The first three rows shows the FSL scores for the known/unknown query and all the elements in the support set, the last row shows the discriminator's confidence between the known/unknown query and the element of the support set with highest FSL score.

head while in the latter the object is moved away from the head. Indeed, the FS part of the model chooses the *Take off Glasses* action because it is the most similar in the support set, but the discriminator correctly rejects the unknown query, showing that the discriminator also learned to deal with temporal information.

In the second example, depicted in Figure 5.4 with results shown in Table 5.4, both the FS and OS parts of the model correctly classify the given known query *Apply cream on face* as one class in the support set, with a high OS-Score. This case is not too difficult for the FS part of the model because the *Apply cream on face* action is the only one of the support set where the arms reach the head's height. For the same reason, the FS part of the model classifies the unknown query *Put on headphones* with high confidence as *Apply cream on face* function, but then the discriminator rejects the prediction.

The third example, depicted in Figure 5.5 with results shown in Table 5.5, is interesting because it highlights the limitation of the 3D skeleton data. As in the other examples, the FSL part of the model chooses the right action in the support set for the known query *Take off glasses* and the discriminator confirms the prediction. Anyway, the unknown query tricks our model. Indeed, the unknown action *Use a fan* is performed only with wrist movements and

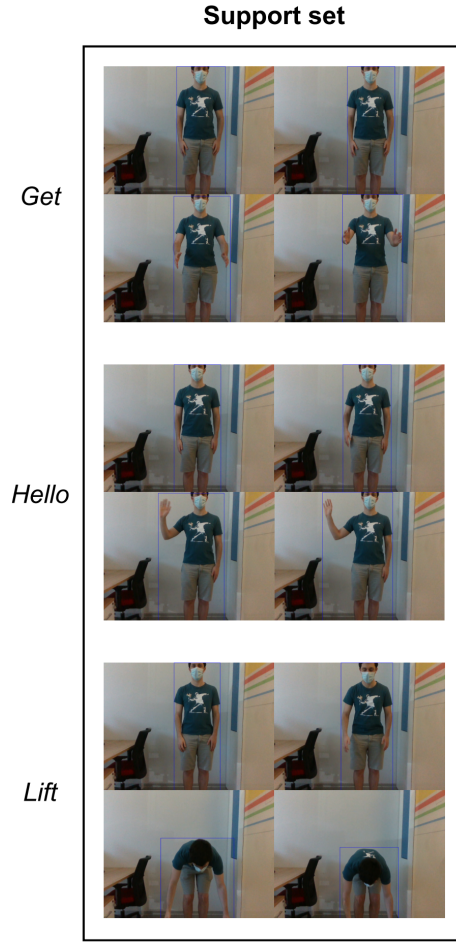


Figure 5.6 Support set as a reference for the two unknown queries.

does not involve movements from the rest of the arm. Since MetrABS does not track hand movements, the sequence of skeletons extracted from the video does not describe correctly the action and, just by looking at them, it is easy to think that the human is extending his arm towards something and nothing else. Because of this, the FS part of the model chooses *Pointing something* as the candidate class and the OS part of the model confirms the wrong prediction since it does not know about the hand movements.

In Figure 5.6 we provide examples with data streamed from a camera. The skeleton are extracted from a input video with MetrABS (Sáráñdi et al., 2020). In our case, we wanted the robot to detect when the human is ready to receive a box (i.e. action “get”), but reject all those cases in which the human is doing different movements. We can see that the desired behavior is obtained for both the two queries in Figure 5.7 and Figure 5.8: the model associates the actions as closer to “get”, however, the decision is rejected based on the the OS score, that is low for both examples. This demonstrates that, without the discriminator and

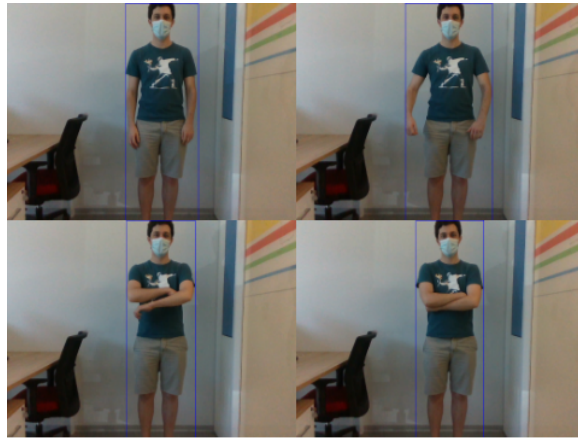


Figure 5.7 : the FS part of the model classifies this query as "get" with 54% of confidence, but the query is rejected because the Discriminator's output is low (0.07).



Figure 5.8 : the FS part of the model classifies this query as "get" with 99% of confidence, but the query is rejected because the Discriminator's output is low (0.01).

the OS score, the system would classify all the unknown queries as actions “Get”, producing unwanted robot behavior.

We showed that our model can efficiently deal with the FSOS problem, avoiding to incorrect classifying queries with the true class that does not belong to the support set. Thus, for the classes tested x_i and the support set S out proposed approach, encoded by $f(x_i, S)$ (in Equation (5.3)) approximates the function $\tilde{f}(x_i)$ (in Equation (5.1)) of problem defined in Section 5.4, and outperforms the baseline.

5.7 Conclusion

In this paper we introduced a novel model to deal with the Few-Shot Open-Set Recognition problem for sequences of 3D skeleton data. A discriminator is added to confirm or “reject” the difference between the query prototype and the query-class prototypes of the most similar element in the support set. We explained how to train the discriminator jointly with the FSL part of the model in such a way to keep the training of the discriminator balanced. Moreover, the discriminator is trained proportionally with the accuracy of the FSAR part of the model. We performed a quantitative analysis with a baseline to demonstrate the importance of the discriminator and qualitative analysis, demonstrating that our model can differentiate similar skeleton sequences that perform different actions.

Chapter 6

A Baseline Study and Benchmark for Few-Shot Open-Set Action Recognition with Feature Residual Discrimination

Abstract

Few-Shot Action Recognition (FS-AR) has shown promising results but is often limited by a closed-set assumption that fails in real-world open-set scenarios. While Few-Shot Open-Set (FSOS) recognition is well-established for images, its extension to spatio-temporal video data remains underexplored. To address this, we propose an architectural extension based on a Feature-Residual Discriminator (FR-Disc), adapting previous work on skeletal data to the more complex video domain. Extensive experiments on five datasets demonstrate that while common open-set techniques provide only marginal gains, our FR-Disc significantly enhances unknown rejection capabilities without compromising closed-set accuracy, setting a new state-of-the-art for FSOS-AR. The project website, code, and benchmark are available at: <https://hsp-iit.github.io/FSOSAR/>.

6.1 Introduction

Deep learning has achieved remarkable success in action recognition by learning complex spatio-temporal representations from large-scale datasets (Pareek and Thakkar, 2021). However, the high cost of video annotation limits these models in real-world scenarios. Few-shot (FS) recognition addresses this by enabling generalization to novel classes with minimal labeled examples (Perrett et al., 2021; Thatipelli et al., 2022; Wang et al., 2023b; 2024c; Wu et al., 2022; Xing et al., 2023). Recent advances in FS-AR have further leveraged multi-modal settings—such as RGB, optical flow, or text—to improve robustness through complementary information (Tang et al., 2024; Wang et al., 2024a; 2021d; 2024b; Wanyan et al., 2023).

A critical yet underexplored challenge in FS-AR is the open-set scenario, where models must reject unknown test samples to avoid false positives. While few-shot open-set (FSOS) recognition is established for images (Deng et al., 2022; Huang et al., 2022; Jeong et al., 2021; Liu et al., 2020; Song et al., 2022; Su et al., 2024; Sun and Chu, 2024; Wang et al., 2023a), existing methods neglect the temporal dynamics of video. To our knowledge, the

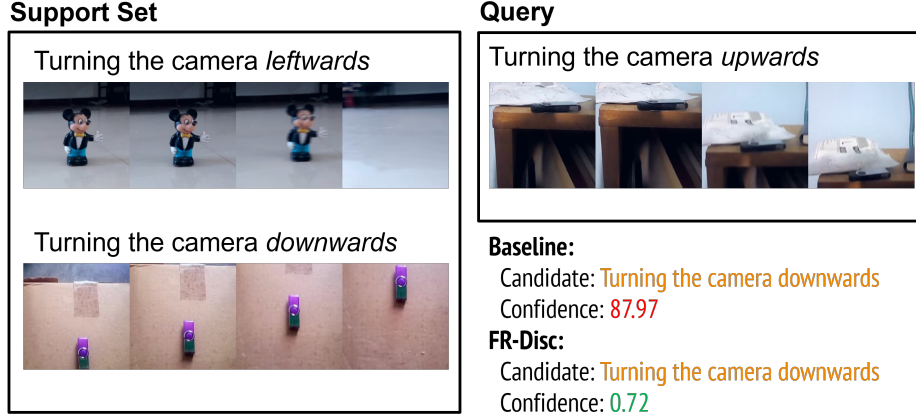


Figure 6.1 The Support Set contains $K = 2$ classes and $N = 1$ videos. The true class of the query is not contained in the Support Set, so it should be rejected through low confidence (e.g., $< 50\%$). The Baseline incorrectly classifies it as known (false positive), while FR-Disc correctly rejects it (true negative). Additional qualitative examples are provided in the Supplementary Material.

only existing approach for FSOS action recognition is our previous work (Berti et al., 2022), which was limited to skeletal data.

Traditional open-set techniques rely either on scoring mechanisms—such as Maximum Softmax Score (MSS) or Maximum Logit Score (MLS) (Dhamija et al., 2018)—or auxiliary classification modules (Ochal et al., 2024; Wang et al., 2023a). Studies in image-based OS recognition (Bisgin et al., 2024; Rody et al., 2020; Vaze et al., 2021) show that MLS often outperforms MSS and more complex methods because the Softmax operation discards logit magnitude, which is vital for uncertainty estimation (Vaze et al., 2021). We investigate these insights in the video-centric FSOS domain by establishing a comprehensive benchmark suite across five datasets. We adapt two distinct FS-AR models, STRM (Thatipelli et al., 2022) and SAFSAR (Tang et al., 2024), with these techniques, to provide a baseline for performance-efficiency trade-off.

Our main contributions are as follows:

- We adapt five state-of-the-art FS-AR video datasets to open-set conditions and establish the first comprehensive set of baselines for FSOS-AR.
- We demonstrate that FS-AR models can be equipped with OS recognition techniques without degrading closed-set accuracy, extending image-based OS findings to the video domain.

- We introduce a specialized architecture by evolving the Feature-Residual Discriminator (FR-Disc) (Berti et al., 2022) for high-dimensional video data. We show that modeling discrepancies between query features and class prototypes is superior to standard logit-based metrics (MLS/MSS) for capturing complex temporal dynamics.

Overall, this work provides a foundational study of FSOS-AR, offering a robust evaluation of varied approaches on public benchmarks.

6.2 Related Work

6.2.1 Few-Shot Action Recognition

FS-AR advances focus on modeling spatio-temporal dynamics to generalize across novel classes. Temporal-relational methods like TRX (Perrett et al., 2021) and STRM (Thatipelli et al., 2022) align query-support patterns via cross-transformers, while motion-aware approaches such as MTFAN (Wu et al., 2022) and MoLo (Wang et al., 2023b) leverage motion-modulated fragments or contrasting contexts. Other frameworks integrate graph-guided matching (GgHm (Xing et al., 2023)) or temporal set alignment (HyRSM++ (Wang et al., 2024c)). Multi-modal trends further enhance robustness by using semantic priors (Wang et al., 2021d), commonsense knowledge prompting (Shi et al., 2022; 2024), multi-modal feature exploration (Wanyan et al., 2023), or CLIP-guided modulation (Wang et al., 2024a; b). Specifically, SAFSAR (Tang et al., 2024) demonstrates state-of-the-art performance in 1-shot settings. Despite their success, these models assume a closed-set environment, limiting real-world utility.

6.2.2 Open-Set Recognition

Images. Early works introduce training-based losses like Entropic Open-Set (EOS) and Objectosphere to regularize feature magnitudes (Dhamija et al., 2018). While foundational baselines relied on Maximum Softmax Scores (MSS) or “garbage” classes, recent studies (Bisgin et al., 2024; Vaze et al., 2021) show that Maximum Logit Scores (MLS) often achieve state-of-the-art results by preserving logit magnitude, outperforming generative (Chen et al., 2021a; Neal et al., 2018) or post-processing methods like OpenMax (Ge et al., 2017). Additionally, while input perturbations (e.g., ODIN) can improve detection, MLS remains a more practical baseline due to its computational efficiency (Roady et al., 2020).

Videos. Video-centric OS recognition is less explored. Recent works evaluate metrics like AUROC on datasets such as FineGym and Diving48 using Regularized-Discriminative OpenMax (Yin et al., 2024), or employ Evidential Deep Learning (Bao et al., 2021) and Information Bottleneck theory (Cen et al., 2023) on UCF101 and HMDB51. Skeleton-based OS-AR has also been recently benchmarked on NTURGBD (Peng et al., 2024).

6.2.3 Few-Shot Open-Set Recognition

Images. FSOS image recognition balances limited labels with unknown rejection through meta-learning (PEELER (Liu et al., 2020)), transformation consistency (SNATCHER (Jeong et al., 2021)), task-adaptive sampling (SEMAN-G (Huang et al., 2022)), and background cues (ProCAM (Song et al., 2022)). Subsequent refinements include relative displacement (Deng et al., 2022), energy-based frameworks (Wang et al., 2023a), and vision-language integration (Chen et al., 2024; Miller et al., 2024; Ochal et al., 2024; Sun and Chu, 2024). These, however, do not encode the temporal information necessary for video.

Videos. To our knowledge, only our previous work (Berti et al., 2022) addresses FSOS-AR, using skeletal data and a feature-residual rejection module. This present work evolves that concept to high-dimensional video data, establishing its effectiveness on our proposed benchmark suite.

6.3 Problem Definition and Metrics

6.3.1 Task Definition

We define Few-Shot Action Recognition (FS-AR) as a classification task where each class represents a different action. Let $D = \{(x_i, y_i) : y_i \in Y\}$ be a video dataset with label set Y , where x_i and y_i represent a video and its class respectively. The label set is split into disjoint training Y^{train} and test Y^{test} sets.

Few-Shot Action Recognition. Following episodic training (Cao et al., 2020; Perrett et al., 2021), we sample N^{train} tasks to form the task set $T_{\text{known}}^{\text{train}} = \{SS_i, Q_i^{\text{known}}\}_{i=1}^{N^{\text{train}}}$. Each support set, defined as $SS_i = \{(x_{ij}, y_{ij}) : y_{ij} \in Y_i^{\text{SS}}\}_{j=1}^{NK}$, contains NK examples, that are N videos for each of the K classes sampled from Y^{train} . The query Q_i^{known} belongs to the same label set Y_i^{SS} . The objective is to train a classifier f such that:

$$P_{\text{FS}} \triangleq f(SS_i, x_i) = \hat{y}_i : \hat{y}_i = y_i. \quad (6.1)$$

At test time, the protocol uses tasks $T_{\text{known}}^{\text{test}}$ sampled from Y^{test} .

Few-Shot Open-Set Action Recognition. We extend this to FSOS-AR by introducing an “unknown” task set $T_{\text{unknown}}^{\text{train}} = \{SS_i, Q_i^{\text{unknown}}\}_{i=1}^{N^T}$, where the query Q_i^{unknown} belongs to classes $Y_i^{\text{unknown}} = Y^{\text{train}} - Y_i^{\text{SS}}$ (or $Y^{\text{test}} - Y_i^{\text{SS}}$ during testing). The full task set is $T = T_{\text{known}} \cup T_{\text{unknown}}$. We first define a binary classification problem P_{OS} to distinguish known from unknown samples:

$$P_{\text{OS}} \triangleq f(SS_i, x_i) = \hat{a}_i : \hat{a}_i = a_i, \quad (6.2)$$

where $a_i = 1$ (accept) for known and $a_i = 0$ (reject) for unknown queries. Ultimately, **FSOS-AR** requires predicting both the acceptance state and, if known, the specific class:

$$P_{\text{FSOS}} \triangleq f(SS_i, x_i) = \begin{cases} \hat{a}_i : \hat{a}_i = a_i \\ \hat{y}_i : \hat{y}_i = y_i & \text{if } a_i = 1 \end{cases} \quad (6.3)$$

6.3.2 Performance Metrics

We evaluate performance using three categories of metrics:

Closed-Set Metrics. Few-shot accuracy (**FS ACC**) evaluates the model’s ability to discriminate among K known classes (P_{FS} , Eq. 6.1), ignoring open-set detection.

Open-Set Metrics. To evaluate the P_{OS} binary task (Eq. 6.2), we report open-set accuracy (**OS ACC**), **AUROC**, and **AUPR**. These are computed over the balanced set $T_{\text{known}}^{\text{test}} \cup T_{\text{unknown}}^{\text{test}}$.

Joint Metrics. The Open-Set Classification Rate (**OSCR**) (Dhamija et al., 2018) measures the trade-off between open-set detection and closed-set accuracy (Eq. 6.3). It computes the area under the Correct Classification Rate (CCR) vs. False Positive Rate (FPR) curve, where $\text{CCR} = \frac{1}{|T_{\text{known}}^{\text{test}}|} \sum_{(SS_i, Q_i) \in T_{\text{known}}^{\text{test}}} \mathbb{I}[\hat{y}_i = y_i] \mathbb{I}[\hat{a}_i = 1]$.

6.4 Methods

6.4.1 Few-Shot Action Recognition Models

We employ two representative closed-set FS-AR baselines: **STRM** (Thatipelli et al., 2022), a 2D transformer-based model, and **SAFSAR** (Tang et al., 2024), a 3D model. STRM enhances 2D features via patch-level and frame-level aggregation modules to capture local and temporal details. SAFSAR utilizes a 3D VideoMAE (Tong et al., 2022) backbone and incorporates text-based semantic priors to guide feature learning, offering superior performance at a higher computational cost. Both models use a standard Cross-Entropy (CE)

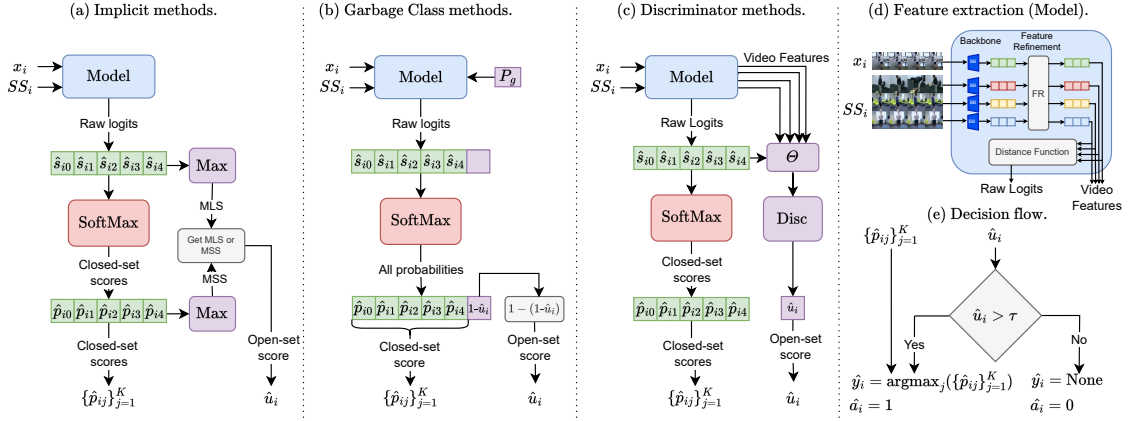


Figure 6.2 for (a) implicit (Softmax, EOS), (b) explicit (Garbage Class), and (c) Discriminator-based methods. For implicit methods, $\hat{u}_i$ is derived via MLS/MSS. In (b), a query is rejected if the garbage prototype P_g yields the highest probability. In (c), $\hat{u}_i$ is computed by an MLP with Sigmoid activation using the Θ function (Eq. 6.6). (d) and (e) illustrate the feature extraction and inference decision flow for SAFSAR and STRM models.

loss on the known task set $T_{\text{known}}^{\text{train}}$:

$$\mathcal{L}_{\text{CE}} = - \sum_{j=1}^K y_{ij} \log(\hat{p}_{ij}) \quad (6.4)$$

where $\hat{p}_{ij}$ is the predicted probability for class j . At inference, the class with the maximum probability is selected.

6.4.2 Open-Set Techniques

We extend these baselines to output a known/unknown score $\hat{u}_i \in [0, 1]$. We categorize these extensions into implicit and explicit techniques (Fig. 6.2).

Implicit OS Techniques

These require no additional parameters, modifying only the scoring or loss function.

- **Softmax Baseline:** Uses the Maximum Logit Score (**MLS**), defined as

$$\hat{u}_i = \text{Max}(\{\hat{s}_{ij}\}_{j=1}^K)$$

or the Maximum Softmax Score (**MSS**), defined as

$$\hat{u}_i = \text{Max}(\text{Softmax}(\{\hat{s}_{ij}\}_{j=1}^K)) = \text{Max}(\{\hat{p}_{ij}\}_{j=1}^K)$$

as the open-set score $\hat{u}_i$. A query is accepted if $\hat{u}_i > \tau$.

- **Entropic Open-Set (EOS)**: Adds an entropy-boosting term to the loss to push predictions for unknown tasks $T_{\text{unknown}}^{\text{train}}$ toward a uniform distribution:

$$\mathcal{L}_{\text{EOS}} = -\frac{1}{K} \sum_{j=1}^K \log(\hat{p}_{ij}).$$

The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha_{\text{EOS}} \mathcal{L}_{\text{EOS}},$$

where α_{EOS} balances the two terms.

Explicit OS Techniques

These methods introduce extra parameters to produce a $(K + 1)$ -way output, where the additional logit represents the “unknown” class.

- **Garbage Class (GC)**: A random learnable “Garbage Prototype” P_g is added to the support set. During training, unknown queries are assigned the label $K + 1$. At inference, the query is rejected if $\hat{p}_{i,K+1}$ is the maximum probability; the open-set score is $\hat{u}_i = 1 - \hat{p}_{i,K+1}$.
- **Feature-Residual Discriminator (FR-Disc)** (Berti et al., 2022): This approach introduces a lightweight auxiliary network (the Discriminator) fed with the “residual” (difference) between the query features and the features of the closest support class. The Discriminator is trained to distinguish between known and unknown queries. During training, positive samples are defined as known queries correctly classified by the closed-set model, while an equal number of negative (unknown) samples are used to maintain class balance. At inference, the Discriminator’s confidence score, $\hat{u}_i$, serves as the known/unknown indicator, which is compared against a threshold τ to accept or reject a query. The training loss for the Discriminator is defined as:

$$\begin{aligned} \mathcal{L}_{\text{Disc}}(\{\hat{p}_{ij}\}_{j=1}^K, y_i) &= \mathbb{I}(\hat{y}_i = y_i) \cdot \text{BCE}(d_i, 1) \\ &+ \mathbb{I}(y_i \in Y_i^{\text{unknown}}) \cdot \text{BCE}(d_i, 0), \end{aligned} \quad (6.5)$$

where $\mathbb{I}$ is the indicator function, BCE is the Binary Cross-Entropy loss, and d_i represents the feature residual:

$$d_i = \Theta(x_i, SS_i, \{\hat{s}_i\}_j) = \phi(x_i) - \phi(SS_{i, \text{argmax}_j(\{\hat{p}_{ij}\}_{j=1}^K)}), \quad (6.6)$$

Here, $\phi(\cdot)$ denotes the feature extractor, x_i is the query, and argmax_j returns the value j corresponding to the maximum value of its argument. The final joint loss is:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha_{Disc} \mathcal{L}_{Disc}. \quad (6.7)$$

where α_{Disc} is a weighting hyperparameter. This approach shares similarities with metric-based methods, such as Siamese Networks, which utilize feature differences to measure similarity and have been shown to enhance performance in few-shot classification (G. Koch, R. Zemel, R. Salakhutdinov, 2015). We adapt this technique to the video domain by defining individual Discriminators for both the SAFSAR and STRM modules, more details are available in Supplementary Material.

6.5 Experimental Setup

6.5.1 FSOS-AR Benchmark Suite

To establish a robust FSOS-AR benchmark, we utilize train/test splits from existing FS-AR literature to generate known and unknown task sets ($T^{\text{train}}, T^{\text{test}}$). We employ five datasets covering diverse domains:

- **HMDB51** (Kuehne et al., 2011) & **UCF101** (Kay et al., 2017): Standard benchmarks for human actions. We use the splits from (Zhang et al., 2020) (HMDB: 31/10 train/test classes; UCF: 70/21).
- **SSv2** (Goyal et al., 2017): Focuses on fine-grained object interactions. We adopt the split from (Cao et al., 2020) (64/24 classes).
- **NTURGBD** (Shahroudy et al., 2016) & **Diving48** (Li et al., 2018): For NTU, we select two-person interactions (82 classes remaining). For both datasets, we generate random 70/30% train/test splits (NTU: 57/25 classes; Diving48: 32/15 classes).

Table 6.1 Combined results for SAFSAR in 5-way 1-shot and 5-shot settings. Green/red values indicate improvement or decline over the Softmax baseline. Values marked with (*) denote models trained for 1K iterations to prevent overfitting.

OS-Method	FS ACC		OS ACC		AUROC		AUPR		OSCR	
	1-s	5-s	1-s	5-s	1-s	5-s	1-s	5-s	1-s	5-s
<i>Diving48</i>										
Softmax	63.49	74.12	64.16	65.36	68.48	71.49	69.71	72.07	59.86	66.15
EOS	64.43	72.80	62.91	64.94	68.64	74.60	69.71	75.06	60.36	67.43
GC	65.01	76.32	64.89	69.36	70.18	75.82	71.38	76.01	60.63	69.13
FR-Disc	68.83	78.58	66.22	70.29	71.28	76.55	72.28	76.32	63.46	71.04
<i>SSv2</i>										
Softmax	62.11	74.08	62.29	69.61	70.39	77.05	73.02	78.80	60.72	69.35
EOS	62.97	73.84	65.08	69.90	71.56	79.60	73.74	81.18	61.53	70.38
GC	62.47	76.24	64.06	70.89	69.20	77.62	70.47	76.92	59.26	70.20
FR-Disc	63.37	77.88	66.56	73.52	72.18	81.56	74.98	82.96	62.12	73.18
<i>NTURGBD</i>										
Softmax	88.31	91.58	79.90	81.45	87.76	91.44	88.35	91.30	81.45	84.83
EOS	87.63	91.86	80.10	82.18	88.17	91.89	88.96	92.19	81.34	85.11
GC	89.07	92.40	81.78	81.51	89.30	89.37	88.94	88.63	81.93	84.03
FR-Disc	89.97	95.54	82.95	86.53	89.78	94.31	89.82	94.26	83.12	88.31
<i>HMDB51</i>										
Softmax*	65.29	79.68	64.44	72.40	70.76	81.79	73.87	83.33	62.23	74.26
EOS*	69.19	80.18	66.23	72.48	74.60	81.45	77.13	83.50	65.74	74.61
GC*	62.85	76.74	60.19	64.30	68.99	76.45	71.87	78.93	60.49	70.42
FR-Disc	72.38	85.17	68.87	76.99	77.48	87.94	80.30	89.75	68.79	80.15
<i>UCF101</i>										
Softmax*	95.04	98.32	80.59	91.25	94.55	98.03	95.04	98.32	88.31	91.57
EOS*	94.84	98.78	81.30	89.32	95.18	98.31	95.73	98.50	88.47	91.97
GC*	79.98	86.33	59.88	57.49	75.15	81.91	77.29	82.99	70.85	77.43
FR-Disc	95.72	99.28	86.82	91.52	95.19	98.89	96.23	99.08	89.01	92.49

6.5.2 Implementation Details

Following standard FS-AR protocols, we evaluate on 5-Way, 1-Shot and 5-Shot tasks, sampling 8 frames per video. We use the original hyperparameters for STRM (Thatipelli et al., 2022). For SAFSAR (Tang et al., 2024), we re-implemented the model using VideoMAE weights pre-trained on Kinetics-400 (Tong et al., 2022). For the explicit thresholding metrics (OS ACC), we fix $\tau = 0.5$. Other metrics (AUROC, OSCR) are threshold-independent. Additional implementation details are provided in the Supplementary Material.

6.6 Results

Table 6.1 and Table 6.2 report the results for SAFSAR and STRM, respectively, across all five datasets and the four open-set techniques. In Section 6.6.1 we discuss the result of the Softmax baseline, while in Section 6.6.2 we analyze the performance of the different Open-Set techniques.

Table 6.2 Results for STRM in 5-way 5-shot settings. Green/red values indicate improvement or decline over the Softmax baseline.

OS-Method	FS ACC	OS ACC	AUROC	AUPR	OSCR
<i>Diving48</i>					
Softmax	74.74	52.91	58.92	63.25	63.61
EOS	73.58	60.51	67.01	69.02	65.63
GC	73.22	53.76	71.71	72.56	66.09
FR-Disc	77.74	64.92	75.73	75.72	70.00
<i>SSv2</i>					
Softmax	65.25	53.98	61.94	66.33	60.77
EOS	63.61	59.91	64.29	69.09	60.65
GC	43.05	50.68	58.15	62.45	47.06
FR-Disc	65.51	58.99	70.82	73.59	62.52
<i>NTURGBD</i>					
Softmax	95.14	50.85	74.69	78.26	80.28
EOS	92.08	75.80	85.74	86.32	83.30
GC	85.71	59.29	80.27	80.02	76.84
FR-Disc	93.28	80.76	92.27	91.95	86.11
<i>HMDB51</i>					
Softmax	75.92	52.18	70.49	65.29	67.34
EOS	75.18	64.99	75.18	70.58	69.33
GC	71.14	51.10	78.85	73.85	65.96
FR-Disc	75.50	56.87	77.59	75.20	69.11
<i>UCF101</i>					
Softmax	95.48	51.20	86.79	89.39	87.40
EOS	95.05	75.71	90.35	92.42	86.87
GC	95.58	63.21	93.35	95.34	83.71
FR-Disc	95.76	63.47	93.60	94.56	88.37

In our experiments, we observed that for SAFSAR, the Softmax Baseline (as well as the EOS and Garbage Class methods) exhibits a performance decline on the HMDB51 and UCF101 datasets if training is extended. This behavior is suggestive of overfitting; while training metrics saturate almost immediately, validation results show a gradual deterioration. This trend may be attributed to the relatively low class cardinality and predominantly spatial nature of HMDB51 and UCF101, combined with the strong action-representation capabilities of the pretrained VideoMAE backbone. In these cases, the model rapidly learns to recognize the training actions and may over-specialize its weights for episodic tasks involving *only* training classes. To ensure a fair evaluation and maintain generalization, we limited the training to 1K iterations for SAFSAR on HMDB51 and UCF101 for the Softmax, EOS, and GC techniques in Table 6.1. In contrast, the FR-Disc technique did not require early stopping, as it exhibited no signs of overfitting in our experiments.

We also do not report performance for STRM on the 1-Shot task in Table 6.2 because in this scenario STRM suffers of feature bias (Qin and Liu, 2024).

6.6.1 Evaluating the Softmax Baseline

Logits normalization worsens open-set recognition. Previous works (Bisgin et al., 2024; Vaze et al., 2021) have shown that MLS outperforms MSS in rejecting unknown queries in the image domain, as the Softmax normalization hides magnitude information. We first investigated this claim in the spatio-temporal setting as well. In Figure 6.3a we compare the open-set performance of MLS and MSS for SAFSAR’s Softmax Baseline (1-Shot) on the SSv2 dataset. The result supports the claim in this domain too: MLS consistently outperforms MSS, underscoring the importance of preserving the logits magnitude for representing prediction confidence. We did not conduct this comparison for STRM because MLS requires bounded logits, such as the cosine similarities produced by SAFSAR. In contrast, STRM produces unbounded logits (negative norms of feature differences). While normalization based on training set ranges is possible, it would necessitate additional threshold tuning; to ensure a fair comparison and avoid extra hyperparameter dependencies, we opt for the most compatible scoring method for each architecture. Consequently, we employ MLS for SAFSAR and MSS for STRM with the Softmax baseline and EOS techniques.

Closed-set models can be strong competitors in open-set tasks. In Table 6.1 and Table 6.2 we report the performance of respectively the SAFSAR and the STRM models combined with the considered OS techniques across datasets. We first observe that the Softmax Baseline achieves competitive open-set metrics. This is somewhat unexpected, as unknown queries can still resemble support examples and lead to confident false positives. However, this also aligns with the claim made by (Vaze et al., 2021), which suggests that a classifier’s ability to make “none-of-the-above” decisions (i.e., to detect unknown queries) is strongly correlated with its accuracy on the closed-set classes, likely due to their ability to learn discriminative, well-separated features. This is particularly evident on the easiest dataset UCF101 where the overall performance is very high, especially for SAFSAR, and the gains offered by OS techniques over the closed-set baseline are smaller.

Are closed- and open-set performance correlated? To investigate this finding in this spatio-temporal task, we conducted an experiment similar to (Vaze et al., 2021). In 6.3b we plot the closed- and open-set metrics in a scatter plot, for SAFSAR’s and STRM’s Softmax Baseline on all datasets. We first notice, for each model, a linear correlation between these two metrics across datasets, with Pearson coefficients of 0.84 for STRM, and 0.99 and 0.97 for 1- and 5-Shot SAFSAR respectively. Moreover, SAFSAR’s equal or better closed-set performance is coupled with a consistently better open-set performance with respect to STRM: this can be noticed when comparing SAFSAR 5-Shot (cross) vs. STRM 5-Shot (triangle) on a same dataset (same color) and noticing that the cross appears to the right

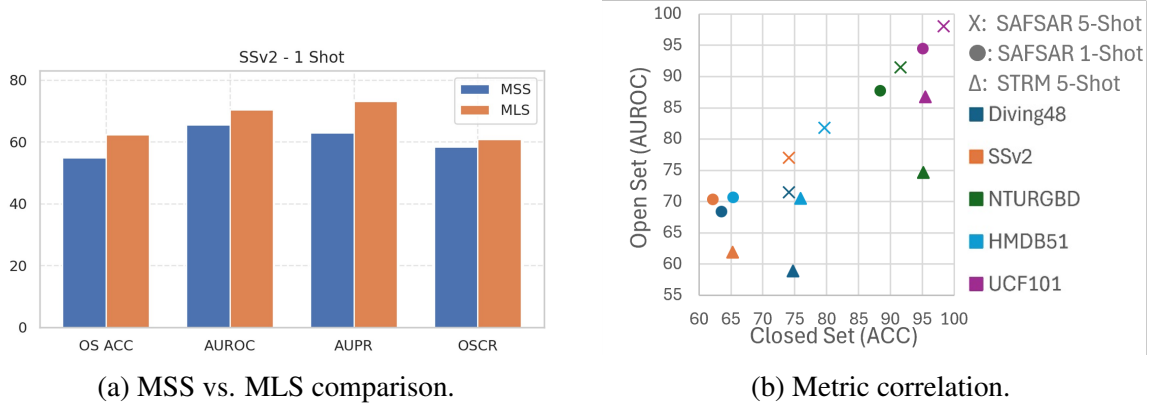


Figure 6.3 (a) Comparison between MSS and MLS for SAFSAR’s Softmax Baseline on SSv2 1-Shot. (b) Global correlation between closed-set and open-set metrics across all five datasets for both SAFSAR and STRM.

or aligned to the triangle for each dataset except NTURGBD, and consistently above the triangle for each dataset. This indicates that even without dedicated open-set mechanisms, a better-trained closed-set classifier can effectively generalize better to distinguish known and unknown classes. We expect this correlation for the prototype-based methods considered here, as both closed- and open-set performance are rooted in the same discriminative feature learning process. Specifically, because the model relies on the same learned representations for both classification and outlier detection, improvements in feature clustering for known classes naturally enhance the ability to isolate unknown samples.

6.6.2 Evaluating Open-Set Techniques

EOS: Benefits of exposing a closed-set model to unknowns. This is a relatively easy technique (requiring only tuning the loss weight σ_{EOS}) to expose the model to unknown queries, where it forces the output probabilities $\{\hat{p}_{ij}\}_{j=1}^K$ to have equal values $1/K$ over all classes. When EOS is applied to SAFAR in Table 6.1, it occasionally causes minor drops in FS ACC (-1.32% on Diving48 5-Shot), but consistently improves OSCR metric across almost all datasets for both 1- and 5-Shot tasks. EOS benefit is larger on the less powerful STRM in Table 6.2, providing improvements of $\sim 10\text{--}25\%$ on the OS ACC on NTURGBD, UCF101 and HMDB51, and an average $\sim 5 - 10\%$ gain on the AUROC and AUPR metrics. As for SAFSAR, the gain on the open-set metrics comes with a minor cost on the FS ACC, with the highest reduction of $\sim 3\%$ on NTURGBD. This result shows that good open-set rejection improvements can be obtained with this simple technique on more

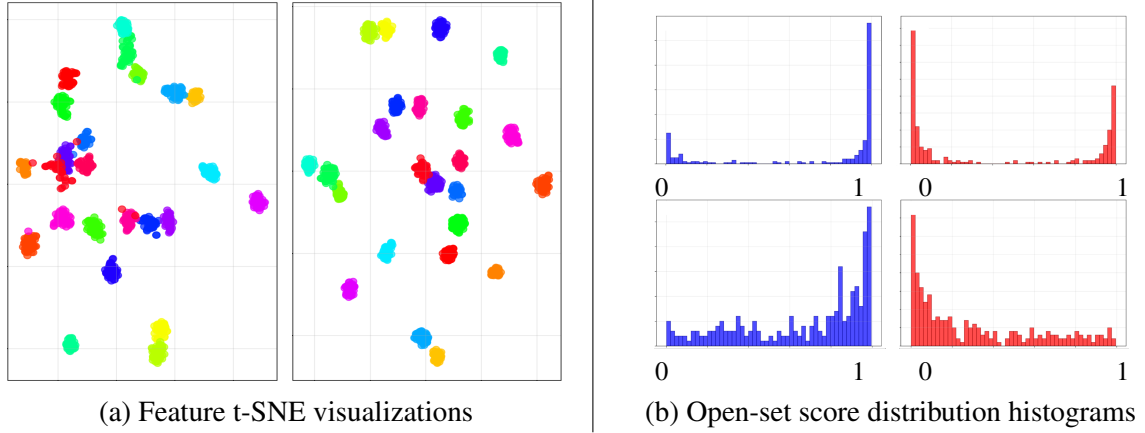


Figure 6.4 (a) t-SNE of video features $\phi(x_i)$ for the baseline (left) and FR-Disc (right). (b) Score histograms for known (blue) and unknown (red) queries, comparing the baseline (top) and FR-Disc (bottom). Our method yields tighter feature clustering and assigns significantly lower confidence to unknown samples.

or less powerful prototype-based models as SAFSAR and STRM, with a minor cost on the closed-set accuracy.

Garbage Class: Difficulty of learning a prototype for unknowns. The Garbage Class (GC) technique exhibits high sensitivity to both dataset characteristics and backbone architecture. For SAFSAR, we observe a clear dichotomy: while GC provides gains on complex spatio-temporal datasets like SSv2 and Diving48, it consistently degrades performance on datasets with strong spatial biases, such as HMDB51 and UCF101. On our experiments we noticed that on these smaller datasets, the VideoMAE backbone rapidly reaches 100% training accuracy, signaling a collapse into overfitting (see Supplementary Materials for the corresponding convergence plots). We hypothesize that in these regimes, the “garbage” category fails to generalize because it only learns to model the “none-of-the-above” samples for that specific training set, memorizing static scene cues rather than a general concept of “unknown” actions.

Unlike SAFSAR, STRM suffers a catastrophic drop in Few-Shot Accuracy (10%-20%) on NTURGBD and SSv2 when the GC is added, alongside a general decline in all open-set metrics. This suggests that the explicit modeling of an auxiliary class interferes with STRM’s specific spatio-temporal alignment mechanism, likely by distorting the prototype space during the meta-learning phase.

FR-Disc: Effectiveness of query-prototype feature residuals. Among the considered techniques, FR-Disc consistently outperforms on all datasets the Softmax Baseline and the other techniques in both 1- and 5-Shot, closed-set and open-set metrics. On SAFSAR

(Table 6.1), FR-Disc improves few-shot accuracy in the 5-Shot task by 3.80%–5.49% (0.96% for UCF101). Regarding open-set performance, the OSCR metric increases by 3.48%–5.89% in the 5-Shot scenario and by 1.40%–6.56% in the 1-Shot scenario, with UCF101 showing consistent gains of approximately 0.70%–0.92%. When applied to STRM, as shown in Table 6.2, this is the only open-set technique that consistently improves the OSCR metric from the Softmax Baseline in all datasets, with a minor cost of respectively 1.86% and 0.42% of FS ACC on NTURGBD and HMDB51, improving the FS ACC for the other datasets. This performance gain may be attributed to the fine-grained discrimination capabilities acquired by the Discriminator. There are achieved by the coupling of (i) the known/unknown classification training protocol, that asks to distinguish a query from the most similar known class, and (ii) the residual-based discrimination which, by emphasizing relative variations in the video pair, helps to catch fine distinctions in the feature representation.

Improved feature representations. Figure 6.4 illustrates how the SAFSAR model benefits from the proposed FR-Disc technique through two qualitative visualizations. First, Figure 6.4a compares t-SNE (Maaten and Hinton, 2008) projections of the video features $\phi(x_i)$, where each color represents a distinct class. To ensure a fair comparison, both the Softmax baseline and FR-Disc are evaluated using the identical set of 1000 SSv2 test queries. Our method produces visibly more compact clusters with significantly improved inter-class separability.

Second, Figure 6.4b displays the open-set score distributions for the same 1000 known (Blue) and 1000 unknown (Red) queries across both configurations. While the Softmax baseline exhibits a high degree of over-confidence—manifesting as a large "acceptance" spike at 1.0 for unknown samples—our FR-Disc method effectively suppresses this peak. The resulting scores are better distributed, allowing for a clearer distinction between known classes and outliers.

6.7 Conclusion

This work addresses a critical gap in evaluating few-shot action recognition (FS-AR) methods under open-world conditions, where models must recognize known classes while rejecting unknowns. We introduce the first standardized benchmark for few-shot open-set action recognition (FSOS-AR), adapting five established datasets to represent open-set tasks. By considering a meta-learning prototype-based approach, we demonstrate that representative closed-set models such as SAFSAR and STRM can be effectively adapted to open-set tasks using techniques from image-based recognition, without sacrificing closed-set accuracy.

Notably, for this category of models, we observe that strong closed-set performance often correlates with open-set robustness. Our evaluation identifies FR-Disc as the most reliable explicit OS technique, consistently outperforming alternatives across datasets. Conversely, we observe unstable behavior in the GC method, which we attribute to the difficulty of learning a representative prototype for unknowns. The implicit EOS technique proved to be a good trade-off between simplicity and effectiveness. Overall, our analysis aims to establish FSOS-AR as an essential research direction to bridge the gap between controlled experiments and real-world action recognition systems.

Additional qualitative examples, per-dataset results, training convergence plots, and the complete dataset splits are provided in Appendix A.

Chapter 7

The Impact of Compositionality in Zero-Shot Multi-Label Action Recognition for Object-Based Tasks

Abstract

Addressing multi-label action recognition in videos represents a significant challenge for robotic applications in dynamic environments, especially when the robot is required to cooperate with humans in tasks that involve objects. Existing methods still struggle to recognize unseen actions or require extensive training data. To overcome these problems, we propose Dual-VCLIP, a unified approach for zero-shot multi-label action recognition. Dual-VCLIP enhances VCLIP, a zero-shot action recognition method, with the DualCoOp method for multi-label image classification. The strength of our method is that at training time it only learns two prompts, and it is therefore much simpler than other methods. We validate our method on the Charades dataset that includes a majority of object-based actions, demonstrating that – despite its simplicity – our method performs favorably with respect to existing methods on the complete dataset, and promising performance when tested on unseen actions. Our contribution emphasizes the impact of verb-object class-splits during robots’ training for new cooperative tasks, highlighting the influence on the performance and giving insights into mitigating biases. Dataset splits and code are publicly available on the project’s repository ¹.

7.1 Introduction

Addressing multi-label action recognition in videos represents a significant challenge for many real-world robotic applications. Robots play a key role in modern workplaces where they are often used for repetitive tasks. To cooperate with humans, robots are expected to actively interact in dynamic and poorly structured environments, where a priori definition of objects and actions is non-trivial (Ayub and Wagner, 2020).

Most recognition methods result impractical for these purposes since they require that all target classes are known beforehand, and an extensive number of examples is provided for each class at training time (Chiatti et al., 2020). For this reason, in many applications, robots

¹<https://github.com/hsp-iit/DualVCLIP>

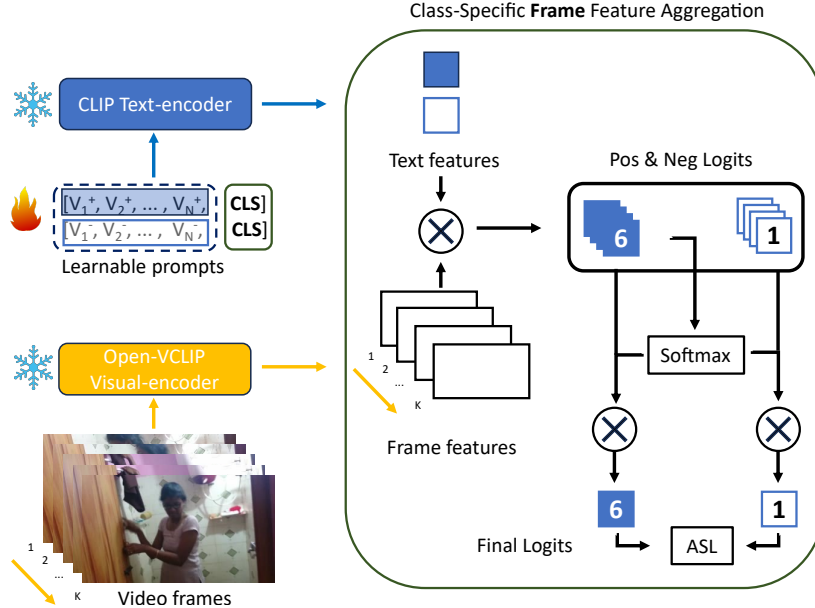


Figure 7.1 **Overview of our proposed Dual-VCLIP for multi-modal multi-label action recognition.** It has two main components: a video encoder and a textual encoder. DualCoOp learns a pair of positive and negative prompts to quickly adapt pretrained vision-text encoders to the Multi Label Recognition (MLR) task. For each class, two prompts generate two contrastive (positive and negative) textual embeddings as the input to the text encoder. Furthermore, we propose Class-Specific Frame Feature Aggregation to first project each frame’s feature to the textual space and then aggregate the temporal logits by the magnitude of class-specific semantic responses. During training, we apply the asymmetric loss from (Sun et al., 2022) to optimize learnable prompts while keeping other network components frozen.

need to be incrementally trained to recognize the specific objects (Ayub and Wagner, 2020; Chiatti et al., 2020) or actions (Wang et al., 2021c) needed for a specific joint task, using only limited set of visual demonstrations provided by the final user. Conventional action recognition methods often focus on learning the visual modality alone, neglecting additional information sources. Recent works leverage large-scale pretrained foundation models due to their adaptability to transfer knowledge on a different set of downstream tasks. Through natural language processing and multimodal information, these methods learn an alignment between textual and visual spaces to compensate for limited labeling. The generalization capability of CLIP-based architectures becomes fundamental for robotics scenarios where training on much smaller datasets while maintaining nearly the same performances becomes desirable (Ayub and Wagner, 2020; Berti et al., 2022; Chiatti et al., 2020; Wang et al., 2021c).

While Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) has demonstrated impressive zero-shot learning abilities for image understanding, there has been yet limited exploration of CLIP’s potential for zero-shot video recognition. To extend image-text models to video, recent work has focused on the visual domain by aggregating temporal information with new modules (Adeli Mosabbebe et al., 2015; Wang et al., 2021b). In addition, most recent action recognition methods focus on identifying single activities, assuming only one action per video-clip and discarding the potential simultaneous occurrence of simultaneous actions (Lentzas et al., 2022), that can be reductive and damaging for an efficient HRI. Instead, robots need to be incrementally trained to recognize the multiple and specific objects (Ayub and Wagner, 2020; Chiatti et al., 2020) or actions (Wang et al., 2021c) needed for a specific joint task, using only limited set of visual demonstrations provided by the final user (Berti et al., 2022), leveraging the adaptability of foundation models to transfer knowledge to different downstream tasks. In this perspective, Compositional Zero-Shot Learning tries to imitate the human ability to reduce complex stimuli to primitives, learn by discrimination, and generalize to recognize representations of unseen compositions composed of seen concepts (Sucholutsky et al., 2024) based on prototype-based categorization (as opposed to exemplar-based) for categorical inferences (Malaviya et al., 2022).

To overcome these limitations, we propose a new approach called Dual-VCLIP which offers a unified solution for different aspects in zero-shot multi-labelling action recognition tasks, and we investigate preliminary results on our solution for multi-labelling action recognition. In this paper, the proposed method, Dual-VCLIP, is an efficient and minimalist approach that extends the zero-shot multi-labelling object recognition method DualCoOp (Sun et al., 2022) to action recognition in video through OpenVCLIP (Weng et al., 2023), a simple yet effective approach that transforms CLIP into a strong zero-shot video classifier that can recognize unseen actions at inference time.

Figure 7.1 shows an overview of our proposed model. Preliminary results suggest that this framework can achieve competitive results on zero-shot action recognition on Charades (Sigurdsson et al., 2016).

Our contribution is summarized as follows: (1) we propose a new method for zero-shot multi-label action recognition, which remains effective while trained only small set of parameters, and compares favorably against existing methods (2) we stress on the relevance of class-splits based on verb-object pairs to analyze bias in the temporal or spatial dimension.

7.2 Related Work

7.2.1 Multi-Label Action Recognition

Recognizing the behaviors of humans in videos is crucial for Human-Robot Interaction (HRI) across different domains, such as services ([Anagnostis et al., 2021](#)), industrial ([Roitberg et al., 2014](#)), healthcare ([Rodomagoulakis et al., 2016](#)). Although most recent action recognition methods focused on identifying single activities, assuming only one action per video-clip and discarding the potential simultaneous occurrence of simultaneous actions ([Lentzas et al., 2022](#)), can be reductive and damaging for an efficient HRI. Many existing approaches primarily concentrate on single-activity recognition and utilize sigmoid as the output activation function to provide multi-label predictions ([Wang et al., 2021b](#)). Some methods solved multi-label recognition at the feature level, by extracting independent feature descriptors for each activity instead of shared feature representations, and thus capturing correlations among different actions ([Zhang et al., 2021](#)). Weakly-supervised learning approaches ([Adeli Mosabbebi et al., 2015](#)) successfully advanced multi-label learning for video analysis, localizing spatio-temporal discriminative regions inferred by clustering regions of common texture and motion features.

7.2.2 Multi-Label Zero-Shot Action Recognition

Robots are expected to actively interact in dynamic and poorly structured environments, where a priori definition of objects and actions is non-trivial ([Ayub and Wagner, 2020](#)). Most recognition methods are impractical for these purposes as they require prior knowledge of all target classes and extensive training examples for each class ([Chiatti et al., 2020](#)). For this reason, in many applications, robots need to be incrementally trained to recognize the specific objects ([Ayub and Wagner, 2020](#); [Chiatti et al., 2020](#)) or actions ([Wang et al., 2021c](#)) needed for a specific joint task, using only limited set of visual demonstrations provided by the final user ([Berti et al., 2022](#)), leveraging the adaptability of foundation models to transfer knowledge to different downstream tasks. In this perspective, zero-shot recognition aims to classify novel categories during testing. However, multi-label zero-shot action recognition remains relatively unexplored. Kerrigan, et al. ([Kerrigan et al., 2021](#)) reformulated multi-label zero-shot action recognition in terms of pairwise scoring function between confidence scores for each class independently learnt, showing strong performances in the prediction of semantically distinct classes. The same can be afforded exploring relationships between multiple labels with graph structure to learn semantic dependencies between words in a class

name and leveraging information propagation mechanism from the semantic label space to predict unseen class labels (Lee et al., 2018). Wang and Chen (Wang and Chen, 2020) achieved multi-label zero-shot recognition by measuring relatedness scores of action labels in the joint latent visual and semantic embedding spaces. Other approaches consider losses with evidence debiasing constraints to reduce the static bias of video representations, via shared multi-attention model to map only relevant segments of an image to a joint visual-label embedding space (Huynh and Elhamifar, 2020) or via evidential neural networks that estimate multi-action uncertainty to detect novel actions (Zhao et al., 2023).

It is worth mentioning that the methods mentioned above do not adapt the foundation model to the distribution of the dataset by looking at a portion of the classes.

7.2.3 Compositionality in Action Recognition

Human activities involve a series of interactions occurring over time and with various objects in space (Baradel et al., 2018; Ji et al., 2020). This is particularly crucial in Human-Robot interaction scenarios where tasks often involve specific objects. However, existing action recognition methods still struggle with novel combinations of known actions and objects (Materzynska et al., 2020), posing challenges for robotic applications. Enhancing current models with compositional generalization could be a crucial and promising avenue for improving action understanding. The work of Sigurdsson et al. (Sigurdsson et al., 2017) provides a comprehensive analysis to identify the key factors contributing to significant improvements in action understanding, emphasizing the role of objects, verbs, and sequential reasoning, providing fine-grained understanding of objects when combined with temporal reasoning. Their analysis sheds light on specific training methodologies, like avoiding treating actions as singular events occurring within videos, leveraging hierarchical event decomposition to facilitate few-shot action recognition (Ji et al., 2020). For instance, Materzynska et al. (Materzynska et al., 2020) introduced a compositional framework wherein there is no overlap between the verb-noun combinations in the training and testing datasets. This compositional approach holds significant promise for the multi-label fine-grained video classification task

7.3 Background

Vision-language models pretrained on large-scale datasets such as CLIP (Radford et al., 2021) align the feature of visual and textual representations, thus allowing for zero-shot

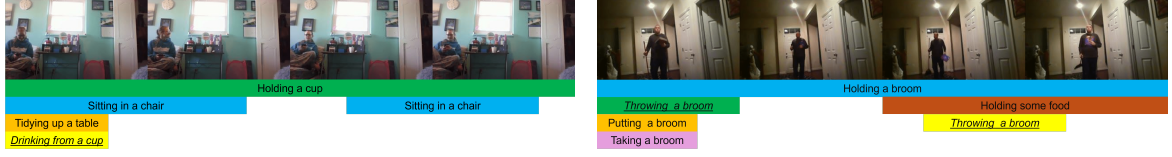


Figure 7.2 **Inference per-frame with our method Dual-VCLIP on two examples from the Charades' test set.** Underlined classes are unseen: they did not appear at training time. Classes that appear only in one frame are discarded.

image classification. Formally, given an image x and a set of classes as textual descriptions y , the probability for each class $y_j \in y$ is:

$$P(y_j|x) = \frac{\exp(\text{sim}(\text{CLIP}_V(x), \text{CLIP}_T(y_i)))}{\sum_j \exp(\text{sim}(\text{CLIP}_V(x), \text{CLIP}_T(y_j)))} \quad (7.1)$$

where CLIP_V is the visual encoder of CLIP, CLIP_T is the textual encoder of CLIP, “sim” is a similarity function such as the cosine similarity and “exp” is the exponential function.

The original image CLIP model does not have the ability to aggregate temporal relations among video frames and thus is not suitable to identify actions. OpenVCLIP (Weng et al., 2023) adapts CLIP framework to sequences of images without additional parameters, but just by expanding the temporal attention view for every self-attention layer. Formally, the self-attention layers of CLIP are computed as follows:

$$y_t = \text{Softmax} \left(\frac{q_t K_t^T}{\sqrt{d}} \right) V_t \quad (7.2)$$

and the modified version for OpenVCLIP as follows:

$$y_t = \text{Softmax} \left(\frac{q_t K_{(t-1) \sim (t+1)}^T}{\sqrt{d}} \right) V_{(t-1) \sim (t+1)} \quad (7.3)$$

where d refers to the dimension of the model, q_t refers to the query vector of the t -th frame, K_t^T is the transpose of the matrix composed of key vectors and V_t is the matrix composed of value vectors in the t -th frame. The subscript $(t-1) \sim (t+1)$ in Eq. (7.3) denotes the concatenation of the key (respectively value) vectors gathered over the local temporal window spanning the frames from $t-1$ to $t+1$; in other words, $K_{(t-1) \sim (t+1)}^T$ and $V_{(t-1) \sim (t+1)}$ extend the attention of frame t to its neighbouring frames, which is precisely how OpenVCLIP injects temporal information without adding new parameters. The global temporal information is obtained from the stacking of the self-attention layers. The resulting

architecture is then fine-tuned on a large-scale action recognition dataset such as Kinetics-400 (Kay et al., 2017). The resulting weights are averaged with the original CLIP weights using a technique named Interpolation Weights Optimization (Ilharco et al., 2022), that allows a model to learn new concepts while preserving original knowledge.

With CLIP we obtain a similarity matrix between images and textual descriptions, that can be used to obtain the probability for a class in a certain image with a softmax function, like in Eq. (7.2). Anyway, the softmax is not designed for the multi-label classification, and a threshold on the logits is not possible because each class may need a different value, decreasing the performance on multi-label zero-shot classification. DualCoOp (Sun et al., 2022) proposes to learn a positive and a negative prompt and to consider a class present in an image only if the similarity between the image and the positive prompt is higher than the similarity between the image and the negative prompt. In the supervised setting, a positive and a negative prompts are learned for each class, while in the zero-shot setting a positive and a negative prompts are learned and given to the textual encoder together with the name of the class. Since a single image can contain different classes, the feature of that image is required to contain information about different classes in different portion of the image, thus obtaining sub-optimal performance since spatial information is reduced and different classes representations are mixed. To overcome this issue, (Sun et al., 2022) reformulate the last multi-headed attention pooling layer of the visual encoder in CLIP: they do not pool the visual attention map, but they project the visual feature of each region to the textual space. This allows to create class-specific region features, that can correctly represent different classes in different portion of the image.

7.4 Problem Definition

In zero-shot setting, the model is trained on a source dataset D_{train} with classes Y_{train} and tested directly on a target dataset D_{test} with classes Y_{test} , where $Y_{train} \cap Y_{test} = \emptyset$.

We also consider the setting where Y_{train} and Y_{test} are part of the same dataset and the name of the classes have the form *subject + object + verb*. This allow to split the classes in a different way, for example by masking out certain objects or verbs at training time to see how this affects the performances of the model and to understand its possible temporal or spatial biases.

7.5 Method

In this paper, we introduce a novel method for zero-shot multi-label action recognition, named DualVCLIP. Our method learns a prompt that adapt the knowledge of the foundation model to the distribution of the target dataset, while exhibiting competitive performance on both seen and unseen classes. DualVCLIP, as DualCoOp (Sun et al., 2022), learns a pair of “prompt” contexts as two learnable sequences of word vectors, to provide positive and negative contexts for a given category name y . In Figure 7.1, we provide a schematic overview of our method.

Formally, to check if an action y is contained in a video x , we extract the features of the video x with OpenVCLIP as $v = \text{OpenVCLIP}(x)$. Then we create the positive and negative class-prompts $\hat{y}_j^+ = [P^+, y_j]$ and $\hat{y}_j^- = [P^-, y_j]$, that are the concatenation of the learnable positive/negative prompts P^+ and P^- and the class name y_j . Then we encode the class-prompts with the textual encoder of CLIP:

$$\hat{t}_j^+ = \text{CLIP}_T(\hat{y}_j^+) \quad (7.4)$$

$$\hat{t}_j^- = \text{CLIP}_T(\hat{y}_j^-) \quad (7.5)$$

and we compute the similarity between the feature extracted from the video and the positive and negative encoded class-prompts:

$$S_{j,t}^+ = \text{sim}(V, \hat{t}_j^+) \quad (7.6)$$

$$S_{j,t}^- = \text{sim}(V, \hat{t}_j^-) \quad (7.7)$$

where t indicates the t -th frame of the sequence and “sim” is a similarity function, e.g., the cosine similarity.

Finally, for each class, we perform aggregation of all temporal logits, in which the weight for each logit is determined by its relative magnitude:

$$S_j^+ = \sum_t (\text{Softmax}(S_{j,t}^+) \times S_{j,t}^+) \quad (7.8)$$

$$S_j^- = \sum_t (\text{Softmax}(S_{j,t}^-) \times S_{j,t}^-) \quad (7.9)$$

We call this processing as *Class-Specific Frame Feature Aggregation*. It can be noted that this process is similar to the one of DualCoOp, but we exploit the temporal dimension instead

of the spatial dimension. At training time, the similarities are given to an asymmetric loss module, as in DualCoOp (Sun et al., 2022), while at inference time a class j is predicted if $S_j^+ > S_j^-$. We freeze the weights of all the components except for the learnable prompts, that has a reduced number of parameters (64x512 in our experiments).

7.5.1 Experiments

Dataset. We conduct experiments on Charades dataset (Sigurdsson et al., 2016) to evaluate zero-shot multi-label action recognition. It contains 66.5k annotations for 157 actions and it is divided into 7985 and 1863 video-clips for training and testing sets, respectively. This dataset is challenging for multi-label action recognition, since it contains clips from hundreds of people recording video-based human activity in their own home. The average duration of video-clips is 30s and an average number of actions involved in a video-clip is 6.8. Since the public split is thought on an instance-level (all the 157 classes appears in both training and testing set) and cannot be directly applied to zero-shot tasks, we define three splits at the class-level. In other words, we define which of the 157 classes the model sees from the training split at training time, such that at testing time we have that the model has never seen a certain number of classes.

We define three different splits for the purposes of this work:

- **Random Split:** we randomly choose half of the classes in the test dataset, irrespective of a fair balance between verbs and objects. Specifically, for this experimental setting, we consider Y_{train} and Y_{test} respectively contain 50% and 50% of the total number of classes in the dataset.
- **Verbs Split:** we first cluster classes according to the verb in the label C_{verb} and then randomly choose 50% of the clusters C_{verb} . All the classes where those verbs appear, belong to the test set Y_{test} , the remaining to the training set Y_{train} .
- **Objects split:** we first cluster classes according to the object in the label C_{obj} and then randomly choose 50% of the clusters C_{obj} . All the classes where those objects appear, belong to the test set Y_{test} , the remaining to the training set Y_{train} .

Table 7.1 reports splits' details on the following evaluation protocols.

Evaluation protocols. We evaluate all methods with both zero-shot setting (only 50% of seen classes) and on all the entire Charades dataset. As done in previous works (Sun et al., 2022), we report the mAP value on the test set for each split.

	<i>Unseen Verbs</i>	<i>Unseen Objects</i>	<i># Unseen Classes</i>
Random Split	26.3% (10/38)	8.1% (3/37)	79
Verb Split	47.3% (18/38)	5.4% (2/37)	84
Object Split	34.2% (13/38)	48.6% (18/37)	82

Table 7.1 **Splits’ Details.** Verbs and objects in testing classes are split into Seen (appeared during training phase) and Unseen (did not appear in training phase). About 50% of classes are used for zero-shot.

Implementation. We use the ViT-B-16 version of CLIP for all the experiments. The OpenVCLIP model is obtained from a weight interpolation with the original CLIP’s weights using a patching ratio of 0.5. The model is trained on 4 TeslaV100 GPUs with a total batch size of 64 until convergence, with one warm-up epoch. We use SGD as optimizer, with a learning rate that starts from 1e-3 and follows a cosine annealing scheduling. We set the number of context vectors to 64. We used 16 frames for inference and training with a resolution of 224.

Baseline. To evaluate the effectiveness of DualVCLIP in the zero-shot setting, we compare it with the following baselines: (1) CLIP-Hitchiker (Bain et al., 2022), (2) VideoCOCA (Yan et al., 2022), (3) OpenVCLIP (Weng et al., 2023) and (4) MSQnet (Mondal et al., 2023). CLIP-Hitchiker and VideoCOCA are two methods that are directly tested on the test set of charades without any additional training. We also tested MSQNet using the code released by the authors. It is a method for zero-shot action recognition that requires to be trained on part of the classes, similarly to our method. Unfortunately, we were not be able to reproduce the performance reported in their work, while achieving consistently lower results on our train-test. We believe this may be due to the fact that we could not find the details of the train-test split they used in their experiment, but further investigation is needed to confirm these experiments. Moreover, we compare with OpenVCLIP adapted to multi-labelling, where we use a threshold of 0.5 on the similarity scores between videos and classes instead of a softmax over all classes.

7.6 Results

7.6.1 Comparison with SOTA

In Table 7.2 we report the mAP results in zero-shot- when only 50% of classes has been seen during training- and on the full Charades testing set, for the different action recognition

methods. The results show that our method has lower performance with respect to the competitor MSQNet. However, these results are not comparable since training the two methods on their splits was not possible (not available), while achieving consistently lower results with MSQNet on our train-test. Instead, when homogeneous comparison is possible (on the full testing dataset), it is remarkable that our method returns competitive mAP values with respect to the SOTA methods. This shows the convenience of adding a learnable context vector to improve the performance in generalized setting (compared to OpenVCLIP). It is noteworthy that our method learns a limited number of parameters with respect to other methods- we learn just two prompts that have 64 context tokens each.

Method	mAP ZSL (50% Seen)	mAP (Full testing set)
CLIP-Hitchhiker* (Bain et al., 2022)	-	21.1
VideoCoCa* (Yan et al., 2022)	-	25.8
MSQNet**	30.9	-
OpenVCLIP adapted	-	18.82
Dual-VCLIP (ours)	16.65	22.98

Table 7.2 **Comparison of different zero-shot action recognition methods on Charades.** (*,** values reported from literature: * unavailable code, ** unavailable splits).

7.6.2 Compositionality

To implement this approach effectively in Human-Robot interaction scenarios, it is crucial to ensure its capability in fine-grained action recognition. This entails the ability to discern between actions that involve the same object, such as “passing a box”, “looking at a box,” and “opening a box.” The Charades dataset serves as a suitable resource for conducting this detailed analysis. Notably, the Charades dataset comprises 157 action classes, with categories that extend beyond atomic actions. Specifically, 146 classes can be deconstructed into “Verb+Object” combinations, while 11 classes involve only a verb. Upon scrutinizing the semantics of these classes, we identified 38 verbs and 37 objects appearing in various combinations. To clarify the knowledge acquired by our method, we calculated the mean Average Precision (mAP) across five distinct subsets of the testing set of the Random Split. These subsets distinguish between scenarios where verbs or objects are either recognized or remain unseen.

The results reported in Table 7.3 clearly show that discriminating among verbs or objects poses differing levels of difficulty for the method. By comparing the mAP in the Object Seen (OS) scenarios, the method achieves higher performance when the verb is unknown (VU). Conversely, when the object is unknown (OU column), the knowledge of the verb does not significantly enhance performance (VS-OU vs. VU-OU). Moreover, comparing scenarios where one of the elements is unknown (VU-OS vs. VS-OU), we observe a significantly higher mAP when object is seen, despite the comparable number of classes in the two scenarios (8 vs. 6). However, in cases where the action does not involve an object (VU-NO), our algorithm demonstrates good generalization, outperforming the other conditions, especially the case VU-OS. This analysis may suggest that our method exhibits strong generalization to new verbs but in scenarios that require compositionality (i.e., discriminating between actions performed on different objects), spatial generalization dominates temporal generalization. It implies that objects play a crucial role in discrimination, as discriminating between actions like “closing a book” and “closing a laptop” is easier than distinguishing between “opening a bag” and “holding a bag”. This hypothesis can be further supported by the results shown in Fig 7.3. We generated the confusion matrix for the unseen classes and conducted an analysis, grouping them based on objects or verbs. It is noteworthy that the method struggles in accurately identifying the verbs associated with the object (as shown in panel (A)), while it effectively exploits the object to distinguish classes with equal temporal dimensions.

	OS	OU	NO
VS	16.41 (59)	9.46 (6)	-
VU	25.21 (8)	7.07 (1)	27.47 (5)

Table 7.3 **mAP for different sub-sets of the testing set.** O-object, V-verb, S-seen, U-unseen, N-not present. The parenthesis contains the respective number of classes.

7.6.3 Spatial vs Temporal Generalization

Based on the findings discussed in the previous section, we wanted to explore a different training approach to assess whether the newly acquired prompt exhibits enhanced or distinct generalization characteristics in spatial versus temporal contexts. Table 7.4 shows interesting results. It clearly demonstrates that (1) the training protocol can impact the precision of the method, and (2) it reaffirms the advantages of exposing the model to more objects (object-augmentation) compared to more verbs (verb-augmentation). However, it remains unclear

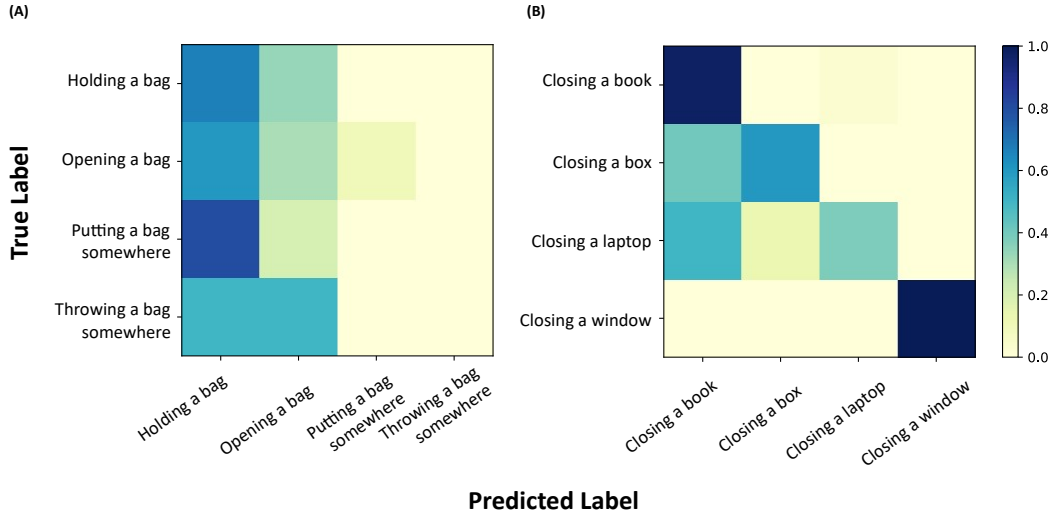


Figure 7.3 **Confusion matrix.** (A) Sub-matrix representing an object-cluster (e.g., 'bag'). (B) Sub-matrix representing a verb-cluster (e.g., 'closing'). For these examples, we binarized the predicted outputs with a threshold equal to 0.5 on the confidence value.

Table 7.4 mAP in zero-shot (*ZSL*) and generalized zero-shot (*GZSL*, testing both on seen and unseen classes) settings for Verb vs Object splits.

Verb Split		Object Split	
<i>ZSL</i>	<i>GZSL</i>	<i>ZSL</i>	<i>GZSL</i>
17.06	21.62	6.84	15.90

whether the focus is primarily on objects or verbs. As our intention is to deploy our module in a robotic application, achieving generalization across different objects is imperative.

7.7 Conclusions

In this study, we introduced a novel unified framework, Dual-VCLIP, designed for zero-shot multi-label action recognition. Leveraging the robust vision-language pretraining from a large-scale dataset, our method addresses multi-label recognition tasks through the incorporation of a lightweight learnable overhead, that is a pair of positive and negative prompts along with the target class name as linguistic input. In addition, to better aggregate temporal visual frame features for each class, we reformulate the original visual attention in the pretraining model as a class-specific frame feature aggregation. Our promising results on the Charades dataset and the comparison with current SOTA methods and baselines demonstrate the effec-

tiveness of our proposed approach. With our analysis on class compositionality, we showed that the learning protocols can play a significant role in the categorization process, especially when the method needs to generalize both in temporal and spatial dimensions, as needed for action recognition. This can suggest the importance of adopting conditional learning protocols, where the prompt is fine-tuned after learning verbs or objects, and the advantage of exploring different temporal aggregation techniques. This aspect holds significant implications for zero-shot/few-shot scenarios in Human-Robot cooperative environments, where robots need to learn new tasks quickly and dynamically. Finding efficient zero-shot/few-shot learning protocols becomes crucial when limited data available for training the robots, when they may encounter new objects, actions, or scenarios that were not part of their initial training data and to use resources in more efficient way. Instead of collecting and labeling large amounts of data for every new task, robots can learn from a few examples or even from textual descriptions, reducing the time and effort required for training. This study also highlights the significance of introducing diagnostic tools and more comprehensive validation mechanisms when assessing these methodologies on intricate datasets such as Charades.

Finally, we acknowledge that this study opens up a number of questions that need further investigation. In the future, we plan to explore advancements to enhance the capabilities of our model. In particular, upcoming work will involve running experiments on multiple splits, validating over additional similar composed datasets, and exploring improvements through conditional learning, by integrating evidence debiasing constraints into the optimization objective to reduce static bias in video representations.

Chapter 8

ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language Tasks

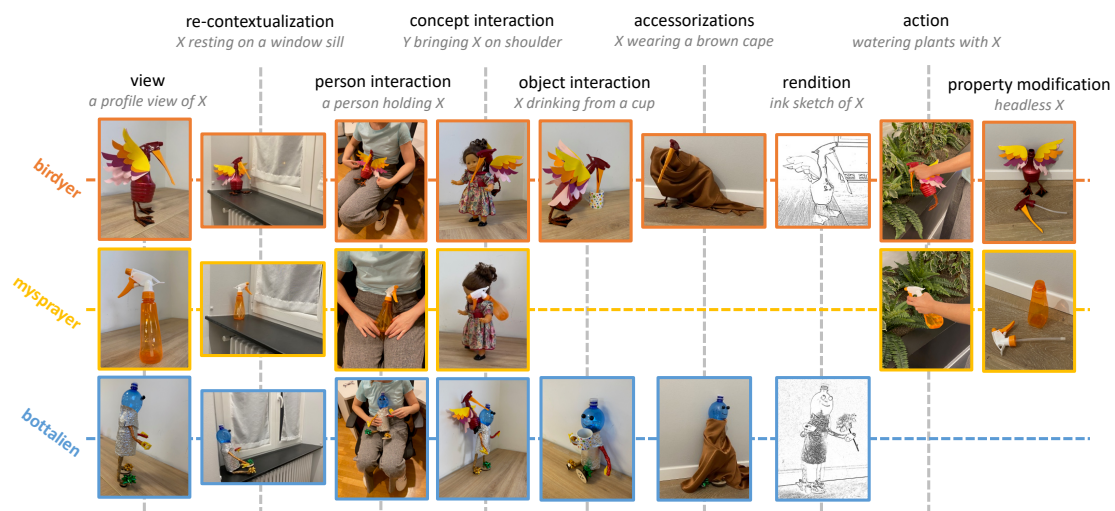


Figure 8.1 **Concept-context structure of the ConCon-Chi dataset.** Excerpt from the dataset structure (see 8.4). Each row represents images of a concept (concept name in Color) and each column images of a context (context kind in Black and context in Grey). The caption for an image is formed by the composition of the context with the concept (replacing X/Y with the concept textual identifier).

Abstract

While recent Vision-Language (VL) models excel at open-vocabulary tasks, it is unclear how to use them with specific or uncommon concepts. Personalized Text-to-Image Retrieval (TIR) or Generation (TIG) are recently introduced tasks that represent this challenge, where the VL model has to learn a concept from few images and respectively discriminate or generate images of the target concept in arbitrary contexts. We identify the ability to learn new meanings and their compositionality with known ones as two key properties of a personalized system. We show that the available benchmarks offer a limited validation of personalized textual concept learning from images with respect to the above properties and introduce ConCon-Chi as a benchmark for both personalized TIR and TIG, designed to fill this gap. We modelled the new-meaning concepts by crafting chimeric objects and formulating a large, varied set of contexts where we photographed each object. To promote the compositionality assessment of the learned concepts with known contexts, we combined different contexts with the same concept, and vice-versa. We carry out a thorough evaluation of state-of-the-art methods on the resulting dataset. Our study suggests that future work on personalized TIR and TIG methods should focus on the above key properties, and we propose principles and a dataset for their performance assessment. Dataset: <https://doi.org/10.48557/QJ1166> and code: https://github.com/hsp-iit/concon-chi_benchmark.

8.1 Introduction

Recent Vision-Language (VL) models for discriminative and generative tasks (Li et al., 2022; Radford et al., 2021; Rombach et al., 2022) excel at associating textual descriptions with images. This resulted in a paradigm shift from closed to open-vocabulary versions of several computer vision tasks. In these settings, the model is not bound to a closed set of predetermined classes but can operate on free-form textual descriptions. However, it can be difficult for a user to formulate a description such that the VL model returns the expected output (Menon and Vondrick, 2023; Stretcu et al., 2023). The problem boils down to finding

a text input whose encoding is close to the visual embedding of the target concept. Since this mapping is typically learned from Web-scale data, crafting effective descriptions for uncommon, novel or specific concepts, is challenging, thus hampering performance of VL models.

To represent this challenge, the “personalized” versions of Text-to-Image Retrieval (Cohen et al., 2022) (TIR) and Text-to-Image Generation (Gal et al., 2023a) (TIG) have been recently proposed. The tasks consist of learning a user-specific *concept* using a few images and then performing retrieval (TIR) or generation (TIG) of such *concept* in a known *context*. Another similar task recently proposed is Zero-Shot Composed Image Retrieval (ZS-CIR) (Baldrati et al., 2023; Saito et al., 2023), the retrieval of a reference image modified according to a relative caption (Vo et al., 2019).

Due to lack of benchmarks for personalization tasks, many works proposed their method alongside a new dataset. However, most of these benchmarks lack two key properties necessary for a thorough evaluation of personalized concept learning: **novel concepts** and **compositional** structure.

Existing datasets use, as concepts, instances of common objects (e.g., in PerVL, clothing items from DeepFashion2 (Ge et al., 2019)). However, these allow to evaluate the learning of new words for known concepts (closely related to synonym matching, or instance identification), rather than new concepts. To simulate a realistic personalization setting, inspired by (Lazaridou et al., 2017), we introduce chimeric concepts: objects created by the union of two unrelated existing concepts (see 8.2).

The number and variability of contexts in which a concept appears is also underrepresented in personalized TIR/TIG benchmarks and this hampers the compositionality assessment of the newly learned concept with known contexts. Conversely, in a personalized TIR dataset multiple concepts should also appear in a same context, to prevent a method from attaining high performance by just attending to the context and disregarding the concept in the query (“context bias”, pointed out also in ZS-CIR benchmarks (Baldrati et al., 2023; Saito et al., 2023)). To avoid both these problems, we formulated a large, varied, set of contexts for each concept, while also ensuring that each context can be composed with several concepts. As a result, the dataset has a concept-context matrix structure (8.1).

Our contributions are as follows:

- We highlight two key properties of personalized textual concept learning from images: learning new meanings and composing them with known ones; we propose the design of ConCon-Chi to model this problem.

- By evaluating on ConCon-Chi we show the limitations of state-of-the-art methods with respect to these properties.
- We release the dataset as a twofold benchmark for personalized TIR *and* TIG.

In the remainder of the paper, we compare related datasets in 8.2; we present ConCon-Chi in 8.3; we present our study for personalized TIR in 8.4 and for personalized TIG in 8.5. We report conclusions in 8.6.

TASK	DATASET	CONCEPT INSTANCES	CONCEPT TYPE	CONCEPTS /QUERY	IMAGES /QUERY	SPLIT	QUERIES	POOL	CONCEPTS	CONTEXTS	CONTEXTS /CONCEPT	CONCEPTS /CONTEXT
pers TIR pers TIG	CONCON-CHI	yes	chimeric & common	≤ 3	6.07	val	42	165	3	20	14.34	2.15
						test-unseen	986	4008	17	101	50.53	8.50
						test	1084	4008	20	101	46.65	9.24
pers TIR	PERVL DF2	yes	fashion	1	1	val	229	229	50	229	5.58	1.00
						test	221	221	50	221	4.42	1.00
pers TIG	DREAMBOOTH CC101	yes	common	1	0	test	750	0	30	35	25	21.43
		yes	common	≤ 2	0	test	3232	0	101	597	36.36	6.15
ZS-CIR	FASHIONIQ	no	fashion	1	1	val (avg)	2005	5179	1442	1994	1.39	1.00
						test (avg)	2039	5179	1454	2030	1.40	1.00
	CIRR	no	open	1	1	val	4184	2297	2165	4157	1.93	1.00
						test	4148	2315	2178	4135	1.90	1.00
	CIRCO	no	open	1	4.53	val	220	123403	220	220	1	1
						test	800	123403	798	796	1.00	1.00

Table 8.1 Comparison of related datasets. PerVL DF2: PerVL DeepFashion2; CC101: CustomConcepts101. For FashionIQ we report the average number of images per split (shirt, dress, toptee). Dataset aspects that fulfill the criteria discussed in 8.3 are in **Bold**. ZS-CIR datasets reported to show the setting difference with personalized TIR/TIG.

8.2 Related Work

Learning Out Of Vocabulary (OOV) words. Learning textual concept representations from images can be seen as the multi-modal version of the NLP task of learning OOV word embeddings from a few examples (Bahdanau et al., 2017) or the word definition (Herbelot and Baroni, 2017; Hill et al., 2016; Ruzzetti et al., 2022). We took inspiration from the Chimera dataset (Lazaridou et al., 2017), which differentiates the problem of learning to associate new words with existing meanings (named entities, synonyms, aliases) from the one of learning new meanings, and models the latter with chimeric words that incorporate two unrelated concepts in a single one.

Personalized Text-to-Image Retrieval (TIR). Cohen *et al.* (Cohen et al., 2022) introduced “Personalized Vision & Language” (PerVL), a setting where the vocabulary of a VL model is expanded with pseudo-tokens whose embeddings are learned from few images of user-specific concepts. The benchmark includes a retrieval and an instance segmentation

dataset obtained by re-annotating images respectively from DeepFashion2 (Ge et al., 2019) and YouTube-VOS (Xu et al., 2018) with concept-context captions. PerVL includes many concept instances, which are however restricted to fashion or common items; moreover, it counts very few contexts per concept and a single concept per context, suffering from context bias. To tackle the proposed benchmark the authors propose PALAVRA, which we evaluate in this study.

The setting of personalized retrieval has been extended to videos (Korbar and Zisserman, 2022; Yeh et al., 2023). Recent work (Parisot et al., 2023) proposed a similar approach to transfer CLIP (Radford et al., 2021) to a downstream image classification task by optimizing the class names.

Zero-Shot Composed Image Retrieval (ZS-CIR). This task can be seen as one-shot personalized TIR, where the reference image is the concept, and the relative caption the context. Thus personalization methods as PALAVRA have been evaluated for ZS-CIR (Baldrati et al., 2023). Similarly, we consider methods as Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023) for evaluation on the proposed personalized TIR benchmark. However, we remark that the two settings are different. In recent ZS-CIR datasets (e.g., CIRR (Liu et al., 2021) and CIRCO (Baldrati et al., 2023)), images are drawn from an open domain (respectively NLVR² (Suhr et al., 2019) and COCO (Lin et al., 2014)), thus concepts are defined at a semantic category level. Moreover, the reference image typically contains multiple elements. This aspect poses the additional challenge of understanding to which image element the relative caption refers to (see, e.g., (Chen et al., 2023)), and whether an image actually represents the same concept or not. Differently, in the considered personalization setting, concepts are typically instances, clearly identifiable in the provided image examples. FashionIQ (Wu et al., 2021) is another CIR dataset focusing on kinds of fashion items.

Personalized Text-to-Image Generation (TIG). This task was proposed in (Gal et al., 2023a) and is the generative counterpart of personalized TIR. The authors presented Textual Inversion (Gal et al., 2023a), a method that expands the vocabulary of a frozen text-to-image model (Latent Diffusion (Rombach et al., 2022)) with user-specific concept embeddings learned from few images. In (Ruiz et al., 2023) the authors present DreamBooth, a method that selects a rarely-used token and binds it to the concept by fine-tuning the text-to-image model on the concept images. Subsequent methods propose improvements over these two (Gal et al., 2023b; Kumari et al., 2023b; Tewel et al., 2023). The datasets introduced by these works (see, e.g., DreamBooth and CustomConcepts101 (Kumari et al., 2023a; b)), are constituted by concept training images and a list of prompts for evaluation, but do not

contain any real image representing such prompts. The validation is thus carried out by comparing a generated image with the training images (to measure concept fidelity) and with the context in the prompt (to measure context fidelity). Differently, typical validation metrics for generative models measure a distance between the population of real and generated samples (see, e.g., (Han et al., 2023; Kynkäänniemi et al., 2019; Naeem et al., 2020)). Since in ConCon-Chi a set of real image realizations is provided for each prompt, we show how the application of these distance measures can improve the validation of personalized TIG methods.

8.3 Concept-Context Chimera Benchmark

We present ConCon-Chi and compare with existing datasets in 8.3.1, then describe its acquisition process in 8.3.2.

8.3.1 Dataset Overview

Since personalized TIR/TIG are few-shot tasks, we compare in 8.1 with related benchmarks in terms of validation/test splits.

ConCon-Chi test split includes 1084 queries and a pool of 4008 images which are treated as ground-truth images for TIR and image realizations for TIG (~ 6 per query). Each query was generated by composing up to 3 concepts and a context from a set of 20 concepts and 101 contexts. The average number of contexts associated to each concept is indicated under CONTEXTS/CONCEPT and gives an idea of the variety of situations in which a concept is required to be retrieved or generated.

In 8.1 we also highlight the difference between personalization and ZS-CIR datasets. These latter do not explicitly deal with concepts and contexts, but each reference image and relative caption that compose a query are counted as a different concept and context. Thus their number of concepts and context is higher than in personalized tasks and the ratios CONCEPTS/CONTEXT and CONTEXTS/CONCEPT are close to 1. Differently, to evaluate the capability of personalization methods to combine the learned concepts with contexts, CONTEXTS/CONCEPT should be high and, to avoid context bias, CONCEPTS/CONTEXT should be high as well.

The most similar dataset to ours is CustomConcepts101, which however does not contain real images for evaluation (see POOL and IMAGES/QUERY in 8.1).

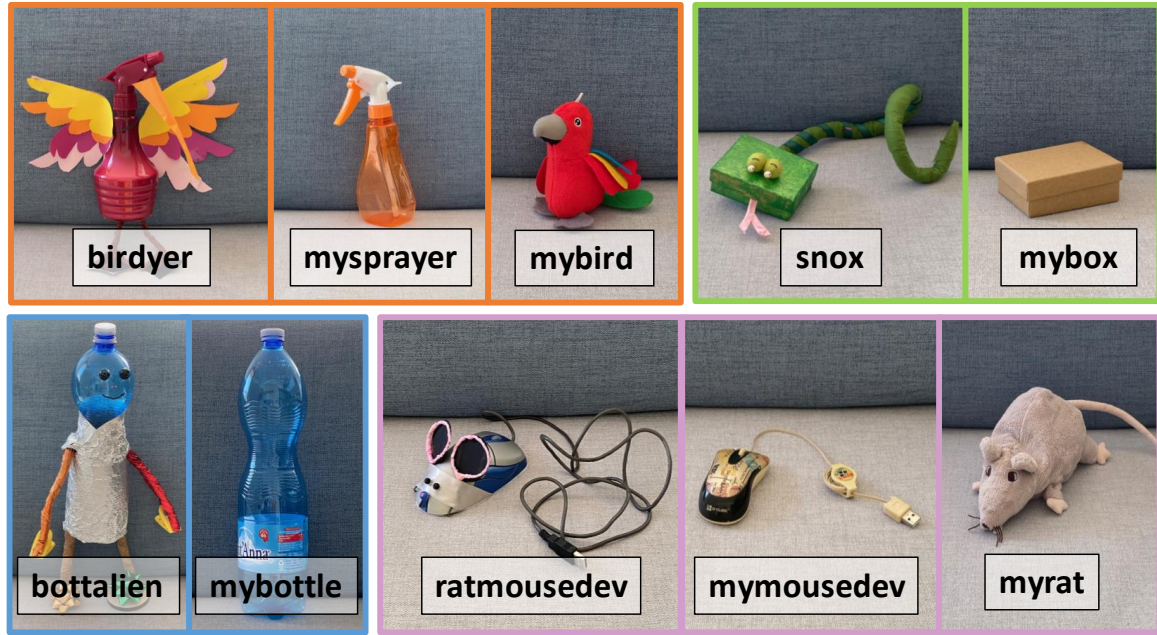


Figure 8.2 **Example concepts in ConCon-Chi¹**. Four chimeric and six common concepts (hard negatives outlined with same color).

Concepts. In 8.2 we show 10 of the 20 concepts in the dataset (the complete set is in supp. material B.1). As in (Ruiz et al., 2023; Tewel et al., 2023), we include animate and inanimate entities (puppets and tools, accessories, clothes) with the aim of creating rich interactions. There are 6 chimeric concepts, which are animal or alien puppets crafted out of tools or accessories, and 14 common concepts. Among these latter, we include instances of the same categories composing the chimeras, which are thus visually and semantically similar. These are hard negatives especially when appearing in the same context (e.g., BIRDYER and MYSPRAYER in 8.1) and enhance the compositionality assessment (Ma et al., 2023).

Concept names are invented and provided for completeness since are not used by the considered methods. Differently, in our evaluation (e.g., 8.2) we consider feeding a pre-trained VL model with a description of the concept as zero-shot baseline, thus we provide the adopted descriptions for reproducibility since words choice was empirical (supp. material B.1). *Discriminative* descriptions were formulated to be a competitive baseline, by thinking to a minimal sentence discriminating the concept from others in the concept set. *Coarse* descriptions are category-level and model the typical case where the concepts to be discriminated from are unknown (e.g., “bag”, but there are two bags in the concept set). *Rich* descriptions aim to verify whether enriching the *Discriminative* with visual details helps.

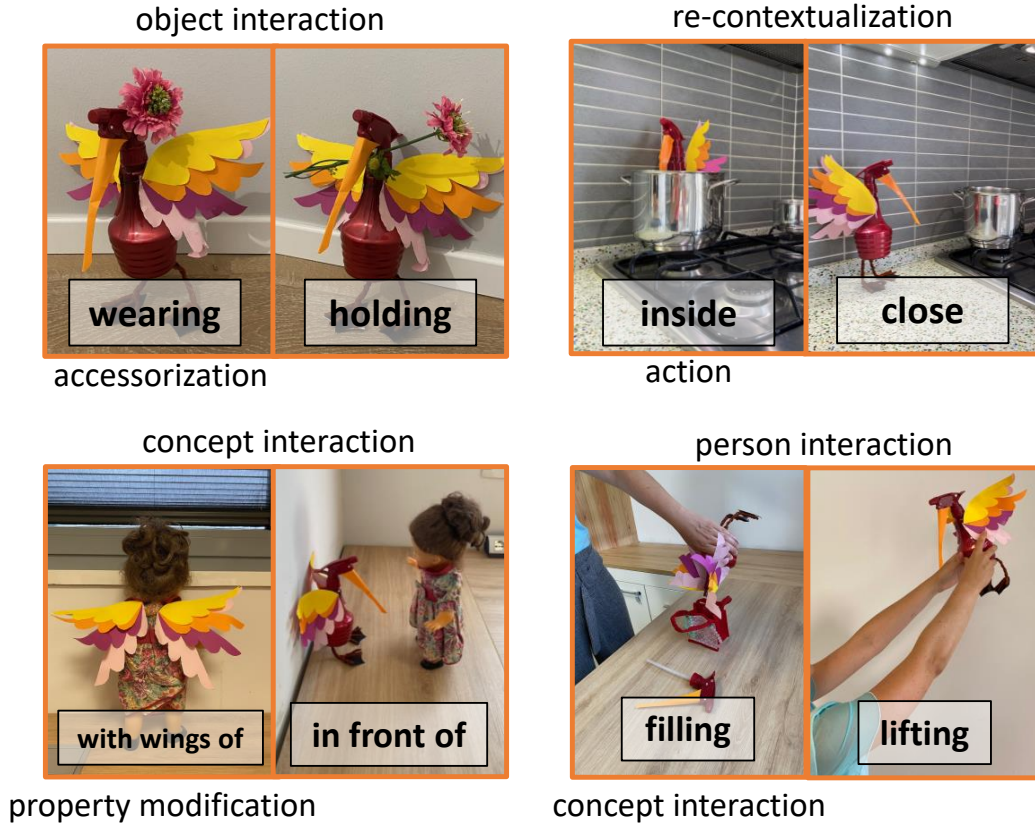


Figure 8.3 **Fine-grained contexts.** Example queries for the concept BIRDYER, in which the recognition of the co-occurrence of elements does not suffice for retrieval.

Contexts. Contexts are grouped into 9 kinds (one example per kind in 8.1). Inspired by (Ruiz et al., 2023), these are structured in concept modifications (accessorizations, property modifications, renditions) and relationships (actions, interactions, re-contextualizations). We describe them and present their distribution in supp. material B.2a.

Similarly to (Ruiz et al., 2023; Tewel et al., 2023) we include general and specific contexts: general contexts are applicable to all or most concepts (e.g. re-contextualization in 8.1) and strongly contribute to avoiding context bias; specific contexts are typical of a

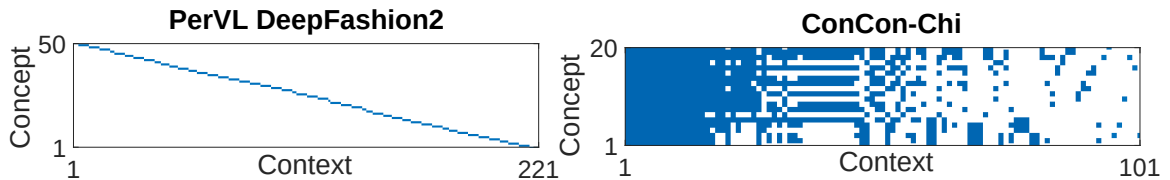


Figure 8.4 **Concept-context matrices.** Blue cells mark the concept-context combination appearing in the dataset (*test split*).

concept category and aim to increase the difficulty of distinguishing between hard negatives (e.g. the action *containing* for bags-like objects). When a context specifies an interaction between a concept and an entity (object, person, another concept), a retrieval method could perform well simply by detecting their co-occurrence. To avoid this we included contexts specifying different interactions with the same entities (see 8.3). A particular interaction is the one between two learned concepts. Concept-concept interactions have been studied in personalized TIG (Kumar et al., 2022; Tewel et al., 2023). In 8.4 we evaluate this aspect in personalized TIR.

Concept-context structure. In 8.4 we compare the concept-context structure of our dataset with PerVL DF2. A matrix cell is Blue when a query formed by the corresponding context and concept exists in the dataset. While our dataset contains approximately half of the concepts and contexts, our concept-context matrix is denser, reflecting a higher number of queries (1084 vs. 221, see 8.1). Specifically, in PerVL DF2 each context is coupled with a single concept, while in our dataset this only occurs for 8 contexts, with 16 contexts combining with every concept. Moreover, in ConCon-Chi each concept is composed with 46.64 contexts on average vs. 4.42 in PerVL DF2 (few cells per row): similarly to context bias, concept bias limits the compositionality assessment since it makes it difficult to determine to what extent the retrieval method understands the context.

8.3.2 Acquisition and Annotation

Acquisition. First we gathered the objects and crafted the chimeric concepts; then we moved to photograph each concept in the designed contexts. We arranged the scene to represent the query and shot multiple pictures by varying the viewpoint. Any private and sensitive information was removed from the scene; images containing humans were cropped to remove the face and make the subject unidentifiable. We provide the distribution of environments in supp. material B.2b and details about the creation of renditions in supp. material B.1.3.

Annotation. Annotations consist of query-image associations. A first step consisted of labelling each image with the represented concepts and context. A second step was carried out to remove false negatives in the annotations. These happen whenever the ground-truth (GT) images of a query that is more specific, are not included among the GT of a query that is

¹Any representation of trademarks, trade names, logos, domain names such as any other distinguishing marks appearing in this dataset is purely random, and it is used exclusively for scientific and non-commercial purposes; therefore, the relevant representation cannot be understood as an expression of an opinion or an indication or a precondition for taking decisions.

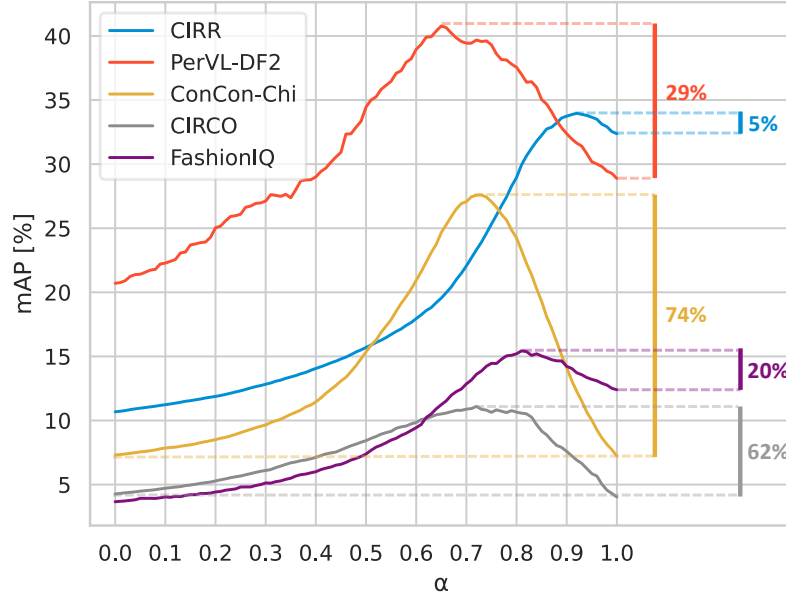


Figure 8.5 **Concept and context biases.** Retrieval performance by varying the relative weight (α) of the concept and the context in the query. Lateral bars: performance drop when using either only the concept ($\alpha=0$) or the context ($\alpha=1$), relative to peak performance.

more generic (query overlap). A similar problem happens when GT images of a query contain content that also corresponds to other queries (image overlap). In benchmarks created by re-annotating existing datasets of images “in the wild” it is not possible to control overlaps and an exhaustive check is unfeasible. To this end, CIRR (Liu et al., 2021) and CIRCO (Baldrati et al., 2023) designed approximated procedures to ensure that false negatives are respectively absent in sub-pools of images or below an estimated percentage. We accounted for the problem since the dataset design by controlling and minimizing overlaps. The GT images were then assigned following concept and context overlap and manually checking those of queries where we were aware of possible image overlap. As a result, to the best of our knowledge, the set of 1084 queries is free from false negatives with respect to the pool of 4008 images. We report the number of GT images per query in supp. material B.3.

8.4 Personalized Text-to-Image Retrieval

We define the TIR benchmark task on ConCon-Chi and compare it with related benchmarks in 8.4.1. We present and analyse results respectively in 8.4.2 and 8.4.3.

8.4.1 Benchmark Task

Train and test splits. Each concept in the set of 20 is trained independently on 1 to 5 images where the concept is standing in front of some background (*train* split, examples in 8.2). At test time, for each query in the *test* split we rank the images in the pool according to their similarity with the query. Training backgrounds do not appear in the *test* split. Since the concepts are not trained jointly, similar concepts play the role of hard negatives. In supp. material B.1.4 we introduce other splits not used in this paper.

Metrics. For each query we evaluate the rank of the first GT image (mean Reciprocal Rank, mRR) and whether this is among the top-k (recall rate, R@k); then we evaluate the rank of all GT images (mean Average Precision (Beitzel et al., 2009), mAP) and of GT images up to the top-k, mAP@k (Baldrati et al., 2023). See supp. material B.2.1 for definitions.

Concept and context biases. We first aim to compare the presented benchmark with existing ones in terms of context and concept biases. To quantify the importance of attending the concept and the context for correct retrieval, we adopt the experimental setup proposed in (Saito et al., 2023) and model the concept-context interaction by computing a query embedding q as weighted sum of c , the CLIP (Radford et al., 2021) embedding of the context (output of CLIP text encoder, ViT-L14 backbone), and i , the average CLIP embedding of the concept training images (output of the vision encoder): $q = (1 - \alpha) \cdot i + \alpha \cdot c$ with $\alpha \in [0, 1]$. In 8.5 we compare the retrieval performance by varying α on ConCon-Chi, PerVL-DF2, CIRCO, CIRRR and FashionIQ. For PerVL DF2 and ConCon-Chi we consider the available training images and for each dataset we use the *test* split if available, the *validation* otherwise (see B.2.2 and 8.1 for details). When in ConCon-Chi multiple concepts appear in the query we average their embeddings.

For three datasets the performance achieved by just putting the context in the query ($\alpha=1$) is higher than 70% of the peak performance, indicating that the probability of retrieving the correct context-concept combination is relatively high even when ignoring the concept (context bias). Differently, in ConCon-Chi we observe low performance for $\alpha=1$, and a drop of 74% from the peak performance with $\alpha=0$ (only the concept in the query). Only CIRCO exhibits a similar trend, with lower peak performance possibly because of the intrinsic difficulty of the proposed ZS-CIR task and the large image pool.

Method	mAP [%]	mRR [%]	R@1 [%]
<i>Coarse</i>	16.83	24.21	14.48
k=0 <i>Discriminative</i>	*30.16*	*43.16*	*31.92*
<i>Rich</i>	*27.65*	*40.58*	*29.98*
PALAVRA	22.56 ± 1.29	34.39 ± 1.68	24.59 ± 1.94
k=1 Pic2Word	25.23 ± 1.20	37.16 ± 1.76	26.35 ± 1.85
SEARLE	28.16 ± 0.55	41.07 ± 0.92	31.16 ± 0.94
PALAVRA	23.59	35.99	26.75
k=5 Pic2Word	26.39	38.62	27.68
SEARLE	30.74	43.83	33.49

Table 8.2 **Personalized TIR benchmark.** Performance of CLIP baselines (k=0) and personalized TIR (PALAVRA) and ZS-CIR methods (Pic2Word, SEARLE) on the retrieval task in ConCon-Chi. The best method per metric is in **Bold**; *Discriminative* and *Rich* baselines (oracles) are highlighted with *asterisks*.

8.4.2 Benchmark Results

Methods. We compare the personalized TIR method PALAVRA (Cohen et al., 2022) with two ZS-CIR methods, Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023), on the defined TIR benchmark. Since, differently from ZS-CIR, in our setting multiple concept example images are available, for the two latter methods we average the generated token embeddings to create the concept embedding. All methods rely on CLIP ViT-L14 backbone with same input pre-processing. The methods learn a textual token embedding that expands CLIP vocabulary and assign it an arbitrary textual identifier such that at inference they replace the learned embedding whenever this is encountered in an input query. We used the code released by the authors (see supp. material B.2.3 for details).

Results. In 8.2 we evaluate each method with k=1 and k=5 training images per concept (for k=1 we report mean and standard deviation over the 5 images). We also report, as baselines that do not use any image (k=0), the performance achieved by feeding CLIP with queries where the concept identifier is replaced with the *Coarse*, *Discriminative* or *Rich* descriptions introduced in 8.3.1.

As expected, *Coarse* descriptions provide a lower bound, since they are shared among the concepts of same category to simulate descriptions that are not tailored to the discriminative task at hand. Differently, *Discriminative* and *Rich* descriptions provide a competitive baseline (higher than PALAVRA and Pic2Word). They represent an oracle, since were formulated by discriminating every concept from the others, thus accessing information which is unavailable

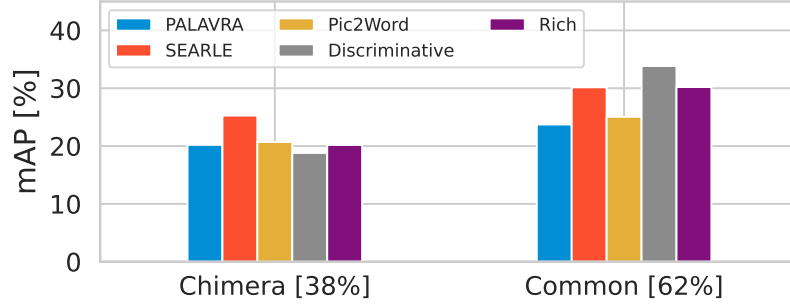


Figure 8.6 **New-meaning learning**. Retrieval performance on queries containing at least one chimeric concept vs. common concepts (percentage of each query kind in the dataset in brackets).

Method	All Concepts	Chimeric Concepts	Common Concepts	Contexts
<i>Coarse</i>	29.38	09.51	37.89	42.11
<i>Discriminative</i>	*66.48*	*60.61*	*68.99*	40.92
<i>Rich</i>	*84.54*	*85.54*	*84.11*	29.81
PALAVRA	91.52	91.15	91.67	28.56
Pic2Word	50.04	51.32	49.49	50.12
SEARLE	76.58	82.25	74.15	40.47

Table 8.3 **Concept-context compositionality**. Performance (F1 score [%]) of recognition of concepts and contexts in the retrieval task of 8.2.

to other methods that learn each concept independently. In this respect, we note that adding visual details degrades performance. Interestingly, SEARLE outperforms the *Discriminative* baseline thus being the best method, also because it exhibits a smaller standard deviation in the 1-shot scenario and a larger gain when more images are provided. This confirms that SEARLE outperforms the other two methods not only in ZS-CIR (Baldrati et al., 2023) but also in the personalized TIR setting. We report more metrics in supp. material B.3.

8.4.3 Analysis of Results

We analyse the performance achieved in the benchmark in terms of new-meaning learning and compositionality. For this study we consider the methods trained on 5 images.

New-meaning learning. In 8.6 we report the mAP of 8.2 separately on queries containing at least one chimeric concept and common ones. We see that all methods and baselines achieve a lower performance when retrieving queries containing a chimeric concept. We then

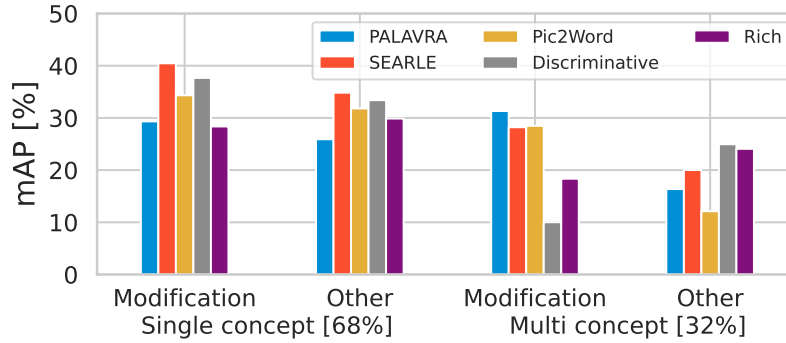


Figure 8.7 **Concept-concept compositionality.** Performance on single- and multi-concept queries, comparing “accessorization” and “property modification” (“modification”) kinds versus the rest of the kinds (“other”).

inspected whether this can be explained with a misclassification of the concept or the context, or their combination. For each query therefore we considered the first N ranked images (N equal to the number of GT images) and marked the concept in the query as ground-truth and the concept in each retrieved image as a prediction. We computed the F1 score for each concept by accumulating the predictions from the queries containing that concept and we report in 8.3 the average scores over all, chimeric and common concepts. Please note that we restricted this analysis to single-concept queries and discarded ranked images containing multiple concepts.

In 8.3 we observe that personalization methods achieve equal or better classification performance for chimeric with respect to common concepts. We then conclude that, while these methods learn to recognize the visual appearance of the chimeras, they struggle more at retrieving these new meanings in composition with known contexts.

Concept-context compositionality. As for the concepts, we computed the classification performance in a similar way for the contexts, and report it in 8.3. We see that PALAVRA achieves the best concept and the lowest context classification performance. We deduce that its low retrieval performance in the benchmark must be due to the poor compositionality properties of the learned tokens. Conversely, Pic2Word excels at recognizing the contexts, but cannot learn discriminative tokens. Thus, the good retrieval performance of the *Discriminative* baseline and of SEARLE in 8.2 seems explained by their capability to trade-off and combine context and concept recognition. This seems confirmed by the fact that the *Rich* descriptions improve concept recognition at the expense of the context, probably since longer descriptions out-weight the rest of the query.

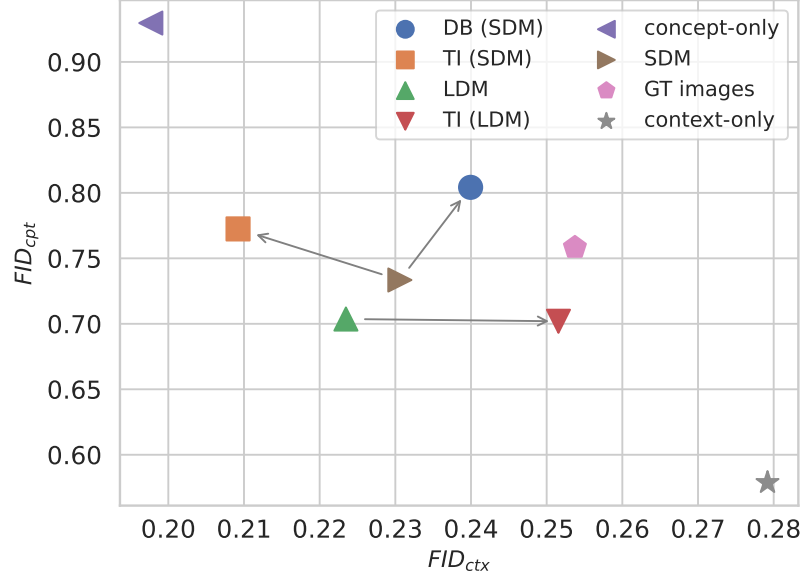


Figure 8.8 **Personalized TIG benchmark**. Scatter plot of the FID_{cpt}/FID_{ctx} trade-off for the considered methods, baselines and upper bounds on the generation task ConCon-Chi.

		k=3		k=10		FID_{cpt}	FID_{ctx}
		Density $\uparrow$	Coverage $\uparrow$	Density $\uparrow$	Coverage $\uparrow$		
SDM	Common	2.66	0.12	8.84	1.33	0.74	0.23
	Chimeric	0.78	0.02	2.40	0.21	0.73	0.23
TI	Common	3.43	0.23	11.56	2.93	0.71	0.25
	Chimeric	2.71	0.15	5.43	0.90	0.69	0.25
DB	Common	6.55	0.28	18.98	2.65	0.80	0.24
	Chimeric	4.81	0.14	11.49	1.09	0.79	0.25

Table 8.4 **Comparing generated and GT images**. We consider TI and DB with SDM. The best for each metric is in **Bold**. Coverage and Density are in [%] (Density not upper bounded by 100).

Concept-concept compositionality. We finally investigated whether performance also depends on the context kind. In 8.7 we report the mAP of 8.2, separately on queries containing a single and multiple concepts and, in each group, containing a concept modification (“property modification” or “accessorization”) or another kind of context. As expected, multi-concept queries are more challenging. On these we observe an interesting trend: description-based baselines drop performance on concept modifications, while the personalization methods drop performance on the other kinds. Multi-concept queries with concept modifications are mostly the ones where a part of a concept is worn or applied to another one (an example in 8.3 where the wings of BIRDYER are applied to MYDOLL). These queries require knowing visual information on the concept parts, thus explaining why the baselines fail in these cases. Conversely, the other kinds of multi-concept queries typically require the detection of the co-occurrence of the concepts and their relationship (e.g., in 8.3, “MYDOLL in front of BIRDYER”). To achieve this, the concept textual representations must exhibit robust compositionality properties. Thus, the tokens learned by the methods do not seem to retain the same compositionality properties of the tokens in the original vocabulary (which form the descriptions).

8.5 Personalized Text-to-Image Generation

We define the TIG benchmark task and present results in 8.5.1. We analyse them in more detail in 8.5.2.

8.5.1 Benchmark Task and Results

Train and test splits. We train each concept on the same 5 images used for TIR and at test time we generate 4 images per prompt. We restrict the evaluation to the single-concept prompts in the *test* split (which are 735).

Metrics. We report the two metrics introduced in (Gal et al., 2023a) and adopted by following works. These rely on CLIP and compute, for each prompt, the average pairwise cosine similarity between the visual embeddings of the generated images and of, respectively, the visual embeddings of the training images (fidelity-concept or FID_{cpt} as in (Ruiz et al., 2023)) and the textual embedding of the context in the prompt (fidelity-context or FID_{ctx}). Personalized TIG methods are expected to trade-off the two metrics by learning to represent the concept appearance while retaining the capability to represent contexts.

Methods. As in personalized TIR, the considered methods assign a textual identifier to the concept and when this is encountered in the prompt they load a corresponding learned token embedding (Textual Inversion, TI (Gal et al., 2023a)) or model (DreamBooth, DB (Ruiz et al., 2023)). For DB we adopted the Diffusers library (von Platen et al., 2022) with a Stable Diffusion Model (SDM) (Rombach et al., 2022) as pre-trained text-to-image model (called DB(SDM)). For TI we adopted the code by the authors with either a Latent Diffusion Model (LDM) (Rombach et al., 2022) or, by following (Tewel et al., 2023), the same SDM as for DB (called TI(LDM) and TI(SDM)). Details in supp. material B.3.2.

Results. In 8.8 we show the fidelity-concept/context trade-off (numbers in supp. material B.4). To quantify these metrics, as in (Gal et al., 2023a) we compare with upper bounds. For the fidelity-concept, the upper bound returns always the 5 training images irrespective of the prompt (concept-only). For the fidelity-context, the upper bound are images generated by the pre-trained SDM when fed as input only with the context of the prompt (context-only). We also evaluate the baseline performance of the LDM or SDM pre-trained models by replacing the concept identifier in the prompt with its *Rich* description. As reference, we report also the metrics for the GT images available in ConCon-Chi.

We first observe that TI(LDM) does not improve the fidelity-concept over LDM, but improves the fidelity-context. This can be explained by considering that the *Rich* description in LDM tends to out-weight the context. Conversely, TI(SDM) improves the fidelity-concept of SDM, while degrading the fidelity-context. This behavior has also been observed in recent work (Tewel et al., 2023) and may be due to the learned token over-fitting the context of the training images. A better learning behavior is provided by DB(SDM), which improves both metrics over SDM. Notably, compared with the GT images, DB shows a higher fidelity-concept and lower fidelity-context.

8.5.2 Analysis of Results

We leverage the availability of GT images to inspect the performance of DB, TI and their pre-trained model SDM with two metrics proposed for evaluation of text-to-image generation (Naeem et al., 2020): Density as measure of realism (fidelity) and Coverage as measure of diversity (how well generated images span the real images manifold). We created one real and one generated manifold per concept as the union of all prompts for that concept (similarly to (Han et al., 2023) Appendix L). In 8.4 we report the metrics for two values of the parameter k , averaged over chimeric and common concepts. For all methods, Density and Coverage are higher on common concepts than on chimeric ones. This is in line with

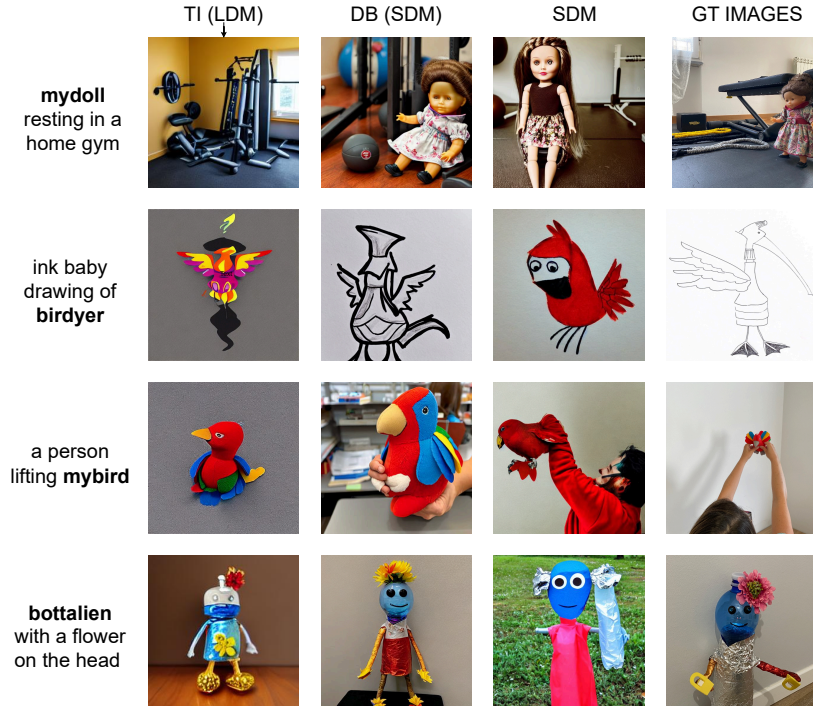


Figure 8.9 Qualitative results for personalized TIG.

what has been observed in (Ruiz et al., 2023; Tewel et al., 2023) that more common objects are easier to generate since these methods leverage the pre-trained knowledge about the concept category. Interestingly, such difference is not evident from the $FID_{\text{cpt/ctx}}$ metrics typically adopted in personalized TIG. We observe that DB has best Density, while in terms of Coverage there is a smaller gap with TI. Thus DB achieves relatively better realism, but diversity remains low. We conclude that leveraging the availability of a real population of images allows for a more thorough evaluation of personalized TIG: there seems to be large room for improvement, with new-meaning concepts posing more challenges.

Qualitative examples. We report some cherry-picked examples in 8.9. We observe that DB represents the concepts more accurately and combines them more nicely with the context than TI, which sometimes forgets the concept (First Row) and sometimes the context (Third Row). It can also be noticed how common concepts are learned very accurately by DB (First and Third Row), while chimeric ones are not. As expected, using a visually *Rich* description (Third Column) is not enough for personalized generation, especially on chimeric concepts (Second and Fourth Row), which lack more distinctive features than the common ones.

8.6 Conclusion

We present a new dataset called ConCon-Chi for the evaluation of personalized TIR and TIG. The dataset models novel concepts as chimeric objects and by adopting a concept-context matrix structure allows to study the learning of new meanings in terms of their compositionality properties with known ones. Our analysis on retrieval showed that current methods struggle at composing these new concepts with known contexts, and also together. A similar issue was observed when evaluating the generated images with respect to real examples by adopting image generation metrics. We hope that this study and the dataset released will help improving current personalized VL methods.

Acknowledgments

We express our sincere gratitude to the members of Sony Interactive Entertainment Inc. teams for their invaluable support and insightful feedback related to our project.

Additional information about the ConCon-Chi dataset, further details on the personalized TIR and TIG experiments, and additional qualitative results are provided in Appendix B.

Part III

Conclusion and Outlook

Chapter 9

Discussion

9.1 From Recognition to Interaction

The three research directions developed in this dissertation can be seen as a progressive improvement for a data-efficient, safe and natural HRI. The first direction starts from action recognition and asks how a robot can remain reliable when labeled data are scarce and the observed action may fall outside the known interactions. The second direction moves beyond fixed label spaces and asks how language-aligned representations can support semantic transfer, compositional reasoning, and user-specific concepts. The third direction then places these perceptual questions inside an embodied interaction loop, showing why the eventual target is not only recognition performance but the ability to turn perception into appropriate robot behavior.

At thesis level, the contributions can be summarized as follows:

- **Few-shot open-set action recognition.** For this task, we formulate FSOSAR as a setting where a robot must recognize actions from a small support set while rejecting unsupported queries, and we propose Feature-Residual Discrimination (FR-Disc) as a residual-based mechanism for episode-level rejection ([Berti et al., 2026](#); [2022](#)).
- **Vision-language generalization and personalization.** For vision-language transfer, we focus on frozen-backbone prompt-based adaptation in two use-cases: DualVCLIP for zero-shot multi-label action recognition with compositional verb–object actions, and ConCon-Chi for evaluating compositional new-meaning learning in personalized retrieval and generation ([Calabrese et al., 2024](#); [Rosasco et al., 2024](#)).

- **Embodied interaction and behavior cloning.** Finally, we use a deployed ergoCub demo to make explicit how perception outputs are used inside an interaction loop, and we reframe the same skills through teleoperation and behavior cloning as an alternative to scaling modular perception interfaces and behavior (Chi et al., 2025; Zhu et al., 2025).

Table 9.1 gives a compact thesis-level view of these three research directions, summarizing for each one the main capability it adds and the questions it leaves open.

Table 9.1 Thesis-level view of the three research directions and of the main capability each one adds to the overall argument of the dissertation.

Research direction	What it adds	What it leaves open
Few-shot open-set action recognition	Action understanding becomes data-efficient and novelty-aware, allowing the robot to classify taught actions while rejecting unknown ones (Berti et al., 2026; 2022).	Recognition is still tied to an episodic support set and does not yet provide broader semantic transfer beyond the current action vocabulary.
Vision-language models for semantic generalization and personalization	Language-aligned representations enlarge the semantic space, enabling zero-shot reasoning over unseen labels and evaluation of user-specific concepts learned from a few images (Calabrese et al., 2024; Rosasco et al., 2024).	Semantic transfer remains uneven: object-level cues are easier than temporal verb structure, and robust new-meaning learning is still far from solved.
Behavior cloning and toward end-to-end policies	Replace the deployed ergoCub modular interaction pipeline with a behavior-cloning policy, highlighting the transition toward demonstration-driven humanoid control (Chi et al., 2025; Zhu et al., 2025).	Learned policies still need reliable uncertainty handling, scalable demonstrations, and a principled way to exploit semantic structure rather than bypass it.

9.2 Cross-Cutting Lessons

Across the three directions, a few recurring points are worth keeping in mind:

- **Data are a constraint at deployment time.** The few-shot and personalization settings studied in this thesis are motivated by the fact that a robot often needs to adapt from a small number of examples, whether those examples are action demonstrations or new concept images.

- **Novelty should be expected.** Unknown actions, unseen verb-object combinations, and user-specific concepts are typical in open-ended HRI; this is why rejection, semantic extensibility, and personalization appear repeatedly across the dissertation.
- **Modularity and policy learning complement each other.** A modular stack supports inspectability and explicit safety logic, while behavior cloning becomes attractive when scaling a hand-designed set of perception interfaces and branching conditions becomes costly.

9.3 Limits and Open Problems

This thesis also points to several concrete directions that are important for moving from robust perception components to more adaptive embodied behavior:

- **Combination of action recognition and behavior cloning at different time scales.** In many embodied humanoid applications, behavior-cloning policies are executed at relatively low rates and rely on short observation histories to maintain inference fast enough. For more fluid and fine-grained interaction, longer temporal context would be desirable, but this may increase latency and computational cost. A natural direction is to use the policy primarily for spatial control cues from the current RGB observation (where to reach, how to move the hands), while using a temporally stronger action-recognition module to provide higher-level temporal cues about what is happening. The resulting temporal summary can then condition policy execution without requiring the policy itself to absorb long histories.
- **Integration of open-set awareness into policy learning.** Open-set recognition is treated explicitly in the first part of the thesis, but behavior-cloned policies typically lack a principled reject/abstain mechanism. Bringing open-set awareness and calibrated uncertainty into policy learning could reduce the frequency of out-of-distribution behavior and support safer fallback strategies at deployment time.
- **Connect VLM personalization and open-vocabulary transfer to policy learning.** Recent humanoid policies increasingly adopt vision-language(-action) backbones. In that setting, techniques for prompt-based adaptation and for learning new concept tokens from a few images can be used to personalize or extend the policy vocabulary, while preserving the compositional structure needed to combine a learned concept with known contexts at inference time.

References

- Adeli Mosabbeeb, E., Cabral, R., De la Torre, F., and Fathy, M. (2015). Multi-label discriminative weakly-supervised human activity recognition and localization. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 241–258. Springer. 95, 96
- Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*. 64, 65
- Akkaladevi, S. C. and Heindl, C. (2015). Action recognition for human robot interaction in industrial applications. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 94–99. IEEE. 2, 7, 57
- Anagnostis, A., Benos, L., Tsaopoulos, D., Tagarakis, A., Tsolakis, N., and Bochtis, D. (2021). Human activity recognition through recurrent neural networks for human–robot interaction in agriculture. *Applied Sciences*, 11(5):2188. 96
- Ayub, A. and Wagner, A. R. (2020). Tell me what this is: Few-shot incremental object learning by a robot. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8344–8350. IEEE. 3, 93, 94, 95, 96
- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*. 63
- Bahdanau, D., Bosc, T., Jastrzebski, S., Grefenstette, E., Vincent, P., and Bengio, Y. (2017). Learning to Compute Word Embeddings On the Fly. *arXiv preprint arXiv:1706.00286*. 110
- Bain, M., Nagrani, A., Varol, G., and Zisserman, A. (2022). A clip-hitchhiker’s guide to long video retrieval. *arXiv preprint arXiv:2205.08508*. 33, 102, 103
- Baldrati, A., Agnolucci, L., Bertini, M., and Del Bimbo, A. (2023). Zero-Shot Composed Image Retrieval with Textual Inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15338–15347. 35, 38, 109, 111, 116, 117, 118, 119, 162, 163
- Bao, W., Yu, Q., and Kong, Y. (2021). Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13349–13358. 80
- Baradel, F., Neverova, N., Wolf, C., Mille, J., and Mori, G. (2018). Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121. 97

- Beitzel, S. M., Jensen, E. C., and Frieder, O. (2009). *MAP*, pages 1691–1692. Springer US, Boston, MA. 117
- Berti, S., Pasquale, G., and Natale, L. (2026). A baseline study and benchmark for few-shot open-set action recognition with feature residual discrimination. *arXiv preprint arXiv:2603.04125*. Accepted at the International Conference on Pattern Recognition (ICPR), 2026. iv, 6, 7, 8, 16, 19, 21, 22, 23, 24, 127, 128
- Berti, S., Rosasco, A., Colledanchise, M., and Natale, L. (2022). One-shot open-set skeleton-based action recognition. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 765–772. IEEE. iv, 6, 7, 8, 16, 19, 22, 23, 24, 78, 79, 80, 83, 94, 95, 96, 127, 128
- Bisgin, H., Palechor, A., Suter, M., and Günther, M. (2024). Large-scale evaluation of open-set image classification techniques. *arXiv preprint arXiv:2406.09112*. 78, 79, 87
- Bo, Y., Lu, Y., and He, W. (2020). Few-shot learning of video action recognition only based on video contents. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 595–604. 57, 60
- Bonasso, R. P., Kortenkamp, D., Miller, D. P., and Slack, M. G. (1997). Experiences with an architecture for intelligent, reactive agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 9(2-3):237–256. 43
- Calabrese, C., Berti, S., Pasquale, G., and Natale, L. (2024). The impact of compositionality in zero-shot multi-label action recognition for object-based tasks. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. iv, 6, 26, 29, 31, 33, 34, 39, 127, 128
- Cao, K., Ji, J., Cao, Z., Chang, C.-Y., and Niebles, J. C. (2020). Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10618–10627. 80, 84
- Careaga, C., Hutchinson, B., Hodas, N., and Phillips, L. (2019). Metric-based few-shot learning for video action recognition. *arXiv preprint arXiv:1909.09602*. 57, 60
- Cen, J., Zhang, S., Wang, X., Pei, Y., Qing, Z., Zhang, Y., and Chen, Q. (2023). Enlarging instance-specific and class-specific information for open-set action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15295–15304. 80
- Chen, G., Peng, P., Wang, X., and Tian, Y. (2021a). Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081. 79
- Chen, W., Si, C., Zhang, Z., Wang, L., Wang, Z., and Tan, T. (2023). Semantic Prompt for Few-Shot Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23581–23591. 111
- Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. (2019). A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*. 12

- Chen, X., Li, Y., and Chen, H. (2024). Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection. *arXiv preprint arXiv:2405.16146*. 80
- Chen, Y., Liu, Z., Xu, H., Darrell, T., and Wang, X. (2021b). Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9062–9071. 12, 14
- Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. (2025). Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11). 46, 48, 128
- Chiatti, A., Bardaro, G., Bastianelli, E., Tiddi, I., Mitra, P., and Motta, E. (2020). Task-agnostic object recognition for mobile robots through few-shot image matching. *Electronics*, 9(3):380. 3, 93, 94, 95, 96
- Cohen, N., Gal, R., Meirom, E. A., Chechik, G., and Atzmon, Y. (2022). "This Is My Unicorn, Fluffy": Personalizing Frozen Vision-Language Representations". In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 558–577, Cham. Springer Nature Switzerland. 34, 35, 38, 109, 110, 118, 162
- Deng, S., Yu, J.-G., Wu, Z., Gao, H., Li, Y., and Yang, Y. (2022). Learning relative feature displacement for few-shot open-set recognition. *IEEE Transactions on Multimedia*, 25:5763–5774. 77, 80
- Dhamija, A. R., Günther, M., and Boulton, T. (2018). Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31. 8, 23, 78, 79, 81
- Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. 14
- G. Koch, R. Zemel, R. Salakhutdinov (2015). Siamese Neural Networks for One-shot Image Recognition. In *ICMLw*. 84
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-or, D. (2023a). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*. 34, 35, 39, 109, 111, 122, 123, 165
- Gal, R., Arar, M., Atzmon, Y., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2023b). Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models. *ACM Trans. Graph.*, 42(4). 111
- Ge, Y., Zhang, R., Wang, X., Tang, X., and Luo, P. (2019). DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5337–5345. 109, 111
- Ge, Z., Demyanov, S., Chen, Z., and Garnavi, R. (2017). Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*. 79

- Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR. 62
- Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850. 23, 84
- Guo, Y. and Cheung, N.-M. (2020). Attentive weights generation for few shot learning via information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13508. 58
- Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. (2023). Rarity Score: A New Metric to Evaluate the Uncommonness of Synthesized Images. In *The Eleventh International Conference on Learning Representations*. 112, 123
- Han, Y., Zhang, P., Zhuo, T., Huang, W., and Zhang, Y. (2018). Going deeper with two-stream convnets for action recognition in video surveillance. *Pattern Recognition Letters*, 107:83–90. 57
- Herbelot, A. and Baroni, M. (2017). High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark. Association for Computational Linguistics. 110
- Hill, F., Cho, K., Korhonen, A., and Bengio, Y. (2016). Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30. 110
- Huang, S., Ma, J., Han, G., and Chang, S.-F. (2022). Task-adaptive negative envision for few-shot open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7171–7180. 77, 80
- Huynh, D. and Elhamifar, E. (2020). A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8776–8786. 97
- Ilharco, G., Wortsman, M., Gadre, S. Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., and Schmidt, L. (2022). Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277. 99
- Jeong, M., Choi, S., and Kim, C. (2021). Few-shot open-set recognition by transformation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12566–12575. 8, 12, 23, 57, 58, 61, 63, 77, 80
- Ji, J., Krishna, R., Fei-Fei, L., and Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247. 97

- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. 84, 99
- Kerrigan, A., Duarte, K., Rawat, Y., and Shah, M. (2021). Reformulating zero-shot action recognition for multi-label actions. *Advances in Neural Information Processing Systems*, 34:25566–25577. 96
- Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., and Abbasi, A. A. (2020). Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, pages 1–27. 57
- Korbar, B. and Zisserman, A. (2022). Personalised CLIP or: how to find your vacation videos. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press. 111
- Kubat, M. (1999). Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, 13(4):409–412. 64
- Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE. 23, 84
- Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. (2022). Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations*. 115
- Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. (2023a). Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488*. 35, 111
- Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. (2023b). Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941. 35, 111
- Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved Precision and Recall Metric for Assessing Generative Models. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 112
- Lazaridou, A., Marelli, M., and Baroni, M. (2017). Multimodal Word Meaning Induction From Minimal Exposure to Natural Text. *Cognitive Science*, 41(S4):677–705. 36, 109, 110
- Lee, C.-W., Fang, W., Yeh, C.-K., and Wang, Y.-C. F. (2018). Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1576–1585. 97
- Lentzas, A., Dalagdi, E., and Vrakas, D. (2022). Multilabel classification methods for human activity recognition: A comparison of algorithms. *Sensors*, 22(6):2353. 95, 96

- Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40. 46
- Li, J., Li, D., Xiong, C., and Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR. 3, 108
- Li, Y., Li, Y., and Vasconcelos, N. (2018). Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528. 23, 84
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing. 111, 162
- Liu, B., Kang, H., Li, H., Hua, G., and Vasconcelos, N. (2020). Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807. 8, 12, 23, 57, 58, 61, 63, 77, 80
- Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701. 22, 23, 67
- Liu, Z., Rodriguez-Opazo, C., Teney, D., and Gould, S. (2021). Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134. 111, 116
- Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. (2023). CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921. 113
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605. 90
- Mahalanobis, P. C. (1930). On test and measures of group divergence: theoretical formulae. *Journal and Proceedings of Asiatic Society of Bengal New series*26, 541-588. 61
- Malaviya, M., Sucholutsky, I., Oktar, K., and Griffiths, T. L. (2022). Can humans do less-than-one-shot learning? *arXiv preprint arXiv:2202.04670*. 95
- Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., and Darrell, T. (2020). Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1049–1059. 23, 97
- Menon, S. and Vondrick, C. (2023). Visual Classification via Description from Large Language Models. In *The Eleventh International Conference on Learning Representations*. 108

- Miller, D., Sünderhauf, N., Kenna, A., and Mason, K. (2024). Open-set recognition in the age of vision-language models. In *European Conference on Computer Vision*, pages 1–18. Springer. 80
- Mondal, A., Nag, S., Prada, J. M., Zhu, X., and Dutta, A. (2023). Actor-agnostic multi-label action recognition with multi-modal query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 784–794. 33, 102
- Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. (2020). Reliable Fidelity and Diversity Metrics for Generative Models. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7176–7185. PMLR. 112, 123, 164
- Neal, L., Olson, M., Fern, X., Wong, W.-K., and Li, F. (2018). Open set learning with counterfactual images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 613–628. 79
- Ochal, M., Patacchiola, M., Boudiaf, M., and Wang, S. (2024). Eol: Transductive few-shot open-set recognition by enhancing outlier logits. *arXiv preprint arXiv:2408.02052*. 78, 80
- Pareek, P. and Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3):2259–2322. 77
- Parisot, S., Yang, Y., and McDonagh, S. (2023). Learning To Name Classes for Vision and Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23477–23486. 111
- Peng, K., Yin, C., Zheng, J., Liu, R., Schneider, D., Zhang, J., Yang, K., Sarfraz, M. S., Stiefelhagen, R., and Roitberg, A. (2024). Navigating open set scenarios for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4487–4496. 80
- Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., and Damen, D. (2021). Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 475–484. 3, 7, 13, 16, 19, 58, 61, 64, 65, 77, 79, 80
- Qin, Y. and Liu, B. (2024). Tfrs: A task-level feature rectification and separation method for few-shot video action recognition. *Neural Networks*, 176:106326. 86
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. 3, 27, 95, 97, 108, 111, 117
- Ren, B., Liu, M., Ding, R., and Liu, H. (2020). A survey on 3d skeleton-based action recognition using learning method. *arXiv preprint arXiv:2002.05907*. 57

- Roady, R., Hayes, T. L., Kemker, R., Gonzales, A., and Kanan, C. (2020). Are open set classification methods effective on large-scale datasets? *Plos one*, 15(9):e0238302. 78, 79
- Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A., and Maragos, P. (2016). Multimodal human action recognition in assistive human-robot interaction. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2702–2706. IEEE. 2, 7, 57, 96
- Roitberg, A., Perzylo, A., Somani, N., Giuliani, M., Rickert, M., and Knoll, A. (2014). Human activity recognition in the context of industrial human-robot interaction. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10. IEEE. 2, 7, 57, 96
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. 3, 108, 111, 123, 165
- Rosasco, A., Berti, S., Pasquale, G., Malafronte, D., Sato, S., Segawa, H., Inada, T., and Natale, L. (2024). Concon-chi: Concept-context chimera benchmark for personalized vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. v, 6, 26, 34, 35, 36, 38, 39, 127, 128
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dream-Booth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510. 35, 39, 111, 113, 114, 122, 123, 124, 164
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252. 163
- Ruzzetti, E. S., Ranaldi, L., Mastromattei, M., Fallucchi, F., Scarpato, N., and Zanzotto, F. M. (2022). Lacking the Embedding of a Word? Look it up into a Traditional Dictionary. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2651–2662, Dublin, Ireland. Association for Computational Linguistics. 110
- Saito, K., Sohn, K., Zhang, X., Li, C.-L., Lee, C.-Y., Saenko, K., and Pfister, T. (2023). Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19305–19314. 35, 38, 109, 111, 117, 118, 163
- Sáráandi, I., Linder, T., Arras, K. O., and Leibe, B. (2020). Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30. 67, 73
- Shahrudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019. 23, 84

- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics. 163
- Shi, Y., Wu, X., and Lin, H. (2022). Knowledge prompting for few-shot action recognition. *arXiv preprint arXiv:2211.12030*. 79
- Shi, Y., Wu, X., Lin, H., and Luo, J. (2024). Commonsense knowledge prompting for few-shot action recognition in videos. *IEEE Transactions on Multimedia*. 79
- Sigurdsson, G. A., Russakovsky, O., and Gupta, A. (2017). What actions are needed for understanding human actions in videos? In *Proceedings of the IEEE international conference on computer vision*, pages 2137–2146. 97
- Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer. 33, 95, 101
- Simmons, R., Goodwin, R., Haigh, K. Z., Koenig, S., and O’Sullivan, J. (1997). A layered architecture for office delivery robots. In *Proceedings of the First International Conference on Autonomous Agents*, pages 245–252. ACM. 43
- Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. 13
- Song, N., Zhang, C., and Lin, G. (2022). Few-shot open-set recognition using background as unknowns. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5970–5979. 77, 80
- Soomro, K. and Zamir, A. R. (2014). Action recognition in realistic sports videos. In *Computer vision in sports*, pages 181–208. Springer. 57
- Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*. 23
- Stretcu, O., Vendrow, E., Hata, K., Viswanathan, K., Ferrari, V., Tavakkol, S., Zhou, W., Avinash, A., Luo, E., Alldrin, N. G., Bateni, M., Berger, G., Bunner, A., Lu, C.-T., Rey, J., DeSalvo, G., Krishna, R., and Fuxman, A. (2023). Agile Modeling: From Concept to Classifier in Minutes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22323–22334. 108
- Su, B., Zhang, H., Li, J., and Zhou, Z. (2024). Toward generalized few-shot open-set object detection. *IEEE Transactions on Image Processing*, 33:1389–1402. 77
- Sucholutsky, I., Zhao, B., and Griffiths, T. L. (2024). Using compositionality to learn many categories from few examples. 95

- Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. (2019). A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics. 111
- Sun, L.-Y. and Chu, W.-T. (2024). Overall positive prototype for few-shot open-set recognition. *Pattern Recognition*, 151:110400. 77, 80
- Sun, X., Hu, P., and Saenko, K. (2022). Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582. 26, 29, 30, 94, 95, 99, 100, 101
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., and Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208. 13
- Tang, Y., Béjar, B., and Vidal, R. (2024). Semantic-aware video representation for few-shot action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6458–6468. 3, 7, 16, 17, 21, 77, 78, 79, 81, 85
- Tewel, Y., Gal, R., Chechik, G., and Atzmon, Y. (2023). Key-Locked Rank One Editing for Text-to-Image Personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH ’23, New York, NY, USA. Association for Computing Machinery. 111, 113, 114, 115, 123, 124, 165, 166
- Thatipelli, A., Narayan, S., Khan, S., Anwer, R. M., Khan, F. S., and Ghanem, B. (2022). Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967. 3, 7, 13, 77, 78, 79, 81, 85
- Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093. 17, 81, 85
- Trivedi, N., Thatipelli, A., and Sarvadevabhatla, R. K. (2021). Ntu-x: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9. 67
- Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2021). Open-set recognition: A good closed-set classifier is all you need? 8, 78, 79, 87
- Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. 13
- Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.-J., Fei-Fei, L., and Hays, J. (2019). Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6439–6448. 109

- von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>. 123, 164
- Wang, H., Pang, G., Wang, P., Zhang, L., Wei, W., and Zhang, Y. (2023a). Glocal energy-based learning for few-shot open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7507–7516. 77, 78, 80
- Wang, L., Liu, J., and Koniusz, P. (2021a). 3d skeleton-based few-shot action recognition with janie is not so naive. *arXiv preprint arXiv:2112.12668*. 58, 60
- Wang, M., Xing, J., and Liu, Y. (2021b). Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*. 95, 96
- Wang, P., Manhardt, F., Minciullo, L., Garattoni, L., Meier, S., Navab, N., and Busam, B. (2021c). Demograsp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5733–5740. IEEE. 3, 94, 95, 96
- Wang, Q. and Chen, K. (2020). Multi-label zero-shot human action recognition via joint latent ranking embedding. *Neural networks*, 122:1–23. 97
- Wang, X., Yan, Y., Hu, H.-M., Li, B., and Wang, H. (2024a). Cross-modal contrastive learning network for few-shot action recognition. *IEEE Transactions on Image Processing*. 77, 79
- Wang, X., Ye, W., Qi, Z., Zhao, X., Wang, G., Shan, Y., and Wang, H. (2021d). Semantic-guided relation propagation network for few-shot action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 816–825. 77, 79
- Wang, X., Zhang, S., Cen, J., Gao, C., Zhang, Y., Zhao, D., and Sang, N. (2024b). Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 132(6):1899–1912. 77, 79
- Wang, X., Zhang, S., Qing, Z., Gao, C., Zhang, Y., Zhao, D., and Sang, N. (2023b). Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18011–18021. 77, 79
- Wang, X., Zhang, S., Qing, Z., Zuo, Z., Gao, C., Jin, R., and Sang, N. (2024c). Hyrsm++: Hybrid relation guided temporal set matching for few-shot action recognition. *Pattern Recognition*, 147:110110. 77, 79
- Wanyan, Y., Yang, X., Chen, C., and Xu, C. (2023). Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502. 77, 79
- Weng, Z., Yang, X., Li, A., Wu, Z., and Jiang, Y.-G. (2023). Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning*, pages 36978–36989. PMLR. 26, 29, 95, 98, 102

- Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., and Feris, R. (2021). Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317. 111
- Wu, J., Zhang, T., Zhang, Z., Wu, F., and Zhang, Y. (2022). Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160. 77, 79
- Wu, L., Wang, Y., Yin, H., Wang, M., and Shao, L. (2019). Few-shot deep adversarial learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 29:1233–1245. 58
- Xing, J., Wang, M., Ruan, Y., Chen, B., Guo, Y., Mu, B., Dai, G., Wang, J., and Liu, Y. (2023). Boosting few-shot action recognition with graph-guided hybrid matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1740–1750. 77, 79
- Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., and Huang, T. (2018). YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 603–619, Cham. Springer International Publishing. 111
- Yan, S., Zhu, T., Wang, Z., Cao, Y., Zhang, M., Ghosh, S., Wu, Y., and Yu, J. (2022). Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*. 33, 102, 103
- Yeh, C.-H., Russell, B., Sivic, J., Heilbron, F. C., and Jenni, S. (2023). Meta-Personalizing Vision-Language Models To Find Named Instances in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19123–19132. 111
- Yin, X., Cao, B., Hu, Q., and Wang, Q. (2024). Rd-openmax: Rethinking openmax for robust realistic open-set recognition. *IEEE Transactions on Neural Networks and Learning Systems*. 80
- Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep Sets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 162
- Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P. H., and Koniusz, P. (2020). Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 525–542. Springer. 84
- Zhang, Y., Li, X., and Marsic, I. (2021). Multi-label activity recognition using activity-specific features and activity correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14625–14635. 96

- Zhao, C., Du, D., Hoogs, A., and Funk, C. (2023). Open set action recognition via multi-label evidential learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22982–22991. 97
- Zhu, G., Xu, C., Huang, Q., Gao, W., and Xing, L. (2006). Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 431–440. 57
- Zhu, Y., Fan, L., and Team, N. G. (2025). Nvidia isaac gr00t n1: An open foundation model for humanoid robots. NVIDIA Research. Accessed: 2026-04-21. 48, 128

Appendix A

Supplementary Material for Chapter 6

A.1 Implementation Details

SAFSAR 5-way 5-shot Since the authors [36] did not release the code, we implemented this method from scratch. We distributed the VideoMAE model over 4 GPUs. We used the pretrained weights of VideoMAE on the Kinetics-400 dataset provided by [38] for all experiments. We used 5 queries for each episodic task T_i . We fixed the weight λ for the $\mathcal{L}_2$ loss of SAFSAR to 0.1. The values for α_{EOS} and α_{Disc} are fixed to 0.1 and 1000 respectively.

SAFSAR 5-way 1-shot In this case we did parallel training over 4 GPUs, so we consider 20 queries for each episodic task T_i . All parameters are set as in the 5-shot case, except for the learning rates and the number of iterations.

STRM 5-way 5-shot As for SAFSAR 5-way 5-shot, we trained the STRM model by distributing the feature extractor over 4 GPUs. The value of δ_{Disc} is fixed to 10. As in [38], for each episode we consider $4 \times K$ queries (4 queries for each class) and we accumulate gradients for 16 iterations before updating the parameters. For standard datasets –SSv2, HMDB, UCF– we used the same learning rates and number of iterations reported in the original paper [38].

Table A.1 Training iterations and learning rates across datasets for SAFSAR 5-way 5-shot

	SSv2	Diving	NTU	HMDB	UCF
Iterations	50K	30K	15K	10K	10K
LR	$4e^{-6}$	$4e^{-6}$	$4e^{-6}$	$4e^{-6}$	$4e^{-6}$

Support Set Classes: • Wipe face • Put object into bag • Arm swings • Drink water • Drop	Unknown query: pick up 	known score (%) SAFSAR: 50.3% SAFSAR FR-Disc: 0.1%
Support Set Classes: • Jump up • Put on telephone • Fold paper • Shake fist • Cross toe touch	Unknown query: ball up paper 	known score (%) SAFSAR: 71.44% SAFSAR FR-Disc: 15.16%
Support Set Classes: • Put on jacket • Fold paper • Take off a shoe • Open bottle • Flick hair	Unknown query: play magik cube 	known score (%) SAFSAR: 61.12% SAFSAR FR-Disc: 3.68%
Support Set Classes: • Bounce ball • Flick hair • Cross arms • Open bottle • Hush	Unknown query: ball up paper 	known score (%) SAFSAR: 61.49% SAFSAR FR-Disc: 0.45%

Figure A.1 (not cherry picked). We represent 4 equidistant frame among the 8 used for inference. We assume that a query is classified as *known* if the known score is above 50% and as *unknown* otherwise. A red label means that the prediction is incorrect, while a green label means that the prediction is correct.

Table A.2 Training iterations and learning rates across datasets for SAFSAR 5-way 1-shot

	SSv2	Diving	NTU	HMDB	UCF
Iterations	50K	30K	15K	10K	10K
LR	$4e^{-7}$	$4e^{-7}$	$4e^{-7}$	$4e^{-8}$	$4e^{-8}$

Table A.3 Training iterations and learning rates across datasets for STRM 5-way 5-shot

	SSv2	Diving	NTU	HMDB	UCF
Iterations	75K	75K	40K	20K	20K
LR	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-4}$	$1e^{-4}$

Support Set Classes:

- Pulling two ends of * so that it gets stretched
- Turning the camera downwards while filming *
- Turning the camera left while filming *
- Moving * and * closer to each other
- Putting * similar to other things that are already on the table

Unknown query: turning the camera upwards while filming *



known score (%)

SAFSAR: 87.76%

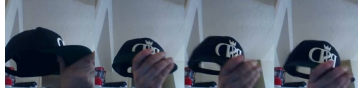
SAFSAR: 0.72%

FR-Disc:

Support Set Classes:

- Tipping * with * in it, so * in it falls out
- Moving away from * with your camera
- Putting * similar to each other things that are already on the table
- Pretending to take * out of *
- Putting number of * onto *

Unknown query: holding *



known score (%)

SAFSAR: 87.96%

SAFSAR: 7.28%

FR-Disc:

Support Set Classes:

- Plugging * into *
- Moving * and * closer to each other
- Turning the camera upwards while filming *
- Moving * across a surface until it falls down
- Pulling two ends of * so that it separates into two pieces

Unknown query: pretending to take * out of *



known score (%)

SAFSAR: 92.90%

SAFSAR: 19.47%

FR-Disc:

Support Set Classes:

- Tilting * with * on it until it falls off
- Pretending to spread air onto *
- Moving * and * closer to each other
- Pouring * onto *
- Spilling * onto *

Known query: pouring * onto *



known score (%)

SAFSAR: 22.42%

SAFSAR: 88.46%

FR-Disc:

Figure A.2 (not cherry picked). We represent 4 equidistant frame among the 8 used for inference. We assume that a query is classified as *known* if the known score is above 50% and as *unknown* otherwise. A red label means that the prediction is incorrect, while a green label means that the prediction is correct.

A.1.1 MLS vs MSS

The importance of not-normalized logits. In Table A.4, we report the performance gain of SAFSAR for open-set metrics using MLS instead of MSS, where the MLS values are reported in the paper. We note that the MLS score consistently outperforms the MSS score for implicit methods. This confirms the result described in [3, 39] for SAFSAR on spatio-temporal data. Indeed, the SoftMax operation discards the magnitude information that is contained in the logits $\{\hat{s}_{ij}\}_{j=1}^K$.

The value $\hat{u}_i$ can be derived from the logits $\{\hat{s}_i\}_j$ for SAFSAR because this model uses the cosine similarity as similarity function, whose range is in $[-1, 1]$ and can be interpreted directly as probability. We cannot apply the same ablation to STRM because it uses the

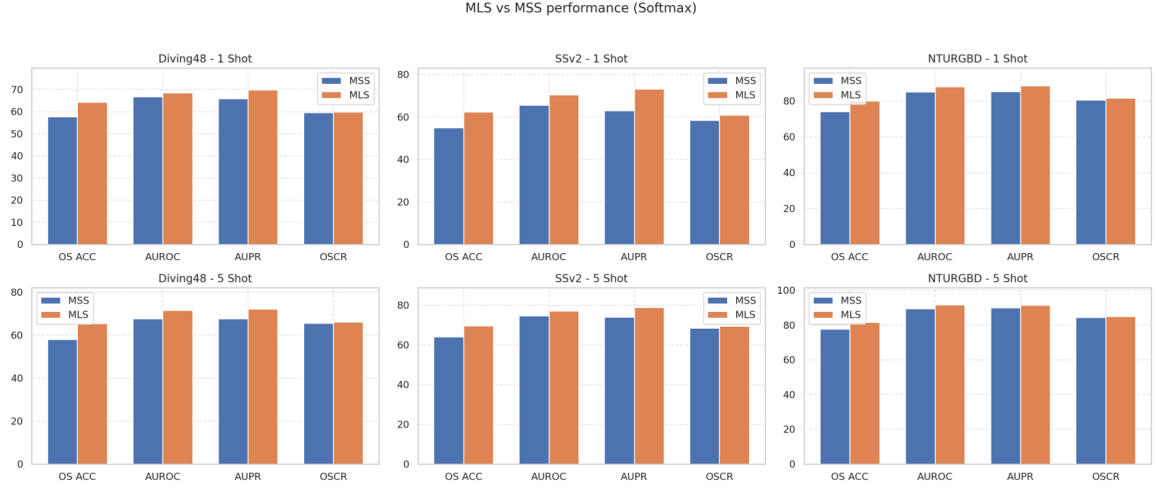


Figure A.3 Open-set performances comparison between MLS and MSS for the SAFSAR model.

Table A.4 The gain for open-set metrics when using MLS instead of MSS for implicit methods on SAFSAR. Gains are colored: green for positive and red for negative. "*" means that we trained that model for 1K iterations to prevent overfitting.

Dataset	OS-Method	OS ACC		AUROC		AuPR		OSCR	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Diving48	Softmax	+6.57	+7.44	+1.80	+3.93	+3.86	+4.54	+0.34	+0.71
	EOS	+3.37	+3.09	+0.04	-0.01	+1.96	+0.96	+0.04	-0.10
SSv2	Softmax	+7.43	+5.57	+4.90	+2.48	+10.05	+4.93	+2.32	+0.90
	EOS	+4.65	+2.28	+0.69	-0.08	+4.62	+2.89	+0.42	+0.30
NTURGBD	Softmax	+5.99	+3.93	+2.83	+2.15	+3.25	+1.40	+0.95	+0.50
	EOS	+3.59	+5.19	+0.58	+2.15	+1.59	+2.40	+0.23	+0.61
HMDB51	Softmax*	+14.44	+22.40	+4.02	+4.10	+3.34	+8.73	+0.44	+1.84
	EOS*	+16.23	+22.48	+5.71	+3.23	+4.67	+8.16	+1.15	+1.57
UCF101	Softmax*	+30.59	+41.25	+1.09	+0.46	+0.79	+1.27	+0.38	+0.34
	EOS*	+31.30	+39.32	+1.29	+0.32	+1.09	+0.70	+0.39	+0.16

negative norm of the difference between the query features and the support set classes features, whose range is in $[-\infty, 0]$ and thus can not be interpreted directly as probability.

A.2 Metrics Definition

A.2.1 AUROC

AUROC (Area Under the ROC Curve) is defined on T^{test} as:

$$\text{AUROC} = \int_0^1 \text{TPR} d\text{FPR},$$

where TPR (True Positive Rate) and FPR (False Positive Rate) are computed for the considered methods by varying the threshold $\tau \in (0, 1)$ (except for the Garbage Class that does not use τ) as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

and true positives TP are known queries classified as known, false positives FP are unknown queries classified as known, true negatives TN are unknown queries classified as unknown and FN are known queries classified as unknown.

A.2.2 AUPR

AUPR (Area Under the Precision-Recall Curve) is defined on T^{test} as:

$$\text{AUPR} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall},$$

where Precision and Recall are computed for the considered methods by varying the threshold $\tau \in (0, 1)$ (except for the Garbage Class that does not use τ) as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \text{TPR}$$

A.2.3 FS ACC

FS ACC (Few-Shot Accuracy) or closed-set accuracy is defined on $T_{\text{known}}^{\text{test}}$ as:

$$\text{FS ACC} = \frac{1}{|T_{\text{known}}^{\text{test}}|} \sum_{(SS_i, Q_i^{\text{known}}) \in T_{\text{known}}^{\text{test}}} \mathbb{I}[\hat{y}_i = y_i] \quad (\text{A.1})$$

where $\hat{y}_i$ is predicted class by $f(SS_i, x_i)$.

A.2.4 OS ACC

OS ACC (Open-Set Accuracy) is defined on T^{test} as:

$$\text{OS ACC} = \frac{1}{|T^{\text{test}}|} \sum_{(SS_i, Q_i) \in T^{\text{test}}} \mathbb{I}[\hat{a}_i = a_i] \quad (\text{A.2})$$

where a_i is the true known/unknown class and $\hat{a}_i$ is the predicted known/unknown class for that sample. For the considered methods (except for GC) the threshold was fixed at $\tau = 0.5$.

A.2.5 OSCR

OSCR (Open-Set Classification Rate) is defined on $T_{\text{known}}^{\text{test}}$ as:

$$\text{CCR} = \frac{1}{|T_{\text{known}}^{\text{test}}|} \sum_{(SS_i, Q_i) \in T_{\text{known}}^{\text{test}}} \mathbb{I}[\hat{y}_i = y_i] \mathbb{I}[\hat{a}_i = a_i] \quad (\text{A.3})$$

$$\text{OSCR} = \int_0^1 \text{CCR} d\text{FPR} \quad (\text{A.4})$$

where CCR is the Correct Classification Rate on the P_{FS} problem. The OSCR thus represents the area under the CCR-FPR curve, by varying the threshold $\tau \in (0, 1)$ for the considered methods (except for GC). This dual evaluation is particularly important for few-shot, where correctly identifying known actions and detecting unknown ones are both crucial requirements.

The integral of the metrics that are computed for all threshold values –AUROC, AUPR, OSCR– is approximated using the trapezoidal rule.

A.3 Discriminator Architecture

A.3.1 SAFSAR

Input shape: 1×768

- Linear Layer $768 \rightarrow 1536$
- ReLU
- Linear Layer $1536 \rightarrow 768$
- ReLU

- Linear Layer $768 \rightarrow 64$
- ReLU
- Linear Layer $64 \rightarrow 1$
- Sigmoid

A.3.2 STRM

Input shape: 28×1152

- Linear Layer $1152 \rightarrow 512$
- BatchNorm1d
- ReLU
- Dropout(0.3)
- Linear Layer $512 \rightarrow 256$
- BatchNorm1d
- ReLU
- Dropout(0.3)
- Linear Layer $256 \rightarrow 128$
- BatchNorm1d
- ReLU
- Dropout(0.3)
- Linear Layer $128 \rightarrow 64$
- ReLU
- Reshape $28 \times 64 \rightarrow 1792$
- Linear Layer $1792 \rightarrow 1$
- Sigmoid

A.4 Splits

A.4.1 Diving48

Training split

Forward_15som_NoTwis_PIKE, Inward_15som_NoTwis_TUCK, Forward_25som_1Twis_PIKE, Back_3som_NoTwis_PIKE, Reverse_35som_NoTwis_TUCK, Forward_25som_NoTwis_TUCK, Forward_45som_NoTwis_TUCK, Reverse_25som_NoTwis_PIKE, Forward_15som_2Twis_FREE, Back_15som_NoTwis_TUCK, Forward_25som_NoTwis_PIKE, Back_35som_NoTwis_TUCK, Reverse_25som_15Twis_PIKE, Reverse_15som_35Twis_FREE, Forward_35som_NoTwis_PIKE, Back_25som_15Twis_PIKE, Inward_35som_NoTwis_TUCK, Inward_15som_NoTwis_PIKE, Back_2som_25Twis_FREE, Back_Dive_NoTwis_PIKE, Back_25som_NoTwis_PIKE, Back_25som_25Twis_PIKE, Inward_25som_NoTwis_PIKE, Reverse_15som_NoTwis_PIKE, Back_15som_15Twis_FREE, Back_25som_NoTwis_TUCK, Forward_1som_NoTwis_PIKE, Back_35som_NoTwis_PIKE, Reverse_15som_05Twis_FREE, Back_3som_NoTwis_TUCK, Back_15som_25Twis_FREE, Back_15som_NoTwis_PIKE

Testing split

Reverse_15som_15Twis_FREE, Back_Dive_NoTwis_TUCK, Reverse_25som_NoTwis_TUCK, Forward_15som_1Twis_FREE, Reverse_Dive_NoTwis_PIKE, Back_15som_05Twis_FREE, Forward_25som_3Twis_PIKE, Reverse_Dive_NoTwis_TUCK, Back_2som_15Twis_FREE, Inward_25som_NoTwis_TUCK, Reverse_15som_25Twis_FREE, Forward_25som_2Twis_PIKE, Inward_Dive_NoTwis_PIKE, Forward_35som_NoTwis_TUCK, Forward_Dive_NoTwis_PIKE

A.4.2 NTURGBD120

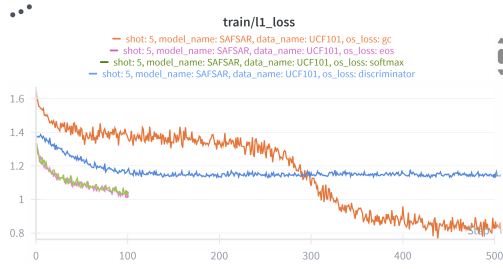
Training split

thumb_down, reach_into_pocket, taking_a_selfie, take_off_glasses, wipe_face, cutting_nails, bounce_ball, phone_call, apply_cream_on_hand, move_heavy_objects, writing, snap_fingers, stand_up, jump_up, thumb_up, side_kick, type_on_a_keyboard, open_a_box, play_magic_cube, reading, juggle_table_tennis_ball, tear_up_paper, put_on_bag, throw, put_on_glasses, arm_swings, shake_head, salute, take_off_bag, kicking_something, throw_up_cap_hat, put_object_into_bag, hush, rub_two_hands, toss_a_coin, apply_cream_on_face, open_bottle, put_on_jacket, cross_toe_touch, ball_up_paper, counting_money, cross_arms,

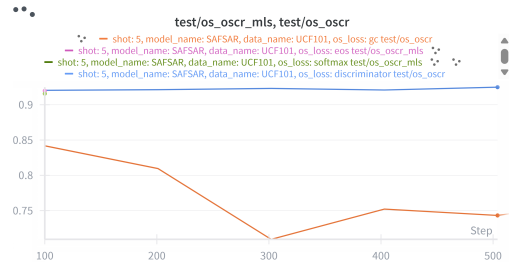
flick_hair,nod_head,make_OK_sign,cross_hands_in_front,shake_fist,take_off_a_shoe,sniff,pick_up,take_object_out_of_bag,drink_water,put_on_headphone,fold_paper,drop,squat_down,take_off_a_hat

Testing split

hopping,play_with_phone,make_victory_sign,brush_teeth,check_time_from_watch,arm_circles,clapping,point_to_something,put_palms_together,take_off_jacket,take_off_headphone,staple_book,sit_down,run_on_the_spot,shoot_at_basket,put_on_a_hat,brush_hair,butt_kicks,hand_waving,capitulate,put_on_a_shoe,cutting_paper,tennis_bat_swing,eat_meal,cheer_up



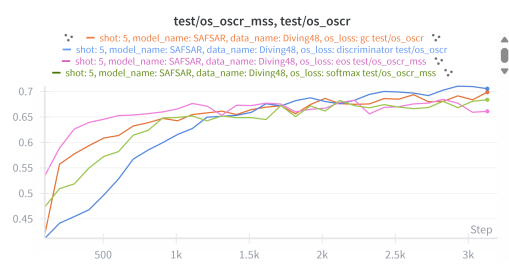
(a) 5-shot SAFSAR on UCF101 - training closed-set loss



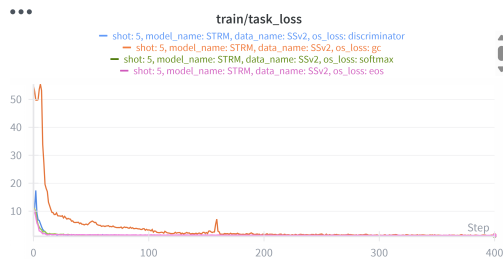
(b) 5-shot SAFSAR on UCF101 - test metric OSCR



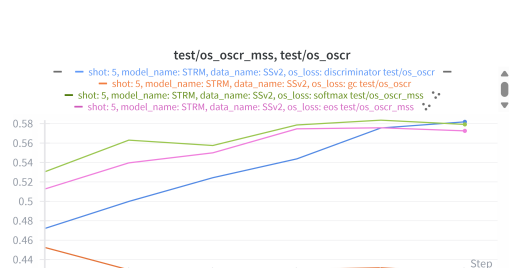
(c) 5-shot SAFSAR on Diving48 - training closed-set loss



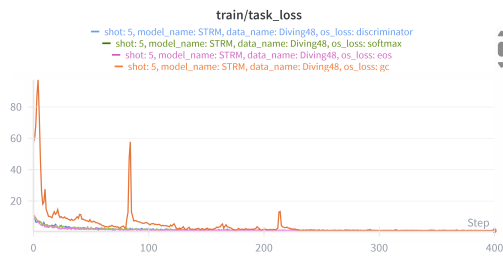
(d) 5-shot SAFSAR on Diving48 - test metric OSCR



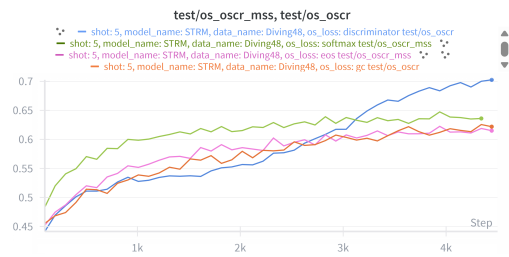
(e) 5-shot STRM on SSv2 - training closed-set loss



(f) 5-shot STRM on SSv2 - test metric OSCR



(g) 5-shot STRM on Diving48 - training closed-set loss



(h) 5-shot STRM on Diving48 - test metric OSCR

Figure A.4, showing both training closed-set loss and test OSCR metrics. We shot that in some cases (Subfigures (a) and (e)) the training Garbage Class loss converges, but the testing OSCR metric decreases (Subfigures (b) and (f)), indicating an overfitting. Instead, in other cases (Subfigures (c) and (g)) the training Garbage Class loss decreases, but the testing OSCR metric increase indicating a correct training. We argue that the Garbage prototype initialization plays an important role in the learning phase, and that this experiment would need different initialization or normalization techniques.

Appendix B

Supplementary Material for Chapter 8

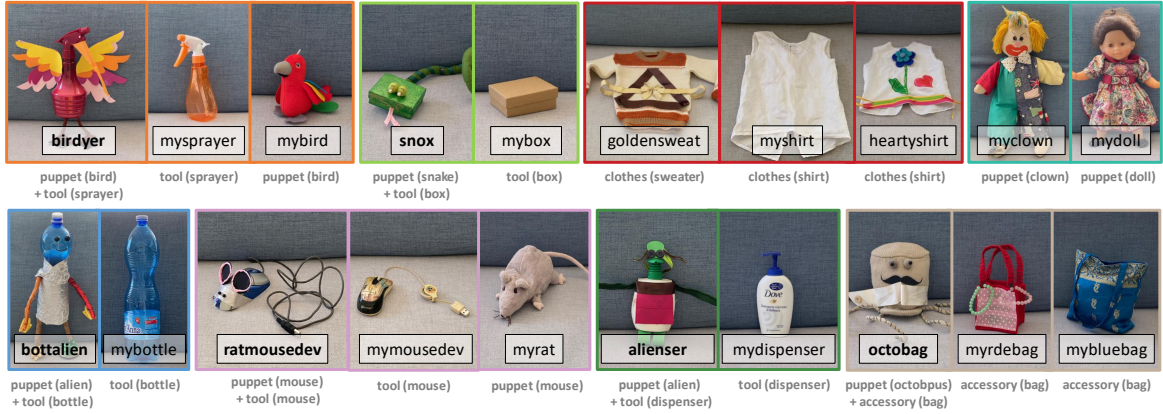


Figure B.1 **Concepts in ConCon-Chi**. Images from one of the five training set environments. The concepts are grouped together with their hard negatives and the chimeric ones are indicated in (**Bold**).

In this supplementary material we first include additional information about the proposed ConCon-Chi dataset (B.1), then provide additional details about the personalized TIR (B.2) and TIG (B.3) experiments.

B.1 Concept-Context Chimera Benchmark

We provide additional information about the concepts (B.1.1) and the contexts (B.1.2) in the dataset; then we provide more details about the acquisition and annotation procedure (B.1.3) and the proposed benchmark splits (B.1.4). Finally we provide examples of queries/prompts and associated GT images in the dataset (B.1.5).

B.1.1 Concepts: Pictures and Descriptions

In B.1 we show all the 20 concepts in ConCon-Chi together with their name and category. Chimeric concepts are indicated with a Bold concept name. Other than the 6 chimeras, 4 other objects are considered “animate” (MYDOLL, MYCLOWN, MYBIRD, MYRAT). This type of concept can combine with contexts that require agency (e.g., “sitting” or “playing”) and with inanimate concepts. In B.1 we report the three types of concept descriptions used for the TIR and TIG baselines.

B.1.2 Contexts: Kinds and Environments

Contexts are grouped into nine kinds:

- **re-contextualization**: the concept appears in a different environment;
- **view**: the concept is photographed by a different view (e.g., back, profile, close-up on details);
- **action**: the concept is carrying out an action (e.g., sitting, lying down);
- **accessorization**: the concept is wearing an accessory (e.g., sunglasses, eye-mask)
- **property modification**: the concept is missing some part (e.g., legs, wings, ears) or some part of the concept is applied to another one;
- **interaction**: the concept is interacting with a person, an object or another concept learned in the same way;
- **rendition**: black and white photography, ink or pencil sketch, baby drawing, blurred painting, dark glow-edge effect.

Contexts can be, and in most of the cases are, tagged with multiple kinds. This is visually depicted in B.2a. From the figure it can be observed that the most frequent kinds are “object interaction” and “recontextualization” which also appear in combination with most of the other kinds. It is also interesting to notice how the 18 contexts tagged as “concept interaction” result in a high number of queries. This is because “concept interaction” contexts allow for numerous concept combinations.

Contexts are also tagged with the environment in which the corresponding ground-truth images were captured. Half of the contexts (48) were captured in a neutral setting (a wooden surface). The other 38 contexts were captured in different environments, coming from two

main location types: houses and office spaces. The environment distribution is shown in B.2b.

B.1.3 Acquisition and Annotation: Procedure Details

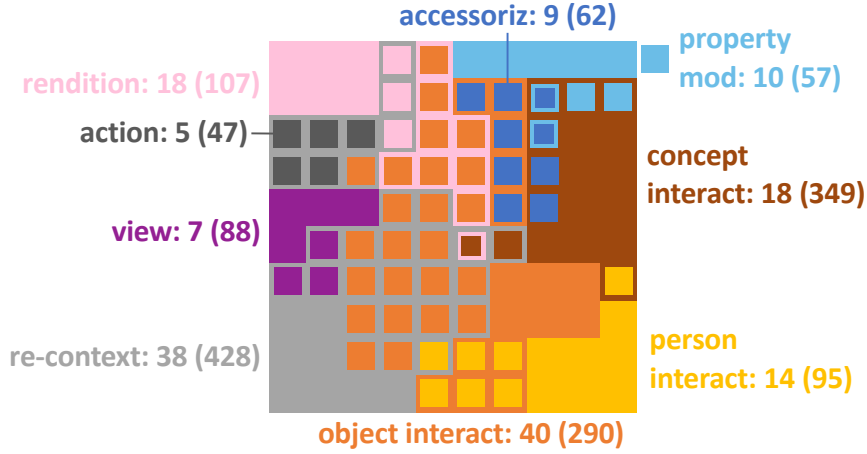
Acquisition. We took the images using an iPhone 11 Pro with resolution 4032×3024 and variable focal length. The released pictures were converted to JPEG at a resolution of 1008×756 . Part of the renditions was obtained by applying digital effects (ink/pencil sketch, grayscale, blur, dark glow edges) to pictures of 11 concepts manually selected in the image pool. The remaining part contains actual pen drawings of the concepts sketched on paper and later photographed.

Annotation. The annotation process was divided in two steps. First we used an annotation GUI developed by us to review the images and assign to each of them the correct concepts/context combination. This step produced, for each query, a list of ground-truth (GT) images. The distribution of the number of GTs per query is shown in B.3 (Blue bars). At this point each of the images is assigned as GT of only one query. However, since some contexts are more general than others, some images needed to be marked as GT for multiple queries. This is necessary to avoid false negatives in the dataset (images that are correct realization of a certain query but are not marked as its GT). To this end, we automatically assigned additional GT images following context overlaps, from more specific to more general queries (e.g., GT images of the query “black and white photography of X resting on a bedside table” are added as GT images for the query “X resting on a bedside table”). Then, the lists of GT images were further extended by manually checking individual GT images of queries which we knew could overlap with others (e.g., we added several images as GT for “X standing on a wooden surface”, where the concept was involved in an action or interaction, but in fact was also standing on the surface). In B.3 we show the resulting number of GT images per query in Magenta.

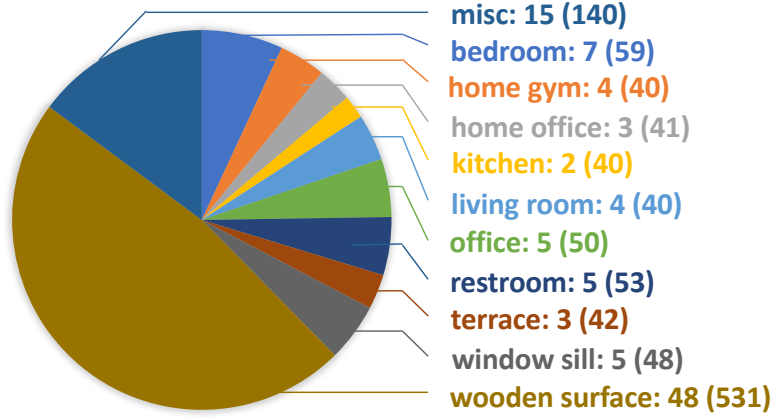
B.1.4 Benchmark Splits

We report below the splits that we introduce for the personalized TIR and TIG benchmarks (notice that in the main paper we use only the *train* and *test* splits):

- **train:** A set of 5 images per concept each with a different background. For the experiments with $k = 1$ in 8.4 we further split the training set into 5 different splits each with one image per concept.



(a) **Context kinds.** Each cell represents a context and is colored by kind (101 contexts arranged in a 10×10 matrix plus 1 cell at the top-right corner). If a concept belongs to more than a kind, 2 or 3 squares are overlapped in the cell with corresponding colors. The total number of contexts per kind is indicated next to the context name and equals the number of cells with the same color, with the corresponding number of queries in parenthesis.



(b) **Context environments.** The number of contexts per environment is indicated to the left with the corresponding number of queries in parenthesis (and they respectively sum to 101 and 1084).

Figure B.2 Context kinds and environments in ConCon-Chi. Panel (a) summarizes context kinds; panel (b) summarizes context environments.

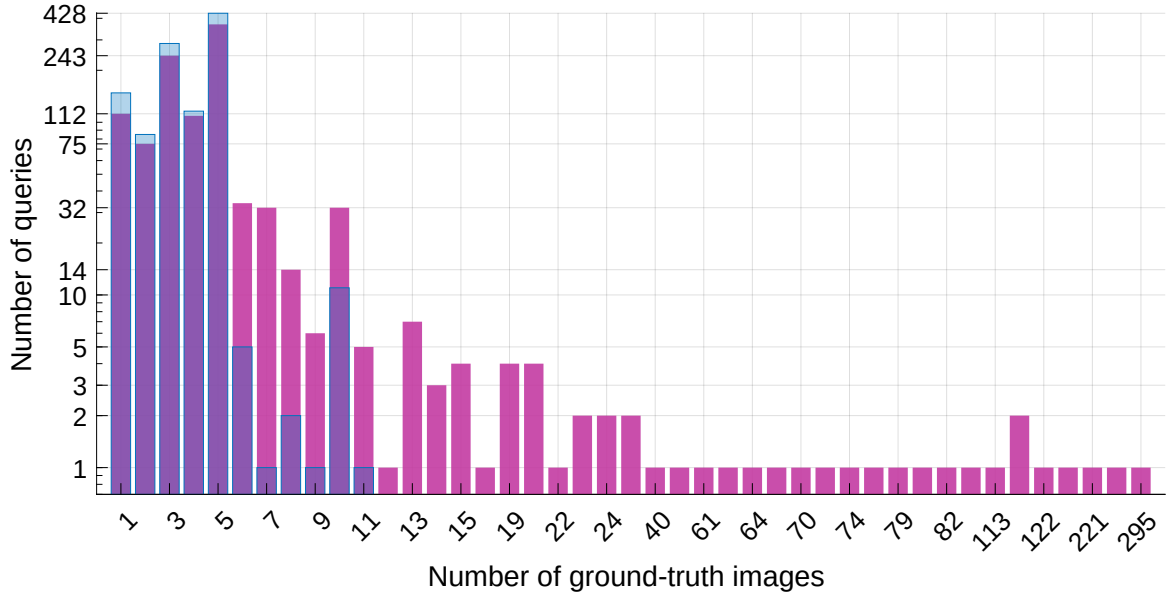


Figure B.3 **Ground-truth images per query.** The number of ground-truth images that were acquired to represent a given query is reported in Blue (with transparency), the number of overall images that were annotated as ground-truth for a given query is in Magenta.

- **val:** A validation set containing 3 concepts (TOOMOUSE, RATMOUSDEV and MYRAT), composed of the (42) queries that contain any of them (but none of the other concepts), set in the OFFICE or WOODEN SURFACE environments. The associated image pool is the union of the ground-truth images of these queries (165 images). Considering the very few-shot nature of the personalization tasks, we do not use the validation set. However, we define and release one since future methods might exploit it.
- **test:** A set that contains all the dataset minus the training images (1084 queries and 4008 images in the pool).
- **test-unseen:** A set from which we removed the queries that contain any of the 3 concepts used for validation (986 queries and 4008 images in the pool).

See 8.1 for the further statistics regarding the validation and the test splits.

B.1.5 Dataset Examples

In B.4 we show 6 examples of caption-image associations present in the dataset, for each of the 6 chimeric concepts (one concept per row). The caption is used as query in TIR and prompt in TIG tasks. In the same way, in B.5 we report 2 examples of caption-image

associations for each of the 14 common concepts (3 concepts per row and 2 in the last row). These two ensembles represent respectively examples of uncommon/novel and common situations, the former more focused on the learning of new-meanings and the latter more focused on the learning of instances of common categories, following related personalization benchmarks.

B.2 Personalized Text-to-Image Retrieval

In this section we provide definition of the TIR metrics adopted in the paper (B.2.1); we report implementation details about the experiment reported in 8.5 (B.2.2) and about the methods compared in the TIR benchmark (B.2.3); we report additional metrics for the TIR benchmark (B.2.4) and additional breakdown of performance (B.2.5) and finally some failure cases (B.2.6).

B.2.1 Metrics Definition

Given a set of queries Q and an image pool I we report below the definition for the adopted metrics:

mean Reciprocal Rank (mRR): the average of the reciprocal rank of the first retrieved ground-truth

$$mRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where $rank_i$ refers to the rank of the first retrieved ground-truth for the i -th query.

mean Average Precision at k (mAP@k): the average of the AP (Average Precision) at k, among the queries

$$mAP@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP_i@k$$

with

$$AP_i@k = \frac{1}{\min(k, |GTs_i|)} \sum_{m=1}^k P_i@m \times rel_i(m)$$

where $P_i@m$ (precision at m) is the fraction of ground-truth images (for the i -th query) among the top m retrieved ones, $rel_i(m)$ is an indicator function which is 1 if the m -th image in I is a ground-truth for the i -th query, and GTs_i is the set of ground-truth images for the i -th query.

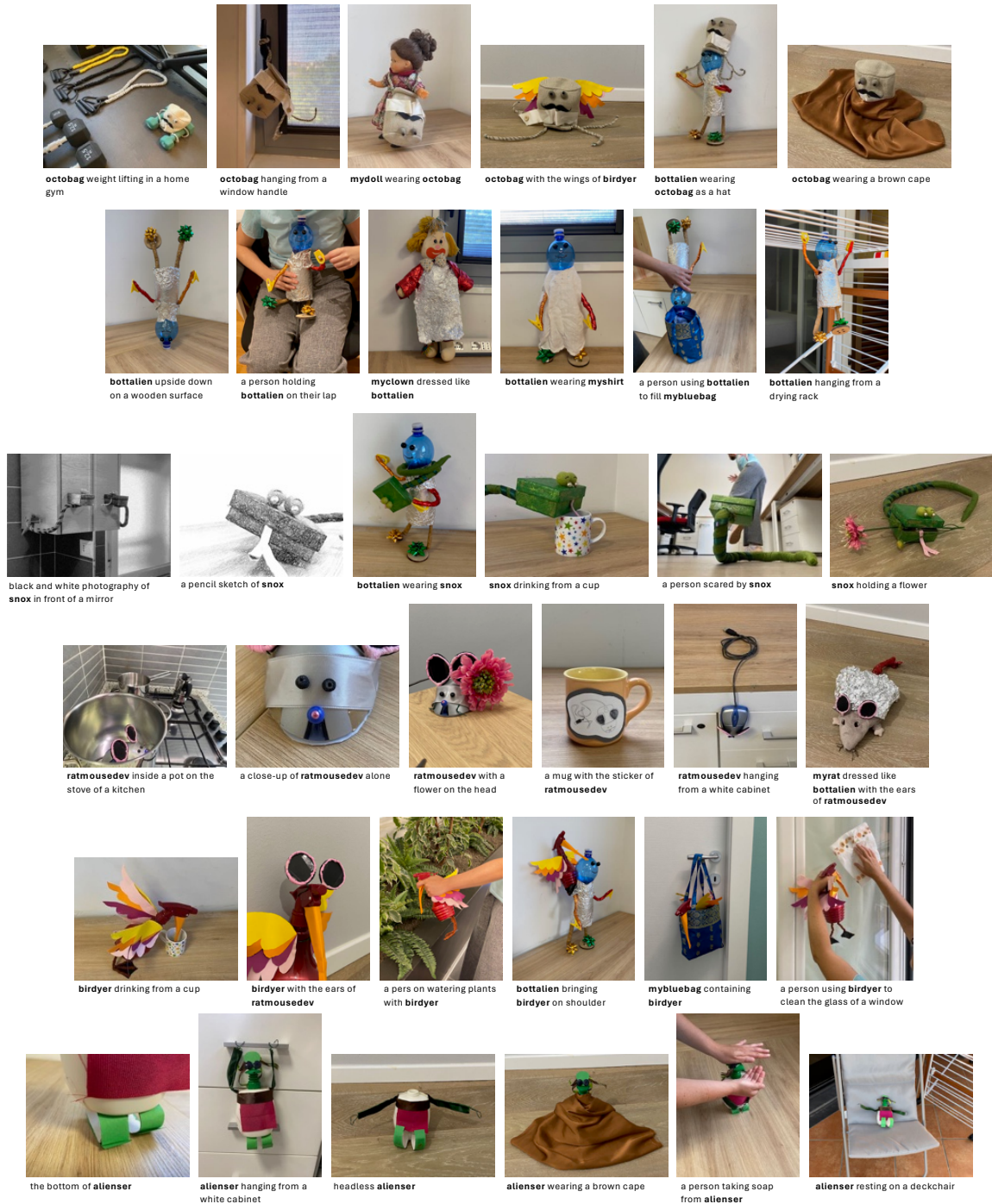


Figure B.4 Examples of caption-image associations involving the 6 chimeric concepts. One concept per row with six examples per concept. The caption is used as query in TIR and prompt in TIG tasks. The concepts are marked in Bold in the captions. Please note that for simplicity we report only one GT image per caption (and only one caption per image).

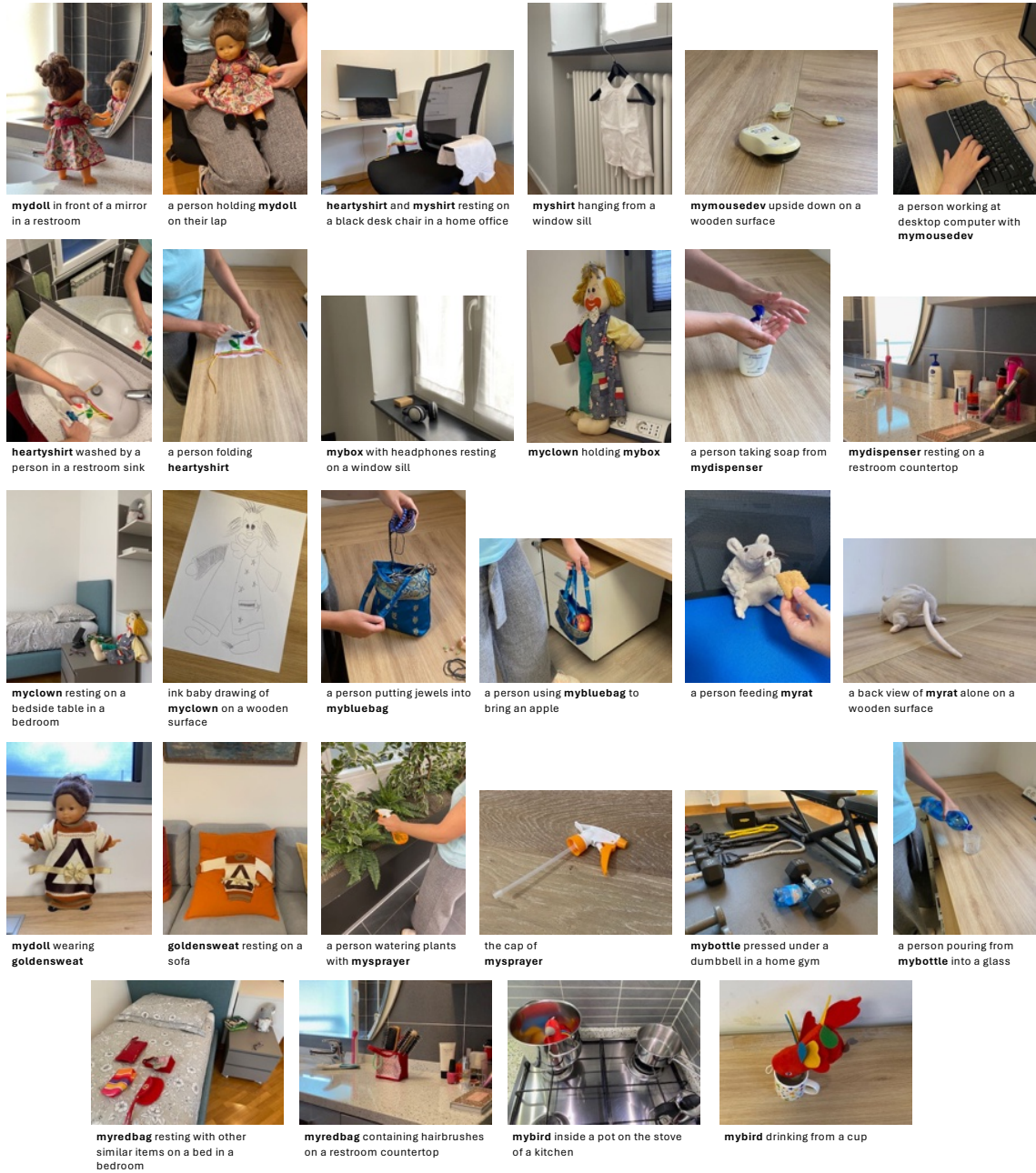


Figure B.5 Examples of caption-image associations involving the 14 common concepts. Three concepts per row and two in the last row. The caption is used as query in TIR and prompt in TIG tasks. The concepts are marked in Bold in the captions. Please note that for simplicity we report only one GT image per caption (and only one caption per image).

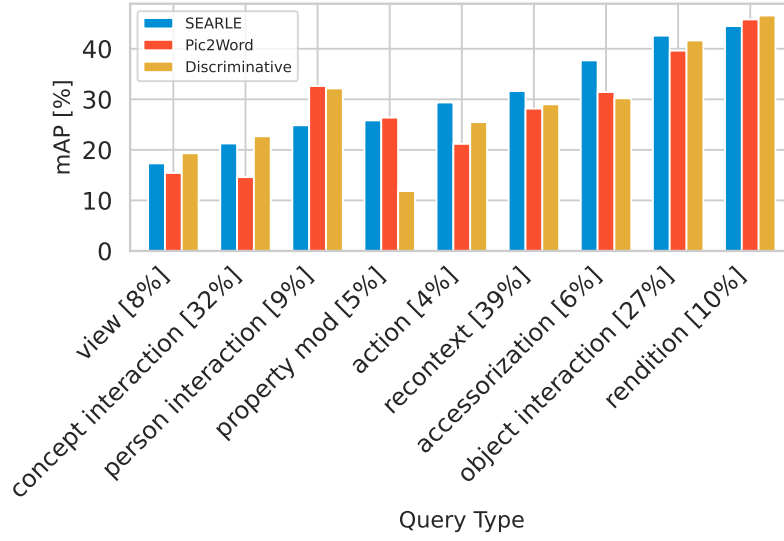


Figure B.6 **Performance per query kind.** Breakdown of the mAP performance in 8.2 per query kind for the two best performing methods and the best text-based baseline.

mean Average Precision (mAP): it is the mAP@k with $k = |I|$.

Recall at k (R@k): fraction of queries for which the rank of the first retrieved image is smaller than k

$$R@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbb{1}_{\{rank_i < k\}}$$

where $\mathbb{1}$ is 1 if $rank_i < k$ and 0 otherwise.

B.2.2 Concept and Context Bias: Implementation Details

Below we provide some details about the datasets compared in 8.5.

For FashionIQ, CIRRR, CIRCO we used the validation set as the test set annotations are not publicly available. For ConCon-Chi and PerVL DF2 we used the test set.

The PerVL DF2 dataset provides two types of captions: short and detailed. In the paper, the authors define the detailed captions as the ones that describe extensive context about the image and can facilitate retrieval, while the short ones describe less detail and therefore are more ambiguous. We used the detailed captions for our analysis.

The FashionIQ dataset provides images as lists of URLs. However, we found some of the links were broken, making it impossible to download the original dataset. In B.2 we report, for the validation set, the original number of URLs and the number of images that we couldn't download.



Figure B.7 **Retrieval Failure Cases.** First retrieved image for each of the 14 worst performing queries in terms of mAP for SEARLE (Left to Right, Top to Bottom). The part of the query successfully retrieved is highlighted in Green, otherwise in Red.

Since FashionIQ provides two relative captions for each reference image, as in (Baldrati et al., 2023) (Appendix A), we concatenate them.

B.2.3 Methods: Implementation Details

For each method we provide a brief description of how it works followed by some details on the choices we took to ensure a fair comparison.

PALAVRA (Cohen et al., 2022). This method uses a Deep Set function (Zaheer et al., 2017) to map the 5 CLIP image embeddings of the concept to a token embedding (inverse mapping). The function is pre-trained on thousands of frequent concepts from COCO (Lin et al., 2014) and used to provide an initial token value for the concept at hand. This is then fine-tuned with a so-called cycle contrastive loss, such that the CLIP text embedding of a template sentence containing the token is pushed closer to the average CLIP image embedding of the 5 images (cycle term) and farther from the embedding of a same sentence with the concept type replacing the token (contrastive term).

We used the code released by the authors ¹ and retrained the inverse mapping following author instructions. When doing so, we replaced the ViT-B/32 CLIP backbone with the larger ViT-L/14 for fair comparison with other methods. When fine-tuning the token embedding, we

¹<https://github.com/NVlabs/PALAVRA>

adopted hyper-parameters suggested by the authors and when using a single image example, we fed the Deep Set function with 5 copies of the single image embedding. As concept types in the fine-tuning, we used the *Coarse* descriptions.

Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023). These are two ZS-CIR methods which, as PALAVRA, learn an inverse mapping function for CLIP. In this case this is a three-layer MLP that maps a CLIP image embedding (input) to a corresponding token embedding (output) and is learned in a self-supervised fashion.

Pic2Word trains the function on 3M images from CC3M (Sharma et al., 2018) with a cycle contrastive loss that uses a given CLIP image embedding as cycle term and others as contrastive term. In SEARLE, first, a token embedding for each of 100K images in the unlabeled test split of ImageNet1K (Russakovsky et al., 2015) is optimized similarly to PALAVRA’s fine-tuning, with a regularization term such that the embedding is also kept close to the textual embeddings of concepts represented in the image. Then, since using this optimization procedure at evaluation time would be computationally expensive, SEARLE learns a mapping network that produces similar embeddings to the ones learnt through back-propagation. This mapping network is trained with a contrastive loss over the 100K optimized tokens, regularized in the same way. At inference time SEARLE architecture is basically identical to Pic2Word.

We use both methods by downloading pre-trained weights of the mapping functions based on CLIP ViT-L14, officially provided by the authors (Pic2Word², SEARLE³). Differently from ZS-CIR, in our setting 5 concept images are available, thus we average the generated token embeddings to create the concept embedding.

B.2.4 Additional Benchmark Metrics

To encourage future comparisons on our benchmark, in B.3 we report some additional retrieval metrics for the experiment reported in 8.2. In particular we show R@k and mAP@k with k=5, 10.

B.2.5 Additional Analysis of Results

In B.6 we report the performance per query kind. SEARLE is comparable or better than the baseline across all kinds but “person interaction”, where it is outperformed also by Pic2Word,

²https://github.com/google-research/composed_image_retrieval

³<https://github.com/miccunifi/SEARLE>

indicating that there may be a limitation specific to its training set. Conversely, the baseline drops heavily on “property modification” and “accessorization”.

B.2.6 TIR Failure Cases

We report in B.7 some failure cases for the top-performing retrieval method, SEARLE. To this end, we sort the queries by decreasing mAP and report the first retrieved image for the last 14 ones. In accordance to the results reported in 8.3, on these worst performing queries, SEARLE is almost always able to retrieve the correct concept, but fails to compose it with the correct context.

B.3 Personalized Text-to-Image Generation

In this section we provide implementation information about the Density and Coverage metrics (B.3.1), and the methods compared in the TIG benchmark (B.3.2); we then report numbers for the FID metrics (B.3.3) and some failure cases (B.3.4).

B.3.1 Metrics: Implementation Details

To compute Density and Coverage we used the code provided by the authors (Naeem et al., 2020) without substantial modifications⁴. We computed the two metrics on the CLIP embeddings of the real and generated images.

B.3.2 Methods: Implementation Details

DreamBooth (DB) (Ruiz et al., 2023). This method fine-tunes a pre-trained TIG model on the example images in order to bind a textual identifier (a chosen rare token followed by the concept class) to the concept appearance. A class-specific prior preservation loss acts as regularizer by forcing the model also to keep generating images of instances of the concept class when the input sentence does not contain the rare token but just the class.

We adopted the implementation provided⁵ in the Diffusers library (von Platen et al., 2022), release 0.18.0, with the Stable Diffusion Model (SDM) checkpoint v1.4⁶. We then

⁴<https://github.com/clovaai/generative-evaluation-prdc>

⁵<https://github.com/huggingface/diffusers/tree/main/examples/dreambooth>

⁶<https://huggingface.co/CompVis/stable-diffusion-v-1-4-original>

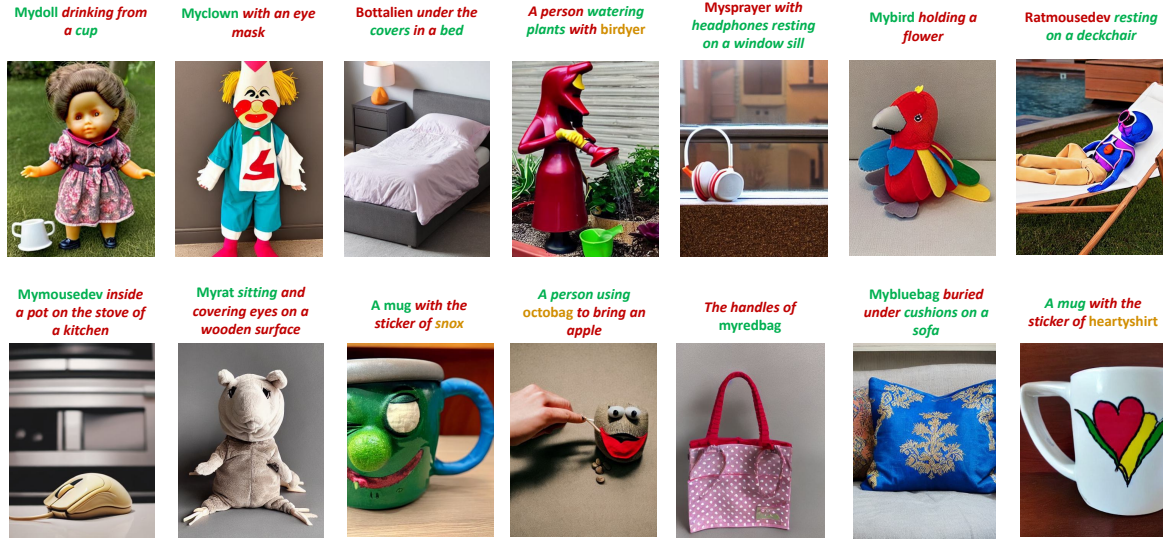


Figure B.8 **Generation Failure Cases.** Some failure cases for DreamBooth. The part of the query successfully generated is highlighted in Green; the part of the query that seems to be contained in the generated image, but is not completely correct, is in Orange; the part that is not represented in the generated images is in Red. While clearly this type of decision can be subjective, we aimed to provide our judgment as additional information.

fine-tuned the text encoder alongside the UNet, using the *Coarse* descriptions at initialization, with a learning rate of $1e-6$.

Textual-Inversion (TI) (Gal et al., 2023a). This method applies an approach similar to PALAVRA and optimizes a token embedding in the vocabulary space of a frozen TIG model (by feeding it with template sentences containing the token and asking it to generate images like the 5 examples). The authors show that a single token suffices to generate a concept faithfully and highlight the advantage of the approach to retain the knowledge of the pre-trained model.

We adopted the officially released code⁷, using the 1-token *Coarse* descriptions for the token initialization. We either adopted the Latent Diffusion Model (LDM) (Rombach et al., 2022)⁸ by following authors' instructions, or, by following more recent work (Tewel et al., 2023), the same SDM used for DB, with the parameters that the authors report for LDM (learning rate of $5e-3$) and by applying the same procedure as for DB to select training steps.

Number of steps in the concept optimization. Since we observed performance variations depending on the number of steps, we devised an automatic procedure to choose the optimal number and applied it to both DB and TI. We created a minimal validation set

⁷https://github.com/rinongal/textual_inversion

⁸<https://github.com/CompVis/latent-diffusion>

composed of 4 prompts ('a photo of * on a beach', 'a photo of * on the moon', 'a photo of * with a cat', 'a photo of a yellow *'). We avoided using the larger proposed *validation* set since, while we defined it for completeness and possible future uses, we preferred preserving the 5-shot nature of the considered personalization setting.

The score to maximize was chosen empirically among different versions, by comparing the best checkpoints selected automatically, with checkpoints selected manually (by qualitatively comparing the generated images at each validation iteration), for each of the 20 concepts. The best score S resulted being $S = \min\text{-}\max(\text{FID}_{\text{ctx}})$, subject to: $\min\text{-}\max(\text{FID}_{\text{ctx}}) < \min\text{-}\max(\text{FID}_{\text{cpt}})$. Namely, we first scale each metric such that it is in $[0, 1]$ along the optimization (min-max). Then, we maximize the capability to generate an image of the context (we maximize FID_{ctx}), subject to the fact that the generated images are also close to the concept image examples (subject to $\text{FID}_{\text{ctx}} < \text{FID}_{\text{cpt}}$). This provided slightly better results than the harmonic mean of the FID_{cpt} and a normalized FID_{ctx} , adopted in (Tewel et al., 2023).

While the ground-truth step annotation was possible for DB, we could not easily identify a learning trend in TI. Thus, we validated the procedure on the former method and then applied it also to the latter.

B.3.3 Benchmark Results: FID Numbers

In B.4 we report the numbers relative to the scatter plot in 8.8.

B.3.4 TIG Failure Cases

In B.8 we report failure cases for the TIG benchmark for the DB method. We first observe that sometimes the method represents the elements mentioned in the prompt, but not in the correct relationship (Top Row picture 1, Bottom Row picture 4). In some cases, the elements in the prompt are “semantically merged” (Top Row picture 4, Bottom Row picture 6). In other cases, some element of the prompt is totally missing (Top Row pictures 2, 3, 6; Bottom Row picture 2) or represented inaccurately (Top Row picture 7, Bottom Row pictures 1, 3, 5, 7).

CONCEPT	CATEGORIES	DESCRIPTIONS		
		COARSE (1-TOKEN)	RICH	DISCRIMINATIVE
BIRDYER	puppet (bird) + tool (sprayer)	puppet	red plastic sprayer bird puppet with colored wings	bird sprayer puppet
MYBIRD	puppet (bird)	plush	red stuffed bird with colored wings	stuffed bird
MYSPRAYER	tool (sprayer)	plastic container (bottle)	orange transparent plastic sprayer with white cap	sprayer
RATMOUSDEV	puppet (mouse) + tool (mouse)	puppet	computer mouse rat puppet with black round ears and cable	rat computer mouse puppet
MYRAT	puppet (mouse)	plush	grey stuffed rat with long pink tail	stuffed rat
MYMOUSEDEV	tool (mouse)	computer mouse (device)	black and beige computer mouse with wheel and cable	computer mouse
BOTTALIEN	puppet (alien) + tool (bottle)	puppet	blue alien plastic bottle puppet with aluminum foil dress	alien plastic bottle puppet
MYBOTTLE	tool (bottle)	plastic container (bottle)	blue plastic bottle with label	plastic bottle
ALIENSER	puppet (alien) + tool (dispenser)	puppet	green alien soap dispenser puppet with pink dress	alien soap dispenser puppet
MYDISPENSER	tool (dispenser)	plastic container (bottle)	white plastic soap dispenser with blue cap	soap dispenser
SNOX	puppet (snake) + tool (box)	puppet	green box headed snake puppet with round eyes and pink tongue	snake box headed puppet
MYBOX	tool (box)	box	cardboard box with lid	box
OCTOBAG	puppet (octopus) + accessory (bag)	puppet	beige woolen octopus bag with moustache and scarf	octopus bag puppet
MYREDBAG	accessory (bag)	bag	red square handbag with pois and bracelets	red bag
MYBLUEBAG	accessory (bag)	bag	blue square silk handbag with gold decorations	blue bag
MYDOLL	puppet (doll)	doll	doll with floral dress and brown hair	doll
MYCLOWN	puppet (clown)	puppet	rag clown with yellow hair and jumpsuit	clown
GOLDENSWEAT	clothes (sweater)	sweater	cream wool sweater with brown triangle and gold bow	sweater
HEARTYSHIRT	clothes (shirt)	shirt	white sleeveless shirt with red heart and blue flower	baby shirt with flower
MYSHIRT	clothes (shirt)	shirt	plain white baby shirt	white baby shirt

Table B.1 **Concept descriptions.** Chimeric concepts (**Bold**) are grouped with their hard negative concepts (Dotted lines).

SPLIT	REFERENCE IMAGES	IMAGE POOL
shirt	2038 - 97	6346 - 164
dress	2017 - 156	3817 - 164
toptee	1961 - 94	5373 - 112

Table B.2 Total number of URLs minus (-) the number of broken ones for each split of FashionIQ.

Method	mAP[%]	mAP@5[%]	mAP@10[%]	mRR[%]	R@1[%]	R@5[%]	R@10[%]
<i>Coarse</i>	16.83	11.74	13.13	24.21	14.48	33.49	43.63
k=0 <i>Discriminative</i>	*30.16*	*23.91*	*26.47*	*43.16*	*31.92*	*55.63*	*66.14*
<i>Rich</i>	*27.65*	*21.62*	*23.57*	*40.58*	*29.98*	*51.75*	*62.55*
PALAVRA	22.56 ± 1.29	17.58 ± 1.08	18.97 ± 1.18	34.39 ± 1.68	24.59 ± 1.94	44.30 ± 1.51	54.85 ± 1.49
k=1 Pic2Word	25.23 ± 1.20	19.52 ± 1.31	21.30 ± 1.18	37.16 ± 1.76	26.35 ± 1.85	48.06 ± 1.91	58.80 ± 1.67
SEARLE	28.16 ± 0.55	23.02 ± 0.65	24.60 ± 0.57	41.07 ± 0.92	31.16 ± 0.94	51.72 ± 0.71	60.85 ± 1.06
PALAVRA	23.59	18.65	20.05	35.99	26.75	45.11	55.08
k=5 Pic2Word	26.39	20.67	22.45	38.62	27.68	50.28	60.61
SEARLE	30.74	25.51	27.11	43.83	33.49	55.54	63.84

Table B.3 **Personalized TIR benchmark**. Additional metrics for the results reported in 8.2.

Method		FID _{cpt}	FID _{ctx}
upper bound	concept-only	*92.97*	19.77
	context-only	57.87	*27.92*
	GT images	75.83	25.37
k=0	LDM	70.37	22.35
	SDM	73.35	23.06
k=5	TI (LDM)	70.19	25.16
	TI (SDM)	77.25	20.92
	DB (SDM)	80.42	23.99

Table B.4 **Personalized TIG benchmark**: Fidelity-concept/context metrics on ConCon-Chi.