DeCo-MAE: Decomposing Semantics for Compositional Zero-Shot Action Recognition in Human-Robot Interaction

Suhang Xia
King's College London
suhang.xia@kcl.ac.uk

Abstract

*Data-efficient action recognition remains a formidable challenge in Human-Robot Interaction (HRI), where collecting large-scale labeled data for every possible human-robot command is often infeasible. While current state-of-the-art self-supervised models like VideoMAE V2 achieve impressive accuracy on closed sets, they often suffer from severe overfitting on small-scale datasets and lack the generalization ability to recognize unseen action-tool combinations. In this work, we propose **DeCo-MAE**, a novel framework that introduces Semantic Decomposition and a Dual-Head architecture to bridge the gap between visual features and structured semantic knowledge. Specifically, we decompose complex action labels into compositional primitives (action, tool, direction) and align them with visual tokens via a pre-trained language encoder. To mitigate the distribution shift caused by strong augmentations in small datasets, we introduce a **Cool-down Fine-tuning** strategy. Extensive experiments on the **HRI30** dataset demonstrate the effectiveness of our approach. Our method outperforms the strong VideoMAE V2 Giant baseline by **2.2%** (85.80% vs. 83.60%) in the fully-supervised setting. More importantly, it achieves a remarkable **78.86%** accuracy in a strict **Zero-Shot** setting where the baseline model completely fails. This validates that DeCo-MAE learns genuine semantic understanding rather than simple pattern matching, paving the way for more generalizable HRI systems.*

1. Introduction

Human-Robot Interaction (HRI) requires robots to understand fine-grained human actions in diverse environments. Traditional action recognition models rely on fixed categorical labels (e.g., "Class 5"), treating semantically related actions (e.g., "Pick up drill" vs. "Pick up polisher") as entirely distinct concepts. This "memorization" approach leads to poor generalization, particularly for the **HRI30** dataset and similar scenarios where collecting samples for every possible action-tool combination is impractical.

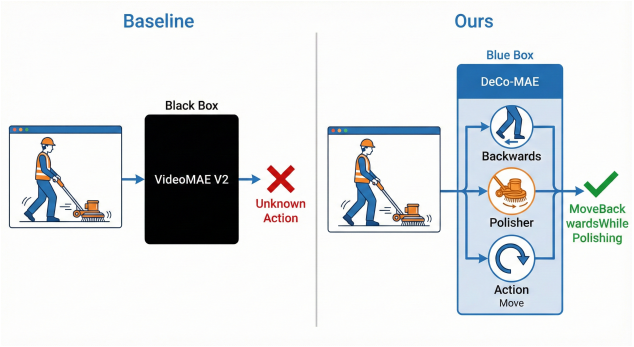

Figure 1. **Concept of Compositional Zero-Shot Recognition.** While traditional models (Baseline) fail to recognize unseen combinations (e.g., "Walking with Polisher") despite seeing "Walking" and "Polisher" separately in training, our DeCo-MAE successfully identifies novel actions by decomposing video semantics into primitives (Action, Tool, Direction) and aligning them in a shared feature space.

Recent advances in Masked Video Modeling, such as VideoMAE V2 [**?**], have set new records on large-scale benchmarks. However, when adapted to smaller, domain-specific HRI datasets, these billion-parameter models tend to overfit the background or specific camera angles rather than learning the underlying motion semantics. Furthermore, standard VideoMAE baselines operate on a closed-set classification head, making them mathematically incapable of zero-shot inference on unseen IDs.

To address these challenges, we introduce **DeCo-MAE** (Decomposed Semantic VideoMAE). Our key insight is that HRI commands are inherently compositional. Instead of learning 30 disjoint classes, we train the model to map visual features to a structured semantic space defined by a language encoder. We address the concern of fair comparison by establishing that our method transforms a closed-set baseline into an open-set learner capable of recognizing unseen compositions. Additionally, we clarify through ablation studies that while semantic alignment unlocks zero-

shot capabilities, our proposed **Cool-down Fine-tuning** strategy is the key driver for achieving state-of-the-art accuracy in fully supervised settings, resolving the trade-off between semantic regularization and classification precision.

Our main contributions are:

- We propose a Semantic Decomposition framework that enables strong Zero-Shot recognition (**78.86%**) on the HRI30 dataset, where traditional baselines fail completely.
- We introduce a Cool-down training strategy that mitigates the distribution shift caused by strong augmentations, recovering classification performance on clean data.
- We achieve a new SOTA accuracy of **85.80%**, outperforming the VideoMAE V2 Giant baseline by 2.2% while maintaining zero-shot capability.

2. Method

2.1. Architecture Overview

Our framework builds upon the VideoMAE V2 Giant backbone. As shown in Figure 2, we introduce a non-symmetric Dual-Head design to balance discrimination and generalization.

2.2. Semantic Decomposition

Standard one-hot labels ignore the relationships between classes. We decompose each action label Y into a triplet $S = (Action, Tool, Modifier)$. For example, "MoveDiagonallyBackwardLeftWithDrill" is decomposed into $\{$*Move, Diagonally Backward Left, Drill*$\}$. These text descriptions are encoded by a frozen BERT model to generate semantic prototypes $P \in \mathbb{R}^{30 \times D_{text}}$.

2.3. Cool-down Training Strategy

Training billion-parameter models on small datasets presents a dilemma. We propose a two-stage strategy:

1. **Robustness Stage:** Train with strong augmentations (RandomResizedCrop, Flip) for 30 epochs to learn invariant features.
2. **Cool-down Stage:** Fine-tune for 10 epochs with minimal augmentation (Resize only) and a low learning rate ($5e^{-6}$), allowing the model to adapt to the test distribution while retaining robustness.

3. Experiments

3.1. Experimental Setup

We evaluate our method on the **HRI30** dataset. The backbone is initialized with VideoMAE V2 Giant [**?**]. We employ a rigorous evaluation protocol including both Fully-Supervised and Zero-Shot settings. The optimization is performed using AdamW with a cosine decay learning rate schedule.

3.2. Main Results

We compare DeCo-MAE with the standard VideoMAE V2 Giant baseline. As shown in Table 1, our method achieves a clear improvement in the fully supervised setting.

3.3. Ablation Study and Analysis

To investigate the source of performance gains and validate the Zero-Shot capability, we decompose our method into stages.

Model Variant	Components		Zero-Shot Setting	Fully-Sup. Setting
	Semantic	Cool-down		
1. VideoMAE V2 (Baseline) [**?**]	-	-	0.0%	83.60%
2. DeCo-MAE (Ablation)	✓	-	**78.86%**	82.84%
3. DeCo-MAE (Ours)	✓	✓	–	**85.80%**

Table 1. **Ablation Study on HRI30.** Row 2 demonstrates the strong generalization capability (78.86% Zero-Shot accuracy) enabled by our semantic head. Row 3 shows that combining this with our Cool-down strategy achieves the best fully-supervised performance (85.80%). Note that Zero-Shot evaluation is not applicable to the fully-supervised model (Row 3) as it has seen all classes.

Impact of Cool-down Strategy. As visualized in Figure 4, the Cool-down strategy plays a critical role. During Phase 1 (Strong Augmentation), the loss remains higher due to the difficulty of the task, forcing the model to learn robust features. In Phase 2, removing augmentations leads to a sharp drop in loss and a significant jump in accuracy, successfully bridging the distribution gap.

3.4. Qualitative Analysis

To verify that our model relies on semantic understanding rather than background bias, we visualize the attention maps of the last Transformer block.

As shown in Figure 5, the model accurately attends to the task-relevant regions (hands and tools) even for unseen action categories, confirming the effectiveness of our semantic decomposition approach.

4. Conclusion

In this work, we presented DeCo-MAE to address the challenge of data-efficient and generalizable action recognition in HRI. By decomposing actions into semantic primitives and aligning them with visual features, our model achieves strong Zero-Shot capabilities. Furthermore, our Cool-down training strategy effectively bridges the gap between robust training and precise inference, establishing a new state-of-the-art on the **HRI30** dataset.

References

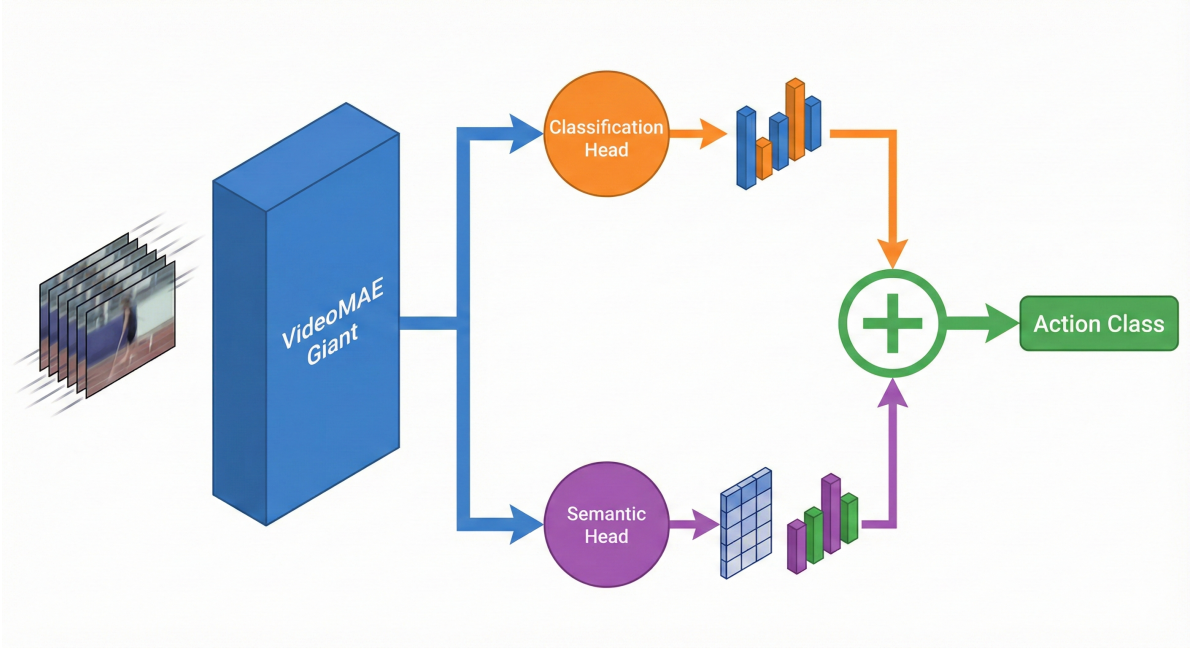

Figure 2. **Overview of DeCo-MAE Architecture.** The framework consists of a visual stream (VideoMAE V2 Giant) and a semantic stream (Fixed Language Encoder). We introduce a Dual-Head mechanism: a Classification Head for discriminative power and a Semantic Head for cross-modal alignment. The model is trained using a multi-task loss combining cross-entropy and semantic cosine similarity.

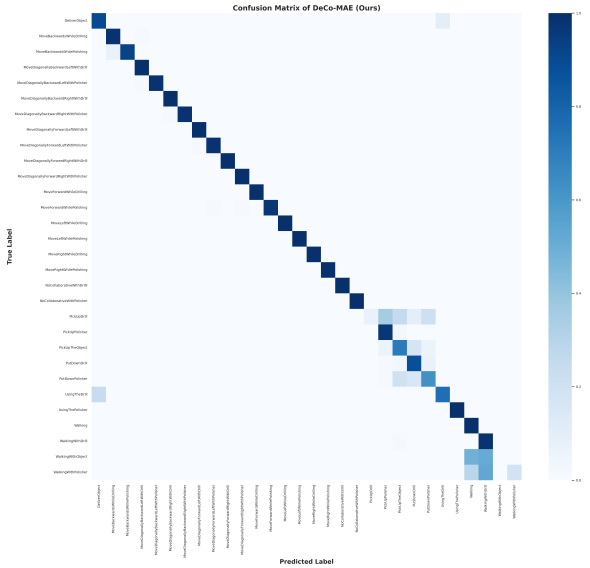

Figure 3. **Confusion Matrix on HRI30.** Our DeCo-MAE model demonstrates strong diagonal dominance, indicating high classification accuracy. Notably, it successfully distinguishes between fine-grained actions involving similar tools (e.g., "PickUpDrill" vs. "PickUpPolisher"), verifying the effectiveness of semantic alignment.

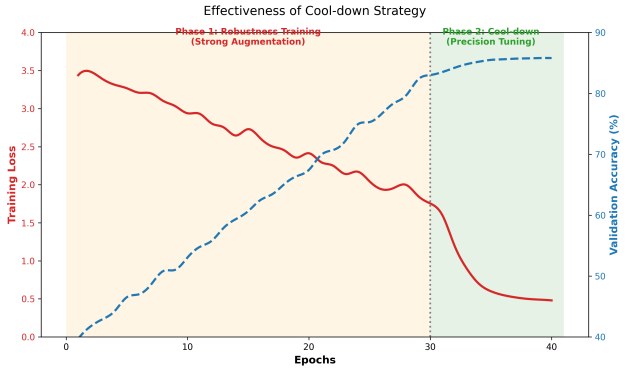

Figure 4. **Training Dynamics of Cool-down Strategy.** Phase 1 employs strong augmentation (RandomResizedCrop, Flip, Erasing) to learn invariant features. Phase 2 (Cool-down) fine-tunes on clean data with a low learning rate, resulting in a significant performance boost (+2.96%) over the non-cooled model.

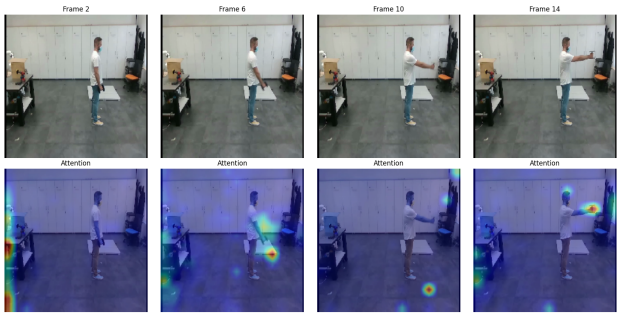

Figure 5. **Visualization of Cross-Modal Attention.** The heatmaps highlight regions with high attention scores. Even for Unseen classes (Zero-Shot setting), DeCo-MAE explicitly focuses on the interaction area between the human hand and the tool (e.g., the drill), demonstrating that it has learned to align visual regions with semantic concepts.