DeCo-MAE: Decomposing Semantics for Compositional Zero-Shot Action Recognition

Suhang Xia, Ruiyi Hu, Muye Yuan

Sensing and Perception Group: ENT

What is DeCo-MAE and why it

What:
Decomposing Semantics for Compositional Zero-Shot Action Recognition in Human-Robot Interaction Classes
Why:
Zero-Shot; explainability;

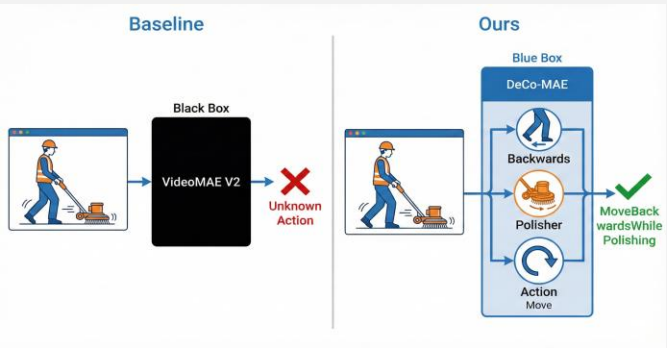
Core Context and Pain Points

Point 1:

Large models (such as VideoMAE V2) tend to overfit on small-scale HRI datasets, memorizing backgrounds or camera angles rather than understanding the actions themselves.

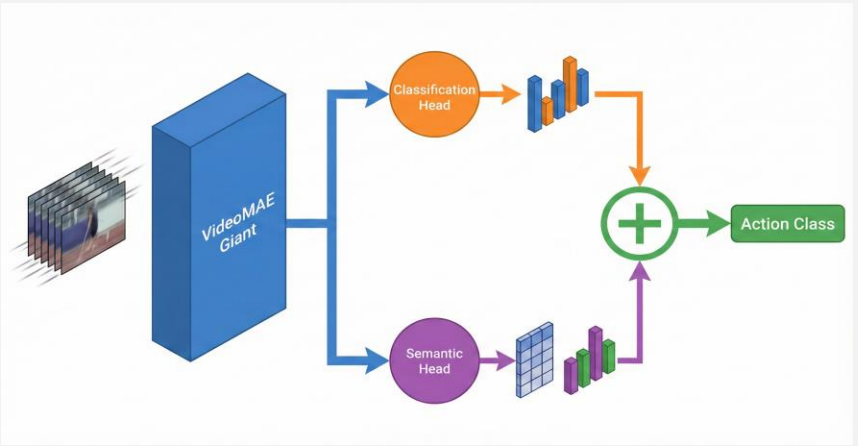
Point 2:

Lack of Generalization Ability: Traditional models treat actions as fixed categories (e.g., "Category 5"). If a model has been trained on 'walking' and "holding a polisher" but never encountered the combination "walking while holding a polisher," it cannot recognize it. This is known as "Zero-Shot" recognition failure.



Codes and pre-trained models are at
https://github.com/SuhangXia/DeCo-MAE

Our architecture: DeCo-MAE (Ours)



Semantic Decomposition

Addressing Point 1

We decompose each action label Y into a set = (Action, Tool, Modifier)
For example,
"MoveDiagonallyBackwardLeftWithDrill" is decomposed
into {Move, Diagonally Backward Left, Drill}.

Cool-down Training Strategy

Addressing Point 2

Training billion-parameter models on small datasets presents a dilemma. We propose a two-stage strategy:
1. Robustness Stage: Train with strong augmentations for 30 epochs to learn invariant features.
2. Cool-down Stage: Fine-tune for 10 epochs with minimal augmentation and a low learning rate (5e−6), allowing the model to adapt to the test distribution while retaining robustness.

Advantages and drawbacks:

Ads: Generalization capability; addresses overfitting and distribution shift issues in large models when trained on small datasets; explainability.
Drawbacks: High computational resource requirements; cumbersome training process

Experiment: comparison of different architectures

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-supervised performance (85.80%).

Model Variant	Semantic	Cool-down	Zero-Shot	Fully Setting
VideoMAE V2 (Baseline)	-	-	0.0%	83.60%
DeCo-MAE (Ablation)	✓	-	78.86%	82.84%
DeCo-MAE (Ours)	✓	✓	-	85.80%

Validation on unseen classes

Need to mention that the unseen classes are not from outside, they are some classes which masked manually. For Unseen classes, DeCo-MAE explicitly focuses on the interaction area between the human hand and the tool, demonstrating that it has learned to align visual regions