InvAct: View and Scene-invariant Atomic Action Learning from Videos

Bahri Batuhan Bilecen [1] **Korrawe Karunratanakul** [1] **Vasileios Choutas** [2] **Thabo Beeler** [2] **Bernt Schiele** [3]
Jan Eric Lenssen [3] **Siyu Tang** [1]

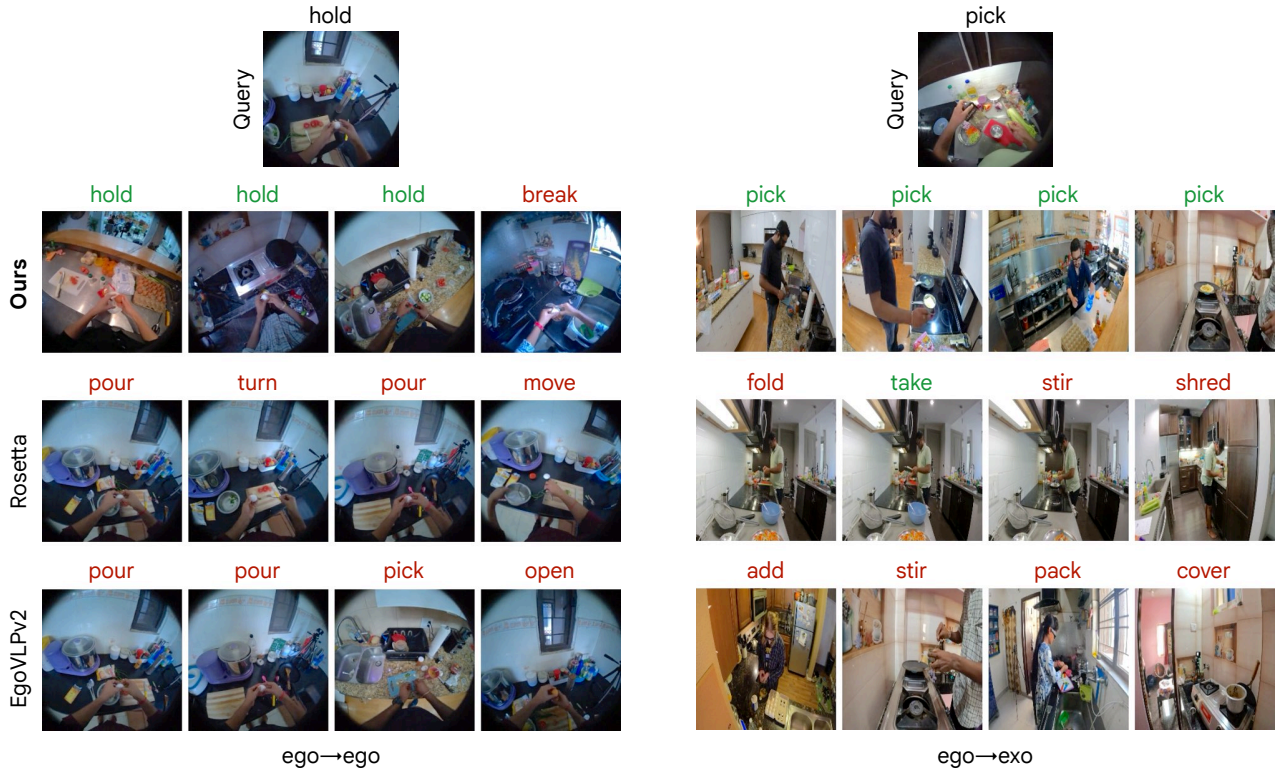

Figure 1. We show the top 4 nearest neighbors retrieved based on the learned latent action space of different methods. Previous methods rank mostly based on static scene context and the nearest retrievals consist of incorrect, different actions in the same environment. Meanwhile, our method successfully ranks based on action similarity, regardless of the viewpoint (ego/exo) or scene difference.

Abstract

Action learning should capture interaction dynamics that generalize between viewpoint and scene changes. Although recent work pursues view-invariant representations, these methods often overfit to scene cues, weakening their ability to model fine interactions. This issue is especially acute for atomic actions, which are short, interaction-centric primitives. We address this with an atomic action embedding model trained to be invariant to both ego–exo viewpoint shifts and scene changes. We learn a latent space such that clips of the same atomic action are pulled together across scenes and views, while different actions from the same scene are pushed apart. We further use language to ground the embeddings in semantics. Experiments show that the proposed representation significantly improves retrieval across cross-view and cross-scene settings, shows strong transfer to unseen datasets, enables longer keystep actions obtained by zero-shot combination of our atomic embeddings, and shows promising results on a preliminary robotics manipulation task. We believe that the proposed approach will benefit robotic and human-understanding downstream tasks.

[1]ETH Zurich, Switzerland [2]Google, Switzerland [3]Max Planck Institute, Germany. Correspondence to: Bahri Batuhan Bilecen <bbilecen@ethz.ch>.

Preprint. February 7, 2026.

1. Introduction

Action representation learning aims to learn action representations that support downstream tasks such as action recognition (Carreira & Zisserman, 2017; Bagad & Zisserman, 2025a), retrieval (Luo et al., 2025), skill transfer (Xu et al., 2023), imitation learning (Kareer et al., 2025; Xiong et al., 2025), pretraining for vision-language action models (VLA) (Bu et al., 2026), and extended reality applications. An ideal action learning paradigm must generalize across changes in viewpoint, scene, and object appearance. In practice, actions can be represented and learned in different ways: as supervised latent embeddings learned from labeled actions (Luo et al., 2025), as unsupervised discovered latents learned without manual labels (Ye et al., 2025), or as tokens in the text domain such as verbs or short phrases aligned with video (Chen et al., 2025).

Actions also vary in granularity. **Keystep actions** *(e.g. checking for damages, cooking omelet, repairing a bike, unboxing a package)* describe longer, higher-level procedures that often correlate strongly with objects, the context of the scene, and appearance (Grauman et al., 2024). These higher-level activities can be decomposed into **atomic actions** *(e.g. pushing, pulling, cutting, placing a box on a table)*, which are short interaction-centric primitives, characterized by fine-grained temporal dynamics and contact patterns. In this work, we present a method to learn a coherent, view and scene-invariant atomic action space from video sequences.

To advance keystep action learning, recent work (Xue & Grauman, 2023; Grauman et al., 2024; Park et al., 2025) has focused on **view-invariance**, which is learning action representations that align across egocentric and exocentric viewpoints, often using large-scale multi-view datasets such as Ego-Exo4D (Grauman et al., 2024). However, when obtaining view-invariant representations becomes the main objective, models can rely on static scene content to obtain representations, performing keypoint matching instead of learning action dynamics. This is worsened by evaluations in narrow single-domain subsets, most commonly `cooking`, where limited scene diversity allows appearance to dominate. Consequently, strong quantitative results may not reflect a real understanding of actions, i.e., models can fail to recover the same action in differently looking scenes.

Figure 1 illustrates this issue by visualizing retrieval behavior in the Ego-Exo4D validation set, with all scenes included in the retrieval pool. It shows that recent baselines are overadapted to appearance cues. Specifically, prior retrieval methods often exploit a shortcut by ranking top-k clips according to shared scene context, such as layout and objects, rather than interaction dynamics. This specific failure mode is most evident in the ego→exo examples retrieved by Viewpoint Rosetta Stone (Luo et al., 2025) in the third row of

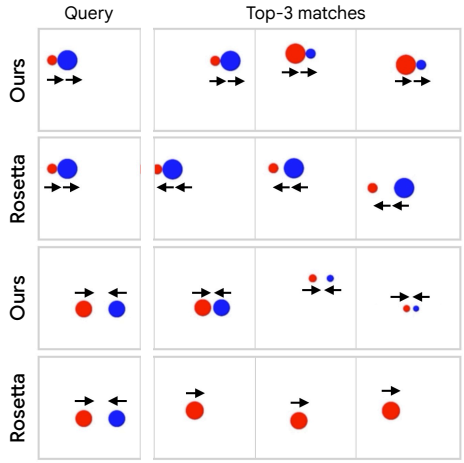

Figure 2. Zero-shot video-video retrieval results of the methods on PhyWorld (Kang et al., 2025) dataset. Arrows denote the movement direction in the video, which are the actions. Our method faithfully represents fine-grained actions as evidenced by the successful retrieval, whereas other methods may opt for color and shape matching without paying attention to movement directions (row 2), or retrieve partially-correct actions (row 4).

Figure 1. The method may return a hit with a superficially similar action, such as *taking*, but it does so by retrieving clips from the same kitchen and matching static cues such as the blue bowl and brown cabinets, rather than recognizing the underlying interaction. As a result, it can miss stronger dynamic matches that occur in different scenes. These observations are further strengthened by Figure 2, which shows similar trends in a simpler dataset in a zero-shot setting. In this case, the baseline also behaves like a near-static visual matcher, prioritizing color and shape similarity over the temporal cues required to distinguish fine-grained actions.

To overcome these issues, we present **InvAct**, the first **atomic-action embedding model** that maps short video clips into a compact action space while being **invariant to both scene changes and ego–exo viewpoint shifts**, yet remaining discriminative for fine-grained interaction dynamics. Unlike prior work that primarily studies viewpoint alignment for keystep actions, our model explicitly targets *both* viewpoint and scene invariance at the level of atomic actions. We further demonstrate that these atomic-action embeddings can be composed to support keystep actions across all ego–exo combinations, enabling substantial transfer across scenes and viewpoints.

Concretely, we train a transformer-based encoder that fuses pretrained visual features with optical flow maps to produce a single-token action embedding. During training, instead of organizing clips directly by visual similarity, we shape the embedding space around atomic actions. Clips depicting the same action are mapped close together even when they come from different scenes or viewpoints, whereas clips from the same scene performing different actions are

pushed apart to counteract scene-specific bias. We enforce this structure with contrastive objectives that emphasize interaction dynamics over appearance. Finally, we leverage sentence and verb-level action labels to anchor the embeddings semantically, facilitating meaningful retrieval.

The effectiveness of our method is shown in Figures 1 and 2, where it retrieves stable, interaction-faithful matches in both complex and diverse scenes and deliberately simplified retrieval settings. Our contributions are as follows:

- To our knowledge, **InvAct** is the first atomic-action learning model to tackle *both* the large viewpoint-invariance and scene-invariance in a single model, demonstrated using the Ego-Exo4D dataset.
- Without additional training, we show that keystep-level representations can be formed by chaining our learned atomic embeddings, allowing the retrieval of keystep action out-of-the-box.
- We outperform existing methods, including egocentric-only models and prior ego-exo alignment approaches, on video-video retrieval for both atomic actions and keystep actions, across averaged cross-view and cross-scene settings, and we show strong generalization to out-of-domain datasets in zero-shot evaluation. We further demonstrate our method's effectiveness by long-sequence probing, latent clustering, visual attention, linguistic alignment analyzes, and a sample downstream task on VLA pretraining for robotics.

We believe this unified approach can benefit video pretraining for both robotics and human action understanding, and help bridge robot-oriented and human-oriented action learning. We will release our code and trained models.

2. Related Work

Generalizable representations. Large pretrained backbones such as CLIP (Radford et al., 2021), DINOv3 (Siméoni et al., 2025), and TimeSFormer-based architectures (Bertasius et al., 2021) provide strong generic representations of large-scale image–text or self-supervised training. A recent video-oriented representation learning work, V-JEPA2 (Assran et al., 2025), further improves temporal modeling, while FlowFeat (Araslanov et al., 2025) emphasizes motion-driven features. LiFT (Bagad & Zisserman, 2025a) converts image features into time-sensitive video descriptors by linearizing feature trajectories. However, these models are not designed to be scene and view-invariant.

Multimodality for action learning. Video–language pretraining (VLP) methods such as LaViLa (Zhao et al., 2023) and X-CLIP (Ma et al., 2022) focus on improving video–text alignment using language supervision and multi-grained contrastive matching between frame-text. HierVL (Ashutosh et al., 2023) complements this by learning hierarchical video–text embeddings that align both clip-level narrations and long-video summaries. Egocentric VLP methods like EgoVLPv2 (Lin et al., 2022; Pramanick et al., 2023) adapt video–text contrastive learning to first-person data, by mining egocentric-aware positives/negatives and using stronger cross-modal fusion. VL-JEPA (Chen et al., 2025) explores predicting text embeddings from videos rather than generating visual tokens. TARA (Bagad & Zisserman, 2025b) adapts multimodal LLMs to time-sensitive video-text embeddings, focusing on temporal order.

These directions complement ours by improving semantic grounding through video-text alignment, similar to how we use atomic labels. However, they do not explicitly enforce scene and viewpoint invariance at the atomic action level, and they often inject text into the model, whereas we use language only through the loss.

View-invariant representation learning. On the ego–exo side, the Ego-Exo4D benchmark formalizes synchronized first- and third-person understanding on scale (Grauman et al., 2024), and works like Viewpoint Rosetta Stone (Luo et al., 2025) and SUM-L (Wang et al., 2023) pursue view-invariant alignment in unpaired or weakly paired settings. AE2 (Xue & Grauman, 2023) learns fine-grained invariance by temporally aligning ego-exo videos without synchronization, BYOV (Park et al., 2025) uses masked ego–exo modeling to encourage cross-view consistency. Seeing Without Pixels (Xue et al., 2025) reveals that actions can be observed through non-appearance signals like camera motion, aligning trajectory cues with language.

In contrast, we tackle scene invariance while still satisfying viewpoint invariance. We also evaluate our method using all subsets of the Ego-Exo4D validation split, going beyond the more constrained subsets or settings used in prior work (Wang et al., 2023; Xue & Grauman, 2023; Luo et al., 2025; Park et al., 2025).

Latent atomic action discovery in robotics. LAPA (Ye et al., 2025) and LAOF (Bu et al., 2025) study latent action discovery in embodied settings, learning action abstractions from video with minimal or no action labels. LAOF also uses optical-flow constraints to stabilize learning. VILA (Jeong et al., 2026) focuses on supervised view-invariant latent actions for policy learning. Ko et al. 2024 show another route by extracting action-relevant structure from videos via flow and depth for policy learning.

These works share our intuition that motion aids in action learning when appearance varies. However, we learn atomic actions that stay consistent across viewpoint shifts with complex environments and grounded by language, a setting that is not commonly studied in such robotics-oriented work.

3. Method

Our aim is to learn a coherent, view and scene-invariant atomic action space from video sequences. The resulting embeddings provide a strong backbone for downstream tasks, as evidenced by retrieval and recognition across viewpoints and environments, while penalizing scene-based shortcuts.

We show that, with appropriate supervision, a single transformer can learn such embeddings and generalize well in zero-shot settings. Moreover, these atomic embeddings can be combined into keystep-action embeddings in zero-shot manner. Our framework uses a combination of contrastive and classification losses between the video and text domains.

3.1. Architecture

Given a video clip $v \in \mathbb{R}^{T\times H\times W\times 3}$, we first compute the optical flow as $f \in \mathbb{R}^{(T-1)\times H\times W\times 3}$. A 3D-CNN tubelet embedder (Arnab et al., 2021) $\phi(\cdot)$ then maps f to spatiotemporal tokens $\phi(f) \in \mathbb{R}^{N\times d}$, where $N = (T-1)\frac{H}{8}\frac{W}{8}$. Then, we prepend a learnable classification token $z_{\text{CLS}} \in \mathbb{R}^d$, and process the sequence with a transformer $\mathcal{T}$ (Vaswani et al., 2017) using rotary positional embeddings (Su et al., 2024). To inject additional semantic cues, we extract frozen features $\mathcal{D}(v)$ from RGB video and fuse them into $\mathcal{T}$ through cross-attention. The final action representation is the output CLS token z_{CLS}, on which all losses are applied:

$$z_{\text{CLS}} = \mathcal{T}\big(\phi(f), z_{\text{CLS}}, \mathcal{D}(v)\big) \tag{1}$$

The architecture and inference scheme are visualized in Figure 3. We compute flow with RAFT (Teed & Deng, 2020) use frame-averaged DINOv3 (Siméoni et al., 2025) for $\mathcal{D}(v)$, and set $T = 8$. We note that previous work (Luo et al., 2025; Pramanick et al., 2023) tends to input the text modality into the inference pipeline, whereas we only utilize language through losses. This means that our model does not require text to extract actions during inference.

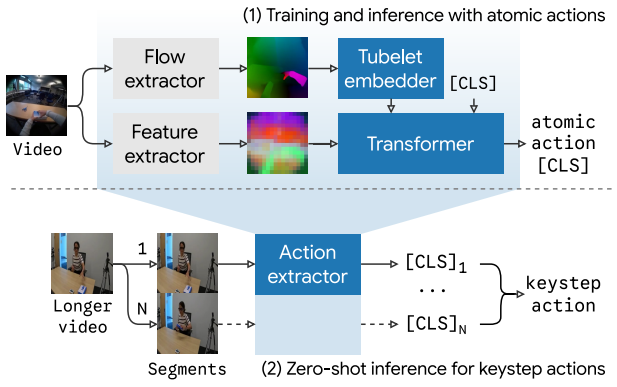

Figure 3. Architecture of our action extractor method. We first only train the tubelet embedder and transformer with atomic-action labels but can later infer atomic-actions and keystep-actions.

3.2. Definitions

Notation. Let $\mathcal{S}$ and $\mathcal{A}$ denote the sets of scene and action labels, respectively. A paired ego-exo clip is associated with a label in the product space $\mathcal{S}\times\mathcal{A}$. We form our minibatches via sampling from this product space. We index paired samples by i and denote their ego and exo embeddings by $g_i \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^d$, respectively, which are the z_{CLS} token outputs of the transformer.

Label types. We train using two kinds of labels: **atomic sentence-level actions** (*e.g., cutting a tomato on a bench*) and **atomic verb-level actions** (*e.g., cut*). For inference, in addition to atomic labels, we also include **keystep sentence-level actions** (*e.g., check for any damage or splits in the tube*). Details are provided in the Appendix.

3.3. Losses

Ego-exo view alignment. For each paired sample i, we align the ego and exo embeddings via the cosine distance:

$$\mathcal{L}_{\text{vv}} = \frac{1}{N}\sum_{i=1}^{N}\Big(1 - \langle g_i, x_i\rangle\Big) \tag{2}$$

Cross-scene action attraction. For each anchor sample i, we form a set of positives $\mathcal{P}(i)$ that share the same action label but come from different scenes, and a set of negatives $\mathcal{N}(i)$ that correspond to different actions. We then apply a multi-positive InfoNCE (Oord et al., 2018) objective in ego:

$$\mathcal{L}_{\text{att}}^{\text{ego}} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{\sum_{p\in\mathcal{P}(i)}\exp\langle g_i, g_p\rangle}{\sum_{k\in\mathcal{P}(i)\cup\mathcal{N}(i)}\exp\langle g_i, g_k\rangle} \tag{3}$$

and define $\mathcal{L}_{\text{att}}^{\text{exo}}$ analogously by replacing g with x. The final attraction loss averages modalities, $\mathcal{L}_{\text{att}} = \frac{1}{2}\big(\mathcal{L}_{\text{att}}^{\text{ego}} + \mathcal{L}_{\text{att}}^{\text{exo}}\big)$.

Same-scene repulsion. For each anchor sample i, we select a negative set $\mathcal{R}(i)$ consisting of clips from the same scene but with a different action label. We penalize their similarity using a 2-way InfoNCE. In ego space:

$$\mathcal{L}_{\text{rep}}^{\text{ego}} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{\exp\langle g_i, g_i\rangle}{\exp\langle g_i, g_i\rangle + \sum_{r\in\mathcal{R}(i)}\exp\langle g_i, g_r\rangle} \tag{4}$$

and $\mathcal{L}_{\text{rep}}^{\text{exo}}$ is defined analogously by replacing g with x. Then both modalities are averaged $\mathcal{L}_{\text{rep}} = \frac{1}{2}\big(\mathcal{L}_{\text{rep}}^{\text{ego}} + \mathcal{L}_{\text{rep}}^{\text{exo}}\big)$.

Cross-scene action attraction and same-scene repulsion together help discouraging learning shortcuts, especially regarding appearance similarity due to scene sharing, and encourage learning a rich atomic action space. In particular, repulsion from the same-scene was not investigated in previous work such as (Luo et al., 2025).

Atomic sentence-level text-video alignment. Let $t_i^{\text{atom}} \in \mathbb{R}^k$ be the atomic sentence-level pooled CLIP (Radford et al., 2021) text embedding paired with sample i, and let $\{g_i, x_i\}$ denote the corresponding video embeddings for ego and exo, respectively. We use a symmetric CLIP-style InfoNCE loss with index-matched positives:

$$\mathcal{L}_{\text{text}}^{\text{ego}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp\langle g_i, t_i^{\text{atom}}\rangle}{\sum_{j=1}^{N}\exp\langle g_i, t_j^{\text{atom}}\rangle} + \log\frac{\exp\langle g_i, t_i^{\text{atom}}\rangle}{\sum_{j=1}^{N}\exp\langle g_j, t_i^{\text{atom}}\rangle}\right] \quad (5)$$

and define $\mathcal{L}_{\text{text}}^{\text{exo}}$ by replacing g with x. We then average two losses, $\mathcal{L}_{\text{text}} = \frac{1}{2}\left(\mathcal{L}_{\text{text}}^{\text{ego}} + \mathcal{L}_{\text{text}}^{\text{exo}}\right)$.

Atomic verb-level text-video alignment. For samples with verb supervision, we predict logits with a shared head $h(\cdot)$ and minimize the cross-entropy over verb classes. Let $y_i \in \{1, \ldots, K\}$ be the verb class index (one-hot $\mathbf{y}_i \in \{0,1\}^K$), and let $\ell_i^g = h(g_i) \in \mathbb{R}^K$, $\ell_i^x = h(x_i) \in \mathbb{R}^K$. Then

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp \ell_{i,y_i}^g}{\sum_{c=1}^{K}\exp \ell_{i,c}^g} + \log\frac{\exp \ell_{i,y_i}^x}{\sum_{c=1}^{K}\exp \ell_{i,c}^x}\right] \quad (6)$$

The final loss is a weighted sum of all terms. More details are reserved to the Appendix.

$$\mathcal{L} = \lambda_{\text{vv}}\mathcal{L}_{\text{vv}} + \lambda_{\text{att}}\mathcal{L}_{\text{att}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} \quad (7)$$

4. Experiments

In this section, we describe the datasets and training setup, the baselines used, and a broad set of experiments including retrieval, clustering, probing, attention visualization, language analysis, and a robotic downstream task.

4.1. Datasets and Training

We utilize Ego-Exo4D (Grauman et al., 2024) for training and main evaluation. Specifically, we leverage the portion `downscaled_takes/448`, which has time-synchronized ego-exo clips with keystep and atomic labels. Keystep labels span longer temporal windows than atomic actions, where each keystep describes the overall activity in a segment, and its timeframe typically contains multiple atomic action labels. Unlike recent cross-view prior works (Luo et al., 2025; Wang et al., 2023), which often report results only on the `cooking` subset, we evaluate the complete Ego-Exo4D validation split combining the {`cooking, health, bike`} subsets, using their shared atomic-action labels. In total, the train set has 30,615, and the validation set has 8,961 videos.

For zero-shot evaluation, we employ Charades-Ego (Sigurdsson et al., 2018), Something-Something v2 (Goyal et al., 2017), PhyWorld (Kang et al., 2025), and NTU RGB+D (Liu et al., 2020; Shahroudy et al., 2016) datasets. Details for datasets and training are given in the Appendix.

4.2. Baselines

All baselines evaluated are listed in Table 1. We include VI Encoder, the official baseline of Ego-Exo4D, and its Ego4D (Grauman et al., 2022) TimeSFormer variant. Viewpoint Rosetta and SUM-L follow a different backbone, add ego–exo and language alignment components, together with pseudo-paired/unpaired data. We also report results for egocentric models (LaViLa, EgoVLP, EgoVLPv2). Finally, we include general-purpose pretrained features (CLIP, DINOv3, VJEPA-2, FlowFeat, LiFT) and robotics-oriented works like LAPA. We omit AE2 and BYOV because their evaluations are scenario-specific, i.e. a single checkpoint per task (e.g. *breaking eggs*, or *pouring milk*, etc.) rather than a unified retrieval setting. More details are in the Appendix.

4.3. Downstream Experiments

To understand whether **InvAct** captures action semantics rather than scene or object cues, we test under progressively harder downstream setups. These experiments assess fine-grained atomic action retrieval across views and scenes, generalization to unseen datasets, composition into higher-level keysteps without additional training, and using the actions in pretraining of VLAs for robotics manipulation.

Atomic action retrieval. Many cross-view methods focus on matching the same action across viewpoints within the same scene and time, where correspondence cues can be highly informative. For atomic actions, we instead emphasize action-level matching between viewpoints and scenes, even when objects, context, and timelines differ. We therefore report cross-view and cross-scene verb-level atomic-action hit-rates and visuals in Table 1 and Figure 4.

Table 1 shows that our method significantly outperforms cross-view alignment methods such as Viewpoint Rosetta, single-view methods such as EgoVLPv2, robotics-focused approaches such as LAPA, and general-purpose feature extractors such as V-JEPA 2 and FlowFeat.

Figure 4 also qualitatively shows that our embedding space supports reliable action retrieval in all ego-exo queries and gallery combinations. In these examples, cross-view methods often over-rely on appearance cues such as desks and papers and omit fine-grained actions, while single-view methods tend to fail when the same action is observed cross-view. General-purpose representations can capture coarse motion patterns but still miss the underlying action.

Zero-shot retrieval. We also evaluate our embeddings on unseen datasets such as SSV2, Charades-Ego, PhyWorld, and NTU RGB+D. Table 2 and Figures 2, 5 and 6 show that

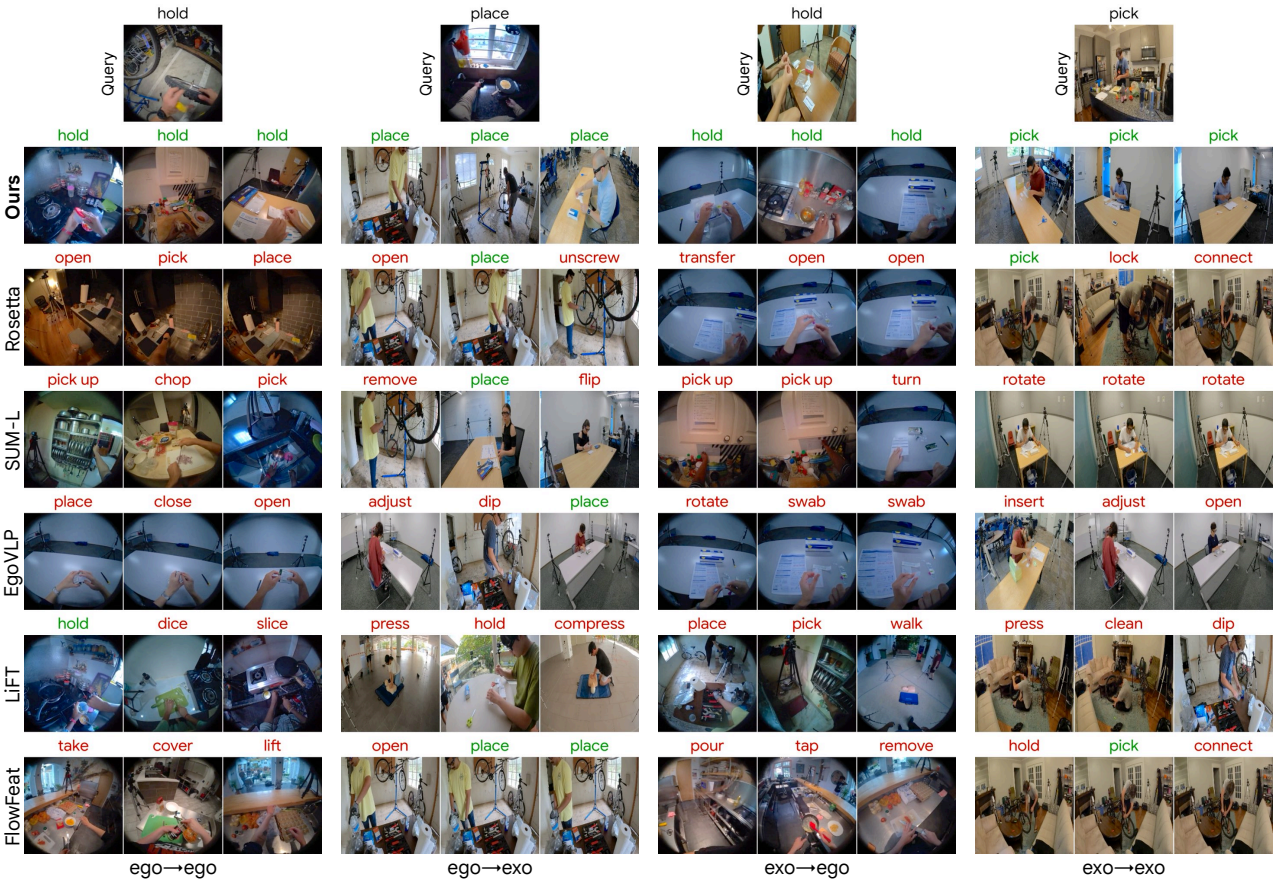

Figure 4. Comparison of different methods on Ego-Exo4D, on atomic-action cross-scene video-video retrieval task. First columns are query videos, and others are top-3 best matches. Correct and incorrect hits are visualized with green and red, respectively.

Table 1. Video-video retrieval hit-rates of methods in the Ego-Exo4D validation set, for both **atomic action** hit-rates and **keystep action** hit-rates. g and x denote ego and exo, respectively. For keystep actions on our method, we re-purpose the atomic action checkpoint without finetuning. The best three results are shown in **bold**, underlined, and in *italic*, respectively.

		Atomic action hit-rates (@10)						Keystep action hit-rates (@10)					
		Cross-scene (↑)			Cross-view (↑)			Cross-scene (↑)			Cross-view (↑)		
	Method	g→g	x→x	avg.	g→x	x→g	avg.	g→g	x→x	avg.	g→x	x→g	avg.
	Random	3.04	3.01	3.02	2.15	3.70	2.92	6.23	7.35	6.79	6.59	6.72	6.65
General embeds	CLIP	16.01	15.33	15.67	14.03	16.25	15.14	26.76	20.47	23.61	17.71	17.68	17.70
	DINOv3	19.08	15.11	17.09	13.60	16.32	14.96	28.50	20.40	24.45	24.43	19.42	21.93
	V-JEPA 2	17.83	14.05	15.94	15.32	18.40	16.86	33.75	21.55	27.65	14.40	17.45	15.93
	FlowFeat	23.56	24.04	23.80	19.70	23.54	21.62	21.25	15.51	18.38	11.84	12.89	12.37
	LiFT	27.84	23.15	25.50	21.71	27.08	24.39	25.88	19.02	22.45	20.79	16.23	18.51
Single view	LAPA	25.24	26.52	25.88	15.96	13.82	14.89	13.58	15.28	14.43	15.09	13.94	14.51
	TimeSFormer	26.21	20.73	23.47	19.55	23.87	21.71	30.50	19.84	25.17	17.55	15.97	16.76
	LaViLa	28.52	*24.25*	*26.39*	23.29	23.84	23.57	41.26	17.68	29.47	20.79	16.43	18.61
	EgoVLP	29.61	21.77	25.69	*22.34*	23.52	22.93	**49.82**	*26.01*	*37.91*	*29.94*	*22.53*	*26.23*
	EgoVLPv2	29.35	22.00	25.67	21.88	*25.30*	*23.59*	48.11	23.84	35.98	27.55	22.43	24.99
Multi view	VI Encoder	19.95	18.64	19.30	19.01	20.40	19.71	19.71	16.17	17.94	16.63	17.22	16.93
	SUM-L	*29.58*	23.33	26.46	22.29	23.27	22.78	36.04	14.73	25.39	15.71	9.81	12.76
	Rosetta	22.38	20.95	21.66	21.66	21.21	21.43	43.82	34.42	39.12	**39.94**	34.77	**37.35**
	Ours	**33.68**	**30.53**	**32.10**	**31.23**	**30.12**	**30.68**	*44.18*	**38.54**	**41.36**	36.34	**35.62**	35.98

our method outperforms other approaches and consistently retrieves the same fine-grained actions across diverse environments, both in ego→ego and exo→exo. The rest of the zero-shot visual results are reserved to the Appendix.

Figure 5. Zero-shot qualitative video-video retrieval results of our method on SSV2. Even though the objects are similar, our method can distinguish between different fine-grained actions.

Table 2. Zero-shot video-video retrieval action top-10 hit-rates of cross-view methods. Ours can generalize to other datasets better.

	SSV2	Charades	NTU
VI Encoder	23.68	15.60	36.61
SUM-L	24.92	13.88	52.70
Rosetta	24.96	**17.98**	58.79
Ours	**29.00**	16.45	**60.69**

Combining atomic actions for keystep retrieval. Each keystep spans a longer time window and typically contains a sequence of atomic actions. We therefore reuse our atomic-action transformer for keystep-level video-video retrieval without additional training. For each keystep segment, we extract the atomic-action video embeddings and temporally average them to obtain a single keystep representation. As shown in Table 1, this simple repurposing is highly effective, matching methods trained directly on keystep labels such as Viewpoint Rosetta and surpassing all other baselines overall. Notably, ours shows the smallest performance change when switching between atomic and keystep representations.

Pretraining VLA with actions for robotics manipulation. We further evaluate our action embeddings on simulated robotic manipulation tasks. Following LAPA, we use latent actions for VLA pretraining and measure task success rates. Specifically, we pretrain OpenVLA (Kim et al., 2024) with atomic actions derived from both LAPA and our method on the SSV2 dataset, and then post-train on LIBERO (Liu et al., 2023) using its task labels. Performance is evaluated on the LIBERO-10 benchmark in simulation.

As shown in Figures 7 and 8, our embedding space improves VLA performance and increases success rates on downstream robotic tasks. Specifically, baseline achieves 34% success rate, LAPA improves +10% and ours improve +18% over the baseline. These results indicate that our representations can help bridge human-centric action understanding and robot-centric action learning. Full details can be found in the Appendix.

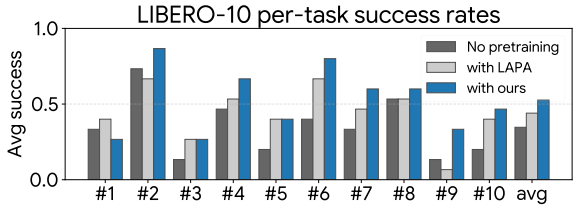

Figure 7. Effect of pretraining with atomic actions on LIBERO-10 tabletop manipulation test suite. Like LAPA, our action latents also improve the task success rate.

Figure 8. A rollout example on LIBERO-10 task #10: *put the 'yellow and white mug' in the microwave and close it*.

4.4. Understanding the Action Representation

Beyond downstream performance, we analyze what information our learned representation encodes and how it structures action semantics. We examine its temporal consistency over long action sequences, invariance to scene and viewpoint, alignment with linguistic verb semantics, and the visual cues it attends to in the transformer layers. Additional details, results, and visualization can be found in the Appendix.

Atomic action probing. We evaluate verb recognition using the trained linear probe with $\mathcal{L}_{\text{CE}}$ loss, applied to longer videos containing sequences of atomic actions (see Figure 9). Although our method performs chunk-wise inference without explicit temporal memory or autoregressive decoding, the probe remains temporally consistent and captures meaningful action transitions. This indicates that our representation supports longer-range atomic action chains.

Figure 9. Atomic action probing results of our method on Ego-Exo4D. Our method can support estimating multiple atomic sequences from longer videos.

Figure 6. Zero-shot video-video retrieval results on NTU RGB+D dataset of our method. Our action embeddings can match the same-labeled actions under different environments and viewpoints.

Latent clustering inspection. We extract atomic-action embeddings from Ego-Exo4D and visualize them using t-SNE (van der Maaten & Hinton, 2008) in Figure 10. EgoVLPv2 clusters embeddings primarily by viewpoint, while Rosetta clusters largely by scene identity. In contrast, ours are less scene-dependent and exhibit more uniformity across scenes.

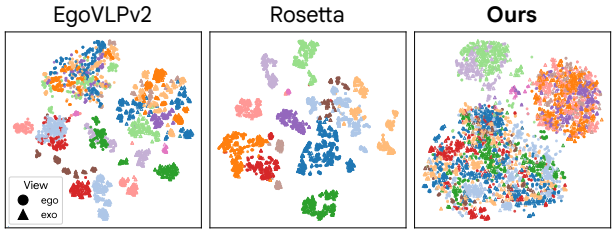

Figure 10. We visualize embeddings in all scenes+all actions. Colors denote different scenes. Prior methods cluster largely by scene or viewpoint, while our embeddings are less scene-dependent and provide a more homogeneous distribution.

Semantic alignment. We assess how well our atomic-action embeddings capture linguistic verb similarity by comparing their induced verb–verb similarity structure with WordNet (Miller, 1995). WordNet provides a lexical taxonomy from which we derive a reference similarity between verbs based on their semantic relatedness. As shown in Figure 11, we group Ego-Exo4D validation embeddings by ground-truth atomic-action verbs and compute verb prototypes. For each method, we then derive a verb–verb similarity matrix and measure its Spearman correlation with the WordNet similarity matrix. Our method achieves the highest alignment, indicating better preservation of relative verb semantics.

Cross-attention inspection. We visualize cross-attention heatmaps for ego-exo inputs in Figure 12. Specifically, we visualize the pooled cross-attention between the learned `CLS` token and the injected DINOv3 tokens, projected onto the corresponding RGB frames. The results show our actions attend primarily to interaction regions and motion cues.

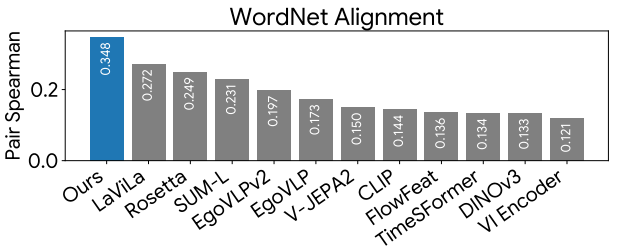

Figure 11. Spearman correlation between each model's pairwise action-similarity matrix and the WordNet ground-truth similarity matrix. Our embeddings are more linguistically aligned.

Figure 12. Cross-attention heatmaps from our model. Brighter regions indicate stronger `CLS` token attention.

5. Conclusion

We present **InvAct**, the first view- and scene-invariant method for atomic action learning. We show that prior view-invariant approaches can still overfit to appearance, and we mitigate this by carefully combining our training objectives and inputs. Our method achieves competitive performance in keystep and atomic action retrieval, supports zero-shot transfer, and we analyze the learned representations via probing, language alignment, attention maps, and a robotic downstream demonstration. We view this as a promising step toward unifying robot-oriented and human-oriented action learning.

Limitations and future work. We currently embed atomic clips independently, so exploring autoregressive variants is a natural next step. It will also be important to scale to larger data (e.g., Action100 (Chen et al., 2026)) and eventually unlabeled video. Finally, beyond retrieval and verb-classifier probing, it would be interesting to adapt token-level video-to-text decoding methods like VL-JEPA (Chen et al., 2025).

6. Potential Broader Impact

This paper advances action representation learning, with the goal of improving generalization across viewpoints and scenes, and aims to close the gap between robotics and human-understanding domains. Better action embeddings can help in applications like robotics, assistive systems, and video understanding. Our work does not involve new data collection or human-subject studies, and we train on existing and verified datasets. We encourage careful use and evaluation.

References

Araslanov, N., Ribic, A., and Cremers, D. FlowFeat: Pixel-dense embedding of motion profiles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Ashutosh, K., Girdhar, R., Torresani, L., and Grauman, K. HierVL: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23066–23078, 2023.

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Bagad, P. and Zisserman, A. Chirality in action: Time-aware video representation learning by latent straightening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Bagad, P. and Zisserman, A. TARA: Simple and efficient time aware retrieval adaptation of mllms for video understanding. *arXiv preprint arXiv:2512.13511*, 2025b.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824, 2021.

Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., and Li, H. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2026.

Bu, X., Lyu, J., Sun, F., Yang, R., Ma, Z., and Li, W. LAOF: Robust latent action learning with optical flow constraints. *arXiv preprint arXiv:2511.16407*, 2025.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., and Fung, P. VL-JEPA: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Chen, D., Kasarla, T., Bang, Y., Shukor, M., Chung, W., Yu, J., Bolourchi, A., Moutakanni, T., and Fung, P. Action100M: A large-scale video action dataset. *arXiv preprint arXiv:2601.10592*, 2026.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5842–5850, 2017.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jeong, Y., Chun, J., and Kim, T. Learning to act robustly with view-invariant latent actions. *arXiv preprint arXiv:2601.02994*, 2026.

Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., and Feng, J. How far is video generation from world model: A physical law perspective. In *Forty-second International Conference on Machine Learning*, 2025.

Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., and Xu, D. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation*, pp. 13226–13233. IEEE, 2025.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E. P., Sanketi, P. R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., and Finn, C. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Ko, P.-C., Mao, J., Du, Y., Sun, S.-H., and Tenenbaum, J. B. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*, 2024.

Lin, K. Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E. Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W., et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. 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, 2020.

Luo, M., Xue, Z., Dimakis, A., and Grauman, K. Viewpoint Rosetta Stone: unlocking unpaired ego-exo videos for view-invariant representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15802–15812, June 2025.

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Miller, G. A. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, November 1995. ISSN 0001-0782.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Park, J., Lee, J., and Sohn, K. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13661–13670, June 2025.

Pramanick, S., Song, Y., Nag, S., Lin, K. Q., Shah, H., Shou, M. Z., Chellappa, R., and Zhang, P. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5285–5297, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016.

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Alahari, K. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Teed, Z. and Deng, J. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer International Publishing, 2020.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wang, Q., Zhao, L., Yuan, L., Liu, T., and Peng, X. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3307–3317, 2023.

Xiong, H., Xu, X., Wu, J., Hou, Y., Bohg, J., and Song, S. Vision in action: Learning active perception from human demonstrations. In *9th Annual Conference on Robot Learning*, 2025.

Xu, M., Xu, Z., Chi, C., Veloso, M., and Song, S. Xskill: Cross embodiment skill discovery. In *7th Annual Conference on Robot Learning*, pp. 3536–3555, 2023.

Xue, Z., Grauman, K., Damen, D., Zisserman, A., and Han, T. Seeing without pixels: Perception from camera trajectories. *arXiv preprint arXiv:2511.21681*, 2025.

Xue, Z. S. and Grauman, K. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.

Ye, S., Jang, J., Jeon, B., Joo, S. J., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.-W., Lin, B. Y., Liden, L., Lee, K., Gao, J., Zettlemoyer, L., Fox, D., and Seo, M. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Appendix

In the Appendix Section A, we provide additional visuals for zero-shot and in-domain retrieval results, action probing, attention maps visualization, and latent clustering. For quantitative results, we discuss same-scene fractions in retrieval sets for the baseline method and ours, and provide an extensive experiment comparing losses, input types, and training paradigms. We also give details about VLA finetuning experiments. In Section B, we detail the training setup, utilized datasets and baseline configurations.

A. Additional Results

A.1. Retrieval and probing

We show additional results in zero-shot retrieval in Figures 13 and 18, elaborate on the performance gap between all-scenes and cross-scenes in Figure 14 for the baseline Viewpoint Rosetta, show verb-level atomic-action probing in Figure 17, and ego-exo retrieval in Figures 19 to 22.

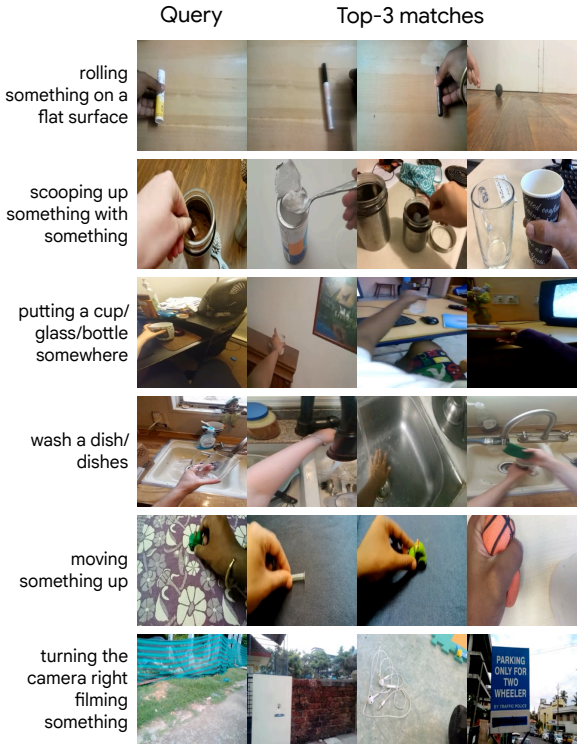

Figure 13. Zero-shot qualitative video-video retrieval results of our method on SSV2 and Charades-Ego datasets. Our method can successfully extract the actions and populate top-3 matches.

A.2. Attention maps and clustering

We provide additional results on attention map visualization in Figure 15, along with more t-SNE visualizations in Figure 16.

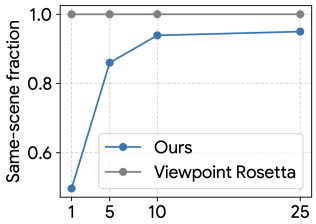

Figure 14. Scores calculated on Ego-Exo4D validation dataset. Same-scene fraction is defined as the proportion of top-k matches drawn from the same scene as the query. The baseline method is heavily biased toward same-scene retrievals across all k, whereas our method substantially reduces this bias while maintaining strong retrieval performance.

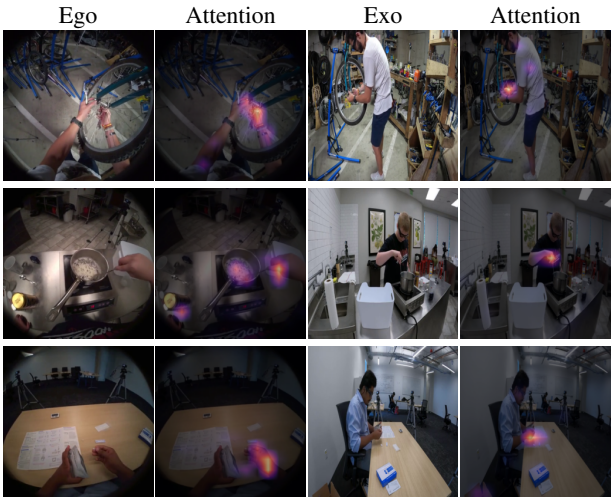

Figure 15. Cross-attention heatmaps from our model. Brighter regions are where `CLS` tokens attend the most. Our action embeddings attend to interaction regions the most.

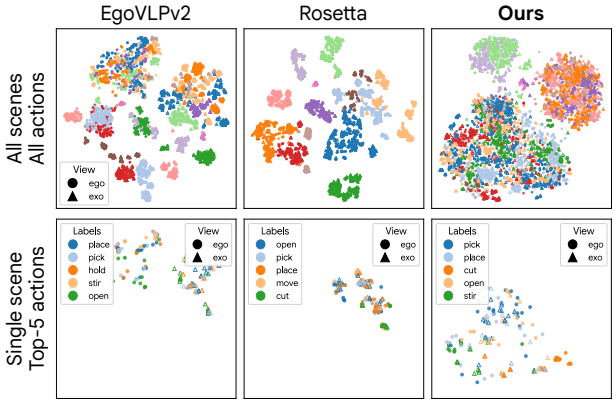

Figure 16. We visualize embeddings in all scenes/all actions (row 1), and a single scene for the top-5 actions (row 2). Different colors denote different scenes in row 1, and different colors denote different verbs in row 2. Prior methods cluster largely by scene or viewpoint, while our embeddings are less scene-dependent and preserve better action grouping across scenes and views.

A.3. VLA pretraining for robotics manipulation task

We follow the same three-stage pipeline as LAPA (Ye et al., 2025): (1) learn a latent action space, (2) pretrain a vision–language–action (VLA) model to predict these latent actions, and (3) post-train the VLA to output robot control actions for simulation. We use OpenVLA (Kim et al., 2024) as the VLA backbone and LIBERO (Liu et al., 2023) for robot post-training and evaluation.

Stage 1: Latent actions. We obtain two latent spaces: LAPA's pretrained `LAQ` model trained on Something-Something V2 (SSV2) (Goyal et al., 2017), and our atomic-action model trained on Ego-Exo4D (Grauman et al., 2024).

Stage 2: VLA pretraining with latents. We finetune OpenVLA using 32-rank adapters (Hu et al., 2022) with batch size 32 on SSV2. Inputs are video frames and the tokenized action text description (e.g., *rolling something on a flat surface*). We attach a single linear head that maps transformer tokens to the latent action vector and train the VLA to match either the non-quantized `LAQ` encoder outputs or our latent actions. Both variants are trained for 25,000 steps on a single H200 GPU.

Stage 3: Robot post-training. We remove the latent-action head, add a new 32-rank LoRA, and finetune on LIBERO to predict continuous robot embodiment actions used by the simulator controller. We add a new trainable output layer that maps transformer tokens to robot actions. In the end, we train three VLA models: (i) no latent pretraining, (ii) latent-pretrained with LAPA, and (iii) latent-pretrained with our embeddings. All use batch size 32 and are trained for 50,000 steps on a single H200 GPU.

Evaluation. We report the average success on the LIBERO-10 test suite with three trained models. For each of the 10 tasks, we run 20 trials and average success rates across tasks. The visuals for the trials per task are given in Figures 23 and 24. Each row in the figures describes a task, totaling up to 10 tasks. The task goals are given in text, and the locations of the objects are randomized per trial.

The task goals are given:

1. `put both the alphabet soup and the tomato sauce in the basket`
2. `put both the cream cheese box and the butter in the basket`
3. `turn on the stove and put the moka pot on it`
4. `put the black bowl in the bottom drawer of the cabinet and close it`
5. `put the white mug on the left plate and put the yellow and white mug on the right plate`
6. `pick up the book and place it in the back compartment of the caddy`
7. `put the white mug on the plate and put the chocolate pudding to the right of the plate`
8. `put both the alphabet soup and the cream cheese box in the basket`
9. `put both moka pots on the stove`
10. `put the yellow and white mug in the microwave and close it`

A.4. Hyperparameter ablation

We perform an extensive ablation with regard to training, input types, and losses in Table 3. We train for fewer iterations for ablation models and train the best model for twice the iterations as our final checkpoint.

Table 3. Video-video retrieval atomic action hit-rates of our method, in all scenes of Ego-Exo4D validation set. CE and CLR denote cross-entropy and contrastive learning, respectively.

	Cross-scene (↑)			Cross-view (↑)		
	g→g	x→x	avg.	g→x	x→g	avg.
Training paradigm						
CLR_{verb}	24.04	23.41	23.72	18.87	22.44	20.65
CE_{verb}	23.70	24.17	23.93	23.47	25.67	24.57
CLR_{atom}	26.82	24.41	25.61	25.99	24.90	25.45
CE_{verb}+CLR_{atom}	27.17	27.75	27.46	27.18	27.74	27.46
Losses						
Text alignment	26.82	24.41	25.61	25.99	24.90	25.45
+View invariance	28.30	24.53	26.41	27.15	27.18	27.16
+Action attract	29.55	24.32	26.93	25.09	25.80	25.45
+Scene repel	29.72	26.67	28.20	27.06	27.23	27.14
Input modality						
RGB	24.44	22.87	23.65	24.20	24.45	24.32
Flow	24.18	23.36	23.77	22.28	23.91	23.05
RGB+flow	25.45	24.46	25.96	24.06	24.88	24.47
CLIP+flow	29.72	26.67	28.20	27.06	27.23	27.14
DINOv3+flow	31.57	28.92	30.25	29.86	29.17	29.52
Trained longer	**33.68**	**30.53**	**32.10**	**31.23**	**30.12**	**30.68**

Training paradigm. We compare verb-level and sentence-level (atomic) supervision under contrastive (CLR) and cross-entropy (CE) objectives. Verb-level CLR performs poorly on cross-view retrieval, suggesting that coarse verb labels provide an unstable contrastive signal at our batch sizes. Switching to verb-level CE substantially improves cross-view performance (+19.0%). Using sentence-level atomic captions with CLR is overall stronger: CLR_{atom} improves the cross-scene and cross-view averages over CE_{verb} by +7.0% and +3.6%, respectively. Finally, combining the two losses CE_{verb} + CLR_{atom} yields the best results, increasing CLR_{atom} by +7.2% (cross-scene) and +7.9% (cross-view), indicating that the two objectives are complementary.

Losses. Starting from text alignment (CLR_{atom}), adding

Figure 17. Qualitative atomic action probing results of our method on Ego-Exo4D validation set, for ego and exo videos. Despite not being autoregressive or not having context memory, our method can successfully support longer-range atomic sequences.

Figure 18. Zero-shot video-video retrieval results on NTU RGB+D dataset of our method. Our action embeddings can match the same-labeled actions, with different environments and camera viewpoints.

view-invariance improves both cross-scene and cross-view retrieval, with a larger gain on cross-view (+6.7%). Adding action attract further improves the cross-scene (+5.2%), but does not improve the cross-view on average. This is expected since action-attract promotes cross-scene consistency by tightening action clusters, yet without an explicit repel component it can over-collapse representations and disturb viewpoint-specific structure. Finally, adding scene repel recovers and yields the strongest overall results, boosting the cross-scene (+10.1%) and the cross-view averages (+6.6%) over the base case.

Input modality. RGB-only features perform well for cross-view retrieval, but they are largely driven by appearance, which transfers less reliably across scenes. Optical flow provides a complementary motion prior, yielding modest gains in cross-scene retrieval (+2.1% on the hardest split, +0.5% on the cross-scene average). However, flow alone degrades cross-view performance (-5.2%), indicating that motion is informative but insufficient for ego–exo matching. Combining RGB and flow mitigates these failure modes (+9.8% cross-scene avg., +0.6% cross-view avg. over RGB). Using frozen foundation embeddings as guidance further improves: CLIP+flow increases the cross-scene and cross-view averages by +19.2% and +11.6%, and DINOv3+flow achieves the best gains (+27.9% and +21.4%). In contrast to approaches that fine-tune pretrained CLIP image encoders as their backbone (Luo et al., 2025; Xue & Grauman, 2023), we only use the output of such pretrained encoders as semantic guidance via cross-attention.

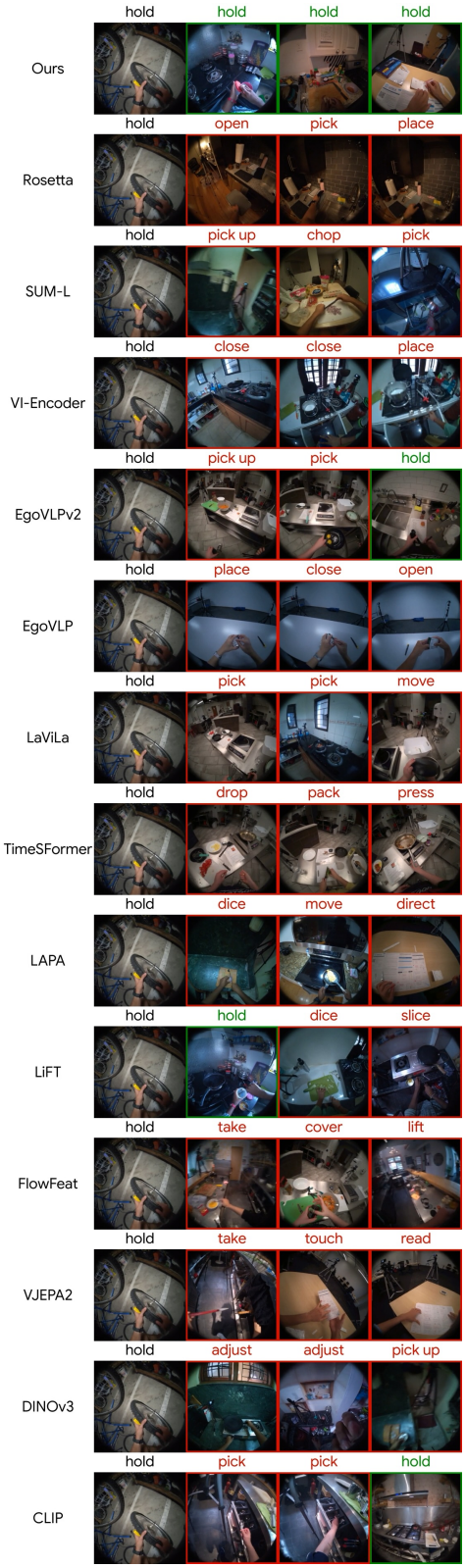

Figure 19. Video-video cross-scene **ego-ego retrieval** results on Ego-Exo4D dataset. First column is the query video, and other columns are top-3 hits.

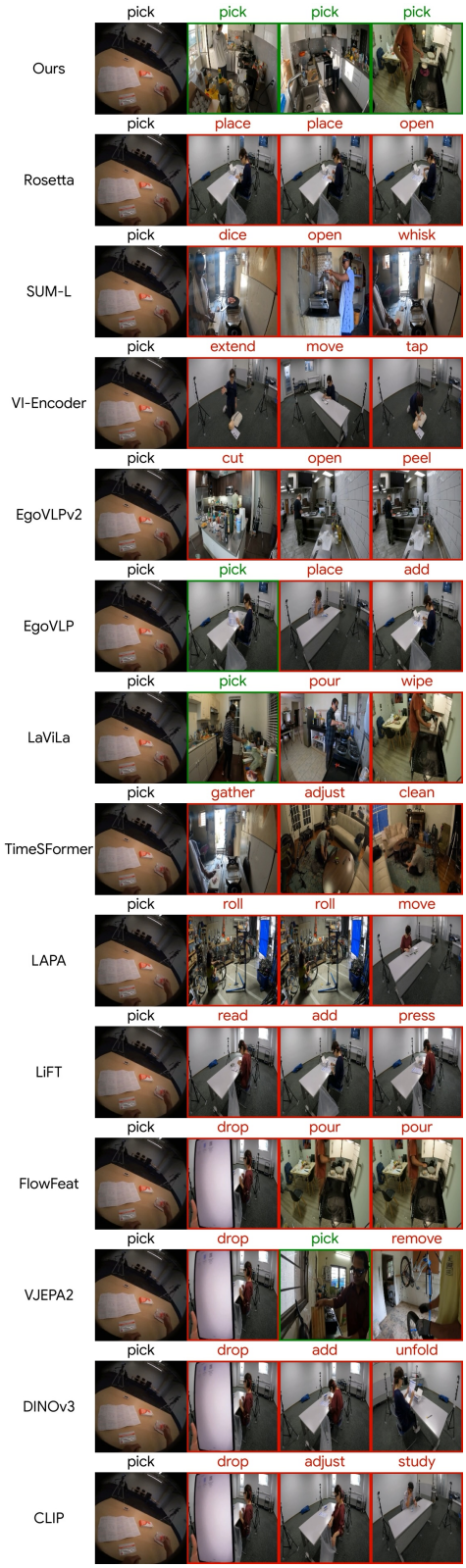

Figure 20. Video-video cross-scene **ego-exo retrieval** results on Ego-Exo4D dataset. First column is the query video, and other columns are top-3 hits.

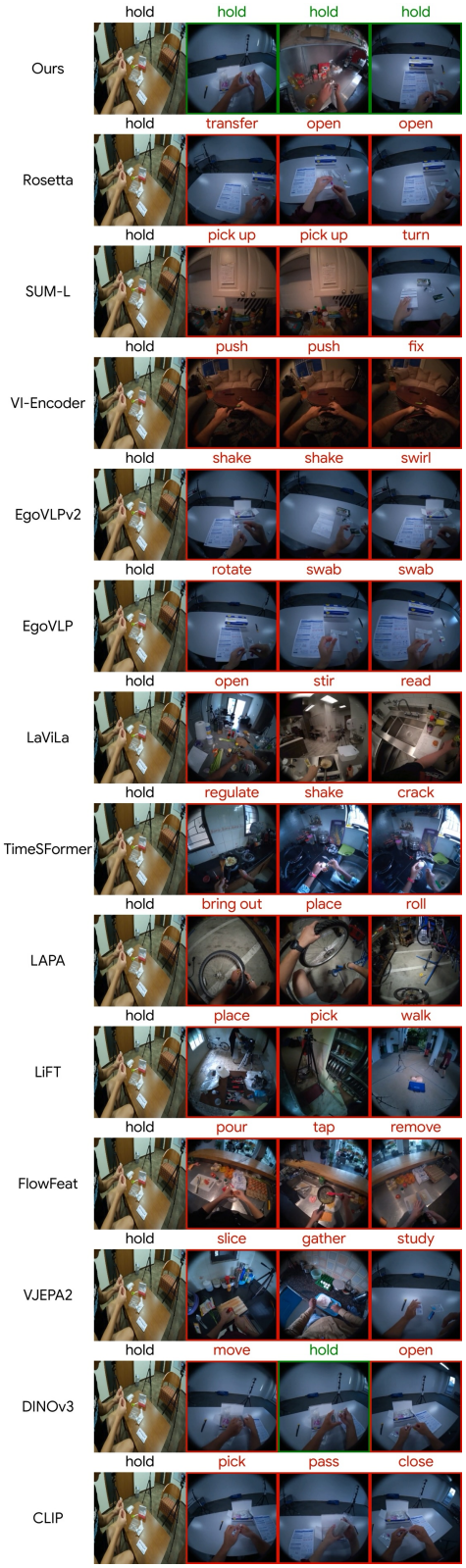

Figure 21. Video-video cross-scene **exo-ego retrieval** results on Ego-Exo4D dataset. First column is the query video, and other columns are top-3 hits.

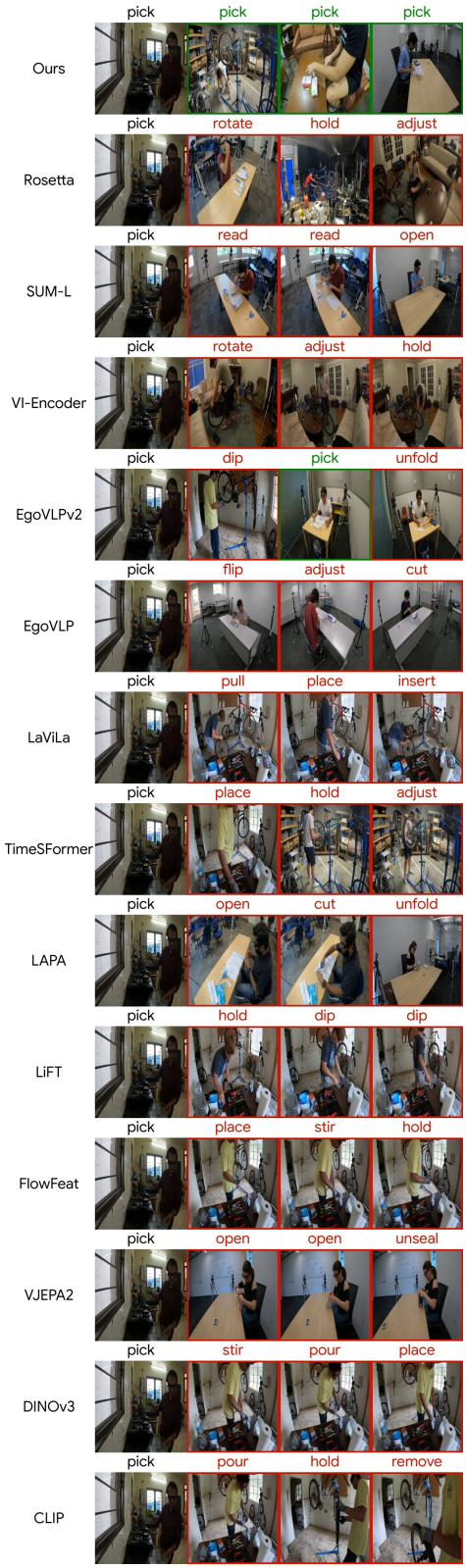

Figure 22. Video-video cross-scene **exo-exo retrieval** results on Ego-Exo4D dataset. First column is the query video, and other columns are top-3 hits.

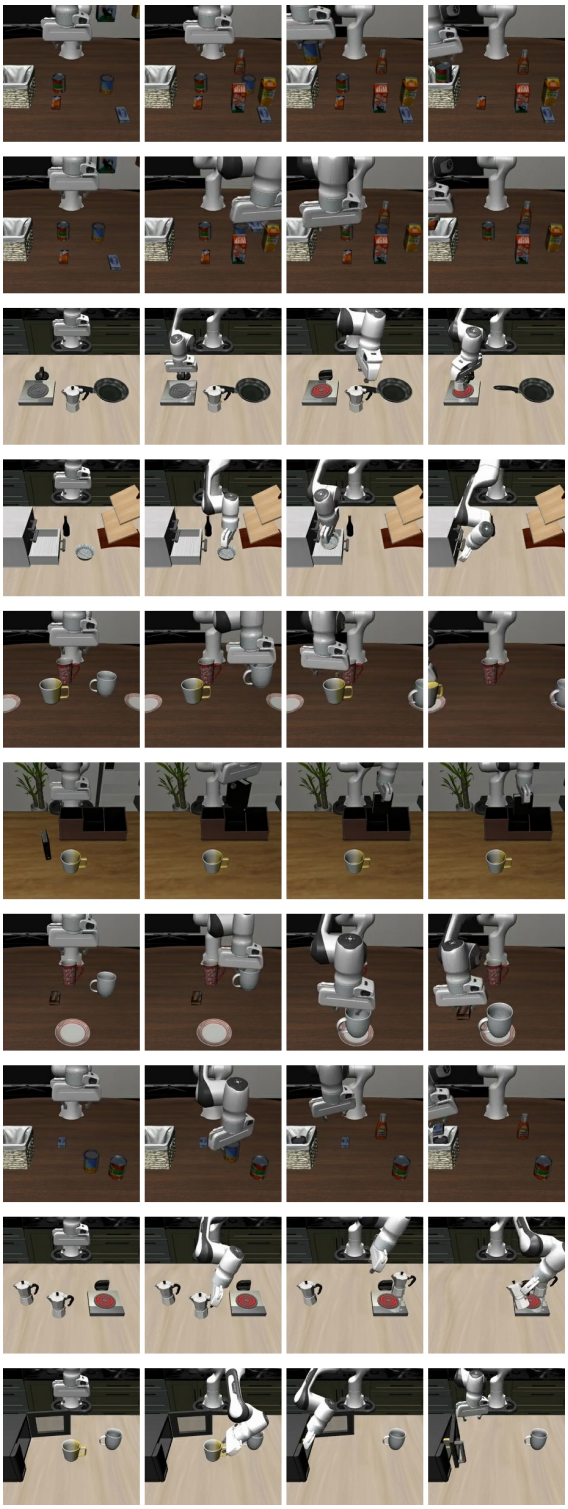

Figure 23. Successful runs on LIBERO-10 evaluation suite. Each row is a task. We display 4 evenly subsampled frames from the simulation output video in each row.

Figure 24. Failed runs on LIBERO-10 evaluation suite. Each row is a task. We display 4 evenly subsampled frames from the simulation output video in each row.

B. Implementation and Evaluation Details

B.1. Training

Training hyperparameters. We train our method, initialized from random weights, distributed on 4×H200 GPUs for 50,000 steps, with `bf16` mixed precision, using AdamW optimizer and constant learning rate of 1e-4. The loss hyperparameters are $\lambda_{\text{vv}} = 0.5$, $\lambda_{\text{CE}} = 0.5$, $\lambda_{\text{att}} = 1.0$, $\lambda_{\text{rep}} = 1.0$.

Batch sampling. Each minibatch has 4 ego-exo time-synchronized video tuples, resulting in 8 videos per minibatch. Concretely, each minibatch is a quadruplet containing two actions from the set $\mathcal{A}$=`{A1,A2}` and three scenes from the set $\mathcal{S}$=`{S1,S2,S3}`. We opt for the layout [`S1A1`, `S2A1`, `S3A1`, `S3A2`] $\subset \mathcal{S} \times \mathcal{A}$. The sets $\mathcal{A}$ and $\mathcal{S}$ are random subsets of the whole Ego-Exo4D training split for each iteration; therefore, the elements in $\mathcal{A}$ and $\mathcal{S}$ differ in each sampling operation. Figure 25 visualizes a minibatch, with atomic verb and atomic sentence labels.

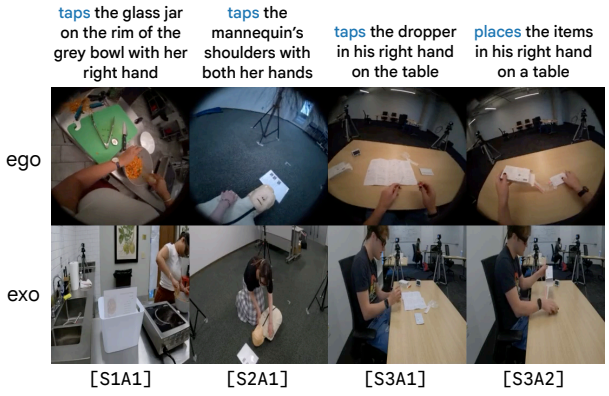

Figure 25. A sample minibatch. Verb-level atomic-action labels are given in blue, and sentence-level atomic-action labels are given in black colors, respectively.

B.2. Datasets

Ego-Exo4D. For our main experiments, we train and evaluate in Ego-Exo4D (Grauman et al., 2024) using the `downscaled_takes/448` subset of time-synchronized ego–exo clips annotated with both keysteps and atomic actions. Keysteps label longer temporal segments that capture the overall activity, and each keystep window typically contains a sequence of finer-grained atomic-action labels. **We train *only* on atomic-action labels and evaluate *both* atomic-action and keystep-action labels.** In contrast to previous cross-view works (Luo et al., 2025; Wang et al., 2023) that often report results only on `cooking`, we evaluate the entire Ego-Exo4D validation split spanning {`cooking, health, bike`}, using their set of shared atomic-action labels. Because the atomic-action labels are provided as free-form sentences rather than a fixed taxonomy, we parse them manually to extract verb classes. We also filter out atomic labels containing typographical errors and discard action clips shorter than 1 s and longer than 5 s. After preprocessing, the dataset provides 30,615 training clips and 8,961 validation clips (each paired with an atomic-action label). In the end, the training split we use includes 365 verb classes and 280 keysteps, while the validation split includes 234 verb classes and 270 keysteps. We obey the training and validation splits of the original `downscaled_takes/448`.

PhyWorld. PhyWorld dataset (Kang et al., 2025) comes with several synthetic subsets (e.g., `collision`, `trajectory`, etc.), and we focus on `collision` for the visual experiments to compare our method and Viewpoint Rosetta. Because everything is rendered in 2D, there is no real "viewpoint" shift like in ego–exo videos. This means that the changes are mostly in appearance and setup. In `collision`, two balls with different colors, sizes and shapes start in different parts of the frame, collide, and then move away, sometimes in different horizontal directions and at different speeds. When we run retrieval on this subset, Viewpoint Rosetta often matches clips by how the balls look (color/size/shape) instead of how they move (direction/speed after the collision). Our method is more driven by the dynamics, therefore indicating better fine-grained atomic-action understanding. This matches what we see on Ego-Exo4D and reinforces our appearance-overfit claim.

Zero-shot datasets. We evaluated zero-shot video-video retrieval in Charades-Ego (Sigurdsson et al., 2018) (egocentric split, 9,338 samples), NTU RGB+D (Shahroudy et al., 2016; Liu et al., 2020) (fully exocentric, 56,880 samples) and the validation split of Something-Something V2 (Goyal et al., 2017) (predominantly egocentric, 24,777 samples). Since PhyWorld does not provide text labels, we report score-based metrics on the latter three datasets.

B.3. Baselines

For all baselines utilized, we use their pretrained checkpoints available in their repositories. The details of their structure, frame numbers T, input resolutions, and output dimensions are given in Table 4.

Why we do not include AE2 and BYOV. AE2 (Xue & Grauman, 2023) and BYOV (Park et al., 2025) focus on learning correspondences between ego and exo videos without requiring temporal alignment. In contrast, Ego-Exo4D provides time-aligned ego–exo pairs, which we explicitly leverage when sampling clips for training. Moreover, AE2 and BYOV are not trained as unified embedding models. Instead, they release separate checkpoints for a small set of specific tasks (e.g., *Break Eggs, Pour Milk, Pour Liquid, Tennis Forehand*), each finetuned for that task. In practice, these methods extract an ego clip embedding and search for its time-synchronized exo segment, or vice versa, rather than learning a single representation that generalizes across

Table 4. Backbone architectures and embedding configurations for all compared methods.

Method	Backbone	T	Res.	Dim
Ours	Transformer + `RAFT` + `DINOv3`	8	256^2	256
Rosetta	`CLIP_OPENAI_TIMESFORMER_BASE` (Radford et al., 2021; Bertasius et al., 2021)	4	224^2	256
LaViLa	`CLIP_OPENAI_TIMESFORMER_BASE` (Radford et al., 2021; Bertasius et al., 2021)	16	224^2	256
SUM-L	`MultiTaskSlowFast R101 (8×8)` (Feichtenhofer et al., 2019)	32	280^2	397
EgoVLPv2	`FrozenInTime base_patch16_224` (Bain et al., 2021)	32	224^2	256
EgoVLP	`FrozenInTime STT base_patch16_224` (Bain et al., 2021; Bertasius et al., 2021)	16	224^2	256
TimeSformer	`TimeSformer vit_base_patch16_224` (Bertasius et al., 2021)	8	224^2	768
LAPA	`LatentActionQuantization (LAQ)` (Ye et al., 2025)	2	256^2	128
LiFT	`dinov2_vits14` (Oquab et al., 2024)	16	224^2	768
FlowFeat	`dinov2_vits14_yt` (Oquab et al., 2024)	8	224^2	768
V-JEPA 2	`vjepa2-vitl-fpc16-256-ssv2` (Assran et al., 2025)	16	256^2	1024
DINOv3	`dinov3-vits16plus` (Siméoni et al., 2025)	8	224^2	384
CLIP	`clip-vit-base-patch32` (Radford et al., 2021)	8	224^2	768

actions. Because they are task-specific and not designed for general-purpose retrieval or atomic-action embedding, we omit them from our comparisons.