INVACT: REDUCING SCENE BIAS IN VIEW-INVARIANT ATOMIC ACTION LEARNING

Anonymous authors
Paper under double-blind review

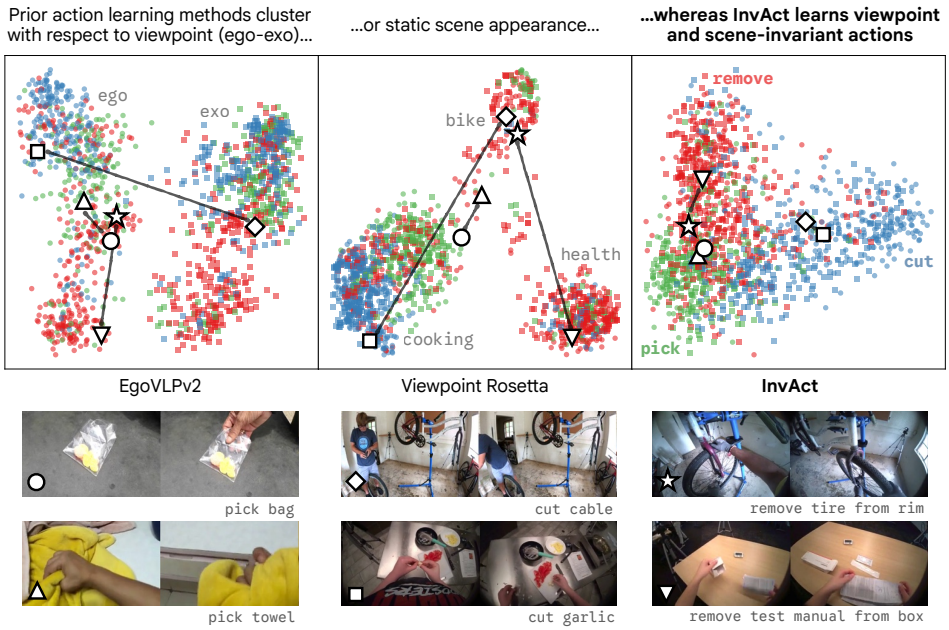

Figure 1: **InvAct groups atomic actions across scenes and viewpoints.** EgoVLPv2 (Pramanick et al., 2023) clusters by viewpoint, while Viewpoint Rosetta (Luo et al., 2025) clusters by scene. It separates `cut cable` from `cut garlic`, yet groups `cut cable` with `remove tire`. InvAct instead aligns matching actions. Colors denote actions. Circles and squares denote ego and exo views. Hollow markers identify clips below, and lines connect matching actions.

ABSTRACT

View-invariant action learning methods aim to match actions from videos across ego-exo views. However, because these views also share the same environment, models can match scenes rather than actions, limiting fine-grained atomic action understanding. We address this with **InvAct**, an atomic action embedding model designed for viewpoint and scene invariance. To reduce reliance on scene cues, we propose structured token commitment, a hard row-wise attention mask that routes each video token to a single special-token slot. Action-relevant tokens align with the class token, while non-action tokens align with register tokens that are discarded at inference. We further introduce a *scene-aware contrastive objective* that attracts same-action clips across scenes while separating different actions within a scene. On Ego-Exo4D, InvAct improves cross-scene and cross-view atomic retrieval while reducing same-scene bias. Importantly, despite training only on atomic actions, combining its embeddings provide strong longer action retrieval without retraining. The same checkpoint transfers to unseen datasets without finetuning and provides useful latent targets for robot-policy pretraining. These results highlight scene dependence as an important consideration in view-invariant action learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

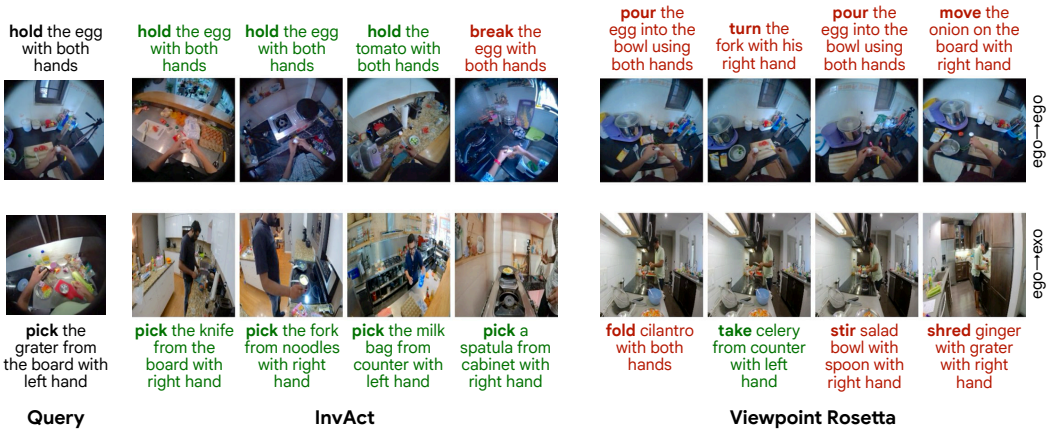

Figure 2: **View alignment can coexist with scene-dependent retrieval.** Baselines often retrieve different actions from the same environment by matching static scene cues. InvAct retrieves matching actions across the viewpoints and scenes shown. Text descriptions are shown only for visualization and are not used during inference.

1 Introduction

Action representation learning aims to capture what is being done even when the camera and surroundings change. Such representations support 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), vision-language-action (VLA) pretraining for robotics (Bu et al., 2026), and conditioning world models (Gao et al., 2026).

Actions also can vary in granularity. **Keystep actions**, such as checking for damage or cooking an omelet, are longer procedures tied to objects and context. They decompose into short, interaction-centric **atomic actions**, such as pushing, pulling, and cutting (Grauman et al., 2024).

Recent works (Xue & Grauman, 2023; Grauman et al., 2024; Park et al., 2025; Luo et al., 2025) emphasize **viewpoint invariance**, aligning representations across egocentric and exocentric views, often at the keystep level. Yet, viewpoint alignment alone does not guarantee action-centric representations. Synchronized ego–exo views share not only the action but also objects, layouts, and backgrounds, allowing models to match views through static scene cues rather than action dynamics. A representation can therefore become view-invariant while remaining scene-dependent, organizing clips by environment rather than action (Fig. 1). This is particularly problematic for fine-grained atomic actions, where different actions can occur within the same scene. Retrieval in restricted settings may therefore hide scene dependence, while cross-scene matching asks whether the same action can be recognized when shared context changes.

Figure 2 makes this failure mode concrete on Ego-Exo4D, with **all scenes included** in the retrieval pool. Baselines rank clips with shared layouts and objects above matching actions elsewhere. In the ego→exo examples, Viewpoint Rosetta Stone (Luo et al., 2025) retrieves the same kitchen, including the blue bowl and brown cabinets, while missing matching actions from other scenes.

To address this, we introduce **InvAct**, a view-invariant atomic-action embedding model designed with scene invariance in mind. Its **scene-aware contrastive objective** matches the same action across scenes while separating different actions within a scene. Its **structured token commitment** uses a hard row-wise attention mask that restricts each video token to one special-token read source in `[CLS,REG]`. With action supervision applied to `CLS`, register tokens provide alternative contextual pathways, implicitly collecting non-action information. Sentence- and verb-level labels further ground the embeddings, with no text required at inference. Our encoder combines optical flow (Teed & Deng, 2020) for motion with frozen pretrained appearance features (Siméoni et al., 2025).

Our primary evidence is on Ego-Exo4D. InvAct improves atomic retrieval across scenes and views while reducing measured same-scene retrieval bias. Controlled ablations isolate the contributions of scene repulsion and token commitment (Table 13 and Section D), and we compare against general-purpose, robot-centric, and human-centric action representations (Table 1).

Importantly, although trained only on atomic actions, InvAct also supports longer actions: combining its atomic embeddings yields strong keystep retrieval without keystep-level training (Table 1). We further evaluate the same checkpoint without finetuning on four unseen datasets (Sigurdsson et al., 2018; Goyal et al., 2017; Liu et al., 2020; Kang et al., 2025) and explore its embeddings as latent targets for VLA pretraining (Ye et al., 2025). These experiments test the broader utility of the learned atomic representation beyond its training objective.

Our contributions are as follows.

- We identify **scene dependence in view-invariant action representations**: aligning ego–exo views can leave embeddings reliant on shared environments. Motivated by this, we introduce **InvAct** to explicitly address scene variation alongside viewpoint invariance in atomic action learning.
- We combine **structured token commitment**, a parameter-free attention mask routing each video token to one special token in `[CLS,REG]`, with a **scene-aware contrastive objective** that attracts same-action clips across scenes and repels different actions within a scene. Controlled ablations isolate their contributions and study training and input choices (Sections 3.2, 3.3 and D and Table 13).
- We show that **atomic embeddings support longer actions**: combining embeddings enable strong keystep retrieval without keystep-level training (Section 4.2 and Table 1).
- We compare InvAct with general-purpose (Radford et al., 2021; Siméoni et al., 2025; Assran et al., 2025; Araslanov et al., 2025; Bagad & Zisserman, 2025a), robot-centric (Ye et al., 2025; Jeong et al., 2026), and human-centric (Grauman et al., 2022; Zhao et al., 2023; Lin et al., 2022; Pramanick et al., 2023; Wang et al., 2023; Xue & Grauman, 2023; Park et al., 2025; Grauman et al., 2024; Luo et al., 2025) representations on atomic and keystep retrieval, and evaluate transfer to four unseen datasets without finetuning (Tables 1 and 2). We further analyze long-sequence predictions, latent structure, attention, scene bias, and linguistic alignment, and explore InvAct latents for VLA pretraining, evaluated on LIBERO-10 benchmark. (Section 4.2)

These results highlight scene dependence as an important consideration in view-invariant action learning. We will release code and trained models.

2 Related Work

Generalizable representations. Large pretrained models like CLIP (Radford et al., 2021), DINOv3 (Siméoni et al., 2025), and TimeSFormer-based architectures (Bertasius et al., 2021) learn useful visual representations from large-scale training. V-JEPA 2 (Assran et al., 2025) further improves temporal modeling of videos, while FlowFeat (Araslanov et al., 2025) emphasizes motion-driven features. LiFT (Bagad & Zisserman, 2025a) converts image features into video descriptors by linearizing feature trajectories. These models do not explicitly target both scene and view invariance.

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 contrastive matching between frames and text at multiple scales. 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 with temporal order. These directions complement ours by improving semantic grounding through video-text alignment, similar to how we use atomic labels. Our focus is on using scene labels to discourage appearance-based action matches. We use language only in the loss.

Viewpoint in action learning. The Ego-Exo4D benchmark supports large-scale learning from synchronized first- and third-person videos (Grauman et al., 2024), and works like Viewpoint Rosetta Stone (Luo et al., 2025) pursue view-invariant alignment in weakly paired settings. SUM-L (Wang et al., 2023) focuses on egocentric learning, but utilizes unpaired ego-exo videos with text label guidance. 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. We jointly target cross-view alignment and reduced scene dependence. We evaluate across the Ego-Exo4D validation categories used in our protocol, beyond the restricted 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), LAOF (Bu et al., 2025), and LAPO (Schmidt & Jiang, 2023) 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. DreamDojo (Gao et al., 2026) uses latent actions in physics-aware world model pretraining and conditioning. Ko *et al.* (Ko et al., 2024) extract action-relevant structure from videos via flow and depth for policy learning. These works share our intuition that motion helps identify actions when appearance varies. We study language-supervised atomic embeddings on Ego-Exo4D, focusing on retrieval across views and scenes alongside an initial evaluation of robot pretraining.

Special tokens and attention re-routing. We consider `CLS` and `REG` (register) tokens as special tokens. Register tokens (Darcet et al., 2024) provide auxiliary workspace in vision transformers, absorbing internal computation while being discarded at the output. Prior works have explored several forms of structured attention routing. Slot Attention (Locatello et al., 2020) used token-to-slot assignment for object-centric decomposition. Mask2Former (Cheng et al., 2022) restricted queries to predicted mask regions. WSSS (Hanna & Borth, 2025) introduced class-specific token masking for weakly supervised segmentation. Marouani *et al.* (Marouani et al., 2026) showed that specialized `CLS`–patch processing paths improve separation between global and local computations. Our hard video-to-special-token routing encourages action-relevant content in `CLS`.

3 METHOD

We learn **atomic action embeddings across views and scenes**. Our transformer combines contrastive and classification losses with structured token commitment to reduce scene bias.

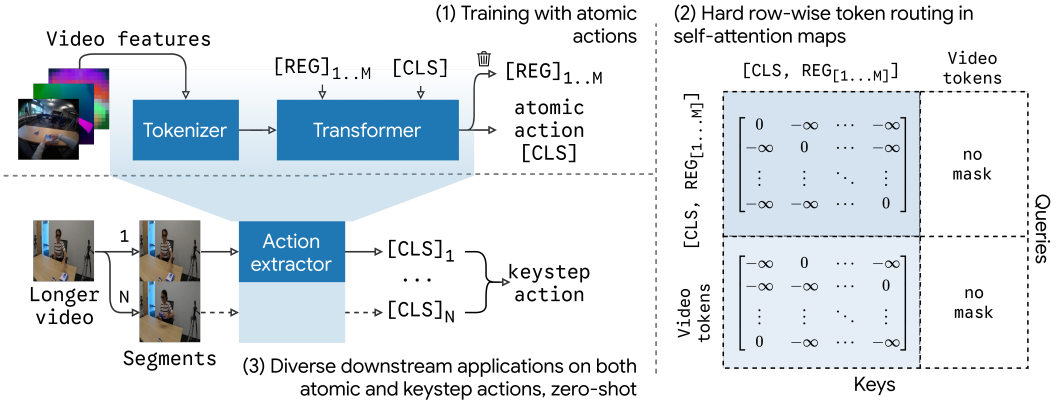

Figure 3: **InvAct learns atomic action embeddings that transfer across scenes and views.** We train on atomic clips with contrastive and semantic losses in block 1 and structured token commitment in block 2. The embeddings support retrieval, action classification, and VLA pretraining in block 3.

3.1 ARCHITECTURE

Figure 3 shows the architecture. A learned tokenizer (Arnab et al., 2021) converts optical flow (Teed & Deng, 2020) into motion tokens, while frozen DINOv3 features (Siméoni et al., 2025) provide appearance cues through cross-attention. A three-layer transformer with rotary positional embeddings processes these inputs using one `CLS` and three `REG` tokens, all 256-dimensional. The attention mask in Section 3.3 encourages action and scene information to separate. At inference, `CLS` serves as the atomic action embedding, while `REG` is discarded. Longer keysteps are represented by averaging their atomic embeddings, without further training. Appendix A.3 provides implementation details, while Appendix D compares the inputs.

3.2 UNIFIED SCENE-AWARE CONTRASTIVE LOSS

Our loss is a systematic extension of InfoNCE (Oord et al., 2018). It pulls together matching actions across scenes and separates different actions within shared scenes, discouraging scene bias. Sentence and verb labels supply additional semantic supervision, as described in Appendix B.

Let a_i, s_i, and v_i denote clip i's action label, scene ID, and CLS embedding. We use Ego-Exo4D take IDs as scene groups, as described in Appendix A.1; however, it can be extended with other datasets with their scene labels. For other batch clips j, we define three sets. Positives $\mathcal{P}(i)$ have the same action, different scene, so $a_j = a_i$ and $s_j \neq s_i$. Negatives $\mathcal{R}(i)$ have different actions, same scene, so $a_j \neq a_i$ and $s_j = s_i$. Negatives $\mathcal{N}(i)$ have different actions, different scenes, so $a_j \neq a_i$ and $s_j \neq s_i$.

$$\mathcal{L}_{\text{scene}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{p\in\mathcal{P}(i)}\exp\langle v_i, v_p\rangle}{\sum_{p\in\mathcal{P}(i)}\exp\langle v_i, v_p\rangle + \sum_{r\in\mathcal{R}(i)}\exp\langle v_i, v_r\rangle + \sum_{n\in\mathcal{N}(i)}\exp\langle v_i, v_n\rangle}. \quad (1)$$

The numerator sums positive-pair scores, while the denominator also includes negatives. We average over N anchors with at least one positive and skip the others in this term. Empty negative sets contribute zero. Without scene labels, action labels alone define the pairs, giving multi-positive InfoNCE (Khosla et al., 2020). With one positive per anchor, the objective reduces to ordinary InfoNCE (Oord et al., 2018). Appendix B.1 further details the loss.

Viewpoint Rosetta (Luo et al., 2025) omits same-scene, different-action negatives, allowing different actions in the same environment to remain similar. This benefits keystep retrieval but hurts atomic retrieval (Fig. 2). InvAct separates these pairs, reducing scene shortcuts while preserving action similarity across scenes. Tables 1 and 13 show improved atomic retrieval and transfer to keysteps.

3.3 STRUCTURED TOKEN COMMITMENT

Action cues are often spatially localized, sometimes to a small region such as a hand, while background cues dominate much of the visual content. This is particularly evident for non-contact actions in Fig. 6, such as jumping and walking apart, where localized motion can distinguish the action. Standard soft attention, however, can mix these action and background cues, encouraging scene-based shortcuts (Fig. 2). Our **hard token re-routing** instead encourages their separation.

Mask design. We achieve this via a mask M (block (2) in Fig. 3) applied to the self-attention logits as $(\mathbf{QK}^\top + \mathrm{M})$ before softmax, where $\mathbf{Q}$ and $\mathbf{K}$ are query and key matrices, respectively. Let $\mathcal{S}$ and $\mathcal{X}$ denote the sets of special [CLS,REG] and video token indices, respectively. The mask enforces two routing rules. First, CLS and REG are blocked from attending to each other, as defined in Eq. (2).

$$\mathrm{M}_{u,v} = \begin{cases} 0 & u = v \\ -\infty & u \neq v \end{cases} \quad u, v \in \mathcal{S}. \quad (2)$$

Second, each video token reads from exactly one special token. We compute the mean attention score $\bar{A}$ from video token u to each special token v across heads, and assign the mask in Eq. (3).

$$k_u^\star = \arg\max_{v\in\mathcal{S}} \bar{A}_{u,v}, \qquad \mathrm{M}_{u,v} = \begin{cases} 0 & v = k_u^\star \\ -\infty & v \neq k_u^\star \end{cases} \quad u \in \mathcal{X},\ v \in \mathcal{S}. \quad (3)$$

The mask is parameter-free. With Eq. (3), each video token must pick a single read source from [CLS,REG] at every layer. Without this restriction, non-action tokens pull in a fraction of CLS's action-aligned content each layer and start to look action-relevant even when they carry no real signal. Forcing one read source blocks this contamination. Since the action losses in Section 3.2 only shape CLS, its limited softmax bandwidth concentrates on genuinely action-discriminative tokens, while the rest receive vanishing attention and their argmax settles on REG. REGs thus act as sinks for non-action content and are discarded at inference. The diagonal mask in Eq. (2) further prevents CLS from reading REG directly, closing another leak.

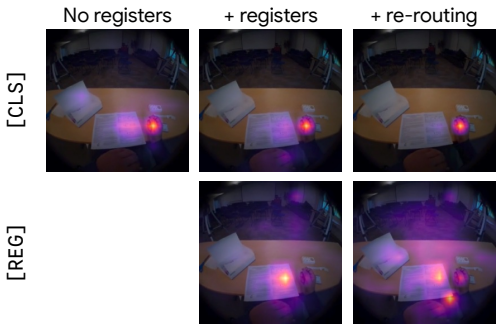

Figure 4: **Token routing focuses attention on the action.** Adding `REG` tokens localizes `CLS` attention. Routing further separates action and background cues.

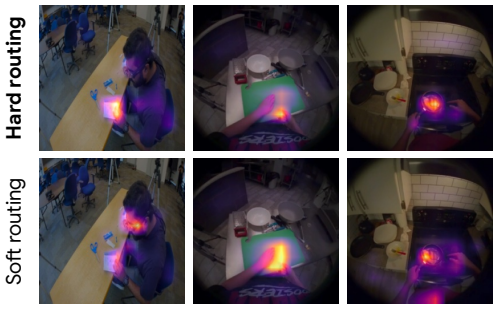

Figure 5: **Hard routing focuses attention on interactions.** Hard masking reduces non-action content in `CLS` compared with soft routing.

Fig. 4 and Table 13 (attention guiding rows) reveal the benefit of our routing. Without the mask, a non-action video token would repeatedly absorb `CLS` content and then be read back by `CLS`, giving `CLS` an indirect path to attend to itself and inflate non-action features into apparent action signal (Fig. 4, first column). Hard-routing such tokens to `REG` closes this path, so `CLS`'s effective receptive field is narrowed to video tokens that already carry genuine action signal (Fig. 4, third column).

We also perform ablations on masking. Column-wise and soft variants perform worse in our ablations. Column-wise masking blocks direct `CLS` access to video tokens assigned elsewhere, which may make routing errors harder to correct. Soft masking allows mixing across special tokens and yields less focused attention in Fig. 5. Appendix D.3 reports the results.

4 Experiments

We first test atomic retrieval and scene dependence on Ego-Exo4D, then examine representation usefulness for longer actions, unseen datasets, and robot-policy pretraining.

4.1 Setup

Label definitions. We train with atomic sentences such as *cutting a tomato* and verbs such as *cut*. For keystep retrieval, labels describe longer procedures such as *check the tube for damage*.

Datasets and training. We use time-synchronized Ego-Exo4D (Grauman et al., 2024) clips with keystep and atomic labels. Unlike cross-view works reporting only on `cooking` (Luo et al., 2025; Wang et al., 2023), we evaluate a processed validation subset spanning `cooking, health, bike` with shared atomic labels. Zero-shot evaluation uses Charades-Ego (Sigurdsson et al., 2018), SSV2 (Goyal et al., 2017), PhyWorld (Kang et al., 2025), and NTU RGB+D (Liu et al., 2020; Shahroudy et al., 2016). Details are given in Appendix A.1.

Baselines. Table 1 compares checkpoints with differing pretraining data, supervision, architectures, and modalities (Appendix A.5). Component attribution relies on the controlled ablations in Table 13.



Figure 6: **InvAct retrieves unseen actions on NTU RGB+D.** The matches share the query action despite changes in environment and viewpoint.

Table 1: **InvAct improves atomic action retrieval across viewpoints and scenes, and transfers to keystep actions with the same checkpoint.** We report video-to-video retrieval hit@10 in percent on the Ego-Exo4D validation set for both atomic and keystep actions. The cross-scene and cross-view columns cover all ego-exo retrieval directions, where g and x denote ego and exo views. For keystep retrieval, InvAct uses its atomic-action checkpoint without finetuning. The best three results are shown in **bold**, underlined, and *italic*, respectively. SV denotes single-view models using ego clips, and MV denotes multi-view models using ego-exo clips.

		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
Domain agnostic	CLIP (Radford et al., 2021)	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 (Siméoni et al., 2025)	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 (Assran et al., 2025)	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 (Araslanov et al., 2025)	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 (Bagad & Zisserman, 2025a)	27.84	23.15	25.50	21.71	*27.08*	*24.39*	25.88	19.02	22.45	20.79	16.23	18.51
Robot centric	LAPA (Ye et al., 2025)	25.24	26.52	25.88	15.96	13.82	14.89	13.58	15.28	14.43	15.09	13.94	14.51
	VILA (Jeong et al., 2026)	19.21	19.51	19.36	*22.59*	29.51	26.05	33.03	29.06	31.04	12.36	7.22	9.79
Human centric (SV)	TimeSFormer (Grauman et al., 2022)	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 (Zhao et al., 2023)	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 (Lin et al., 2022)	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 (Pramanick et al., 2023)	29.35	22.00	25.67	21.88	23.54	23.59	48.11	23.84	35.98	27.55	22.43	24.99
	SUM-L (Wang et al., 2023)	*29.58*	23.33	26.46	22.29	23.27	22.78	36.04	14.73	25.39	15.71	9.81	12.76
Human centric (MV)	AE2 (Xue & Grauman, 2023)	14.88	14.06	14.47	15.47	15.39	15.43	29.35	33.13	31.24	11.22	11.41	11.32
	BYOV (Park et al., 2025)	16.84	15.94	16.39	14.57	14.33	14.45	41.06	37.49	39.27	20.47	21.06	20.75
	VI Encoder (Grauman et al., 2024)	19.95	18.64	19.30	19.01	20.40	19.71	19.71	16.17	17.94	16.63	17.22	16.93
	Rosetta (Luo et al., 2025)	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	**34.03**	**31.03**	**32.53**	**31.55**	**32.02**	**31.79**	*44.25*	**38.70**	**41.47**	37.98	32.21	35.09



Figure 7: **We visually compare cross-scene atomic retrieval on Ego-Exo4D**. The first row shows queries, and later rows show each method's top-three matches. Green and red mark correct and incorrect matches.

4.2 DOWNSTREAMING THE REPRESENTATION

Atomic action retrieval. Following prior work (Oquab et al., 2024; Siméoni et al., 2025; Chen et al., 2025; Assran et al., 2025), we measure verb-level retrieval across scenes and views (Table 1 and Fig. 7).

InvAct leads all four atomic retrieval directions, averaging 32.53% across scenes and 31.79% across views. The comparisons span view-invariant, single-view, robot-centric, and general-purpose representations. In Fig. 7, baselines often retrieve similar objects or backgrounds with different actions, while InvAct matches the action across environments.

When retrieval must match both the verb and the object, InvAct achieves 10.03% across scenes and 9.90% across views. The best corresponding baselines score 8.46% and 7.28% (Section C.1).

Keystep action retrieval. We next test whether emphasizing atomic action similarity preserves information useful for longer activities. Without keystep training, we combine the atomic embeddings within each keystep to retrieve longer procedures. InvAct achieves the highest cross-scene average of 41.47% and ranks second across views at 35.09% (Table 1). Viewpoint Rosetta, trained with keystep labels on Ego-Exo4D, remains stronger across views, whereas InvAct reuses its atomic-action checkpoint. Thus, reducing scene dependence at the atomic level remains compatible with strong retrieval at a longer action timescale.

Table 2: **InvAct improves retrieval on SSV2 and NTU.** We report hit@10 in percent. Columns marked * exclude training verbs from queries and gallery, as detailed in Appendix A.2.

	SSV2	*	NTU	*
VI Encoder	23.68	44.83	36.61	25.99
SUM-L	24.92	53.71	52.70	65.76
Rosetta	24.96	48.91	58.79	73.86
InvAct	**29.00**	**55.75**	**60.69**	**74.23**

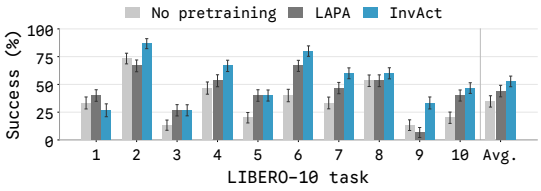

Figure 8: **InvAct latent targets improve mean LIBERO-10 success.** Error bars show reported evaluation variability (Table 16).

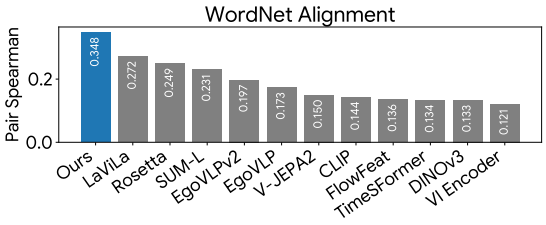

Figure 9: **InvAct best preserves semantic relationships between actions.** Spearman correlation compares pairwise similarities of learned action embeddings with WordNet similarities across 75 verbs. Higher correlation means better alignment.

Unseen-dataset retrieval. We test whether the learned atomic representation remains useful outside Ego-Exo4D. Following prior work (Zhao et al., 2023; Luo et al., 2025; Pramanick et al., 2023), we evaluate the same checkpoint without finetuning on SSV2 and NTU RGB+D. Their short action sequences align with our atomic-action formulation, since matching the action need not depend on a shared scene. InvAct leads on both datasets (Table 2 and Fig. 6). Charades-Ego and PhyWorld results are also presented in Appendices C.4 and F.2, respectively.

Latent action pretraining for robot manipulation. Following LAPA (Ye et al., 2025), we pretrain OpenVLA (Kim et al., 2024) on SSV2 with LAPA or InvAct latents, then post-train on LIBERO (Liu et al., 2023) and evaluate on LIBERO-10. InvAct improves mean success to 52.7%, compared with 44.0% for LAPA and 34.7% without latent pretraining (Fig. 8). LAPA and InvAct use the same SSV2 pretraining data and number of steps, isolating the choice of latent representation in this stage. InvAct uses its Ego-Exo4D checkpoint to supply these latents. Appendix E provides detailed information.

4.3 SCENE DEPENDENCE AND ABLATIONS

Scene bias. Although cross-scene and cross-view retrieval of atomic actions provides evidence about scene bias, we additionally measure it directly by testing whether improved action retrieval coincides with reduced reliance on scene cues. SSF@k measures the fraction of queries whose top-k results contain a clip from the query's scene, *independent of action correctness*; lower values therefore indicate less scene-dependent retrieval. As shown in Table 3, InvAct reduces top-1 SSF to 0.52,

Table 3: **InvAct reduces same-scene retrieval.** Lower SSF indicates less scene dependence.

Method	T1	T5	T10	T25
Data prior	0.39	0.89	0.90	0.95
DINOv3	0.99	0.99	1.00	1.00
LAPA	0.60	0.86	0.99	1.00
VI Encoder	0.99	0.99	1.00	1.00
SUM-L	0.94	0.98	0.99	0.99
Rosetta	0.99	0.99	1.00	1.00
InvAct	**0.52**	**0.86**	**0.92**	**0.93**

Table 4: **Scene repulsion complements action attraction, and structured token commitment further strengthens it.** We report Ego-Exo4D hit@10 (%).

	Cross-scene (↑)			Cross-view (↑)		
	g→g	x→x	Avg.	g→x	x→g	Avg.
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
+ Tok. commitment	**30.55**	**27.15**	**28.85**	**27.19**	**27.89**	**27.54**

compared with 0.60 for LAPA and 0.99 for Rosetta. At top-5 and top-25, InvAct also falls below the random-gallery data prior, while remaining slightly above it at top-10. Thus, InvAct reduces, rather than eliminates, same-scene preference. Appendix C.2 provides further details.

Semantic alignment. We test whether the embedding space captures relationships between action meanings beyond matching identical verbs. We average embeddings for 75 verbs and compare pairwise cosine similarities with WordNet (Miller, 1995) using Spearman correlation. InvAct achieves the highest correlation at 0.348, compared with 0.272 for LaViLa (Fig. 9), indicating stronger alignment with verb semantics. Appendix C.3 provides details.

Component ablations. We next examine our proposed components. The Viewpoint Rosetta objective (Luo et al., 2025), used as our starting point, does not explicitly repel different actions from the same scene. As shown in Table 4, adding cross-scene action attraction alone improves the cross-scene average but reduces cross-view retrieval for atomic actions. Adding same-scene action repulsion raises the cross-scene average from 26.93% to 28.20% and the cross-view average from 25.45% to 27.14%. Structured token commitment further improves these averages to 28.85% and 27.54%, respectively. Since the Ego-Exo4D validation set is quite large (around 9,000 atomic clips), these gains correspond to nearly **a thousand of additional successful retrievals** across four viewpoint combinations, and both components significantly improve the measured averages ($p < 0.03$). Appendix D.1 reports the full statistical analysis.

Robustness to input perturbations. Table 5 stress-tests our structured token commitment under noise and camera shake added to the input signal. Commitment reduces degradation in three of four comparisons. It cuts the cross-view drop under noise from 1.50 to 0.26 percentage points, and the cross-scene drop under camera shake from 2.41 to 0.62. These results support improved robustness beyond clean-input gains. Appendix D.4 provides further details.

Table 5: **Structured token commitment (STC) reduces degradation in most comparisons.** We report clean hit@10 (%) and changes in percentage points.

	Cross-scene		Cross-view	
Input	STC	No STC	STC	No STC
Clean	32.53	31.47	31.79	31.05
Gaussian noise	-0.73	-1.13	-0.26	-1.50
Camera shake	-0.62	-2.41	-1.74	-1.06

Further analyses. Appendix D.2 compares objectives, inputs, registers, and longer training. Appendix D.3 compares routing variants. Appendix F.2 provides non-contact examples, including jumping and walking apart. Appendix A.4 reports latency and memory costs.

5 Conclusion

We identify scene bias as an overlooked challenge in view-invariant action representation learning: aligning viewpoints alone can still leave representations reliant on shared backgrounds rather than action dynamics. To address this, we introduce **InvAct**, combining scene-aware contrastive learning with structured token commitment to separate action cues from scene context. InvAct improves retrieval on Ego-Exo4D and transfers well to zero-shot retrieval on SSV2 and NTU RGB+D. Its atomic embeddings further support keystep retrieval without additional training, and improve robotic manipulation benchmark LIBERO-10 success through latent pretraining compared to robot-centric latent action representations. Together, these results highlight the importance of reducing scene bias for learning action representations that generalize across scenes, viewpoints, and downstream tasks.

REFERENCES

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

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

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

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 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.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2026.

Xizhou Bu, Jiexi Lyu, Fulei Sun, Ruichen Yang, Zhiqiang Ma, and Wei Li. LAOF: Robust latent action learning with optical flow constraints. *arXiv preprint arXiv:2511.16407*, 2025.

Joao Carreira and Andrew Zisserman. 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.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, 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.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, 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.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, 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.

Joëlle Hanna and Damian Borth. Know your attention maps: Class-specific token masking for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23763–23772, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Youngjoon Jeong, Junha Chun, and Taesup Kim. Learning to act robustly with view-invariant latent actions. *arXiv preprint arXiv:2601.02994*, 2026.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Forty-second International Conference on Machine Learning*, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation*, pp. 13226–13233. IEEE, 2025.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 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.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. 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.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 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.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy V Vo. Revisiting [cls] and patch token interaction in vision transformers. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 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.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 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.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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

Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. 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.

Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *9th Annual Conference on Robot Learning*, 2025.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *7th Annual Conference on Robot Learning*, pp. 3536–3555, 2023.

Zihui Xue, Kristen Grauman, Dima Damen, Andrew Zisserman, and Tengda Han. Seeing without pixels: Perception from camera trajectories. *arXiv preprint arXiv:2511.21681*, 2025.

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

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Technical Appendices for InvAct

We first describe the experimental setup and training objectives. We then provide retrieval analyses, ablations, robot experiments, and qualitative examples that support the main paper.

A **Experimental setup** 15
- A.1 Training data and label definitions 15
- A.2 Zero-shot datasets and unseen-verb evaluation 15
- A.3 Architecture, optimization, and batch sampling 16
- A.4 Inference cost 16
- A.5 Baseline configurations 16

B **Training objectives** 18
- B.1 Scene-aware contrastive learning 18
- B.2 View alignment and semantic supervision 18
- B.3 Attraction and repulsion ablations 19

C **Retrieval and representation analysis** 20
- C.1 Object-sensitive atomic retrieval 20
- C.2 Measuring scene bias 20
- C.3 Semantic alignment with WordNet 21
- C.4 Scene-dependent datasets 21

D **Ablations and robustness** 23
- D.1 Controlled three-seed comparison 23
- D.2 Training objectives, inputs, and register tokens 23
- D.3 Alternative attention routing 23
- D.4 Robustness to noise and camera shake 24

E **Robot policy pretraining** 25
- E.1 Latent pretraining and robot adaptation 25
- E.2 Per-task success rates 25
- E.3 Task goals and rollout examples 26

F **Qualitative results** 27
- F.1 Cross-scene and cross-view retrieval 27
- F.2 Zero-shot retrieval 29
- F.3 Action predictions over time 30
- F.4 Attention maps and embedding structure 30

G **Limitations and future work** 31

H **AI disclosure** 32

A EXPERIMENTAL SETUP

This section provides the data, model settings, and evaluation protocols used in Section 4. Robot policy training is described separately in Section E.

A.1 TRAINING DATA AND LABEL DEFINITIONS

Ego-Exo4D. For the experiments in Section 4.1, 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 processed Ego-Exo4D validation subset spanning {`cooking`, `health`, `bike`}, using their set of shared atomic-action labels. We keep clips of 1–5 s and extract verbs from their free-form atomic annotations with ChatGPT 5.5, followed by human verification. We discard malformed strings (e.g., `adjsust`, `stirr`, `uncrew`), ambiguous extractions, and non-verbs rather than correcting or reassigning them. This removes 82/30,697 training clips (0.27%) and 30/8,991 validation clips (0.33%). 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`.

Scene and action identities. We use the take/sub-class identifiers provided by Ego-Exo4D as scene labels in Section 3.2, (`uniandes_cooking_009_6`, `sfu_covid_009_14`, etc.) from categories such as `cooking`, `bike`, and `health`. Each identifier defines an operational scene group, without assuming that different identifiers correspond to visually distinct environments. Action identities are given by the verified verb labels.

We will release the parsed labels alongside the code.

A.2 ZERO-SHOT DATASETS AND UNSEEN-VERB EVALUATION

PhyWorld (zero-shot). PhyWorld dataset (Kang et al., 2025) comes with several synthetic subsets (e.g., `collision`, `trajectory`, etc.), and we use `collision` to compare InvAct with Viewpoint Rosetta in Fig. 14. 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). These examples are consistent with matching motion rather than appearance. They illustrate the usefulness of the learned representation in another setting, without measuring whether reduced scene dependence transfers.

Other zero-shot datasets. The evaluations in Tables 2 and 11 cover 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.

Unseen-verb evaluation. Unstarred results in Table 2 test unseen datasets, which can still share verbs with Ego-Exo4D. For each starred subset, we remove samples whose verbs occur in the training vocabulary from both queries and gallery, then recompute hit@10. Table 6 records the filtered evaluation sizes. Because filtering changes both label composition and gallery size, starred and full-set scores are not directly comparable measures of difficulty. The NTU subset includes whole-body actions such as sitting down, jumping, kicking, saluting, sneezing, and hugging, beyond the predominantly hand-object training vocabulary.

We cite all assets and respect their licenses. Ego-Exo4D uses a custom signed license that allows some commercial/product-development use. SSV2, Charades-Ego, and NTU RGB+D are research-only/non-commercial unless separately permitted. PhyWorld uses CC-BY-4.0, requiring attribution.

Table 6: **Starred subsets exclude training verbs from both query and gallery.** We report removed classes, shared verbs, and remaining clips for the unseen-verb evaluation.

Dataset	Classes removed	Shared verbs	Clips remaining
SSV2	121/174	38	687
Charades-Ego	111/157	16	3439
NTU RGB+D	28/60	22	30336

A.3 Architecture, optimization, and batch sampling

We train from random initialization for 50,000 steps on four H200 GPUs. We use AdamW with a constant learning rate of 10^{-4} and `bf16` mixed precision.

Each minibatch contains four synchronized ego-exo pairs, giving eight videos. We sample two actions `A` and three scenes `S` with the layout `[S1A1, S2A1, S3A1, S3A2]`. This provides matching actions across scenes and different actions within one scene. We resample the actions and scenes at each iteration. Anchors without cross-scene positives, such as `S3A2`, are omitted from the scene loss but remain available to the other losses.

The transformer in Section 3.1 and Fig. 3 has three layers and hidden dimension 256, with one learnable `CLS` and three learnable `REG` tokens. Each clip supplies eight 256×256 frames. Flow and special tokens form cross-attention queries. Frozen DINOv3 frame features supply keys and values in every layer. Rotary positional embeddings encode the spatio-temporal token positions. Hard re-routing is also applied in every layer.

A.4 Inference cost

We measure inference cost on one H200 using eight 256×256 frames per clip. The full InvAct pipeline takes 185 ms per clip with 39.5M parameters, compared with 12 ms/177.0M for Viewpoint Rosetta and 57 ms/308.6M for EgoVLPv2. Most of InvAct's latency comes from optical flow extraction with RAFT. Because components overlap during execution, the full-pipeline latency is lower than the sum of their individual timings. Peak memory is measured separately for each configuration.

Table 7: **Optical flow dominates InvAct inference time.** We report mean latency per clip, parameter count, and peak memory on an H200. Components overlap in the full pipeline.

Model or component	Time (ms)	Params (M)	Peak VRAM (GB)
RAFT	155	5.3	0.25
DINOv3	6	28.7	0.20
InvAct transformer	43	5.5	4.68
Full pipeline	185	39.5	4.80
RGB-only ablation	30	4.2	4.68
Viewpoint Rosetta	12	177.0	0.80
EgoVLPv2	57	308.6	2.41

For retrieval, gallery embeddings can be cached, so the full pipeline is evaluated only on new query clips. In the LIBERO-10 experiments in Fig. 8, InvAct embeddings are computed offline as latent pretraining targets; neither InvAct nor RAFT is used during policy evaluation in simulation. The RGB-only ablation reduces inference to 30 ms per clip, at the cost of lower retrieval accuracy (Table 14).

A.5 Baseline configurations

For the comparisons in Tables 1 and 2, we use each baseline's released pretrained checkpoint. Table 8 lists architectures, frame counts T, input resolutions, and output dimensions.

Table 8: **Baselines use their released pretrained configurations.** We list each backbone, frame count, input resolution, and embedding dimension.

Method	Backbone	T	Res.	Dim
InvAct	Transformer with RoPE + `RAFT` + `DINOv3`	8	256^2	256
Rosetta	`CLIP_OPENAI_TIMESFORMER_BASE` (Radford et al., 2021; Bertasius et al., 2021)	4	224^2	256
BYOV	`CLIP_OPENAI_TIMESFORMER_BASE` (Park et al., 2025; Radford et al., 2021)	32/20	224^2	256
AE2	`ResNet50 + 3D Conv + MLP` (Xue & Grauman, 2023)	32	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
VILA	`VILA stage-1 IDM (LAOM (Schmidt & Jiang, 2023)-style CNN)` (Jeong et al., 2026)	2	64^2	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

B TRAINING OBJECTIVES

Let g_i and x_i denote the ego and exo embeddings of paired clip i. We expand the losses in Section 3.2, using action and scene labels to form the contrastive sets in Eq. (1).

B.1 SCENE-AWARE CONTRASTIVE LEARNING

Our scene-aware objective extends multi-positive InfoNCE (Oord et al., 2018) with action and scene labels to encourage scene-invariant action representations. Let a_i and s_i denote the action and scene labels of clip i. For each anchor i, we define

$$\mathcal{P}(i) = \{j \mid a_j = a_i,\ s_j \neq s_i\}, \tag{4}$$

$$\mathcal{R}(i) = \{j \mid a_j \neq a_i,\ s_j = s_i\}, \tag{5}$$

$$\mathcal{N}(i) = \{j \mid a_j \neq a_i,\ s_j \neq s_i\}. \tag{6}$$

Here, $\mathcal{P}(i)$ contains the same action performed in different scenes, encouraging action similarity across scene changes. Conversely, $\mathcal{R}(i)$ contains different actions from the same scene, explicitly discouraging similarity based on shared scene appearance. Finally, $\mathcal{N}(i)$ contains different actions from different scenes and serves as the remaining negative set. These sets are disjoint, and same-action, same-scene clips are excluded from this contrastive term.

Let $q_i(j)$ denote the softmax probability assigned to clip j over $\mathcal{A}(i) = \mathcal{P}(i) \cup \mathcal{R}(i) \cup \mathcal{N}(i)$. We minimize the negative log probability mass assigned to the positive set:

$$\ell_i = -\log \sum_{p \in \mathcal{P}(i)} q_i(p). \tag{7}$$

Unlike objectives that average losses over individual positives, this formulation encourages the positive set to receive high probability mass without requiring every positive to be equally similar. Anchors without a valid positive retain their other training losses but are excluded from this term; if a batch contains no valid anchors, the term is zero.

Relation to InfoNCE. The key extension over multi-positive InfoNCE is therefore the scene-aware construction of contrastive relationships. When scene labels are unavailable, the scene constraints can be removed, yielding $\mathcal{P}(i) = \{j \mid a_j = a_i\}$ and a single negative set containing clips with different actions. Equation 7 then reduces to action-supervised multi-positive InfoNCE, and further to ordinary InfoNCE when $|\mathcal{P}(i)| = 1$.

B.2 VIEW ALIGNMENT AND SEMANTIC SUPERVISION

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

$$\mathcal{L}_{\text{vv}} = \frac{1}{N} \sum_{i=1}^{N} \Big(1 - \langle g_i, x_i \rangle\Big) \tag{8}$$

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.

$$\mathcal{L}_{\text{text}}^{\text{ego}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp\langle g_i, t_i^{\text{atom}} \rangle}{\sum_{j=1}^{N} \exp\langle g_i, t_j^{\text{atom}} \rangle} + \log \frac{\exp\langle g_i, t_i^{\text{atom}} \rangle}{\sum_{j=1}^{N} \exp\langle g_j, t_i^{\text{atom}} \rangle} \right] \tag{9}$$

We define $\mathcal{L}_{\text{text}}^{\text{exo}}$ by replacing g with x. We then average the two losses to obtain $\mathcal{L}_{\text{text}} = \frac{1}{2}\big(\mathcal{L}_{\text{text}}^{\text{ego}} + \mathcal{L}_{\text{text}}^{\text{exo}}\big)$.

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

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{\exp \ell_{i,y_i}^g}{\sum_{c=1}^{K} \exp \ell_{i,c}^g} + \log \frac{\exp \ell_{i,y_i}^x}{\sum_{c=1}^{K} \exp \ell_{i,c}^x} \right] \tag{10}$$

The final loss sums all terms with weights $\lambda_{\text{vv}} = \lambda_{\text{CE}} = 0.5$ and 1.0 otherwise.

$$\mathcal{L} = \lambda_{\text{vv}}\mathcal{L}_{\text{vv}} + \lambda_{\text{scene}}\mathcal{L}_{\text{scene}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} \tag{11}$$

B.3 Attraction and repulsion ablations

Unified loss and component ablations. The final model uses $\mathcal{L}_{\text{scene}}$ in Eq. (1). The attraction-only and repulsion-only expressions below describe the component ablations in Table 14. The unified objective combines both negative sets in one denominator, so it differs from summing the two separate losses. The *Scene repel* row uses the unified objective.

Separated 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 $\mathcal{N}(i)$ with different action and scene labels. We then apply a multi-positive InfoNCE (Oord et al., 2018) objective for ego embeddings.

$$\mathcal{L}_{\text{att}}^{\text{ego}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\sum_{p\in\mathcal{P}(i)} \exp\langle g_i, g_p\rangle}{\sum_{k\in\mathcal{P}(i)\cup\mathcal{N}(i)} \exp\langle g_i, g_k\rangle} \tag{12}$$

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

Separated 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 ego-space similarity with a self-anchored contrastive term.

$$\mathcal{L}_{\text{rep}}^{\text{ego}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\langle g_i, g_i\rangle}{\exp\langle g_i, g_i\rangle + \sum_{r\in\mathcal{R}(i)} \exp\langle g_i, g_r\rangle} \tag{13}$$

The exo loss $\mathcal{L}_{\text{rep}}^{\text{exo}}$ is defined analogously by replacing g with x. We then average both views to obtain $\mathcal{L}_{\text{rep}} = \frac{1}{2}\big(\mathcal{L}_{\text{rep}}^{\text{ego}} + \mathcal{L}_{\text{rep}}^{\text{exo}}\big)$.

C RETRIEVAL AND REPRESENTATION ANALYSIS

These analyses extend Sections 4.2 and 4.3 by testing object-sensitive retrieval, scene dependence, and alignment with verb meanings.

C.1 OBJECT-SENSITIVE ATOMIC RETRIEVAL

Verb labels can group actions involving different objects. We therefore repeat the retrieval evaluation of Table 1, counting a match only when both the verb and object agree. Table 9 shows that InvAct leads the cross-scene and cross-view averages, while EgoVLPv2 performs best for ego-to-ego retrieval. Scores are lower because this matching criterion is stricter.

Table 9: **InvAct retains its advantage with object-sensitive labels.** We report verb+object hit@10 in percent on Ego-Exo4D. g and x denote ego and exo views. CS and CV are cross-scene and cross-view averages.

Method	g→g	x→x	CS	g→x	x→g	CV
InvAct	10.87	9.19	**10.03**	**10.81**	8.99	**9.90**
Rosetta	6.51	**9.26**	7.89	8.59	5.97	7.28
EgoVLPv2	**12.55**	4.36	8.46	3.42	**9.53**	6.48
DINOv3	7.65	3.49	5.57	5.17	3.76	4.46
SUM-L	6.58	4.56	5.57	1.88	6.98	4.43
FlowFeat	5.84	4.97	5.40	3.49	4.30	3.89
V-JEPA 2	8.99	2.48	5.74	2.28	2.89	2.58
AE2	5.10	2.62	3.86	2.82	2.89	2.58
BYOV	5.03	4.09	4.56	1.88	1.74	1.81
Random	3.11	3.06	3.09	3.56	3.50	3.53

C.2 MEASURING SCENE BIAS

We expand the scene-bias analysis in Section 4.3 and Table 3 using the same-scene fraction, SSF@k. It counts the fraction of queries with at least one clip from the same scene among their top-k results, regardless of whether the action matches. Lower values therefore indicate less preference for the query's scene.

To account for the gallery composition, we also compute the probability of this event under uniform sampling. For query q, let N_q be the gallery size and M_q the number of clips from its scene. The reference probability is

$$\text{Prior@}k(q) = 1 - \frac{\binom{N_q - M_q}{k}}{\binom{N_q}{k}}. \tag{14}$$

We average this probability over queries. Table 10 shows that InvAct has the lowest SSF at every k, tying LAPA at top-5. Its top-1 value is 0.52, compared with 0.60 for LAPA and 0.99 for Rosetta. InvAct still exceeds the prior at top-1 and top-10. Therefore, these results show reduced scene bias, rather than complete scene independence across every data distribution.

Table 10: **InvAct has the lowest same-scene fraction, tying LAPA at top-5.** Lower SSF@k indicates less scene dependence. The prior is the reference probability under uniform gallery sampling.

Model	Top-1	Top-5	Top-10	Top-25
Data prior	0.39	0.89	0.90	0.95
CLIP	0.99	1.00	1.00	1.00
DINOv3	0.99	0.99	1.00	1.00
V-JEPA 2	0.99	0.99	0.99	0.99
FlowFeat	0.99	1.00	1.00	1.00
LiFT	0.99	0.99	0.99	0.99
LAPA	0.60	0.86	0.99	1.00
TimeSFormer	0.99	0.99	0.99	1.00
LaViLa	0.86	0.96	0.97	0.99
EgoVLP	0.99	0.99	0.99	0.99
EgoVLPv2	0.99	0.99	0.99	0.99
VI Encoder	0.99	0.99	1.00	1.00
SUM-L	0.94	0.98	0.99	0.99
Rosetta	0.99	0.99	1.00	1.00
InvAct	**0.52**	**0.86**	**0.92**	**0.93**

C.3 Semantic alignment with WordNet

The analysis in Section 4.3 and Fig. 9 tests whether learned action similarities agree with relationships between verb meanings in WordNet (Miller, 1995). For example, related verbs should have more similar embeddings than unrelated verbs.

We use 75 verbs with at least 20 paired clips each. For each clip, we average its ego and exo embeddings and normalize the result. Then we average these features for each verb and normalize again, giving one representative embedding per verb. Cosine similarity between these embeddings gives the visual similarities. The corresponding WordNet score uses the closest pair of verb senses in its hierarchy.

We compare the rankings of all 2,775 distinct verb pairs using Spearman correlation, assigning average ranks to ties. Higher correlation means that the visual and lexical similarities order verb pairs more consistently. InvAct scores 0.348, compared with 0.272 for LaViLa and 0.249 for Viewpoint Rosetta. This supports better alignment with verb meanings.

C.4 Scene-dependent datasets

Charades-Ego. This dataset complements the SSV2 and NTU results in Table 2. It includes actions tied to objects and scene context, making those cues useful for distinguishing action labels. InvAct's emphasis on matching atomic actions across scenes can suppress some of this useful context. This offers an explanation for its slightly lower retrieval score than Viewpoint Rosetta, 16.45% versus 17.98% (Table 11). The comparison is consistent with this interpretation, without isolating scene dependence as its cause. Starred results exclude Ego-Exo4D training verbs from both queries and gallery, as described in Appendix A.2.

Table 11: **Rosetta leads on Charades-Ego.** We report hit@10 (%); * excludes training verbs.

Method	Charades-Ego	*
VI Encoder	15.60	39.34
SUM-L	13.88	36.21
Rosetta	**17.98**	**39.91**
InvAct	16.45	39.85

Table 12: **Static context improves Charades-Ego retrieval.** We report hit@10 (%). Rosetta is included for comparison.

Representation	Ego-Exo4D	Charades-Ego
Rosetta	21.54	17.98
InvAct	32.16	16.45
+ DINOv3	29.70	19.72
Only DINOv3	16.03	19.52

Adding frozen DINOv3 features raises InvAct's Charades-Ego retrieval from 16.45% to 19.72%, above Rosetta's 17.98%, but lowers Ego-Exo4D atomic retrieval from 32.16% to 29.70% (Table 12). These results highlight a tradeoff. Reducing reliance on shared scene cues helps cross-scene atomic

matching, while static information can remain useful for context-dependent action labels. PhyWorld examples are provided in Appendix F.2.

Figure 10 contrasts Charades-Ego examples involving objects and their context with the more atomic-aligned actions in SSV2.



Figure 10: **Unseen datasets emphasize different aspects of actions.** Putting a glass, reading, and washing a dish are Charades-Ego examples. The remaining examples are from SSV2 and more closely match atomic actions. We show the top-three retrievals.

D ABLATIONS AND ROBUSTNESS

We first isolate the contributions of scene repulsion and token commitment, then compare training choices, routing variants, and robustness to input perturbations. These experiments expand the ablations in Section 4.3.

D.1 CONTROLLED THREE-SEED COMPARISON

We compare three configurations under the same training budget, each with three random seeds. The base uses DINOv3 and flow with register tokens, but without commitment or same-scene repulsion. We then add repulsion, followed by commitment. Table 13 reports the mean and standard error for each retrieval direction and supports the component analysis in Section 4.3.

Repulsion improves all four directions by 1.15–2.31 percentage points, with $p < 0.01$ in two-sample t-tests. Commitment adds 0.55–0.91 points in three directions, with $p < 0.05$, and 0.15 points for ego-to-exo retrieval. Both additions improve the cross-scene and cross-view averages, with $p < 0.03$. Therefore, commitment provides a smaller but consistent gain across these runs.

Table 13: **Scene repulsion and token commitment improve retrieval.** We report hit@10 in percent as mean ± standard error over three seeds. Rows sequentially add components to the baseline. CS and CV are cross-scene and cross-view averages. g and x denote ego and exo views.

Configuration	g→g	x→x	CS	g→x	x→g	CV
Base	31.45±0.12	28.89±0.15	30.17±0.10	29.50±0.10	28.74±0.09	29.12±0.07
+ scene repulsion	32.60±0.18	30.53±0.09	31.57±0.10	31.04±0.04	31.05±0.21	31.05±0.11
+ token commitment	33.51±0.14	31.08±0.05	32.30±0.07	31.19±0.05	31.80±0.13	31.50±0.07

D.2 TRAINING OBJECTIVES, INPUTS, AND REGISTER TOKENS

Table 14 gives the full ablations under a shorter schedule. Its measured loss ablations also appear in Table 4. The other groups compare training paradigms, inputs, and routing. The final row doubles the training steps for the selected configuration. Darker blue cells indicate stronger results within each ablation group.

Training objectives. Verb classification and sentence-level contrastive learning work best together. Their combination reaches 27.46% on both retrieval averages, compared with 25.61% and 25.45% for sentence-level contrastive learning alone. Hence, coarse action labels and detailed captions provide complementary supervision.

Loss components. View alignment improves both averages over text alignment alone. Cross-scene action attraction then helps cross-scene retrieval but reduces cross-view retrieval. Adding same-scene repulsion recovers cross-view performance and gives the strongest cross-scene result in this group. This supports explicitly separating different actions that share a scene.

Input features. Combining RGB and flow improves both averages over either input alone. Frozen pretrained visual features provide further gains, with DINOv3 and flow performing best. These features supply semantic information through cross-attention while the flow stream captures motion.

Attention routing and longer training. Adding register tokens improves both retrieval averages. Hard routing raises them from 31.62% and 31.04% to 32.28% and 31.50%, alongside the more focused attention in Fig. 4. Longer training yields the final scores of 32.53% and 31.79%.

D.3 ALTERNATIVE ATTENTION ROUTING

We compare alternatives to the routing in Section 3.3, keeping the rest of the model fixed in Table 15. Darker blue cells indicate higher retrieval scores.

Row-wise and column-wise routing. Our row-wise mask lets each video token read from one special token, while `CLS` can still read every video token. Column-wise masking instead lets each video token be read by only one special token. Therefore, a token assigned to `REG` loses its direct connection to

Table 14: **The ablations compare supervision, inputs, and attention routing.** We report atomic retrieval on Ego-Exo4D. CE and CLR denote cross-entropy and contrastive learning. Darker cells indicate higher scores within each group.

	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
Attention guiding						
Only `CLS` token	32.52	30.28	31.40	31.01	30.15	30.58
+`REG` tokens	32.66	30.59	31.62	31.16	30.91	31.04
+Attn re-routing	33.55	31.01	32.28	31.23	31.76	31.50
Trained longer	**34.03**	**31.03**	**32.53**	**31.55**	**32.02**	**31.79**

Table 15: **Hard row-wise routing gives the strongest retrieval results.** We extend the attention-routing comparison in Table 14 on Ego-Exo4D.

	Cross-scene (↑)			Cross-view (↑)		
	g→g	x→x	avg.	g→x	x→g	avg.
Only `CLS` token = (A)	32.52	30.28	31.40	31.01	30.15	30.58
(A) + `REG` tokens = (B)	32.66	30.59	31.62	31.16	30.91	31.04
(B) + column-wise routing	24.74	23.54	24.14	22.44	23.69	23.06
(B) + soft row-wise routing	29.66	27.10	28.38	29.42	30.06	29.74
(B) + hard row-wise routing	33.55	31.01	32.28	31.23	31.76	31.50

`CLS`. This may make early routing mistakes harder to correct and help explain its weaker retrieval scores.

Soft and hard routing. Soft routing adds head-averaged attention logits as biases, allowing each video token to mix information from several special tokens. Hard routing keeps only the special token with the largest score. It performs better in all four retrieval directions and produces more focused attention maps, as shown in Fig. 5. We apply hard routing in all three layers with three register tokens and no warm-up schedule.

D.4 Robustness to noise and camera shake

We add Gaussian noise with $\sigma = 1$ or camera shake of up to ± 32 pixels to 256-pixel RGB frames, then recompute flow. Both input streams are therefore affected. Table 5 compares each model with its own clean-input scores.

Commitment reduces the drop under Gaussian noise and cross-scene camera shake. Cross-view camera shake is the exception, with a 1.74-point drop versus 1.06 without commitment. Its final score nevertheless remains slightly higher, at 30.05% versus 29.99%.

E ROBOT POLICY PRETRAINING

This exploratory evaluation tests the usefulness of InvAct embeddings as latent targets, rather than the transfer of scene invariance. For the results in Fig. 8, we follow the three-stage pipeline of LAPA (Ye et al., 2025), using OpenVLA (Kim et al., 2024) and LIBERO-10 (Liu et al., 2023).

E.1 LATENT PRETRAINING AND ROBOT ADAPTATION

Latent action learning. We compare LAPA's released `LAQ` encoder, trained on SSV2 (Goyal et al., 2017), with InvAct trained on Ego-Exo4D (Grauman et al., 2024). Their upstream data and supervision differ, so matching the downstream procedure does not isolate scene dependence as the source of performance differences.

Latent pretraining. We train OpenVLA on SSV2 frames and action descriptions to predict either encoder's latent actions through a linear head. Both variants use rank-32 LoRA adapters (Hu et al., 2022), batch size 32, and 25,000 steps on one H200. We use LAPA's continuous encoder outputs before quantization.

Robot post-training. We replace the latent head with a robot-action head and train fresh rank-32 adapters on LIBERO. All three variants, including the baseline without latent pretraining, use batch size 32 and 50,000 steps on one H200.

E.2 PER-TASK SUCCESS RATES

We test the ten tasks below with randomized object positions. Table 16 reports per-task success and evaluation variability, while Fig. 11 shows successful and failed runs. InvAct improves the mean over LAPA by 8.7 percentage points under matched pretraining data and steps. Its 18.0-point gain over no pretraining also includes the benefit of additional SSV2 training. The reported deviations describe evaluation variability rather than differences across independently trained models.

Table 16: **InvAct improves the mean success rate over both baselines.** We report LIBERO-10 success rates in percent with evaluation variability, following the task order below. These are evaluation deviations, not standard errors across VLA training seeds.

Task	No pretraining	LAPA	InvAct
Task 1	33.3 ± 5.4	40.0 ± 5.2	26.7 ± 5.8
Task 2	73.3 ± 4.8	66.7 ± 5.3	86.7 ± 4.5
Task 3	13.3 ± 4.6	26.7 ± 5.1	26.7 ± 5.0
Task 4	46.7 ± 5.5	53.3 ± 5.4	66.7 ± 5.1
Task 5	20.0 ± 4.7	40.0 ± 5.2	40.0 ± 5.0
Task 6	40.0 ± 5.6	66.7 ± 5.0	80.0 ± 4.7
Task 7	33.3 ± 5.3	46.7 ± 5.1	60.0 ± 4.9
Task 8	53.3 ± 5.2	53.3 ± 5.3	60.0 ± 5.0
Task 9	13.3 ± 4.8	6.7 ± 4.5	33.3 ± 5.4
Task 10	20.0 ± 5.1	40.0 ± 5.0	46.7 ± 4.8
Mean	34.7 ± 5.1	44.0 ± 5.2	52.7 ± 4.8

E.3 TASK GOALS AND ROLLOUT EXAMPLES

The LIBERO-10 task goals follow the order in Table 16.

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`

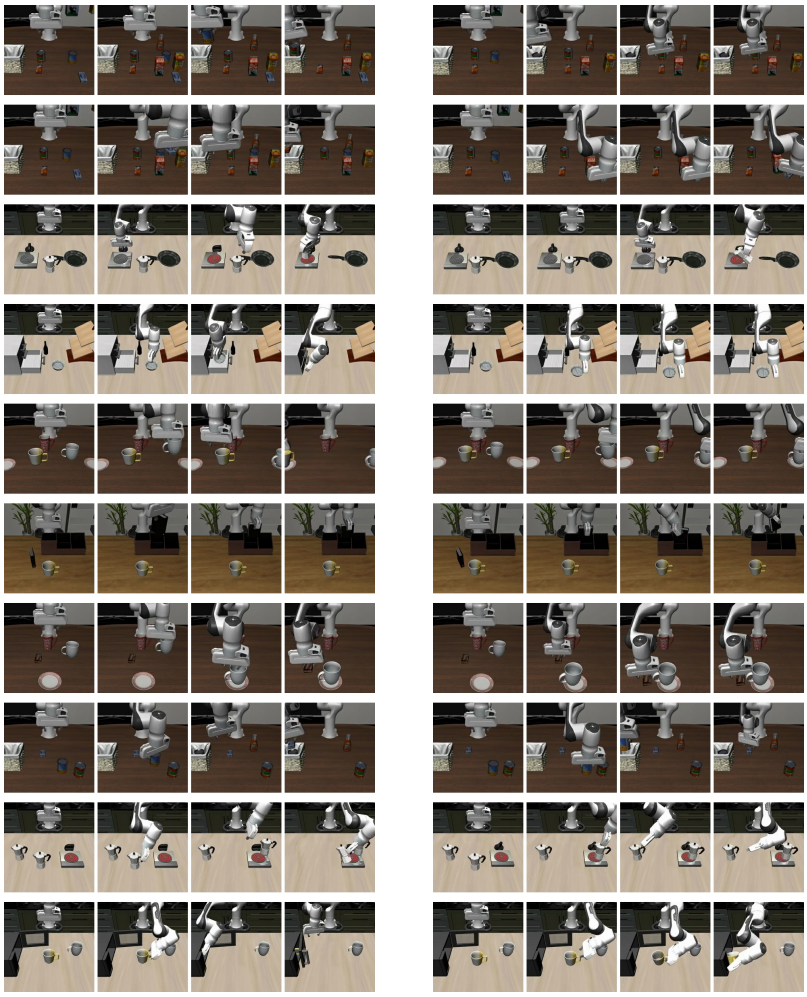

Figure 11: **LIBERO-10 rollouts include successful and failed attempts.** Successes appear on the left and failures on the right, with four frames per task.

F QUALITATIVE RESULTS

The following examples complement the retrieval and representation measurements in Sections 4.2 and 4.3. We group them by the behavior being inspected.

F.1 CROSS-SCENE AND CROSS-VIEW RETRIEVAL

The examples in Figs. 12 and 13 extend the Ego-Exo4D comparisons in Figs. 2 and 7. We group the retrieval directions by the query viewpoint, showing ego queries first and exo queries afterwards.



Figure 12: **InvAct retrieves actions across scenes from ego queries.** Ego-ego results appear on the left and ego-exo results on the right. Queries precede their top-three matches on Ego-Exo4D.

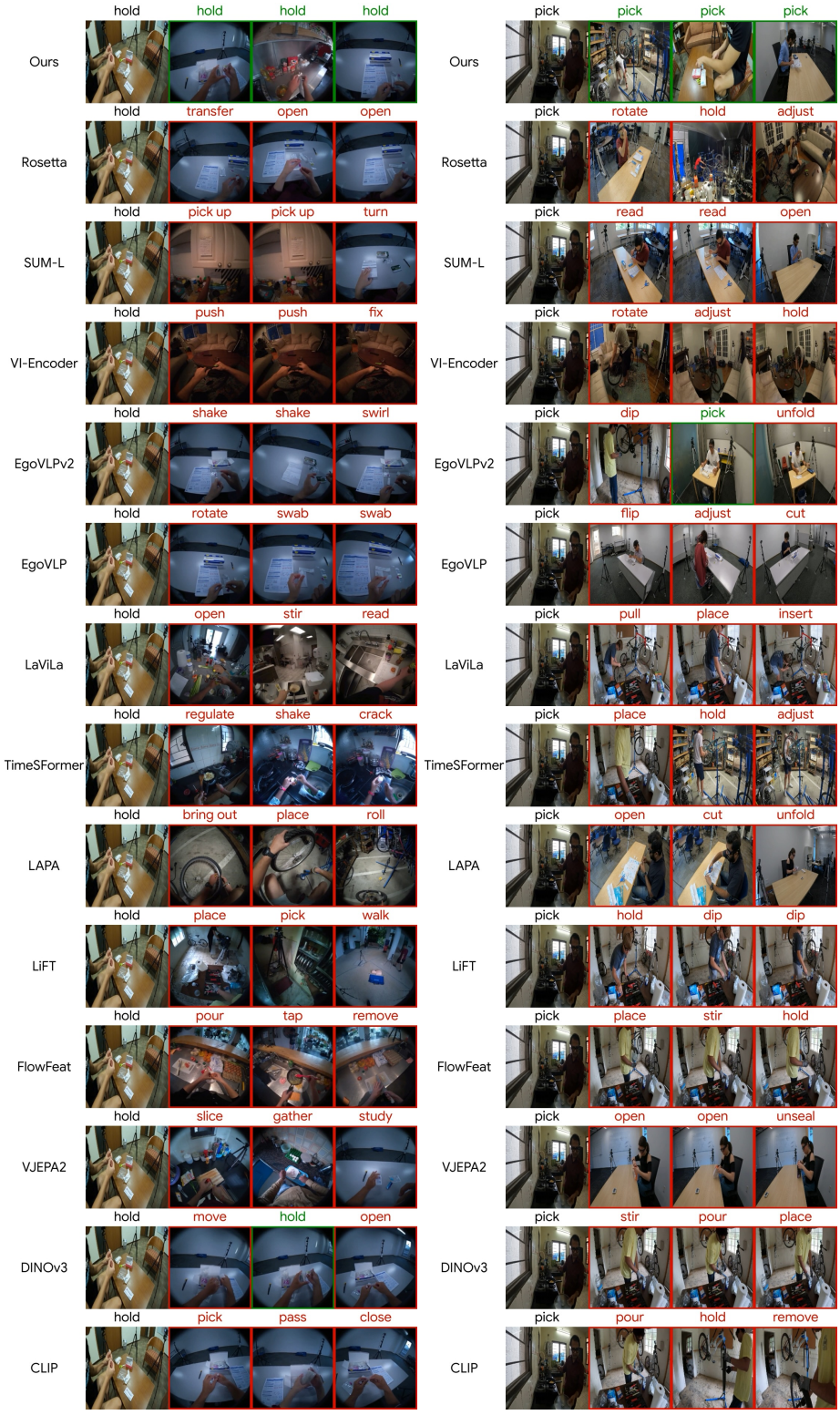

Figure 13: **InvAct retrieves actions across scenes from exo queries.** Exo-ego results appear on the left and exo-exo results on the right. Queries precede their top-three matches on Ego-Exo4D.

F.2 ZERO-SHOT RETRIEVAL

These examples complement the zero-shot evaluation in Section 4.2 and Table 2. We compare motion matching in PhyWorld and action retrieval on NTU RGB+D without additional training. The SSV2 and Charades-Ego examples appear in Section C.4 and Fig. 10.

The NTU examples in Fig. 6 show non-contact actions such as jumping and walking apart. Routing does not require contact or detected hands.

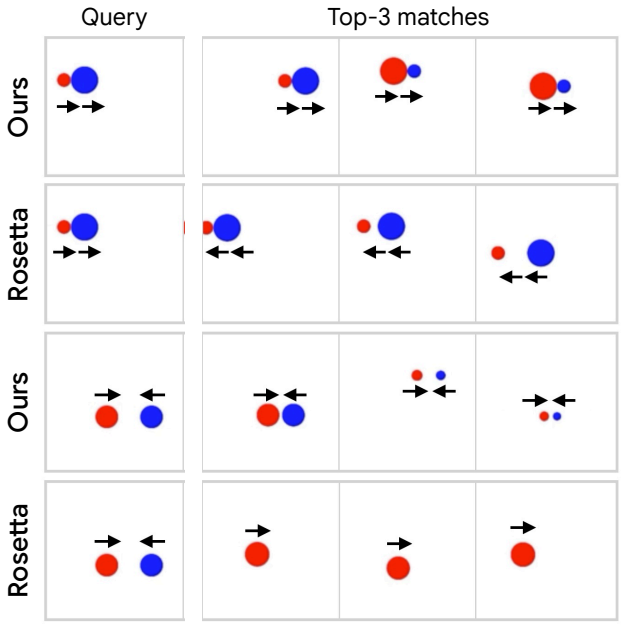

Figure 14: **InvAct retrieves matching motion in PhyWorld.** We compare zero-shot retrieval on synthetic collisions (Kang et al., 2025). Arrows show motion directions. InvAct retrieves fine-grained actions, while other methods match color and shape (row 2) or only part of the action (row 4).



Figure 15: **InvAct retrieves matching actions on NTU RGB+D without additional training.** These examples extend Fig. 6 across environments and viewpoints.

F.3 Action predictions over time

Figure 16 compares predictions over time and across synchronized views, extending the keystep evaluation in Section 4.2. The model preserves frame order within each clip but embeds clips independently. Averaging their embeddings for keystep retrieval discards their order.

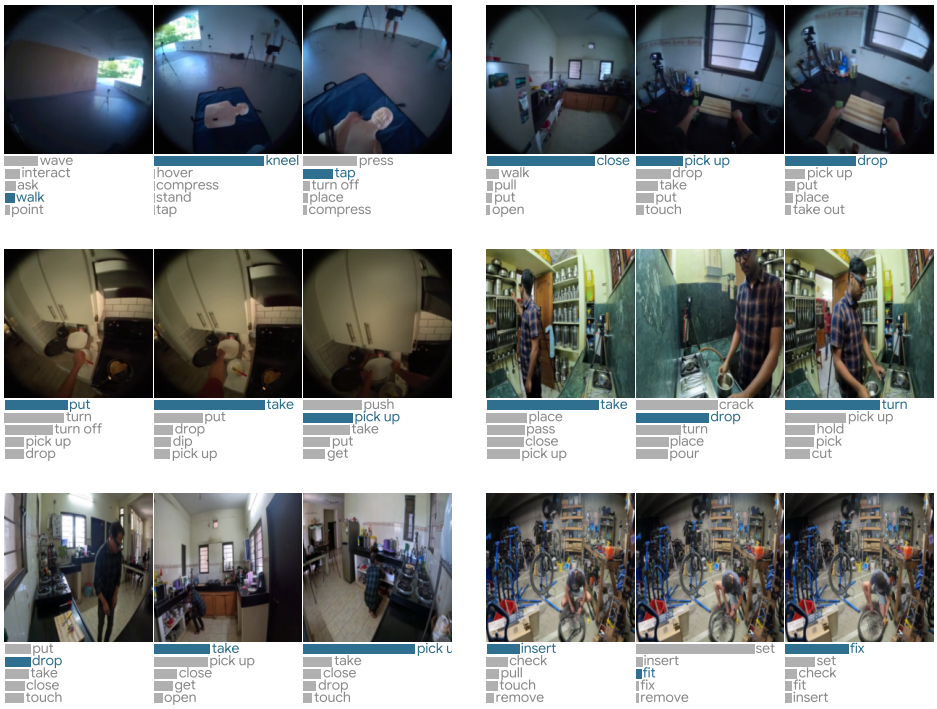

Figure 16: **Atomic-action predictions remain consistent across views.** We probe ego and exo videos from the Ego-Exo4D validation set. The predictions remain consistent across longer sequences without autoregressive decoding or context memory.

F.4 Attention maps and embedding structure

The attention maps in Fig. 17 show where CLS reads visual information. The embedding projections in Figs. 18 and 19 extend Fig. 1 by comparing action, scene, and viewpoint groupings.



Figure 17: **Cross-attention concentrates on interactions.** Video frames appear above their heatmaps. Brighter regions receive more CLS attention.

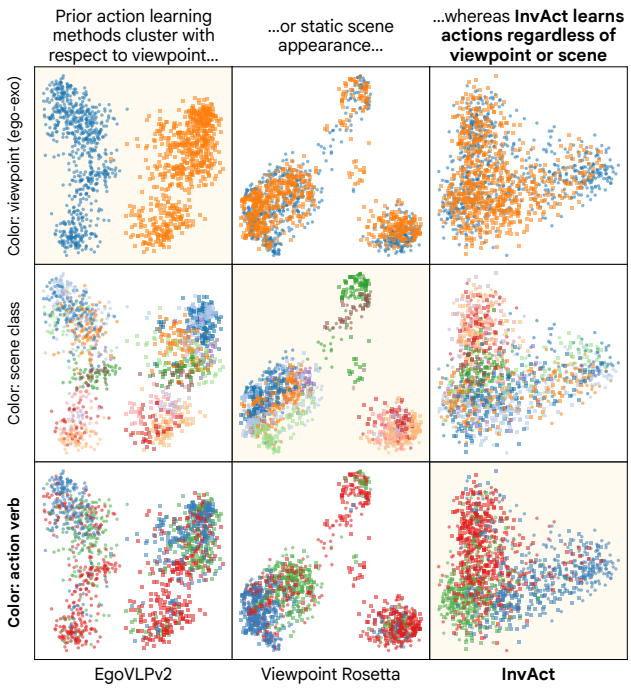

Figure 18: **InvAct groups actions across scenes and viewpoints.** Extending Fig. 1, PCA shows the embeddings colored by viewpoint, scene, and action in successive rows. EgoVLPv2 groups viewpoints, while Rosetta groups scenes. In the bottom row, green, red, and blue denote `pick`, `remove`, and `cut`.

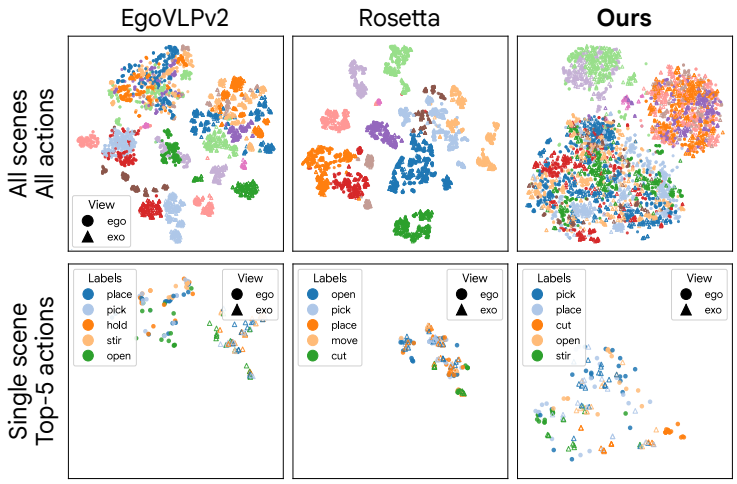

Figure 19: **InvAct preserves action groups with less scene dependence.** The top row shows t-SNE embeddings across all scenes and actions, colored by scene. The bottom row shows the five most frequent actions within one scene, colored by verb.

G Limitations and Future Work

For keystep retrieval in Section 4.2, long videos are divided into atomic segments and embedded independently. Averaging these embeddings ignores their order, so successful retrieval does not establish procedural-order understanding. There is also no enforced temporal consistency between consecutive embeddings from the same long clip. Despite this, action probing remains qualitatively consistent in practice but produces occasional incorrect hits (see Fig. 16). Autoregressive variants could improve temporal consistency.

Scaling to unlabeled data, further robotics deployments, and adapting token-level video-to-text decoding (Chen et al., 2025) are also promising and planned directions.

H AI DISCLOSURE

We used generative AI tools to assist with code, figure preparation, and the writing and layout of this submission. We also used ChatGPT to extract verb labels, followed by human verification as described in Section A.1. We reviewed the AI-assisted work and take responsibility for the final content, including the text, claims, and artifacts produced with these tools.