LARY: A Latent Action Representation Yielding Benchmark

Dujun Nie, Fengjiao Chen*, Qi Lv, Jun Kuang,
Xiaoyu Li, and Xuezhi Cao

Meituan, Beijing, China
https://meituan-longcat.github.io/LARYBench/

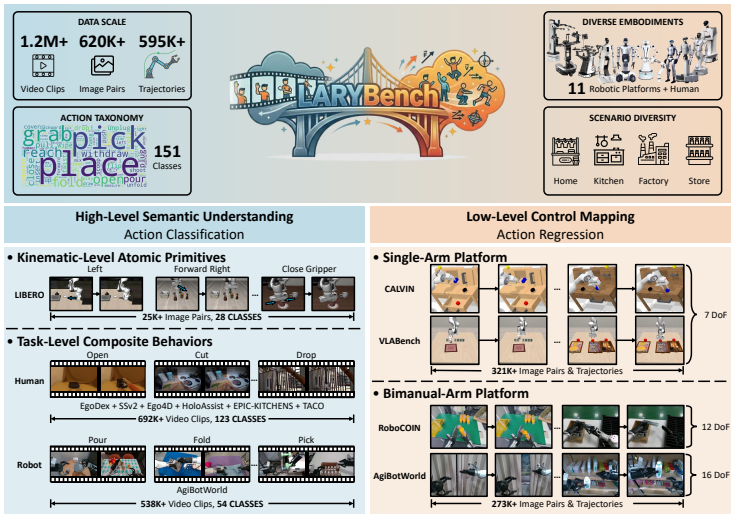

Fig. 1: LARYBench evaluates vision-to-action transformation on both action generalization and robotic control.

Abstract. While the shortage of explicit action data limits Vision-Language-Action (VLA) models, human action videos offer a scalable yet unlabeled data source. A critical challenge in utilizing large-scale human video datasets lies in transforming visual signals into ontology-independent representations, known as latent actions. However, the capacity of latent action representation to derive robust control from visual observations has yet to be rigorously evaluated. We introduce the Latent Action Representation Yielding (LARY) Benchmark, a unified framework for evaluating latent action representations on both high-level semantic actions (*what to do*) and low-level robotic control (*how to do*). The comprehensively curated dataset encompasses over one million videos (1,000 hours) spanning 151 action categories, alongside 620K image pairs and 595K motion trajectories across diverse embodiments and environments. Our experiments reveal two crucial insights: (i) General visual foundation models, trained without any action supervision, consistently outperform specialized embodied latent action models. (ii)

* Corresponding author (chenfengjiao02@meituan.com)

Latent-based visual space is fundamentally better aligned to physical action space than pixel-based space. These results suggest that general visual representations inherently encode action-relevant knowledge for physical control, and that semantic-level abstraction serves as a fundamentally more effective pathway from vision to action than pixel-level reconstruction.

Keywords: Latent Action · Embodied AI · Benchmark

1 Introduction

The paradigm of learning from large-scale, unlabeled human video data has emerged as a promising solution to the "data island" problem in robotics, where diverse action-labeled datasets remain too scarce to train generalist foundation models. The pivotal challenge in leveraging human video data lies in the transformation of raw visual signals into ontology-agnostic action representations, commonly referred to as latent actions [4,6–8,10,26,36,37]. Pioneering Latent Action Models (LAM) such as Latent Action Pretraining from Videos (LAPA) [37], Moto [8], and LAPO [29] introduced unsupervised frameworks that tokenize visual changes between video frames into discrete latent action tokens, analogous to word pieces in natural language processing. Similarly, IGOR [6] treats the compression of visual changes between initial and goal images as atomic control units, facilitating semantic consistency across human and robot embodiments.

Despite recent progress in transforming vision to latent action, it lacks a thorough and rigorous framework for assessing the quality and effectiveness of latent action representations. Existing evaluations primarily rely on downstream manipulation task performance or qualitative methods such as cluster visualization [4, 6, 8, 11, 26, 28, 29, 37]. These approaches fail to decouple the assessment of the VLA components from the latent action quality itself. Furthermore, there is a notable absence of evaluation methods that span different entities, tasks, and granularities, making it difficult to assess the generalization capabilities of action representations. Crucially, the impact of different training architectures, strategies, and usage paradigms on action representation remains underexplored.

To bridge this gap, we introduce the **L**atent **A**ction **R**epresentation **Y**ielding (**LARY**) Benchmark, a quantitative framework designed to rigorously evaluate latent action representations. Our objective is to establish a standardized metric that assesses both embodied capabilities spanning cross-agent and cross-scenario applications and video understanding ability.

Actions in embodied intelligence inherently span two complementary levels: high-level semantic intent that specifies *what to do*, and low-level physical control that determines *how to do it*. LARYBench evaluates latent action representations along both dimensions. For High-Level Semantic Understanding, we assess whether representations can distinguish atomic primitives (e.g., "move up", "close gripper") and composite behaviors (e.g., "pick up", "place", "twist"), drawing on diverse human and robot manipulation videos that cover a wide range of scenarios. For Low-Level Control Mapping, we examine whether representations

preserve sufficient physical detail to reconstruct end-effector trajectories, using embodied datasets with fine-grained robot motion annotations.

Since current embodied video datasets [9, 12–14, 19, 34] often suffer from imprecise temporal boundaries and inconsistent action annotations, we develop an automated data engine to re-segment and re-annotate a large-scale corpus. The resulting benchmark comprises over 1.2M short videos (totaling more than 1,000 hours) spanning 151 unique action categories, alongside 620K image pairs and 595K motion trajectories. It covers both human and robotic agents, captured from egocentric and exocentric perspectives across simulated and real-world environments. Building on this diverse foundation, LARYBench instantiates the two evaluation dimensions as concrete tasks: Semantic Action Classification and Low-level Control Regression.

To systematically assess latent action quality, we benchmark three families of models: (i) **Embodied LAMs** (e.g., LAPA [37], UniVLA [4], villa-X [7]), which are purpose-built for robot manipulation; (ii) **General Vision Encoders**, including both semantic-level (DINOv3 [30], V-JEPA 2 [3]) and pixel-level (FLUX.2-dev [16], Wan2.2 [33]) backbones, evaluated for their inherent capacity to encode action-relevant features without explicit action supervision; and (iii) **General LAMs**, a new class of models we propose by grafting the LAM training paradigm onto frozen general vision backbones (e.g., LAPA-DINOv2, LAPA-DINOv3, LAPA-SigLIP2, LAPA-MAGVIT2). Across 11 models, our evaluation reveals a consistent hierarchy: off-the-shelf General Vision Encoders outperform General LAMs, which in turn surpass Embodied LAMs, suggesting that current embodied-specific training not only fails to leverage powerful visual priors but may actually constrain representation quality.

Our contributions are as follows:

- We introduce LARYBench, a comprehensive benchmark that first decouples the evaluation of latent action representations from downstream policy performance. LARYBench probes representations along two complementary dimensions, high-level semantic action (*what to do*) encoding and the low-level physical dynamics required for robotic control (*how to do it*), enabling direct, standardized measurement of representation quality itself.
- To support rigorous evaluation, we develop an automated data engine to re-segment and re-annotate a large-scale corpus, yielding 1.2M videos, 620K image pairs, and 595K trajectories across 151 action categories and 11 robotic embodiments, covering both human and robotic agents from egocentric and exocentric perspectives in simulated and real-world environments.
- Through systematic evaluation of 11 models, we reveal two consistent findings: (i) action-relevant features can emerge from large-scale visual pretraining without explicit action supervision, and (ii) latent-based feature spaces tend to align with robotic control better than pixel-based ones. These results suggest that future VLA systems may benefit more from leveraging general visual representations than from learning action spaces solely on scarce robotic data.

2 Related Work

Latent Action Representation from Videos. To alleviate reliance on teleoperated data, recent research extracts latent control signals from unlabeled videos via Inverse Dynamics Models (IDMs). Existing approaches diverge into discrete and continuous paradigms. Discrete methods, such as LAPA [37] and Moto [8], employ Vector Quantization (VQ) to facilitate autoregressive behavior cloning, though often at the cost of fine-grained information loss. Conversely, continuous approaches like CoMo [36] preserve motion fidelity but risk shortcut learning from background cues. To improve physical grounding and mitigate such artifacts, recent works incorporate semantic constraints via language or saliency (UniVLA [4], IGOR [6]) and integrate physical priors, like real robot trajectories [7, 17, 25].

Latent Actions in World Models and VLAs. Latent actions function as a unification interface in generalist systems. In Vision-Language-Action (VLA) architectures like GR00T [26], they decouple high-frequency control from low-frequency reasoning. Simultaneously, World Foundation Models (WFMs), including Cosmos [2], VideoWorld [28], AdaWorld [11], and V-JEPA 2 [3] utilize latent actions to condition future-frame prediction, enabling agents to internalize physical rules from passive observation. Expanding on the data sources for these models, DreamDojo [10] leverages large-scale human videos to construct a generalist robot world model, demonstrating the efficacy of extracting physical dynamics and actionable representations directly from human demonstrations. Advanced frameworks such as F1 [21] and InternVLA-A1 [5] further unify understanding, generation, and action within a single VLA framework, with the former emphasizing visual foresight via predictive inverse dynamics modeling.

Evaluation of Latent Action Representations. Despite the proliferation of LAMs, quantitative evaluation remains challenging. Standard reconstruction metrics often fail to distinguish action dynamics from environmental noise. While benchmarks like EWMBENCH [38] and LAWM [31] utilize trajectory consistency or Canonical Correlation Analysis (CCA) for alignment, recent diagnostic studies [15, 24, 39] reveal that many models struggle with distractor robustness and the extraction of task-relevant semantics. Our work extends these inquiries by employing attentive probing and regression to rigorously test the semantic separability and embodied ability of latent action representations.

3 The LARY Benchmark

To enable a comprehensive assessment of latent action representations, we propose the Latent Action Representation Yielding Benchmark (LARYBench), which unifies the evaluation of both high-level semantic action encoding and the low-level physical dynamics required for robotic control. Formally, given a sequence of visual observations $o_{1:T}$, the motion information is extracted by a latent action model (LAM) as the latents $z \in \mathcal{Z}$. LARYBench evaluates the efficacy of $\mathcal{Z}$ through two tasks: $f_{sem} : \mathcal{Z} \to \mathcal{C}$ for semantic action decoding via classification task, and $f_{dyn} : \mathcal{Z} \to \mathcal{A}$ for robotic control construction via regression task.

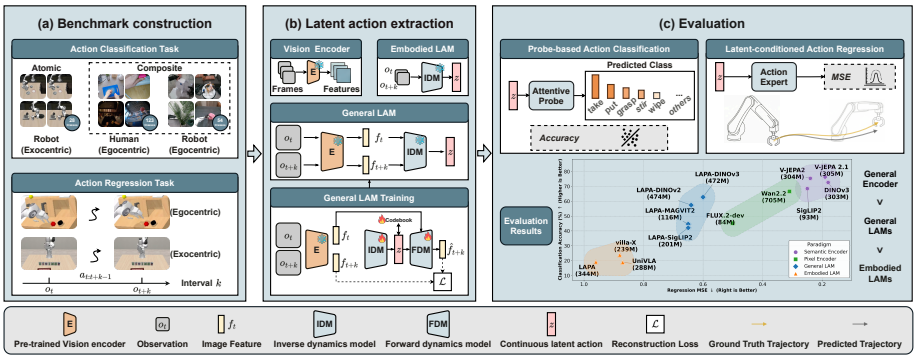

Fig. 2: Overall pipeline of LARYBench. First, (a) we construct a comprehensive dataset featuring both atomic and composite actions across varied domains to support classification and regression tasks. Next, (b) we extract the continuous latent action z using various representation paradigms, also highlighting the integration of pre-trained general vision encoders within the VQ-VAE training architecture. Finally, (c) we evaluate these representations utilizing probe-based classification and latent-conditioned regression to quantify their semantic separability (Accuracy) and physical dynamics modeling capabilities (MSE), respectively.

As shown in Figure 1, LARYBench is curated from a massive amount of multi-embodiment datasets and human datasets, encompassing 151 meticulously defined actions (including 28 atomic actions and 145 composite actions), and corresponding 1.2M annotated samples. The dataset covers a large range of human activities from the frequent "pick" and "place", to the long-tail "shovel" (snow) and "float" (balloon). To ensure morphological diversity, the dataset spans 11 distinct robotic embodiments, ranging from widely used single-arm manipulators such as Franka to complex bimanual and semi-humanoid platforms such as the AgiBot G1, Agilex Cobot, and Realman series, and includes extensive human-ego-centric interaction data. To ensure environmental diversity, the dataset captures thousands of unique object manipulations across a broad spectrum of unstructured environments, including simulated tabletops, authentic residential kitchens, commercial spaces, and industrial scenes.

With the diverse dataset spanning actions, embodiments, and objects, the evaluation framework of LARYBench is depicted in Figure 2. Next, we introduce the construction pipeline for the semantic action classification task and the robotic control regression task, respectively.

3.1 Hierarchical Semantic Probing Protocol

Evaluating the semantic richness of latent action representations requires disentangling spatio-temporal complexities. We formalize this through a multi-granularity semantic probing protocol.

Kinematic-Level Atomic Primitives At the most granular semantic level, representations must capture instantaneous state variations. We define the *Atomic*

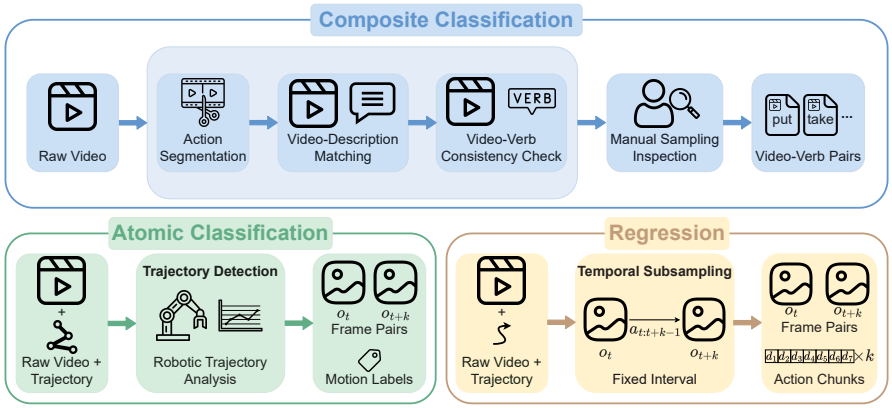

Fig. 3: Data curation process for LARYBench. To efficiently transform diverse raw videos into standardized evaluation tasks, we integrate a Vision-Language Model (VLM) with robust spatio-temporal video understanding capabilities. The VLM serves as the core reasoning agent for temporal video segmentation and semantic action alignment.

Robot task by decomposing the robot's end-effector actions into 28 discrete kinematic primitives, comprising finely resolved directional translations and binary gripper states. Leveraging exo-centric demonstrations from the LIBERO suite [18], we apply detailed data preprocessing, such as thresholding motion along the z-axis relative to static orthogonal directions, to extract 25,940 high-quality image pairs with trajectories.

Task-Level Composite Behaviors To assess the capacity for abstracting complex behavioral semantics across diverse embodiments and scenarios, we introduce the *Composite Human* and *Composite Robot* tasks. To handle diverse data quality across existing datasets [9,12–14,34], we develop an automated and scalable data engine to perform precise temporal segmentation and semantic alignment (see Figure 3). Comprehensive details of this engine's architecture are provided in the Appendix. By deploying this system across a diverse corpus of ego-centric human datasets (HoloAssist [34], Ego4D [13], Something-Somethingv2 [12], TACO [19], EPIC-KITCHENS [9], EgoDex [14]) and realistic bimanual robot demonstrations (AgiBotWorld-Beta [1]), we extract 692,297 human clips and 538,423 robot clips under a unified taxonomy of 145 composite behavior classes. Beyond this initial curation, our automated engine holds the potential to continuously process future data streams, ensuring the ever-growing nature of the dataset.

3.2 Physical Execution Mapping Assessment

While semantic features encode *what to do*, robotic manipulation ultimately demands *how to do*. Our trajectory regression task evaluates the latent space's physical grounding by directly decoding continuous end-effector actions.

This protocol spans diverse hardware morphologies and action spaces. For single-arm exo-centric scenarios, we utilize CALVIN [22] and VLABench [40] featuring Franka arms. For bimanual ego-centric scenarios, we incorporate RoboCOIN [35] across 10 diverse platforms and AgiBotWorld-Beta featuring the AgiBot G1. The action space heterogeneity is meticulously preserved: AgiBotWorld-Beta targets a 16-DoF space containing absolute position, quaternions, and gripper state, whereas RoboCOIN maps to a 12-DoF space of position and Euler angles. We deliberately mask the dexterous hand joint data in RoboCOIN to focus the evaluation on macroscopic arm displacements, as fine-grained finger articulation remains an ill-posed inverse problem for current visual encoders.

4 Experiments

In this section, we conduct comprehensive experiments and try to answer the following questions.

1. Do Latent Actions Capture Diverse Actions?
2. Do Latent Actions Encode Enough for Control?
3. What Constitutes an Effective Latent Action Model?

4.1 Taxonomy of Action Representation Paradigms

To establish a comprehensive baseline, we categorize latent action models into four distinct learning paradigms:

- **Embodied Latent Action Models:** Architectures natively designed for VLAs, such as LAPA [37], UniVLA [4], and villa-X [7], which explicitly entangle temporal dynamics with robotic control via forward/inverse dynamics modeling.
- **General Semantic Encoders:** High-capacity visual backbones like DINOv3 [30] and V-JEPA-2 [3] optimized for contrastive alignment or latent-level reconstruction, evaluated here for their zero-shot temporal modeling capacity.
- **Generative Pixel Encoders:** Video synthesis models such as Wan2.2 VAE [33] and FLUX.2-dev VAE [16] that leverage spatial-temporal compression and pixel-level reconstruction to implicitly capture motion priors.
- **General Latent Action Models:** To leverage the extensive knowledge embedded in pre-trained visual backbones, we explore modeling visual motion at the feature level. Specifically, we substitute the default encoder in the LAPA framework with various pre-trained backbones—ranging from semantic encoders like DINOv2 [27], DINOv3 [30], and SigLIP 2 [32] to pixel ones such as MAGVIT2 [20]. These hybrid models, trained on internet data containing human motions, robot motions, and environment motions, are included to assess how pre-trained visual priors enhance latent action learning.

4.2 Evaluation Settings

Data Split and Evaluation Metrics To maintain consistent class distributions across classification tasks, we apply randomized stratified sampling, partitioning atomic and composite tasks at 75:25 and 70:30 ratios, respectively. Regression tasks on VLABench [40] adhere to a standard 75:25 train-validation split. For cross-environment evaluation on CALVIN [22], environments A, B, and C are allocated for training, with D designated for validation. RoboCOIN [35] and AgiBotWorld-Beta [1] are also divided at a 75:25 ratio for training and validation. For semantic action classification, we report the Top-1 Accuracy averaged across all action categories. For low-level control regression, we utilize several metrics to measure physical fidelity. In Table 2, we report the Normalized MSE, which calculates the element-wise MSE in the normalized action space. To provide a comprehensive evaluation, we also define three unnormalized physical metrics, detailed results for which are provided in the Appendix: (1) **phys_vector_mse**: the element-wise MSE over the entire unnormalized action vector, averaged across all elements; (2) **phys_se3_mse**: the sum of the Euclidean squared error for position and the SO(3) angular squared error for orientation (excluding the gripper state); and (3) **phys_balanced_mse**: the MSE calculated separately for the position, orientation, and gripper semantic blocks, which are then averaged to ensure balanced weighting.

Sampling and Latent Action Extraction For the Atomic Robot and regression tasks, we mainly rely on curated image pairs, which allow us to obtain latent action representations from the LAM directly. To retain as much information as possible, we work with the continuous latent embeddings instead of discretized codebook indices, thereby avoiding quantization-related information loss. In contrast, extracting representations from video clips in the composite classification tasks is more difficult because the source datasets differ in FPS and exhibit non-uniform motion speeds. Simple uniform sampling generally fails to capture the true temporal dynamics and often yields latent actions that are effectively insensitive to the underlying motion. To address this, we employ the Motion-Guided Sampler (MGSampler) [41] to select an effective sequence of frames that exhibits sufficient dynamic variation. Latent action features are subsequently extracted across all adjacent sampled frames to ensure the representation effectively encapsulates the transition dynamics.

Probing Protocol for Semantic Action Classification We uniformly sample 9 frames from each video clip (typically lasting under 5 seconds) and resize every frame to 224×224. To assess the semantic expressiveness of the learned latent actions, we employ a probe-based classification task. Following the architecture in V-JEPA [3], we use a 4-layer attentive probe as the classifier. Given the varying latent dimensions between general vision encoders and LAMs, we incorporate a projector to align the continuous latent actions into a uniform dimensional space prior to classification, thereby ensuring a fair comparison. The probe is trained for 20 epochs using bfloat16 precision. We apply a multi-head

optimization strategy with learning rates from 3×10^{-4} to 5×10^{-3} and weight decay values from 0.01 to 0.8, to ensure a robust evaluation across different representation scales.

Action Experts for Low-level Control Regression The latent actions are derived from image pairs separated by a frame interval of $s = 5$. To assess the physical grounding of latent actions, we implement an easy MLP-based Action Experts to regress absolute end-effector trajectories. It adopts a standard MLP with residual connections architecture featuring 2 residual blocks and a hidden dimension of 4096. This model directly maps latent actions to an action chunk of size s (7-DoF/12-DoF/16-DoF per chunk).

Feature-Level Latent Action Model Training To leverage the pretrained vision encoders, we develop a family of Feature-Level Latent Action Models by substituting the visual representation from pixels to features generated by vision encoders in the LAPA framework, while retaining its VQ-VAE structure. During training, the encoder weights remain frozen, and the latent action representation is learned from scratch on an internal video dataset. Detailed training configurations are provided in the Appendix.

4.3 Benchmark Results

Do Latent Actions Capture Diverse Actions? As shown in Table 1, the average (Avg.) accuracy is computed across composite human and robot actions to evaluate high-level semantic understanding. General vision foundation models (e.g., V-JEPA series [3,23], DINOv3 [30], SigLIP2 [32], and Wan2.2 [33]) deliver surprisingly strong performance, even though they do not appear to perform any explicit motion extraction. **This phenomenon demonstrates that the visual self-supervised training with large-scale data inherently yields general semantic action representations covering both robot and human data.** In particular, V-JEPA 2.1 [23] learns directly from visual latent features instead of pixel features (the representations commonly used in world models), and achieves the best performance by a substantial margin. **This supports the hypothesis that actions can be derived from latent features without needing to be explicitly represented in the pixel space.** In contrast, existing Embodied LAMs (e.g., LAPA [37], UniVLA [4] and villa-X [7]) exhibit restricted generalization on diverse actions (averaging 18.82%–23.85%), due to the limited amount of training data or the early constraints to low-level actions [4,7].

We try to combine these two kinds of techniques, which extract motion using the self-supervised method in LAPA [37] on general vision foundation models, named as **General LAMs**. The training dataset is curated from the internet data, containing both human motions and non-human motions (e.g., car driving), which is very different from the test dataset. As shown in Table 1, the General LAMs perform better than existing Embodied LAMs, even on capturing

the robot action. Although sharing the same vision foundation model, LAPA-DINOv2 significantly outperforms UniVLA (57.41% vs. 18.82%) by leveraging more diverse training data and imposing fewer constraints. Although the General LAMs achieve relatively high accuracy on human actions, their performance declines on robot actions due to limited data scale and diversity. We leave addressing this issue to future work.

In conclusion, vision foundation models can capture semantic action with self-supervised learning from large-scale of internet visual data. Learning based on latent features performances better than pixel features. More data diversity and fewer task-specific constrains may be the key to generalization.

Do Latent Actions Encode Enough for Control? We evaluate robotic control under two action formulations. (1) *Absolute Regression* predicts exact end-effector poses, and (2) *Delta Regression* predicts relative pose changes. We report Normalized MSE for both (Table 2). Since the absolute and delta targets are normalized within their own action spaces, the two formulations are not directly comparable in raw MSE. We therefore compare models within each formulation rather than across them.

Under **Absolute Regression**, semantic encoders dominate. DINOv3 [30] and V-JEPA 2.1 [23] attain the best average MSE (0.18 and 0.19). On each dataset, the strongest semantic encoder leads the strongest General LAM (LAPA-DINOv3) by a large margin, for example 0.18 versus 0.50 on CALVIN. Latent-based encoders also capture absolute control far better than pixel encoders such as Wan2.2 [33] and FLUX.2-dev [16]. This indicates that per-frame semantic features align naturally with absolute end-effector positioning.

Under **Delta Regression**, the separation between paradigms narrows sharply. V-JEPA 2.1 remains the single strongest model on average (0.50). The best General LAM, however, now trails it only slightly (0.57), a far smaller margin than under Absolute Regression. Such behavior is consistent with the LAM construction, which extracts latents from consecutive frame pairs and thus encodes transition-like quantities.

By contrast, both Embodied LAMs (avg 0.82–0.94) and pixel encoders (avg 0.84) remain weak under Delta Regression. This indicates that neither the embodied-specific training nor the pixel-reconstruction objective extracts relative motion as effectively as the General-LAM paradigm.

In summary, the most effective representation depends on the action formulation. For absolute spatial positioning, semantic visual features such as DINOv3 and V-JEPA 2.1 are clearly preferable. For relative increments, semantic encoders remain a strong default, since V-JEPA 2.1 still leads on average, but the General-LAM paradigm closes most of the gap and becomes competitive.

What Constitutes an Effective Latent Action Model? To systematically build a robust latent action model, we conduct ablations under the LAPA [37] framework. Specifically, we focus our ablations on the absolute action space. Because predicting absolute actions is inherently more challenging and highly sensitive to

Table 1: Action Classification results of latent action representations. Avg. denotes the mean accuracy across Composite Human and Composite Robot. Best results are in **bold** and second best are underlined.

Model	Paradigm	Params(M)	Classification Accuracy↑			
			Avg.	*Atomic Robot*	*Composite Human*	*Composite Robot*
V-JEPA2 [3]	Semantic Encoder	303.89	75.39	79.09	80.35	70.43
V-JEPA 2.1 [23]		304.68	**76.29**	**84.19**	**81.14**	**71.44**
DINOv3 [30]		303.13	72.63	60.79	76.19	69.06
SigLIP2 [32]		92.88	68.49	52.84	70.40	66.58
Wan2.2 [33]	Pixel Encoder	704.69	66.58	14.91	67.77	65.39
FLUX.2-dev [16]		84.05	45.24	51.95	46.12	44.36
LAPA [37]	Embodied LAM	343.80	19.13	22.27	14.61	23.64
UniVLA [4]		287.75	18.82	18.62	19.08	18.56
villa-X [7]		238.71	23.85	15.00	17.80	29.90
LAPA-MAGVIT2	General LAM	116.40	44.62	33.12	59.70	29.53
LAPA-SigLIP2		200.83	42.09	46.83	54.74	29.44
LAPA-DINOv2		473.69	57.41	58.04	55.86	58.96
LAPA-DINOv3		472.45	62.73	56.27	64.19	61.26

Table 2: Action Regression results of latent action representations Avg. is the mean over available dataset results. Best results are in **bold** and second best are underlined.

Model	Paradigm	Params(M)	Normalized MSE↓									
			Absolute					Delta				
			Avg.	CALVIN	VLABench	RoboCOIN	AgiBotWorld -Beta	**Avg.**	CALVIN	VLABench	RoboCOIN	AgiBotWorld -Beta
V-JEPA 2 [3]	Semantic Encoder	303.89	0.24	0.25	0.06	0.33	0.33	0.57	0.35	0.36	0.80	0.76
V-JEPA 2.1 [23]		304.68	0.19	**0.18**	**0.04**	0.27	0.28	**0.50**	**0.27**	0.43	**0.69**	0.59
DINOv3 [30]		303.13	**0.18**	0.20	0.06	**0.22**	**0.24**	0.63	0.51	0.46	0.74	0.79
SigLIP2 [32]		92.88	0.25	0.29	0.05	0.30	0.34	0.60	0.45	**0.34**	0.78	0.84
Wan2.2 [33]	Pixel Encoder	704.69	0.31	0.38	0.09	0.36	0.40	0.84	0.86	0.59	0.98	0.94
FLUX.2-dev [16]		84.05	0.50	0.56	0.32	0.49	0.63	0.84	0.68	0.81	0.97	0.88
LAPA [37]	Embodied LAM	343.80	0.96	0.94	0.90	0.99	1.00	0.94	0.79	0.98	1.03	0.95
UniVLA [4]		287.75	0.87	0.82	0.73	0.97	0.97	0.82	0.60	0.82	0.99	0.88
villa-X [7]		238.71	0.88	0.86	0.71	0.97	0.97	0.82	0.59	0.82	0.98	0.87
LAPA-MAGVIT2	General LAM	116.40	0.65	0.59	0.36	0.80	0.83	0.70	0.60	0.58	0.88	0.74
LAPA-SigLIP2		200.83	0.65	0.57	0.30	0.86	0.88	0.67	0.52	0.53	0.91	0.71
LAPA-DINOv2		473.69	0.64	0.55	0.27	0.88	0.84	0.60	0.43	0.52	0.87	0.58
LAPA-DINOv3		472.45	0.60	0.50	0.22	0.85	0.84	0.57	0.43	0.46	0.83	**0.57**

spatial-temporal precision than relative delta-actions, it serves as a more rigorous testbed for evaluating representation capacity. Figure 4 maps a performance evolution path bridging the gap between the worst baseline (LAPA) and V-JEPA 2. First, self-supervised visual encoders (e.g., DINOv3 [30]) consistently construct superior LAMs compared to reconstruction-based and vision-language contrastive models like MAGVIT2 [20] and SigLIP2 [32] (Tables 1-2), indicating that contrastive priors better capture fine-grained spatial-temporal correspondences. Setting LAPA-DINOv3 ($cs = 8, sl = 16, dim = 32$) as our foundation, we sequentially optimize its quantization bottleneck. Second, expanding codebook capacity improves downstream regression, but overly large sizes (e.g., 256) cause codebook utilization drops without further gains (Table 3), making a moderate size ($cs = 64$) optimal for dense representations. Third, sequence length critically dictates temporal diversity; short sequences ($sl = 16$) trigger catastrophic codebook collapse (1.6% utilization), whereas a moderate length ($sl = 49$) en-

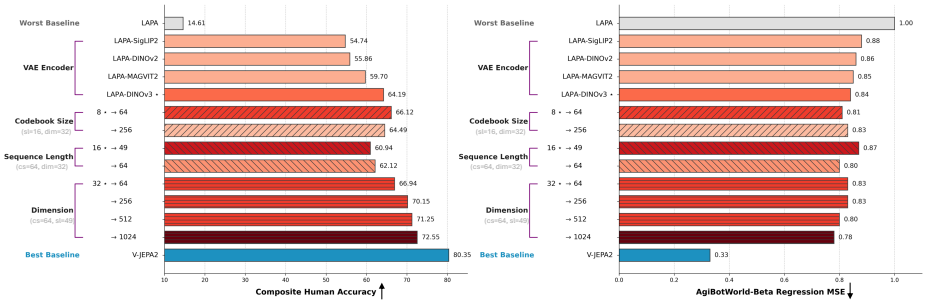

Fig. 4: Performance Evolution of Latent Action Models. The (*) denotes the default quantization settings for LAPA-DINOv3 ($cs = 8, sl = 16, dim = 32$). cs, sl, and dim represent Codebook Size, Sequence Length and Latent Dimension respectively. *Composite Human* classification on the left and AgiBotWorld-Beta [1] regression on the right.

Table 3: Ablation Studies. Default settings are: Codebook Size = 64, Sequence Length = 49, Latent Dimension = 256. Best results are in **bold** and second best are underlined within each section.

Ablated Setting	Recon Loss	Codebook Utilization (%)	Accuracy (Composite task)↑		MSE (AgiBotWorld-Beta)↓
			Human	Robot	
Codebook Size					
8	0.00808	**100.0**	**71.31**	**64.89**	0.88
64	**0.00751**	**100.0**	70.15	64.04	**0.83**
256	0.00758	89.5	69.84	63.82	0.85
Sequence Length					
16	0.00908	1.6	69.23	63.33	0.79
49	**0.00751**	**100.0**	70.15	64.04	0.83
64	0.00773	79.7	**71.37**	**64.70**	**0.72**
Latent Dimension					
32	0.01141	3.1	60.94	57.69	0.87
64	0.00967	**100.0**	66.94	63.07	0.83
256	0.00751	**100.0**	70.15	64.04	0.83
512	0.00732	1.6	71.25	64.97	**0.80**
1024	**0.00670**	84.4	**72.55**	**65.78**	0.81

sures 100% utilization and robust generalization (Table 3). Finally, scaling the latent dimension improves theoretical capacity but introduces quantization instability, evidenced by sudden utilization collapses at intermediate sizes (Table 3). Thus, $dim = 256$ strikes the best capacity-stability balance. **Ultimately, beyond dataset quality, an effective latent action space requires two key components: robust self-supervised visual priors to capture precise spatial-temporal dynamics, and a strictly regularized quantization bottleneck to maintain stability and dense utilization**.

5 Error Analysis

To provide deeper insights into the limitations of latent action representations, we conduct a fine-grained error decomposition.

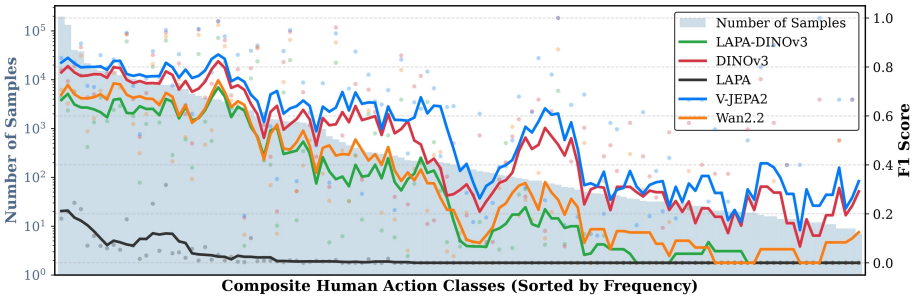

Fig. 5: Action classification performance across the long-tail distribution of the *Composite Human* dataset. The background histogram indicates the number of samples per action class, sorted by descending frequency. Solid lines represent the moving average F1 scores, while transparent scatters denote exact class-wise scores.

5.1 The Long-Tail Dilemma and Mid-Frequency Semantic Aliasing

Figure 5 illustrates action classification performance across class frequencies on the *Composite Human* dataset. While the baseline LAPA [37] model exhibits uniformly poor performance, LAPA-DINOv3 closely mirrors the continuous DINOv3 [30] encoder, demonstrating a direct inheritance of both its representational strengths and specific vulnerabilities. Stronger models generally outperform weaker ones across most actions, suggesting that the LARYBench provides a stable and reliable evaluation of model capabilities. As the frequency decreases, the performance gap between strong and weak models widens, indicating that strong models exhibit better generalization capabilities in long-tail scenarios.

5.2 Spatiotemporal Grounding and Action-Centric Attention

Figure 6 visualizes the cross-attention maps (additional heatmap visualizations are provided in the Appendix), revealing a stark contrast in how different models ground physical interactions. First, among general visual encoders, self-supervised models demonstrate superior spatiotemporal grounding. Notably, V-JEPA 2 [3] exhibits the most accurate attention distribution, sharply localizing on the exact interaction points between both the left and right hands and the manipulated object (bowl). DINOv3 [30] similarly maintains a precise, geometry-aware focus on the active end-effector. Conversely, generative priors Flux2-dev [16] and Wan2.2 [33] exhibit highly dispersed, unfocused attention distributions, indicating an inherent bias toward global scene understanding rather than localized physical interactions. Second, regarding LAMs versus general encoders, standard Embodied LAMs, like LAPA [37], completely fail to localize meaningful features, producing broad, uninformative attention blobs. However, our LAPA-DINOv2 effectively inherits the strong localization capabilities of its backbone. Despite the severe spatial quantization bottleneck (e.g., $SL = 16$, producing coarse 4×4 patches), it successfully anchors its attention to the active object. Finally, within the LAM architectures, our General LAMs consistently exhibit

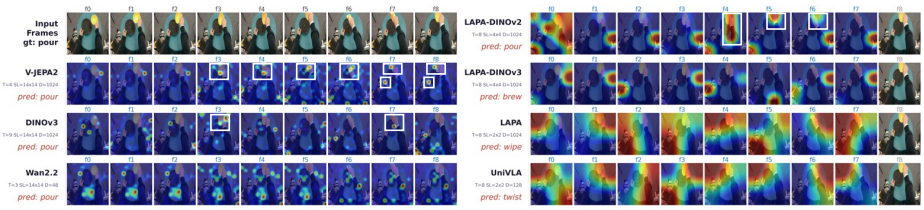

Fig. 6: Cross-attention heatmaps of the temporal pooler across various models for a 9-frame *pour* sequence. The spatial sequence length (SL) dictates visual granularity: $SL = 14 * 14$ yield fine-grained 14×14 heatmaps, whereas LAMs ($SL \in \{4, 16\}$) produce blocky 2×2 or 4×4 attention regions. Temporally, $T = 9$ models compute frame-by-frame attention; $T = 8$ models extract pairwise latent actions (leaving the 9th frame unmodified); and models with temporal compression ($T \in \{3, 4\}$) duplicate attention maps across their corresponding frame groups.

sharper object-centric grounding compared to other Embodied LAMs (e.g., UniVLA [4], villa-X [7]), which suffer from severe attention diffusion. **Ultimately, this demonstrates that predictive failure fundamentally stems from an inability to concentrate attention on the corresponding dynamically interacting objects**.

5.3 Stride Ablation: Latent Action Spaces Encode Robust Dynamic Trajectories

To investigate the temporal robustness of various representations, we ablate the sampling stride (stride=5, 15, 30) between input frames on VLABench, as shown in Table 4. Crucially, as the stride increases, the number of actions to regress scales linearly, significantly amplifying the dimensionality and complexity of the task. Pure spatial generative encoders, such as FLUX.2-dev [16], excel at extremely short horizons (achieving 0.04 MSE at stride=5) but fail catastrophically under this increased temporal dimensionality (MSE spiking to 0.62 at stride=30). In contrast, Latent Action Models (LAMs)—encompassing both Embodied LAMs and our General LAMs—exhibit consistent stability across all strides. This shared characteristic proves that the latent action paradigm does not merely encode static spatial alignments; instead, it successfully captures and preserves underlying dynamic action trajectories. While traditional Embodied LAMs suffer from a high error (about 0.70 average MSE), introducing general visual priors as in our General LAMs effectively lowers this error margin while fully inheriting the paradigm's temporal stability. **In conclusion, although a performance gap remains when compared to general vision encoders, the fundamental mechanism of mapping visual observations into a latent action space provides a robust representation for long-horizon intents.**

Table 4: Ablation study on VLABench [40] sampling stride. We report the regression MSE with sampling strides of 5, 15, and 30, and their average. Best results are in **bold** and second best are underlined.

Model	Paradigm	Params(M)	Regression MSE↓			
			stride=5	stride=15	stride=30	**Avg.**
V-JEPA 2 [3]	Semantic Encoder	303.89	0.07	**0.13**	**0.16**	**0.12**
DINOv3 [30]		303.13	0.06	0.20	0.25	0.17
Wan2.2 [33]	Pixel Encoder	704.69	0.09	0.16	0.24	0.16
FLUX.2-dev [16]		84.05	**0.04**	0.57	0.62	0.41
LAPA [37]	Embodied LAM	343.80	0.95	0.87	0.77	0.86
UniVLA [4]		287.75	0.74	0.68	0.69	0.70
villa-X [7]		238.71	0.72	0.70	0.66	0.69
LAPA-MAGVIT2	General LAM	116.40	0.36	0.32	0.33	0.34
LAPA-SigLIP2		200.83	0.30	0.25	0.24	0.26
LAPA-DINOv2		473.69	0.26	0.23	0.37	0.29
LAPA-DINOv3		472.45	0.25	0.20	0.26	0.24

6 Conclusion

We introduce LARYBench to evaluate latent action representations across kinematic and semantic granularities. Our empirical results reveal that general visual foundation models consistently outperform specialized Embodied Latent Action Models (LAMs). Specifically, visual understanding models dominate in semantic tasks, while general visual encoders surprisingly achieve superior regression MSE without domain-specific training. This indicates that effective latent actions naturally emerge from large-scale visual pre-training, whereas specialized Embodied LAMs often suffer from representation collapse due to data scarcity or premature constraints to domain-specific low-level control. Consequently, we advocate for a paradigm shift in Vision-Language-Action (VLA) design: instead of learning action spaces from limited robotic data, future research should focus on aligning control policies with the robust feature spaces of general vision models. Addressing architectural challenges such as continuous signal decoding and feature alignment will be pivotal to unlocking these universal priors for data-efficient embodied agents.

Acknowledgement

We hereby express our appreciation to the LongCat Team EVA Committee for their valuable assistance, guidance, and suggestions throughout the course of this work.

References

1. AgiBot-World-Contributors, Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Huang, X., Jiang, S., Jiang, Y., Jing, C., Li, H., Li, J., Liu, C.,

Liu, Y., Lu, Y., Luo, J., Luo, P., Mu, Y., Niu, Y., Pan, Y., Pang, J., Qiao, Y., Ren, G., Ruan, C., Shan, J., Shen, Y., Shi, C., Shi, M., Shi, M., Sima, C., Song, J., Wang, H., Wang, W., Wei, D., Xie, C., Xu, G., Yan, J., Yang, C., Yang, L., Yang, S., Yao, M., Zeng, J., Zhang, C., Zhang, Q., Zhao, B., Zhao, C., Zhao, J., Zhu, J.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv: 2503.06669 (2025)
2. et. al., N.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
3. 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)
4. Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., Li, H.: Univla: Learning to act anywhere with task-centric latent actions (2025), https://arxiv.org/abs/2505.06111
5. Cai, J., Cai, Z., Cao, J., Chen, Y., He, Z., Jiang, L., Li, H., Li, H., Li, Y., Liu, Y., Lu, Y., Lv, Q., Ma, H., Pang, J., Qiao, Y., Qiu, Z., Shen, Y., Shi, X., Tian, Y., Wang, B., Wang, H., Wang, J., Wang, T., Wei, X., Wu, C., Xie, Y., Xing, B., Yang, Y., Yang, Y., Yu, Q., Yuan, F., Zeng, J., Zhang, J., Zhang, S., Zhang, S., Zhaxi, Z., Zhou, B., Zhou, Y., Zhou, Y., Zhu, H., Zhu, Y., Zhu, Y.: Internvla-a1: Unifying understanding, generation and action for robotic manipulation (2026), https://arxiv.org/abs/2601.02456
6. Chen, X., Guo, J., He, T., Zhang, C., Zhang, P., Yang, D.C., Zhao, L., Bian, J.: Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. arXiv preprint arXiv:2411.00785 (2024)
7. Chen, X., Wei, H., Zhang, P., Zhang, C., Wang, K., Guo, Y., Yang, R., Wang, Y., Xiao, X., Zhao, L., et al.: Villa-x: enhancing latent action modeling in vision-language-action models. arXiv preprint arXiv:2507.23682 (2025)
8. Chen, Y., Ge, Y., Li, Y., Ge, Y., Ding, M., Shan, Y., Liu, X.: Moto: Latent motion token as the bridging language for robot manipulation. arXiv preprint arXiv: 2412.04445 (2024)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 4125–4141 (2020)
10. Gao, S., Liang, W., Zheng, K., Malik, A., Ye, S., Yu, S., Tseng, W.C., Dong, Y., Mo, K., Lin, C.H., et al.: Dreamdojo: A generalist robot world model from large-scale human videos. arXiv preprint arXiv:2602.06949 (2026)
11. Gao, S., Zhou, S., Du, Y., Zhang, J., Gan, C.: Adaworld: Learning adaptable world models with latent actions. arXiv preprint arXiv:2503.18938 (2025)
12. Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense (2017), https://arxiv.org/abs/1706.04261
13. 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)
14. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)

15. Klepach, A., Nikulin, A., Zisman, I., Tarasov, D., Derevyagin, A., Polubarov, A., Lyubaykin, N., Kiselev, I., Kurenkov, V.: Object-centric latent action learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 22626–22634 (2026)
16. Labs, B.F.: Flux.2. https://blackforestlabs.ai/ (2024)
17. Li, Z., Gao, X., Wang, X., Fu, J.: Latbot: Distilling universal latent actions for vision-language-action models. arXiv preprint arXiv:2511.23034 (2025)
18. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. arXiv preprint arXiv:2306.03310 (2023)
19. Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., Yi, L.: Taco: Benchmarking generalizable bimanual tool-action-object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21740–21751 (2024)
20. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410 (2024)
21. Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., Song, H., Chen, Q., Deng, X., Pang, J.: F1: A vision-language-action model bridging understanding and generation to actions (2025), https://arxiv.org/abs/2509.06951
22. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022). https://doi.org/10.1109/LRA.2022.3180108
23. Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., Bardes, A.: V-jepa 2.1: Unlocking dense features in video self-supervised learning. arXiv preprint arXiv:2603.14482 (2026)
24. Nikulin, A., Zisman, I., Klepach, A., Tarasov, D., Derevyagin, A., Polubarov, A., Nikita, L., Kurenkov, V.: Vision-language models unlock task-centric latent actions. arXiv preprint arXiv:2601.22714 (2026)
25. Nikulin, A., Zisman, I., Tarasov, D., Lyubaykin, N., Polubarov, A., Kiselev, I., Kurenkov, V.: Latent action learning requires supervision in the presence of distractors. arXiv preprint arXiv:2502.00379 (2025)
26. NVIDIA, :, Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L.J., Fang, Y., Fox, D., Hu, F., Huang, S., Jang, J., Jiang, Z., Kautz, J., Kundalia, K., Lao, L., Li, Z., Lin, Z., Lin, K., Liu, G., Llontop, E., Magne, L., Mandlekar, A., Narayan, A., Nasiriany, S., Reed, S., Tan, Y.L., Wang, G., Wang, Z., Wang, J., Wang, Q., Xiang, J., Xie, Y., Xu, Y., Xu, Z., Ye, S., Yu, Z., Zhang, A., Zhang, H., Zhao, Y., Zheng, R., Zhu, Y.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv: 2503.14734 (2025)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Ren, Z., Wei, Y., Guo, X., Zhao, Y., Kang, B., Feng, J., Jin, X.: Videoworld: Exploring knowledge learning from unlabeled videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29029–29039 (2025)
29. Schmidt, D., Jiang, M.: Learning to act without actions. arXiv preprint arXiv:2312.10812 (2023)
30. 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)

31. Tharwat, B., Nasser, Y., Abouzeid, A., Reid, I.: Latent action pretraining through world modeling. arXiv preprint arXiv:2509.18428 (2025)
32. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
33. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
34. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20270–20281 (October 2023)
35. Wu, S., Liu, X., Xie, S., Wang, P., Li, X., Yang, B., Li, Z., Zhu, K., Wu, H., Liu, Y., et al.: Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. arXiv preprint arXiv:2511.17441 (2025)
36. Yang, J., Shi, Y., Zhu, H., Liu, M., Ma, K., Wang, Y., Wu, G., He, T., Wang, L.: Como: Learning continuous latent motion from internet videos for scalable robot learning (2025), https://arxiv.org/abs/2505.17006
37. Ye, S., Jang, J., Jeon, B., Joo, S., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.W., Lin, B.Y., Liden, L., Lee, K., Gao, J., Zettlemoyer, L., Fox, D., Seo, M.: Latent action pretraining from videos. arXiv preprint arXiv: 2410.11758 (2024)
38. Yue, H., Huang, S., Liao, Y., Chen, S., Zhou, P., Chen, L., Yao, M., Ren, G.: Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. arXiv preprint arXiv:2505.09694 (2025)
39. Zhang, C., Pearce, T., Zhang, P., Wang, K., Chen, X., Shen, W., Zhao, L., Bian, J.: What do latent action models actually learn? arXiv preprint arXiv:2506.15691 (2025)
40. Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.G., et al.: Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11142–11152 (2025)
41. Zhi, Y., Tong, Z., Wang, L., Wu, G.: Mgsampler: An explainable sampling strategy for video action recognition. In: Proceedings of the IEEE/CVF International conference on Computer Vision. pp. 1513–1522 (2021)