VLA-JEPA: Enhancing Vision-Language-Action Model with Latent World Model

Jingwen Sun[1,2]*, Wenyao Zhang[3,5]*, Zekun Qi[4], Shaojie Ren[2,6], Zezhi Liu[2,7], Hanxin Zhu[1], Guangzhong Sun[1], Xin Jin[2,5]†, and Zhibo Chen[1,2]†

[1] University of Science and Technology of China
[2] Zhongguancun Academy, Beijing, China
[3] Shanghai Jiao Tong University
[4] Tsinghua University
[5] Eastern Institute of Technology, Ningbo
[6] University of Chinese Academy of Sciences
[7] Nankai University

Abstract. Pretraining Vision-Language-Action (VLA) policies on internet-scale video is appealing, yet current latent-action objectives often learn the wrong thing: they remain anchored to pixel variation rather than action-relevant state transitions, making them vulnerable to appearance bias, nuisance motion, and information leakage. We introduce VLA-JEPA, a JEPA-style pretraining framework that sidesteps these pitfalls by design. The key idea is *leakage-free state prediction*: a target encoder produces latent representations from future frames, while the student pathway sees only the current observation—future information is used solely as supervision targets, never as input. By predicting in latent space rather than pixel space, VLA-JEPA learns dynamics abstractions that are robust to camera motion and irrelevant background changes. This yields a simple two-stage recipe—JEPA pretraining followed by action-head fine-tuning—without the multi-stage complexity of prior latent-action pipelines. Experiments on LIBERO, LIBERO-Plus, SimplerEnv and real-world manipulation tasks show that VLA-JEPA achieves consistent gains in generalization and robustness over existing methods. Code is available at https://github.com/ginwind/VLA-JEPA.

Keywords: Vision-language-action model · Latent Action · World Model

1 Introduction

Learning visuomotor policies from internet-scale video has become an increasingly attractive route for robot learning (32; 33; 71). Compared to robot interaction data, which are costly and narrow in coverage (22; 47), unlabeled videos are abundant and diverse, offering rich demonstrations of temporally extended change. This has motivated a growing body of *latent-action pretraining* for Vision–Language–Action (VLA) (9; 10; 35; 41): rather than learning actions only from scarce

*Equal contribution. †Corresponding author.

control trajectories, they first learn representations and transition structure from video, then adapt them to downstream control (13; 21; 79).

However, we argue that today's "latent action from video" objectives often do not learn what we actually need for control. For an embodied agent, the most useful notion of "action" is not a compact descriptor of pixel differences; it is a variable that captures how the underlying state will evolve under interaction, i.e., *action-relevant state transition semantics*. When pretraining is misaligned with this goal, the downstream policy inherits representations that are temporally predictive yet weakly tied to controllable structure—leading to brittle behavior, poor transfer, and inefficient fine-tuning.

Why latent-action pretraining often drifts away from action semantics?

We identify four failure modes that repeatedly appear in latent-action pretraining pipelines built on unlabeled video:

1. **Pixel-level objectives bias representations toward appearance, not action.** A common strategy is to use "the future" as supervision—either by predicting future pixels directly or by compressing frame-to-frame changes into a latent variable interpreted as an action (13; 79; 86). Even with compression mechanisms such as VQ-VAE (69), the supervision signal is frequently dominated by what changes visually: texture, illumination, background clutter, and viewpoint. These factors are high-variance but low-control; they are easy to predict yet only weakly connected to the controllable degrees of freedom that a policy must master.
2. **Real-world videos amplify noisy motion.** This mismatch becomes especially pronounced on human videos and in-the-wild footage (33; 34), where camera motion and non-causal background changes can be stronger than interaction-induced state changes. Frame-difference-based latent-action objectives are then incentivized to encode these dominant signals, turning the latent action into a delta-frame encoder of nuisance motion rather than a representation of meaningful transition dynamics.
3. **Information leakage makes *latent action* collapse into a shortcut.** Several latent-action pipelines model transitions by feeding both the current observation and future observation into the same module, or by allowing future context to influence the learned action variable during training (13; 79). This design creates an easy shortcut: the latent action can simply encode the future itself instead of capturing how state transitions should be explained (82). The resulting "action" becomes semantically empty—useful for matching training losses, but not a meaningful factor for control.
4. **Multi-stage training pipelines are complex and fragile.** To stabilize training and mitigate the above issues, many approaches rely on three-stage (or more) procedures: representation pretraining, latent-action learning/alignment, then policy learning (20; 28; 79). These pipelines increase engineering complexity, introduce stage-wise inconsistencies, and make it harder to train and evaluate methods cleanly.

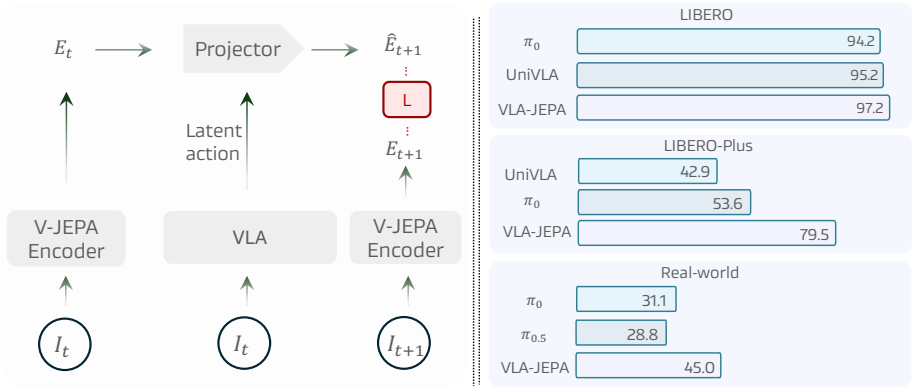


Fig. 1: Left: VLA-JEPA model architecture; Right: performance on LIBERO, LIBERO-Plus, and real-world benchmarks, compared with previous SOTA methods.

Together, these issues point to a single underlying problem: many latent-action objectives remain implicitly anchored to pixel variation, and thus learn dynamics that are predictive but not necessarily action-centric (82). For embodied control, we want a *value-bearing latent state* that discards nuisance appearance while preserving factors governing state evolution under interaction (34; 37; 74). This motivates a key principle: *predict future latent states that reflect action-relevant transition structure, while preventing future information from leaking into the predictor.*

This aligns naturally with JEPA (Joint-Embedding Predictive Architectures) (1; 2; 4; 17; 24), which replaces pixel reconstruction with latent-space alignment. By predicting representations rather than pixels, JEPA is inherently robust to low-level noise and encourages semantic abstraction (31).

Motivated by this perspective, we introduce VLA-JEPA, a JEPA-style pretraining framework tailored to VLA policies. As shown in Figure 1, the key design is *leakage-free state prediction*: during pretraining, a target video encoder produces latent targets from future context (e.g., a short clip), while the latent action pathway receives only the current observation through a VLM backbone. Compared with V-JEPA2-AC (2), a VLM backbone leverages world knowledge from large-scale text-annotated data, while direct action prediction eliminates the need for energy-minimization-based planning, which relies on a goal image to generate actions. Furthermore, a predictor maps the history latent states and the latent action representations to future latent states, which is trained as a latent world model using a JEPA alignment loss. Crucially, future frames are never provided as inputs to the VLM backbone; instead, they are used solely to construct the training targets, eliminating the shortcut that causes latent-action collapse. This yields two benefits: (1) semantic robustness to camera motion and background changes, since supervision operates in latent space rather than pixel space; (2) a streamlined two-stage pipeline—JEPA pretraining followed by

action-head fine-tuning—without introducing auxiliary modules or redefining the learned representations. This work makes three contributions:

- **Analysis of latent-action pretraining pitfalls.** We analyze why many future-supervision objectives (including compressed frame-difference latent actions) remain pixel-tethered, becoming biased toward appearance, vulnerable to nuisance motion in real-world videos, and prone to information leakage when future context enters the learner.
- **VLA-JEPA: leakage-free, state-level JEPA pretraining.** We propose a JEPA-style latent predictive alignment scheme that learns action-relevant transition semantics by predicting and aligning future latent states—without pixel reconstruction, information leakage and only one-stage pretraining pipeline.
- **Improved and robustness with a simpler workflow.** Across embodied control benchmarks (LIBERO, LIBERO-Plus, SimplerEnv) and real-world settings, VLA-JEPA yields consistent gains in robustness, and generalization, while simplifying training relative to prior multi-stage latent-action pipelines.

2 Related Works

2.1 Vision-Language-Action Models

With the rapid advancement of Large Language Models (LLMs) (6; 25; 38; 51; 62; 77) and large-scale robot datasets (23; 26; 39; 60), Vision-Language-Action (VLA) models have become a dominant paradigm in robot learning (9; 41; 43; 76). The RT series (5; 10; 92) pioneered fine-tuning multimodal LLMs on robot demonstrations, and subsequent works further improved manipulation and navigation performance (9; 41; 46; 64; 75; 84; 85). However, most VLA methods rely heavily on large-scale action-labeled robot data, which is costly and difficult to scale (72). To reduce dependence on explicit action supervision, recent works introduce multimodal Chain-of-Thought signals, such as hierarchical planning (5; 36; 48; 63; 91), subgoal or rollout prediction (18; 49; 54; 67; 73; 80; 87; 88), object-centric conditioning (15; 23; 35; 55; 65; 83; 90) and latent future embeddings or actions (12; 53; 56; 86; 89). Nevertheless, these approaches suffer from relying on action-labeled data. In contrast, VLA-JEPA learns action-centric representations via latent predictive alignment, avoiding explicit future reconstruction and reducing the need for large-scale action supervision.

2.2 Latent action learning for robotics

To leverage large-scale videos without action labels, ILPO (27), LAPO (66) and Genie (11) propose using latent action for video games. For robot learning, LAPA (79), IGOR (19), UniVLA (13), MotoGPT (21), Adaworld (30), CoMo (78) and StaMo (52) extract discrete or continuous motion tokens from frame transitions, and pretrain VLA to predict these latent actions before mapping them to real robot controls. To align the latent action to the real action space, villa-x (20), XR-1 (28), CLAP (81) and VITA (57) propose that extract latent action from both robot and human video and use a unified codebook. However,

since latent actions are often learned directly from adjacent frames, models may exploit pixel-level shortcuts and encode future-frame leakage. Although some methods attempt to mitigate the above issues by constraining the latent action space—through optical-flow (7; 14), or object-centric constraints (42), among others (58; 59)—such constraints inevitably bias the learned latent actions toward aligning with hand-crafted, visually defined priors rather than the true underlying controllable factors. Consequently, when such priors become mismatched in novel environments, the learned latent actions tend to fail systematically, and the latent spaces may align more with visual deltas than with actionable control signals, necessitating multi-stage training pipelines and additional alignment mechanisms (82). On the contrary, our proposed VLA-JEPA learns action-relevant representations without relying on delta frame info extraction, avoiding leakage and pixel shortcuts while enabling single-stage end-to-end pretraining.

3 Methodology

To address the limitations of existing approaches, we introduce VLA-JEPA, a unified framework that enables joint pre-training on both action-free and action-labeled data. For action-free human videos, VLA-JEPA extracts latent actions from vision-language prior representations by optimizing a world-model-based state transition objective. Building upon this foundation, we further integrate a flow-matching-based action generator for robot demonstrations to support precise end-effector trajectory generation. During fine-tuning, VLA-JEPA enables end-to-end fusion of the two objectives, allowing the learned state-transition dynamics to be effectively leveraged for downstream robotic control.

3.1 Model Backbone

As shown in Figure 2, we adopt Qwen3-VL (3) as the core large vision-language model (VLM) in our framework. This VLM is built upon the Qwen3 (77) and uses SigLIP-2 (68) as its vision encoder. It is well established that the world knowledge acquired by VLMs during large-scale pretraining, including image understanding and key object detection, can be transferred to robot control tasks. To distill continuous latent action representations from the VLM, which encode state transition information for world modeling, we introduce a set of learnable tokens that denoted as $\langle latent_i \rangle$ and $\langle action \rangle$, where i denotes the time step. For instance, $\langle latent_0 \rangle$ represents the state transition between s_0 and s_1. To model state transitions in continuous videos, we encode pixel-level human videos into time-step–aware feature sequences using an encoder architecture trained with V-JEPA2 (2), and employ a time-causal attention mechanism to capture the correlations between the feature sequences and the latent actions.

3.2 Learning from Human Videos

To enable VLM to learn from human demonstration videos, we design a training framework that explicitly injects environment dynamics into the latent action

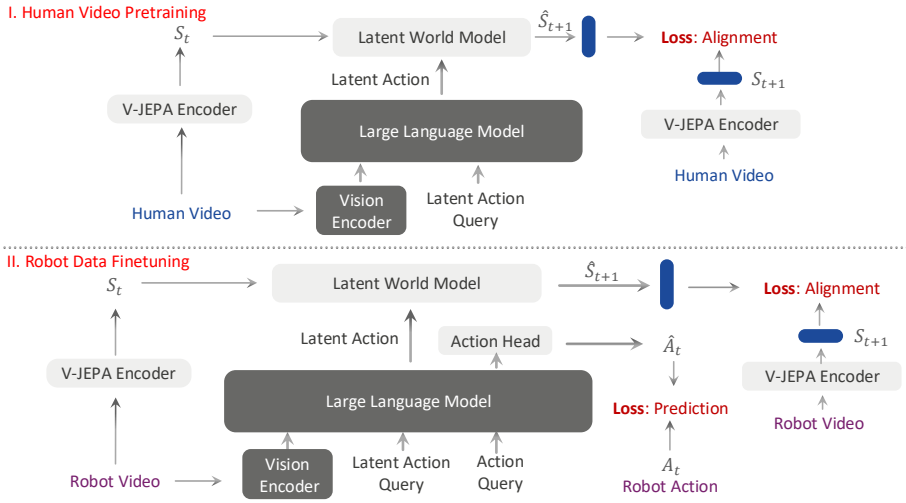


Fig. 2: VLA-JEPA supports cross-domain training on both human videos and robot data, where human videos are trained using an alignment loss under the latent world modeling objective, while robot data are trained with a joint objective consisting of an alignment loss and a robot action prediction loss.

tokens. Specifically, we consider a human video dataset $D = \{(O_0, O_1, ..., O_v, \ell)\}$, where ℓ denotes the associated language instruction, and each O_v represents a video captured from viewpoint v. Formally, $O_v = (I_{v,t_0}, I_{v,t_1}, ..., I_{v,t_n})$, where I_{v,t_i} denotes the video frame at time step t_i from view v, and n is the total number of frames within a video.

World State Encoder. Unlike traditional single-view video representation approaches, we encode diverse observations from the same episode into a unified world state representation through a world state encoder. Specifically, we adopt a self-supervised V-JEPA2 encoder as the single-view video state representation and integrate representations from multiple viewpoints via a concatenation operator. The encoding process for each view and the subsequent aggregation across views are formulated as follows:

$$s_{t_i} = \|_v F(I_{v,t_i}) \tag{1}$$

where $F(\cdot)$ is the V-JEPA2 single-view video encoder, and $\|$ denotes the vector concatenation operator. The resulting s_{t_i} is the unified world-state representation.

Latent Action Pretraining through World Modeling. To encourage learnable latent action tokens to model state dynamics, we introduce a state prediction objective based on an auto-regressive transformer-based world model.

Formally, the VLM takes as input the multi-view observations at the initial time step t_0, together with the language instruction ℓ. Conditioned on these inputs, the VLM maps a set of special learnable tokens $\langle \text{latent}_i \rangle$ into latent representations that summarize the underlying world dynamics:

$$z_{t_i} = p_\theta^{VLM}\left(\langle \text{latent}_i \rangle \mid \{I_{j,t_0}\}_{j=0}^{v}, \ell\right), \tag{2}$$

where z_{t_i} denotes the latent action representation at time step t_i.

Subsequently, the unified representation z_{t_i} is used to condition the world model, providing additional context for state prediction. Formally, given the sequence of encoded world states $s_{t_{0:i}}$ and the corresponding conditioning variables $z_{t_{0:i}}$, the world model predicts the next chunk of states as:

$$\hat{s_{t_{1:i+1}}} = p_\theta^{WM}(s_{t_{0:i}}, z_{t_{0:i}}), \tag{3}$$

where $\hat{s_{t_{1:i+1}}}$ is the predicted world state chunk over $[t_1, t_{i+1}]$.

In practice, each special latent token $\langle \text{latent}_i \rangle$ is replicated K times in the input sequence to enable variable-length latent action encoding, where K is a tunable hyperparameter. The world model employs a time-causal attention mechanism. Within each time step, all latent action tokens and world state tokens attend to each other via bidirectional full attention. Across different time steps, attention is strictly causal: tokens at time step t are allowed to attend only to tokens from time steps up to and including t, while attention to future time steps is masked out.

Finally, we optimize the combined WM and VLM using a teacher-forcing objective. This enables the unified representation z_{t_i} informed by the world knowledge encoded in the VLM, in order to effectively characterize world state transitions. The world modeling loss is defined as:

$$\mathcal{L}_{\text{WM}} = \sum_{k=1}^{T} \mathbb{E}_{s_{t_k} \sim F(\cdot)}(\hat{s}_{t_k} - s_{t_k}), \tag{4}$$

where s_{t_k} denotes the ground-truth world state at time t_k, $\hat{s}_{t_k}$ is the corresponding prediction, and T is the video prediction horizon.

3.3 Action Prediction with Joint Optimization Objectives

Action Token Conditioning. To leverage the latent action representations learned from video data for guiding action prediction, we design joint optimization objectives. Specifically, for multi-view RGB videos in the robot dataset, we adopt the same training objective as in Equation 4 to fine-tune the latent action representations within the robot data domain. For practical action prediction, we intend the latent action to serve as a conditioning signal for embodied action generation, analogous to the roles of the initial image observation I_{v,t_0} and the language instruction. Therefore, we append a set of learnable embodied action tokens $\langle \text{action} \rangle$ after the latent action tokens. Leveraging the causal attention mechanism of the VLM, the model captures the dependencies among $\langle \text{action} \rangle$, the latent action tokens, the initial visual observation, and the language instruction.

Formally, given the visual tokens $\{I_{i,t_0}\}_{i=0}^{v}$ at the initial time step t_0, the language instruction l, and a sequence of latent action tokens $\langle \text{latent}_i \rangle$, we obtain a global action-conditioning representation

$$z_a = p_\theta^{VLM}(\langle \text{action} \rangle \mid \{I_{i,t_0}\}_{i=0}^{v}, \ell, \langle \text{latent}_i \rangle), \tag{5}$$

where z_a serves as an additional conditioning signal for the action head.

Conditional Flow-Matching Action Head. We adopt conditional flow matching to model a distribution over continuous action trajectories. Specifically, let $a_{0:H}$ denote the ground-truth action sequence over a horizon H, and $\hat{a}_{0:H}$ denote the predicted action sequence generated by the learned flow. Following standard flow-matching formulations, we define a time-dependent interpolation

$$a_t = (1-t)\,\epsilon + t\,a_{0:H}, \quad t \sim \mathcal{U}(0,1), \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise. The action head parameterizes a vector field $v_\theta(a_t, t \mid z_a)$, conditioned on z_a, which is trained to match the ground-truth conditional flow. The flow-matching objective is given by

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{a_{0:H},\,\epsilon,\,t}\Big[\big\|v_\theta(a_t, t \mid z_a) - (a_{0:H} - \epsilon)\big\|_2^2\Big], \tag{7}$$

where $v_\theta(\cdot)$ denotes the predicted velocity field, $(a_{0:H} - \epsilon)$ is the target velocity induced by the linear interpolation, and $\|\cdot\|_2$ is the ℓ_2 norm. At inference time, the learned vector field is integrated from noise to data space to obtain the predicted action trajectory $\hat{a}_{0:H}$, conditioned on the action tokens z_a.

The overall training objective for action-labeled robot data is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \beta \mathcal{L}_{\mathrm{WM}}, \tag{8}$$

where β is a tunable hyperparameter.

4 Experiments

To evaluate the generalization and robustness of VLA-JEPA, we perform comprehensive simulation experiments utilizing two distinct simulation environments and three public benchmarks, Simpler-Env, LIBERO, along with real-world experiments covering both in-distribution and out-of-distribution settings as shown in Fig. 3. Furthermore, we compare VLA-JEPA against the latest VLA baselines that support pretraining with human videos as well as VLA baselines pretrained solely on robot-collected data. For a fair comparison, all baselines are fine-tuned on the corresponding benchmark datasets, including LIBERO (50), BridgeV2 (70), and our self-collected real-world dataset.

We first pretrain VLA-JEPA on the large-scale human hand–object interaction dataset Something-Something-V2 (32) and the large-scale action-labeled robot dataset DROID (39). We then fine-tune VLA-JEPA on each simulation and real-world environment using its environment-specific dataset. All experiments are conducted on 8 NVIDIA A100 GPUs. Detailed implementations are provided in the supplementary materials.

4.1 Simulation Setup and Baseline

Benchmarks. We conduct generalization experiments on the LIBERO (50) and SimplerEnv (45) benchmarks. The LIBERO benchmark employs the Franka

Emika Panda arm and comprises four task suites designed to facilitate research on lifelong learning in robotic manipulation. SimplerEnv features WidowX and Google Robot setups, offering diverse manipulation scenarios with varying lighting, colors, textures, and robot camera poses, thereby bridging the visual appearance gap between real and simulated environments. Furthermore, we conduct robustness experiments on LIBERO-Plus (29), a large-scale benchmark designed to systematically stress-test VLAs through perturbed tasks across seven dimensions.

The three simulation environments correspond to (i) in-distribution scenarios in which policies, trained using simulated expert data, are validated on in-distribution tasks in simulation (LIBERO), (ii) out-of-distribution scenarios involving a real-to-sim gap, in which policies are trained using real-world data and evaluated in simulation (SimplerEnv), and (iii) out-of-distribution scenarios in which policies, trained using simulated expert data, are validated on out-of-distribution tasks in simulation (LIBERO-Plus).

Baselines. We compare VLA-JEPA with previous latent-action VLAs, future-prediction VLAs and state-of-the-art VLAs including Moto (21), LAPA (79), UniVLA (13), villa-X (20), CoT-VLA (87), WorldVLA (16), RoboVLMs (44), GR00T N1 (8), OpenVLA-OFT (40), DreamVLA (86), F1 (55), π_0 (9), π_0-`Fast` (61), and $\pi_{0.5}$ (35).

Table 1: Comparison on the LIBERO benchmark. We report the task success rate for each suite and the average across all tasks. Each task is evaluated over 50 episodes (500 episodes per suite), and success rates are computed over these trials. **Bold** denotes the best performance, and italics denotes the second best.

Method	LIBERO				
	Spatial	Object	Goal	LIBERO-10	Avg
LAPA (79)	73.8	74.6	58.8	55.4	65.7
UniVLA (13)	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT (40)	*97.6*	98.4	*97.9*	*94.5*	*97.1*
π_0 (9)	96.8	*98.8*	95.8	85.2	94.2
π_0-`Fast` (61)	96.4	96.8	88.6	60.2	85.5
CoT-VLA (87)	87.5	91.6	87.6	69.0	81.1
WorldVLA (16)	87.6	96.2	83.4	60.0	81.8
villa-X (20)	97.5	97.0	91.5	74.5	90.1
GR00T N1 (8)	94.4	97.6	93.0	90.6	93.9
DreamVLA (86)	97.5	94.0	89.5	89.5	92.6
F1 (55)	98.2	97.8	95.4	91.3	95.7
$\pi_{0.5}$ (35)	**98.8**	98.2	**98.0**	92.4	96.9
QwenVLA	95.0	99.4	96.8	92.6	96.0
+ ViT	94.2	**99.6**	97.0	94.6	96.4
+ VQGAN	91.8	99.0	96.2	87.6	93.7
+ LAM	94.8	99.2	96.4	93.0	95.9
VLA-JEPA	96.2	**99.6**	97.2	**95.8**	**97.2**
w/o human video	94.8	**99.6**	95.8	94.0	96.1

In addition, we construct backbone-controlled VLA baselines, termed QwenVLA, by equipping Qwen3-VL-2B with a flow-matching-based action head. We further evaluate three QwenVLA variants that differ in their prediction targets: (1) **"+ViT"**, which predicts continuous future representations encoded by the Qwen3-VL ViT; (2) **"+VQGAN"**, which predicts discrete future tokens encoded by VQGAN; and (3) **"+LAM"**, which predicts latent actions encoded by a latent action model. All QwenVLA variants are trained using the same data split as VLA-JEPA. Since QwenVLA does not support training on human videos, it only uses the robot data adopted by VLA-JEPA.

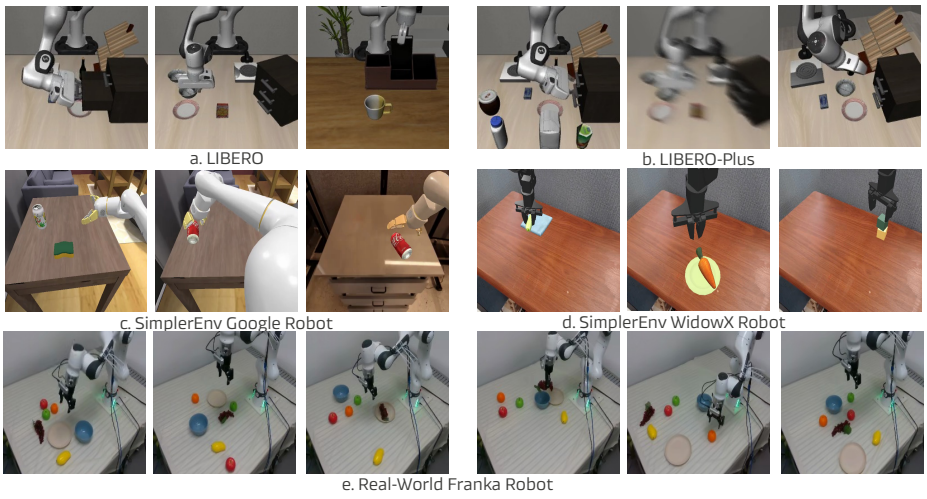


Fig. 3: Experiments setup on LIBERO, LIBERO-Plus, SimplerEnv and real-world Franka Robot.

4.2 Simulation Evaluation

LIBERO. As presented in Table 1, VLA-JEPA achieves state-of-the-art performance on 2 out of the 4 task suites within the LIBERO benchmark, securing the highest average success rate overall. We observe that other top-performing models on LIBERO, such as OpenVLA-OFT and $\pi_{0.5}$, rely on extensive robot datasets for pre-training. In contrast, VLA-JEPA achieves better performance using less training data. Furthermore, previous latent-action-based VLAs and VLAs trained on human videos, such as UniVLA, villa-X, LAPA, and CoT-VLA, consistently underperform VLA-JEPA, which corroborates the limitations of such approaches discussed in Section 1.

SimplerEnv. Table 2 reports the results on SimplerEnv. VLA-JEPA achieves the best performance both on the Google Robot and the WidowX Robot. Since SimplerEnv does not provide expert demonstrations, we observe that the quality of robot data plays a crucial role in performance on this benchmark. For example, LAPA extracts only successful rollouts in SimplerEnv and uses them as expert

Table 2: Comparison on the SimplerEnv benchmark. We report the average success rate for each individual task under the visual matching setting in both SimplerEnv-WidowX Robot and SimplerEnv-Google Robot. * denotes that the method is trained on in-distribution expert demonstrations collected in simulation environment. All other methods are trained on subsets of the OXE dataset.

Method	Google Robot					WidowX Robot				
	Pick	Move	Drawer	Place	Avg	Spoon	Carrot	Block	Eggplant	Avg
LAPA* (79)	-	-	-	-	-	70.8	45.8	**54.2**	58.3	**57.3**
villa-x (20)	81.7	55.4	38.4	4.2	44.9	48.3	24.2	19.2	71.7	40.8
UniVLA (13)	-	-	-	-	-	-	-	-	-	42.7
RoboVLMs (44)	77.3	61.7	43.5	24.1	51.7	45.8	20.8	4.2	**79.2**	37.5
GR00T N1 (8)	0.7	1.9	2.9	0.0	1.4	1.4	0.0	0.0	13.9	3.8
MoTo (21)	74.0	60.4	43.1	-	-	-	-	-	-	-
OpenVLA-OFT (40)	-	-	-	-	-	34.2	30.0	30.0	72.5	41.8
π_0 (9)	72.7	65.3	38.3	-	-	29.1	0	16.6	62.5	40.1
π_0-Fast (61)	75.3	67.5	42.9	-	-	29.1	21.9	10.8	66.7	48.3
VLA-JEPA	**88.3**	64.1	59.3	49.1	65.2	**75.0**	**70.8**	12.5	70.8	**57.3**
w/o human video	85.3	66.7	**75.5**	**86.1**	**78.4**	**75.0**	54.2	20.8	**79.2**	**57.3**

demonstrations for training, and achieves the second-highest success rate on the WidowX Robot with merely 100 rollouts. This is mainly because training on successful rollouts effectively mitigates the real-to-sim gap. In comparison, VLA-JEPA achieves the strongest performance among methods trained on subsets of the OXE dataset.

Table 3: Comparison on the LIBERO-Plus benchmark. We report the average success rate across each perturbation dimension, where each perturbation includes the four task suites from the original LIBERO benchmark.

Method	LIBERO-Plus							
	Camera	Robot	Language	Light	Background	Noise	Layout	Total
UniVLA (13)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
OpenVLA-OFT (40)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0 (9)	13.8	6.0	58.8	85.0	81.4	**79.0**	68.9	53.6
π_0-Fast (61)	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	61.6
WorldVLA (16)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
QwenVLA	53.0	62.5	82.8	**96.8**	90.1	40.0	77.2	69.6
+ViT	60.3	64.5	81.4	91.8	90.3	48.0	83.8	72.6
+VQGAN	12.8	49.8	70.8	68.2	47.9	37.1	72.6	50.5
VLA-JEPA	63.3	**67.1**	**85.4**	95.6	**93.6**	66.3	**85.1**	**78.1**
w/o human video	40.3	55.7	72.9	88.2	70.5	38.2	74.6	62.9

LIBERO-Plus. As shown in Table 3, VLA-JEPA achieves the best performance on 5 out of 7 perturbations in LIBERO-Plus. It demonstrates a substantial advantage over UniVLA, which is trained on human videos, and OpenVLA-OFT, π_0, and other methods trained on large amounts of robot data. This indicates that latent actions, learned through our pretraining approach, possess a level of world knowledge representation comparable to that encoded by text data. In particular, we find that VLA-JEPA demonstrates a significant advantage over all baselines under perturbations in Language, Light, Background, and Layout, which verifies that our latent action can effectively handle task-agnostic disturbances, thereby achieving a more robust and generalized policy.

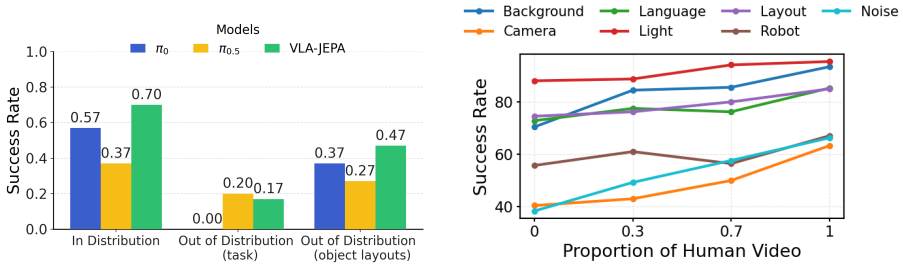
4.3 Real-world Experiments

For real-world experiments, we designed table-top manipulation tasks using a Franka Research 3 arm equipped with a Robotiq 2F-85 gripper. We collected 100 human demonstration trajectories for training, which include 3 picking and placing tasks. In alignment with the simulation experiments and extending beyond in-distribution task evaluation, we introduce two OOD experimental protocols to rigorously assess the model's generalization capability and robustness for physical deployment. The first OOD protocol involves executing tasks not present in the training data, validating the model's ability to acquire and transfer fundamental skills. The second protocol requires executing tasks seen during training but with randomized object layouts, thereby emulating the cluttered scenarios typical of real-world table-top manipulation tasks. For fair comparison, we finetune π_0, $\pi_{0.5}$, and VLA-JEPA on collected demonstration datasets.

As shown in Figure 4a, VLA-JEPA achieves state-of-the-art performance under both the ID and the object layout OOD settings, demonstrating a clear advantage over both π_0 and $\pi_{0.5}$. In the task OOD setting, it attains the second-best result. During real-world deployment, we observed that the generalization capability of VLA-JEPA is less robust than that of $\pi_{0.5}$, yet it produces execution trajectories that are significantly more stable and reliable. Specifically, $\pi_{0.5}$ more accurately follows instructions to contact the target object compared to VLA-JEPA. However, its position motion controller frequently violates the safety boundaries of the robot arm, leading to execution failures. In contrast, VLA-JEPA rarely breaches the robot arm's safety constraints. In addition, VLA-JEPA acquires the skill of repeated grasping, i.e., reopening the gripper to attempt another grasp after a failure, which is not observed in π_0 or $\pi_{0.5}$. We attribute this to the abundance of repeated-grasping knowledge present in human videos, whereas this skill does not require additional physical dynamics knowledge. In contrast, robot data rarely contain demonstrations specifically designed for repeated grasping. More details are provided in the supplementary.

4.4 Further Analysis and Ablation Study

In this section, we investigate the following questions under the multi-view setting to thoroughly evaluate the ability of our model:



(a) Real-world experimental results.

(b) Effect of the proportion of human video data in pre-training on success rates across different perturbation dimensions on the LIBERO-Plus.

Question 1. **Does JEPA-style latent world modeling improve VLA performance?** By comparing QwenVLA with VLA-JEPA in Tables 1 and 3, we show that the performance gains stem from the JEPA-style world modeling objective rather than the choice of model backbone. Moreover, comparing the "+ViT" and "+VQGAN" variants with VLA-JEPA shows that the JEPA-style objective is more effective than pixel-level future prediction. Similarly, the comparison between the "+LAM" variant and VLA-JEPA indicates that, when controlling for the model backbone, information leakage in LAM-based latent actions can adversely affect downstream VLA performance, a limitation that VLA-JEPA effectively mitigates.

Question 2. **What impact does human video have?** Tables 1, 2 show that removing human video pretraining does not yield a significant performance drop on either LIBERO or SimplerEnv. This further indicates that, for both ID settings and OOD scenarios with a real-to-sim gap, high-quality expert demonstrations are more critical to model performance than action-free human videos. Table 3 present that action-free human videos provide substantial performance gains on the LIBERO-Plus benchmark.

Our analysis suggests that human videos lack actionable information about action trajectories, which prevents VLA models from directly learning the physical dynamics of robotic actions. Instead, the primary benefit lies in enhancing the robustness and stability of the model's pre-existing skills. This process closely resembles human skill acquisition from observation: when a person learns a skill by watching a video, the first attempt is unlikely to succeed, and proficiency is achieved only after repeated trials that establish the correspondence between the observed video knowledge and the underlying physical dynamics, ultimately leading to a stable and reliable policy.

Figure 4b illustrates how the success rates across different perturbation dimensions on the LIBERO-Plus benchmark vary as the proportion of human videos in the pre-training data increases. These results further corroborate our hypothesis that human video data primarily enhances the robustness and stability of the VLA model by strengthening its existing skill repertoire, rather than

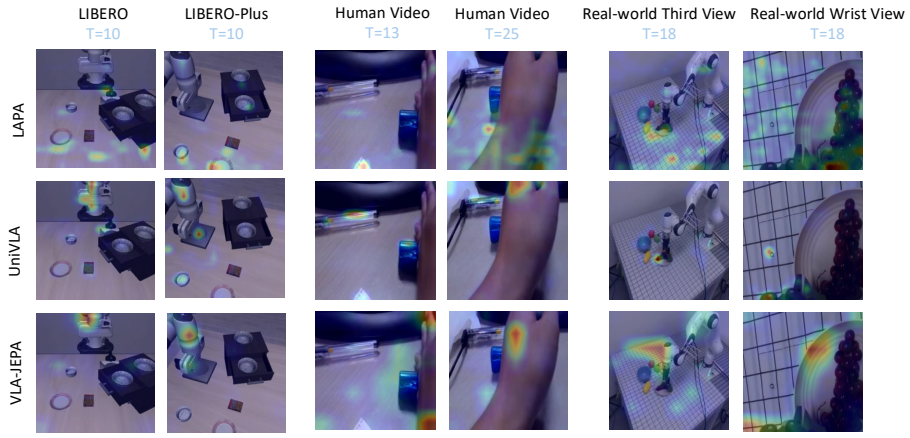


Fig. 5: Visualization of the attention weight matrix of latent action tokens attending to image tokens.

introducing new action execution capabilities. Moreover, as the scale of human video data increases, the robustness of the resulting policy consistently improves.

Question 3. **What is the impact of unified pretraining?** According to Section 4.2, the experimental results demonstrate that our unified pretraining method consistently outperforms the previous two-stage pretraining paradigm. In addition, to more clearly analyze the information captured by latent action tokens, we visualize the attention weights from latent action tokens to image tokens in the internal attention maps of three VLAs—LAPA, UniVLA, and VLA-JEPA—under simulated, human-video, and real-world robot image inputs. For a fair comparison, all methods are evaluated using pretrained-only checkpoints, without any fine-tuning on either simulated or real data.

As shown in Figure 5, the latent actions of LAPA focus on dense visual information, resulting in the inclusion of too many operation-irrelevant details, such as unrelated objects on the desktop. We attribute this to the leakage of information during the pretraining stage, leading to the degradation of latent actions into compressed representations of the target images, a point further analyzed in Section 1. In contrast, UniVLA alleviates this issue via task-relevant textual guidance, yet its overemphasis on semantics causes attention to operation-irrelevant background elements in the images, such as the stationary pen in human videos or the texture of the tablecloth in real-world wrist-view recordings. By comparison, VLA-JEPA focuses more precisely on the operation, for instance, the robotic arm, the hand, and the objects to be manipulated. This shows that the unified pretraining approach simplifies the training pipeline while improving training effectiveness by reducing the impact of task-irrelevant information.

Question 4. **How does the number of video horizon affect performance?** We vary the number of future video horizon $T \in \{4, 8, 16\}$ to investigate the

impact of dynamic information on the performance. For a fair comparison, all models are directly fine-tuned on the LIBERO dataset, with all other hyperparameters kept the same in the comparison. As shown in Table 4, the model achieves its best when the video horizon is close to the predefined action horizon 8, which indicates that latent actions are effective for embodied action generation. When T is too small, the encoded information is insufficient, leading to inferior performance, particularly on long-horizon tasks. In contrast, an excessively large T introduces redundant information. It yields the best results on the goal-oriented task suite with relatively simple objectives, but performs worst on the spatial task suite that requires fine-grained manipulation.

Table 4: Comparison of VLA-JEPA on the LIBERO benchmark with different numbers of future video horizon.

T	Spatial	Object	Goal	LIBERO-10	Avg
4	**95.0**	99.2	95.8	89.0	94.8
8	94.8	**99.8**	95.8	**94.0**	**96.1**
16	92.8	98.8	**98.0**	92.2	95.5

5 Conclusion

This paper proposes VLA-JEPA, a unified pretraining framework that learns latent actions from both human and robot videos via latent world modeling. Extensive experiments and analyses reveal that existing latent action pretraining methods suffer from information leakage and representation degeneration, which hinder the learning of meaningful temporal dynamics. In contrast, VLA-JEPA effectively mitigates these issues and enables the model to capture genuine inter-frame dynamics. As a result, VLA-JEPA achieves competitive performance in both simulation benchmarks and real-world robotic experiments. We further believe that the human-video pretraining paradigm introduced by VLA-JEPA is highly scalable, and can be naturally extended by incorporating robot data and text-based reasoning data, thereby further improving the generalization and robustness of VLA models.

Acknowledgements

This work is supported by the Zhongguancun Academy under Grant No.s C20250302.

Bibliography

[1] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture (2023)

[2] 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)

[3] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. CoRR **abs/2511.21631** (2025)

[4] Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: V-jepa: Latent video prediction for visual representation learning (2023)

[5] Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D.: Rt-h: Action hierarchies using language. arXiv preprint (2024)

[6] Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint (2024)

[7] Bi, H., Tan, H., Xie, S., Wang, Z., Huang, S., Liu, H., Zhao, R., Feng, Y., Xiang, C., Rong, Y., et al.: Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030 (2025)

[8] Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint (2025)

[9] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi0: A vision-language-action flow model for general robot control. arXiv preprint (2024)

[10] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M.S., Salazar, G., Sanketi, P.R., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H.T., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: RT-1:

robotics transformer for real-world control at scale. In: Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023 (2023)
[11] Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
[12] Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Huang, X., Jiang, S., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint (2025)
[13] 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. arXiv preprint (2025)
[14] Bu, X., Lyu, J., Sun, F., Yang, R., Ma, Z., Li, W.: Laof: Robust latent action learning with optical flow constraints. arXiv preprint arXiv:2511.16407 (2025)
[15] Cai, J., Cai, Z., Cao, J., Chen, Y., He, Z., Jiang, L., Li, H., Li, H., Li, Y., Liu, Y., et al.: Internvla-a1: Unifying understanding, generation and action for robotic manipulation. arXiv preprint arXiv:2601.02456 (2026)
[16] Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint (2025)
[17] Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., Fung, P.: VL-JEPA: joint embedding predictive architecture for vision-language. CoRR **abs/2512.10942** (2025)
[18] Chen, J., Song, W., Ding, P., Zhou, Z., Zhao, H., Tang, F., Wang, D., Li, H.: Unified diffusion vla: Vision-language-action model via joint discrete denoising diffusion process. arXiv preprint arXiv:2511.01718 (2025)
[19] 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)
[20] 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 (2025)
[21] 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 (2024)
[22] Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., Tedrake, R., Song, S.: Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In: Proceedings of Robotics: Science and Systems (RSS) (2024)
[23] Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Zhang, Z., Yang, T., Zhang, X., Cui, H., et al.: Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. arXiv preprint (2025)
[24] Destrade, M., Bounou, O., Lidec, Q.L., Ponce, J., LeCun, Y.: Value-guided action planning with jepa world models (2025)

[25] Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: DreamLLM: Synergistic multimodal comprehension and creation. In: ICLR (2024)
[26] Ebert, F., Yang, Y., Schmeckpeper, K., Bucher, B., Georgakis, G., Daniilidis, K., Finn, C., Levine, S.: Bridge data: Boosting generalization of robotic skills with cross-domain datasets. arXiv preprint (2021)
[27] Edwards, A., Sahni, H., Schroecker, Y., Isbell, C.: Imitating latent policies from observation. In: International conference on machine learning. pp. 1755–1763. PMLR (2019)
[28] Fan, S., Wu, K., Che, Z., Wang, X., Wu, D., Liao, F., Liu, N., Zhang, Y., Zhao, Z., Xu, Z., et al.: Xr-1: Towards versatile vision-language-action models via learning unified vision-motion representations. arXiv preprint arXiv:2511.02776 (2025)
[29] Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., Fu, J., Gong, J., Qiu, X.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626 (2025)
[30] Gao, S., Zhou, S., Du, Y., Zhang, J., Gan, C.: Adaworld: Learning adaptable world models with latent actions. In: Forty-second International Conference on Machine Learning
[31] Garrido, Q., Nagarajan, T., Terver, B., Ballas, N., LeCun, Y., Rabbat, M.: Learning latent action world models in the wild. arXiv preprint arXiv:2601.05230 (2026)
[32] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)
[33] 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: CVPR (2022)
[34] Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. arXiv preprint (2024)
[35] Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: pi0.5: a vision-language-action model with open-world generalization. arXiv preprint (2025)
[36] Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., Xue, X., Su, Q., Lyu, H., Zheng, X., Liu, J., Wang, Z., Zhang, S.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 1724–1734. Computer Vision Foundation / IEEE (2025)
[37] Jia, Y., Liu, J., Liu, S., Zhou, R., Yu, W., Yan, Y., Chi, X., Guo, Y., Shi, B., Zhang, S.: Video2act: A dual-system video diffusion policy with robotic spatio-motional modeling (2025)
[38] Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., Sadigh, D.: Prismatic vlms: Investigating the design space of visually-conditioned language models. arXiv preprint (2024)

[39] Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint (2024)
[40] Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint (2025)
[41] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint (2024)
[42] Klepach, A., Nikulin, A., Zisman, I., Tarasov, D., Derevyagin, A., Polubarov, A., Lyubaykin, N., Kurenkov, V.: Object-centric latent action learning. CoRR **abs/2502.09680** (2025)
[43] Li, H., Zuo, Y., Yu, J., Zhang, Y., Yang, Z., Zhang, K., Zhu, X., Zhang, Y., Chen, T., Cui, G., Wang, D., Luo, D., Fan, Y., Sun, Y., Zeng, J., Pang, J., Zhang, S., Wang, Y., Mu, Y., Zhou, B., Ding, N.: Simplevla-rl: Scaling VLA training via reinforcement learning. CoRR **abs/2509.09674** (2025)
[44] Li, X., Li, P., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Kong, T., Zhang, H., Liu, H.: Towards generalist robot policies: What matters in building vision-language-action models. arXiv preprint (2024)
[45] Li, X., Hsu, K., Gu, J., Mees, O., Pertsch, K., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., Xiao, T.: Evaluating real-world robot manipulation policies in simulation. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) CoRL (2024)
[46] Liang, Z., Li, Y., Yang, T., Wu, C., Mao, S., Pei, L., Yang, X., Pang, J., Mu, Y., Luo, P.: Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. arXiv preprint (2025)
[47] Lin, F., Hu, Y., Sheng, P., Wen, C., You, J., Gao, Y.: Data scaling laws in imitation learning for robotic manipulation. arXiv preprint (2024)
[48] Lin, F., Nai, R., Hu, Y., You, J., Zhao, J., Gao, Y.: Onetwovla: A unified vision-language-action model with adaptive reasoning. arXiv preprint (2025)
[49] Lin, M., Ding, P., Wang, S., Zhuang, Z., Liu, Y., Tong, X., Song, W., Lyu, S., Huang, S., Wang, D.: Hif-vla: Hindsight, insight and foresight through motion representation for vision-language-action models. arXiv preprint arXiv:2512.09928 (2025)
[50] Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: LIBERO: benchmarking knowledge transfer for lifelong robot learning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS (2023)
[51] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
[52] Liu, M., Shu, J., Chen, H., Li, Z., Zhao, C., Yang, J., Gao, S., Chen, H., Shen, C.: Stamo: Unsupervised learning of generalizable robot motion from compact state representation. arXiv preprint arXiv:2510.05057 (2025)
[53] Liu, Z., Liu, J., Chen, H., Guo, Z., Hou, C., Gu, C., Yu, J., Mi, X., Zhang, R., Che, Z., et al.: Last_{0}: Latent spatio-temporal chain-of-thought for robotic vision-language-action model. arXiv preprint arXiv:2601.05248 (2026)
[54] Liu, Z., Liu, J., Xu, J., Han, N., Gu, C., Chen, H., Zhou, K., Zhang, R., Hsieh, K.C., Wu, K., et al.: Mla: A multisensory language-action model for

multimodal understanding and forecasting in robotic manipulation. arXiv preprint arXiv:2509.26642 (2025)
[55] 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. arXiv preprint arXiv:2509.06951 (2025)
[56] Lyu, J., Li, Z., Shi, X., Xu, C., Wang, Y., Wang, H.: Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. arXiv preprint (2025)
[57] Ma, X., Xing, L., Zhang, H., Li, W., Lu, S.: Unifying perception and action: A hybrid-modality pipeline with implicit visual chain-of-thought for robotic action generation. arXiv preprint arXiv:2511.19859 (2025)
[58] Nikulin, A., Zisman, I., Klepach, A., Tarasov, D., Derevyagin, A., Polubarov, A., Nikita, L., Kurenkov, V.: Vision-language models unlock task-centric latent actions (2026)
[59] Nikulin, A., Zisman, I., Tarasov, D., Lyubaykin, N., Polubarov, A., Kiselev, I., Kurenkov, V.: Latent action learning requires supervision in the presence of distractors. In: Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025. Proceedings of Machine Learning Research, vol. 267. PMLR / OpenReview.net (2025)
[60] O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint (2023)
[61] Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., Levine, S.: Fast: Efficient action tokenization for vision-language-action models. arXiv preprint (2025)
[62] Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. In: ECCV (2024)
[63] Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., et al.: Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. arXiv preprint (2025)
[64] Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint (2025)
[65] Ranasinghe, K., Li, X., Mata, C., Park, J., Ryoo, M.S.: Pixel motion as universal representation for robot control. arXiv preprint (2025)
[66] Schmidt, D., Jiang, M.: Learning to act without actions. In: The Twelfth International Conference on Learning Representations
[67] Tian, Y., Yang, S., Zeng, J., Wang, P., Lin, D., Dong, H., Pang, J.: Predictive inverse dynamics models are scalable learners for robotic manipulation. ICLR (2024)
[68] 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)

[69] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS (2017)
[70] Walke, H.R., Black, K., Zhao, T.Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A.W., Myers, V., Kim, M.J., Du, M., et al.: Bridgedata v2: A dataset for robot learning at scale. In: CoRL (2023)
[71] Wang, Q., Wu, M., Zhang, Y., Yuan, M., Zhang, W., You, H., Wang, Y., Jin, X., Yang, X., Zeng, W.: Goal-driven reward by video diffusion models for reinforcement learning. arXiv preprint arXiv:2512.00961 (2025)
[72] Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., et al.: Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. arXiv preprint arXiv:2509.09372 (2025)
[73] Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., Zhang, Z.: Unified vision-language-action model. arXiv preprint (2025)
[74] Wen, Y., Lin, J., Zhu, Y., Han, J., Xu, H., Zhao, S., Liang, X.: Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. NeurIPS (2024)
[75] Wu, W., Lu, F., Wang, Y., Yang, S., Liu, S., Wang, F., Zhu, Q., Sun, H., Wang, Y., Ma, S., et al.: A pragmatic vla foundation model. arXiv preprint arXiv:2601.18692 (2026)
[76] Xu, K., Zhu, Z., Chen, A., Zhao, S., Huang, Q., Yang, Y., Lu, H., Xiong, R., Tomizuka, M., Wang, Y.: Seeing to act, prompting to specify: A bayesian factorization of vision language action policy. arXiv preprint arXiv:2512.11218 (2025)
[77] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
[78] 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. arXiv preprint (2025)
[79] Ye, S., Jang, J., Jeon, B., Joo, S., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.W., Lin, B.Y., et al.: Latent action pretraining from videos. arXiv preprint (2024)
[80] Zhang, B., Song, N., Li, J., Zhu, X., Deng, J., Zhang, L.: Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. arXiv preprint arXiv:2510.11092 (2025)
[81] Zhang, C., Wang, J., Gao, Z., Su, Y., Dai, T., Zhou, C., Lu, J., Tang, Y.: Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos (2026)
[82] 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)
[83] Zhang, J., Guo, Y., Hu, Y., Chen, X., Zhu, X., Chen, J.: Up-vla: A unified understanding and prediction model for embodied agent. arXiv preprint (2025)
[84] Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint (2024)

[85] Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. Robotics: Science and Systems (2024)
[86] Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., et al.: Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. arXiv preprint (2025)
[87] Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. arXiv preprint (2025)
[88] Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3d-vla: A 3d vision-language-action generative world model. arXiv preprint (2024)
[89] Zhong, L., Liu, Y., Wei, Y., Xiong, Z., Yao, M., Liu, S., Ren, G.: Acot-vla: Action chain-of-thought for vision-language-action models. arXiv preprint arXiv:2601.11404 (2026)
[90] Zhong, Z., Yan, H., Li, J., Liu, X., Gong, X., Zhang, T., Song, W., Chen, J., Zheng, X., Wang, H., et al.: Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models. arXiv preprint arXiv:2508.18269 (2025)
[91] Zhou, Z., Zhu, Y., Zhu, M., Wen, J., Liu, N., Xu, Z., Meng, W., Cheng, R., Peng, Y., Shen, C., et al.: Chatvla: Unified multimodal understanding and robot control with vision-language-action model. arXiv preprint (2025)
[92] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H.T., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)