PAVXPLORERL: PHYSICAL-ACTION-VISUAL WORLD MODEL REINFORCEMENT LEARNING WITH ACTION EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Action-conditioned world models are a key component of embodied AI, serving as scalable policy evaluators that reduce reliance on expensive real-world rollouts. To accurately capture diverse action-induced dynamics, such models should satisfy three key objectives—Physical Plausibility (P), Action Adherence (A), and Visual Fidelity (V), collectively referred to as PAV—while remaining robust to both in-distribution (ID) expert demonstrations and out-of-distribution (OOD) actions. However, existing methods primarily rely on ID action–video pairs and pixel-level reconstruction losses, which do not explicitly optimize PAV objectives and generalize poorly beyond expert data. To address this, we propose **PAVXploreRL**, a reinforcement learning framework built on a pretrained latent world model that explicitly optimizes PAV objectives through reward-driven training. To improve action generalization, our method jointly leverages ID trajectories and noise-driven OOD action exploration, without paired video supervision. Experiments show that PAVXploreRL consistently outperforms pretrained baselines, achieving a 5.6% average gain across benchmarks and producing higher-quality PAV properties. As a policy evaluator, it also yields more reliable performance estimates and reduces the overestimation bias of prior expert-only world models such as Ctrl-World.

1 INTRODUCTION

With the advent of large-scale video generation models, controllable world models have emerged as a promising paradigm for embodied AI, predicting future observations from initial states and control signals to better emulate real-world dynamics. (Guo et al., 2025; Gao et al., 2026; Chen et al., 2026; Liao et al., 2025; Zhu et al., 2024; Quevedo et al., 2025). They can be broadly categorized into text-conditioned models, which function as action policies, and action-conditioned models, which directly take robot actions as input to simulate future observations under real-world dynamics. The latter are commonly used as virtual environments for policy evaluation and development, reducing the cost and risk of real-world deployment. This work focuses on action-conditioned world models.

Action-conditioned world models are expected to satisfy the Physical-Action-Visual (PAV) targets (Peng et al., 2026; Wang et al., 2026a). First, **Physical Plausibility**: they must generate physically plausible trajectories that respect real-world dynamics, including object permanence, consistent robot–object contacts, and valid rigid or deformable body behavior, as violations lead to unrealistic simulations. Second, **Action Adherence**: models should faithfully follow conditioning actions, as weak controllability undermines their reliability as policy evaluators. Third, **Visual Fidelity**: outputs should maintain high visual fidelity, preserving sufficient detail for human interpretation and downstream tasks. Most existing approaches, however, rely primarily on pixel-level similarity targets or non-adapted general VLM scores, providing limited explicit supervision for PAV targets.

Another limitation of current world models is their reliance on in-distribution (ID) action–video pairs collected from human experts, which predominantly consist of successful task-completion trajectories(Wu et al., 2026; Bardhan et al., 2026; Zhang et al., 2025c). As a result, current world models may overestimate the effectiveness of out-of-distribution (OOD) actions from suboptimal policies, inflating policy success rates. Reliable world models must generalize across diverse policies and

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

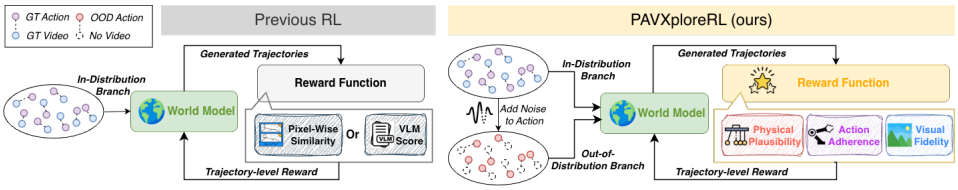


Figure 1: Comparison between **PAVXploreRL** and prior RL: unlike in-distribution-only expert-based approaches, it enables both in- and out-of-distribution action exploration and replaces pixel/VLM rewards with a domain-adapted latent world model for direct PAV objective evaluation.

accurately predict future outcomes under unseen actions; however, the absence of corresponding videos for such OOD actions limits their integration into standard training pipelines.

To address these challenges, we propose **PAVXploreRL** (Fig. 1, right), a reinforcement-learning method that enhances the PAV properties of action-conditioned world models via action exploration. Built upon the latent world model VJEPA-2 (Assran et al., 2025), which demonstrates strong physical understanding compared to other latent or vision-language models, we design three complementary rewards targeting a distinct PAV component: (1) a raw VJEPA reward for physical consistency, (2) an embodiment-focused VJEPA reward for accurate action execution, and (3) a visual fidelity reward perceptually realistic frames. Leveraging VJEPA-2's latent predictive capabilities, these rewards can be computed for both ID action–video pairs and OOD action generated via perturbations of ID actions with control noise, where paired videos are unavailable. To enable the above function, we retrain the action-conditioned VJEPA-2 on our robotic dataset using both raw and embodiment-focused frames, enabling multi-step autoregressive prediction over the whole action chunk.

To validate our approach, we build an action-conditioned world model by augmenting a state-of-the-art video diffusion model with action encoding and temporal history modeling. After supervised fine-tuning (SFT) on robotic datasets to obtain the pretrained baseline, we further post-train the model using our PAVXploreRL framework. Experiments yield consistent gains over the pretrained baseline, averaging 5.6% across all benchmarks. Qualitative results further indicate improved physical plausibility, action adherence, and visual realism. Experiments as policy evaluator also show reduced policy overestimation after our OOD action exploration. We summarize our contributions: **(1).** We propose a unified reward function that explicitly optimize the PAV targets-Physical Plausibility, Action Adherence, and Visual Fidelity. **(2).** We develop an RL framework that leverages both ID action–video pairs and OOD actions without paired video supervision. **(3).** We show that PAVXploreRL consistently improves pretrained world models, yielding more reliable policy evaluation.

2 Related Works

World Model: Recently, advances in large-scale video diffusion models (Agarwal et al., 2025b;a; Wan et al., 2025; Chi et al., 2025) have made generated videos increasingly realistic and controllable. Building on these, robotic world models have also progressed rapidly. According to the control signal, existing methods can be categorized into two groups. The first is text-conditioned models, which either implicitly (Jang et al., 2025; Bharadhwaj et al., 2024; G. et al., 2025) or explicitly (Du et al., 2024; Hu et al., 2024; Liang et al., 2024; Liao et al., 2025; Tan et al., 2025; Feng et al., 2025; Ye et al., 2025; Chen et al., 2026) function as action policy. Some recent works further unify policy learning and world modeling through co-training (Zhao et al., 2025; Li et al., 2025; Zhu et al., 2025; Guo et al., 2024; Gao et al., 2024; Zhang et al., 2025b; Zheng et al., 2025b; Zhong et al., 2025; Ye et al., 2026; Bi et al., 2026). The second is action-conditioned world models, which take robot actions as input and predict future observations consistent with real world interactions (Zhu et al., 2024; Bardhan et al., 2026; Guo et al., 2025; Gao et al., 2026; Zhang et al., 2025a; Team et al., 2026). These models can serve as policy evaluators. Moreover, integrating action-conditioned world models into policy training reduces reliance on real-robot rollouts and improves task success rates. (Guo et al., 2025; Zhang et al., 2025a; 2026b; Jiang et al., 2026; Sharma et al., 2026).

Physical-Action-Visual Targets Evaluation: Previous works have identified key criteria for world models, including physical realism, task completion, embodiment plausibility, and visual quality (Peng et al., 2026; Wang et al., 2026a). Building on this, we propose the Physical-Action-Visual

(PAV) targets for action-conditioned world models. **Physical Plausibility** is often assessed by pretrained VLMs (Peng et al., 2026; Chen et al., 2026), but their physical understanding is limited without domain-specific training. Alternatively, VJEPA—a latent world model—has demonstrated superior physical understanding compared to VLM-based methods (Bardes et al., 2024; Assran et al., 2025; Garrido et al., 2025) and improves video generation quality during both inference (Yuan et al., 2026) and training (Zhang et al., 2025c). To align with our task, we train an action-conditioned VJEPA-2 (Assran et al., 2025) on our robot data and use it to measure the physical plausibility. **Action adherence** is commonly evaluated using inverse dynamics models (IDMs) (Tian et al., 2025; Chi et al., 2025). However, IDMs infer actions from frame differences, inducing an inverse information flow relative to forward world models. To maintain consistency, we use the embodiment segmentation strategy of (Chi et al., 2025) to adapt the action-conditioned VJEPA-2 to explicitly measure alignment between robot motion and actions. **Visual fidelity** is commonly assessed via pixel-wise reconstruction loss. As our out-of-distribution (OOD) branch lacks paired videos, we instead design a reference-free reward integrating perceptual, statistical, and temporal metrics, representative metrics such as NIQE (Mittal et al., 2013), SSIM (Wang et al., 2004), and temporal consistency (Lai et al., 2018).

Reinforcement Learning of World Models. Prior RL approaches for world models, such as Flow-GRPO (Liu et al., 2026), rely on log-likelihood estimation and storing full denoising trajectories, resulting in high memory usage and prolonged training. More efficient methods—FPO (McAllister et al., 2025) and Diffusion-NFT (Zheng et al., 2025a)—avoid log-likelihood computation and have been integrated into recent world model training pipelines (Wang et al., 2026b; Bardhan et al., 2026). So our paper would adopt the Diffusion-NFT (Zheng et al., 2025a). However, previous action-conditioned world model RL still depends on ground-truth (GT) video for reward evaluation (Wu et al., 2026; Bardhan et al., 2026; Zhang et al., 2025c), limiting exploration of OOD actions. We overcome this by introducing a GT-free reward based on latent future predictions capability of our main reward evaluator, VJEPA-2, enabling RL with real action exploration including OOD actions.

3 METHOD

An overview of our framework is shown in Figure 2. We discuss our model architecture and training in Section 3.1, the reward function design in Section 3.2, and the VJEPA-2 training in Section 3.3.

3.1 WORLD MODEL ARCHITECTURE AND TRAINING

We define the world model π as a conditional video predictor. At step $t = i$, given the first frame I_0 of video, history frames $I_{t\in(i-2T,i]}$, and future actions $A_{t\in[i,i+T)}$, the model predicts future frames $\hat{I}_{t\in(i,i+T]}$. The policy is trained to minimize the discrepancy between predicted and real frames:

$$\pi\big(I_0, I_{t\in(i-2T,i]}, A_{t\in[i,i+T)}\big) = \hat{I}_{t\in(i,i+T]} \approx I_{t\in(i,i+T]} \tag{1}$$

We build our action-conditioned world model by adapting a pretrained video diffusion model with two modifications: incorporating action conditioning and maintaining a compact history representation for efficient long-horizon interaction. First, we introduce action conditioning by injecting action embeddings into the temporal pathway via adaptive layer normalization (AdaLN) and time embeddings, enabling action controllability. Second, to capture long-range temporal context, we adopt a Framepack (Zhang et al., 2026a) strategy that samples and compresses history frames in $0 \wedge (i - 2T, i]$ unevenly—distant frames are sampled densely and heavily compressed. The model is first trained with supervised fine-tuning (SFT) using a reconstruction MSE loss; upon convergence (no further gains on the validation set with additional training), it is further optimized via reinforcement learning (RL) by updating only the action encoder and LoRA parameters.

Our RL training pipeline is illustrated in the top of Figure 2. We adopt the DiffusionNFT (Zheng et al., 2025a) framework to efficiently fine-tune diffusion-based world models. During rollouts, we generate G groups of future frames using a first-order stochastic differential equation (SDE) solver conditioned on historical frames and future actions. The generated frames are evaluated with our reward function R, yielding a scalar trajectory-level reward $r \in [0, 1]$. For training, we maintain two model snapshots: (i) current policy π_θ and (ii) EMA-updated policy π_{old}. Implicit positive and negative denoising targets are constructed as:

$$\hat{\mathbf{x}}_0^+ = (1-\beta)\hat{\mathbf{x}}_0^{\text{old}} + \beta\hat{\mathbf{x}}_{0,\theta}, \qquad \hat{\mathbf{x}}_0^- = (1+\beta)\hat{\mathbf{x}}_0^{\text{old}} - \beta\hat{\mathbf{x}}_{0,\theta} \tag{2}$$

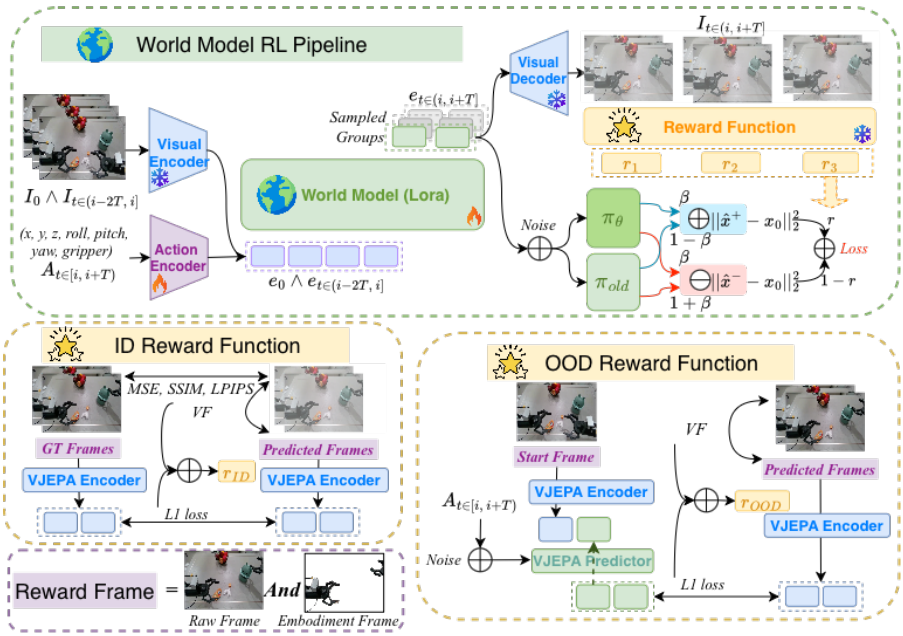


Figure 2: Overview of the **PAVXploreRL** framework. **World Model Reinforcement Learning Pipeline:** Given historical frames and actions, predicted future frames are evaluated by the reward function, and the resulting rewards weight the diffusion policy optimization loss. **In-distribution Reward Function:** For in-distribution ground-truth (GT) action–video pairs, rewards combine pixel-level metrics (MSE, SSIM, and LPIPS), VJEPA encoder-based metrics, and the proposed visual fidelity (VF) metrics. **Out-of-distribution Reward Function:** For out-of-distribution actions without GT videos, rewards are computed using VJEPA predictor-based metrics and VF metrics. **Reward Frame:** VJEPA-based rewards are assessed on both raw and embodiment-focused frames.

where $\hat{\mathbf{x}}_0^{\text{old}} \sim \pi_{\text{old}}$ and $\hat{\mathbf{x}}_{0,\theta} \sim \pi_\theta$. Positive targets encourage reward-aligned paths, while negative targets suppress deviations toward suboptimal paths. The training objective is a reward-weighted denoising loss:

$$\mathcal{L}(\theta) = \mathbb{E}\Big[r\,\|\hat{\mathbf{x}}_0^+ - \mathbf{x}_0\|_2^2 + (1-r)\,\|\hat{\mathbf{x}}_0^- - \mathbf{x}_0\|_2^2\Big] \tag{3}$$

where $x_0 = \hat{I}_{t\in(i,i+T]}$ is the predicted frames in rollout stage.

3.2 Reward Function Design

We define a unified reward function $\mathcal{R}$ that maps a predicted frame sequence to a scalar score $r \in [0,1]$, reflecting video generation quality. As illustrated in the bottom of Figure 2, the reward formulation depends on supervision availability, combining an in-distribution (ID) branch r_{ID} with ground-truth action–video pairs and an out-of-distribution (OOD) branch r_{OOD} with only perturbed actions without video pairs. The branch is selected via a Bernoulli-like sampling scheme:

$$r = \begin{cases} r_{\text{OOD}}, & \text{if } \epsilon \le \alpha_{\text{OOD}}, \\ r_{\text{ID}}, & \text{otherwise}, \end{cases} \qquad \epsilon \sim \text{Uniform}(0,1). \tag{4}$$

where α_{OOD} determines the proportion of OOD trajectories. To generate suboptimal actions without incurring the computational cost of policy inference during training, we construct OOD actions by applying three perturbations to non-static GT action segments exceeding a motion threshold, treating each hand and action type (position, rotation, and gripper) independently: **(1) Temporal shift:** segment boundaries (s, e) are randomly perturbed within ± 0.6 s; **(2) Structured magnitude bias:** chunk-wise multiplicative scaling induces consistent amplitude deviations of ± 0.15; and **(3) Smooth temporal noise:** Gaussian noise ($\sigma = 0.02$) is added at each timestep and temporally accumulated to produce smooth drift. Human inspection indicates that the resulting trajectories deviate substantially from GT actions.

The unified reward comprises five components, each targeting a different aspect of video quality:

Pixel-level Reward r_{pixel}: This reward measures image similarity between predicted and GT frames using MSE, SSIM, and LPIPS, providing low-level supervision via direct comparison. After group-level normalization, the three metrics (with MSE and LPIPS negated) are averaged to obtain the pixel-level reward:

$$r_{\text{pixel}} = \frac{1}{3}\left(r_{\text{MSE}} + r_{\text{SSIM}} + r_{\text{LPIPS}}\right) \tag{5}$$

Normalization formula to $[0, 1]$ is given as follows and applied to all forward reward normalization.

$$r_i^{\text{norm}} = 0.5 + 0.5 \cdot \text{clip}\left(\frac{r_i - \mu_G}{\sigma_G + \epsilon}, -1, 1\right), \quad \mu_G = \text{mean}_{j\in G}(r_j), \qquad \sigma_G = \text{std}_{j\in G}(r_j). \tag{6}$$

VJEPA Encoder Reward $r_{\text{VJEPA-E}}$: We leverage VJEPA-2's strong physical understanding capability to supervise high-level properties. For GT action-video pairs, we compute an encoder-based reward using the VJEPA encoder $E(\cdot)$ to measure the similarity between predicted and GT frames:

$$r_{\text{VJEPA-E}} = -\left\| E(\hat{I}_{t\in(i,i+T]}) - E(I_{t\in(i,i+T]}) \right\|_1 \tag{7}$$

VJEPA Predictor Reward $r_{\text{VJEPA-P}}$: To evaluate physical plausibility without GT frames, we use an action-conditioned VJEPA-2's predictor $P(\cdot)$, which predicts future latent representations from the start frame I_i and actions $A = A_{t\in[i,i+T)}$. We compute the predictor-based reward by comparing the world model predicted frames with the VJEPA-based latent prediction:

$$r_{\text{VJEPA-P}}(A) = -\left\| E(\hat{I}_{t\in(i,i+T]}) - P\big(E(I_i), A_{t\in[i,i+T)}\big) \right\|_1 \tag{8}$$

However, latent prediction reward tends to favor static but high-quality predictions. To mitigate this, we introduce a static-motion regularization term using a zero-action sequence A_0. Specifically, the latent representation predicted under A_0 should remain static and dissimilar to the world model generated videos. The final predictor-based reward is given by:

$$r_{\text{VJEPA-P}} = r_{\text{VJEPA-P}}(A) - \lambda_{\text{static}} r_{\text{VJEPA-P}}(A_0) \tag{9}$$

where $\lambda_{\text{static}} \in [0, 1]$ controls the strength of the static regularization.

Embodiment-focused VJEPA Reward $r^{\text{emb}}_{\text{VJEPA-E/P}}$: To better assess action adherence, both encoder- and predictor-based rewards are computed on two complementary visual streams: (1) raw frames and (2) embodiment-focused frames. While raw-frame rewards may be influenced by background dynamics and object interactions, embodiment-focused rewards emphasize robot motion consistency with executed actions. The final rewards are combinations of both streams:

$$r_{\text{VJEPA-E/P}} = r^{\text{raw}}_{\text{VJEPA-E/P}} + \lambda_{\text{emb}} r^{\text{emb}}_{\text{VJEPA-E/P}} \tag{10}$$

where $\lambda_{\text{emb}} \in [0, 1]$ controls the contribution of the embodiment-focused reward.

Visual Fidelity Reward r_{VF}: As a latent model, VJEPA-2 does not explicitly preserve fine image details. To complement this, we introduce r_{VF}, which combines frame-level and temporal no-reference metrics to capture sharpness, perceptual quality, motion consistency, structural coherence, and color stability. (Details of each component is in Appendix A.1). After normalization, all components are averaged to obtain the visual fidelity reward.

Final Reward Combinations After group-level normalization, the final reward is defined as:

$$r_{\text{ID}} = \lambda_{\text{pixel}}\, r_{\text{pixel}} + \lambda_{\text{VJEPA}}\, r_{\text{VJEPA-E}} + \lambda_{\text{VF}}\, r_{\text{VF}}, \quad r_{\text{OOD}} = (1 - \lambda_{\text{VF}})\, r_{\text{VJEPA-P}} + \lambda_{\text{VF}}\, r_{\text{VF}} \tag{11}$$

where λ_{VJEPA} and λ_{VF} control the VJEPA and VF contributions, and $\lambda_{\text{pixel}} = 1 - \lambda_{\text{VJEPA}} - \lambda_{\text{VF}}$.

3.3 VJEPA-2 Training Pipeline

To adapt VJEPA-2 to the robotic domain, we fine-tune its action-conditioned variant on robotic data with two modifications. First, while standard VJEPA-2 captures holistic scene-level representations, embodied settings require explicit modeling of the agent's structure (e.g., grippers) for accurate dynamics. We therefore introduce two complementary inputs: (i) raw images and (ii) embodiment-focused images obtained via segmentation, retaining only robot-specific regions. Second, VJEPA-2's original one-step prediction limits temporal consistency and robustness over longer horizons.

Table 1: Quantitative comparison of interactive long-horizon trajectory generation on the Agibot and Droid validation set. Our models generate 9.6-second trajectories autoregressively from a start frame conditioned on three sequential 3-step action chunks. Other baseline models follow their respective training settings. *: Reproduced results. PT: Pretrained; RL: Reinforcement Learning; CS: Control Signal; T: Text; AC: Action.

Dataset	Method	CS	Visual Quality			Flow Consistency		VJEPA Loss	
			PSNR ↑	SSIM ↑	LPIPS ↓	EPE ↓	COS ↑	Enc ↓	Pred ↓
Agibot	Genie-Envisioner*	AC	17.44	0.730	0.159	0.514	0.021	0.545	-
	ABot-PhysWorld*	T	18.24	0.772	0.189	0.516	0.073	0.526	-
	DreamDojo*	AC	19.69	0.790	0.161	0.437	0.155	0.528	-
	PT (ours)	AC	20.52	0.793	0.133	0.346	0.203	0.524	0.379
	RL (ours)	AC	**21.20**	**0.812**	**0.127**	**0.306**	**0.231**	**0.521**	**0.373**
Droid	WPE	AC	20.33	0.772	0.131	-	-	-	-
	IRASIM	AC	21.36	0.774	0.117	-	-	-	-
	Ctrl-World	AC	21.27	0.793	0.110	-	-	-	-
	Ctrl-World*	AC	21.30	0.827	0.097	**0.222**	0.162	**0.509**	-
	PT (ours)	AC	20.79	0.814	0.097	0.252	0.168	0.511	0.373
	RL (ours)	AC	**21.71**	**0.828**	**0.092**	0.227	**0.197**	**0.509**	**0.369**

Table 2: Reward-function ablations on the AgiBot validation set after 1K RL steps. We separately evaluate the ID and OOD reward branches and remove individual auxiliary components from the full PAV reward.

Setting	**Reward Branch**		**Visual Quality**			**Flow Consistency**		**VJEPA Loss**	
	ID	OOD	PSNR ↑	SSIM ↑	LPIPS ↓	EPE ↓	COS ↑	Enc ↓	Pred ↓
Pixel only	×	×	20.80	0.806	0.145	0.333	0.204	0.526	0.374
Pixel + OOD	×	✓	20.84	0.805	0.131	0.325	0.227	0.523	0.375
Pixel + ID	✓	×	20.93	0.807	0.135	0.324	0.216	0.524	0.374
Full w/o Static	✓	✓	20.85	0.806	0.140	0.328	0.212	0.526	0.409
Full w/o Emb	✓	✓	20.89	0.807	0.139	0.326	0.213	0.525	0.374
Full w/o VF	✓	✓	20.94	0.807	0.128	0.318	0.218	0.525	0.374
Full PAV reward	✓	✓	**21.05**	**0.810**	**0.126**	**0.310**	**0.230**	**0.521**	**0.373**

While accumulative training exists in the original implementation, it is still insufficient for long-horizon prediction. We therefore adopt an accumulative loss over multi-step predictions across the entire predicted frame chunk T, enabling the model to predict future frames given only the start frame and corresponding actions, akin to standard world models.

To preserve the VJEPA encoder's physical understanding, we adopt its action-conditioned training setup, freezing the encoder and training an action-conditioned predictor from scratch. And to accelerate training, we split the training into two stages, the pipeline is as shown in Appendix A.2. In Stage 1, following standard action-conditioned VJEPA-2 training, the model performs one-step future prediction, using frames $I_i, \ldots, I_{i+T-1}$ and actions to predict only I_{i+T} embedding. In Stage 2, we introduce two modifications: (i) random sampling from raw and embodiment-focused frames, and (ii) autoregressive rollout from start frame I_i and $A_{t\in[i,i+T)}$ with accumulated loss over horizon T, improving long-horizon consistency and embodiment sensitivity.

4 EXPERIMENTS

Datasets: Our primary dataset is Agibot Alpha (Bu et al., 2025), a dual-hand manipulation dataset with 92K trajectories (590 h) across 36 tasks. To evaluate dataset robustness, we also use DROID dataset (Khazatsky et al., 2024), a single-hand dataset with 76K trajectories (350 h) over 86 tasks.

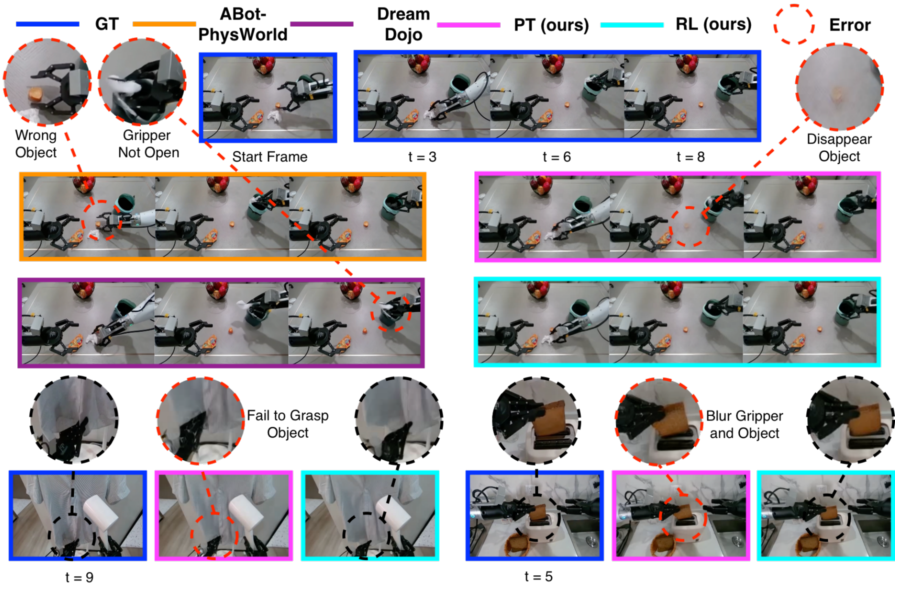


Figure 3: Qualitative comparison of long-horizon rollouts on Agibot validation set.

Training Details: We adopt Wan2.2-TI2V-5B (Wan et al., 2025) as the base model, conditioned on the first frame and 16 sparse history frames (6.4 s) to predict future $T = 16$ frames at 5 FPS (3.2 s). Input resolutions are 288×384 (Agibot) and 192×320 (Droid). The model uses single-view (Agibot: head view; Droid: third-person view) observations and 14D action inputs, consisting of 6D end-effector poses (x, y, z, roll, pitch, yaw) and gripper states per hand, with single-hand action zero-padded. The unified reward function uses $\alpha_{\text{OOD}} = 0.1$, $\lambda_{\text{VJEPA}} = 0.2$, $\lambda_{\text{static}} = 0.8$, $\lambda_{\text{emb}} = 0.5$, and $\lambda_{\text{VF}} = 0.1$, selected based on the ablation study in Section 4.2. Full supervised fine-tuning (SFT) is conducted on 8×H200 GPUs for 80K steps (batch size 32, 4 epochs, 3 days). Reinforcement learning (RL) further updates the action encoder and LoRA modules on the same hardware for 5K steps (group size 12, batch size 8, 5 days). Validation performance over training steps is shown in Appendix A.3. Stage 1 VJEPA-2 (Assran et al., 2025) training requires 60K steps (batch size 64, 3 days), while Stage 2 requires 12K steps (batch size 64, 5 days). To extract the embodiment-only frame, we uses GroundingDINO (Liu et al., 2024) for robot detection and Segment Anything Model (SAM) (Kirillov et al., 2023) for robot segmentation.

Baselines and Evaluation Metrics: For Agibot dataset, we evaluate single-view models with published weights, having Genie-Envisioner (Liao et al., 2025), ABot-PhysWorld (Chen et al., 2026), and DreamDojo (Gao et al., 2026); for methods without reported results, we re-run evaluations. We hold out 1% of each task from training and sample three trajectories per task for validation, yielding approximately 220 clips of 9.6 s each. For DROID dataset, we follow the Ctrl-World validation set and report results from prior single-view methods, including WPE (Quevedo et al., 2025), IRASIM (Zhu et al., 2024), and Ctrl-World (Guo et al., 2025), with Ctrl-World reproduced. Evaluation metrics include visual quality metrics (PSNR (Hore & Ziou, 2010), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018)) and flow-based metrics (Avg EPE and COS (Zhang et al., 2025a)), where EPE measures motion magnitude error and COS measures directional consistency between predicted and ground-truth frames. We also report VJEPA-2 encoder loss under ground-truth actions and predictor loss under suboptimal actions, the latter only for our models that support such conditioning.

4.1 Main Results

As shown in Table 1, our RL pipeline consistently improves upon the pretrained (PT) model, particularly in SSIM, LPIPS, and flow consistency, indicating enhanced structural fidelity, motion alignment and action controllability. Across both Agibot and DROID, our method achieves the best overall performance among all single-view baselines. Notably, although the PT model trails sev-

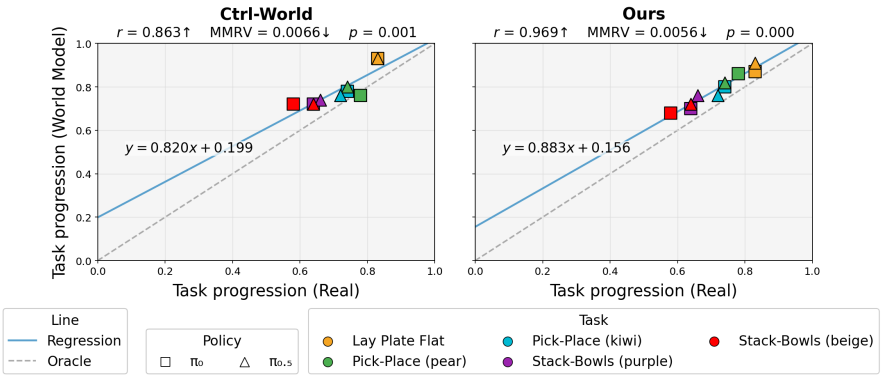


Figure 4: Quantitative correlations between real-world and world-model rollouts. Our policy predicts 32-action chunks at 30 fps, which are downsampled or extended to match each world model training setting. We evaluate 20 interactions between the policy and the world model.

eral baselines on DROID, RL training enables it to surpass them, highlighting the robustness and generalizability of our pipeline across datasets.

We further present qualitative comparisons in Figure 3 on the Agibot dataset using four representative models. As shown in the top example, despite its strong visual realism enabled by a 14B-parameter scale, ABot-PhysWorld (Chen et al., 2026) frequently fails to follow the text instructions, such as grasping incorrect objects. DreamDojo (Gao et al., 2026) exhibits weak long-horizon consistency, where action execution deteriorates during autoregressive rollouts, leading to delayed or missing gripper motions. Our pretrained (PT) model demonstrates reasonable action adherence but still suffers from limitations. In the top example, it violates physical plausibility—object permanence—having objects to disappear during gripper-induced occlusions. In the bottom-left example, it violates action adherence, as fine-grained manipulation actions, such as grasping a specific corner of a cloth, are occasionally missed. In the bottom-right example, visual fidelity degrades due to motion blur during object interaction. In contrast, our RL-enhanced model consistently achieves superior physical plausibility, action adherence, and visual fidelity across all examples, demonstrating the effectiveness of the proposed unified reward design.

A blinded human evaluation on 100 randomly sampled AgiBot trajectories further confirms that PAVXploreRL is preferred over the pretrained model in physical plausibility, action adherence, visual fidelity, and overall quality; detailed protocols and results are provided in Appendix A.4.

4.2 Ablation Study

Table 2 evaluates the reward branches and auxiliary components. Compared with pixel-only RL, independently adding the ID VJEPA encoder reward or OOD VJEPA predictor reward improves overall performance, while their combination performs best. Ablations of λ_{VJEPA} and α_{OOD} are provided in Appendix A.5.

Removing static regularization, embodiment-focused rewards, or the visual fidelity reward consistently degrades performance, confirming their complementary contributions. Further analyses of λ_{static}, λ_{emb}, and λ_{VF} are reported in Appendix A.6.

4.3 VLA Evaluation

To evaluate our world model as a policy evaluator, we test two Droid-focused policies—π_0, and $\pi_{0.5}$. We compare our model against Ctrl-world across five tasks: (1) lay plate flat, (2) pick-place (pear), (3) pick-place (kiwi), (4) stack-bowls (purple) and (5) stack-bowls (beige). Each task comprises around expert 300 trajectories to posttrain the policy, the world model and VJEPA-2. Starting from our Droid-pretrained checkpoint, we apply 5k SFT and 1k RL updates to our model and 6k fine-tuning updates to Ctrl-World; RL additionally incurs rollout generation and reward computation.

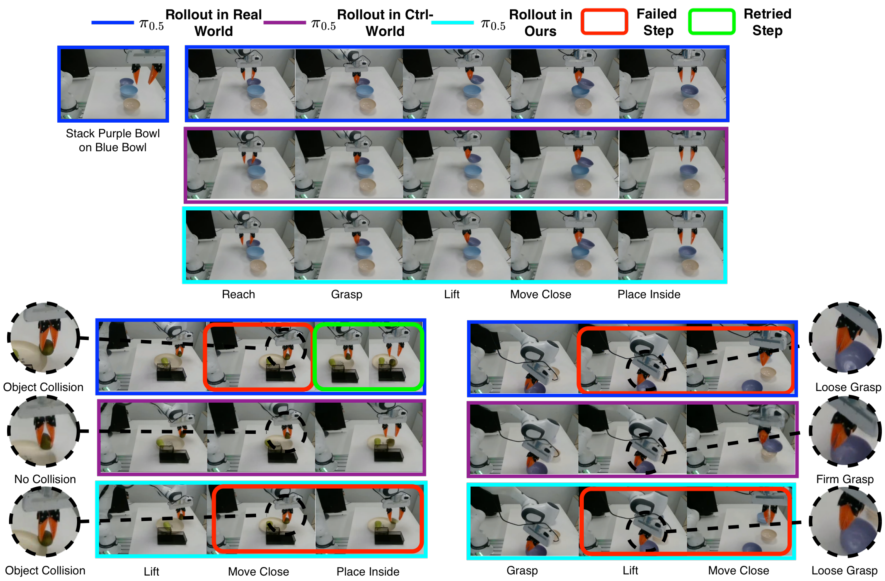


Figure 5: Qualitative comparison of $\pi_{0.5}$ rollouts in the real world and world model.

After training, we evaluate each task over 10 episodes in both the real world and the learned world model. The overall progress is divided into N steps, where successfully completing each step in sequence contributes $1/N$ to the total progress. The lay plate flat task is decomposed into three steps: (1) reach, (2) touch, and (3) flatten. The pick-place task consists of five steps: (1) reach, (2) grasp, (3) lift, (4) move close, and (5) place inside. The stack-bowls task follows the same five-step structure, but differs in the measurement of grasp, and placement-in-bowl phases due to the distinct geometry of bowls compared to fruits.

We report a quantitative comparison of policy rollouts between real world and the world model in Figure 4. Results show that, compared to Ctrl-World, our model exhibits reduced overestimation of success rates and provides more accurate estimates of real-world performance. Qualitative results are shown in Figure 5, where Ctrl-world often predicts task success even when the policy fails from the same initial observation in real world. While our model faithfully reflects real-world dynamics (e.g., object collisions and loose grasps), these results demonstrate robustness in handling suboptimal policy failures under RL with OOD actions.

5 Conclusion

We present **PAVXploreRL**, a model-agnostic reinforcement learning (RL) pipeline featuring a unified reward function that jointly optimizes the PAV targets: Physical Plausibility, Action Adherence, and Visual Fidelity. The pipeline trains on both in-distribution and out-of-distribution actions without requiring paired video supervision. Empirical results on our pretrained action-conditioned world model demonstrate consistent gains in both quantitative metrics and qualitative evaluations, as well as improved its capability as a reliable action policy evaluator.

A limitation of this work is that our focus is primarily on RL design, while the base world model does not explicitly incorporate components such as multi-view consistency. Another limitation is the long training time; despite using DiffusionNFT as a more efficient RL pipeline, the long rollouts and reward computation still incur substantial computational cost, making it suitable primarily as a post-training method. Nevertheless, the proposed RL framework is orthogonal to these design choices and can be readily extended to future world model architectures. Future work could extend our method to advanced pretrained world models to further test its generality and robustness.

AI USE STATEMENT

Generative AI tools were used to support idea discussion, language polishing, and limited code execution and debugging. All research decisions, implementations, experiments, analyses, and conclusions were performed or verified by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

We provide the implementation as supplementary material. The accompanying README includes instructions for training and inference, while additional implementation details and experimental settings are documented in the paper and appendix. Upon acceptance, we will publicly release the trained model weights.

ETHICS STATEMENT

This work focuses on action-conditioned video generation for robotics research. Our implementation builds upon publicly available code, pretrained models, and research datasets, subject to their respective licenses and usage terms. We conduct a blinded human evaluation but do not collect personal or sensitive participant data. Although we do not anticipate direct harmful impacts, generated videos should be treated as simulated predictions rather than guaranteed representations of real-world outcomes, particularly in safety-critical robotics applications.

REFERENCES

N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, others, and A. Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025a.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, and et al. Cosmos world: Foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025b.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *CoRR*, abs/2506.09985, June 2025. URL https://doi.org/10.48550/arXiv.2506.09985.

A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, ..., and N. Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

J. Bardhan, P. Drozdik, J. Sivic, and V. Petrik. Persistent robot world models: Stabilizing multi-step rollouts via reinforcement learning. *arXiv preprint arXiv:2603.25685*, 2026.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

H. Bi, H. Tan, S. Xie, Z. Wang, S. Huang, H. Liu, ..., and J. Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, J. Zhu, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Y. Chen, R. Chen, D. Huo, Y. Yang, D. Qi, H. Liu, et al. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, and et al. Wow: Towards a world omniscient model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Team G., Ye A., Wang B., Ni C., Huang G., Zhao G., Zhu Z., and et al. Gigaworld-0: World models as data engine to empower embodied ai. *arXiv preprint arXiv:2511.19861*, 2025.

Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Zhehao Cai, and Lin Shao. Flip: Flow-centric generative planning as a general-purpose manipulation world model. *arXiv preprint arXiv:2412.08261*, 2024.

S. Gao, W. Liang, K. Zheng, A. Malik, S. Ye, S. Yu, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, others, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Y. Guo, L. X. Shi, J. Chen, and C. Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369. IEEE, 2010.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Björck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, and et al. Dreamgen: Unlocking generalization in robot learning through neural trajectories. *arXiv preprint arXiv:2505.xxxxx*, 2025.

Z. Jiang, S. Zhou, Y. Jiang, Z. Huang, M. Wei, Y. Chen, ..., and D. Zhao. Wovr: World models as reliable simulators for post-training vla policies with rl. *arXiv preprint arXiv:2602.13977*, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pp. 170–185, 2018.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, and et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

J. Liu, G. Liu, J. Liang, Y. Li, J. Liu, X. Wang, ..., and W. Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in Neural Information Processing Systems*, 38:40783–40818, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection, 2024. URL `https://openreview.net/forum?id=DS5qRs0tQz`.

D. McAllister, S. Ge, B. Yi, C. M. Kim, E. Weber, H. Choi, and A. Kanazawa. Flow matching policy gradients. *arXiv preprint arXiv:2507.21053*, 2025.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. In *IEEE Signal Processing Letters*, volume 20, pp. 209–212, 2013.

Bohao Peng, Wanyi Zhang, Liang Xu, Zekun Qi, Jiayi Zhang, Hao Liu, and Xin Jin. Reworld: Multi-dimensional reward modeling for embodied world models. *arXiv preprint arXiv:2601.12428*, 2026.

Julian Quevedo, Percy Liang, and Sherry Yang. Evaluating robot policies in a world model. *arXiv preprint arXiv:2506.00613*, 2025.

A. K. Sharma, Y. Sun, N. Lu, Y. Zhang, J. Liu, and S. Yang. World-gymnast: Training robots with reinforcement learning in a world model. *arXiv preprint arXiv:2602.02454*, 2026.

Hengkai Tan, Yao Feng, Xinyi Mao, Shuhe Huang, Guodong Liu, Zhongkai Hao, Hang Su, and Jun Zhu. Anypos: Automated task-agnostic actions for bimanual manipulation. *arXiv preprint arXiv:2507.12768*, 2025.

R. Team, Z. Gao, Q. Wang, Y. Zeng, J. Zhu, K. L. Cheng, ..., and H. Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations (ICLR)*, pp. 92033–92052, 2025.

T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C. W. Xie, others, and Z. Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

S. Wang, J. Shi, Z. Fu, X. He, F. Liu, C. Yang, others, and Y. G. Jiang. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zhou Wang, Tian Wang, Hao Zhang, Xin Zuo, Jiaya Wu, Han Wang, ..., and Zhen Zhao. Worldcompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026b.

Ji Wu, Shuo Yin, Nannan Feng, and Meng Long. Rlvr-world: Training world models with reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 125312–125350, 2026.

G. Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, Peng Li, Qiuping Deng, Runqi Ouyang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yang Wang, Yifan Li, Yilong Li, Yiran Ding, Yuan Xu, Yun Ye, Yukun Zhou, Zhehao Dong, Zhenan Wang, Zhichao Liu, and Zheng Zhu. Gigabrain-0: A world model-powered vision-language-action model. *ArXiv*, abs/2510.19430, 2025. URL https://api.semanticscholar.org/CorpusID:282272674.

S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, others, and J. Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

J. Yuan, X. Zhang, F. Friedrich, N. Beltran-Velez, M. Hall, R. Askari-Hemmat, ..., and A. Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

J. Zhang, Z. Huang, C. Gu, Z. Ma, and L. Zhang. Reinforcing action policies by prophesying. *arXiv preprint arXiv:2511.20633*, 2025a.

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agents. *arXiv preprint arXiv:2501.18867*, 2025b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=J8JCF64aEn.

R. Zhang, M. Zhang, J. Zhou, Z. Guo, X. Liu, Z. Xu, and X. Li. Mind-V: Hierarchical video generation for long-horizon robotic manipulation with RL-based physical alignment. *arXiv preprint arXiv:2512.XXXXX*, 2025c.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Z. Zhang, H. Ren, Y. Sun, Y. Sheng, H. Wang, H. Lin, ..., and Y. Yu. Towards practical world model-based reinforcement learning for vision-language-action models. *arXiv preprint arXiv:2603.20607*, 2026b.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, and et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025.

K. Zheng, H. Chen, H. Ye, H. Wang, Q. Zhang, K. Jiang, ..., and M. Y. Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025a.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Björck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, and et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025b.

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Wenxuan Song, Jiayi Chen, and Haoang Li. Flowvla: Thinking in motion with a visual chain of thought. *arXiv preprint arXiv:2508.18269*, 2025.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.

A Appendix

A.1 Visual Fidelity Reward Components

To evaluate visual fidelity without ground-truth video references, we employ a set of non-reference metrics summarized in Table 3. The overall reward $r_{r_{VF}}$ combines multiple complementary objectives that jointly regularize spatial sharpness, perceptual realism, and temporal consistency.

Laplacian sharpness encourages high-frequency detail by penalizing low variance in the Laplacian response of grayscale frames, thereby discouraging blurriness. **NIQE**-inspired naturalness measures deviations of MSCN statistics from Gaussianity, capturing perceptual artifacts without reference supervision. **Flow Distribution** regularizes the statistics of frame-wise intensity changes, penalizing static collapse, excessive dynamics, and degenerate motion patterns to promote physically plausible temporal evolution.

For temporal consistency, **Temporal Gradient** penalizes second-order temporal derivatives to suppress flickering, while **Interframe SSIM** enforces structural coherence between adjacent frames via $1 - \text{SSIM}$. In the frequency domain, **Phase Coherence** constrains temporal drift of Fourier phase representations, preserving stable spatial structure over time.

Appearance stability is further enforced via **Color Stability**, which penalizes temporal shifts in histogram-based color distributions. Finally, **Gradient Consistency** minimizes temporal variance in Sobel gradient magnitudes, reducing texture flickering and preserving edge stability across frames.

Table 3: Summary of Visual Fidelity reward components and their effects.

Metric	Effect
Laplacian	Captures frame sharpness
NIQE	Measures perceptual quality (no reference)
Flow Distribution	Ensures optical flow consistency
Temporal Gradient	Detects sudden motion changes
Interframe SSIM	Evaluates frame-to-frame similarity
Phase Coherence	Captures frequency-domain motion consistency
Color Stability	Maintains stable colors
Gradient Consistency	Preserves spatial edge consistency over time

A.2 VJEPA-2 Training Pipeline

This is a supplementary illustration of the VJEPA training pipeline shown in Figure 6. The full pipeline is described in Section 3.3.

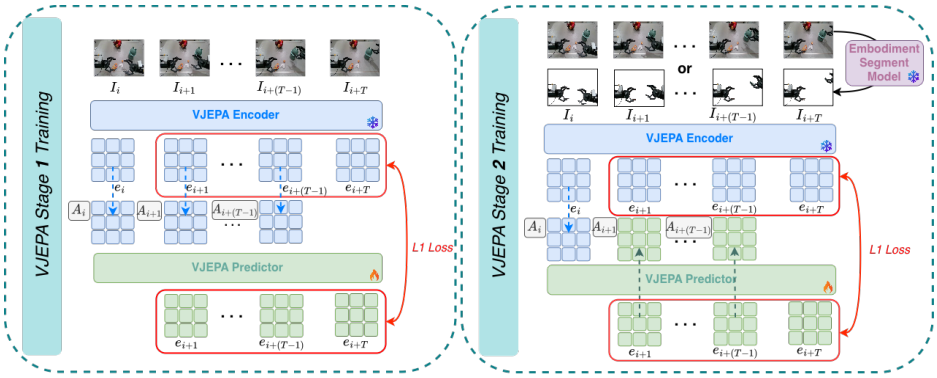


Figure 6: VJEPA-2 training pipeline. **Stage 1:** VJEPA-2 is trained on raw frames with one-step prediction. **Stage 2:** It extends to autoregressive multi-step prediction over a horizon T, iteratively propagating from an initial frame using frames sampled from raw or embodiment-focused views.

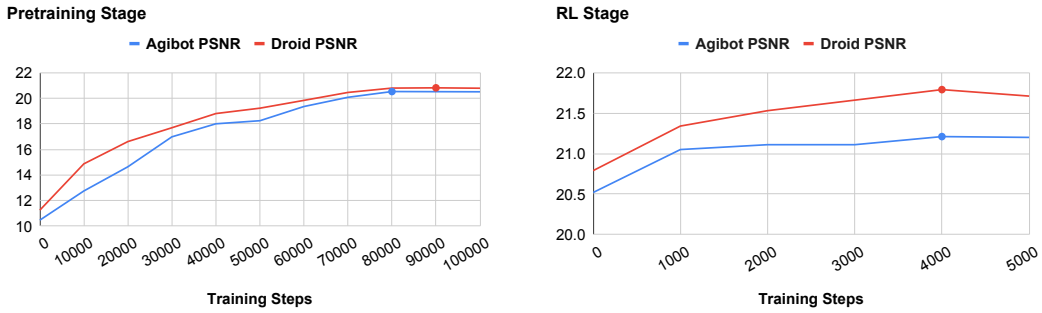


Figure 7: PSNR of model on validation sets in pretraining and reinforcement learning stage over training steps for Agibot and Droid datasets.

Table 4: Blinded human evaluation of the pretrained model (PT) and PAVXploreRL (RL) on 100 randomly sampled AgiBot validation trajectories. Each trajectory is evaluated by two independent evaluators. We report trajectory-level average preference rates.

Criterion	RL Preferred ↑	Tie	PT Preferred ↓	Cohen's κ ↑
Physical Plausibility	**60.5%**	17.0%	22.5%	0.36
Action Adherence	**65.0%**	16.5%	18.5%	0.49
Visual Fidelity	**68.0%**	14.0%	18.0%	0.53
Overall Quality	**69.0%**	16.0%	15.0%	0.52

A.3 Model Performance over Training Steps

We report PSNR on the validation set across training steps for both datasets during the pretraining and reinforcement learning (RL) stages in Figure 7. We firstly observe that pretraining converges at approximately 80K steps, with further Supervised Fine-Tuning (SFT) yielding marginal improvements, indicating that the RL posttraining gains do not stem from additional training alone. Based on these results, we select 80K steps for pretraining and 5K steps for RL, achieving a balance between performance and efficiency.

A.4 Human Evaluation

We conduct a blinded human evaluation on the 100 trajectories sampled from the AgiBot validation set. For each trajectory, two independent evaluators compare PT and RL predictions, presented in randomized order alongside the ground-truth future. They assess physical plausibility, action adherence, visual fidelity, and overall quality, selecting PT, RL, or a tie for each criterion. We report trajectory-level average preference rates and Cohen's κ for inter-rater agreement. As shown in Table 4, the results provide an evaluation of PAV quality independent of the VJEPA-based metrics.

A.5 Ablation Study of VJEPA Reward

As discussed in Section 4.2, the contribution of VJEPA-E is governed by λ_{VJEPA}, while VJEPA-P, due to its critical role in the out-of-distribution (OOD) branch, is effectively controlled by the OOD branch proportion α_{OOD}. We conduct a detailed ablation over these coefficients in Table 5. We observe that moderate values yield optimal performance, whereas excessively large values can degrade performance. The best performance is achieved with $\lambda_{\text{VJEPA}} = 0.2$ and $\alpha_{\text{OOD}} = 0.1$.

A.6 Reward Function Coefficient Selection

To avoid the high computational cost of RL training, we first perform reward-function coefficient selection through an offline diagnostic benchmark. The underlying assumption is that a well-designed

Table 5: Ablation study on the effect of VJEPA based metrics ratios in reward function, evaluated on Agibot with 1K-step reinforcement learning.

Setting	Visual Quality			Flow Consistency		VJEPA Loss	
	PSNR ↑	SSIM ↑	LPIPS ↓	EPE ↓	COS ↑	Enc ↓	Pred ↓
$\lambda_{\text{VJEPA}} = 0, \alpha_{\text{OOD}} = 0$	20.80	0.806	0.145	0.333	0.204	0.526	0.374
VJEPA Predictor Only							
$\alpha_{\text{OOD}} = 0.05$	20.84	0.805	**0.131**	**0.325**	**0.227**	**0.523**	0.375
$\alpha_{\text{OOD}} = 0.1$	**20.94**	**0.807**	0.139	0.329	0.206	0.526	0.375
$\alpha_{\text{OOD}} = 0.2$	20.90	0.806	0.141	0.329	0.205	0.526	**0.374**
VJEPA Encoder Only							
$\lambda_{\text{VJEPA}} = 0.1$	20.86	0.806	0.141	0.333	0.204	0.526	0.375
$\lambda_{\text{VJEPA}} = 0.2$	**20.93**	**0.807**	**0.135**	**0.324**	**0.216**	**0.524**	**0.374**
$\lambda_{\text{VJEPA}} = 0.3$	20.82	0.805	**0.135**	0.325	0.215	**0.524**	**0.374**
VJEPA Encoder and Predictor							
$\lambda_{\text{VJEPA}} = 0.2, \alpha_{\text{OOD}} = 0.05$	20.99	0.806	0.138	0.327	0.210	0.525	0.375
$\lambda_{\text{VJEPA}} = 0.2, \alpha_{\text{OOD}} = 0.1$	**21.05**	**0.810**	**0.126**	**0.310**	**0.230**	**0.521**	**0.373**

Table 6: Ablation of reward function coefficients on the AgiBot validation set, reporting the accuracy of correctly assigning the highest reward to the ground-truth (GT) video among 12 evaluated video variants.

(a) λ_{static}

Value	0.0	0.2	0.4	0.5	0.8	1.0
GT ratio	0%	11%	39%	67%	**82%**	67%

(b) λ_{emb}

Value	0.00	0.25	0.50	0.75	1.00
GT ratio	79%	81%	**82%**	77%	75%

(c) λ_{VF}

Value	0.0	0.1	0.2
GT ratio	75%	**82%**	71%

reward function should consistently assign the highest reward to high-quality videos while penalizing degraded ones.

Specifically, we construct 12 video variants, including: (1) ground-truth (GT) videos, (2) videos predicted by the pretrained world model, (3) noisy videos with global or motion-specific perturbations, (4) temporally truncated videos with the first 0.2, 0.4, or 0.8 seconds removed, (5) temporally slowed videos (2× slower) with different slowdown onset points corresponding to 100%, 75%, 50%, and 25% of the original trajectory, (6) temporally static videos repeating the first frame.

For each coefficient setting, we treat the reward function as a classifier and report accuracy as the proportion of cases in which the ground-truth (GT) video attains the highest reward among 12 variants. This metric is used to ablate the coefficients λ_{static}, λ_{emb}, and λ_{VF}, with higher accuracy indicating stronger alignment between the reward function and video quality.

The results are reported in Table 6. We observe that λ_{static} has the most significant impact and is essential; removing it leads to cases where ground-truth videos consistently receive lower rewards than temporally static videos. Both λ_{emb} and λ_{VF} further improve performance, indicating that embodiment consistency and visual fidelity act as complementary signals to the main reward, and that moderate weighting improves classifier accuracy. Based on these results, we set $\lambda_{static} = 0.8$, $\lambda_{emb} = 0.5$, and $\lambda_{VF} = 0.1$.