PHR-VLA: Planning Horizon Reasoning for Vision-Language-Action Models

Davood Soleymanzadeh[1], Kaidi Zhang[2], Zhiyuan Zhang[2], Bihao Zhang[1], Xiao Liang[3], Yu She[2], Minghui Zheng[1]

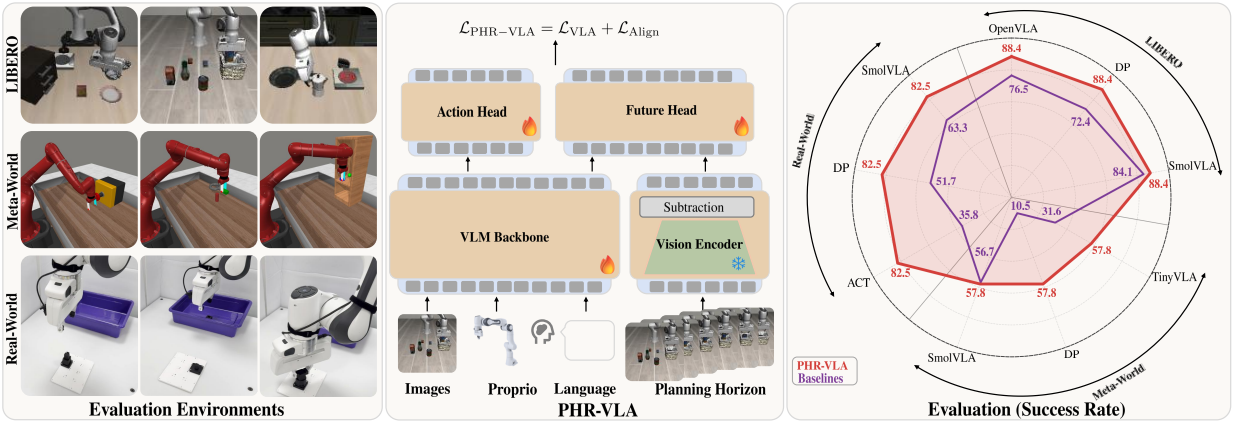

Fig. 1: **Method Overview.** *Left:* The three evaluation domains: LIBERO [1], Meta-World [2], and real-world disassembly. *Center:* PHR-VLA fine-tunes a base VLA policy with an auxiliary future head that predicts planning-horizon latent dynamics from a frozen vision encoder, jointly optimizing $\mathcal{L}_{\text{PHR-VLA}} = \mathcal{L}_{\text{VLA}} + \lambda\mathcal{L}_{\text{Align}}$. The future head and vision encoder are used only during training. *Right*: Across all three domains, PHR-VLA (red) consistently outperforms SmolVLA [3], and other baseline policies (purple), including OpenVLA [4], Diffusion Policy (DP) [5], TinyVLA [6], and ACT [7].

***Abstract*—Vision-language-action models (VLAs) have shown strong promise for general-purpose robotic manipulation by mapping language instructions and vision observations directly to actions. However, most VLAs primarily condition action prediction on current observations and lack an explicit mechanism for reasoning over future task dynamics, which is particularly important for fine-grained, contact-rich manipulation. We present PHR-VLA, a framework that enables planning-horizon reasoning in VLAs through privileged latent representations of future dynamics. PHR-VLA introduces a lightweight auxiliary future head that, during training, aligns the VLA's internal representations with latent dynamics extracted from future observations. Evaluation results demonstrate that local, contact-centric, patch-level latent dynamics supervision from the wrist camera improves success rate on LIBERO from 84.1% to 88.4% and on real-world disassembly tasks from 63.3% to 82.5%. Patch-level supervision from a third-person camera also improves performance on Meta-World from 56.70% to 57.8%. These results demonstrate that privileged latent dynamics alignment provides an effective training signal for improving anticipatory reasoning in VLA policies. Project website: https://davoodsz.github.io/PHR-VLA.github.io/**

***Index Terms*—Vision-Language-Action Models, Planning-Horizon Reasoning, Privileged Representations**

I. Introduction

Vision-language-action (VLA) models have demonstrated strong potential for general-purpose robotic manipulation [4], [8]–[10]. By extending pretrained vision-language models (VLMs) to robot action prediction, VLAs ground language instructions and visual observations into executable control commands using large-scale robot demonstration datasets [11]. Despite this progress, many VLAs remain largely reactive: they predict actions from current time-step observations without explicitly modeling how the task and scene may evolve over the planning horizon. This limitation can hinder temporally coordinated behavior in manipulation tasks that require anticipation of future dynamics [12].

Recent studies have begun to address this limitation by incorporating memory and anticipatory reasoning into VLA policies. Memory-based approaches provide additional temporal context by conditioning the policy on a dense history of observations [13]–[15]. However, processing dense, long-horizon observation histories can increase computational cost

[1]Davood Soleymanzadeh, Bihao Zhang, and Minghui Zheng are with the J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA (e-mail: davoodso@tamu.edu; bhzhang@tamu.edu; mhzheng@tamu.edu).

[2]Kaidi Zhang, Zhiyuan Zhang, and Yu She are with the Department of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: zhan5896@purdue.edu; zhan5570@purdue.edu; shey@purdue.edu).

[3]Xiao Liang is with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xliang@tamu.edu).

This work was partially supported by the USA National Science Foundation under Grant No. 2527316, No. 2422826 and No. 2423068. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

and inference latency, which is undesirable for real-time robotic control. To reduce this overhead, recent methods have explored compact memory representations, including proprioceptive states memory [16], 2D point tracks [17], selected keyframes [18], and natural-language summaries [19].

In parallel, forecasting-based approaches seek to improve embodied reasoning by predicting future-oriented representations before or during action generation. These representations include language descriptions, latent features, subgoal images, or visual foresight signals [10], [12], [20], [21]. Large video prediction and world-model-based methods further suggest that future prediction objectives can improve policy generalization to novel scenes and objects [22], [23]. However, these approaches often require explicit future prediction, autoregressive reasoning, or additional inference modules. Such components can increase computational cost and complicate integration with pretrained VLAs for real-time control [21].

These observations motivate the following question: can a VLA learn future-aware planning representations without explicitly generating future observations or rollouts at inference time [24]? To address this question, we introduce **PHR-VLA**, a framework for planning-horizon reasoning in VLA models through privileged latent dynamics representations. Rather than adding an explicit world model at deployment, PHR-VLA uses latent representations of future observations as privileged supervision during training. Specifically, a lightweight auxiliary future head aligns current internal embeddings of the VLA with latent representations of future task dynamics over the planning horizon. This objective encourages the policy to encode information relevant to future task evolution while preserving the original action-generation pipeline at inference time.

Our main contributions are:

- We introduce PHR-VLA, a framework that leverages privileged latent dynamics to enable planning-horizon reasoning in VLAs. PHR-VLA encourages the policy to encode future task dynamics without requiring explicit future rollouts or world-model inference during deployment.
- We systematically investigate different camera viewpoints, representation granularities, latent target formulations, and encoder choice for privileged future supervision. Our results show that short-horizon, local, and contact-centric signals are particularly effective. Patch-level wrist-camera supervision improves the success rate on LIBERO [1] from **84.1**% to **88.4**% and on real-world disassembly tasks from **63.3**% to **82.5**%. Patch-level third-person-camera supervision also improves performance on Meta-World [2] from **56.7**% to **57.8**%.

The remainder of this paper is organized as follows. Section II reviews related work on VLAs, reasoning in robotic foundation models, and world models. Section III presents the proposed PHR-VLA framework. Section IV describes the experimental evaluation. Finally, Section V concludes the paper.

II. Related Work

We review prior work on vision-language-action models, temporal reasoning through memory and forecasting, and world models for robotic planning.

Vision-Language-Action Models (VLAs). Vision-language-action models (VLAs) have emerged as a promising approach to general-purpose robotic manipulation by fine-tuning pretrained vision-language models (VLMs) on robot demonstration data [4], [8]–[11], [25]. These models use either discrete action decoders [4] or continuous policy heads [9] to generate robot actions. By leveraging the semantic knowledge acquired during large-scale VLM pretraining, VLAs can generalize across tasks, objects, and environments. However, most VLAs primarily generate actions from the current time-step visual language context, with limited explicit representation of how the task may evolve over the planning horizon [26], [27]. In contrast, PHR-VLA improves the internal planning horizon representations of a VLA while preserving its original action-generation interface.

Improving VLAs with Memory and Forecasting. A growing body of work has investigated the use of temporal context to improve VLA policies [13], [28]. One line of research conditions the policy on a dense history of observations [14], [15]. Although dense observation histories provide useful temporal information, they also increase input sequence length, computational cost, and inference latency [13]. Recent methods have therefore explored more compact memory representations, including proprioceptive state memory [16], 2D point tracks [17], selected keyframes [18], and natural-language summaries [19].

Another line of research introduces explicit reasoning or forecasting stages before action generation. These methods generate intermediate representations such as language reasoning traces, feature embeddings, subgoal images, visual foresight signals, or future action representations [10], [12], [20], [21], [29], [30]. Such representations can improve task-level planning and temporal action consistency by encouraging the policy to reason beyond the current observation. However, many of these methods require additional inference modules, autoregressive reasoning steps, or explicit prediction stages, which can increase latency and complicate their integration with pretrained VLA architectures. PHR-VLA leverages privileged future dynamics reasoning only during training and does not require explicit future rollouts or additional reasoning stages during deployment.

World Models in Robotics. World models and video prediction models support future-aware decision-making by explicitly modeling how the environment evolves over time [22], [23], [31]. By learning predictive latent representations or generating future visual rollouts, these methods can support planning, policy learning, and generalization to novel scenes and object configurations. Recent studies have shown that coupling future prediction objectives with action prediction can improve the generalization of robotic policies [22], [23].

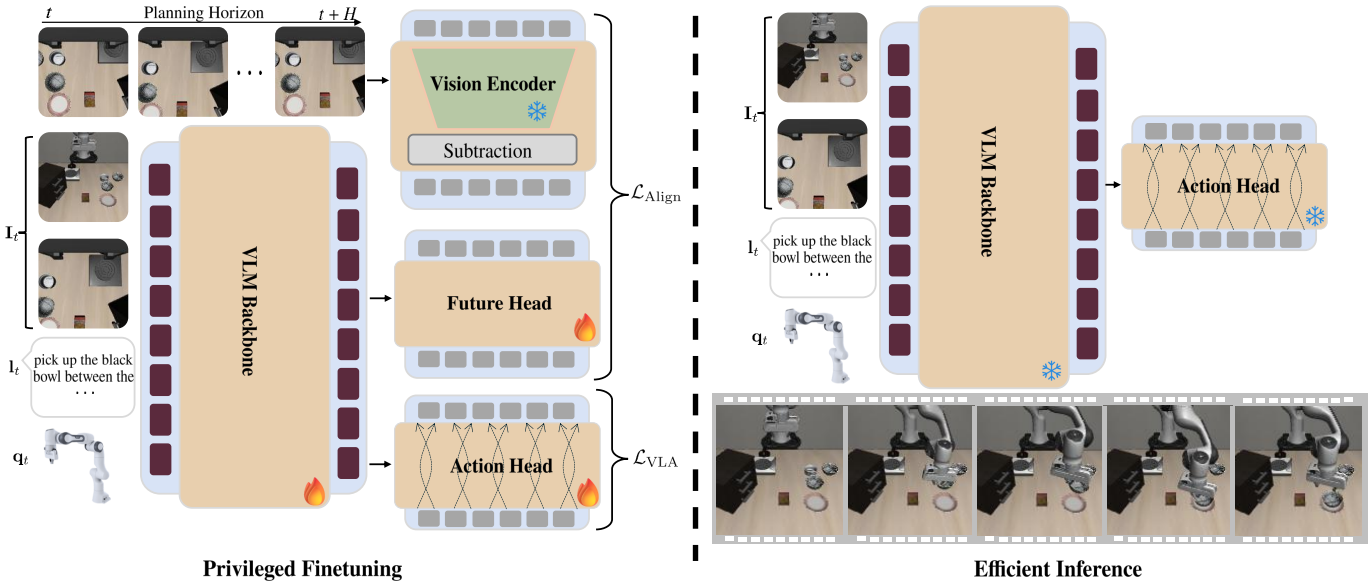

Fig. 2: **PHR-VLA Framework. Privileged Finetuning:** In addition to the current observation $\mathbf{I}_t$, language instruction $\mathbf{l}_t$, and proprioceptive state $\mathbf{q}_t$ used by the standard VLA backbone, privileged planning horizon frames $\mathbf{I}_{t+1:t+H}$ are passed through a frozen vision encoder to compute latent-dynamics targets. The action-token latents produced by VLM backbone feed two heads: the standard flow matching action head, supervised by $\mathcal{L}_{\text{VLA}}$, and the auxiliary future head which is supervised by $\mathcal{L}_{\text{Align}}$ against the latent dynamics targets. **Efficient Inference:** The action head runs exactly as in standard VLA backbone, so PHR-VLA adds zero latency or memory cost at deployment.

However, world-model- and video-prediction-based policies often rely on future rollout generation, autoregressive latent prediction, or large generative architectures at inference time [21]. These components can introduce substantial computational cost and latency, particularly in real-time manipulation settings. PHR-VLA draws on the representational benefits of future dynamics modeling while avoiding explicit world-model inference during deployment [24]. Rather than generating future observations or rollouts online, PHR-VLA uses a lightweight alignment objective to transfer privileged future dynamics information into the internal representations of the VLA during training. Therefore, PHR-VLA complements world-model-based approaches by improving future-aware action generation while maintaining the inference pipeline and computational efficiency of the underlying VLA.

III. PHR-VLA

We introduce PHR-VLA, a training-time auxiliary supervision framework that uses privileged future latent dynamics to improve planning horizon reasoning in VLAs.

A. VLAs

A standard VLA framework receives an observation $\mathbf{o}_t \triangleq (\mathbf{I}_t, \mathbf{l}_t, \mathbf{q}_t)$ at each timestep t and predicts an action chunk $\mathbf{a}_{t:t+H}$ over a planning horizon H. Here, $\mathbf{I}_t = \{I_t^i\}_{i=1}^N$ denotes a set of N camera images, $\mathbf{l}_t$ is the language instruction, and $\mathbf{q}_t$ is the robot proprioceptive state. The visual observations and proprioceptive state are encoded using modality-specific encoders and projected into the token embedding space of the VLM backbone, together with the tokenized language instruction. The backbone produces multimodal representations, which are processed by an action head to parameterize an action distribution [32]. The model is trained on a demonstration dataset $\mathcal{D}$ with the standard log-likelihood objective:

$$\max_{\theta} \; \mathbb{E}_{(\mathbf{a}_{t:t+H}, \mathbf{o}_t)\sim\mathcal{D}} \left[\log p_\theta(\mathbf{a}_{t:t+H} \mid \mathbf{o}_t)\right]. \tag{1}$$

B. VLA Baseline

We adopt SmolVLA (0.45B) [3] as our VLA baseline. SmolVLA couples a compact SmolVLM-2 backbone with a flow-matching action expert that predicts continuous action chunks given visual, language, and proprioceptive inputs. We choose SmolVLA because it provides strong VLA representations while remaining computationally efficient. We can finetune it on a single A100 GPU, which enables controlled ablations of our planning horizon dynamics-aware alignment.

C. Planning Horizon Latent Dynamics

PHR-VLA uses future planning horizon observations available in the demonstration dataset as privileged training information. We encode the observations over the planning horizon $\mathbf{o}_{t:t+H}$ using a frozen visual encoder $\mathcal{E}_{\text{encoder}}$ [33], [34]. Let

$$\mathbf{s}^c_{t:t+H} = \mathcal{E}_{\text{encoder}}(\mathbf{I}^c_{t:t+H}) \tag{2}$$

denote the visual latents of the planning horizon observation, where c denotes the camera used for the auxiliary target. Rather than supervising the policy to match absolute planning horizon visual latents, the main formulation supervises planning horizon latent dynamics:

$$\mathbf{y}^c_{t:t+H} = \mathbf{s}^c_{t+1:t+H} - \mathbf{s}^c_{t:t+H-1}. \tag{3}$$

This target emphasizes how the scene representation changes over the planning horizon, while reducing the need

to predict static visual content already present in the current timestep observation.

TABLE I: **Results on LIBERO.** Success rates are reported across four task suites (300 trials per suite). Entries marked with * are taken from [35] and may use different model sclaes, training setups, and computational budgets. The rest are trained under the same a single A100 GPU setting. Best results are **bold**, and second-best results are underlined.

Method	Spatial ↑	Object ↑	Goal ↑	Long ↑	Average ↑
ACT [7]	30.7%	48.4%	02.0%	15.7%	24.2%
Diffusion Policy [5]*	78.3%	92.5%	68.3%	50.5%	72.4%
Octo [15]*	78.9%	85.7%	84.6%	51.1%	75.1%
DiT Policy [36]*	84.2%	96.3%	85.4%	63.8%	82.4%
OpenVLA [4]*	84.7%	88.4%	79.2%	53.7%	76.5%
SmolVLA [3]	86.0%	97.7%	86.7%	66.0%	84.1%
PHR-VLA (ours)	**88.7%**	**98.0%**	**92.4%**	**74.4%**	**88.4%**

D. Planning Horizon Reasoning

Given the current observation and language instruction, SmolVLA [3] produces an action chunk together with an internal action-token representation $\mathbf{z}_t$. PHR-VLA attaches a lightweight future head g_ϕ that predicts planning horizon latent dynamics from these action-token representations:

$$\hat{\mathbf{y}}^c_{t,t+H} = g_\phi(\mathbf{z}_t). \tag{4}$$

The head encourages the planned-action representation to encode how the visual scene is expected to evolve across the action horizon. The predicted and privileged latent dynamics are aligned alongside the standard VLA training objective as follows:

$$\mathcal{L}_{\text{PHR}-\text{VLA}} = \mathcal{L}_{\text{VLA}} + \lambda \mathcal{L}_{\text{Align}}, \tag{5}$$

where the alignment loss $\mathcal{L}_{\text{Align}}$ is computed as the mean-squared error between predicted and privileged planning-horizon latent dynamics and λ controls the strength of planning horizon dynamics supervision.

TABLE II: **Results on Meta-World.** Success rates are reported across all difficulty levels. Entries marked with * are taken from [3]and may use different model sclaes, training setups, and computational budgets. The rest are trained under the same a single A100 GPU setting. Best results are **bold**, and second-best results are underlined.

Method	Easy ↑	Medium ↑	Hard ↑	VeryHard ↑	Average ↑
ACT [7]	68.2%	49.7%	28.9%	49.4%	49.0%
Diffusion Policy [5]*	23.1%	10.7%	01.9%	06.1%	10.5%
TinyVLA [6]*	77.6%	21.5%	11.4%	15.8%	31.6%
SmolVLA [3]	82.0%	54.2%	43.9%	**46.7%**	56.7%
PHR-VLA (ours)	**85.2%**	**55.1%**	**45.5%**	45.3%	**57.7%**

E. Inference

PHR-VLA uses future observations over the planning horizon and the auxiliary future head during training. At inference time, the future head is discarded, and the policy executes the original SmolVLA [3] action given current timestep observations.

IV. Results and Discussion

We evaluate whether PHR-VLA improves VLA fine-tuning by supervising the policy with planning-horizon latent scene dynamics. Specifically, we address the following questions: (1) Does privileged future-latent dynamics supervision improve standard VLA fine-tuning? (2) Is supervising changes in the visual latent space more effective than supervising absolute future visual embeddings? (3) Which camera view is most effective for planning-horizon reasoning?

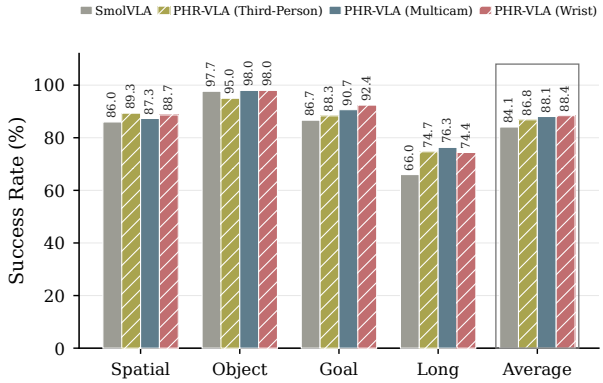

Fig. 3: **Ablation: Camera View.** Success rate of SmolVLA and PHR-VLA with wrist-mounted, fixed third-person, and multi-camera auxiliary supervision.

A. Baselines & Evaluation Protocol

Baselines. We compare PHR-VLA against recent VLA and imitation-learning baselines: ACT [7], Diffusion Policy [5], Octo [15], DiT Policy [36], OpenVLA [4], TinyVLA [6], and SmolVLA [3].

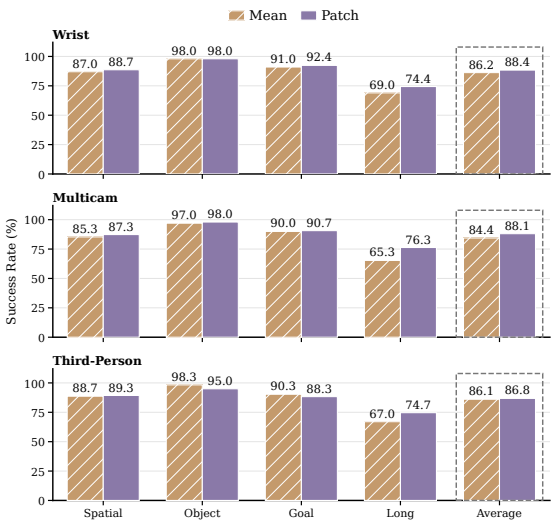

Fig. 4: **Ablation: Target Granularity.** Success rate of PHR-VLA with various target supervision granularity across wrist-mounted, fixed-third person, and multi-camera.

Evaluation Protocol. We evaluate PHR-VLA on two multi-task simulation benchmarks: LIBERO [1], and Meta-

World [2]. LIBERO is a language-conditioned tabletop manipulation benchmark that evaluates multi-task generalization across task families with varying spatial layouts, target goals, and object configurations. Meta-World is a simulated robotic manipulation benchmark that evaluates generalization across diverse manipulation tasks with varying object configurations and goal conditions.

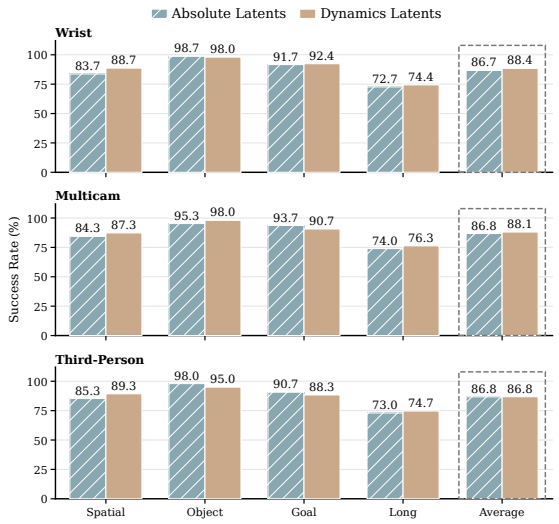

Fig. 5: **Ablation: Absolute Latent vs. Dynamics Latent Supervision.** Success rate of PHR-VLA with planning-horizon absolute latents and dynamics latent supervision during training.

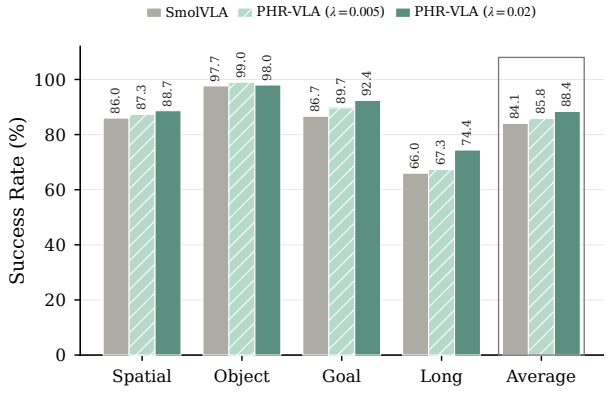

Fig. 6: **Ablation: λ.** Success rate of SmolVLA and PHR-VLA with various auxiliary loss weights (λ).

B. Results

LIBERO. As shown in Table I, under the controlled single-A100 fine-tuning setting, PHR-VLA consistently improves the SmolVLA baseline across all four LIBERO suites.. More specifically, PHR-VLA improves standard SmolVLA fine-tuning when the auxiliary objective supervises latent scene dynamics over the planning horizon. PHR-VLA increases the average LIBERO success rate from 84.1% to 88.4%. The gains are most pronounced on LIBERO-Long, where PHR-VLA improves performance by +8 points. This suggests that future-latent fine-grained dynamics supervision through wrist camera is especially useful in settings where the policy's success depends on long-horizon reasoning and manipulation.

TABLE III: **Ablation: Encoder Choice, SigLIP vs. JEPA on LIBERO.** Success rates are reported across four task suites (300 trials per task) for SigLIP and JEPA encoders auxiliary supervision. "**y**" denotes planning-horizon latent dynamics supervision, and "**s**" denotes planning-horizon absolute latent supervision. Best results are **bold**, and second-best results are underlined.

Method	View	Target	Type	Spatial ↑	Object ↑	Goal ↑	Long ↑	Average ↑
smolVLA [3]	–	–	–	86.0%	97.7%	86.7%	66.0%	84.1%
PHR-VLA (SigLIP)	Wrist	Patch	y	88.7%	98.0%	92.4%	74.4%	**88.4%**
	Wrist	Patch	s	83.7%	**98.7%**	91.7%	72.7%	86.7%
	Wrist	Mean	y	87.0%	98.0%	91.0%	69.0%	86.2%
	Wrist	Mean	s	84.0%	98.0%	89.0%	63.3%	83.6%
	Multicam	Patch	y	87.3%	98.0%	90.7%	**76.3%**	88.1%
	Multicam	Patch	s	84.3%	95.3%	**93.7%**	74.0%	86.8%
	Multicam	Mean	y	85.3%	97.0%	90.0%	65.3%	84.4%
	Multicam	Mean	s	87.0%	97.0%	91.0%	65.7%	85.2%
	Third-Person	Patch	y	89.3%	95.0%	88.3%	74.7%	86.8%
	Third-Person	Patch	s	85.3%	98.0%	90.7%	73.0%	86.8%
	Third-Person	Mean	y	88.7%	98.3%	90.3%	67.0%	86.1%
	Third-Person	Mean	s	84.0%	96.3%	90.0%	71.7%	85.5%
PHR-VLA (JEPA)	Wrist	Patch	y	87.3%	95.7%	91.3%	70.7%	86.2%
	Wrist	Patch	s	83.7%	97.3%	88.0%	69.0%	84.5%
	Wrist	Mean	y	82.0%	95.3%	89.0%	64.7%	82.8%
	Wrist	Mean	s	88.7%	94.7%	88.7%	66.0%	84.5%
	Multicam	Patch	y	87.3%	95.7%	90.0%	70.7%	85.9%
	Multicam	Patch	s	89.0%	95.0%	85.3%	67.7%	84.2%
	Multicam	Mean	y	89.3%	95.7%	92.0%	65.3%	85.6%
	Multicam	Mean	s	**91.7%**	95.3%	92.0%	63.0%	85.5%
	Third-Person	Patch	y	85.0%	95.7%	87.7%	64.3%	83.2%
	Third-Person	Patch	s	87.7%	95.0%	87.7%	63.0%	83.3%
	Third-Person	Mean	y	89.3%	95.3%	90.7%	62.3%	84.4%
	Third-Person	Mean	s	90.3%	95.3%	92.7%	63.7%	85.5%

TABLE IV: **Ablation: Encoder Choice, SigLIP vs. JEPA on Meta-World.** Success rates are reported across all difficulty levels for SigLIP and JEPA encoders auxiliary supervision. "**y**" denotes planning-horizon latent dynamics supervision, and "**s**" denotes planning-horizon absolute latent supervision. Best results are **bold**, and second-best results are underlined.

Method	Target	Type	Easy ↑	Medium ↑	Hard ↑	Very Hard ↑	Average ↑
smolVLA [3]	–	–	82.0%	54.2%	43.9%	46.7%	56.7%
PHR-VLA (SigLIP)	Patch	y	85.2%	55.1%	45.5%	45.3%	57.8%
	Patch	s	84.2%	**57.6%**	38.3%	47.3%	56.9%
	Mean	y	84.3%	55.8%	50.0%	44.7%	58.7%
	Mean	s	84.2%	**57.6%**	38.3%	47.3%	56.9%
PHR-VLA (JEPA)	Patch	y	82.0%	51.5%	48.3%	48.0%	57.5%
	Patch	s	85.2%	56.1%	43.3%	**50.0%**	58.7%
	Mean	y	**85.7%**	53.9%	51.7%	48.0%	**59.8%**
	Mean	s	85.1%	53.0%	**53.3%**	46.7%	59.5%

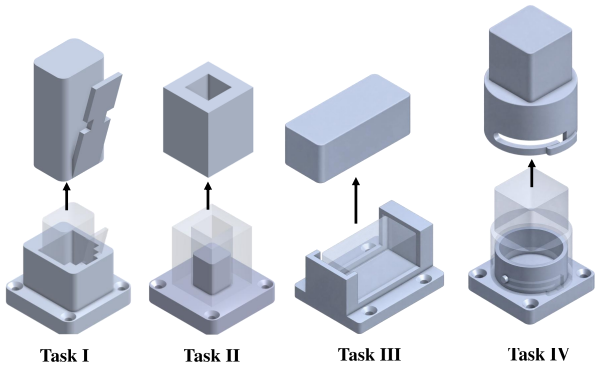

Fig. 7: **Disassembly Tasks.** Contact-rich disassembly tasks for real-world evaluation of PHR-VLA and baseline policies.

Meta-World. As shown in Table II, PHR-VLA improves the

TABLE V: **Real-world Deployment.** Success rate of PHR-VLA and baseline methods on real-world disassembly tasks. Best results are **bold**, and second-best results are underlined.

Task	Prompt	ACT [7]	Diffusion Policy [5]	SmolVLA [3]	PHR-VLA	PHR-VLA (JEPA)
Task I	Press the clip to release the object, remove it, then place it in the purple container.	19/30	21/30	26/30	28/30	**29/30**
Task II	Remove the object from the loose shaft, then place it in the purple container.	14/30	19/30	20/30	26/30	**29/30**
Task III	Slide the object inward, pull it out, then place it in the purple container.	10/30	14/30	24/30	22/30	**26/30**
Task IV	Twist the object 90 degrees and pull it out from the shaft, then place it in the purple container.	00/30	08/30	06/30	**23/30**	09/30
Total [#] ↑	–	43/120	62/120	76/120	**99/120**	93/120
Total [%] ↑	–	35.8%	51.7%	63.3%	**82.5%**	77.5%

overall average Meta-World success rate compared to the considered baseline. More specifically, PHR-VLA improves standard SmolVLA fine-tuning when the auxiliary objective supervises latent scene dynamics over the planning horizon. However, given that Meta-World simulation environment only has third-view camera, the planning-horizon reasoning through this camera cannot provide fine-grained supervision during training.

C. Ablation Studies

To identify which design choices in PHR-VLA drive the gains reported in Section IV-B, we conduct five controlled ablations that each isolate a single factor. the camera view supplying the auxiliary target, the spatial granularity of the latent target, the choice of the frozen encoder $\mathcal{E}_{\text{encoder}}$, the latent-dynamics versus absolute-latent target formulation, and the auxiliary loss weight λ. Unless otherwise stated, all ablations use LIBERO with patch-level, wrist-camera, latent-dynamics supervision at $\lambda = 0.02$ as the reference configuration, matching our main results in Table I.

Camera View. We compare three views for the auxiliary target: wrist-mounted, fixed-third person, and a multi-camera variant that concatenates latents from both, using patch-level, latent dynamics ($\mathbf{y}$) supervision at $\lambda = 0.02$. All three views improve over the SmolVLA [3] baseline (84.1%), but the wrist camera yields the largest gain (88.4%) ahead of the multi-camera configuration (88.1%), and the third-person view alone (86.8%). This is due to the fact that the wrist-camera captures the short-range, contact-centric interactions which are the most task-relevant scene dynamics over the planning horizon. The fixed third-person view instead spends capacity on background regions that evolve less informatively as illustrated in Figure 3. This finding is corroborated by our real-world disassembly results in Table V, where wrist camera supervision alone raises success rate from 63.3% to 82.5%.

Target Granularity: Patch vs. Mean. We compare patch-level against mean-pooled latent targets under latent-dynamics ($\mathbf{y}$) supervision at $\lambda = 0.02$ across all three camera views. Patch-level supervision outperforms mean pooling in every view: 88.4% vs. 86.3% on the wrist camera, 88.1% vs. 84.4 on the multi-camera configuration, and 86.8% vs. 86.1% on the third-person view as shown in Figure 4. Mean-pooling discards where the scene changes, whereas patch-level targets retain per-region dynamics, giving the future head a denser and more spatially grounded supervision signal.

Latent Dynamics vs. Absolute Latent Supervision. We compare our latent-dynamics target $\mathbf{y}$ against directly supervising absolute future latents $\mathbf{s}$ holding view (wrist), granularity (patch), and λ (0.02) fixed. The dynamics formulation reaches 88.4% versus 86.7% for the absolute variant, with the same ordering, though a narrower margin, on the third-person view (86.8% vs. 86.7%). Supervising change rather than absolute content relieves the future head of reconstructing static, task-irrelevant scene content already present in the current observation, concentrating the training signal on action relevant scene evolution as shown in Figure 5.

Loss Weight λ. We also sweep the auxiliary loss weight λ using patch-level, wrist-camera, latent-dynamics supervision. Increasing λ from 0.005 to 0.02 improves the average success rate from 85.8% to 88.4% as shown in Figure 6. We adopted $\lambda = 0.02$ as our default auxiliary loss.

Encoder Choice: SigLIP [33] vs. JEPA [34]. PHR-VLA is agnostic to the frozen encoder $\mathcal{E}_{\text{encoder}}$ used to compute planning-horizon targets. We compare SigLIP [33], a vision-language-aligned image encoder, against V-JEPA 2 [34], a self-supervised video encoder trained explicitly for dynamics prediction, across camera view, target granularity, and target formulation. On LIBERO, SigLIP is the stronger encoder in every matched configuration we evaluated, e.g., 88.4% vs. 86.3% at patch, wrist, latent future dynamics, as reported in Table III. On Meta-World, JEPA is the stronger encoder, reaching 59.8% vs. 57.8% for SigLIP, as reported in Table IV. On real-world disassembly, SigLIP supervision reaches 82.5% average success versus 77.5% for JEPA-supervision as listed in Table V. Both encoders consistently exceed their respective no-future-supervision baselines across all three benchmarks.

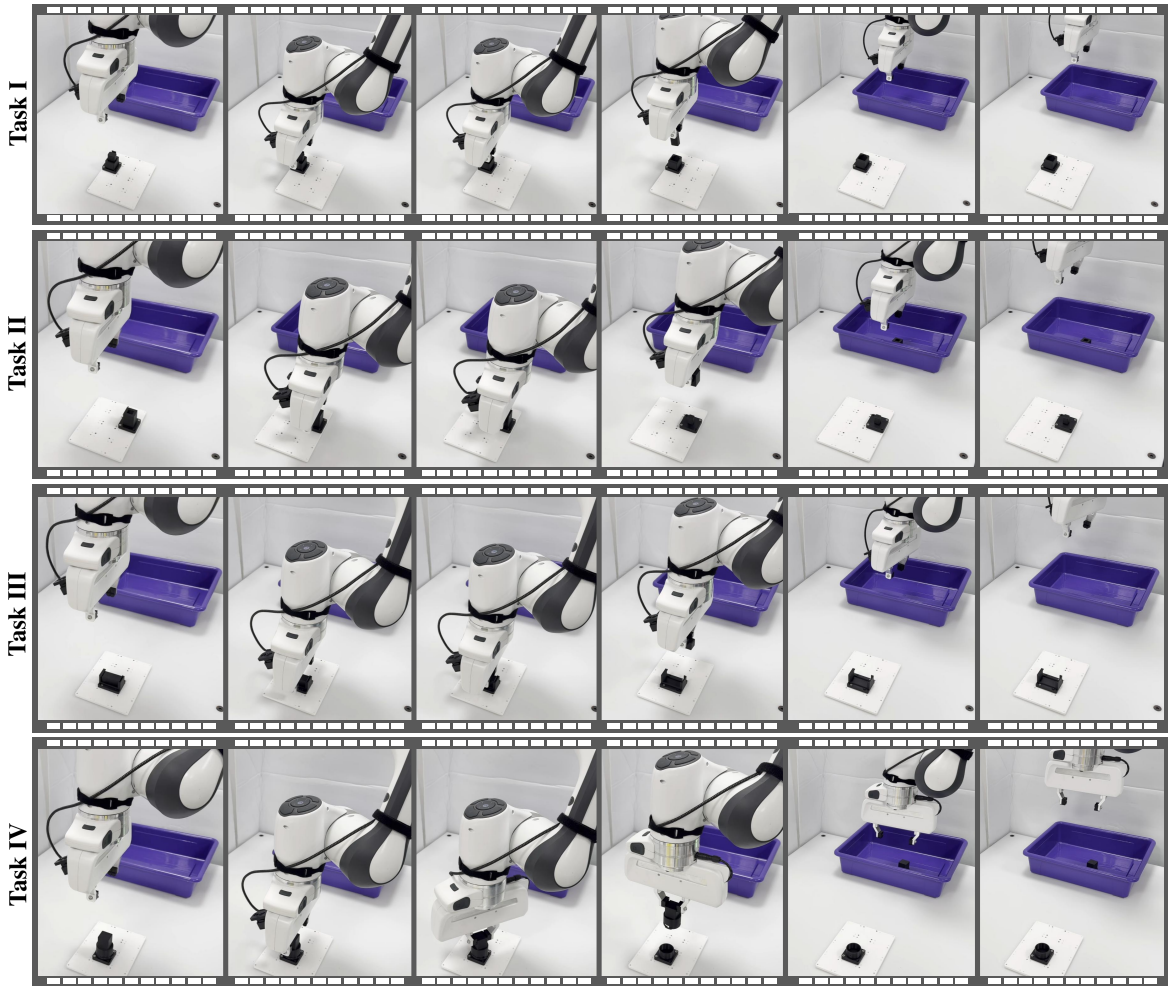

Fig. 8: **PHR-VLA Deployment in Real-world.** Examples of PHR-VLA deployment across the considered disassembly tasks, illustrating the policy execution across subsequent frames.

D. Real-world Deployment

Experimental Setup. Our experiment platform consists of a 7-DoF Franka Emika Panda robotic arm equipped with parallel grippers. For visual perception, we employ two RGB cameras: a fixed third-view camera providing a global scene view, and a wrist mounted camera offering close-range, fine-grained observations.

Disassembly Tasks. We design a set of contact-rich disassembly tasks where each scenario requires fine-grained adjustments for successful disassembly as shown in Figure 7. The specific disassembly instructions/prompts for each task are listed in Table V.

To train and evaluate PHR-VLA, we collect a real-world dataset on these disassembly tasks. The dataset contains 100 demonstrations for each task collected by human teleoperation. Visual data consists of RGB images captured from a fixed third-view and wrist-mounted camera. Language data provides prompt instructions for each task as shown in Table V. All modality streams are temporally aligned and synchronized to the frequency of 10 Hz at each timestep.

Results and Analysis. Table V reports the success rates across four disassembly tasks among our models (PHR-VLA, PHR-VLA (JEPA)), and ACT [7], Diffusion Policy [5], and SmolVLA [3] baselines, and Figure 8 illustrates examples of successful deployment of PHR-VLA across the disassembly tasks. PHR-VLA and PHR-VLA (JEPA) achieve the best overall performance, with an average success rate of 82.5% and 77.5% outperforming ACT (35.8%), Diffusion Policy (51.7%), and baseline SmolVLA (63.3%). These results demonstrate that reasoning over planning horizon through the wrist-mounted camera provides high quality supervision signal for contact-rich disassembly tasks.

V. Conclusions

We presented PHR-VLA, a training-time planning-horizon supervision framework for VLA policies. PHR-VLA aligns action-token representations with privileged latent dynamics computed from planning-horizon demonstration frames, then discards the future head and auxiliary encoder at inference. PHR-VLA improves the success rate consistently across standard manipulation benchmarks and real-world disassembly

tasks. Additionally, our ablations demonstrate that target structure, not merely the presence of a future-prediction objective, drives the gain. Patch-level targets outperform mean-pooled targets, wrist-camera supervision outperforms third-person and multi-camera supervision, and supervising latent dynamics outperforms supervising absolute future latents, consistently across configurations we tested.

One of PHR-VLA's limitations is that it is an auxiliary training objective, not an inference-time planner or world model: it only shapes the policy's representations during fine-tuning but does not perform test-time correction. Extending PHR-VLA with tactile or force-aware future targets, and object-centric patch supervision are promising directions for future work.

REFERENCES

[1] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[2] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[3] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[4] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[6] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen *et al.*, "Tinyvla: toward fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3988–3995, 2025.

[7] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[8] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[9] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[10] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[11] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[12] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.

[13] M. Torne, K. Pertsch, H. Walke, K. Vedder, S. Nair, B. Ichter, A. Z. Ren, H. Wang, J. Tang, K. Stachowicz *et al.*, "Mem: Multi-scale embodied memory for vision language action models," *arXiv preprint arXiv:2603.03596*, 2026.

[14] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior generation with latent actions," *arXiv preprint arXiv:2403.03181*, 2024.

[15] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[16] Z. Zhang, H. Xu, Z. Yang, C. Yue, Z. Lin, H.-a. Gao, Z. Wang, and H. Zhao, "Ta-vla: Elucidating the design space of torque-aware vision-language-action models," *arXiv preprint arXiv:2509.07962*, 2025.

[17] J. Chen, H. Fang, C. Wang, S. Wang, and C. Lu, "History-aware visuomotor policy learning via point tracking," *arXiv preprint arXiv:2509.17141*, 2025.

[18] M. S. Mark, J. Liang, M. Attarian, C. Fu, D. Dwibedi, D. Shah, and A. Kumar, "Bpp: Long-context robot imitation learning by focusing on key history frames," *arXiv preprint arXiv:2602.15010*, 2026.

[19] F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao, "Onetwovla: A unified vision-language-action model with adaptive reasoning," *arXiv preprint arXiv:2505.11917*, 2025.

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

[21] Z. Zhang, S. Yang, Q. Hu, L. J. Huang, J. Hou, Y. Sun, Y. Lu, and S. Han, "Foreact: Steering your vla with efficient visual foresight planning," *arXiv preprint arXiv:2602.12322*, 2026.

[22] L. Li, Q. Zhang, Y. Luo, S. Yang, R. Wang, F. Han, M. Yu, Z. Gao, N. Xue, X. Zhu *et al.*, "Causal world modeling for robot control," *arXiv preprint arXiv:2601.21998*, 2026.

[23] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang *et al.*, "World action models are zero-shot policies," *arXiv preprint arXiv:2602.15922*, 2026.

[24] T. Yuan, Z. Dong, Y. Liu, and H. Zhao, "Fast-wam: Do world action models need test-time future imagination?" *arXiv preprint arXiv:2603.16666*, 2026.

[25] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[26] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne *et al.*, "Flare: Robot learning with implicit world modeling," *arXiv preprint arXiv:2505.15659*, 2025.

[27] H. Fang, J. Duan, D. Clay, S. Wang, S. Liu, W. Huang, X. Fan, W.-C. Tsai, S. Chen, Y. R. Wang *et al.*, "Molmoact2: Action reasoning models for real-world deployment," *arXiv preprint arXiv:2605.02881*, 2026.

[28] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 54 277–54 296.

[29] P. Intelligence, A. Amin, R. Aniceto, A. Balakrishna, K. Black, K. Conley, G. Connors, J. Darpinian, K. Dhabalia, J. DiCarlo *et al.*, "$\pi_{0.6}$: a vla that learns from experience," *arXiv preprint arXiv:2511.14759*, 2025.

[30] P. Intelligence, B. Ai, A. Amin, R. Aniceto, A. Balakrishna, G. Balke, K. Black, G. Bokinsky, S. Cao, T. Charbonnier *et al.*, "$\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities," *arXiv preprint arXiv:2604.15483*, 2026.

[31] Y. Guo, T. Lee, L. X. Shi, J. Chen, P. Liang, and C. Finn, "Vlaw: Iterative co-improvement of vision-language-action policy and world model," *arXiv preprint arXiv:2602.12063*, 2026.

[32] C. Xu, J. T. Springenberg, M. Equi, A. Amin, A. Esmail, S. Levine, and L. Ke, "Rl token: Bootstrapping online rl with vision-language-action models," *arXiv preprint arXiv:2604.23073*, 2026.

[33] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

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

[35] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[36] Z. Hou, T. Zhang, Y. Xiong, H. Pu, C. Zhao, R. Tong, Y. Qiao, J. Dai, and Y. Chen, "Diffusion transformer policy," *arXiv preprint arXiv:2410.15959*, 2024.