S-VAM: Shortcut Video-Action Model by Self-Distilling Geometric and Semantic Foresight

Haodong Yan[1,*], Zhide Zhong[1,*], Jiaguan Zhu[1], Junjie He[1], Weilin Yuan[1], Wenxuan Song[1], Xin Gong[1], Yingjie Cai[2], Guanyi Zhao[2], Xu Yan[2], Bingbing Liu[2], Ying-Cong Chen[1], and Haoang Li[1] (✉)

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Huawei Foundation Model Department
haoangli@hkust-gz.edu.cn

Abstract. Video action models (VAMs) have emerged as a promising paradigm for robot learning, owing to their powerful visual foresight for complex manipulation tasks. However, current VAMs, typically relying on either slow multi-step video generation or noisy one-step feature extraction, cannot simultaneously guarantee real-time inference and high-fidelity foresight. To address this limitation, we propose S-VAM, a shortcut video-action model that foresees coherent geometric and semantic representations via a single forward pass. Serving as a stable blueprint, these foreseen representations significantly simplify the action prediction. To enable this efficient shortcut, we introduce a novel self-distillation strategy that condenses structured generative priors of multi-step denoising into one-step inference. Specifically, vision foundation model (VFM) representations extracted from the diffusion model's own multi-step generated videos provide teacher targets. Lightweight decouplers, as students, learn to directly map noisy one-step features to these targets. Extensive experiments in simulation and the real world demonstrate that our S-VAM outperforms state-of-the-art methods, enabling efficient and precise manipulation in complex environments. Our project page is https://haodong-yan.github.io/S-VAM/.

Keywords: Robot Learning · Video-Action Models · Vision Foundation Models

1 Introduction

Vision-Language-Action (VLA) models [3, 15, 18, 37, 44, 50, 52] have emerged as a dominant paradigm for robotic manipulation. By equipping pretrained vision-language models (VLMs) with action heads and finetuning on robot action data, these models learn to predict executable actions from visual observations and instructions. However, standard VLMs are pre-trained on static image-text pairs, inherently lacking the spatiotemporal foresight essential for dynamic interaction. Consequently, VLA models must learn physical dynamics directly from robot

* Equal contribution. ✉ Corresponding author.

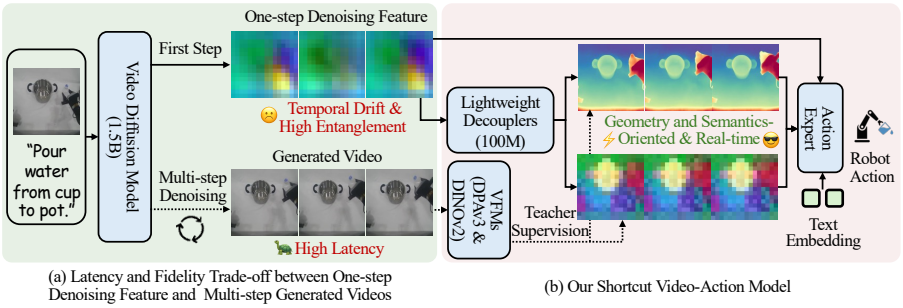


Fig. 1: Motivation and Overview of our Shortcut Video-Action Model. (a) Current video-action models struggle with a trade-off: *one-step feature extraction* is fast but yields noisy and entangled representations, whereas *multi-step video generation* predicts precise future states but is too slow for real-time control. (b) To address this, we propose a shortcut video-action model that foresees coherent geometric and semantic representations via a single forward pass. Specifically, we introduce a self-distillation strategy that extracts vision foundation model (VFM) representations (DPAv3 [25] and DINOv2 [28]) from the diffusion model's own multi-step generated videos to serve as teacher supervision exclusively during training (dashed path). By employing lightweight decouplers as students to map entangled one-step features to these geometry and semantics-oriented targets, our approach condenses structured generative priors of multi-step denoising into one-step inference, thereby enabling real-time and precise action prediction.

action data. This reliance causes a critical bottleneck, as high-quality robot action data is significantly more expensive to collect compared to abundant web-scale video data, thereby severely limiting the scalability of direct VLA approaches [20].

To address the above challenges, recent work [4,8,10,11,14,22,34,43] has increasingly adopted a video-action model (VAM) pipeline: a video diffusion model (VDM) synthesizes a visual plan conditioned on the current observation, in tandem with an action expert (AE) predicting the corresponding control signals. This design leverages the predictive priors learned from abundant internet-scale videos within the VDM, which enables the AE to learn precise policies while relying on substantially less robot action data.

As illustrated in Fig. 1(a), current VAMs mainly fall into two paradigms. The first paradigm [4,10,11,17,45] relies on generating future videos to guide control. While these videos provide detailed planning priors, the prohibitive latency of multi-step denoising renders it unsuitable for high-frequency closed-loop control. To enable efficient inference, the second paradigm (e.g., Video Prediction Policy (VPP) [14]) utilizes one-step denoising features, driven by the insight that initial denoising steps focus on establishing a global future blueprint, while later steps refine local details. However, these one-step features remain noisy and highly entangled, exhibiting poor temporal coherence (e.g., the representation of gripper drifts across frames shown in Fig. 1(a)). Such entangled representations lack

geometry-oriented cues needed to compensate for monocular depth ambiguity and the semantic distinctiveness required to distinguish task-relevant objects from irrelevant elements in the scene.

To overcome the above dilemma between inference latency and foresight fidelity in existing VAMs, we propose S-VAM, a shortcut video-action model that foresees coherent geometric and semantic representations via a single forward pass for guiding precise action generation. As shown in Fig. 1(b), our S-VAM employs relatively lightweight decouplers to explicitly disentangle noisy one-step diffusion features into specialized vision foundation model (VFM) representations. Crucially, to enable this shortcut, we introduce a novel self-distillation strategy. During training, the decouplers serve as student networks, directly supervised by stable VFM representation targets extracted from videos generated by the VDM via multi-step denoising. Specifically, we select DPAv3 [25] to provide dynamic geometry-oriented targets, and DINOv2 [28] to yield patch-level semantic targets. This self-distillation process condenses structured generative priors of multi-step denoising into one-step inference to drive efficient and precise robotic manipulation. To summarize, our contributions are as follows:

- We propose S-VAM, a shortcut video-action model that foresees coherent geometric and semantic representations in a single feed-forward pass to drive precise action generation.
- To enable this efficient shortcut, we introduce a self-distillation strategy where lightweight decouplers serve as students to directly map entangled one-step denoising features to geometry and semantics-oriented VFM teacher targets derived from multi-step generated videos.
- Extensive experiments in simulation and the real world demonstrate that our S-VAM outperforms state-of-the-art methods, enabling efficient and precise manipulation in complex scenarios.

2 Related Works

2.1 Vision-Language-Action Models

Vision-Language-Action models [6, 18, 24, 26, 36, 37, 52] establish an end-to-end mapping from text instructions and visual observations to executable actions. A prevailing paradigm in recent VLA research involves fine-tuning pre-trained VLMs on robot action data. Representative frameworks such as RT-2 [52], OpenVLA [18], π_0 [3], $\pi_{0.5}$ [15], and CogVLA [23] leverage this strategy to achieve promising performance across diverse manipulation tasks. However, these methods inherently struggle to capture the spatiotemporal foresight essential for dynamic interaction, as their underlying VLMs are primarily pretrained on static image-text pairs. Distinct from standard VLA paradigms, we augment the action model with pre-trained video generation models that inherently encode rich spatiotemporal dynamics and physical laws from internet-scale video datasets.

2.2 Video Generation Models for Robot Learning

Recent research [4, 10–12, 14, 17, 29, 43, 49, 51] has increasingly explored leveraging video generation models for their capability of visual foresight. A common approach [1, 4, 10, 11] guides action generation by synthesizing future videos. Under this framework, a VDM drafts a visual plan from the observation alongside an action expert that outputs control signals. Despite their effectiveness, these methods suffer from prohibitive inference latency due to the multi-step denoising iterations required for high-quality generation. To address this issue, some works [14, 43] directly extract one-step denoising features from video generation models to serve as future representations. However, these one-step features are noisy and highly entangled, which hinders precise action prediction. To unlock the full potential of these informative but entangled features, we introduce a shortcut that maps the one-step features to coherent geometric and semantic representations, achieving efficient inference without compromising the high-fidelity planning priors essential for precise control.

2.3 Vision Foundation Models in Robot Learning

VFMs have been extensively adopted in the field of robot learning. A prevalent paradigm [3, 18, 26, 37] leverages DINO [28, 35] and SigLIP [48] encoders to capture fine-grained local and global semantic context from observation frames. To further enhance geometric understanding, several approaches incorporate explicit 3D cues. For instance, SpatialVLA [30] utilizes off-the-shelf depth estimators to generate pseudo point clouds as input, and DepthVLA [47] leverages estimated depth maps to train a dedicated depth expert for geometric reasoning. Distinct from these explicit injection methods, Spatial Forcing [19] proposes a simple yet effective alignment strategy that implicitly compels VLA models to acquire geometry-aware knowledge. However, these methods are typically restricted to encoding or aligning the *current* static observation, missing the critical capability of *future* visual foresight. In contrast, our shortcut video-action model directly foresees future geometric and semantic VFM representations.

3 Method

As shown in Fig. 2, we propose S-VAM, a shortcut video-action model that foresees geometric and semantic representations for action generation without multi-step denoising. Specifically, we first introduce the preliminaries in Sec. 3.1. We then detail geometric and semantic decouplers in Sec. 3.2, which disentangle noisy one-step denoising features into future VFM representations. Finally, in Sec. 3.3, we describe how the Uni-Perceiver module fuses these representations with original diffusion features to condition the downstream diffusion policy.

3.1 Preliminaries

Video Diffusion Models. Our framework leverages Stable Video Diffusion (SVD) [5] for the video generation backbone. SVD utilizes a VAE to encode

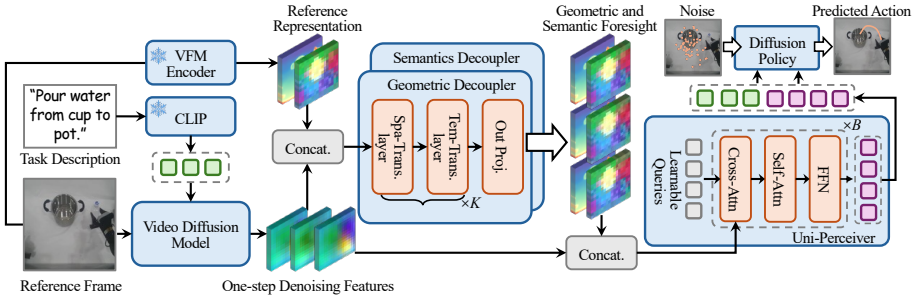


Fig. 2: Architecture of our S-VAM. The core technical novelty lies in establishing a **shortcut** that bypasses the prohibitive latency of iterative video generation. Specifically, specialized decouplers disentangle highly entangled one-step diffusion features into coherent geometric and semantic foresight. This foresight is then aggregated with original features by a Uni-Perceiver, providing a holistic condition context for the downstream diffusion policy to predict precise robot action.

the input video $V \in \mathbb{R}^{T\times 3\times H\times W}$ into a compressed latent representation $z \in \mathbb{R}^{T\times C\times h\times w}$. Operating within this latent space, the model generates a video by iteratively removing Gaussian noise from a completely noisy latent $z_S \sim \mathcal{N}(0, \mathbf{I})$ through a reverse denoising process:

$$z_{s-1} = \frac{1}{\sqrt{\alpha_s}}\left(z_s - \frac{1-\alpha_s}{\sqrt{1-\bar{\alpha}_s}}\epsilon_\theta(z_s, s, P, I)\right) + \sigma_s\epsilon, \tag{1}$$

where $\epsilon_\theta(\cdot)$ is a learnable noise estimator and $\epsilon_\theta(z_s, s, P, I)$ is the predicted noise conditioned on the observed frame I and task description P. However, such multi-step sampling leads to high inference latency that hinders real-time control.

Vision Foundation Models. Pre-trained on internet-scale datasets, VFMs have demonstrated remarkable capabilities in learning highly structured and regularized representation spaces. We leverage frozen VFMs to extract these representations from the model's own multi-step generated video frames $\hat{V}$ to serve as teacher targets for self-distilling geometric and semantic foresight (see Sec. 3.2). Formally, given a VFM, we extract the target representation $Y \in \mathbb{R}^{T\times C_{\text{VFM}}\times h\times w}$ as:

$$Y = \Phi(\text{Interpolate}(\hat{V})), \tag{2}$$

where $\Phi(\cdot)$ denotes the VFM encoder and Interpolate$(\cdot)$ is the linear interpolation operation that aligns the output feature maps to the target spatial resolution $h \times w$. Here, C_{VFM} denotes the output channel dimension determined by the specific VFM.

3.2 Geometric and Semantic Foresight Distillation

One-step Denoising Features. To enable efficient inference, we extract predictive features via a single denoising forward pass following [14]. Specifically, we

leverage multi-layer intermediate features from the up-sampling blocks. While inherently noisy and entangled, these internal representations are highly informative, thereby providing essential cues to be distilled into coherent geometric and semantic foresight.

Let $F_l \in \mathbb{R}^{T\times C_l\times h_l\times w_l}$ denote the feature map from the l^{th} up-sampling layer of denoising network $\epsilon_\theta(Z_S, S, P, I)$ at the first denoising step S. To effectively aggregate these features across different resolutions, we align each layer to a unified target spatial dimension $h \times w$ via linear interpolation Interpolate$(\cdot)$:

$$F_l' = \text{Interpolate}(F_l), \quad F_l' \in \mathbb{R}^{T\times C_l\times h\times w}. \tag{3}$$

The final comprehensive one-step feature F is constructed by concatenating these aligned layers along the channel dimension:

$$F = \text{Concat}((F_0', \ldots, F_L'), \text{dim} = 1), \quad F \in \mathbb{R}^{T\times C_\Sigma\times h\times w}, \tag{4}$$

where $C_\Sigma = \sum_{l=0}^{L} C_l$ denotes the total channel dimension of the aggregated features.

Geometric and Semantic Decouplers. To exploit the rich but entangled dynamics within one-step features F, we introduce the geometric and the semantic decouplers that directly disentangle these noisy one-step features into coherent geometric and semantic VFM representations. Specifically, the geometric branch targets the representation space of DPAv3 [25], and the semantic branch targets that of DINOv2 [28]. This design choice is motivated by their complementary strengths: DPAv3 provides dynamic geometric structures, and DINOv2 offers patch-level semantic distinctiveness.

In terms of architecture, we instantiate each decoupler as a spatio-temporal transformer. For each branch $i \in \{\text{geo}, \text{sem}\}$, instead of inferring future states solely from the noisy diffusion features, the module leverages the reference representation of the current observation, $Y_i^{\text{ref}} = \Phi_i(\text{Interpolate}(I_{\text{obs}})) \in \mathbb{R}^{1\times C_i\times h\times w}$, as a strong conditioning anchor. This anchor eases the optimization burden, effectively guiding the network to map unstable diffusion features into coherent VFM representations. Formally, we temporally replicate the reference representation Y_{ref}^i to match the sequence length of the one-step feature volume F, and concatenate them along the channel dimension to obtain the fused input representation:

$$\tilde{F}_i^0 = \text{Concat}((F, \text{Repeat}(Y_i^{\text{ref}})), \text{dim} = 1), \quad i \in \{\text{geo}, \text{sem}\}. \tag{5}$$

Then, this fused representation $\tilde{F}_i^0$ is first projected into a compact latent space via a linear embedding layer ($\mathbb{R}^{T\times(C_i+C_\Sigma)\times h\times w} \to \mathbb{R}^{T\times C_{\text{hidden}}\times h\times w}$). The compressed features are then contextualized through a stack of K factorized spatio-temporal blocks. Specifically, for each block k ($1 \le k \le K$), the features are processed sequentially by spatial and temporal aggregation:

$$\tilde{F}_i^k = \mathcal{T}_i^k\left(\mathcal{S}_i^k(\tilde{F}_i^{k-1})\right), \quad i \in \{\text{geo}, \text{sem}\}, \tag{6}$$

where $\mathcal{S}(\cdot)$ and $\mathcal{T}(\cdot)$ denote the spatial and temporal transformer layers, respectively. Finally, the output from the K^{th} block, $\tilde{F}_i^K$, is mapped back to the target VFM dimension via an output projection ($\mathbb{R}^{T\times C_{\text{hidden}}\times h\times w} \rightarrow \mathbb{R}^{T\times C_i\times h\times w}$).

Self-Distillation Optimization Objectives. As indicated by the dashed lines in Fig. 1(b), we apply specific supervision to each branch by introducing a self-distillation mechanism, where VFM representations extracted from the diffusion model's own multi-step generated video $\hat{V}$ serve as teacher signals. The core advantage of this self-distillation is that the generated video originates from the same diffusion trajectory as the one-step features. In contrast, relying on targets extracted from ground-truth future frames would introduce an inherent trajectory misalignment that hinders optimization. Given that geometric and semantic branches target distinct representation spaces, we optimize them independently. Utilizing the extraction process in Eq. (2) to obtain the teacher representations Y_i, we formulate the optimization objective for each branch $i \in \{\text{geo}, \text{sem}\}$:

$$\mathcal{L}_i = \|\tilde{F}_i^K - Y_i\|_2^2, \quad i \in \{\text{geo}, \text{sem}\}. \tag{7}$$

By minimizing $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{sem}}$, our S-VAM effectively condenses the multi-step generation capability into our single-step shortcut, yielding high-fidelity foresight without the latency of iterative denoising.

3.3 Action Expert

To bridge the gap between decoupled geometric and semantic foresight and physical control, our action expert employs a Uni-Perceiver to aggregate a holistic context, guiding a diffusion policy to predict actions.

Uni-Perceiver. We first construct a holistic conditioning context, denoted as $\mathcal{C} = \text{Concat}(\tilde{F}_{\text{geo}}^K, \tilde{F}_{\text{sem}}^K, F) \in \mathbb{R}^{T\times C_{\text{hol}}\times h\times w}$, by concatenating the geometric and semantic foresight with the original one-step diffusion feature along the feature channel dimension. Retaining the original feature F is critical, as it preserves residual global context that complements the explicit geometric and semantic foresight. To mitigate the curse of dimensionality and the associated computational overhead introduced by $\mathcal{C}$, we compress it into a highly condensed token representation $F_{\text{agg}} \in \mathbb{R}^{N\times C_{\text{agg}}}$. This condensation effectively filters out redundant information to facilitate the diffusion policy in learning to predict precise actions. Specifically, we initialize a compact set of N learnable latent queries $\mathcal{Q}$ that aggregate spatiotemporal information through a QFormer-style [16, 21] token condensation module:

$$F_{\text{agg}} = \text{FFN}(\text{SelfAttn}(\text{CrossAttn}(\mathcal{Q}, \mathcal{C}))). \tag{8}$$

Here, the cross-attention layer $\text{CrossAttn}(\cdot)$ aggregates salient spatiotemporal features from the holistic context $\mathcal{C}$ into latent queries $\mathcal{Q}$, while the subsequent self-attention layer $\text{SelfAttn}(\cdot)$ models the internal interactions among these N condensed tokens. Finally, a feed-forward network $\text{FFN}(\cdot)$ further enhances the expressiveness of the condensed tokens.

Diffusion Policy. Finally, we incorporate the aggregated representation F_{agg} and the text embedding E of the task description into a Diffusion Transformer (DiT) architecture via cross-attention layers. The policy is trained to reconstruct the ground-truth action sequence a_0 from the pure noise a_J by predicting the added noise ϵ. The training objective is formulated as follows:

$$\mathcal{L}_A = \mathbb{E}_{j,a_j,\epsilon}\left[\|\epsilon - \epsilon_\phi(a_j, F_{\text{agg}}, E, j)\|_2^2\right], \tag{9}$$

where a_j is the noisy action at diffusion timestep j, and $\epsilon_\phi(\cdot)$ denotes the noise prediction network.

4 Experiment

We evaluate our S-VAM on both simulated and real-world robotic tasks to answer the following questions: (**Q1**) Can S-VAM outperform state-of-the-art methods? (**Q2**) How does each key component of S-VAM (e.g., geometric distillation, semantic distillation, and Uni-Perceiver) contribute to the overall performance? (**Q3**) How do alternative VFM representations affect distillation and downstream control? (**Q4**) Can S-VAM adapt to complex real-world scenarios?

4.1 Experimental Setup

Simulated Benchmarks. We evaluate our S-VAM on two challenging manipulation benchmarks. 1) CALVIN [27] ABC→D includes four distinct tabletop environments (ABCD). Policies are trained on ABC and evaluated on unseen D to assess generalization across consecutive long-horizon tasks. 2) MetaWorld [46] features a Sawyer robot performing 50 manipulation tasks across varying difficulty levels as defined in [33]. The training set consists of 50 demonstrations for each task. For fair comparison, all experiments are conducted following [14] under a third-view setup using only the primary monocular RGB camera.

Implementation Details. Our pipeline consists of a three-stage training process. In the first stage, we fine-tune the SVD backbone, initialized from weights pretrained on embodied scenes [14]. We train for 100k steps on MetaWorld and 40k steps on real-world tasks using 4 NVIDIA H100 GPUs. For CALVIN, we directly adopt the fine-tuned model from [14] without additional training. In the second stage, we freeze the video generation model and train the geometric and semantic decouplers to distill coherent foresight from noisy one-step features. This self-distillation is efficient, requiring only 50k steps on a single NVIDIA H100 GPU for all benchmarks. In the final stage, while keeping the SVD backbone and decouplers frozen, we exclusively train the action expert to decode control signals. We train this stage for 60k steps on CALVIN and 40k steps on the remaining benchmarks, using four NVIDIA H100 GPUs. For inference, we use a single NVIDIA RTX 3090 (24GB) GPU. In qualitative analysis, we additionally train a DPT head [32] that maps geometric foresight to depth maps following a standard probing protocol [19, 41].

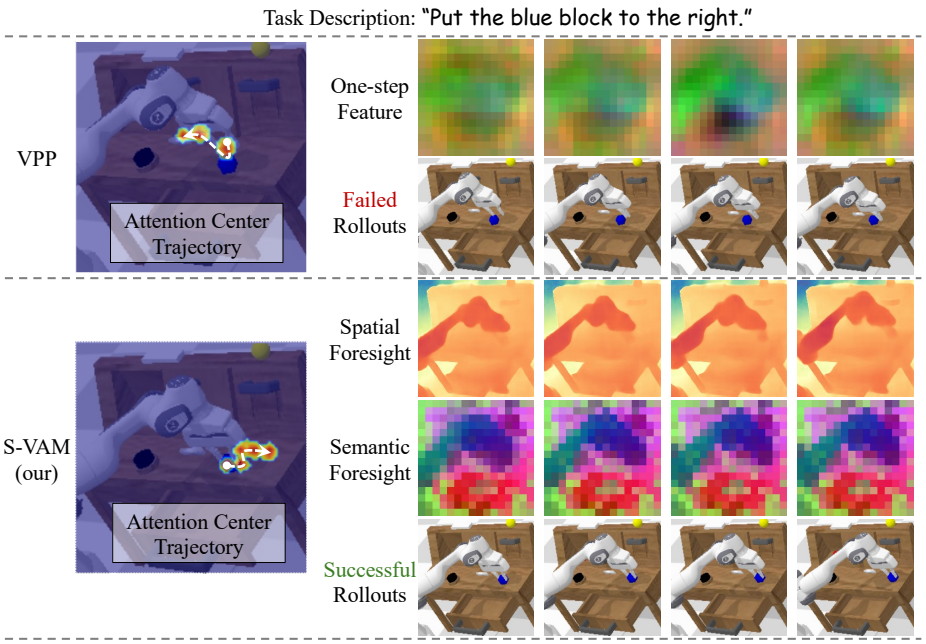


Fig. 3: Qualitative comparison on CALVIN [27]. VPP [14] utilizes entangled one-step features, resulting in an erratic attention trajectory that explicitly contradicts the language instruction and leads to failed actuation. In contrast, our S-VAM foresees geometric and semantic representations. This decoupled future blueprint enables the action expert to anchor a coherent attention trajectory that perfectly aligns with the language instruction, thereby ensuring successful execution. *Note: Geometric foresight is visualized as probe-based depth maps* [19, 41]*; one-step and semantic features are visualized via PCA computed globally across the entire sequence.*

Methods for Comparison. We compare our S-VAM against state-of-the-art baselines, broadly categorized into two paradigms. **Direct action learning** methods (e.g., RT-1 [6], Diffusion Policy [9], OpenVLA [18], CLOVER [7], π_0 [3], and Spatial Forcing [19]) directly map observations to actions without explicitly modeling future dynamics, whereas **predictive methods** (e.g., SuSIE [4], GR-1 [42], VPP [14], Uni-VLA [40], and HiF-VLA [26]) predict future states or intermediate representations to guide action generation.

4.2 Evaluations Results on Simulated Benchmarks (Q1)

Results on CALVIN [27].

As demonstrated in Tab. 1, our S-VAM achieves a state-of-the-art average length of **4.16**, outperforming both direct action learning and predictive baselines. Notably, our S-VAM surpasses our direct baseline, VPP [14], by a significant margin of 0.58. As shown in Fig. 3, VPP [14] relies on entangled one-step

Table 1: Quantitative comparison on CALVIN [27]. We report stage-wise success rates (1-5) and average sequence length (Avg. Len.). **Best** and second-best results are highlighted accordingly.

Category	Method	i^{th} Task Success Rate					Avg. Len. ↑
		1	2	3	4	5	
Direct Action Learning Methods	OpenVLA [18]	91.3	77.8	62.0	52.1	43.5	3.27
	CLOVER [7]	**96.0**	83.5	70.8	57.5	45.4	3.53
	π_0 [3]	93.7	83.2	74.0	62.9	51.0	3.65
	Spatial Forcing [19]	93.6	85.8	78.4	72.0	64.6	3.94
Predictive Methods	SuSIE [4]	87.0	69.0	49.0	38.0	26.0	2.69
	VPP [14]	90.9	81.5	71.3	62.0	51.8	3.58
	Uni-VLA [40]	95.5	85.8	74.8	66.9	56.5	3.80
	HiF-VLA [26]	93.5	87.4	81.4	75.9	**69.4**	4.08
	S-VAM(our)	95.8	**90.7**	**83.7**	**77.0**	68.9	**4.16**

features, which leads to an erratic attention center trajectory[3] that explicitly contradicts the language instruction, resulting in imprecise actuation. In contrast, by distilling decoupled geometric and semantic foresight, our S-VAM anchors a coherent attention trajectory that perfectly aligns with the language instruction. Our execution rollouts strictly follow this high-fidelity blueprint, confirming that the action expert successfully translates these future representations into precise manipulation. Furthermore, our S-VAM outperforms Spatial Forcing [19] by distilling future states instead of merely aligning representations on observation frames, enabling the agent to better anticipate complex interactions. Additionally, our S-VAM surpasses other predictive HiF-VLA [26] and Uni-VLA [40]. Although they also employ predictive mechanisms, their underlying VLMs inherently lack the rich spatiotemporal dynamics naturally captured by our video backbone.

Results on MetaWorld [46]. As shown in Tab. 2, our S-VAM achieves the highest average success rate of **72.8%**, surpassing the previous state-of-the-art (SOTA) method VPP [14] by a clear margin. The advantage of our S-VAM is particularly demonstrated in hard tasks, where it achieves a success rate of 68.4%, significantly outperforming VPP (52.6%). Hard tasks in MetaWorld often involve complex object interactions and fine-grained geometric constraints, such as the precise assembly visualized in Fig. 4. In this task, the failure of VPP [14] largely stems from its entangled one-step features, which produce a diverging attention trajectory that completely misses the target "nut". In contrast, the decoupled geometric and semantic foresight enables our S-VAM to anchor a coherent attention trajectory that accurately traces a path to the instructed target for a precise grasp. Furthermore, S-VAM surpasses leading non-VAM meth-

[3] Using head-averaged attention maps from the final cross-attention layer of the token condensation module, we extract this trajectory by tracking the maximum activation (and its 3-pixel neighborhood) across frames.

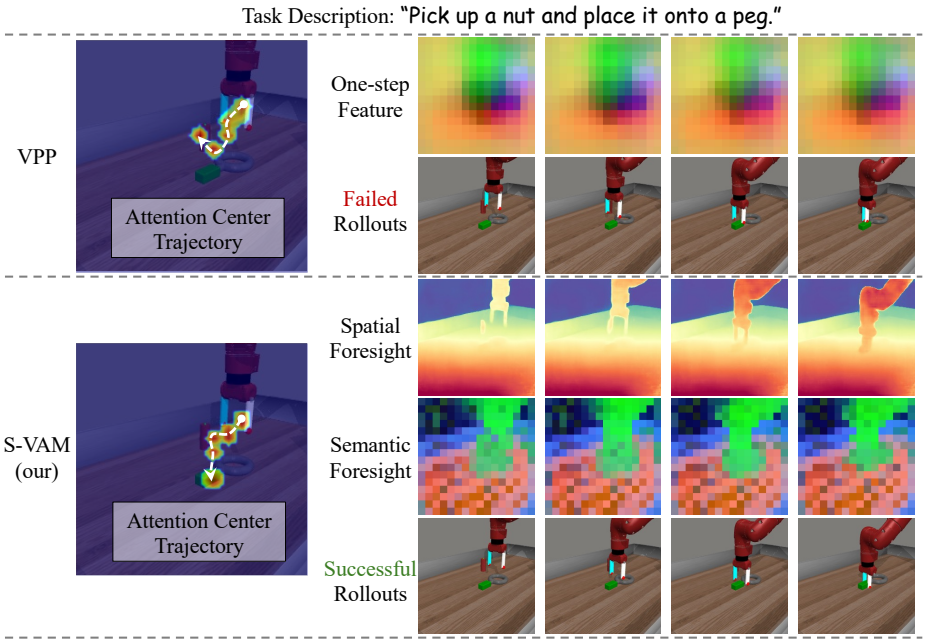


Fig. 4: Qualitative comparison on MetaWorld [46]. VPP [14] utilizes entangled one-step features, resulting in a diverging attention trajectory that completely misses the target "nut". In contrast, our S-VAM foresees explicit geometric and semantic representations. This decoupled future blueprint enables the action expert to anchor a coherent attention trajectory that accurately guides to the instructed target, ensuring successful grasping. *Note: Visualization settings follow Fig. 3.*

Table 2: Quantitative comparison on Metaworld [46]. We report success rates on easy, middle, and hard tasks, along with the overall average. **Best** and second-best results are highlighted accordingly.

Category	Method	Easy (28 tasks)	Middle (11 tasks)	Hard (11 tasks)	Average ↑ (50 tasks)
Direct Action Learning Methods	RT-1 [6]	0.605	0.042	0.015	0.346
	Diffusion Policy [9]	0.442	0.062	0.095	0.279
	Spatial Forcing [19]	0.737	0.436	0.451	0.609
Predictive Methods	SuSIE [4]	0.560	0.196	0.255	0.410
	GR-1 [42]	0.725	0.327	0.451	0.574
	HiF-VLA [26]	0.729	0.364	0.404	0.577
	VPP [14]	**0.818**	0.493	0.526	0.682
	S-VAM (our)	0.793	**0.607**	**0.684**	**0.728**

ods under the 50-demonstration setting, outperforming Spatial Forcing [19] and HiF-VLA [26] by 11.9% and 15.1%, respectively. This underscores the profound

Table 3: Ablation study on key components of our S-VAM.

Method Variant	i^{th} Task Success Rate					Avg. Len. ↑
	1	2	3	4	5	
w/o Geometric Distillation	94.1	87.1	79.3	73.5	66.5	4.01
w/o Semantic Distillation	94.0	87.1	80.4	73.2	64.1	3.99
w/o Self-Distillation	94.2	85.2	75.9	67.8	59.0	3.82
w/o Uni-Perceiver	94.0	84.6	74.7	65.1	53.8	3.72
w/o Original Diffusion Feature	95.3	86.6	77.6	70.9	62.5	3.93
S-VAM (Full)	**95.8**	**90.7**	**83.7**	**77.0**	**68.9**	**4.16**

advantage of leveraging dynamic priors from the video generation model in data-scarce scenarios.

4.3 Ablation Study (Q2)

To validate the effectiveness of key components in our S-VAM, we conduct an ablation study on the CALVIN [27] benchmark. Results are summarized in Tab. 3.
Geometric and Semantic Distillation. Removing either the geometric (4.16 → 4.01) or semantic (4.16 → 3.99) distillation distinctly degrades performance. Lacking semantic foresight particularly harms long-horizon execution (the 5^{th} task success rate drops from 68.9% to 64.1%), highlighting its critical role in maintaining consistent object identities to mitigate compounding errors.
Self-Distillation. To evaluate the necessity of the proposed mechanism, we compare our approach against a w/o Self-Distillation variant that uses VFM supervision extracted from ground-truth (GT) future frames instead of the model's own multi-step predictions. This variant degrades performance (4.16 → 3.82). The resulting decline stems from the inherent trajectory misalignment between noisy one-step features and GT future representations, highlighting that self-distilling foresight from the model's own predictions is crucial for preserving a shared diffusion trajectory.
Uni-Perceiver. Removing the Uni-Perceiver causes a severe performance drop (4.16 → 3.72). This demonstrates that adaptively condensing high-dimensional multi-modal features into a compact representation is essential for providing a holistic condition context to the action expert.
Original Diffusion Features. Relying exclusively on distilled foresight without the original one-step feature F yields sub-optimal results (4.16 → 3.93). The raw diffusion features are indispensable, as they supplement structural VFM representations by providing the action expert with residual global context.

4.4 Alternative Vision Foundation Representations (Q3)

To demonstrate the effectiveness of our chosen combination of DINOv2 [28] and DPAv3 as VFM targets, we evaluate a diverse set of alternative VFMs on the

Table 4: Quantitative comparison of different VFM representations as teacher targets.

Type	Representation Source	i^{th} Task Success Rate					Avg. Len. ↑
		1	2	3	4	5	
Semantic Targets	CLIP [31]	94.1	82.0	72.5	65.0	58.0	3.72
	SigLIP [48]	94.5	84.8	74.7	65.5	57.0	3.77
	DINOv2 [28]	94.1	87.1	79.3	73.5	66.5	4.01
	DINOv3 [35]	95.6	87.7	78.8	70.3	62.1	3.95
Geometric Targets	DPAv3 [25]	94.0	87.1	80.4	73.2	64.1	3.99
	VGGT [38]	93.5	84.3	74.8	65.6	55.8	3.74
Motion-aware Targets	VideoMAEv2 [39]	94.6	85.9	78.0	70.2	60.8	3.90
	V-JEPA2 [2]	93.1	84.3	74.3	65.6	56.4	3.74
Synergistic Targets	SigLIP & DPAv3	**96.2**	88.3	80.7	74.2	66.1	4.06
	DINOv2 & VGGT	95.7	88.8	80.9	73.9	64.9	4.04
	DINOv2 & DPAv3 (our)	95.8	**90.7**	**83.7**	**77.0**	**68.9**	**4.16**

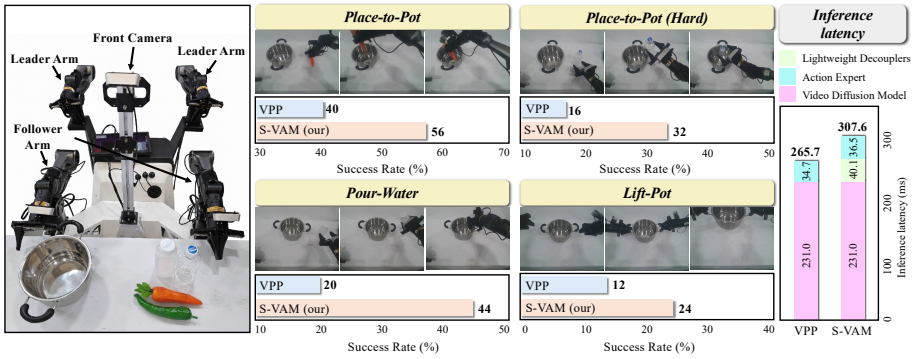
CALVIN [27] benchmark. We categorize them into semantic, geometric, motion-aware, and synergistic targets. The results are detailed in Tab. 4.

Semantic Targets (Dense vs. Global). By extracting dense patch-level representations, DINOv2 [28] and DINOv3 [35] significantly outperform global semantic models CLIP [31] and SigLIP [48] (e.g., 4.01 for DINOv2 vs. 3.72 for CLIP in average length). This gap demonstrates that while global semantics provide scene-level context, precise manipulation inherently relies on dense semantic distinctiveness for the action expert to attend to task-relevant affordances.

Geometric Targets (Static vs. Dynamic). DPAv3 [25] significantly outperforms VGGT [38] (3.99 vs. 3.74). We attribute this performance margin to DPAv3 being trained to adapt to dynamic video streams, enabling it to extract dynamic geometric structures. Conversely, VGGT focuses on static scene reconstruction, which makes it fundamentally unsuited to handle the dynamic object interactions inherent in robotic tasks.

Motion-aware Targets. We also explore using the representations extracted by video foundation models, VideoMAEv2 [39] and V-JEPA2 [2], as motion-aware targets. However, they demonstrate sub-optimal results. We hypothesize that current video models still lag behind specialized image counterparts (e.g., DINOv2) in fine-grained feature fidelity.

Synergistic Targets. The best performance is achieved by fusing complementary modalities. Our choice (DINOv2 & DPAv3) achieves the highest average sequence length of 4.16. Replacing DPAv3 with VGGT (DINOv2 & VGGT) results in a performance drop to 4.04, confirming that static-scene geometric priors are insufficient for dynamic manipulation tasks. Second, substituting DINOv2 with SigLIP (SigLIP & DPAv3) leads to a decline to 4.06, demonstrating that dense and patch-level representations are more critical for low-level control than global-level language-aligned features.



(a) Robotic Arm Platform (b) Performance of VPP *vs.* S-VAM (our)

Fig. 5: Multi-task real-world experiments. (a) We deploy our system on a dual-arm Cobot using only monocular front-camera observations. (b) Our S-VAM demonstrates a significant success rate improvement over VPP [14] on all tasks without compromising real-time control capabilities.

4.5 Multi-task Experiments in the Real World (Q4)

Task setups. As shown in Fig. 5(a), we conduct our real-world experiments on a Cobot dual-arm robot manufactured by AgileX Robotics, which adopts the Mobile ALOHA system design [13]. Each arm possesses 7 degrees of freedom and is equipped with a parallel gripper. To maintain consistency with our simulated setup, all experiments are conducted using only the front camera to capture monocular RGB observations. We design four manipulation tasks to evaluate S-VAM's effectiveness: (1) *Place-to-Pot*: picking an object and placing it into a pot; (2) *Place-to-Pot (Hard)*: picking a transparent object and placing it into a pot; (3) *Pour-Water*: pouring water from a transparent cup into a pot; and (4) *Lift-Pot*: lifting the pot with both arms. We train a unified multi-task model on roughly 50 human demonstrations per task. During evaluation, we report the average success rate over 25 trials for each task.

Results Analysis. As shown in Fig. 5(b), our S-VAM outperforms VPP [14] across all tasks. The performance gain is most evident in the *Place-to-Pot (Hard)* task involving transparent objects. In this task, VPP achieves only 16% success, as its noisy one-step features lack the coherent and stable foresight needed to resolve visual ambiguity. In contrast, our S-VAM benefits from the synergy of both decouplers: the geometric decoupler distills precise geometric priors to resolve the depth ambiguity of transparent surfaces, while the semantic decoupler extracts discriminative object features to maintain a consistent representation of the transparent entity. This distilled foresight enables the action expert to reliably anticipate the target's future states and plan precise action trajectories, boosting the success rate to 32%. Furthermore, our S-VAM takes 307.6 ms per forward pass (video diffusion backbone: 231.0 ms; decouplers: 40.1 ms; action expert: 36.5 ms), introducing a modest 15.8% overhead over VPP (265.7 ms).

By predicting an action chunk of 8 per forward pass, our S-VAM achieves an effective control frequency of 25 Hz.

5 Conclusion

In this work, we present S-VAM, a novel framework that establishes a direct shortcut to resolve the dilemma between inference efficiency and foresight fidelity inherent to the video-action paradigm. Specifically, through a novel self-distillation strategy, we demonstrate that noisy one-step diffusion features can be successfully distilled into coherent geometric and semantic foresight, supervised by stable visual foundation representations extracted from multi-step generated videos. Crucially, this shortcut mechanism allows the policy to inherit high-fidelity planning while strictly maintaining the efficiency of single-step inference. Extensive experiments validate that S-VAM not only achieves state-of-the-art performance on simulated benchmarks but also unlocks efficient and precise manipulation capabilities in complex real-world scenarios.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 62403401, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011992 and Grant 2026A1515012323, and in part by the Guangdong Provincial Project under Grant 2024QN11X127.

References

1. Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., Jaakkola, T., Tenenbaum, J., Kaelbling, L., Srivastava, A., Agrawal, P.: Compositional foundation models for hierarchical planning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 22304–22325. Curran Associates, Inc. (2023)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Mojtaba, Komeili, Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning (2025), https://arxiv.org/abs/2506.09985
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
4. Black, K., Nakamoto, M., Atreya, P., Walke, H.R., Finn, C., Kumar, A., Levine, S.: Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In: The Twelfth International Conference on Learning Representations (2024)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

6. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
7. Bu, Q., Zeng, J., Chen, L., Yang, Y., Zhou, G., Yan, J., Luo, P., Cui, H., Ma, Y., Li, H.: Closed-loop visuomotor control with generative expectation for robotic manipulation. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
8. Chen, B., Zhang, T., Geng, H., Zhang, C., Li, P., Song, K., Freeman, W.T., Malik, J., Abbeel, P., Tedrake, R., et al.: Large video planner enables generalizable robot control. arXiv preprint arXiv:2512.15840 (2025)
9. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025)
10. Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. Advances in neural information processing systems **36**, 9156–9172 (2023)
11. Feng, Y., Tan, H., Mao, X., Xiang, C., Liu, G., Huang, S., Su, H., Zhu, J.: Vidar: Embodied video diffusion model for generalist manipulation. arXiv preprint arXiv:2507.12898 (2025)
12. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. Advances in neural information processing systems **29** (2016)
13. Fu, Z., Zhao, T.Z., Finn, C.: Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117 (2024)
14. Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. In: Forty-second International Conference on Machine Learning (2025)
15. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
16. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O.J., Botvinick, M., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver IO: A general architecture for structured inputs & outputs. In: International Conference on Learning Representations (2022)
17. Kim, M.J., Gao, Y., Lin, T.Y., Lin, Y.C., Ge, Y., Lam, G., Liang, P., Song, S., Liu, M.Y., Finn, C., et al.: Cosmos policy: Fine-tuning video models for visuomotor control and planning. arXiv preprint arXiv:2601.16163 (2026)
18. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: Conference on Robot Learning. pp. 2679–2713. PMLR (2025)
19. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., ZENG, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In: The Fourteenth International Conference on Learning Representations (2026)
20. Li, H., Zhang, I., Ouyang, R., Wang, X., Zhu, Z., Yang, Z., Zhang, Z., Wang, B., Ni, C., Qin, W., et al.: Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. arXiv preprint arXiv:2509.22199 (2025)
21. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

22. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., Shen, Y., Xu, Y.: Causal world modeling for robot control. arXiv preprint arXiv:2601.21998 (2026)
23. Li, W., Zhang, R., Shao, R., He, J., Nie, L.: CogVLA: Cognition-aligned vision-language-action models via instruction-driven routing & sparsification. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
24. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., Li, H., Kong, T.: Vision-language foundation models as effective robot imitators. In: The Twelfth International Conference on Learning Representations (2024)
25. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
26. Lin, M., Ding, P., Wang, S., Zhuang, Z., Liu, Y., Tong, X., Song, W., Lyu, S., Huang, S., Wang, D.: Hif-vla: Hindsight, insight and foresight through motion representation for vision-language-action models. arXiv preprint arXiv:2512.09928 (2025)
27. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters (RA-L) **7**(3), 7327–7334 (2022)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
29. Pai, J., Achenbach, L., Montesinos, V., Forrai, B., Mees, O., Nava, E.: mimic-video: Video-action models for generalizable robot control beyond vlas. arXiv preprint arXiv:2512.15692 (2025)
30. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 (2025)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021)
33. Seo, Y., Hafner, D., Liu, H., Liu, F., James, S., Lee, K., Abbeel, P.: Masked world models for visual control. In: Conference on Robot Learning. pp. 1332–1344. PMLR (2023)
34. Shen, Y., Wei, F., Du, Z., Liang, Y., Lu, Y., Yang, J., Zheng, N., Guo, B.: VideoVLA: Video generators can be generalizable robot manipulators. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Song, W., Chen, J., Ding, P., Zhao, H., Zhao, W., Zhong, Z., Ge, Z., Li, Z., Wang, D., Ma, J., et al.: Pd-vla: Accelerating vision-language-action model integrated with action chunking via parallel decoding. arXiv preprint arXiv:2503.02310 (2025)

37. Song, W., Zhou, Z., Zhao, H., Chen, J., Ding, P., Yan, H., Huang, Y., Tang, F., Wang, D., Li, H.: Reconvla: Reconstructive vision-language-action model as effective robot perceiver. arXiv preprint arXiv:2508.10333 (2025)
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
39. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14549–14560 (2023)
40. Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., Zhang, Z.: Unified vision-language-action model. arXiv preprint arXiv:2506.19850 (2025)
41. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In: The Fourteenth International Conference on Learning Representations (2026)
42. Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., Kong, T.: Unleashing large-scale video generative pre-training for visual robot manipulation. In: International Conference on Learning Representations (2024)
43. Xie, S., Cao, H., Weng, Z., Xing, Z., Chen, H., Shen, S., Leng, J., Wu, Z., Jiang, Y.G.: Human2robot: Learning robot actions from paired human-robot videos. arXiv preprint arXiv:2502.16587 (2025)
44. Ye, A., Zhang, Z., Wang, B., Wang, X., Zhang, D., Zhu, Z.: Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623 (2025)
45. Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y.L., Zhu, C., Xiang, J., et al.: World action models are zero-shot policies. arXiv preprint arXiv:2602.15922 (2026)
46. Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., Levine, S.: Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Conference on robot learning. pp. 1094–1100. PMLR (2020)
47. Yuan, T., Liu, Y., Lu, C., Chen, Z., Jiang, T., Zhao, H.: Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. arXiv preprint arXiv:2510.13375 (2025)
48. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
49. Zhao, G., Wang, Y., Wang, X., Zhu, Z., Yu, T., Huang, G., Zai, Y., Jiao, J., Xue, C., Wang, X., et al.: Unidrivedreamer: A single-stage multimodal world model for autonomous driving. arXiv preprint arXiv:2602.02002 (2026)
50. Zhong, Z., Li, J., He, J., Yan, H., Gong, X., Zhao, G., Cai, Y., Gao, J., Yan, X., Liu, B., et al.: Dualcot-vla: Visual-linguistic chain of thought via parallel reasoning for vision-language-action models. arXiv preprint arXiv:2603.22280 (2026)
51. Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.Y., Gan, C.: Robodreamer: Learning compositional world models for robot imagination. In: International Conference on Machine Learning. pp. 61885–61896. PMLR (2024)
52. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)