DREAM-LIFT-VLA: GENERATIVE 2D-TO-3D WORLD DREAMING FOR VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models offer a promising path toward generalist robot policies. A key challenge is that sparse action supervision provides limited guidance for capturing fine-grained future scene changes. To address this limitation, recent methods incorporate world modeling into VLAs by jointly predicting future 2D videos and 3D geometry. However, parallel training imposes 3D grounding before 2D scene evolution has been sufficiently learned. We argue that VLAs should first establish a clear semantic understanding of 2D scene evolution and then lift the evolution into 3D. We propose **Dream-Lift-VLA**, a progressive VLA learning framework that includes two phases: **Dream** and **Lift**. Specifically, the Dream phase uses world embeddings to condition a pretrained video diffusion model for future-video generation, enabling the VLA to learn fine-grained scene dynamics and manipulation semantics. Building upon these learned dynamics, the Lift phase further conditions a pretrained depth diffusion model on the world embeddings to generate future depth, thereby equipping the VLA with future spatial awareness. Empowered by this progressive 2D-to-3D world dreaming, our policy develops stable representations of both future manipulation semantics and 3D geometry. Extensive experiments on diverse simulation and real-world tasks demonstrate that Dream-Lift-VLA consistently outperforms state-of-the-art approaches. The code will be released soon.

1 INTRODUCTION

Vision-language-action (VLA) models unify visual-language understanding and action generation, offering a promising path toward generalist robot policies (Brohan et al., 2023; Kim et al., 2024). Early approaches directly use ground-truth robot actions (Intelligence et al., 2025) or latent actions (Ye et al., 2025) as supervision for VLA learning. However, these sparse action signals provide limited supervision for capturing fine-grained scene evolution.

To address this limitation, recent work incorporates world modeling into VLAs, as illustrated in Fig. 1. One line of work jointly models future 2D information and robot actions by generating future videos in pixel space (Cen et al., 2025; Zhao et al., 2025) or predicting future representations in latent space (Su et al., 2026; Luo et al., 2026b;a). Although these methods demonstrate that world modeling benefits action generation, their future imagination remains confined to 2D visual space and provides limited spatial support for 3D action generation. Another line of work further predicts future 3D geometry (Qian et al., 2026; Xu et al., 2026a). However, directly mapping current 2D observations to future 3D space bypasses explicit modeling of the scene's 2D evolution, which reveals the interaction process and manipulation semantics that underpin subsequent action generation.

Only a few studies jointly model future 2D information and 3D geometry (Zhang et al., 2026; He et al., 2025), but they treat the two as parallel prediction targets and overlook the inherent connection between the two forms of supervision. Future 2D prediction learns how the current observation evolves into a future scene within the same modality; building on this, future 3D prediction grounds this evolution across modalities in 3D space. Parallel training ignores this dependency, forcing the model to perform future 3D grounding before the underlying 2D scene evolution has been sufficiently learned. This insight motivates us to study how to build on future 2D scene evolution to model the future 3D world, progressively shaping VLA representations with both manipulation semantics and spatial awareness.

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.

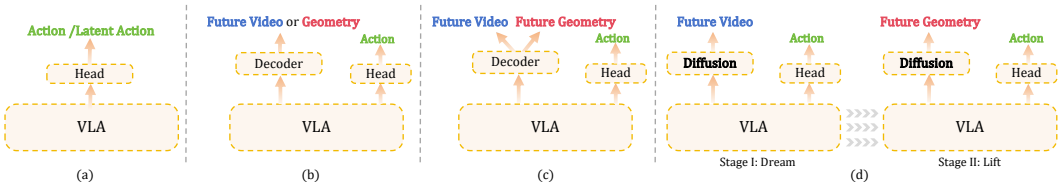


Figure 1: **Comparison of different world modeling paradigms in VLAs.** (a) Vanilla VLAs directly learn actions or latent actions from visual observations and language instructions. (b) Some VLAs additionally model future world information, predicting either future 2D videos or future 3D geometry to support action generation. (c) Other VLAs model both forms of future information, using video generation and 3D geometry prediction as parallel objectives. (d) Our proposed Dream-Lift-VLA first dreams future 2D videos and then lifts the learned 2D dynamics into future 3D geometry, progressively shaping VLA representations for action generation.

Consequently, we propose **Dream-Lift-VLA**, a progressive VLA learning framework that advances from future 2D scene evolution to the future 3D world. Specifically, the **Dream phase** trains the VLA to imagine future 2D scene evolution conditioned on the current observation and language instruction, thereby strengthening its understanding of manipulation semantics while capturing the 2D scene dynamics needed for subsequent 3D lifting. To realize this, we first introduce world and action queries into the VLM, whose encoded outputs form world and action embeddings, respectively. We then use the world embeddings as the condition for a pretrained video diffusion model to imagine future videos. This process enhances the VLA's understanding of manipulation semantics, endowing the policy with more fine-grained knowledge of future dynamics.

Building on the 2D future dynamics learned in the **Dream phase**, the **Lift phase** equips the VLA with 3D spatial awareness to further support action generation. While the Dream phase captures how the scene evolves during manipulation, its supervision remains confined to 2D space. We therefore initialize the Lift phase from the Dream-phase checkpoint and augment the native conditioning pathway of a pretrained depth diffusion model with the world embeddings to generate future depth. By grounding the inherited scene dynamics in future geometry, the Lift phase enhances the model's capacity for spatial foresight. In both phases, the world and action embeddings jointly condition action generation, allowing the learned future representations to directly support robotic control.

To the best of our knowledge, Dream-Lift-VLA is the first framework that jointly models future 2D dynamics and future 3D geometry in a progressive manner. By learning fine-grained semantics of future scene evolution in the **Dream phase** and anticipating the spatial consequences of such evolution in the **Lift phase**, our policy generates more accurate and robust actions. Extensive experiments demonstrate that Dream-Lift-VLA outperforms state-of-the-art approaches on various simulation and real-world tasks, while systematic ablations validate the effectiveness of its two-stage design.

2 Related Works

2.1 Vision–Language–Action Models

Vision–language–action (VLA) models learn generalist robot policies by integrating visual perception, language understanding, and action generation. According to the source of action supervision, existing approaches can be broadly divided into two categories. The first category directly learns from ground-truth robot actions, represented as either discrete tokens or continuous action sequences (Brohan et al., 2023; Kim et al., 2024; Team et al., 2024; Black et al., 2024; Intelligence et al., 2025). The second category infers latent actions from visual transitions, enabling policy pretraining on large-scale human and robot videos without explicit action annotations (Ye et al., 2025; Chen et al., 2025; Bu et al., 2025; Luo et al., 2026a). Although these methods effectively learn action-relevant representations, their sparse action signals provide limited guidance for modeling future scene evolution, which reveals the physical changes caused by robot actions and the fine-grained interaction dynamics essential for precise manipulation.

2.2 World Modeling in VLAs

To provide richer supervision beyond robot actions, recent approaches incorporate world modeling into VLAs. Existing methods can be broadly categorized according to the space in which future information is modeled. The first category models future evolution in the 2D space. Some methods explicitly predict future images or videos together with robot actions (Cen et al., 2025; Zhao et al., 2025). Recent world-action models further unify future visual generation and action prediction within a shared generative architecture (Kim et al., 2026; Bi et al., 2026; Ye et al., 2026), while other approaches retain video modeling during training without performing visual rollout at inference (Yuan et al., 2026; Team et al., 2026). To avoid reconstructing redundant visual details, another line of work predicts future states in a compact latent space rather than reconstructing complete observations (Su et al., 2026; Sun et al., 2026; Luo et al., 2026b;a). However, the above methods typically model the future in 2D space, providing limited spatial support for 3D action generation.

The second category introduces future 3D supervision, such as future depth or point clouds (Qian et al., 2026; Xu et al., 2026a), to provide explicit spatial support for action learning. However, directly predicting future 3D geometry from current 2D observations leaves the scene's 2D evolution insufficiently modeled, even though the manipulation semantics and interaction process revealed by this evolution are important for subsequent action generation. A few methods jointly predict future 2D information and 3D geometry (Zhang et al., 2026; He et al., 2025). Nevertheless, they optimize both forms of supervision in parallel, overlooking that future 3D grounding in fact relies on scene-evolution prior first established through same-modality 2D world prediction.

In this paper, we also use future 2D and 3D information as supervision for VLA representation learning. Unlike existing methods that learn them in parallel, we first capture stable manipulation semantics and scene dynamics through 2D world dreaming and then lift the learned dynamics into 3D, thereby equipping the VLA with both future dynamics and spatial awareness.

3 Methods

3.1 Problem Formulation and Overview

We consider learning a language-conditioned visuomotor policy from expert demonstrations. At each timestep t, we retain an RGB image I_t, the robot's proprioceptive state $\mathbf{p}_t$, and the corresponding language instruction ℓ. Given this multimodal context, the policy π_Θ, parameterized by Θ, models the conditional distribution $\pi_\Theta(\mathbf{a}_{t:t+H-1} \mid I_t, \ell, \mathbf{p}_t)$ over the H-step action chunk $\mathbf{a}_{t:t+H-1} = [\mathbf{a}_t, \ldots, \mathbf{a}_{t+H-1}]$.

As shown in Fig. 2, Dream-Lift-VLA follows a two-stage training scheme. In the **Dream** phase, future-video generation shapes the VLA representations to capture manipulation semantics and fine-grained scene dynamics. The **Lift** phase is initialized from the Dream-phase checkpoint and uses future-depth generation to ground the learned scene evolution in 3D, further shaping the VLA representations with spatial foresight. Finally, shaped by this two-stage process, the policy can better anticipate how manipulation will unfold in the physical world and generate more precise actions.

3.2 Stage I: Dreaming Future Dynamics in 2D

In Stage I, we use a pretrained Stable Video Diffusion (SVD) model (Blattmann et al., 2023) for future-video generation, with the world embeddings $\mathbf{W}_t \in \mathbb{R}^{N_w \times d_w}$ serving as the interface for absorbing 2D world knowledge, where N_w and d_w denote the number of tokens and the VLM embedding dimension, respectively. Given the current image I_t, the frozen CLIP encoder of SVD first produces its native image context $\mathbf{C}_t \in \mathbb{R}^{N_c \times d_c}$. We then map the embedding dimension of $\mathbf{W}_t$ from d_w to d_c and concatenate the projected embeddings with $\mathbf{C}_t$ along the token dimension, yielding the combined condition $\widetilde{\mathbf{C}}_t$, which is injected into the SVD through cross-attention.

Given the future video clip $I_{t+1:t+H}$, the frozen VAE encoder of SVD encodes the H future frames into a latent sequence $\mathbf{z}_{t+1:t+H}$ and separately encodes the current image I_t into the current-image latent $\mathbf{z}_t$. At noise level σ, we sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and construct the noised future-video latent sequence $\mathbf{z}^\sigma_{t+1:t+H} = \mathbf{z}_{t+1:t+H} + \sigma\boldsymbol{\epsilon}$. Following the native SVD denoising objective, the denoiser $\mathcal{D}_\phi(\cdot)$ predicts the clean future-video latent from $\mathbf{z}^\sigma_{t+1:t+H}$, conditioned on $\mathbf{z}_t$ and $\widetilde{\mathbf{C}}_t$.

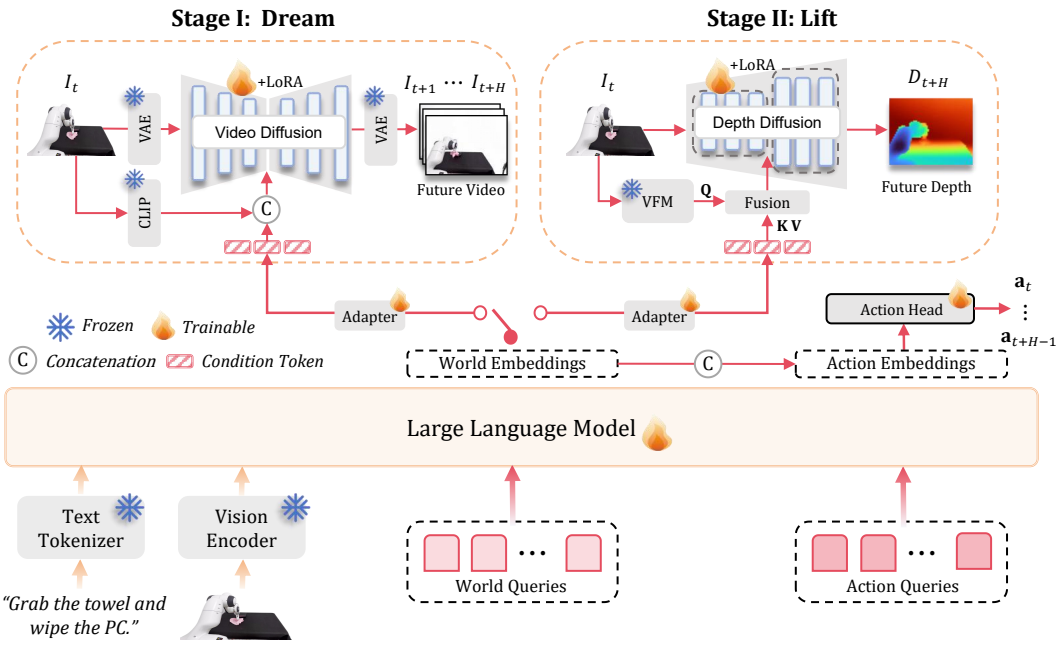


Figure 2: **Overview of Dream-Lift-VLA.** We introduce world and action queries into the VLM, with their resulting final-layer embeddings serving as interfaces for world modeling and action generation, respectively. In Stage I, the world embeddings condition a pretrained video diffusion model to dream future 2D scene evolution. Building on Stage I, Stage II fuses the world embeddings with the native conditioning features extracted by the visual foundation model (VFM) of a pretrained depth diffusion model to anticipate future 3D geometry, thereby lifting the learned 2D scene dynamics into 3D. Both generative branches are removed during inference.

The corresponding objective is:

$$\mathcal{L}_{\text{2D}} = \mathbb{E}_{\sigma,\boldsymbol{\epsilon}}\left[\left\|\mathcal{D}_\phi\left(\mathbf{z}^{\sigma}_{t+1:t+H}, \sigma, \mathbf{z}_t, \widetilde{\mathbf{C}}_t\right) - \mathbf{z}_{t+1:t+H}\right\|_2^2\right]. \tag{1}$$

To preserve the pretrained video prior, we freeze the SVD backbone and adapt it using Low-Rank Adaptation (LoRA) (Hu et al., 2021). In parallel, we concatenate $\mathbf{W}_t$, the action embeddings $\mathbf{A}_t$ produced by the action queries, and the encoded proprioceptive state to condition action generation. The Stage I objective is as follows:

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{act}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}}, \tag{2}$$

where $\mathcal{L}_{\text{act}}$ denotes the flow-matching (Lipman et al., 2022) objective for action generation, following prior work (Su et al., 2026), and λ_{2D} balances the two objectives. Through this joint objective, the VLA learns to anticipate how the scene evolves during manipulation, thereby strengthening its understanding of manipulation semantics.

3.3 Stage II: Lifting Future Dynamics into 3D

Stage II is initialized from the Stage I model and replaces future-video generation with future-depth generation using a pretrained Pixel-Perfect Depth (PPD) model (Xu et al., 2026b). In its native formulation, PPD employs a frozen visual foundation model (VFM) to extract semantic features $\mathbf{S}_t \in \mathbb{R}^{N_s \times d_p}$ from the current image I_t as an additional semantic condition for depth generation, where N_s and d_p denote the number of tokens and their embedding dimension, respectively.

First, we map the embedding dimension of $\mathbf{W}_t$ from d_w to d_p, obtaining the world condition tokens $\mathbf{W}_t^c \in \mathbb{R}^{N_w \times d_p}$. We then apply separate linear projections to obtain queries $\mathbf{Q}_t$ from $\mathbf{S}_t$ and keys $\mathbf{K}_t$ and values $\mathbf{V}_t$ from $\mathbf{W}_t^c$. The fused semantic condition is computed using multi-head cross-attention:

$$\widetilde{\mathbf{S}}_t = \text{MHCA}\left(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t\right), \tag{3}$$

Table 1: **Performance comparison on MetaWorld.** We report success rates (%) on six **hard** and **very hard** manipulation tasks and their average. Best results are highlighted in bold.

Method	Pick Place	Disassemble	Stick Pull	Assembly	Shelf Place	Hand Insert	Avg.
π_0 (Black et al., 2024)	**80.0**	52.0	68.0	80.0	32.0	68.0	63.3
$\pi_{0.5}$ (Intelligence et al., 2025)	68.0	**92.0**	60.0	96.0	48.0	64.0	71.3
DreamVLA (Zhang et al., 2026)	60.0	68.0	80.0	70.0	56.0	64.0	66.3
VLA-JEPA (Sun et al., 2026)	56.0	64.0	84.0	96.0	64.0	**76.0**	73.3
WoG (Su et al., 2026)	28.0	60.0	80.0	24.0	64.0	60.0	52.7
Fast-WAM (Yuan et al., 2026)	44.0	56.0	16.0	88.0	32.0	60.0	49.3
Dream-Lift-VLA (Ours)	72.0	**92.0**	**88.0**	**100.0**	**68.0**	72.0	**82.0**

where $\mathrm{MHCA}(\cdot)$ denotes multi-head cross-attention (Vaswani et al., 2017). The resulting features $\widetilde{\mathbf{S}}_t \in \mathbb{R}^{N_s \times d_p}$ replace $\mathbf{S}_t$ as PPD's semantic condition.

Let D_{t+H} denote the normalized and centered depth map corresponding to the future observation I_{t+H}. Given flow timestep $\tau \in [0, 1]$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct the noised depth $D^{\tau}_{t+H} = (1-\tau)D_{t+H} + \tau\boldsymbol{\epsilon}$ and target velocity $v^* = \boldsymbol{\epsilon} - D_{t+H}$. Following PPD, we employ the velocity predictor v_θ under the rectified-flow formulation (Liu et al., 2022) to predict the velocity field:

$$\mathcal{L}_{\mathrm{3D}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\left[\left\| v_\theta\left(D^{\tau}_{t+H}, \tau, I_t, \widetilde{\mathbf{S}}_t\right) - v^* \right\|_2^2\right]. \tag{4}$$

We freeze the original PPD parameters and jointly optimize its LoRA modules and the introduced cross-attention module. Stage II retains the action-generation objective from Stage I and combines it with future-depth supervision as follows:

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{act}} + \lambda_{\mathrm{3D}}\mathcal{L}_{\mathrm{3D}}, \tag{5}$$

where λ_{3D} balances the action-generation and future-depth objectives. By incorporating the world embeddings $\mathbf{W}_t$ into PPD's semantic condition and shifting its native prediction target from current to future depth, Stage II grounds the scene dynamics learned in Stage I in 3D, further shaping the VLA representations with spatial foresight.

4 Simulation Experiments

4.1 Setup

We evaluate Dream-Lift-VLA on three simulation benchmarks. On MetaWorld (Yu et al., 2020), we train a multitask policy on six hard and very hard manipulation tasks and evaluate each task over 25 episodes. On LIBERO (Liu et al., 2023), we jointly train the policy across 40 tasks from the Spatial, Object, Goal, and Long suites and evaluate each task over 50 trials. We further assess out-of-distribution robustness on LIBERO-Plus (Fei et al., 2025), which introduces perturbations in camera viewpoints, robot initial states, language instructions, lighting, backgrounds, sensor noise, and object layouts. We directly evaluate the LIBERO checkpoint on all 10,030 perturbed task instances without additional adaptation. More details are provided in **Appendix A.2.** We compare Dream-Lift-VLA against representative standard VLAs (e.g., $\pi_{0.5}$ (Intelligence et al., 2025)), world-modeling VLAs (e.g., DreamVLA (Zhang et al., 2026)), and World-Action Models (e.g., Fast-WAM (Yuan et al., 2026)); further details are provided in **Appendix A.4**.

4.2 Results

As shown in Tables 1–3, Dream-Lift-VLA consistently achieves the strongest overall performance across all three simulation benchmarks. On MetaWorld, it achieves an average success rate of 82.0% across six tasks. This exceeds VLA-JEPA's 73.3% by an absolute margin of 8.7% and achieves the best or tied-best performance on four tasks. On LIBERO, Dream-Lift-VLA achieves the highest average success rate of 99.0% across four suites, surpassing Motus at 97.7% by 1.3%. It also obtains the best or tied-best performance on every suite. The stage-wise results show that future 2D supervision in Stage I improves the action-only baseline from 96.8% to 98.0%, while future-depth grounding in Stage II further raises it to 99.0%, validating both stages of the progressive design. On

Table 2: **Performance comparison on LIBERO.** We report success rates (%) on the four LIBERO suites and their average. Best results are highlighted in bold.

Method	Spatial	Object	Goal	Long	Avg.
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
$\pi_{0.5}$ (Intelligence et al., 2025)	98.8	98.2	98.0	92.4	96.9
DreamVLA (Zhang et al., 2026)	97.5	94.0	89.5	89.5	92.6
VLA-JEPA (Sun et al., 2026)	96.2	99.6	97.2	95.8	97.2
WoG (Su et al., 2026)	98.2	96.8	96.4	89.2	95.2
Motus (Bi et al., 2026)	96.8	99.8	96.6	**97.6**	97.7
Fast-WAM (Yuan et al., 2026)	98.2	**100.0**	97.0	95.2	97.6
Dream-Lift-VLA	**99.2**	**100.0**	**99.2**	**97.6**	**99.0**
− Stage II	98.6	99.6	97.6	96.0	98.0
− Stage II, − Stage I	97.4	98.2	96.8	94.8	96.8

Table 3: **Performance comparison on LIBERO-Plus.** We report success rates (%) across seven perturbation dimensions. Since the categories contain different numbers of tasks, Avg. is computed over all evaluation tasks; see **Appendix A.2** for details. Best results are highlighted in bold.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
π_0 (Black et al., 2024)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$ (Intelligence et al., 2025)	64.8	**71.8**	83.0	93.5	92.2	78.8	**85.5**	80.3
DreamVLA (Zhang et al., 2026)	26.2	17.6	67.0	77.5	71.5	53.6	43.5	48.8
VLA-JEPA (Sun et al., 2026)	63.3	67.1	85.4	95.6	93.6	66.3	85.1	78.0
WoG (Su et al., 2026)	11.9	52.6	76.2	58.2	66.3	17.7	70.8	49.0
Fast-WAM (Yuan et al., 2026)	16.4	44.5	68.9	78.2	53.7	37.7	60.7	50.0
Dream-Lift-VLA	**77.9**	60.8	**87.5**	**96.0**	**94.5**	**92.2**	83.3	**83.7**
− Stage II	70.4	59.3	85.7	94.9	91.9	84.2	81.4	80.0
− Stage II, − Stage I	59.6	46.6	84.4	93.2	89.2	78.7	78.2	74.3

Table 4: Stage-wise contributions of Dream-Lift-VLA. Numbers in parentheses denote the total training steps. Best results are highlighted in bold.

Model	2D Dreaming	3D Lifting	Avg. Success Rate (%) ↑
Action-only (60k)	×	×	68.0
Stage I (40k)	✓	×	72.0
Stage I (60k)	✓	×	71.3
Stage I + Stage II (60k)	✓	✓	**82.0**

LIBERO-Plus, Dream-Lift-VLA achieves the highest overall success rate of 83.7%, outperforming $\pi_{0.5}$ (80.3%) and VLA-JEPA (78.0%) by 3.4% and 5.7%, respectively. It ranks first in five of seven perturbation categories, with particularly strong gains over $\pi_{0.5}$ under camera-viewpoint changes (77.9% vs. 64.8%) and sensor noise (92.2% vs. 78.8%). The stage-wise gains from the action-only baseline (74.3%) to Stage I (80.0%) and Stage II (83.7%) further demonstrate that progressive 2D-to-3D world dreaming improves the policy's robustness to distribution shifts.

4.3 Ablation Studies

We conduct controlled ablations on the same six MetaWorld tasks reported in Table 1. All variants share the same evaluation protocol, and we report the average success rate across the six tasks. SVD and PPD denote Stable Video Diffusion and Pixel-Perfect Depth, respectively.

Progressive World Dreaming. We first examine the contribution of the two-stage design. As shown in Table 4, Stage I improves the action-only baseline from 68.0% to 72.0% by learning future 2D scene dynamics. Extending Stage I from 40k to 60k steps yields no further improvement (72.0% vs. 71.3%), showing that longer training alone does not account for the gain. In contrast, introducing 3D lifting in Stage II raises the success rate to 82.0%. These results demonstrate that grounding the 2D dynamics learned in Stage I with future 3D geometry provides substantial additional benefit.

Table 5 compares different training orders and combinations of 2D dreaming and 3D lifting. Direct 3D dreaming achieves 69.3%, whereas jointly optimizing 2D and 3D dreaming from scratch reduces performance to 64.7%. Longer training provides no meaningful improvement in either setting,

Table 5: Ablation of the training curriculum.

Training Curriculum	Stage I Objective	Stage II Objective	Avg. SR (%) ↑
Direct 3D dreaming	PPD	–	69.3
Direct 3D dreaming (longer)	PPD	–	68.0
2D and 3D dreaming	SVD + PPD	–	64.7
2D and 3D dreaming (longer)	SVD + PPD	–	65.3
3D to 2D dreaming	PPD	SVD	68.7
3D to 2D and 3D dreaming	PPD	SVD+PPD	69.3
2D to 2D and 3D dreaming	SVD	SVD+PPD	72.0
2D to 3D dreaming	**SVD**	**PPD**	**82.0**

Table 6: World modeling interface.

Stage I Configuration	Stage I Avg. SR (%)
Action emb. → action	68.0
Action emb. → action and SVD	69.3
Action emb. → action; World emb. → SVD	**72.0**

Table 7: Stage II depth target.

Depth Target	Avg. SR (%)
No depth grounding	72.0
Current depth D_t	68.7
Future depth D_{t+H}	**82.0**

Table 8: 3D grounding formulation.

Grounding	Avg. SR (%)
Image emb. → depth decoder	75.3
World emb. → depth decoder	77.3
World emb. → PPD	**82.0**

Table 9: World condition injection of PPD.

Injection	Avg. SR (%)
Single cross-attn.	76.7
3-layer cross-attn.	78.0
Fuse VFM	**82.0**

Table 10: Generative-prior adaptation.

Prior	Frozen	Full FT	LoRA
SVD (Stage I)	67.3	57.3	**72.0**
PPD (Stage II)	75.3	77.3	**82.0**

Table 11: Different generative models.

Video Model	Stage I	+ PPD	+ DepthFM
SVD	72.0	**82.0**	78.7
Wan2.2	76.0	80.7	81.3

yielding 68.0% and 65.3%, respectively. Starting with 3D dreaming and then introducing either 2D dreaming alone or both objectives is similarly ineffective, reaching only 68.7% and 69.3%. When Stage I first learns 2D scene evolution with SVD, retaining both SVD and PPD in Stage II achieves only 72.0%, indicating that joint optimization prevents Stage II from focusing on spatial refinement. Removing the SVD objective and dedicating Stage II to PPD-based 3D lifting raises performance to 82.0%. These results suggest that Stage I must first establish stable future representations of 2D scene evolution, allowing Stage II to incorporate geometric foresight and better capture how manipulation unfolds in the physical world.

Moreover, we examine whether future imagination requires dedicated world queries through a controlled ablation of Stage I. As shown in Table 6, using the action embeddings only for action generation yields 68.0%. Reusing them to also condition SVD provides a limited gain to 69.3%, whereas introducing separate world embeddings for SVD raises performance to 72.0%. This shows that dedicated world queries provide a more effective interface for absorbing future-world supervision.

Design Choices for 3D Lifting. Table 7 compares the depth targets used in Stage II. The Stage I model without depth grounding achieves 72.0%. Using current depth as the Stage II target reduces performance to 68.7%, indicating that current-state grounding disrupts the future dynamics learned in Stage I. In contrast, future-depth supervision preserves these future-oriented representations while grounding the predicted scene evolution in 3D, raising performance to 82.0%. Table 8 compares different formulations for 3D grounding in Stage II. Directly decoding future depth from image embeddings achieves 75.3%, while using world embeddings improves performance to 77.3%. Conditioning PPD on the world embeddings further raises performance to 82.0%, showing that consistent generative supervision across both stages more effectively shapes VLA representations from 2D dynamics to 3D spatial knowledge. Implementation details of the standalone decoders are provided in **Appendix A.5**. Finally, Table 9 compares different strategies for injecting world embeddings into PPD. Injecting world embeddings via cross-attention into a single middle PPD denoising block achieves 76.7%, while uniformly injecting them into three denoising blocks improves the result to 78.0%. In contrast, allowing PPD's VFM features to attend to the world embeddings and feeding the fused features through its native semantic-conditioning pathway achieves 82.0%.

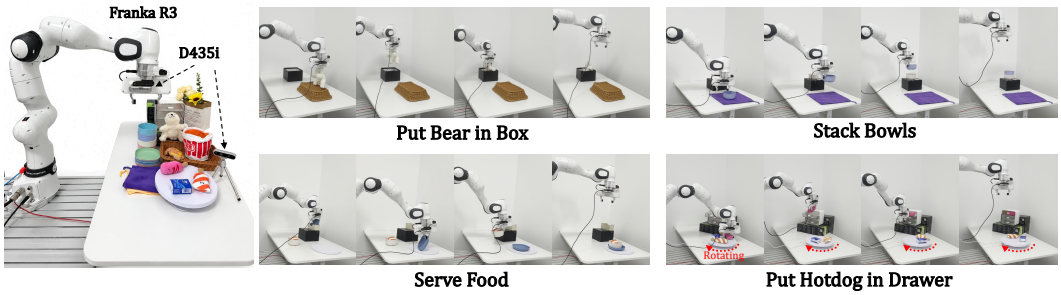


Figure 3: **Overview of the real-world experiments.** The figure shows our robotic platform and sensors on the left and demonstrations of the four tasks under the in-distribution setting on the right.

This design better leverages PPD's pretrained semantic and depth-generation priors to shape VLA representations.

Adaptation and Generality of Generative Models. Table 10 compares different strategies for adapting the pretrained generative models. In Stage I, LoRA (Hu et al., 2021) adaptation of SVD achieves 72.0%, compared with 67.3% when frozen and 57.3% with full fine-tuning. In Stage II, starting from the same LoRA-adapted Stage I checkpoint, adapting PPD with LoRA achieves 82.0%, outperforming frozen PPD at 75.3% and full fine-tuning at 77.3%. These results indicate that LoRA better preserves the pretrained generative priors than full fine-tuning, while adapting more effectively to robot observations than freezing the generative models.

To examine whether our progressive framework depends on particular generative backbones, we additionally use Wan2.2 (Wan et al., 2025) and DepthFM (Gui et al., 2025) as alternatives to SVD (Blattmann et al., 2023) and PPD (Xu et al., 2026b), respectively. More implementation details are provided in **Appendix A.5**. Table 11 compares different combinations of these generative models. Each row reports the performance of a Stage I policy trained with the corresponding video model, followed by Stage II using either PPD or DepthFM. Both depth models consistently improve the SVD- and Wan2.2-based Stage I policies. SVD followed by PPD achieves the best result of 82.0%, while Wan2.2 followed by DepthFM reaches a comparable 81.3%. These consistent gains demonstrate that our progressive 2D-to-3D framework is not tied to a specific pair of generative models.

5 Real-World Experiments

5.1 Setup

Platform. For our real-world experiments, we use a 7-DoF Franka Research 3 robot equipped with a 1-DoF parallel gripper. RGB observations are captured by two Intel RealSense D435i cameras, including an external-view camera and a wrist-mounted camera. Model inference and robot control run on a workstation equipped with an NVIDIA RTX 5060 Ti GPU.

In-Distribution (ID) and Out-of-Distribution (OOD) Setup. We design four real-world manipulation tasks: *Put Bear in Box* requires placing a toy bear into a box; *Stack Bowls* requires stacking one bowl onto another; *Serve Food* requires placing a plate on the table and then placing a food-containing bowl onto it; and *Put Hotdog in Drawer* requires grasping a hotdog from a rotating turntable, placing it into a drawer, and closing the drawer. For training, we collect 100 expert demonstrations per task exclusively under the ID setting, with varying initial robot and object poses. We evaluate OOD generalization using task-specific variations in background, lighting, object identity, and height that are unseen during training, with details provided in **Appendix A.3**. For both ID and OOD settings, we evaluate each task over 20 trials and report the average success rate.

5.2 Results

We compare Dream-Lift-VLA with $\pi_{0.5}$ (Intelligence et al., 2025), a strong generalist VLA, and DreamVLA (Zhang et al., 2026), which jointly predicts future dynamic regions, depth, and semantic features. **Under the ID settings**, Dream-Lift-VLA achieves the best performance on all four tasks.

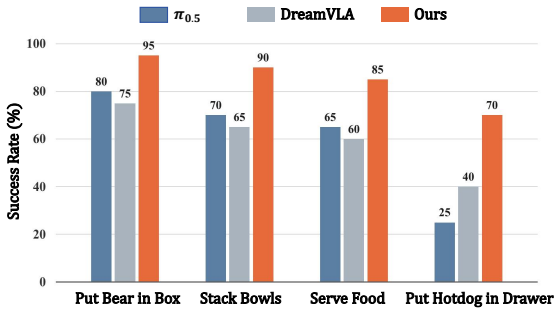


Figure 4: Performance comparison on the ID settings.

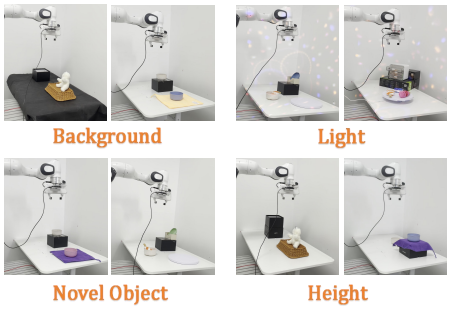


Figure 5: Four OOD settings across tasks.

Table 12: **Real-world generalization performance.** Each entry reports the ID to OOD success rates (%). Best results are highlighted in bold.

Method	Background		Light		Novel Object		Height		OOD Avg.
	Bear	Bowls	Serve	Hotdog	Bowls	Serve	Bear	Bowls	
$\pi_{0.5}$	80.0→50.0	70.0→65.0	65.0→60.0	25.0→15.0	70.0→55.0	65.0→25.0	80.0→30.0	70.0→35.0	41.9
DreamVLA	75.0→55.0	65.0→55.0	60.0→50.0	40.0→20.0	65.0→50.0	60.0→30.0	75.0→45.0	65.0→35.0	42.5
Ours	**95.0→80.0**	**90.0→90.0**	**85.0→80.0**	**70.0→60.0**	**90.0→85.0**	**85.0→65.0**	**95.0→85.0**	**90.0→80.0**	**78.1**

In particular, it reaches 70.0% on *Put Hotdog in Drawer*, exceeding $\pi_{0.5}$ and DreamVLA by 45.0% and 30.0%, respectively. This task requires anticipating the hotdog's motion on the rotating turntable for precise grasping, followed by accurate placement and drawer closure. **Under the OOD settings,** Dream-Lift-VLA achieves the highest average success rate of 78.1%, substantially outperforming $\pi_{0.5}$ (41.9%) and DreamVLA (42.5%). Under the novel-object variation of *Stack Bowls*, our method retains 85.0% success with only a 5.0% drop, whereas $\pi_{0.5}$ and DreamVLA drop by 15.0% to 55.0% and 50.0%, respectively. The advantage is particularly pronounced under the height variation of *Put Bear in Box*, where our method retains 85.0% with a 10.0% drop, whereas $\pi_{0.5}$ and DreamVLA drop by 50.0% and 30.0% to 30.0% and 45.0%, respectively. These results suggest that our two-stage design first develops robust manipulation semantics through 2D dreaming and then strengthens spatial reasoning through 3D lifting, leading to greater real-world robustness across semantic and geometric variations. Real-world evaluation videos are provided in the **Supplementary Material.**

6 More Analyses

Given the limited space, we provide additional details and analyses in the **Appendix**, including implementation and simulation and real-world protocols (**Appendix A.1–A.3**), baseline and ablation details (**Appendix A.4 and A.5**), limitations and future directions (**Appendix A.6**), further studies of future-supervision designs (**Appendix B.1–B.3**), query configurations and the effect of 3D lifting on real-world performance (**Appendix B.4 and B.5**), inference efficiency (**Appendix B.6**), and visualizations of attention, future predictions, and real-world ID/OOD trajectories (**Appendix C**).

7 Conclusion

In this paper, we introduce Dream-Lift-VLA, which learns future 2D scene evolution before grounding it in future 3D geometry. By conditioning future-video generation and subsequent future-depth generation on world embeddings, our framework progressively shapes VLA representations for action generation. Extensive experiments across simulation and real-world tasks demonstrate consistent improvements over state-of-the-art baselines. Looking forward, through progressive future-world imagination, our method opens a new avenue for effectively modeling future 2D and 3D information in VLAs, with the potential to support broader robotic manipulation scenarios. Building on this direction, future work could strengthen the Dream phase through robot-centric video pretraining, extend the Lift phase to point-cloud and 3D scene-flow prediction from multi-view observations, and pretrain both stages on large-scale human videos using inferred latent actions before robot-specific adaptation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and presentation. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final content of the paper.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19752–19763. IEEE, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Yuxin He, Ruihao Zhang, Xianzu Wu, Zhiyuan Zhang, Cheng Ding, and Qiang Nie. 3d dynamics-aware manipulation: Endowing manipulation policies with 3d foresight. *arXiv preprint arXiv:2502.10028*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable vla pretraining in the wild. *arXiv preprint arXiv:2602.21736*, 2026a.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Jingjing Qian, Boyao Han, Chen Shi, Lei Xiao, Long Yang, Shaoshuai Shi, and Li Jiang. Geopredict: Leveraging predictive kinematics and 3d gaussian geometry for precise vla manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13529–13539, 2026.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Conference on Robot Learning*, pp. 1332–1344. PMLR, 2023.

Yue Su, Sijin Chen, Haixin Shi, Mingyu Liu, Zhengshen Zhang, Ningyuan Huang, Weiheng Zhong, Zhengbang Zhu, Yuxiao Liu, and Xihui Liu. World guidance: World modeling in condition space for action generation. *arXiv preprint arXiv:2602.22010*, 2026.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

GigaWorld Team, Angen Ye, Angyuan Ma, Boyuan Wang, Chaojun Ni, Fangzheng Ye, Guan Huang, Guo Li, Guosheng Zhao, Haodong Yan, et al. Gigaworld-policy-0.5: A faster and stronger wam empowered by autoresearch. *arXiv preprint arXiv:2607.13960*, 2026.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, et al. ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Chongyang Xu, Haipeng Li, Shen Cheng, Haoqiang Fan, Ziliang Feng, and Shuaicheng Liu. Action-geometry prediction with 3d geometric prior for bimanual manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35036–35046, 2026a.

Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, et al. Pixel-perfect depth with semantics-prompted diffusion transformers. *Advances in Neural Information Processing Systems*, 38:174731–174755, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Yi Yang, Zhihong Liu, Siqi Kou, Yiyang Chen, Yanzhe Hu, Jianbo Zhou, Boyuan Zhao, Zhijie Wei, Xiao Xia, Xueqi Li, et al. World-language-action model for unified world modeling, language reasoning, and action synthesis. *arXiv preprint arXiv:2606.05979*, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026.

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

Appendix Contents

A. More Details 13
A.1 Implementation Details 13
A.2 Simulation Details 13
A.3 Real-World Details 14
A.4 Baseline Details 14
A.5 Ablation Details 14
A.6 Limitations and Future Work 15

B. More Results 15
B.1 Single-Frame vs. Multi-Frame Supervision 15
B.2 Comparison with Future-Image Editing 16
B.3 Alignment-Based 2D-to-3D Learning 16
B.4 Number of Queries 17
B.5 Effect of 3D Lifting on Real-World Performance 17
B.6 Inference Efficiency 17

C. More Visualizations 18

A More Details

A.1 Implementation Details

We build Dream-Lift-VLA upon the pretrained InternVL3-2B (Chen et al., 2024), which comprises an InternViT-300M vision encoder and a Qwen2.5-1.5B (Yang et al., 2025) language model. Following World Guidance (Su et al., 2026), we adopt a flow-matching DiT (Peebles & Xie, 2023) action head with 12 Transformer blocks, a hidden dimension of 768, and 12 attention heads. For Metaworld, we use six world queries and six action queries and predict action chunks with a horizon of $H = 4$. In the Dream phase, we initialize the video branch from pretrained SVD (Blattmann et al., 2023) and generate the next four frames. The Lift phase initializes the VLA from the Dream-phase checkpoint and replaces the video branch with a depth branch initialized from pretrained PPD (Xu et al., 2026b) to predict future depth at $t + 4$. For LIBERO, we retain the same query configuration and set both the action and video horizons to $H = 8$, and the Lift phase predicts future depth at $t + 8$. For both benchmarks, future-depth targets are generated offline using the frozen pretrained PPD. PPD employs a frozen Depth Anything V2 (Yang et al., 2024) as its visual foundation model, whose features query the world embeddings through an eight-head cross-attention module. Both SVD and PPD are frozen and adapted with LoRA (Hu et al., 2021), using rank $r = 4$, scaling parameter $\alpha = 8$. The Stage I and Stage II loss weights are $\lambda_{2D} = \lambda_{3D} = 0.5$.

A.2 Simulation Details

Training. For Metaworld, we optimize both stages using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 1×10^{-5}. Stage I is trained for 40k steps, while Stage II is initialized from the Stage I checkpoint and trained for another 20k steps. Training is conducted on two NVIDIA A800 GPUs with 80 GB of memory each, using a batch size of 16 per GPU (32 in total). For LIBERO, we train Stage I for 60k steps and Stage II for an additional 30k steps. We use initial learning rates of 1×10^{-5} for the VLM and the LoRA modules in SVD and PPD, and 1×10^{-4} for the action DiT. Both parameter groups follow a cosine schedule with a minimum learning rate of 1×10^{-6}.

Evaluation. Our MetaWorld (Yu et al., 2020) evaluation comprises six hard and very hard (Seo et al., 2023) manipulation tasks: Pick Place, Disassemble, Stick Pull, Assembly, Shelf Place, and Hand Insert. We use 25 expert trajectories per task and evaluate each task over 25 rollouts, with a random seed sampled for each rollout and success recorded if the task is completed within 175 environment steps. LIBERO (Liu et al., 2023) contains four suites, Spatial, Object, Goal, and Long, each comprising 10 tasks with 50 expert trajectories per task. Following prior methods (Yuan et al., 2026; Sun et al., 2026), we evaluate each task over 50 rollouts. LIBERO-Plus (Fei et al., 2025) expands the 40 LIBERO tasks into 10,030 perturbed instances across Camera (1,599), Robot (1,550), Language (1,537), Light (1,142), Background (1,076), Noise (1,601), and Layout (1,525). We evaluate the LIBERO-trained checkpoint without additional adaptation, allowing only one execution attempt for each test instance. The overall success rate is computed across all 10,030 test instances.

A.3 Real-World Details

Training and Evaluation. We follow the LIBERO training configuration, except that we set both the action and video horizons to $H = 16$. The Lift phase predicts future depth at $t + 16$. Following Ψ_0 (Wei et al., 2026), we further adopt training-time real-time chunking (RTC) to enable smooth and continuous action execution during asynchronous inference. For both ID and OOD settings, we evaluate each task over 20 trials.

OOD Evaluation Details. For background variation, we introduce a black tablecloth in *Put Bear in Box* and replace the purple towel under the bowls with an orange one in *Stack Bowls*. **For lighting variation**, temporally changing illumination patterns are projected onto the workspace in *Serve Food* and *Put Hotdog in Drawer*. **For height variation**, we double the box height in *Put Bear in Box*; in *Stack Bowls*, we raise the white bowl, raise the grasped bowl while lowering the white bowl, or place both bowls at the same height, with 7, 7, and 6 trials, respectively. **For novel-object variation** in *Stack Bowls*, the purple grasped bowl is replaced with a pink or blue bowl, with 10 trials each. In *Serve Food*, the 20 trials are evenly divided between a blue dish with toy chicken feet and a green dish with toy crab sticks. The corresponding language instructions are updated to reflect the changed dish color and food type. These height and novel-object configurations are illustrated in Figure 6.

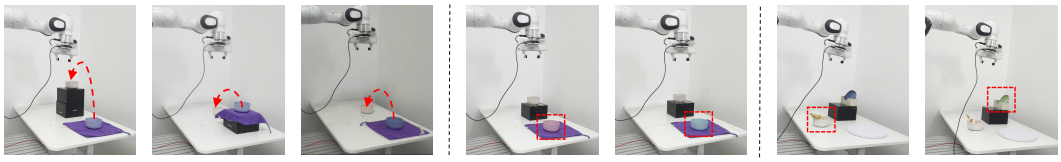


Figure 6: Height and novel-object variations in real-world evaluation.

A.4 Baseline Details

We compare Dream-Lift-VLA with three groups of representative methods. π_0 (Black et al., 2024) and $\pi_{0.5}$ (Intelligence et al., 2025) are standard VLAs that predict continuous action chunks via flow matching. Among VLAs incorporating future-world supervision, DreamVLA (Zhang et al., 2026) jointly predicts future dynamics, depth, and semantic cues, VLA-JEPA (Sun et al., 2026) predicts future latent representations, and WoG (Su et al., 2026) models future observations in a compact condition space. We also include Motus (Bi et al., 2026) and Fast-WAM (Yuan et al., 2026), two World-Action Models that couple video and action learning; Motus unifies video and action generation, whereas Fast-WAM retains video co-training but predicts actions without explicit future-video generation at inference. On MetaWorld, we train each method on the same expert demonstrations following its original downstream training recipe and evaluate all methods under the same protocol. For LIBERO and LIBERO-Plus, we report the published or officially released results.

A.5 Ablation Details

Depth Decoding. To examine whether the improvement merely results from dense depth reconstruction, we replace PPD in Stage II with standalone decoders using the same Stage-I checkpoint and $t{+}H$ depth target. For the image-embedding variant, the final-layer VLM image embeddings are arranged into a $14{\times}14$ grid, projected to 512 dimensions, and processed by an eight-layer Transformer decoder (He et al., 2022). Each token predicts a $16{\times}16$ patch, which is folded into a $224{\times}224$ depth map. For the world-embedding variant, $14{\times}14$ learned spatial queries retrieve information from the six world embeddings through cross-attention (Jaegle et al., 2021), followed by the same decoder. Both variants are supervised using a scale-normalized mean-squared error loss, the same as in DreamVLA (Zhang et al., 2026).

Alternative Generative Backbones. For the alternative generative backbones used in our ablation, we retain their native conditioning pathways. For Wan2.2 (Wan et al., 2025), the current image provides visual context, while projected world embeddings condition future-frame generation through its DiT cross-attention layers. Likewise, for DepthFM (Gui et al., 2025), the current image provides

spatial context, while projected world embeddings condition future-depth generation through its UNet cross-attention layers. Both models are adapted using LoRA (Hu et al., 2021).

A.6 Limitations and Future Work

Our results demonstrate the effectiveness of progressively shaping VLA representations from 2D scene evolution to 3D spatial grounding. Building on this foundation, several promising directions remain for extending Dream-Lift-VLA.

First, although SVD (Blattmann et al., 2023) provides a powerful general video prior for modeling scene evolution, robotic manipulation often involves subtle, action-relevant changes. Incorporating robot interaction data into video pretraining could further enhance its ability to capture these fine-grained dynamics, making robot-centric video models a promising direction for future work.

Second, the current Lift phase grounds the learned dynamics in 3D using future depth. Future work could leverage multi-view observations to jointly predict future point clouds and 3D scene flow, capturing scene geometry and its evolution more comprehensively. Such richer geometric supervision could further strengthen spatial grounding in VLA representations.

Third, large-scale human videos provide a scalable source of diverse manipulation experience. Recent methods show that latent actions can be extracted from videos without robot action annotations (Ye et al., 2025; Luo et al., 2026a;b). These inferred actions could enable both Dream and Lift to be pretrained on human videos before adaptation to robot-specific actions and embodiments.

B More Results

B.1 Single-Frame vs. Multi-Frame Supervision

We compare single-frame and multi-frame supervision in the Dream phase. The Decoder variant adopts the same eight-layer Transformer as the standalone depth decoder described in Section A.5, but predicts future RGB frames instead of depth. As shown in Table 13, supervising the full future sequence $\mathbf{I}_{t+1:t+H}$ rather than only the endpoint frame I_{t+H} raises the average success rate of the Decoder-based policy from 60.0% to 64.0% and that of the SVD-based policy from 66.0% to 72.0%. These gains show that intermediate frames provide useful supervision for learning scene evolution. Under multi-frame supervision, SVD further outperforms direct decoding by 8.0%, indicating that a pretrained video model more effectively shapes VLA representations with continuous scene-evolution priors.

Table 13: **Single-frame vs. Multi-frame future prediction.** Best results are highlighted in bold.

Predictor	Future Target	Avg. SR (%)
Decoder	I_{t+H}	60.0
Decoder	$\mathbf{I}_{t+1:t+H}$	64.0
SVD	I_{t+H}	66.0
SVD (Ours)	$\mathbf{I}_{t+1:t+H}$	**72.0**

We next examine the choice of future-depth supervision in the Lift phase. All variants are initialized from the same SVD-trained Dream-phase checkpoint (72.0% Avg. SR), and Table 14 compares endpoint and multi-depth supervision. For direct decoding, multi-depth supervision raises the average success rate from 77.3% to 79.3%, as the intermediate targets provide denser geometric guidance to a decoder trained from scratch. In contrast, the PPD-based policy performs best with endpoint supervision (82.0% vs. 81.3%). This suggests that, once continuous scene evolution has been learned in the Dream phase, the endpoint depth is sufficient for the pretrained PPD to ground this evolution in 3D, while intermediate depth targets offer limited additional benefit. Accordingly, we adopt PPD in Stage II to predict the endpoint depth at $t + H$ for effective and efficient 3D grounding.

Table 14: **Single-frame vs. Multi-frame future-depth prediction.**

Predictor	Future Target	Avg. SR (%)
Decoder	D_{t+H}	77.3
Decoder	$D_{t+1:t+H}$	79.3
PPD (Ours)	D_{t+H}	**82.0**
PPD	$D_{t+1:t+H}$	81.3

B.2 Comparison with Future-Image Editing

Recent work explores future-image editing with image diffusion models for policy learning (Yang et al., 2026). Following this approach, we replace SVD with SANA (Xie et al., 2024), conditioning it on the current image and world embeddings to predict the future endpoint frame I_{t+H}. This raises the action-only baseline from 68.0% to 71.3%, close to the 72.0% achieved by SVD. However, adding PPD (Xu et al., 2026b) after SANA reduces performance to 69.3%, whereas our SVD-to-PPD training reaches 82.0%. Single-frame SANA provides only endpoint information, whereas SVD shapes the world embeddings with continuous scene evolution. Since PPD is pretrained for current-depth generation and lacks a future-evolution prior, the dynamics learned under SVD supervision provide richer context for future-depth prediction. The future-depth supervision then grounds these dynamics in 3D, encouraging the VLA representations to capture both action-relevant scene evolution and its future spatial consequence.

Table 15: **Image editing vs. Video generation.**

Stage I Objective	Stage II Objective	Avg. SR (%)
Action only	–	68.0
SANA (I_{t+H})	–	71.3
SANA (I_{t+H})	PPD (D_{t+H})	69.3
SVD ($\mathbf{I}_{t+1:t+H}$)	–	72.0
SVD ($\mathbf{I}_{t+1:t+H}$)	**PPD (D_{t+H})**	**82.0**

B.3 Alignment-Based 2D-to-3D Learning

Recent methods use future-feature alignment to inject predictive knowledge into VLAs (Luo et al., 2026b; Su et al., 2026). To examine our two-stage design under this paradigm, we replace SVD in Stage I with feature distillation from a frozen V-JEPA (Assran et al., 2025). As shown in Table 16, this Stage I variant achieves 71.3%. In Stage II, a frozen Depth Anything V2 (DA-V2) (Yang et al., 2024) extracts depth features from future frames as distillation targets. Distilling consecutive-frame or endpoint features improves performance to 76.7% and 77.3%, respectively. Direct depth decoding improves only to 73.3%, while PPD generation reduces performance to 67.3%, suggesting that its generative objective is less compatible with the feature-aligned representations learned in Stage I. These results show that progressive 2D-to-3D supervision also benefits feature-based representation learning when the supervision across stages is appropriately matched.

Table 16: **Feature-based 2D-to-3D supervision.** All Stage II variants inherit the same V-JEPA-distilled Stage I checkpoint.

Stage I Supervision	Stage II Supervision	Avg. SR (%)
V-JEPA features	–	71.3
V-JEPA features	Depth decoder (D_{t+4})	73.3
V-JEPA features	PPD generation (D_{t+4})	67.3
V-JEPA features	DA-V2 features ($t+1{:}t+4$)	76.7
V-JEPA features	**DA-V2 features ($t+4$)**	**77.3**

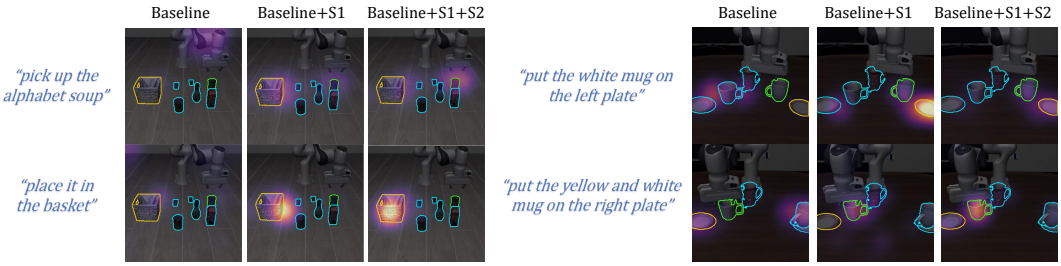


Figure 7: **Attention maps for fine-grained instruction grounding.** Rows correspond to step-specific sub-instructions, while columns compare the action-only baseline, baseline+Stage I (S1), and baseline+Stage I (S1)+Stage II (S2). Green and yellow contours mark task-relevant objects, blue contours mark distractors, and warmer regions indicate higher attention.

B.4 Number of Queries

We investigate how the numbers of world and action queries affect progressive 2D-to-3D learning. With six action queries fixed, three world queries provide insufficient representational capacity to preserve the scene-evolution cues required for subsequent 3D lifting. Both six and twelve world queries benefit from Stage II, while six achieves the best final performance by providing a compact yet sufficiently expressive world representation. When world and action queries are varied together, Stage II consistently improves all three configurations, with a moderate query count of $6/6$ achieving the best overall performance.

Table 17: **World Queries.**

World Queries	Stage I SR (%)	Stage II SR (%)
3	70.0	68.0
6	72.0	**82.0**
12	**74.0**	78.0

Table 18: **World and Action Queries.**

Queries	Stage I SR (%)	Stage II SR (%)
3	59.3	72.7
6	**72.0**	**82.0**
12	68.0	74.7

B.5 Effect of 3D Lifting on Real-World Performance

Table 19 compares Stage I with the complete Stage I+II model on real-world tasks under in-distribution (ID) settings and two selected out-of-distribution (OOD) settings: novel objects and height variations. Stage II raises the average ID success rate from 76.3% to 85.0% and the average over these OOD evaluations from 65.0% to 78.8%, improving all four evaluated OOD cases. The gain is particularly pronounced under height variations, where the average success rate increases from 60.0% to 82.5%, including improvements from 50.0% to 85.0% on *Put Bear in Box* and from 70.0% to 80.0% on *Stack Bowls*. These results show that future-depth supervision in Stage II strengthens spatial grounding, allowing the policy to better handle changes in object and target height.

Table 19: **Effect of progressive 3D lifting on real-world performance.** Stage I uses future-video supervision, while Stage I+II further introduces future-depth supervision.

	ID				ID Avg.	Novel Object		Height		OOD Avg.
Training Stage	**Bear**	**Bowls**	**Serve**	**Hotdog**	**SR (%)**	**Bowls**	**Serve**	**Bear**	**Bowls**	**SR (%)**
Stage I	85.0	**90.0**	75.0	55.0	76.3	80.0	60.0	50.0	70.0	65.0
Stage I+II (Ours)	**95.0**	**90.0**	**85.0**	**70.0**	**85.0**	**85.0**	**65.0**	**85.0**	**80.0**	**78.8**

B.6 Inference Efficiency

Table 20 compares inference latency and FLOPs on an NVIDIA RTX 5060 Ti GPU. Dream-Lift-VLA achieves the lowest latency and computational cost, requiring 76.8 ms and 0.9162 TFLOPs,

compared with 182.1 ms/4.7567 TFLOPs for $\pi_{0.5}$, 107.2 ms/1.0435 TFLOPs for DreamVLA, and 179.0 ms/2.5451 TFLOPs for Fast-WAM. Dream-Lift-VLA enhances policy learning with future-video and future-depth supervision in the Dream and Lift phases, while removing both generative branches at inference, thereby combining strong performance with high efficiency.

Table 20: **Inference efficiency.** TFLOPs and latency are measured per inference request. Lower is better.

Method	**TFLOPs**	**Latency (ms)**
Dream-Lift-VLA (Ours)	**0.9162**	**76.8**
$\pi_{0.5}$ (Intelligence et al., 2025)	4.7567	182.1
DreamVLA (Zhang et al., 2026)	1.0435	107.2
Fast-WAM (Yuan et al., 2026)	2.5451	179.0

C More Visualizations

To examine whether future supervision improves fine-grained semantic grounding, we split each compound instruction into two sub-instructions describing individual manipulation steps and visualize their attention separately in Fig. 7. Compared with the baseline, Stage I generally produces more concentrated attention on the object or target specified by each sub-instruction, with less attention to distractors. Stage II preserves these patterns, suggesting that 2D future supervision improves fine-grained instruction grounding and that subsequent 3D lifting maintains this benefit.

In addition, Figure 8 provides a qualitative comparison of future predictions, showing that the model reasonably captures future scene evolution and spatial structure. Figures 9–12 illustrate successful task execution across ID scenes and OOD variations in background, lighting, object identity, and height. Figure 13 further demonstrates continuous long-horizon operation under repeated human interventions and scene changes.

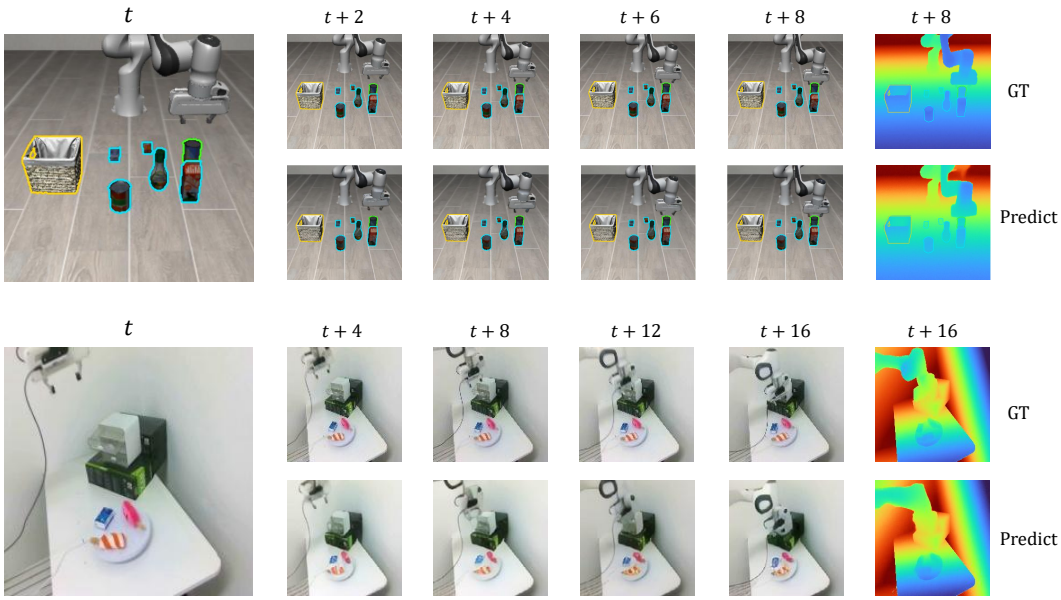


Figure 8: **Qualitative comparison of future predictions.** The upper and lower panels show simulation and real-world examples, respectively. Given the current observation at t, each panel presents the ground-truth future RGB frames and endpoint depth in the upper row and the corresponding predictions in the lower row.

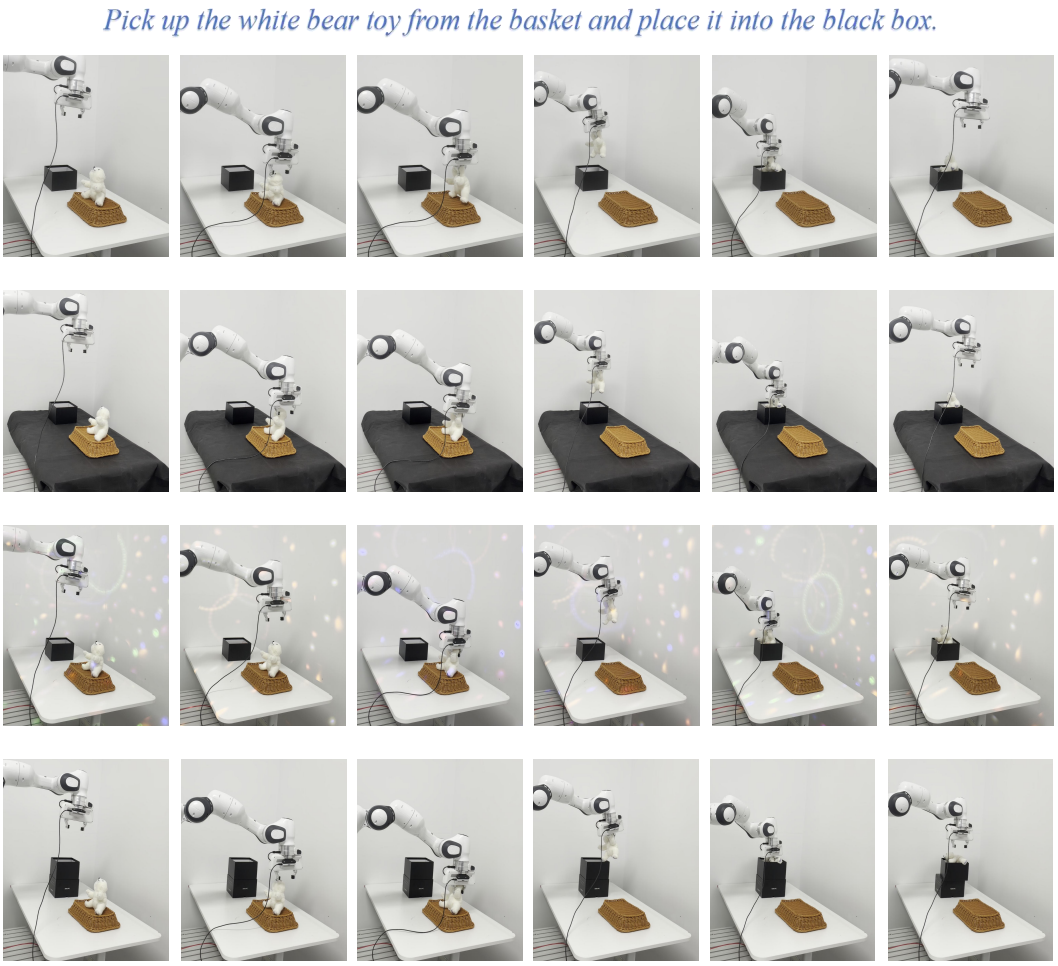


Figure 9: **ID and OOD trajectories for *Put Bear in Box*.** Row 1 shows the ID setting, Rows 2 and 3 show background and lighting variations, respectively, and Row 4 shows the height variation.



Figure 10: **ID and OOD trajectories for *Serve Food*.** Row 1 shows the ID setting, Rows 2 and 3 show background and lighting variations, respectively, and Rows 4–5 show novel-object variations. For the novel-object variations, "blue" and "chicken feet" in the original language instruction are updated to match the corresponding dish color and food type.

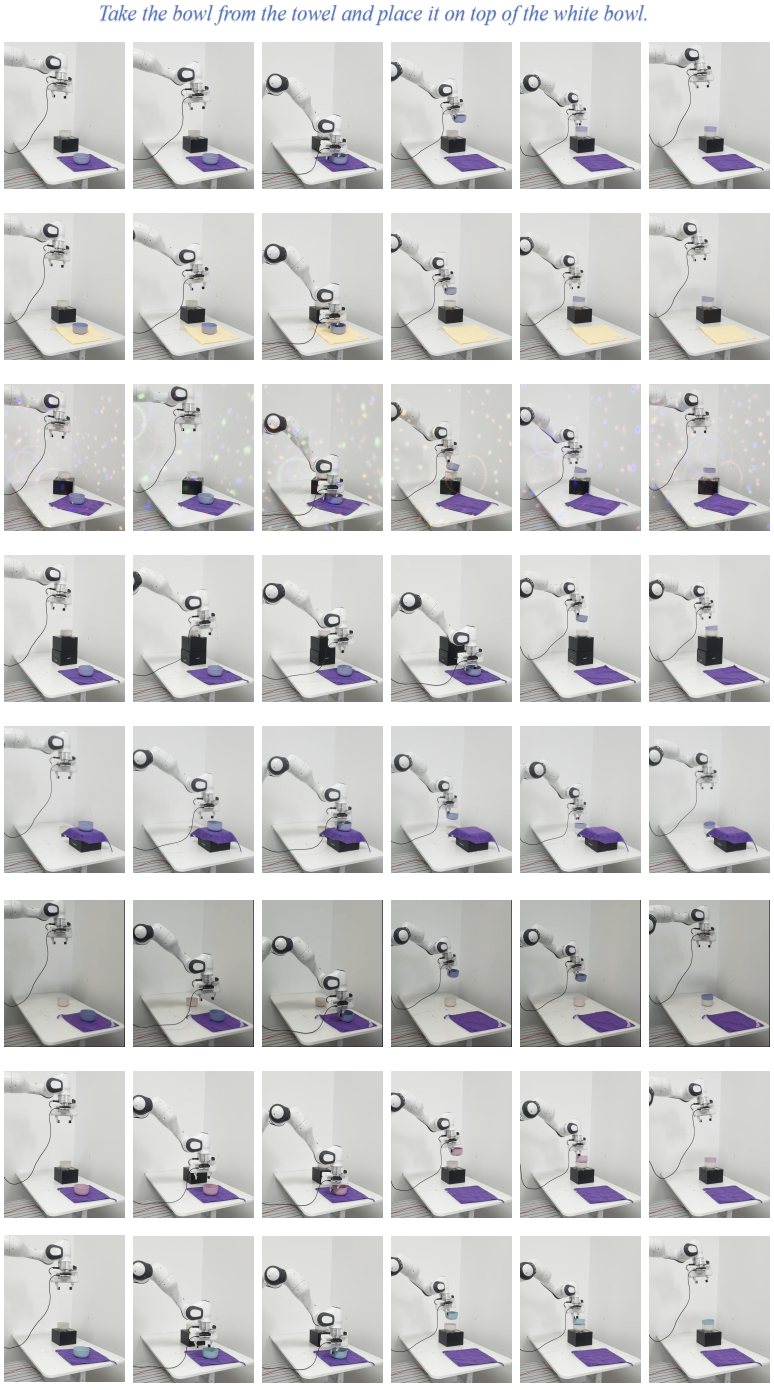


Figure 11: **ID and OOD trajectories for *Stack Bowls*.** Row 1 shows the ID setting, Rows 2 and 3 show background and lighting variations, Rows 4–6 show height variations, and Rows 7–8 show novel-object variations.

Take the sausage toy from the turntable, place it inside the drawer, and then close the drawer.

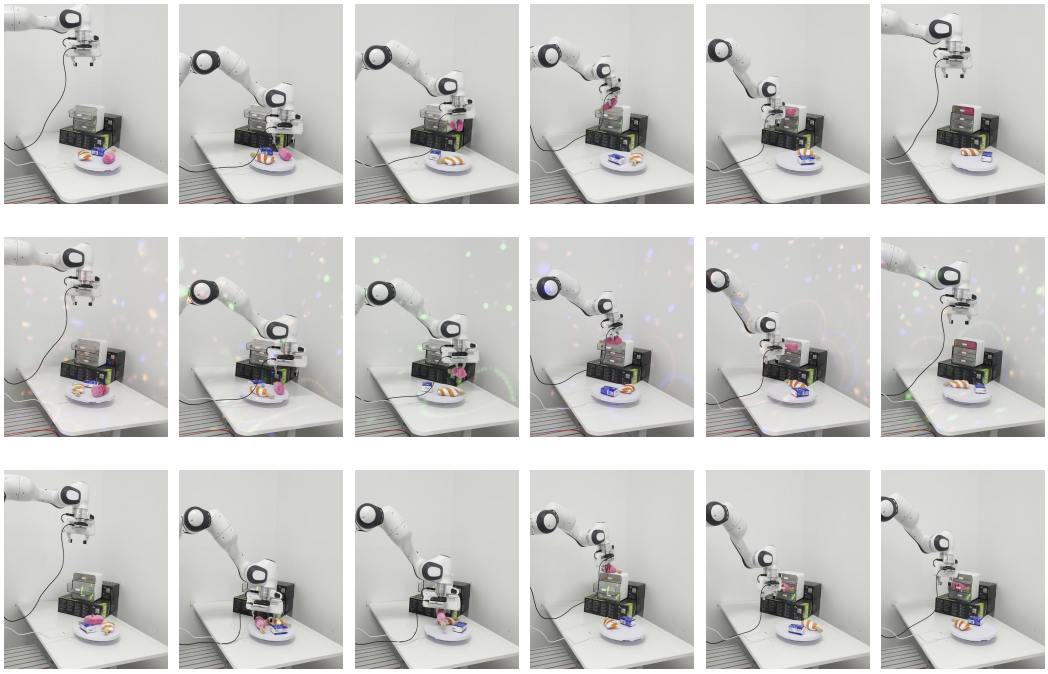


Figure 12: **ID and OOD trajectories for *Put Hotdog in Drawer*.** Row 1 shows the ID setting, while Rows 2 and 3 show lighting and drawer-height variations, respectively.

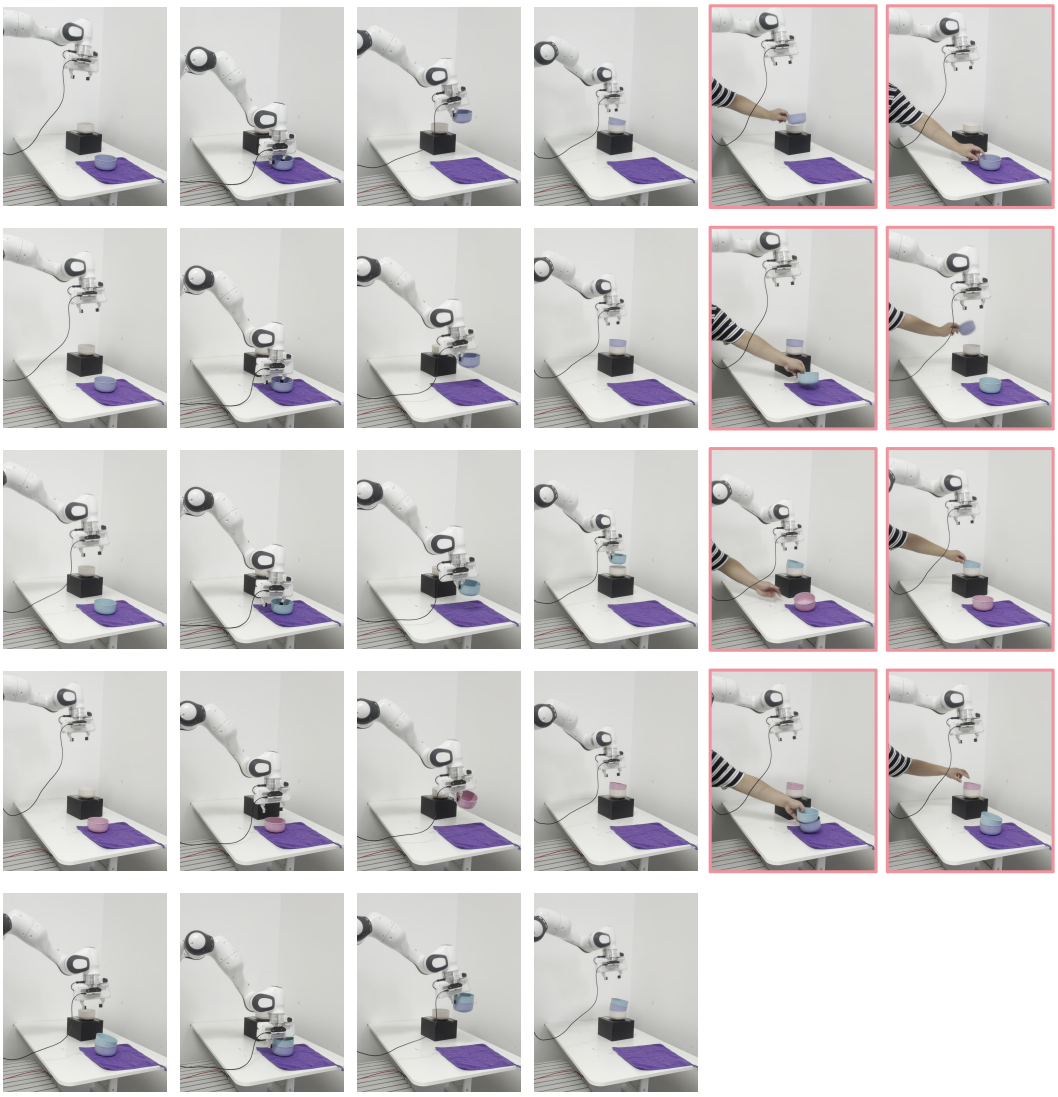


Figure 13: **Continuous human–robot interaction.** Dream-Lift-VLA completes successive *Stack Bowls* tasks through continuous long-horizon inference. Pink-bordered frames indicate human interventions: resetting the scene in Row 1, replacing the bowl with different-colored instances in Rows 2–3, and introducing two bowls simultaneously in Row 4.