DIDO: DISTILLING INTERACTION-CENTRIC DYNAMICS INTO ONE-STEP DENOISING FOR WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) use video generation models to predict future visual dynamics for robotic manipulation, but iterative denoising introduces additional latency for closed-loop control. We empirically find that visual content converges at different rates during denoising. Static background structure forms early, whereas the gripper and manipulated object remain blurry after the first step, with their interaction dynamics emerging only through subsequent denoising. Consequently, naively truncating a multi-step video model to one step preserves scene structure but loses the interaction-centric dynamics most critical for manipulation. To address this issue, we propose DIDO, which distills the converged dynamics of a multi-step video model into a single denoising step. DIDO combines distribution matching distillation with interaction-centric representation guidance. Beyond compressing multi-step generation into one forward pass, DIDO explicitly models the gripper, manipulated object, and their interaction using supervised bounding-box visual reasoning tokens. Additionally, DIDO aligns the target object's representations across multiple model layers with features from a pretrained DINOv3 encoder. This interaction-centric guidance helps the distilled model preserve both the relevant entities and their future dynamics in a single step, while substantially reducing inference latency. DIDO achieves an average success rate of 99.0% on LIBERO, 76.6% on LIBERO-Plus, and 92.0% on RoboTwin, while also demonstrating effective transfer to long-horizon and generalization tasks in real-world robotic manipulation.

1 INTRODUCTION

Vision-Language-Action Models (VLAs) have achieved strong performance in robotic manipulation by leveraging large-scale vision-language pretraining to follow language instructions across diverse tasks and embodiments (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025b; Physical Intelligence et al., 2025). Most VLA policies predict actions directly from the current observation and language instruction, without explicitly modeling how the environment will evolve under the robot's actions. Recently, World Action Models (WAMs) have explored a complementary direction by using video generation models to predict future visual dynamics (Bai et al., 2026b; Wu et al., 2024; Hu et al., 2025; Ye et al., 2026b). These predictions provide the policy with explicit foresight into how the robot, manipulated objects, and their interactions evolve over time. However, realizing such foresight at inference time remains computationally expensive because video generation typically relies on iterative denoising.

A natural way to reduce this cost is to truncate the denoising process and extract predictive representations from only one or a few early steps (Hu et al., 2025; Chi et al., 2025). This substantially improves inference efficiency and has been shown to retain useful predictive information. Yet it remains unclear how much of the dynamics represented by the fully denoised video model are already present at these early steps. In particular, when denoising is reduced to a single step, what dynamics are preserved, and what dynamics are still missing?

To answer this question, we conduct an empirical study of the denoising trajectory in Figure 1, examining both decoded videos and latent representations. We observe a clear spatial and temporal

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.

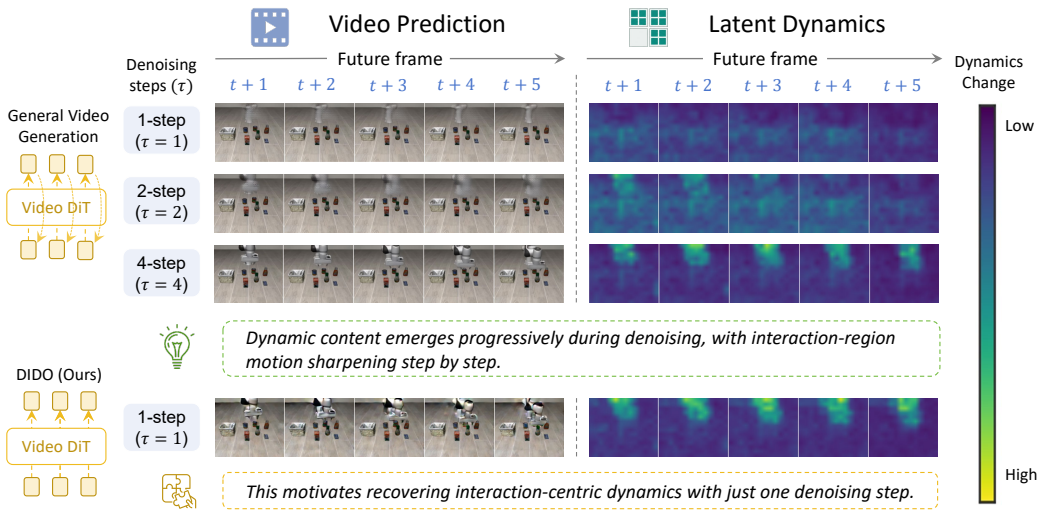


Figure 1: Interaction-centric dynamics emerge progressively during denoising and can be recovered in a single step. We compare a four-step teacher fine-tuned on robot data (top; rows show predictions after 1, 2, and 4 denoising steps) with DIDO (bottom). The four-step robot-domain teacher shows that scene structure forms early, while gripper-object dynamics sharpen across denoising steps. DIDO recovers these dynamics in a single step.

asymmetry. After a single step, static background structure is already well formed, whereas the gripper and manipulated object remain blurry and poorly resolved. The same pattern appears in latent space: inter-frame changes in these interaction regions are weak at early steps but progressively strengthen and become spatially localized as denoising proceeds. This behavior is consistent with the coarse-to-fine nature of diffusion models, where global structure forms early while finer content is refined over subsequent steps (Choi et al., 2022; Balaji et al., 2022). These findings show that naive single-step truncation largely preserves scene structure but misses the interaction-centric dynamics that emerge later in denoising.

We instantiate this insight in DIDO, a World Action Model that learns interaction-centric future dynamics within a single denoising step. In Stage I, DIDO jointly combines distribution matching distillation (DMD) (Yin et al., 2024) with interaction-centric visual reasoning. DMD transfers the converged dynamics of a multi-step video model, while dedicated object, interaction, gripper, and alignment tokens explicitly ground the distilled dynamics in manipulation-relevant entities. The object and gripper tokens are supervised through future bounding-box trajectories to capture their spatial and temporal evolution, while the alignment tokens are aligned with target-object features from a pretrained DINOv3 encoder across multiple transformer layers (Siméoni et al., 2026), reinforcing discriminative visual information throughout the video model. In Stage II, the resulting single-step world model is coupled with an action expert through a mixture-of-transformers architecture, allowing the policy to directly access its predictive representations without decoding future frames. Together, these designs enable DIDO to capture interaction-centric future dynamics in one forward pass and efficiently exploit them for closed-loop control.

The main contributions are threefold. First, we reveal that static scene structure forms early during video denoising, while manipulation-relevant dynamics around the gripper and target object emerge progressively at later steps. Second, we propose DIDO, which distills multi-step dynamics into a single denoising step and introduces interaction-centric visual reasoning with trajectory supervision and multi-layer visual feature alignment. Third, DIDO substantially improves the efficiency of future dynamics modeling for closed-loop control while achieving 99.0% on LIBERO, 76.6% on LIBERO-Plus, and 92.0% on RoboTwin, with effective transfer to real-world manipulation.

2 Related Work

World Action Models for Robotic Manipulation. Vision-Language-Action Models (VLAs) have achieved strong performance in robotic manipulation (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025b; Physical Intelligence et al., 2025; Bai et al., 2026a). World Action Models

(WAMs) explore a complementary direction by incorporating video-based future modeling into action prediction (Ye et al., 2026b; Li et al., 2026). Early approaches synthesize future frames or subgoal images to guide action generation (Du et al., 2023; Black et al., 2024; Tian et al., 2025; Wu et al., 2024), while subsequent methods condition policies on predictive representations (Hu et al., 2025; Wang et al., 2026b; Lyu et al., 2026). Recent WAMs further couple video and action through joint generative modeling, latent action representations, or multi-stream architectures (Cheang et al., 2024; Bi et al., 2026; Team et al., 2026b). Action-centered variants also use future visual dynamics primarily as training supervision or optional inference-time guidance (Ye et al., 2026a; Team et al., 2026a). For WAMs that explicitly generate future visual dynamics, inference typically relies on multi-step denoising. DIDO instead distills these dynamics into a single denoising step while explicitly grounding them in gripper-object interactions.

Efficient World Action Models. Recent work improves WAM efficiency through several strategies. One line removes test-time future imagination entirely. Fast-WAM predicts actions without generating future observations (Yuan et al., 2026), eliminating the cost of explicit future prediction. A second line retains future modeling but truncates denoising, conditioning the policy on representations extracted from only one or a few early steps (Chi et al., 2025; Hu et al., 2025; Zhao et al., 2026). Other approaches preserve future prediction while reducing architectural or conditioning overhead, for example by simplifying the action module, reusing future features across action denoising steps (Ma et al., 2026), or moving auxiliary geometric supervision to training time (Zhang et al., 2026). Closest to DIDO are methods that distill multi-step video or video-action predictions into one-step generation or predictive representations (Akbari et al., 2026; Yan et al., 2026). DIDO follows this one-step distillation direction, but goes beyond generic compression by introducing interaction-centric visual reasoning tokens, supervised by gripper-object trajectories and multi-layer visual alignment, to recover the dynamics that early denoising fails to capture.

Object- and Interaction-Centric Representations for Manipulation. Object-centric VLA methods structure policy representations around task-relevant entities using visual prompts, object-conditioned features, slots and relations, or 3D object representations (Li et al., 2025c;a; Hanyu et al., 2026; Liu et al., 2026b). Recent WAMs extend this idea to future prediction by modeling object masks (Yu et al., 2026; Lou et al., 2026), temporally persistent robot/object slots (Liu et al., 2026a), or structured semantic, geometric, and kinematic foresight (Yan et al., 2026; Jin et al., 2026; Zheng et al., 2026b; Ji et al., 2025). While these approaches make future prediction increasingly object- and structure-aware, robot–object interactions are still largely represented implicitly. DIDO instead adopts interaction-centric visual reasoning, explicitly integrating gripper–object trajectories with visual representations during video generation to capture manipulation-relevant dynamics.

3 METHOD

DIDO learns interaction-centric future dynamics in a single denoising step and directly exploits them for action prediction. It combines one-step dynamics distillation (Section 3.1), interaction-centric visual reasoning (Section 3.2), and interaction-aware action prediction (Section 3.3). The overall training strategy is described in Section 3.4. Additional architectural details, including token configurations, action-expert design, and shared attention, are provided in Appendix B.

3.1 ONE-STEP WORLD DYNAMICS DISTILLATION

Given a current observation o and language instruction ℓ, the video model predicts a future latent z describing how the scene evolves under the robot's actions. We build the video model on Wan2.2-TI2V-5B (Wan et al., 2025) and initialize it from its publicly released four-step distilled variant (Li et al., 2025b). Before the two-stage training of DIDO, we first adapt this four-step generator to robot manipulation demonstrations, yielding the robot-domain teacher used for distillation.

We distill the four-step teacher into a single-step generator using distribution matching distillation (DMD) (Yin et al., 2024). Let $p_{\mathrm{T},t}$ and $p_{\theta,t}$ denote the perturbed distributions induced by the four-step teacher and single-step generator at diffusion timestep t, respectively. DMD optimizes

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_t \left[D_{\mathrm{KL}} \left(p_{\theta,t} \parallel p_{\mathrm{T},t} \right) \right], \tag{1}$$

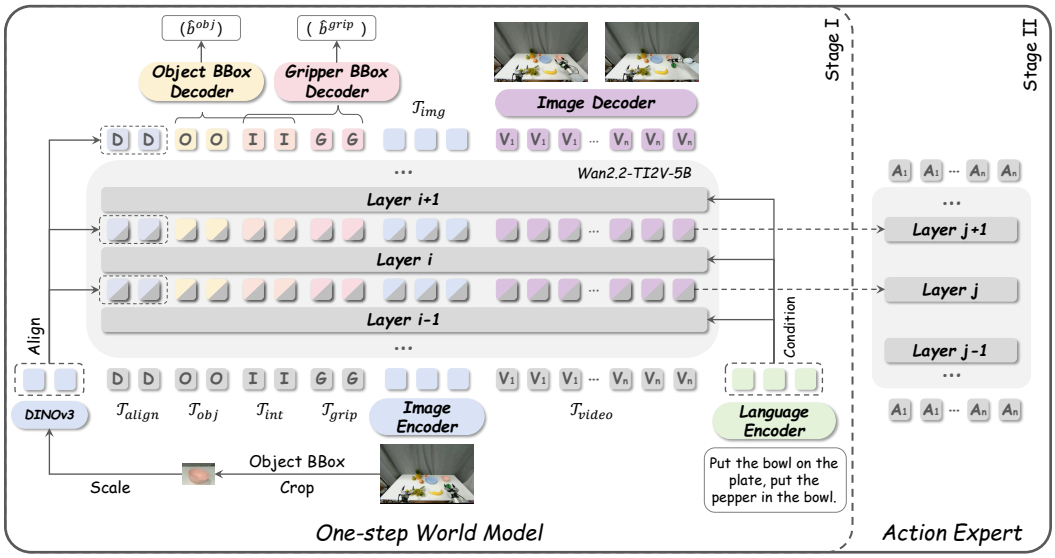


Figure 2: **Overview of DIDO.** In Stage I, the one-step world model augments the video transformer with four interaction-centric token groups, including object $\mathcal{T}_{\text{obj}}$, interaction $\mathcal{T}_{\text{int}}$, gripper $\mathcal{T}_{\text{grip}}$, and alignment $\mathcal{T}_{\text{align}}$ tokens. The object and gripper branches predict future bounding-box trajectories $\hat{b}^{\text{obj}}$ and $\hat{b}^{\text{grip}}$ using the shared interaction tokens, while the alignment tokens are supervised by DINOv3 features extracted from the target object crop. Together with the image-conditioning and video representations $\mathcal{T}_{\text{video}}$, these tokens evolve through the video transformer to capture interaction-centric future dynamics. In Stage II, an action expert is coupled to the world model through a mixture-of-transformers architecture and directly accesses its intermediate representations for action prediction without decoding and re-encoding future video.

whose gradient is estimated using real and fake score networks. DMD trains the generator to match the output distribution of the multi-step teacher, enabling a single forward pass to approximate the teacher's converged future dynamics. Additional training details are provided in Appendix B.1. While the distilled generator can recover the gripper and manipulated object in the predicted future, the distillation objective does not explicitly encourage modeling their interaction. We therefore introduce interaction-centric visual reasoning to further capture their coupled dynamics.

3.2 Interaction-Centric Visual Reasoning

To explicitly ground the distilled dynamics in manipulation-relevant entities, we introduce four interaction-centric token groups

$$\mathcal{T}_{\text{inter}} = [\mathcal{T}_{\text{obj}}; \mathcal{T}_{\text{int}}; \mathcal{T}_{\text{grip}}; \mathcal{T}_{\text{align}}], \tag{2}$$

representing the target object, gripper-object interaction, gripper, and object-centric visual alignment, respectively. These tokens are embedded directly into the video transformer and evolve jointly with the video representations. We supervise them through complementary trajectory and visual alignment objectives.

Trajectory Supervision. We supervise the future spatial evolution of the gripper and target object over the action horizon. Let H_a denote the action chunk size. The gripper branch $[\mathcal{T}_{\text{grip}}; \mathcal{T}_{\text{int}}]$ predicts a sequence of H_a future bounding boxes $\hat{b}^{\text{grip}} = \{\hat{b}_h^{\text{grip}}\}_{h=1}^{H_a}$, while the object branch $[\mathcal{T}_{\text{obj}}; \mathcal{T}_{\text{int}}]$ predicts $\hat{b}^{\text{obj}} = \{\hat{b}_h^{\text{obj}}\}_{h=1}^{H_a}$. Sharing $\mathcal{T}_{\text{int}}$ across both branches encourages it to capture their relative spatial and temporal evolution.

For a predicted bounding-box sequence $\hat{b} = \{\hat{b}_h\}_{h=1}^{H_a}$ and its target $b = \{b_h\}_{h=1}^{H_a}$, we define

$$\mathcal{L}_{\text{box}}(\hat{b}, b) = \frac{1}{H_a} \sum_{h=1}^{H_a} \left[\text{smooth}_{L_1}(\hat{b}_h, b_h) + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}(\hat{b}_h, b_h) \right]. \tag{3}$$

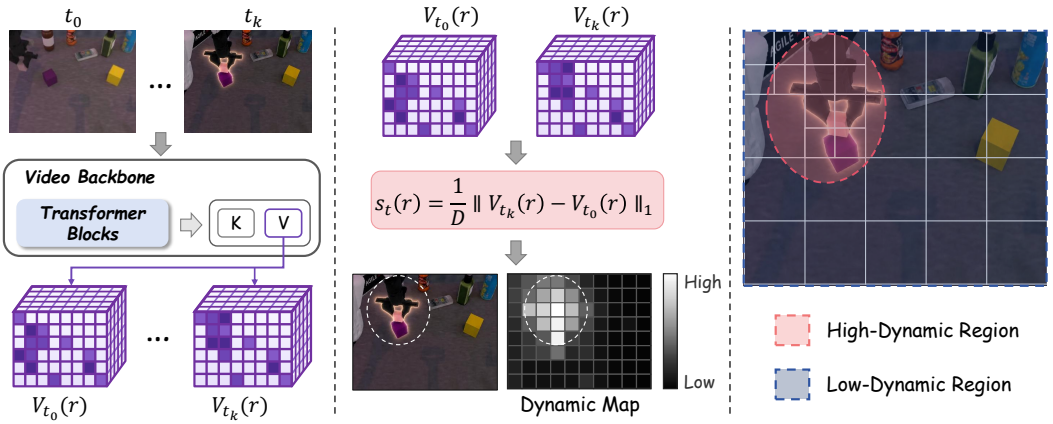


Figure 3: **Dynamics-based token refinement.** Given the current timestamp t_0 and a future timestamp t_k, we obtain their corresponding value features V_{t_0} and V_{t_k} from the video backbone and compute cell-wise differences to construct a dynamic map. The spatial feature map is partitioned into a grid, with each region containing 2×2 fine-grained cells. We rank the regions by the averaged dynamics of their four cells, retain the top-3 high-dynamic regions at full token resolution, and apply 2×2 average pooling to the remaining regions. This preserves fine-grained interaction-relevant dynamics while substantially reducing redundant future-video tokens.

The interaction trajectory objective is

$$\mathcal{L}_{\text{inter}} = \lambda_{\text{obj}} \mathcal{L}_{\text{box}} \left(\hat{b}^{\text{obj}}, b^{\text{obj}} \right) + \lambda_{\text{grip}} \mathcal{L}_{\text{box}} \left(\hat{b}^{\text{grip}}, b^{\text{grip}} \right). \tag{4}$$

Multi-Layer Visual Alignment. Trajectory supervision provides spatial and temporal grounding, but bounding-box prediction alone does not ensure that the video model preserves discriminative visual information about the manipulated object. We therefore supervise $\mathcal{T}_{\text{align}}$ with object-centric features extracted by a frozen DINOv3 encoder (Siméoni et al., 2026). Specifically, we crop the manipulated-object region using its bounding box, resize the crop to the DINOv3 input resolution, and encode it into a sequence of visual tokens. At selected video-transformer layers, the intermediate representations of $\mathcal{T}_{\text{align}}$ are aligned with these frozen DINOv3 tokens.

Let $\mathcal{S}$ denote the transformer layers used for alignment, $\mathcal{T}^l_{\text{align}}$ the alignment-token representations at layer l, and F_{DINO} the frozen DINOv3 tokens extracted from the object crop. We define

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \left\| P_l \left(\mathcal{T}^l_{\text{align}} \right) - F_{\text{DINO}} \right\|_1, \tag{5}$$

where P_l projects the alignment-token representations into the DINOv3 feature space.

The required bounding-box trajectories and visual targets are automatically curated from the original demonstrations without manual labeling; details are provided in Appendix A.

3.3 Interaction-Aware Action Prediction

DIDO couples the single-step video model with an action expert through a mixture-of-transformers architecture, allowing the action expert to directly attend to the video model's internal representations without decoding and re-encoding future video. As illustrated in Figure 2, the action expert accesses the complete latent stream of the video model, including image-conditioning, interaction-centric, and future-video representations. Only the future-video representations are selectively refined before being exposed to the action expert.

Dynamics-Based Token Refinement. Future-video representations contain a large number of tokens and substantial spatial redundancy, while manipulation-relevant changes are typically concentrated in a few local regions. We therefore preserve fine-grained tokens in regions with strong predicted dynamics and compress the remaining regions. Since the configuration is adapted to the resolution of input images, the concrete values may differ. Here we illustrate with RoboTwin.

As illustrated in Figure 3, let t_0 denote the current frame and t_k the k-th future frame, with V_{t_0} and V_{t_k} denoting their corresponding value features, $V_{t_0}, V_{t_k} \in \mathbb{R}^{H\times W\times D}$. We partition the spatial feature map into a 4×5 grid of 20 regions, each containing a 2×2 set of fine-grained cells. For each cell $r = (h, w)$, we measure its predicted change as

$$s_{t_k}(r) = \frac{1}{D} \left\| \mathbf{V}_{t_k}(r,:) - \mathbf{V}_{t_0}(r,:) \right\|_1 . \tag{6}$$

For each region R_m, we average the scores of its four cells:

$$S_{t_k}(R_m) = \frac{1}{|R_m|} \sum_{r\in R_m} s_{t_k}(r), \qquad |R_m| = 4. \tag{7}$$

We rank the 20 regions by $S_{t_k}(R_m)$ and retain all fine-grained tokens in the top-3 regions. For each of the remaining 17 regions, 2×2 average pooling merges its four cells into a single token. The same spatial refinement is applied to the corresponding key and value features before being accessed by the action expert, while all other video-model latents remain unchanged.

Flow-Matching Action Generation. Conditioned on the video-model representations, the action expert generates an action chunk of length H_a via flow matching. Let $a = \{a_h\}_{h=1}^{H_a}$ denote the ground-truth action chunk. We sample $\tau \sim U[0, 1]$ and $\epsilon \sim \mathcal{N}(0, I)$ with the same shape as a, and construct $a_\tau = (1 - \tau)\epsilon + \tau a$ with target velocity $u = a - \epsilon$. The action expert v_θ is optimized by

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{a,\tau,\epsilon} \left[\left\| v_\theta \left(a_\tau, \tau \right) - u \right\|_2^2 \right]. \tag{8}$$

3.4 Training Strategy

DIDO is trained in two stages after preparing the robot-domain four-step teacher described in Section 3.1; detailed optimization settings are provided in Appendix C.2.

Stage I: One-Step Interaction-Centric Dynamics Learning. The first stage jointly optimizes one-step distillation, interaction trajectory supervision, and visual alignment:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{DM}} + \mathcal{L}_{\text{inter}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{9}$$

Here, $\mathcal{L}_{\text{DM}}$ transfers the teacher's converged future distribution, $\mathcal{L}_{\text{inter}}$ grounds the prediction in the gripper and target-object trajectories, and $\mathcal{L}_{\text{align}}$ preserves object-centric visual information.

Stage II: Joint Policy Learning. We then initialize the action expert and jointly optimize it with the single-step video model. To preserve predictive dynamics during policy learning, we retain the video-generation objective $\mathcal{L}_{\text{video}}$ (Yuan et al., 2026) together with the interaction-centric objectives:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{video}}\mathcal{L}_{\text{video}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \mathcal{L}_{\text{inter}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{10}$$

The video-model backbone, interaction-centric tokens, prediction heads, and action expert are jointly optimized, while the language encoder, VAE, and DINOv3 encoder remain frozen. DMD is used only in Stage I; the video generator remains single-step throughout Stage II and inference.

4 Experiments

Our experiments are designed to address the following questions: **RQ1**: Does DIDO achieve state-of-the-art performance on standard manipulation benchmarks? (Section 4.2) **RQ2**: How well does DIDO generalize under systematic perturbations in simulation and to real-world manipulation? (Section 4.3) **RQ3**: What drives DIDO's performance gains, and how do its components affect accuracy, inference efficiency, and token utilization? (Section 4.4)

4.1 Experimental Setup

Benchmarks. We evaluate DIDO on two standard manipulation benchmarks, one simulation benchmark for generalization, and a real-robot setting. LIBERO (Liu et al., 2023) contains four

Table 1: Performance comparison on LIBERO and RoboTwin. Results are reported as success rates (%). PT indicates the use of pretraining. The best and second-best average results on each benchmark are shown in **bold** and underlined, respectively.

		LIBERO					RoboTwin		
Method	Emb.PT	Spatial	Object	Goal	Long	Avg.	Clean	Rand.	Avg.
Vision-Language-Action Models									
π_0 (Black et al., 2025)	✓	96.8	98.8	95.8	85.2	94.1	65.9	58.4	62.2
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	✓	98.8	98.2	98.0	92.4	96.9	82.7	76.8	79.8
World Action Models									
LingBot-VA (Li et al., 2026)	✓	98.5	99.6	97.2	98.5	98.5	92.9	91.5	**92.2**
Motus (Bi et al., 2026)	✓	96.8	99.8	96.6	97.6	97.7	88.7	87.0	87.8
Faster-WAM (Ma et al., 2026)	✗	98.4	100.0	97.0	97.8	98.5	89.7	88.6	89.2
Flash-WAM (Akbari et al., 2026)	✗	97.0	92.8	96.4	98.0	96.1	88.4	82.7	85.5
Fast-WAM (No co-video)	✗	89.2	99.2	95.4	90.0	93.5	82.8	84.8	83.8
Fast-WAM (Yuan et al., 2026)	✗	98.2	100.0	97.0	95.2	97.6	91.9	91.8	91.8
DIDO (No co-video, Ours)	✗	97.8	99.4	97.2	96.4	97.7	89.6	89.0	89.3
DIDO (Ours)	✗	99.8	99.6	98.6	98.0	**99.0**	92.0	92.0	92.0

suites, Spatial, Object, Goal, and Long, covering spatial reasoning, object generalization, goal understanding, and long-horizon execution. RoboTwin (Chen et al., 2025) is a dual-arm simulation benchmark with multi-camera observations and contact-rich bimanual tasks. We additionally use LIBERO-Plus (Fei et al., 2025) to evaluate robustness under controlled distribution shifts, and four real-world manipulation tasks to assess transfer beyond simulation. Simulation performance is reported as success rate, while real-world performance is measured by normalized task-progress scores. Detailed real-world tasks and scoring protocols are provided in Appendix C.3.

Baselines. We compare DIDO with two groups of methods. The first includes VLA models that predict actions without explicit future visual modeling, including π_0 (Black et al., 2025), $\pi_{0.5}$ (Physical Intelligence et al., 2025), and, on LIBERO-Plus, OpenVLA-OFT (Kim et al., 2025a) and X-VLA (Zheng et al., 2026a). The second includes WAMs that incorporate future visual prediction into action generation, including LingBot-VA (Li et al., 2026) and Motus (Bi et al., 2026) with multi-step future prediction, as well as Fast-WAM (Yuan et al., 2026), Faster-WAM (Ma et al., 2026), Flash-WAM (Akbari et al., 2026), and ST-WAM (Wang et al., 2026a).

4.2 RQ1: Performance on Standard Benchmarks

LIBERO. Table 1 reports success rates across the four LIBERO suites. DIDO achieves an average success rate of 99.0%, outperforming the strongest prior methods, LingBot-VA and Faster-WAM, by 0.5 percentage points. Compared with Fast-WAM, DIDO improves the average success rate by 1.4 points. Notably, DIDO also exceeds Flash-WAM, another single-step distilled WAM, by 2.9 points, demonstrating that DIDO retains stronger manipulation-relevant predictive representations under single-step generation.

RoboTwin. On RoboTwin, Table 1 shows that DIDO achieves an average success rate of 92.0%, only 0.2 percentage points below the overall best result of LingBot-VA at 92.2%. Unlike LingBot-VA, DIDO does not rely on embodied pretraining, and achieves the best performance among methods trained without such pretraining. DIDO further outperforms Fast-WAM and Flash-WAM by 0.2 and 6.5 percentage points, respectively. These results show that the advantages of DIDO extend from single-arm manipulation in LIBERO to contact-rich bimanual tasks in RoboTwin.

4.3 RQ2: Robustness and Generalization

LIBERO-Plus. Table 2 evaluates robustness under seven controlled distribution shifts. DIDO achieves an average success rate of 76.6%, outperforming Faster-WAM by 1.6%. It ranks first under robot initialization, language phrasing, lighting, and object layout perturbations, reaching 74.3%, 94.4%, 96.6%, and 83.3%, respectively. DIDO outperforms the pretrained X-VLA which is

Table 2: Robustness evaluation on LIBERO-Plus under systematic perturbations. Results are reported as success rates (%). Emb. PT. denotes embodied pretraining. The best and second-best results are shown in **bold** and underlined, respectively.

Method	Emb. PT.	Camera	Robot	Lang.	Light	Backg.	Noise	Layout	Avg.
Vision-Language-Action Models									
OpenVLA-OFT (Kim et al., 2025a)	✓	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0 (Black et al., 2025)	✓	81.4	13.8	58.8	85.0	68.9	6.0	79.0	53.6
X-VLA (Zheng et al., 2026a)	✓	96.0	23.4	75.7	88.2	71.8	89.7	62.7	71.4
World Action Models									
Fast-WAM (Yuan et al., 2026)	✗	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
ST-WAM (Wang et al., 2026a)	✗	55.4	60.1	79.3	93.0	74.2	79.5	74.3	72.8
Faster-WAM (Ma et al., 2026)	✗	67.9	49.0	92.1	94.3	57.0	82.3	82.7	75.0
DIDO (Ours)	✗	45.3	74.3	94.4	96.6	72.7	75.0	83.3	**76.6**

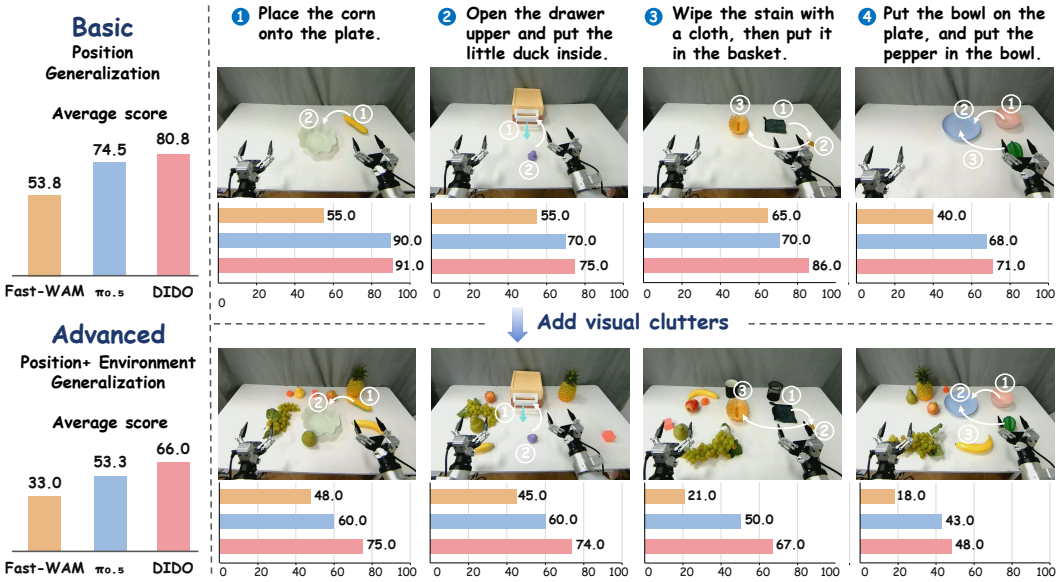


Figure 4: Real-world generalization on Galbot G1. The basic protocol evaluates unseen target positions, while the advanced protocol additionally introduces visual clutter. Bars show normalized task scores for Fast-WAM, $\pi_{0.5}$, and DIDO, with the average across four tasks.

pretrained on large scale robot datas by 5.2 points. Overall, these results demonstrate DIDO's strong robustness across diverse distribution shifts.

Real-World Generalization. We further evaluate DIDO on Galbot G1 across four manipulation tasks, with 150 demonstrations per task. The *basic* protocol evaluates position generalization using target configurations unseen during training, while the *advanced* protocol additionally introduces task-irrelevant visual clutter to test robustness to simultaneous changes in target position and scene appearance. DIDO achieves the highest average score under both protocols and performs best on all four tasks. In the basic setting, DIDO reaches 80.8, compared with 74.5 for $\pi_{0.5}$ and 53.8 for Fast-WAM. Under the more challenging advanced setting, DIDO maintains 66.0, versus 53.3 and 33.0, respectively. The average degradation from basic to advanced is 14.8 points for DIDO, smaller than 21.2 for $\pi_{0.5}$ and 20.8 for Fast-WAM. Since the same policies are evaluated without additional adaptation, these results demonstrate stronger generalization under combined shifts in target position and scene appearance.

4.4 RQ3: Ablation and Analysis

Building on the observation in Figure 1 that interaction-centric dynamics emerge progressively during denoising, we first examine whether later denoising steps provide more useful information for

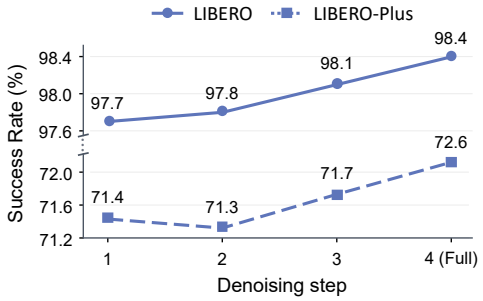


Figure 5: Effect of denoising steps on downstream action prediction using a four-step video model and action expert.

Table 3: **Component ablation.** Starting from the direct one-step truncation baseline, we cumulatively enable One-Step DMD Distillation, Interaction-Centric visual Reasoning, and Dynamics-Based token Refinement to quantify the contribution of each component on LIBERO and LIBERO-Plus.

1-Step Distill.	Interact. Reason.	Dyn. Refine.	LIBERO	LIBERO-Plus
			97.7	71.4
✓			98.3	72.3
✓	✓		98.9	75.9
✓	✓	✓	**99.0**	**76.6**

action prediction. We then conduct controlled ablations on LIBERO and LIBERO-Plus to isolate the contribution of each DIDO component and analyze its inference efficiency.

Dynamics Emerge Progressively During Denoising. We first examine whether later denoising steps provide more useful predictive representations for downstream control. To isolate this effect from the proposed DIDO components, we use the four-step Wan video model before one-step distillation and couple it directly with an Action DiT expert, without interaction-centric tokens or dynamics-based token refinement. The video model and action expert are jointly trained for policy learning, and at evaluation time the action expert is conditioned on the intermediate video representation after denoising step $t \in \{1, 2, 3, 4\}$. All model parameters and evaluation settings are fixed, with only the denoising step used for action conditioning varied. As shown in Figure 5, performance generally improves as denoising proceeds, increasing from 97.7% to 98.4% on LIBERO and from 71.4% to 72.6% on LIBERO-Plus. This confirms that later denoising steps expose additional action-relevant dynamics, motivating the need to recover such information within a single step.

Contribution of Each Component. We construct DIDO cumulatively from the one-step truncation baseline, as shown in Table 3. Naively truncating the fine-tuned four-step video model to one denoising step achieves 97.7% on LIBERO and 71.4% on LIBERO-Plus. One-Step Dynamics Distillation improves performance to 98.3% and 72.3%, indicating that distillation preserves more of the converged multi-step dynamics within a single step. Adding Interaction-Centric Visual Reasoning further raises performance to 98.9% and 75.9%, demonstrating the benefit of explicitly modeling the target object, gripper, and their interaction. Additional analysis of interaction-centric tokens is provided in Appendix D.2. Finally, Dynamics-Based Token Refinement achieves 99.0% and 76.6%, while reducing the conditioning context and end-to-end latency by approximately 30 ms. Its larger improvement on LIBERO-Plus suggests that selectively compressing near-static regions is particularly beneficial under visual distribution shifts.

Inference Cost. On an NVIDIA H100 with batch size 1, DIDO predicts an action chunk in 384 ms end to end, compared with 562 ms for the four-step teacher, yielding a 32% latency reduction while improving LIBERO performance from 98.4% to 99.0%. Compared with Fast-WAM, which removes future imagination at inference, DIDO introduces only 28 ms of additional latency, while improving performance by 1.4 points on LIBERO and 25.1 points on LIBERO-Plus.

5 Conclusion

World Action Models offer a promising way to leverage predictive video dynamics for robotic manipulation, but iterative denoising introduces substantial inference latency, while naive single-step truncation can lose the interaction-centric dynamics that emerge later during denoising. We introduced DIDO, which distills converged multi-step dynamics into a single denoising step with interaction-centric reasoning over the gripper and object. DIDO achieves 99.0% on LIBERO, 92.0% on RoboTwin, and 76.6% on LIBERO-Plus, while also demonstrating strong generalization in real-world manipulation. These results show that efficient predictive dynamics can retain manipulation-relevant interaction information, enabling WAMs to support fast and effective closed-loop control.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability and clarity of the manuscript and to assist with parts of the implementation, evaluation, and plotting code. Generative AI was not used to formulate research ideas or hypotheses, design the proposed method or experiments, interpret experimental results, or generate synthetic data. All AI-assisted text and code were reviewed and verified by the authors, and every result reported in the paper was obtained from our own executed experiments. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the paper and appendix. The model architecture and training procedure are described in the main text of Section 3, with additional architectural details provided in Appendix B. Data curation and annotation procedures, including the construction of interaction-centric supervision, are described in Appendix A. Training configurations, optimization hyperparameters, loss weights, and evaluation protocols are provided in Appendix C. We will publicly release the code upon acceptance.

REFERENCES

Arman Akbari, Ci Zhang, Arash Akbari, Lin Zhao, Yixiao Chen, Weiwei Chen, Xuan Zhang, Geng Yuan, and Yanzhi Wang. Flash-wam: Modality-aware distillation for world action models. *arXiv preprint arXiv:2606.05254*, 2026.

Shuanghao Bai, Jing Lyu, Wanqi Zhou, Zhe Li, Dakai Wang, Lei Xing, Xiaoguang Zhao, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Badong Chen, and Shanghang Zhang. Latent reasoning vla: Latent thinking and prediction for vision-language-action models. In *Forty-third International Conference on Machine Learning*, 2026a.

Shuanghao Bai, Wenxuan Song, Jiayi Chen, Yuheng Ji, Zhide Zhong, Jin Yang, Han Zhao, Wanqi Zhou, Zhe Li, Pengxiang Ding, et al. Embodied robot manipulation in the era of foundation models: Planning and learning perspectives. *IEEE Transactions on Robotics*, 2026b.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

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, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 33431–33452, 2024.

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. In *Robotics: Science and Systems (RSS)*, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems*, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Xiaowei Chi, Kuangzhi Ge, Jiaming Liu, Siyuan Zhou, Peidong Jia, Zichen He, Yuzhen Liu, Tingguang Li, Lei Han, Sirui Han, et al. Mind: Learning a dual-system world model for real-time planning and implicit risk analysis. *arXiv preprint arXiv:2506.18897*, 2025.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11462–11471. IEEE, 2022.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in neural information processing systems*, volume 36, pp. 9156–9172, 2023.

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

Taisei Hanyu, Nhat Chung, Huy Le, Toan Nguyen, Yuki Ikebe, Anthony Gunderman, Duy Ho Minh Nguyen, Khoa Vo, Tung Kieu, Kashu Yamazaki, Chase Rainwater, Anh Nguyen, and Ngan Le. SlotVLA: Towards modeling of object-relation representations in robotic manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 24328–24346, 2025.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1724–1734, 2025. doi: 10.1109/CVPR52734.2025.00168.

Yuheng Ji, Yuyang Liu, Huajie Tan, Xuchuan Huang, Fanding Huang, Yijie Xu, Cheng Chi, Yuting Zhao, Huaihai Lyu, Peterson Co, Mingyu Cao, Qiongyu Zhang, Zhe Li, Enshen Zhou, Pengwei Wang, Zhongyuan Wang, Shanghang Zhang, and Xiaolong Zheng. Prm-as-a-judge: A dense evaluation paradigm for fine-grained robotic auditing. *arXiv preprint arXiv:2603.21669*, 2026.

Minghao Jin, Mozheng Liao, Mingfei Han, Zhihui Li, and Xiaojun Chang. Beyond dense futures: World models as structured planners for robotic manipulation. In *European Conference on Computer Vision*, 2026.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems*, 2025a.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713. PMLR, 2025b.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. In *Robotics: Science and Systems*, 2026.

Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. In *Proceedings of The 9th Conference on Robot Learning*, pp. 1898–1913, 2025a.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12112–12123. IEEE, 2025b.

Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27638–27648. IEEE, 2025c.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. Oa-wam: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*, 2026a.

Zonghe Liu, Shanyuan Jie, Xiaoquan Sun, Chen Cao, Zetian Xu, Zongsheng Liu, and Jiayu Chen. Sam3d-guided object-centric representation alignment for vision-language-action models. *arXiv preprint arXiv:2607.25912*, 2026b.

Yunfan Lou, Xiaowei Chi, Xiaojie Zhang, Zezhong Qian, Chengxuan Li, Rongyu Zhang, Yaoxu Lyu, Guoyu Song, Chuyao Fu, Haoxuan Xu, Pengwei Wang, and Shanghang Zhang. Mask world model: Predicting what matters for robust robot policy learning. In *Proceedings of the International Conference on Machine Learning*, 2026.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. In *Robotics: Science and Systems*, 2026.

Liheng Ma, Rui Heng Yang, Zhanguang Zhang, Mateo Clemente, Ziwen Hu, Tongtong Cao, and Yingxue Zhang. Faster-wam: Do world action models need deep action modules? *arXiv preprint arXiv:2608.02365*, 2026.

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. In *Conference on Robot Learning*, pp. 17–40. PMLR, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *Transactions on Machine Learning Research*, 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*, 2026a.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026b.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 92033–92052, 2025.

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.

Mingxin Wang, Bin Hu, Bin Qian, Kaitao Jiang, Haoning Wu, Feng Yan, Bowen Jing, Ruiyang Hao, Enyi Wang, Kangning Niu, et al. St-wam: Semantic-temporal world action model for robust manipulation under visual distribution shifts. *arXiv preprint arXiv:2607.28993*, 2026a.

Yihao Wang, Yuheng Ji, Mingyu Cao, Yanqing Shen, Runze Xiao, Huaihai Lyu, Senwei Xie, Euan Liu, Klara Tian, Tianfeng Long, et al. Orca: The world is in your mind. *arXiv preprint arXiv:2606.30534*, 2026b.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Haodong Yan, Zhide Zhong, Jiaguan Zhu, Junjie He, Weilin Yuan, Wenxuan Song, Xin Gong, Yingjie Cai, Guanyi Zhao, Xu Yan, et al. S-vam: Shortcut video-action model by self-distilling geometric and semantic foresight. In *European Conference on Computer Vision*, 2026.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

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*, 2026b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in neural information processing systems*, volume 37, pp. 47455–47487, 2024.

Hanyang Yu, Haitao Lin, Jingbo Zhang, Wenyao Zhang, Chenghao Gu, Heng Li, and Ping Tan. Maskwam: Unifying mask prompting and prediction for world-action models. *arXiv preprint arXiv:2606.13515*, 2026.

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.

Jianjun Zhang, Jian Zhu, Taiyi Su, Chong Ma, Zitai Huang, Yi Xu, and Hanli Wang. Learning 4d geometric priors for inference-efficient world action models. *arXiv preprint arXiv:2607.05468*, 2026.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-wam: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026a.

Yixin Zheng, Jiangran Lyu, Yifan Zhang, Jiayi Chen, Mi Yan, Yuntian Deng, Xuesong Shi, Xiaoguang Zhao, Yizhou Wang, Zhizheng Zhang, and He Wang. Emerging extrinsic dexterity in cluttered scenes via dynamics-aware policy learning. In *Robotics: Science and Systems*, 2026b.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A Data Curation and Annotation

Interaction-centric visual reasoning requires dense bounding-box trajectories for the gripper and manipulated object. Since the original datasets do not provide these annotations, we automatically curate them from the demonstrations without manual labeling.

A.1 Automatic Interaction Annotation

Entity Trajectory Annotation. For the manipulated object, we follow the automatic annotation pipeline of LaRA-VLA (Bai et al., 2026a). The target entity is first parsed from the language instruction and localized in a reference frame using an open-vocabulary detector. A video tracker then propagates the target through the sequence, producing dense frame-level bounding boxes.

For the gripper, LIBERO and RoboTwin provide its pose through simulator state, which is projected into image space to obtain the corresponding bounding boxes. For Galbot G1 in real-world experiments, where such state information is unavailable, the gripper is detected using Grounding DINO (Liu et al., 2024) and propagated through the sequence with SAM 3 (Carion et al., 2026). Tracking is performed in both temporal directions, and the higher-confidence prediction is retained at each frame. Missing intermediate boxes are filled by linear interpolation between valid detections.

Visual Alignment Targets. The manipulated-object annotations are also used to construct the targets for multi-layer visual alignment. For each frame, we crop the target-object region using its annotated bounding box, enlarge the crop by a fixed margin to retain local context, and resize it to the DINOv3 input resolution. A frozen DINOv3 encoder (Siméoni et al., 2026) then extracts a sequence of object-centric visual tokens, which serve as fixed targets for the dedicated alignment tokens $\mathcal{T}_{\text{align}}$. Cropping before encoding keeps the supervision object-specific and reduces interference from static background content.

Annotation Quality. We audit the automatically generated bounding-box annotations at the trajectory level. On LIBERO, the target-object boxes remain correct throughout 95 of 100 sampled trajectories, while 5 exhibit tracking drift; the gripper boxes are correct in all sampled trajectories. On RoboTwin, the target-object boxes are correct in 181 of 200 sampled trajectories, with 9 detection failures and 10 tracking drifts, while the gripper boxes are correct in all cases. On Galbot G1, all 40 sampled trajectories, with 10 per task, contain correct annotations for both the target object and gripper. These results indicate that the automatic pipeline provides reliable supervision for interaction-centric visual reasoning.

B Additional Model Details

This section provides architectural details omitted from the main text.

B.1 One-Step Distribution Matching Distillation

Video Model. DIDO builds on the pretrained Wan2.2-TI2V-5B video generation model (Wan et al., 2025) and is initialized from its publicly released four-step distilled checkpoint (Li et al., 2025b). We further adapt this checkpoint to robot manipulation demonstrations, yielding a robot-domain four-step teacher used for distribution matching in Stage I. The VAE decoder is used only to visualize predicted future frames and is not required during policy inference.

We distill the robot-domain four-step teacher into a one-step generator following the distribution-matching principle of DMD (Yin et al., 2024). We denote the student generator by G_θ, the frozen teacher score model by f_{T}, and the online fake-score model by f_{F}. All three networks are initialized from the same robot-domain four-step checkpoint. During distillation, f_{T} remains frozen, while G_θ and f_{F} are optimized alternately. Importantly, the frozen teacher is queried at a single perturbed student sample during each DMD update; it does not execute the complete four-step denoising trajectory.

Given conditioning information c, an observed first-frame latent x_{obs}, and Gaussian noise z_{future} for the future latent frames, the student predicts the future-video latent in a single network evaluation:

$$x_g = G_\theta\left(z_{\text{future}}, x_{\text{obs}}, c\right). \tag{11}$$

The observed first-frame latent is kept fixed, while the future latent frames are initialized from Gaussian noise.

Distribution Matching Let p_θ denote the distribution induced by the one-step student and p_{T} the target distribution represented by the frozen teacher. Rather than reproducing the teacher's multi-step sampling trajectory, we match their perturbed distributions:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_t\left[D_{\text{KL}}\left(p_{\theta,t} \,\|\, p_{\text{T},t}\right)\right], \tag{12}$$

where $p_{\theta,t}$ and $p_{\text{T},t}$ denote the corresponding distributions at noise level t.

For a noisy distribution p_t, its score is defined as $s(x_t, t) = \nabla_{x_t} \log p_t(x_t)$. The gradient of Eq. (12) with respect to the student parameters can be expressed using the difference between the fake and target scores:

$$\nabla_\theta \mathcal{L}_{\text{DM}} \propto \mathbb{E}\left[\left(s_{\text{F}}(x_t, t, c) - s_{\text{T}}(x_t, t, c)\right) \frac{\partial G_\theta}{\partial \theta}\right], \tag{13}$$

where timestep-dependent positive weighting factors are omitted for clarity. The score difference therefore provides a distribution-level direction for aligning the student with the teacher without explicitly evaluating probability densities.

Teacher-Aligned Score Estimation We sample a nominal timestep τ and map it to the corresponding scheduler noise level:

$$\sigma_t = w(\tau), \tag{14}$$

where $w(\cdot)$ denotes the scheduler's timestep mapping. The mapping is applied by the Flow Matching scheduler and is not expanded here into a separate score-sampling schedule.

Given the student output x_g, we construct a noisy sample using the flow-matching forward process:

$$x_t = (1 - \sigma_t)x_g + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{15}$$

The same noisy sample is evaluated by the frozen teacher and the online fake-score model:

$$\hat{x}_0^{\text{T}} = f_{\text{T}}(x_t, t, c), \qquad \hat{x}_0^{\text{F}} = f_{\text{F}}(x_t, t, c), \tag{16}$$

where the original Wan flow predictions are converted to clean-latent predictions before computing the distribution-matching signal. For the frozen teacher, the conditional and negative-prompt predictions are combined with classifier-free guidance; in our implementation the guidance scale is 6.0. The fake-score model uses a conditional prediction without an additional CFG branch.

For the corruption process in Eq. (15), an x_0 predictor induces the score

$$s(x_t, t) = \frac{(1 - \sigma_t)\hat{x}_0 - x_t}{\sigma_t^2}. \tag{17}$$

Therefore, for $0 < \sigma_t < 1$, the fake and teacher score difference becomes

$$s_{\text{F}}(x_t, t, c) - s_{\text{T}}(x_t, t, c) = \frac{1 - \sigma_t}{\sigma_t^2}\left(\hat{x}_0^{\text{F}} - \hat{x}_0^{\text{T}}\right). \tag{18}$$

Thus, for interior noise levels with $0 < \sigma_t < 1$, the difference between their clean-latent predictions provides the same gradient direction as the theoretical score difference up to a timestep-dependent positive scaling factor. At the endpoint $\sigma_t = 1$, this proportionality degenerates because the perturbed sample is pure noise; we retain the normalized clean-latent difference below as a practical endpoint surrogate.

Following the practical DMD formulation, we use the normalized gradient surrogate

$$g_{\text{DMD}} = \frac{\hat{x}_0^{\text{F}} - \hat{x}_0^{\text{T}}}{\text{mean}\left|x_g - \hat{x}_0^{\text{T}}\right|}, \tag{19}$$

which reduces variation in gradient magnitude across noise levels. We inject g_{DMD} into G_θ through a stop-gradient surrogate objective such that the resulting gradient with respect to x_g is proportional to g_{DMD}. This aligns the one-step student with the distribution represented by the four-step teacher without requiring the student to reproduce its sampling trajectory.

Online Fake-Score Learning As the student distribution evolves during training, the fake-score model f_{F} is updated online to track the current generator distribution. Given a student sample x_g, we independently sample a score timestep, perturb x_g according to Eq. (15), and optimize

$$\mathcal{L}_{\text{fake}} = \mathbb{E}\left[\|f_{\mathrm{F}}(x_t, t, c) - \mathrm{sg}(x_g)\|_2^2\right], \tag{20}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Following DMD2, f_{F} is updated more frequently than G_θ so that it can track the evolving student distribution and provide a stable distribution-matching signal.

In our object-aware implementation, the generator additionally receives auxiliary object, gripper, and DINO supervision. These auxiliary losses are applied only to the generator and are not used to train the frozen teacher or the fake-score model. Real robot videos provide the observed first-frame condition and the object-aware supervision targets; the DMD configuration does not apply a full future-video latent reconstruction loss.

During training, the frozen four-step teacher provides the target score estimate, while f_{F} estimates the score of the current one-step student distribution. At inference time, both score models are discarded and only G_θ is retained, reducing four-step future prediction to a single network evaluation.

B.2 Other Model Details

Interaction-Centric Tokens. DIDO uses 29 interaction-centric tokens divided into four groups: 4 object tokens $\mathcal{T}_{\text{obj}}$, 4 interaction tokens $\mathcal{T}_{\text{int}}$, 4 gripper tokens $\mathcal{T}_{\text{grip}}$, and 17 alignment tokens $\mathcal{T}_{\text{align}}$. These tokens are inserted into the video transformer and evolve jointly with its latent representations.

For trajectory prediction, $[\mathcal{T}_{\text{obj}}; \mathcal{T}_{\text{int}}]$ and $[\mathcal{T}_{\text{grip}}; \mathcal{T}_{\text{int}}]$ predict future object and gripper boxes, respectively. Each branch uses a two-layer MLP with GELU activations and hidden width equal to the video-transformer width, predicting normalized bounding-box coordinates for $H_a = 32$ future steps. For visual alignment, the representations $\mathcal{T}^l_{\text{align}}$ at selected transformer layers are mapped by lightweight projections P_l to the DINOv3 feature space and supervised by Equation 5. DINOv3 is used only to construct training targets and is removed at inference.

Action Expert. The action expert has the same depth as the video model and one quarter of its hidden width. It is coupled to the video backbone through the mixture-of-transformers architecture described in Section 3.3. Its output projection matches the action dimension of each embodiment, which is 7 for LIBERO, 14 for dual-arm RoboTwin, and 14 for Galbot G1. The action expert predicts a chunk of $H_a = 32$ actions per forward pass. During inference, the flow field is integrated with 10 Euler steps, and the first 10 actions on LIBERO and 24 actions on RoboTwin are executed before replanning.

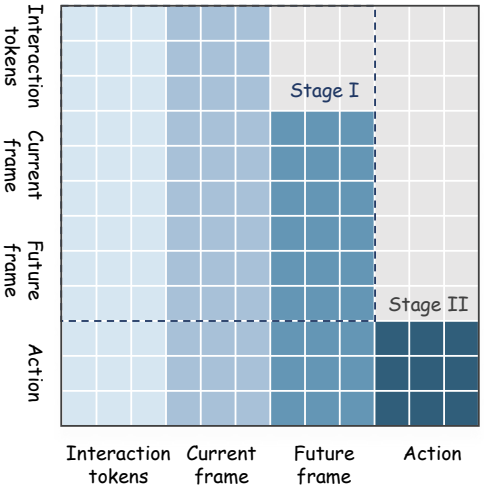


Figure 6: Shared attention between the world model and action expert.

Shared Attention. The video model and action expert are coupled through shared attention at corresponding transformer layers, as shown in Figure 6. In Stage I, interaction-centric, current-frame, and future-frame tokens attend within the world-model stream. In Stage II, action tokens are introduced and can attend to all world-model representations and action tokens. World-model tokens remain masked from the action stream, preserving the predictive pathway learned in Stage I. This asymmetric attention gives the action expert direct access to intermediate world-model representations without decoding future observations.

Table 4: Hyperparameter settings for teacher preparation and the two training stages of DIDO on LIBERO, RoboTwin, and real-world experiments.

Hyperparameter	LIBERO	RoboTwin	Real World
Shared settings			
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.01	0.01	0.01
Precision	bf16	bf16	bf16
Teacher preparation			
Learning rate	3×10^{-6}	3×10^{-6}	3×10^{-6}
Betas	$(0.0, 0.999)$	$(0.0, 0.999)$	$(0.0, 0.999)$
Schedule	500 warm-up, const	500 warm-up, const	100 warm-up, const
Training steps	3K	4K	1K
Global batch size	384	2304	384
λ_{obj} / λ_{grip} / λ_{giou}	0.05 / 0.05 / 0.5	0.05 / 0.05 / 0.5	0.05 / 0.05 / 0.5
λ_{align}	0.02	0.02	0.02
Stage I: one-step dynamics learning			
Generator LR	1×10^{-6}	1×10^{-6}	1×10^{-6}
Fake score LR	1×10^{-7}	1×10^{-7}	1×10^{-7}
Betas	$(0.0, 0.999)$	$(0.0, 0.999)$	$(0.0, 0.999)$
Schedule	500 warm-up, const	500 warm-up, const	100 warm-up, const
Generator update interval	5 iters	5 iters	5 iters
Training steps	3K	3K	1K
Global batch size	128	1536	128
λ_{obj} / λ_{grip} / λ_{giou}	0.03 / 0.03 / 0.5	0.045 / 0.045 / 0.5	0.03 / 0.03 / 0.5
λ_{align}	0.01	0.01	0.01
Stage II: joint policy learning			
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Betas	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Schedule	5% warm-up, cosine	5% warm-up, cosine	5% warm-up, cosine
Training steps	14,480	90K	6K
Global batch size	384	384	96
λ_{obj} / λ_{grip} / λ_{giou}	0.03 / 0.03 / 1.0	0.03 / 0.03 / 1.0	0.03 / 0.03 / 1.0
λ_{align}	0.02	0.02	0.02
λ_{video} / λ_{act}	0.5 / 1.0	0.5 / 1.0	0.5 / 1.0

C Additional Experimental Setup

This section summarizes the training data (Section C.1), optimization details and hyperparameters (Section C.2), and evaluation protocols (Section C.3).

C.1 Training Data

Beyond the pretrained Wan2.2 initialization, we use only the robot demonstrations provided by the corresponding evaluation settings, without additional robot video or cross-embodiment data. For LIBERO, both Stage I and Stage II are jointly trained on demonstrations from all four suites, and a single model is evaluated across all suites. RoboTwin follows the same setting, with both stages trained on the full training set and a single model evaluated across all tasks. For the real-world experiments, Stage I uses all 600 demonstrations across four tasks, with 150 demonstrations per task, while Stage II trains a separate policy for each task using its corresponding 150 demonstrations. LIBERO, RoboTwin, and the real-world experiments are trained independently and share no policy checkpoints.

C.2 Implementation Details

Training consists of three optimization phases: teacher preparation, Stage I, and Stage II. All phases use AdamW with weight decay 0.01 and bf16 mixed precision. Detailed hyperparameters are summarized in Table 4.

Teacher Preparation. Before the two-stage training of DIDO, we adapt the publicly released four-step Wan2.2-TI2V-5B checkpoint to robot manipulation videos. Interaction-centric supervision is enabled during this adaptation, with the corresponding warm-up schedules specified in Table 4. The resulting robot-domain four-step model serves as the teacher for Stage I

Stage I Optimization. We initialize the one-step generator and both score networks from the robot-domain teacher. The teacher score network is frozen; the fake-score network is updated every iteration, and the generator every fifth iteration. We use classifier-free guidance with scale 7 for the teacher score and no guidance for the fake-score network. The DMD objective and normalized gradient are described in Appendix B.1. Stage-I learning rates, update intervals, and training steps are listed in Table 4.

Stage II Optimization. In Stage II, the text encoder, VAE, and DINOv3 encoder remain frozen, while the video-transformer backbone, interaction-centric tokens, their prediction and projection heads, and action expert are jointly optimized. The DMD objective is removed and replaced by the video-generation objective following Fast-WAM (Yuan et al., 2026), retaining future prediction as a training signal during policy learning.

Let z denote the ground-truth future video latents encoded by the frozen VAE, and let $\mathcal{C}_{\mathrm{vid}}$ denote the video-model conditioning. We follow the general flow matching objective, with $\tau = 0$ enforced as our one-step generator does not require intermediate timesteps. For $\epsilon_z \sim \mathcal{N}(0, I)$, we construct $z_\tau = (1-\tau)\epsilon_z + \tau z$ with target velocity $u_z = z - \epsilon_z$. The video-generation objective is

$$\mathcal{L}_{\mathrm{video}} = \mathbb{E}_{z,\tau,\epsilon_z}\left[\left\|v_\phi\left(z_\tau, \tau \mid \mathcal{C}_{\mathrm{vid}}\right) - u_z\right\|_2^2\right], \tag{21}$$

where v_ϕ denotes the video model. Video generation, trajectory supervision, visual alignment, and action prediction are jointly optimized during Stage II.

Loss Weights and Hyperparameters. For each predicted bounding box, the smooth-L_1 term uses $\beta = 1.0$, while the GIoU term is weighted by λ_{giou}. The resulting object and gripper box losses are further scaled by λ_{obj} and λ_{grip}, respectively, to form $\mathcal{L}_{\mathrm{inter}}$. The multi-layer visual alignment loss $\mathcal{L}_{\mathrm{align}}$ is weighted by λ_{align}. In Stage II, the video-generation loss $\mathcal{L}_{\mathrm{video}}$ and action loss $\mathcal{L}_{\mathrm{act}}$ are additionally weighted by λ_{video} and λ_{act}, respectively. All loss weights and optimization hyperparameters are summarized in Table 4.

Compute. All experiments are trained on NVIDIA H100 GPUs with the global batch sizes reported in Table 4. For the simulation benchmarks, LIBERO requires approximately 15 hours for teacher preparation, 5 hours for Stage I, and 7 hours for Stage II. RoboTwin requires approximately 25 hours, 12 hours, and 100 hours for the three phases, respectively, primarily due to its larger demonstration set and optimization budget.

C.3 Evaluation Protocol

Simulation Evaluation. LIBERO, RoboTwin, and LIBERO-Plus follow their standard evaluation protocols. Performance is reported as task success rate averaged over evaluation episodes.

Real-World Evaluation. As shown in Figure 7, we evaluate DIDO on four real-world manipulation tasks: (i) Corn on Plate, placing the corn onto the plate; (ii) Bowl on Plate and Pepper in Bowl, placing the bowl on the plate and then putting the pepper in the bowl; (iii) Wipe Stain and Store Cloth, wiping the stain with a cloth and then putting the cloth in the basket; and (iv) Open Drawer and Place Duck, opening the upper drawer and putting the little duck inside. The first is a single-goal task evaluated with a three-stage rubric, while the remaining three are long-horizon tasks evaluated with four-stage rubrics.

Binary success does not distinguish between failures occurring at different stages of a task (Ji et al., 2026). We therefore use a staged progress score, where each rollout receives the score associated with the furthest completion stage reached. Each task is scored on a 10-point scale: the single-goal task uses stages worth 3, 6, and 10 points, while the long-horizon tasks use stages worth 2, 5, 7, and 10 points. Figure 7 illustrates the target state of each stage. We conduct 10 rollouts per task under

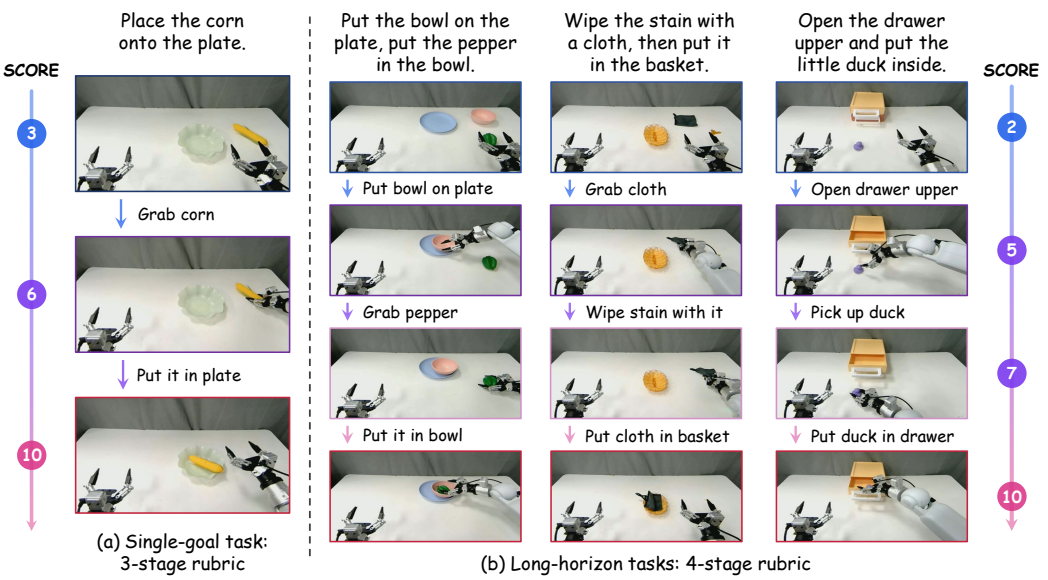


Figure 7: Staged scoring for real-world tasks.

each evaluation protocol, resulting in 80 rollouts in total, and report the mean score as a percentage of the maximum possible score.

Real-World Baseline Training. Since public results are unavailable for our real-robot platform, all real-world baselines are trained using the same demonstrations, observation and action spaces, and training budget as DIDO. Each baseline is trained separately for each task using its released implementation and recommended hyperparameters. $\pi_{0.5}$ (Physical Intelligence et al., 2025) is initialized from its official pretrained weights, while Fast-WAM (Yuan et al., 2026) is trained on our robot demonstrations. This ensures a consistent data and evaluation setting across methods.

D Supplementary Experiments

This section provides additional task-level results and analyses supporting the main experiments.

D.1 Per-Task RoboTwin Results

Table 5 reports per-task success rates on RoboTwin under both clean and domain-randomized settings. We compare the full DIDO model with a variant that removes the video-generation objective during Stage II while retaining the action and interaction-centric objectives. Overall, co-video supervision improves the average success rate from 89.56% to 92.00% in the clean setting and from 89.04% to 91.96% under domain randomization.

The gains are distributed unevenly across individual tasks, which is expected under multi-task joint training, where improvements on some tasks may be accompanied by small degradations on others. Several tasks exhibit particularly large gains, including `handover_block`, `open_microwave`, `put_object_cabinet`, and `pick_dual_bottles`. Despite occasional task-level regressions, the consistent improvement in overall performance indicates that retaining future-video supervision provides an effective training signal for the multi-task policy.

D.2 Analysis of Interaction-Centric Tokens

We further investigate whether DIDO's interaction-centric tokens are effectively utilized by the action expert. We refer to all tokens accessible to the action expert as **cache tokens** and group them by semantic source into *Object*, *Interaction*, *Gripper*, *Alignment*, *Current Visual*, and *Future Visual*. The first four groups constitute the interaction-centric tokens. For each cache token i, we quantify its contribution to action generation using the token-level action V-attribution score a_i.

Table 5: Per-task success rates (%) on RoboTwin. We compare DIDO without co-video supervision in Stage II with the full model under clean and domain-randomized settings. Each result is averaged over 50 evaluation episodes, with the better result shown in **bold**.

	DIDO (No co-video)		DIDO	
Task	Clean	Rand.	Clean	Rand.
adjust_bottle	100	98	100	100
beat_block_hammer	98	98	100	96
blocks_ranking_rgb	98	98	98	98
blocks_ranking_size	78	78	78	78
click_alarmclock	100	100	98	100
click_bell	100	100	98	100
dump_bin_bigbin	96	98	90	90
grab_roller	100	100	100	98
handover_block	82	74	96	90
handover_mic	98	98	100	98
lift_pot	100	100	100	100
move_can_pot	100	96	78	98
move_playingcard_away	100	100	100	100
move_stapler_pad	70	64	76	80
hanging_mug	48	30	40	46
open_laptop	98	100	98	98
open_microwave	54	50	68	64
pick_diverse_bottles	66	80	86	78
pick_dual_bottles	86	86	98	94
place_a2b_left	96	92	98	100
place_a2b_right	94	98	98	98
place_bread_basket	90	88	94	94
place_bread_skillet	94	88	94	88
place_can_basket	82	62	80	54
place_cans_plasticbox	98	100	100	96
place_container_plate	100	100	96	100
place_dual_shoes	78	82	88	92
place_empty_cup	100	100	100	100
place_fan	90	96	100	100
place_burger_fries	94	98	94	98
place_mouse_pad	92	86	96	96
place_object_basket	86	92	94	92
place_object_scale	96	90	100	96
place_object_stand	96	96	92	98
place_phone_stand	94	92	94	98
move_pillbottle_pad	96	92	96	96
place_shoe	90	94	90	94
press_stapler	66	94	66	84
put_bottles_dustbin	92	84	98	92
put_object_cabinet	78	84	94	92
rotate_qrcode	86	84	96	94
scan_object	94	82	92	94
shake_bottle	100	100	100	100
shake_bottle_horizontally	100	98	100	100
stack_blocks_three	94	98	100	98
stack_blocks_two	100	98	100	100
stack_bowls_three	86	88	86	86
stack_bowls_two	96	96	100	92
stamp_seal	84	92	92	98
turn_switch	64	60	70	72
Average	89.56	89.04	**92.00**	**91.96**

Metrics. For each token group g, we first compute its share of total action attribution as

$$A_g = \frac{\sum_{i \in g} a_i}{\sum_j a_j}, \tag{22}$$

where a larger A_g indicates that the group contributes more strongly to action generation. Since groups contain different numbers of tokens, however, A_g alone cannot distinguish token importance

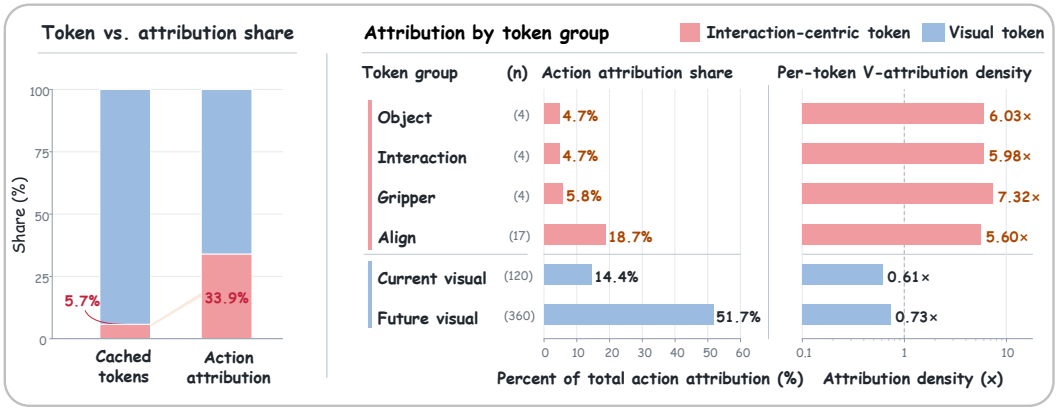


Figure 8: **Token contributions to action generation.** *Left*: Interaction-centric tokens occupy only 5.74% of the cache tokens for action generation but account for 33.9% of total action attribution. *Right*: Attribution is further decomposed across token groups. After normalization by token frequency, all interaction-centric groups exhibit substantially higher per-token attribution than the $1\times$ baseline, whereas current and future visual tokens remain below it.

from group size. We therefore normalize the attribution share by the token-frequency share and define the per-token attribution enrichment as

$$E_g = \frac{A_g}{N_g/N}, \tag{23}$$

where N_g is the number of tokens in group g and N is the total number of tokens. $E_g = 1$ indicates attribution proportional to token frequency, while $E_g > 1$ indicates that tokens in group g contribute more to action generation per token than expected from their frequency. We report means and 95% confidence intervals using task-level bootstrap with 10K resamples over 50 clips and 31 tasks.

Results. Figure 8 shows that action attribution is highly concentrated in the interaction-centric tokens. Although they constitute only 5.74% of the cache, they account for 33.9% of total action attribution, corresponding to approximately $5.9\times$ greater attribution per token than expected from their frequency. This indicates that the action expert places disproportionate emphasis on these compact interaction representations.

This pattern is consistent across the individual interaction-centric groups. Object, Interaction, Gripper, and Alignment tokens account for 4.7%, 4.7%, 5.8%, and 18.7% of total action attribution, respectively, with per-token enrichments of $6.03\times$, $5.98\times$, $7.32\times$, and $5.60\times$. In contrast, Current Visual and Future Visual tokens achieve enrichments of only $0.61\times$ and $0.73\times$, despite comprising most of the cache. These results suggest that the interaction-centric tokens concentrate action-relevant information into a small token budget, complementing the broader scene and future context provided by the visual tokens.

E LIMITATIONS

DIDO has several limitations. First, its interaction-centric reasoning relies on bounding-box supervision for the gripper and target object. Although effective in our settings, such coarse object-centric supervision may be insufficient for multi-object interaction, deformable objects, or fine-grained contact geometry. The required annotations are generated automatically, but detection and tracking failures under severe occlusion or ambiguous object references can still introduce noisy supervision.

Second, DIDO introduces additional training complexity and computational cost. Training involves three separate optimization phases, including robot-domain teacher preparation, one-step distillation, and joint policy learning, making the overall pipeline more involved than directly training an action policy. Finally, the single-step world model is distilled from a robot-adapted four-step teacher, so its predictive capability remains dependent on the quality and coverage of the teacher. Extending DIDO to richer interaction representations, simpler training pipelines, and broader embodiments and environments remains an important direction.