WORLD TOKENS: ENHANCING EMBODIED POLICIES WITH TRAINING-TIME WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models are efficient at inference but do not explicitly model how physical scenes evolve as a task unfolds, whereas world-action models (WAMs) capture this evolution with pretrained video world models at the cost of keeping future generation or a large video backbone in the control loop. We introduce World Tokens, an embodied policy architecture that uses world modeling only during training, through an exclusive predictive interface. A World Adapter transforms VLM features into a fixed set of world tokens, which condition a jointly fine-tuned future-video denoiser and serve as the action expert's sole visual-language context. The denoiser receives only a coarse layout cue of the current scene, so neither it nor the action expert can bypass the world tokens, and future-video supervision directly shapes the representation used for control. At deployment, the world-model branch is removed, leaving only the VLM, World Adapter, and action expert. We evaluate World Tokens on LIBERO, the zero-shot perturbations of LIBERO-Plus, SIMPLER, and real-world manipulation with a Galaxea R1 Pro robot. With a 2B backbone and no embodied action pretraining, it performs strongly across these benchmarks and consistently improves over a matched action-only baseline, while running at VLA-level latency.

1 INTRODUCTION

Vision-language-action (VLA) policies pair a pretrained vision-language model (VLM) with an action expert that maps visual observations and language instructions to continuous robot commands (Kim et al., 2025b; Black et al., 2025b; NVIDIA, 2025a; Kim et al., 2025a; Black et al., 2025a). The VLM provides semantic recognition and instruction grounding, while the action expert generates actions without any video generation in the loop. However, image–text pretraining provides little direct supervision for action-relevant temporal transitions, such as contact, occlusion, and object displacement. Because appropriate action selection often depends on how the scene is likely to evolve rather than on current appearance alone, a VLA must learn these transitions primarily from robot demonstrations, which are typically much more limited than its visual-language pretraining data.

Pretrained video world models offer complementary supervision by predicting how scenes change over time. World-action models (WAMs) incorporate this capability by jointly generating videos and actions, conditioning actions on predicted futures, or extracting features from a video-denoising backbone (Bi et al., 2026; Ye et al., 2026b; Kim et al., 2026; Ma et al., 2026; Li et al., 2026). These architectures have shown promising long-horizon and out-of-distribution manipulation performance, but iterative video denoising or a retained multi-billion-parameter backbone can substantially increase control latency.

Recent evidence suggests that the representational benefit of predictive training can be separated from the cost of online video inference. Fast-WAM shows that video co-training can contribute more than explicit future synthesis alone, yet its deployed controller still retains a video DiT together with a large companion action DiT and language encoder (Yuan et al., 2026). In parallel, lighter predictive objectives can improve VLAs without retaining a video backbone at deployment (Fang et al., 2025; Sun et al., 2026). Training-time prediction is thus an established idea, but how it should reach the controller is less settled: predictive supervision typically acts only through a parallel branch, while the action expert still conditions on the full VLM context. The controller can then simply ignore the subspace shaped by world modeling, and prediction and action can drift apart in representation space.

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.

We therefore ask: *how can world modeling better enhance the representation used for embodied control while the world model itself is removed completely after training?*

We address this question with a World Adapter (Figure 1), an intermediate module that bridges task understanding, future prediction, and action generation (i.e., the VLM, a video world model, and an action DiT). During training, VLM features are expressed through the World Adapter as a fixed set of world tokens, which jointly condition a fine-tuned future-video denoiser and a flow-matching action expert. The action expert has no parallel access to the full VLM sequence; all visual and linguistic context must pass through these tokens. This exclusive routing prevents the controller from bypassing the future-supervised representation. To keep the video branch dependent on these tokens, we give it only a simple layout cue of the current scene—an edge map—rather than the full RGB frame. With the full frame, the video model can extrapolate the future largely from appearance and need not consult the tokens. Given only the layout, the world tokens must supply the content from which the future is imagined. They thus form an exclusive predictive interface that neither the policy nor the world model can bypass. Because the video world model only supervises the world tokens, it is removed at deployment: inference runs through the VLM, World Adapter, and action expert, whose representation has already been shaped by world modeling, at a latency close to that of a VLA.

Our contributions are summarized as follows:

- **An exclusive predictive interface.** We propose World Tokens, in which a World Adapter maps the VLM sequence to 256 world tokens that are the action expert's sole visual-language context and condition a jointly fine-tuned video world model given only a coarse layout cue of the scene. Future-video supervision therefore acts on exactly the representation the policy consumes, and the entire video branch is removed at deployment, leaving a policy that adds only a $\sim$0.5B adapter to a standard VLA.
- **An analysis of when predictive supervision helps control.** Controlled ablations show that video supervision helps control when neither side can bypass the learned world tokens: letting the action expert also read the full VLM sequence, giving the video branch the RGB first frame, or replacing the learned queries with a feed-forward extractor each scores below a variant trained without any video supervision. Attention analysis further shows that supervised world tokens attend more sharply and track the current subtask.
- **Strong performance at VLA-level latency.** With a 2B VLM backbone and no embodied action pretraining, World Tokens improves over the matched action-only baseline on LIBERO (98.2% vs. 96.9%), zero-shot LIBERO-Plus (75.3% vs. 69.1%), SIMPLER-WidowX (71.5% vs. 66.7%), and a real robot (76.0% vs. 59.4%). It runs at 62 ms per action chunk, within the 43–95 ms range of the compared VLAs and at least 2.9$\times$ faster than the compared WAMs.

2 Related Work

Vision-language-action policies. Large-scale robot transformers established the benefits of scaling policy capacity and data diversity across tasks and embodiments (Brohan et al., 2023b; Ghosh et al., 2024), while RT-2 transferred web-scale visual-language knowledge into robotic control (Brohan et al., 2023a). Complementary action-generation approaches model multimodal behavior through latent action tokens or diffusion (Lee et al., 2024; Chi et al., 2023; Liu et al., 2025). Modern VLA models adapt multimodal foundation models to robot control through autoregressive action decoding (Kim et al., 2025b;a) or flow-matching and diffusion action experts conditioned on VLM features (Black et al., 2025b; NVIDIA, 2025a; Black et al., 2025a). We retain this dual-system organization but, rather than conditioning the action expert on the full VLM sequence, route control through a fixed-size token interface jointly constrained by action prediction and future-video denoising.

WAMs. Earlier work connected video modeling to control by jointly predicting future images and actions, extracting predictive video representations, learning task-centric latent actions from heterogeneous videos, or generating synthetic robot trajectories (Wu et al., 2024; Hu et al., 2025; Bu et al., 2025; Jang et al., 2025). WAMs more tightly couple video prediction and control through a shared generative backbone. Motus and DreamZero jointly model video and action with large video DiTs (Bi et al., 2026; Ye et al., 2026b); other systems adapt video foundation models for control and planning (Kim et al., 2026; Ma et al., 2026) or execute causal future rollouts asynchronously (Li et al.,

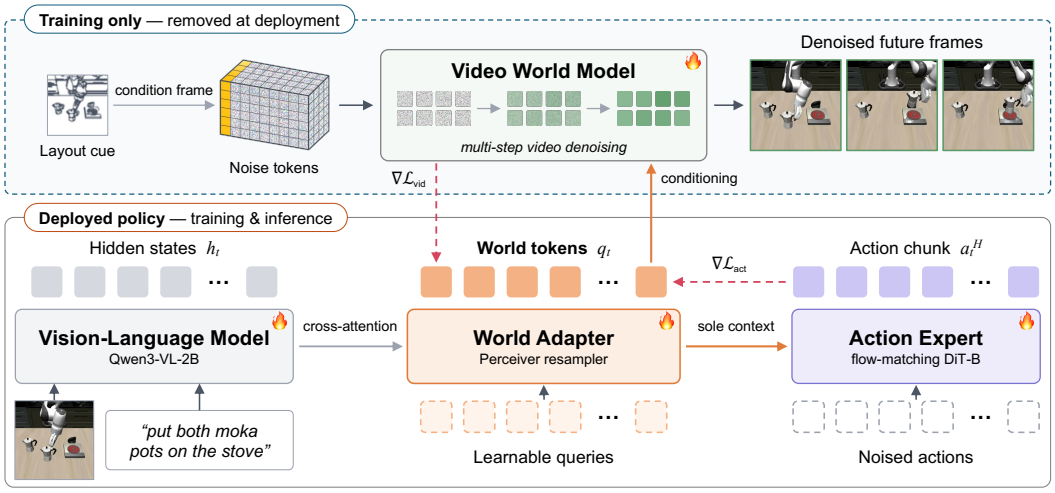

Figure 1: **Overview of World Tokens.** *Deployed policy (bottom):* the VLM encodes the observation o_t and instruction ℓ into hidden states h_t; the $\sim$0.5B World Adapter cross-attends from K=256 learnable queries to h_t to produce the world tokens q_t, which are the action expert's sole visual-language context for denoising the action chunk. *Training only (top):* q_t also conditions a pretrained video world model that iteratively denoises the future-clip tokens, with a VAE-encoded edge map of o_t as a coarse layout cue in the condition frame. All modules are trained jointly with $\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_w \mathcal{L}_{\text{vid}}$, and the gradients of both losses (dashed) flow back into q_t. At deployment the top panel is removed, so no video is generated in the control loop. Frames are from a LIBERO-Long demonstration; the denoised frames are illustrative.

2026). These methods differ in whether they decode future RGB frames, expose denoising features to the action module, or plan over predicted futures, but all retain substantial video-model computation at deployment. Fast-WAM removes explicit future synthesis yet still evaluates a pretrained video DiT as a world encoder (Yuan et al., 2026).

Predictive world modeling for VLA policies. Predictive VLA methods vary in both the target they model and the computation they retain at inference. WorldVLA and UVA jointly generate images and actions (Cen et al., 2025; Li et al., 2025a), while DreamVLA predicts dynamic, spatial, and semantic world knowledge before action generation (Zhang et al., 2025). Other methods use lighter training signals, including future-latent alignment (Zheng et al., 2025), motion-image diffusion (Fang et al., 2025), and latent-state prediction of future images (Sun et al., 2026). Related token formulations either roll compact visual states forward in time or encode discrete world changes (Tang et al., 2026; Zhu et al., 2026). World Tokens differs in two respects: the supervised tokens are the sole conditioning input to the policy and represent the current context rather than a predicted future, and predictive supervision is sourced from a pretrained video world model rather than a lightweight proxy.

3 Method

3.1 Problem Formulation

We consider vision-language-conditioned robotic manipulation. At control step t, the robot observes RGB images from V views, $o_t = (o_t^1, \ldots, o_t^V)$, and receives a language instruction ℓ. The policy predicts an H-step action chunk $a_t^H \equiv (a_t, \ldots, a_{t+H-1})$. During training, each demonstration additionally provides a T-frame future RGB clip from the primary view, $v_t \equiv (o_{t+1}^1, \ldots, o_{t+T}^1)$. The clip is used only as a prediction target: it is never provided to the action policy and is unavailable at evaluation. Our goal is to use v_t to shape the control representation while preserving the standard deployed mapping from (o_t, ℓ) to a_t^H.

3.2 Architecture Overview

As shown in Figure 1, the deployed policy (lower path) contains a VLM, a World Adapter, and an action expert, and a video world model (upper path) is added only for training. A fixed set of world tokens connects the two paths: they are the action expert's sole visual-language context and, after projection, condition the video model as it denoises the demonstrated future clip, given only an edge map of the current frame (Canny, 1986) as a layout cue.

3.3 World Adapter

Observation and language encoding. A pretrained VLM f_ϕ jointly encodes all image views and the instruction into a variable-length sequence $h_t \in \mathbb{R}^{N_t \times d_h}$ of patch and language tokens. We insert a World Adapter b_ψ that maps this sequence into a fixed-size representation:

$$h_t = f_\phi(o_t, \ell), \qquad q_t = b_\psi(h_t) \in \mathbb{R}^{K \times d}. \tag{1}$$

We term the K output vectors of q_t the world tokens; the name reflects their predictive training objective rather than any predefined slot semantics.

Query-based token extraction. The adapter follows the Perceiver resampler design (Jaegle et al., 2021; Alayrac et al., 2022): it maintains K learned queries $Q^{(0)} \in \mathbb{R}^{K \times d}$, and each of its L blocks cross-attends from the current queries to h_t, exchanges information among the queries through self-attention, and applies a feed-forward update. Denoting the j-th block by $B_\psi^{(j)}$, we write

$$Q_t^{(0)} = Q^{(0)}, \qquad Q_t^{(j+1)} = B_\psi^{(j)}(Q_t^{(j)}, h_t), \quad j = 0, \ldots, L-1, \qquad b_\psi(h_t) = Q_t^{(L)}. \tag{2}$$

The action and video objectives jointly determine what information the queries retain. With $K{=}256$ queries and $L{=}12$ blocks at width $d{=}2048$, the adapter has roughly 0.5B parameters (Appendix A).

Exclusive action pathway. The action expert g_θ is a DiT flow-matching model (Peebles & Xie, 2023; Lipman et al., 2023) that generates the action chunk a_t^H conditioned solely on the world tokens:

$$a_t^H \sim g_\theta(\cdot \mid q_t). \tag{3}$$

3.4 Training-Time World Modeling

Video denoiser. The training-only video branch is a pretrained video diffusion model, denoted by $\mathcal{W}_\omega$. A lightweight projection p_ρ adapts the world tokens to its cross-attention dimension:

$$q_t^{\mathrm{wm}} = p_\rho(q_t).$$

A frozen VAE encoder E maps the demonstrated future clip to the latent target $z_t = E(v_t)$. During training, $\mathcal{W}_\omega$ denoises noisy versions of z_t, using q_t^{wm} as cross-attention context and the layout cue defined below as its first-frame condition.

Layout cue. The video model conditions on a first frame, and what that frame contains determines how much the video branch has to rely on the world tokens. Given the full RGB frame, the denoiser can largely extrapolate the future from appearance persistence and has little need for the tokens; given no first frame at all, it must also generate the static scene from the tokens, and in our preliminary experiments it converged poorly. We therefore give it only a simple layout cue—a Canny edge map of the primary view, encoded by the same frozen VAE encoder:

$$c_t = E\big(\mathrm{Canny}(o_t^1)\big).$$

The edge map indicates where things are but carries little color or texture. The video branch thus imagines the future on top of this layout, and the content it needs, namely which objects matter for the instruction and how the scene will change, has to come from the world tokens.

Table 1: LIBERO success rate (%) and latency per action chunk (ms). Each suite has 10 tasks; our results average three environment seeds (1,500 episodes per suite), and Avg. is the unweighted suite mean. Emb. PT denotes embodied action pretraining; Qwen-DiT is our matched baseline. Latency is measured by us end-to-end on one H100, including preprocessing (Appendix B). Bold marks the best value per column (lowest latency), separately among VLAs and among WAMs.

Vision-language-action policies (VLAs)							
Method	Emb. PT	Latency (ms)	Spatial	Object	Goal	Long	Avg.
π_0 (Black et al., 2025b)	✓	**43**	96.8	98.8	95.8	85.2	94.2
$\pi_{0.5}$ (Black et al., 2025a)	✓	47	98.8	98.2	98.0	92.4	96.9
OpenVLA-OFT (Kim et al., 2025a)	✓	76	97.6	98.4	97.9	94.5	97.1
VLA-JEPA (Sun et al., 2026)	✓	95	96.2	99.6	97.2	95.8	97.2
StarVLA (Ye et al., 2026a)	✗	88	98.7	**99.7**	**98.6**	94.2	97.8
Qwen-DiT	✗	52	97.6	98.8	98.2	92.8	96.9
World Tokens (ours)	✗	62	**99.6**	98.8	97.4	**97.0**	**98.2**
WAMs							
WorldVLA (Cen et al., 2025)	✗	570	87.6	96.2	83.4	60.0	81.8
Fast-WAM (Yuan et al., 2026)	✗	**182**	**98.2**	**100.0**	97.0	95.2	97.6
Motus (Bi et al., 2026)	✓	2,033	96.8	99.8	96.6	**97.6**	97.7
Cosmos Policy (Kim et al., 2026)	✗	454	98.1	**100.0**	**98.2**	**97.6**	**98.5**

3.5 Joint Training and Deployment

Training objective. Both branches are trained with a conditional denoising objective. For a target y, we sample base noise $\epsilon \sim \mathcal{N}(0, I)$ and a noise level τ, and write the noised input as y_τ and the regression target as $u_\tau(y, \epsilon)$. For a conditional prediction network F, we define

$$\mathcal{L}_{\text{den}}(F; y \mid c) = \mathbb{E}_{\tau,\epsilon}\left[\|F(y_\tau, \tau \mid c) - u_\tau(y, \epsilon)\|_2^2\right]. \tag{4}$$

The two branches differ only in parameterization: the action expert uses flow matching on a linear interpolation path, with $\tau \sim \mathcal{U}[0, 1]$, $y_\tau = (1-\tau)\epsilon + \tau y$, and $u_\tau(y, \epsilon) = y - \epsilon$, while the video branch retains the native diffusion formulation of the underlying video model (NVIDIA, 2025b). We apply this objective to the demonstrated action chunk and the future-video latent:

$$\mathcal{L}_{\text{act}} = \mathcal{L}_{\text{den}}\big(g_\theta; a_t^H \mid q_t\big), \quad \mathcal{L}_{\text{vid}} = \mathcal{L}_{\text{den}}(\mathcal{W}_\omega; z_t \mid q_t^{\text{wm}}, c_t), \quad \mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_w \mathcal{L}_{\text{vid}}, \quad \lambda_w = 0.5. \tag{5}$$

Gradient paths. The action loss updates f_ϕ, b_ψ, and g_θ, forcing the world tokens to retain information needed for control; the video loss updates f_ϕ, b_ψ, p_ρ, and $\mathcal{W}_\omega$, encouraging the same tokens to explain the demonstrated scene evolution. Because both gradient paths meet at q_t, predictive supervision directly shapes the representation consumed by the policy rather than a detachable auxiliary feature.

Initialization. The VLM and the video denoiser are initialized from pretrained checkpoints and fine-tuned jointly. The World Adapter, world-model projection, and action expert are trained from scratch. We fine-tune the video denoiser because robot demonstrations differ from its pretraining data in viewpoint, embodiment, and contact dynamics; a frozen denoiser could provide poorly matched supervision. The tokenizer E remains frozen to keep the target latent space stationary.

Deployment. At deployment, the video branch (E, p_ρ, and $\mathcal{W}_\omega$) is removed, leaving the VLM f_ϕ, the World Adapter b_ψ, and the action expert g_θ; neither v_t nor the edge map is needed, and each action chunk is generated with four flow-integration steps. World modeling therefore shapes the control representation during training without adding video-model computation at deployment.

Table 2: Controlled LIBERO ablations under the protocol of Table 1. Each factor column marks one design choice of World Tokens: future-video supervision from the jointly fine-tuned video denoiser (*Video sup.*), exclusive action conditioning on the world tokens q_t rather than the full VLM sequence (*Excl. routing*), query-based extraction rather than a parameter-matched FFN (*Learned query*), and Canny first-frame conditioning rather than native RGB (*Edge cue*). Each variant changes one factor of the full method. Avg. is the mean over all four suites; Spatial, Object, and Goal are omitted because every variant reaches at least 97.2% on them. Bold marks the best value in each column; the World Tokens row reproduces Table 1.

	Design factor				Success (%)	
Variant	Video sup.	Excl. routing	Learned query	Edge cue	Long	Avg.
w/o wm	✗	✓	✓	–	95.0	97.4
w/ VLM bypass	✓	✗	✓	✓	94.1	97.2
FFN adapter	✓	✓	✗	✓	93.4	97.0
RGB cue	✓	✓	✓	✗	91.5	96.2
World Tokens (ours)	✓	✓	✓	✓	**97.0**	**98.2**

4 EXPERIMENTS

Our evaluation addresses three questions: Does training-time world modeling help once the video branch is removed at deployment? Which components drive the improvement? Does the policy remain effective beyond LIBERO?

4.1 EXPERIMENTAL SETUP

Implementation. All experiments use the same configuration: Qwen3-VL-2B-Instruct (Qwen Team, 2025) as the VLM, a $\sim$0.5B World Adapter with $K{=}256$ learned queries, a DiT-B action expert, and Cosmos Predict2.5-2B (NVIDIA, 2025b) as the training-only video denoiser, with eight-step action chunks and eight-frame future targets. Optimization and training-scale details are given in Appendix A.

Benchmarks. LIBERO comprises four language-conditioned manipulation suites—Spatial, Object, Goal, and Long—with 10 tasks and 500 demonstrations each (Liu et al., 2023); we train a single policy on all 40 tasks and evaluate it with three environment seeds and 50 episodes per task and seed, i.e., 1,500 episodes per suite and 6,000 in total. To test robustness beyond the training distribution, we evaluate the same policy zero-shot on LIBERO-Plus (Fei et al., 2025), which perturbs the four suites along seven dimensions—camera viewpoint, robot initial state, language instruction, lighting, background texture, sensor noise, and object layout—in 10,030 test tasks. For SIMPLER, we train on the BridgeV2 (Walke et al., 2023) and Fractal (Brohan et al., 2023b) real-robot datasets and evaluate in the visually matched simulator (Li et al., 2025b) under the official WidowX protocol; Google Robot results are in Appendix F.

R1 Pro real-world setup. We control the right arm of a Galaxea R1 Pro with an 8-dimensional joint-and-gripper action from two views: a cropped head-mounted fisheye view, whose crop removes large background regions but keeps the arm and workspace, and a wrist view. The task is language-conditioned pick-and-place: the instruction names one fruit (banana, mango, strawberry, or lemon), and a trial succeeds if the robot picks that fruit and places it in a basket. A single policy is trained on 400 demonstrations, 100 per task. We run 24 trials per task, for 96 evaluation trials per policy, randomizing object and basket positions and the order in which policies are evaluated.

Baselines and protocol. Our matched VLA baseline is Qwen-DiT: the same Qwen3-VL backbone and DiT action expert as World Tokens, but with the action expert reading h_t directly and with neither a World Adapter nor a training-time video branch. All models we train ourselves—World Tokens, Qwen-DiT, and every ablation variant—share the same training data, optimizer and schedule, number of steps, and batch size (Appendix A), so each comparison among them varies only the architecture. On R1 Pro, both $\pi_{0.5}$ variants (Black et al., 2025a) are fine-tuned on the same demonstrations; with

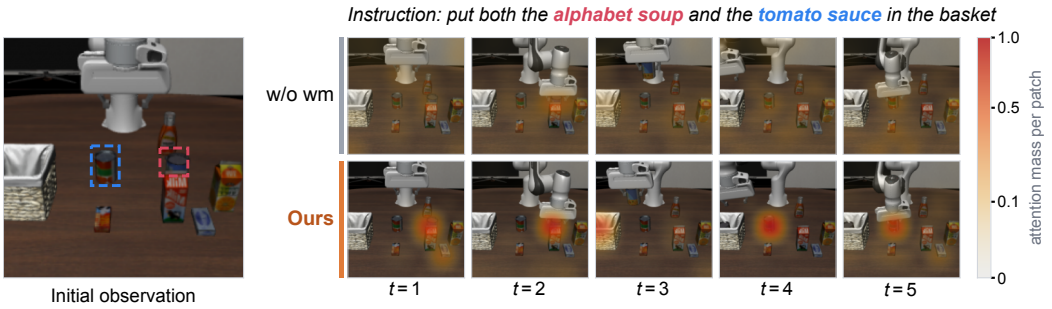

Figure 2: **Task-phase-dependent attention from world tokens.** Main-view World Adapter cross-attention on a LIBERO-Long episode. Each panel is one head/world-token map of the final adapter block, chosen by the same rule for both models: the map with the most attention on the current subtask object (the can being approached, or the basket while a can is carried). Shared square-root color scale; the selected maps place 60–100% of their mass on that object for World Tokens and at most 25% for *w/o wm* (Appendix D).

and without embodied pretraining, they contrast large-scale pretraining followed by fine-tuning on limited demonstrations against training on those demonstrations alone. Success rates of all other entries in Tables 1 and 3 are quoted from their source papers under the official protocols. Success rate is the primary metric; significance tests are given in Appendix G.

4.2 LIBERO

World Tokens averages 98.2% across the four LIBERO suites (Table 1), against 96.9% for the matched Qwen-DiT, with the gap concentrated on Long (97.0% vs. 92.8%). Without embodied pretraining, World Tokens is on par with the 4B StarVLA (97.8%) using a 2B backbone. The only WAM with a higher average, Cosmos Policy (98.5%), takes 454 ms per chunk versus 62 ms for World Tokens, and every WAM is at least 2.9× slower. Among published VLAs, only π_0 and $\pi_{0.5}$ are faster (43 and 47 ms), and World Tokens exceeds their suite averages by 4.0 and 1.3 points.

The Long suite is more informative, since its multi-stage tasks give sequential errors room to accumulate. World Tokens attains 97.0%, above the other VLAs in Table 1 and 0.6 points below Motus and Cosmos Policy (97.6%), which both keep a video backbone at inference.

4.3 Ablation Studies

Table 2 varies four design factors of World Tokens under the LIBERO protocol, one at a time. *w/o wm* removes the video branch and trains with the action loss only. The other variants keep video supervision but alter the token interface: *w/ VLM bypass* lets the action expert read around the world tokens, *RGB cue* lets the video branch predict without them, and *FFN adapter* replaces the learned queries that form the interface.

Mechanistic ablations. Removing video supervision lowers Long from 97.0% to 95.0% (*w/o wm*; p=0.005, Appendix G). *w/ VLM bypass* (94.1%) and *FFN adapter* (93.4%) keep world-model supervision, yet both fall below *w/o wm*, which has none, so predictive supervision does not help unconditionally. When the action expert can bypass the supervised representation, or extraction lacks learned queries, the video objective competes with action prediction instead of shaping it. We conjecture that learned queries, trained from scratch under both objectives, form a shared interface for prediction and control, whereas an FFN only reshapes VLM features the action expert already uses.

Layout-cue comparison. With the native RGB first frame instead of the edge map, *RGB cue* falls from 97.0% to 91.5% on Long, below even Qwen-DiT (92.8%). This matches the shortcut that motivates the layout cue: given the full frame, the video model can predict the future without the world tokens, so its gradient no longer shapes them usefully for control. Because *RGB cue* deploys exactly the same network as World Tokens, the gap stems from the training signal alone.

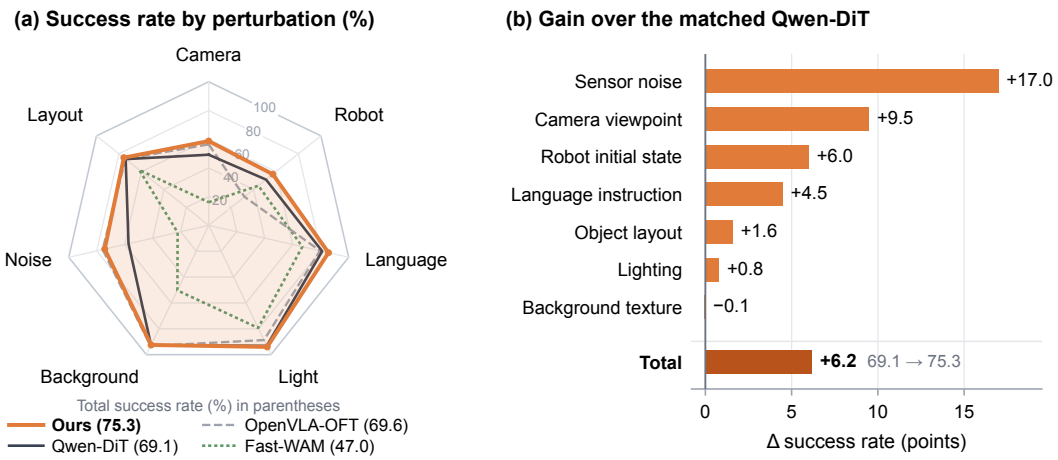

Figure 3: **Zero-shot robustness on LIBERO-Plus** (Fei et al., 2025). All policies are trained only on LIBERO and tested on the 10,030 perturbed tasks of LIBERO-Plus. **(a)** Success rate per perturbation dimension, with Total success in the legend; OpenVLA-OFT is quoted from Fei et al. (2025), the WAM Fast-WAM (Yuan et al., 2026) is evaluated by us, and Qwen-DiT is our matched baseline. **(b)** Gain of World Tokens over Qwen-DiT. All published baselines are in Table S1.

Attention analysis. Figure 2 examines how future-video supervision shapes the world tokens on a LIBERO-Long episode, comparing World Tokens with *w/o wm*, a variant trained with the action loss only. For World Tokens, the selected maps follow the task: they fall on the alphabet soup grasped first (t=1, 2), on the basket while it is carried (t=3), and on the tomato sauce next (t=4, 5). The selected map of *w/o wm* puts at most 25% of its mass on the same object. Over the 32 control steps of this episode, the mean entropy of the per-head, per-token main-view attention over the 8×8 patch grid is 3.29 bits for World Tokens versus 5.22 for *w/o wm* (uniform: 6), and is lower at every step, as on a perturbed LIBERO-Plus episode (Appendix D). Video supervision thus gives some world tokens, which form the action expert's only visual-language context, a focus that tracks the current subtask; *w/o wm* has no such tokens.

4.4 Zero-Shot Robustness on LIBERO-Plus

LIBERO success is near its ceiling, so we additionally ask whether the gains persist when the test scenes depart from the training distribution. Figure 3 evaluates the LIBERO policy of Section 4.1 on LIBERO-Plus without any further training. World Tokens succeeds on 75.3% of the 10,030 perturbed tasks, against 69.1% for the matched Qwen-DiT, a gap several times larger than the 1.3-point gap on the unperturbed suites (Table 1). It also compares favorably with the published policies evaluated under the same protocol, including the 7B OpenVLA-OFT (69.6%), which relies on embodied pretraining (Table S1).

The improvement over the matched baseline is largest for sensor noise (+17.0 points), camera viewpoint (+9.5), and robot initial state (+6.0), which change how the scene is observed or where the robot starts rather than what the scene contains. Language perturbations add +4.5, while lighting, background texture, and object layout change by at most 1.6 points. One reading is that, because the video branch sees only the layout and the action expert sees only the world tokens, the tokens come to encode the task-relevant state of the scene, such as which objects are involved and where the arm must go, rather than view-specific appearance that a policy conditioned directly on VLM features can more easily overfit; in a perturbed episode, the world tokens also remain focused on the task objects (Appendix D).

4.5 SIMPLER

The four SIMPLER-WidowX tasks (Table 3) put a spoon on a towel, a carrot on a plate, a green block on a yellow block, and an eggplant into a basket. World Tokens averages 71.5%, 4.8 points above the

Table 3: SIMPLER visual-matching success (%) on WidowX (Ye et al., 2026a; Sun et al., 2026; Qu et al., 2025). Qwen-DiT is our matched baseline. Bold marks the best value in each column.

Method	Spoon	Carrot	Block	Eggplant	Avg.
π_0 (Black et al., 2025b)	29.1	0.0	16.6	62.5	27.1
GR00T-N1.5 (NVIDIA, 2025a)	75.3	54.3	**57.0**	61.3	62.0
OpenVLA-OFT (Kim et al., 2025a)	34.2	30.0	30.0	72.5	41.7
SpatialVLA (Qu et al., 2025)	16.7	25.0	29.2	**100.0**	42.7
VLA-JEPA (Sun et al., 2026)	75.0	70.8	12.5	70.8	57.3
$\pi_{0.5}$ (Black et al., 2025a)	49.3	64.7	44.7	69.7	57.1
StarVLA (Ye et al., 2026a)	**79.7**	59.8	22.8	98.5	65.2
Qwen-DiT	75.0	66.7	29.2	95.8	66.7
World Tokens (ours)	74.0	**85.0**	32.0	95.0	**71.5**

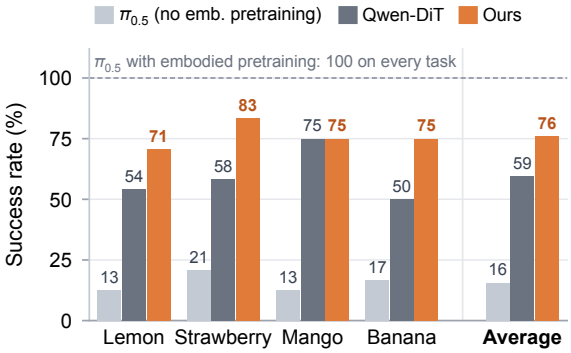

Figure 4: R1 Pro closed-loop success (%), 24 trials per task; Average pools all 96 trials. All policies are trained on the same demonstrations, and object and basket positions and the order in which policies are evaluated are randomized across trials. World Tokens improves on the matched Qwen-DiT for lemon, strawberry, and banana and ties it on mango. The dashed line is the embodied-pretrained $\pi_{0.5}$, shown as context. Rollouts are in Appendix H.

matched Qwen-DiT, mainly through carrot-on-plate (+18.3), and 6.3 points above StarVLA (65.2%). World Tokens also compares favorably on Google Robot (Appendix F).

4.6 Real-World Manipulation

Pooled over the four tasks, World Tokens succeeds on 76.0% of trials against 59.4% for Qwen-DiT (Figure 4; p=0.02, Appendix G). The gain comes from strawberry and banana (+25.0 points each) and lemon (+16.7); on mango, where Qwen-DiT is strongest, the two tie. Since each task has only 24 trials, we base the claim on the 96 pooled trials. The $\pi_{0.5}$ results are context rather than a matched comparison: trained only on the same demonstrations it reaches 15.6%, while its embodied-pretrained checkpoint solves every trial, as its pretraining already covers such near-in-distribution pick-and-place.

Qualitatively, Qwen-DiT tends to home in on a wrong object or lose the fruit to an insecure grasp, whereas World Tokens localizes the instructed fruit early and completes the placement (Appendix H).

5 Conclusion

We presented World Tokens, an embodied policy architecture that obtains the representational benefits of world modeling without its inference cost through an exclusive predictive interface. Its world tokens condition future-video denoising during training and are the action expert's only input; the video branch is discarded at deployment. Across LIBERO, the zero-shot perturbations of LIBERO-Plus, SIMPLER, and real-robot R1 Pro experiments, it consistently improves over a matched VLA baseline. Our analysis indicates that predictive supervision improves control most when it shapes the representation the policy actually consumes: exclusive routing keeps the action expert from bypassing the world tokens, and the layout cue keeps the video branch from doing so. The study has limitations: the world model adds considerable training-time cost, the edge-map layout cue is hand-designed, the real-robot evaluation covers four pick-and-place tasks with 24 trials each, and the attention analysis is descriptive. Future work will explore cheaper world-model supervision, learned layout cues, and large-scale joint training of the World Adapter across embodiments and tasks.

AI USE STATEMENT

We used generative AI tools to polish the writing and check grammar, to format the manuscript and references, to write plotting scripts for figures based on our results, and to assist with descriptive data analysis (recomputing reported averages, the attention statistics in Appendix D, and the significance tests in Appendix G). All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies robotic manipulation policies trained on public benchmark datasets (LIBERO, BridgeV2, Fractal) and on demonstrations of fruit pick-and-place collected on our own R1 Pro robot. No human subjects or personally identifiable data are involved. Real-robot experiments were conducted in a controlled laboratory setting. We are not aware of other ethical concerns specific to this work beyond the general dual-use considerations of more capable robot policies.

REPRODUCIBILITY STATEMENT

Section 3 and Section 4.1 specify the architecture, training objective, backbones, and evaluation protocols. Appendix A gives the optimizer, per-group learning rates, edge-map parameters, training scale (steps, batch size, GPUs), and World Adapter dimensions; Appendix B details the latency protocol and hardware; Appendix D describes how attention maps and entropies are computed. LIBERO, SIMPLER, BridgeV2, and Fractal are publicly available. Source code will be released upon publication.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bi_Motus_A_Unified_Latent_Action_World_Model_CVPR_2026_paper.html.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40, 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and

Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025b. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Marcin Choromanski, Tianli Ding, Danny Driess, Kumar Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montserrat Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023a. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H. Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023b. doi: 10.15607/RSS.2023.XIX.025. URL https://roboticsproceedings.org/rss19/p025.html.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025. URL https://arxiv.org/abs/2506.21539.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Yu Fang, Kanchana Ranasinghe, Le Xue, Honglu Zhou, Juntao Tan, Ran Xu, Shelby Heinecke, Caiming Xiong, Silvio Savarese, Daniel Szafir, Mingyu Ding, Michael S. Ryoo, and Juan Carlos Niebles. Robotic VLA benefits from joint learning with motion image diffusion. *arXiv preprint arXiv:2512.18007*, 2025. URL https://arxiv.org/abs/2512.18007.

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.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*,

2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24328–24346, 2025. URL https://proceedings.mlr.press/v267/hu25g.html.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through video world models. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194, 2025. URL https://proceedings.mlr.press/v305/jang25a.html.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, 2025a. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wPEIStHxYH.

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26991–27008, 2024. URL https://proceedings.mlr.press/v235/lee24y.html.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. URL https://arxiv.org/abs/2601.21998.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Proceedings of Robotics: Science and Systems*, 2025a. doi: 10.15607/RSS.2025.XXI.074. URL https://www.roboticsproceedings.org/rss21/p074.html.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3705–3728, 2025b. URL https://proceedings.mlr.press/v270/li25c.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yAzN4tz7oI.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. DiT4DiT: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026. URL https://arxiv.org/abs/2603.10448.

NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025a. URL https://arxiv.org/abs/2503.14734.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025b. URL https://arxiv.org/abs/2501.03575.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for visual-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.011. URL https://www.roboticsproceedings.org/rss21/p011.html.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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. URL https://arxiv.org/abs/2602.10098.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025.

Zuojin Tang, Shengchao Yuan, Xiaoxin Bai, Zhiyuan Jing, De Ma, Gang Pan, and Bin Liu. One token per frame: Reconsidering visual bandwidth in world models for VLA policy. *arXiv preprint arXiv:2605.07931*, 2026. URL https://arxiv.org/abs/2605.07931.

Homer Rich Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023. URL https://arxiv.org/abs/2308.12952.

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*, 2024. URL https://openreview.net/forum?id=NxoFmGgWC9.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. StarVLA-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026a. URL https://arxiv.org/abs/2604.11757.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b. URL https://arxiv.org/abs/2602.15922.

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. URL https://arxiv.org/abs/2603.16666.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/22d4f952efa13970f0b1ffb22170d416-Abstract-Conference.html.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3952–3971, 2025. URL https://proceedings.mlr.press/v305/zheng25a.html.

Yijie Zhu, Jie He, Rui Shao, Kaishen Yuan, Tao Tan, Xiaochen Yuan, and Zitong Yu. δVLA: Prior-guided vision-language-action models via world knowledge variation. *arXiv preprint arXiv:2603.08361*, 2026. URL https://arxiv.org/abs/2603.08361.

A Implementation Details

Optimization. All benchmarks share one optimization recipe. We use AdamW with $\beta = (0.9,\ 0.95)$, $\epsilon = 10^{-8}$, and weight decay 10^{-8}, under a cosine schedule with linear warmup. Learning rates are assigned per parameter group according to whether the group is adapted from a pretrained checkpoint or trained from initialization: 1×10^{-5} for the pretrained vision-language model, 2×10^{-5} for the jointly fine-tuned video denoiser, and 1×10^{-4} for the World Adapter, the world-model projection, and the action expert. The smallest rate on the backbone limits drift from its pretrained visual-language grounding; the video denoiser is likewise adapted rather than trained from scratch and takes an intermediate rate; the from-scratch modules take the largest rate.

Layout cue. The edge map is computed on the current frame of the primary view, which serves as the video branch's first frame, at the video branch's own resolution. We convert the frame to grayscale, apply the Canny detector with hysteresis thresholds 60 and 140, blur the resulting binary edge map with an isotropic Gaussian of $\sigma = 2.5$, and replicate the single channel to RGB before encoding it as the first-frame latent.

Training scale. All policies are trained on H100 GPUs with 16 samples per device and no gradient accumulation. The LIBERO policy is trained for 70K steps at an effective batch size of 128 across 8 GPUs; the SIMPLER policy, which covers the larger BridgeV2 and Fractal mixture, for 140K steps at an effective batch size of 256 across 16 GPUs; and the R1 Pro policy for 30K steps at an effective batch size of 128 across 8 GPUs. On all three benchmarks the matched Qwen-DiT baseline is trained at an identical step count and batch size, so every comparison against it holds the data and optimization budget fixed and varies only the architecture.

Deployment footprint. The World Adapter uses $K{=}256$ world tokens and $L{=}12$ blocks at width $d{=}2048$ with 8 attention heads. Each block applies cross-attention to the vision-language sequence, self-attention over the world tokens, and a feed-forward layer whose inner dimension equals the width, which is $10d^2$ parameters per block and approximately 0.5B in total. It is the only component the deployed policy adds over the matched Qwen-DiT baseline; its runtime cost is measured in Appendix B.

B Latency Measurement

Protocol. We measure every latency in Table 1 ourselves, on a single H100 GPU, as the end-to-end time of one policy call that returns one action chunk. It covers the full inference path from the raw camera images and instruction to the action chunk, including image and language preprocessing, and excludes environment stepping and action execution. For World Tokens, the call runs the vision-language model, the World Adapter, and the action expert with four flow-integration steps, the setting behind every result in the paper, with a compiled graph and cached conditioning; the video branch is not loaded. We report the steady-state median, 61.85 ms (62 ms in Table 1), an action throughput of roughly 129 actions per second.

Matched denoising steps. Per-chunk latency also depends on the number of denoising steps and on graph compilation. Fast-WAM denoises in ten steps without compilation; re-timed in that same configuration, World Tokens takes 119.72 ms per chunk against Fast-WAM's 182 ms, a factor of 1.5. This ten-step setting is a like-for-like control, not a configuration we deploy.

World Adapter overhead. The matched Qwen-DiT baseline, which shares the same vision-language model and action expert but computes no adapter, takes 52 ms per chunk (Table 1), which attributes roughly 10 ms of World Tokens's 61.85 ms to the World Adapter, both policies being timed on the same H100 with two 224×224 views and batch size one. The vision-language forward pass and the four-step flow integration are identical between the two policies, so the difference is attributable to the adapter; at 0.5B parameters, it is not the deployment bottleneck.

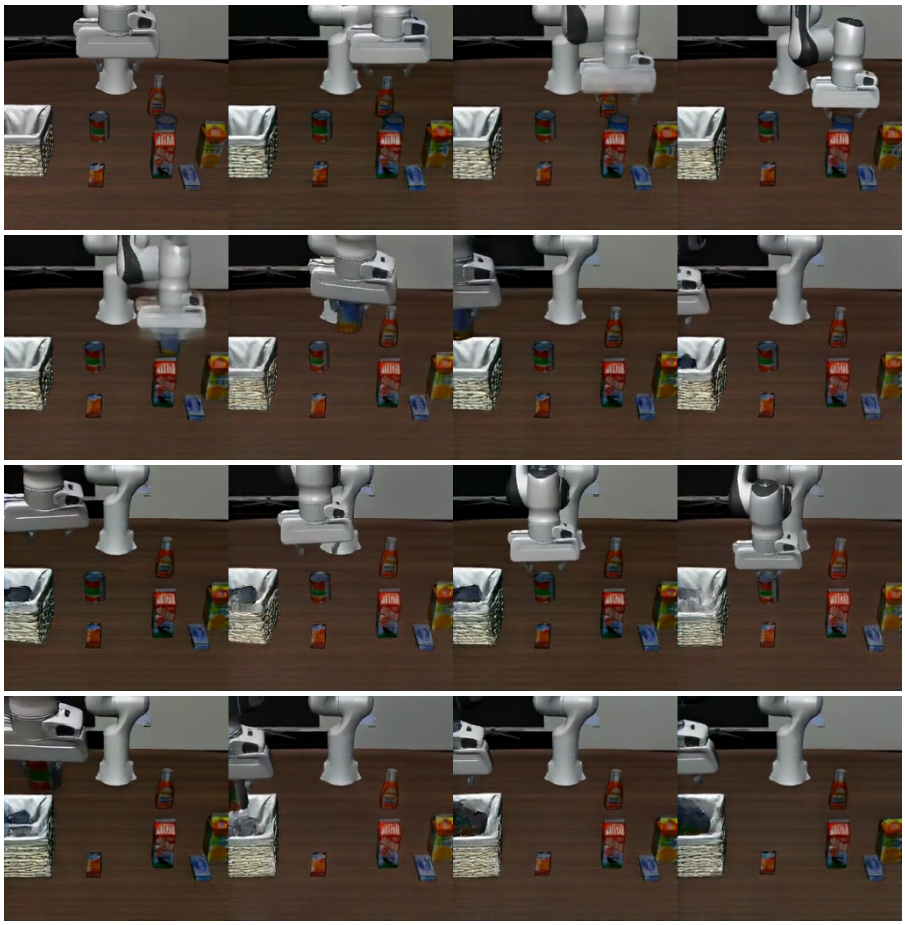

Figure S1: **Autoregressive world-model rollout.** Sixteen frames sampled uniformly from a video rolled out autoregressively by the world-model branch from a single observed frame, on the same LIBERO-Long episode as Figure 2. Frames read left to right, top row first.

C World-Model Rollout

Figure S1 shows the world-model branch rolled out autoregressively from a single observed frame, on the same LIBERO-Long episode (*put both the alphabet soup and the tomato sauce in the basket*) used for the attention maps in Figure 2 and in Appendix D. Frames are sampled uniformly over the rollout. The scene layout and the arm stay coherent across the rollout, and the arm moves through repeated reaching and retracting motions between the objects and the basket, consistent with the branch having fit the scene dynamics of the task rather than its static appearance alone. This branch is used only during training and is discarded at deployment.

D Attention Analysis

Procedure. For a given control step we take the cross-attention weights of the World Adapter's final block from each of its 8 heads and each world token to the vision-language sequence, and restrict them to the image-token positions of one camera view. Each view contributes an 8×8 grid of 64 patch tokens; renormalizing gives, for every head and world token, a distribution over 64 spatial positions whose Shannon entropy is at most $\log_2 64 = 6$ bits. We report the mean of this entropy over heads and world tokens. A control step is one policy invocation: the World Adapter recomputes the world tokens once per action chunk, so the 32 control steps of the episode in Figure S2 span 256 environment steps. Except in the last paragraph, all statistics in this section are computed over these 32 control steps of the visualized LIBERO-Long episode. World Tokens and *w/o wm* (trained with

the action loss only) are evaluated under the same environment seed, so the two trajectories start from the same state and are compared step by step.

Map selection in Figure 2. For each model and control step, Figure 2 shows the single head/world-token map with the largest attention mass on the object of the current subtask: the alphabet soup at $t{=}1, 2$, the basket at $t{=}3$ while the soup is carried, and the tomato sauce at $t{=}4, 5$, each given by the one to four patches it covers. The same objects are used for both models. The maps share one square-root color scale, with no per-map renormalization; for display they are upsampled bicubically and smoothed with a Gaussian of one quarter patch width. The selected maps place 60–100% of their mass on that object for World Tokens, against 9–25% for *w/o wm*; at every step, 2–82 head/world-token maps of World Tokens put more than half of their mass on it, and no map of *w/o wm* does.

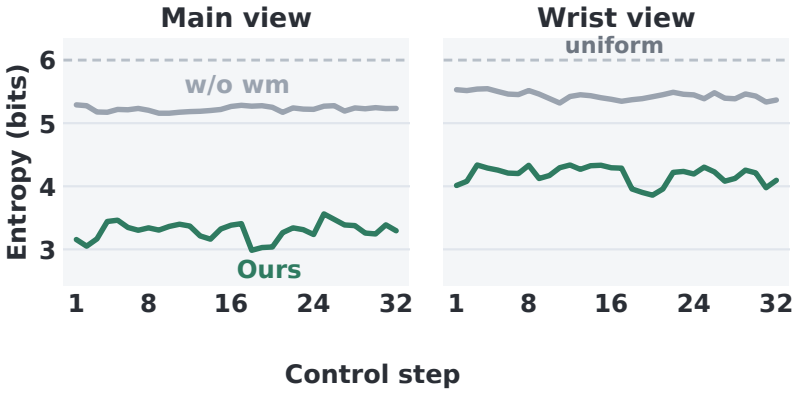

Figure S2: Spatial entropy of World Adapter attention at each control step of the LIBERO-Long episode visualized in Figure 2. The dashed line is the uniform-attention bound $\log_2 64 = 6$ bits. World Tokens has the lower entropy at every step on both views.

Additional metrics. Section 4.3 reports the main-view entropy contrast. Resolved per control step on the visualized episode (Figure S2), World Tokens attends more sharply at all 32 control steps on both views, so the gap is not an artifact of averaging. The gap also holds on the wrist view, where entropy averages 4.18 bits for World Tokens against 5.44 for *w/o wm*, a smaller but same-signed gap, consistent with the wrist view carrying less task-discriminative layout information. Other concentration measures agree: on the main view the single most-attended patch of a head/world-token map receives on average 0.42 of its mass under World Tokens versus 0.10 without world-model supervision, and its top four of the 64 patches receive 0.63 versus 0.29. Part of the concentration of World Tokens comes from maps that place most of their mass on the four corner patches, a common attention-sink pattern: 23.3% of its head/world-token maps put more than half of their mass there, against 6.2% for *w/o wm*. Excluding these maps, the mean main-view entropy is 3.99 versus 5.37 bits, so the gap persists.

A perturbed LIBERO-Plus episode. Figure S3 applies the map selection of Figure 2 to a LIBERO-Plus test episode of the task *put both the cream cheese box and the butter in the basket*. Both models are evaluated on the observations of the same successful rollout, which spans 30 control steps. The object of the current subtask is the cream cheese box at $t{=}1, 2$ (approach and grasp), the basket at $t{=}3$, the butter at $t{=}4$, and the basket at $t{=}5$; the five steps are chosen by task phase. Over all 30 steps, the mean entropy is 3.94 versus 5.22 bits on the main view and 4.73 versus 5.45 bits on the wrist view, lower for World Tokens at every step on both views. The selected map places on average 77% of its mass on the current object for World Tokens and 37% for *w/o wm*, and more than half at 26 and 6 of the 30 steps, respectively.

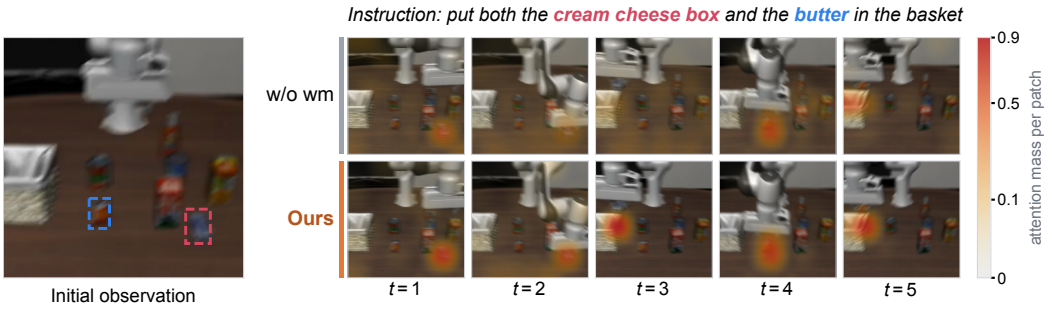

Figure S3: **World Adapter attention on a perturbed LIBERO-Plus episode.** Layout, map selection, and square-root color scale as in Figure 2; main view. The selected maps place 70, 48, 100, 97, and 100% of their mass on the current object for World Tokens and 40, 28, 29, 67, and 84% for *w/o wm*.

E Full LIBERO-Plus Results

Table S1 lists the per-dimension results behind Figure 3 for all eight published baselines evaluated by Fei et al. (2025), together with Fast-WAM, which we evaluate ourselves under the same protocol.

Table S1: Zero-shot success rate (%) on LIBERO-Plus (Fei et al., 2025). All policies are trained only on the original LIBERO data and evaluated on the 10,030 perturbed test tasks, grouped by perturbation dimension. Total is the success rate over all tasks, i.e., the per-dimension rates weighted by task count, following Fei et al. (2025). Rows in the first group are quoted from Fei et al. (2025); OpenVLA-OFT$_m$ is its variant trained on all four suites jointly. Fast-WAM is evaluated by us; its run covers 10,026 of the 10,030 tasks. Qwen-DiT is our matched baseline, trained with the same recipe as World Tokens. Bold marks the best value in each column.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
OpenVLA (Kim et al., 2025b)	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
WorldVLA (Cen et al., 2025)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
UniVLA (Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0 (Black et al., 2025b)	13.8	6.0	58.8	85.0	81.4	**79.0**	68.9	53.6
π_0-FAST (Pertsch et al., 2025)	**65.1**	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA (Tan et al., 2025)	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT (Kim et al., 2025a)	56.4	31.9	79.5	88.7	**93.3**	75.8	74.2	69.6
OpenVLA-OFT$_m$ (Kim et al., 2025a)	55.6	21.7	81.0	92.7	91.0	78.6	68.7	67.9
Fast-WAM (Yuan et al., 2026)	16.1	44.3	66.9	79.3	50.3	22.2	61.9	47.0
Qwen-DiT	49.2	51.1	81.1	93.2	92.6	57.3	74.0	69.1
World Tokens (ours)	58.7	**57.1**	**85.6**	**94.0**	92.5	74.3	**75.6**	**75.3**

F SIMPLER Google Robot Results

Table S2 reports the Google Robot visual-matching tasks for the SIMPLER policy of Section 4.1. Published Google Robot averages differ in which tasks they include, so we recompute every average from the per-task values shown, once over the three tasks that all listed methods report and once over all four.

G Statistical Tests

Table S3 tests the differences between World Tokens and the models we train ourselves. Success counts are recovered from the reported rates and the number of evaluation episodes per model: 1,500 per LIBERO suite, 6,000 over the four suites, 10,030 LIBERO-Plus tasks, and 96 R1 Pro trials. We use a two-sided two-proportion z-test with a Wald 95% confidence interval for the difference and, for the small R1 Pro sample, Fisher's exact test. The tests treat evaluation episodes as independent trials.

Table S2: SIMPLER visual-matching success (%) on Google Robot (Ye et al., 2026a; Sun et al., 2026; Qu et al., 2025). Tasks are Pick Coke Can (Pick), Move Near (Move), Open/Close Drawer (Drawer), and Open Top Drawer and Place Apple (Place). Avg. (3) averages Pick, Move, and Drawer; Avg. (4) averages all four tasks and is marked – when a method does not report Place. Methods that report only an aggregate Google Robot average (OpenVLA-OFT and $\pi_{0.5}$) are omitted. Bold marks the best value in each column.

Method	Pick	Move	Drawer	Place	Avg. (3)	Avg. (4)
π_0 (Black et al., 2025b)	72.7	65.3	38.3	–	58.8	–
GR00T-N1.5 (NVIDIA, 2025a)	51.7	54.0	27.8	7.4	44.5	35.2
SpatialVLA (Qu et al., 2025)	86.0	77.9	57.4	–	73.8	–
VLA-JEPA (Sun et al., 2026)	88.3	64.1	59.3	49.1	70.6	65.2
StarVLA (Ye et al., 2026a)	90.1	**82.6**	56.3	68.7	76.3	74.4
World Tokens (ours)	**91.0**	65.0	**78.7**	**93.5**	**78.2**	**82.1**

Table S3: Differences between World Tokens and models we train, in success-rate points computed from success counts, with 95% confidence intervals and two-sided p-values. n is the number of evaluation episodes per model.

Benchmark	Comparison	n	Success (%)	Difference [95% CI]	p
LIBERO, Avg.	vs. Qwen-DiT	6,000	98.2 vs. 96.9	+1.3 [0.7, 1.9]	4×10^{-6}
LIBERO, Avg.	vs. w/o wm	6,000	98.2 vs. 97.4	+0.8 [0.3, 1.3]	3×10^{-3}
LIBERO-Long	vs. Qwen-DiT	1,500	97.0 vs. 92.8	+4.2 [2.6, 5.8]	2×10^{-7}
LIBERO-Long	vs. w/o wm	1,500	97.0 vs. 95.0	+2.0 [0.6, 3.4]	5×10^{-3}
LIBERO-Long	vs. w/ VLM bypass	1,500	97.0 vs. 94.1	+2.9 [1.4, 4.3]	1×10^{-4}
LIBERO-Long	vs. FFN adapter	1,500	97.0 vs. 93.4	+3.6 [2.1, 5.1]	4×10^{-6}
LIBERO-Long	vs. RGB cue	1,500	97.0 vs. 91.5	+5.5 [3.9, 7.2]	8×10^{-11}
LIBERO-Plus, Total	vs. Qwen-DiT	10,030	75.3 vs. 69.1	+6.2 [5.0, 7.4]	$<10^{-20}$
R1 Pro, pooled	vs. Qwen-DiT	96	76.0 vs. 59.4	+16.7 [3.7, 29.7]	0.02

H REAL-WORLD ROLLOUTS

Figure S4 shows one episode pair for each R1 Pro task. Within each pair, World Tokens and Qwen-DiT start from the same object arrangement, and all episodes come from the same evaluation session.

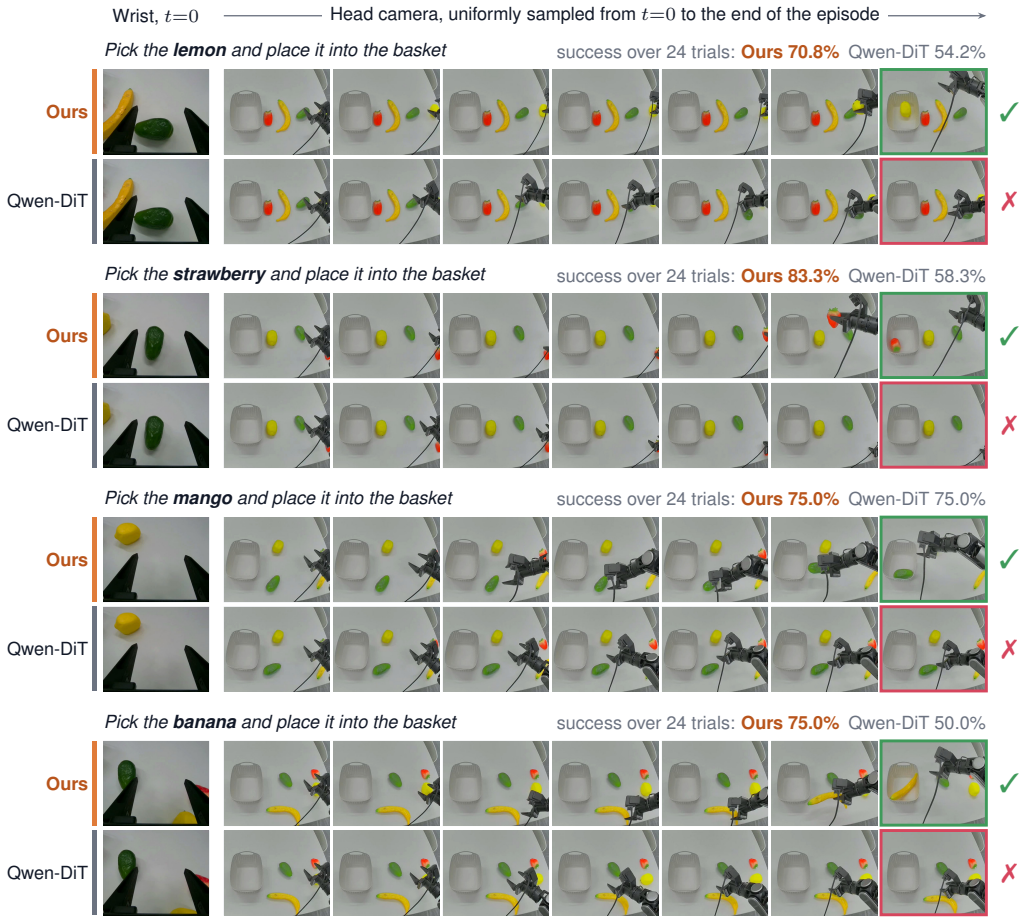

Figure S4: **R1 Pro rollouts.** Each row shows the wrist view at $t=0$ followed by seven head-camera frames sampled uniformly over the episode; the border of the last frame and the mark on the right show whether the named fruit ends in the basket. The success rates in each header are over 24 trials (Figure 4). The pairs are chosen to show differing outcomes and are illustrative only; on mango, both policies succeed in 75.0% of trials. Frames are cropped to the workspace and share the same mild contrast adjustment.