Worlds Teach, Actions Execute: Decoupled World-to-Action Distillation for Robot Manipulation

Yajie Li[1,*], Bozhou Zhang[1,*], Chun Gu[1,2], Xiaofei Wei[1,2],
Nan Song[1], Xiatian Zhu[3], Li Zhang[1,2,†]

https://logosintelligencelab.github.io/DAWN/

***Abstract*— World Action Models (WAMs) exploit spatiotemporal knowledge from video foundation models to jointly predict future visual states and actions, but inference-time video generation incurs substantial computational cost. Recent approaches accelerate inference by simplifying or bypassing video generation, yet retain the video backbone for action prediction. Distilling this future-predictive knowledge into a standalone action model could remove the inference dependency. However, directly matching world-model features can also encourage the action model to encode appearance details irrelevant to control, such as background textures. We introduce DAWN (Decoupled Action from World kNowledge), which addresses this challenge through dynamics-appearance disentangled distillation. During training, a frozen world model provides future features, while an auxiliary appearance generation pathway models their appearance content, encouraging the action model to learn complementary action-relevant dynamics. At inference, the world model and distillation module are discarded, leaving only the action model to predict actions from current observations and language instructions. Extensive experiments on LIBERO, LIBERO-Plus, RoboTwin 2.0, and real-world tasks show that DAWN matches or surpasses state-of-the-art baselines without inference-time video prediction.**

Keywords: World Action Models, Knowledge Distillation, Robot Manipulation

I. Introduction

Robotic manipulation requires grounding visual observations and language instructions into executable actions, motivating Vision-Language-Action (VLA) models [1]–[3]. While benefiting from large-scale vision-language pretraining, their reliance on static image-text data limits exposure to temporal dynamics governing physical interaction. World Action Models (WAMs) [4]–[6] enrich policy learning with spatiotemporal foresight by jointly predicting future visual states and actions, inheriting rich priors from video foundation models [7], [8] and leveraging video prediction as visual planning [9], [10].

Existing WAMs follow two paradigms: cascaded architectures [8], [11], where a world branch synthesizes a visual plan before a separate action branch decodes commands; and joint architectures [4]–[6] that unify future state prediction and action generation. Both typically couple video prediction with action generation at inference, and iterative denoising for video synthesis introduces considerable computational overhead, challenging real-time closed-loop control.

Recent work makes progress toward reconciling WAM effectiveness with inference efficiency. Fast-WAM [12] retains video co-training while removing explicit future generation at inference, repurposing the video backbone as a single-pass encoder. However, it still relies on the video backbone to encode observations for action prediction at inference. S-VAM [13] compresses multi-step video denoising into single-step inference via self-distillation, yet still uses the video backbone and feature decouplers at inference. These observations suggest an ideal solution should enable the action model to inherit the world model's future-predictive knowledge while operating independently at inference.

As illustrated in Figure 1, we advocate a different perspective: the world model and action model should assume entirely *asymmetric* roles—the world model *teaches* during training by providing structured future dynamics as a supervisory signal, while the action model independently *executes* at inference without generative overhead. This motivates three desiderata: **(i) Fast**—no video prediction or auxiliary processing at inference; **(ii) Informed**—the action model should encode future transition dynamics captured by the world model; and **(iii) Focused**—representations should prioritize action-relevant dynamics over visual appearance details.

A natural approach is to distill the world model's intermediate features into the action model via forward dynamics prediction. However, these features deeply entangle visual appearance—textures, lighting, backgrounds—with action-relevant dynamics—object displacements, gripper movements, contact changes. Aligning to such entangled features forces the action model to expend capacity on appearance information irrelevant for control, diluting the action-predictive signal.

Building on this, we introduce **DAWN** (**D**ecoupled **A**ction from **W**orld k**N**owledge), a framework connecting a dedicated world model and an action model through **dynamics-appearance disentangled distillation**. An auxiliary appearance generation pathway absorbs visual appearance from the world model's future features, encouraging the action model to specialize in action-relevant dynamics critical for manipulation. During training, the frozen world model transfers *action-centric structural priors*; at inference, the world model and distillation module are discarded, and

*Equal contribution. †Corresponding author.
[1]School of Data Science, Fudan University.
[2]Shanghai Innovation Institute.
[3]University of Surrey.

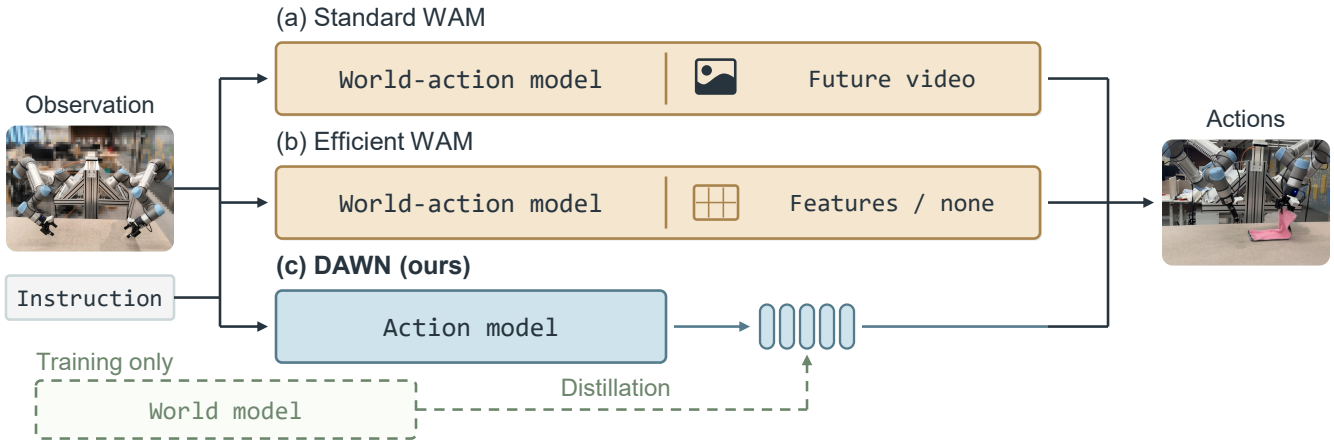


Fig. 1. **World Action Model inference paradigms.** (a) Standard WAMs synthesize explicit future video at inference, providing spatiotemporal context but incurring substantial cost from iterative denoising. (b) Efficient WAMs eliminate full video generation by repurposing the backbone as a feature encoder or compressing denoising, yet still retain auxiliary world-model components at inference. (c) *Ours*: the world model serves solely as a training-time teacher, distilling action-centric structural priors into the action model via dynamics-appearance disentangled distillation; at inference, only the standalone action model operates with no generative overhead.

only the action model operates. Experiments demonstrate that DAWN achieves performance comparable to WAM-based approaches while eliminating inference-time video generation cost.

Our contributions are summarized as follows:

- We identify the tension in World Action Models between inference efficiency and future-informed action prediction, and propose a training-inference asymmetric paradigm where the world model teaches during training and only the action model executes at inference.
- We introduce dynamics-appearance disentangled distillation, leveraging complementary learning with an auxiliary appearance pathway to encourage the action model to internalize action-relevant dynamics while filtering out irrelevant visual appearance.
- Experiments on LIBERO, LIBERO-Plus, RoboTwin 2.0, and real-world tasks demonstrate that DAWN achieves WAM-level performance with only the action model at inference, eliminating the computational overhead of video generation.

A comprehensive review of related work is provided in Appendix A.

II. Method

A. Problem Formulation and Overview

a) Task Paradigm: Given a visual observation o_t and a language instruction l, the goal is to predict an action sequence $a_{t:t+H-1}$ over an action horizon H. Robot manipulation tasks are challenging because the policy must ground high-level language goals in the current scene, identify task-relevant objects and spatial relations, and produce temporally coherent low-level controls under contact-rich interactions, occlusions, and changing object states.

b) Framework Overview: DAWN addresses this challenge through a training-inference asymmetry with three components, illustrated in Figure 2:

(1) A *frozen world model*—built on the LTX-Video [14] video diffusion transformer, pretrained on large-scale video data and adapted to multi-view robot observations, with parameters ϕ frozen after adaptation. Given o_t and l, it generates K future frames and provides intermediate layer features $\mathbf{f}_{t+1:t+K}$ as structural priors.

(2) A *dynamics-appearance disentangled distillation module* that selectively transfers action-centric priors from these features.

(3) An *action model* that predicts control commands independently of the video generation model. Since the world model is completely removed at inference time, the action model cannot rely on world-model features as inputs; it must encode the raw observation and instruction on its own while absorbing the structural priors provided during training. We therefore instantiate the model with the Qwen3-VL [15] vision-language backbone, whose strong capacity supports standalone vision-language understanding and the distillation of future-scene priors, and attach a lightweight MLP-based action head to predict $\hat{a}_{t:t+H-1}$.

During training, all three components interact; at inference, only the action model operates:

$$p_\theta(a_{t:t+H-1} \mid o_t, l) = p_\theta(a_{t:t+H-1} \mid \mathbf{z}_a(o_t, l)), \qquad (1)$$

where $\mathbf{z}_a$ denotes the action-predictive embedding produced from the current observation and language instruction, enriched during training with distilled action-centric structural priors.

B. Dynamics-Appearance Disentangled Distillation

The future features $\mathbf{f}_{t+1:t+K}$ entangle *action-relevant dynamics* (object displacements, gripper changes, contacts) with *visual appearance* (textures, lighting, backgrounds). Direct alignment to such entangled features wastes action representation capacity on control-irrelevant appearance. To address this, as illustrated in Figure 3, we introduce an

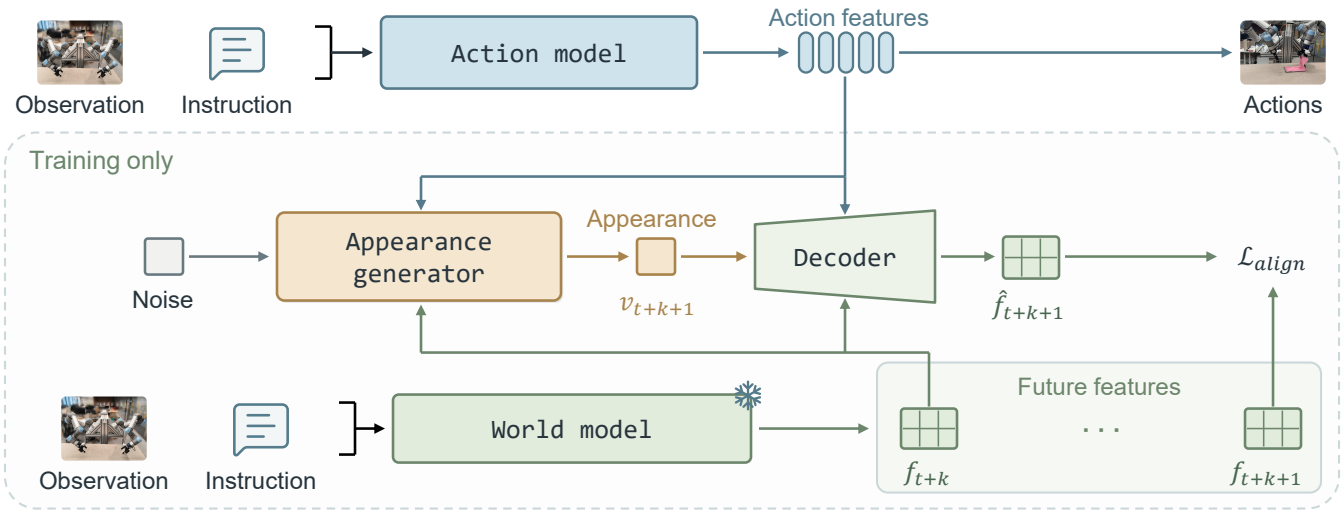


Fig. 2. **Overview of DAWN.** During training, a frozen world model generates future structural features from the current observation and language instruction. The dynamics-appearance disentangled distillation module separates action-relevant dynamics from visual appearance: an auxiliary appearance generator absorbs appearance information, while the action model's representations are driven to encode the complementary dynamics. Blue connections carry action features, while green connections show teacher conditioning and feature alignment. At inference, the world model and distillation module are discarded; only the action model operates.

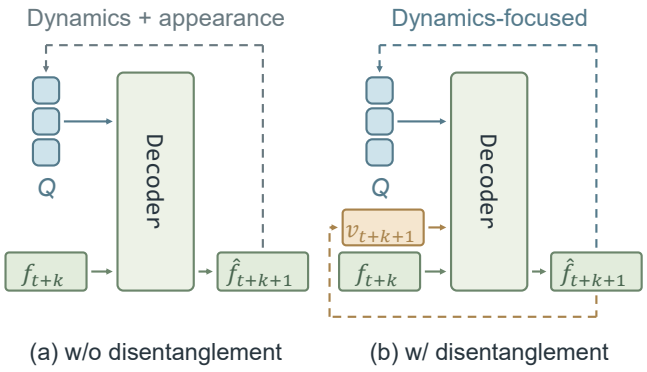


Fig. 3. **Dynamics-appearance disentangled distillation.** Blue tokens are action features Q; the yellow box is the appearance representation v_{t+k+1} from Figure 2. Solid arrows show decoder inputs and reconstruction. Dashed arrows indicate schematic learning pressure, not separate losses: (a) entangled supervision of dynamics and appearance; (b) appearance modeling through v_{t+k+1} encourages Q to focus on dynamics.

auxiliary appearance generation pathway built upon InstructPix2Pix [16], a conditional image editing diffusion model fine-tuned with LoRA [17]. Its pretrained image-to-image generation capacity naturally fits modeling frame-to-frame appearance transitions, supporting the separation of dynamics from appearance.

The world model's future features are organized into adjacent frame transition pairs following a forward dynamics formulation, modeling $\mathbf{f}_{t+k} \to \mathbf{f}_{t+k+1}$. The source frame features are projected into a compact latent space as the image condition, while the action model's intermediate representations Q are projected into the editing model's conditioning space via a learnable projector. The editing model generates visual appearance information for the future frame:

$$v_{t+k+1} = \mathscr{E}(\text{Proj}_{\text{img}}(\mathbf{f}_{t+k}),\ \text{Proj}_{\text{act}}(Q)), \tag{2}$$

where $\mathscr{E}$ is the pretrained editing model, Proj_{img} projects source features into latent space, and Proj_{act} maps action representations into the conditioning space. A decoder then reconstructs the next-frame structural features jointly conditioned on the source features, generated appearance, and action embeddings:

$$\hat{\mathbf{f}}_{t+k+1} = \text{Decode}(\mathbf{f}_{t+k},\ v_{t+k+1},\ Q). \tag{3}$$

Since the pretrained editing model inherently handles visual appearance, the action embeddings are driven to encode the complementary residual—action-relevant dynamics. This encourages the action representations to specialize in task-relevant dynamics while the auxiliary pathway absorbs appearance.

C. Training and Inference

Stage 1: World Model Preparation. The world model is pretrained and fine-tuned as described above, after which the parameters ϕ are frozen. Full training data composition is provided in Appendix B-B.

Stage 2: Distillation and Action Model Training. The distillation module and the action model are trained end-to-end with the frozen world model providing structural supervision. The total training objective $\mathscr{L} = \mathscr{L}_{\text{act}} + \lambda \mathscr{L}_{\text{align}}$ combines a *structural prior alignment loss* $\mathscr{L}_{\text{align}}$ that aligns the reconstructed features with the world model's future structural features, and an *action prediction loss* $\mathscr{L}_{\text{act}}$ that supervises the final action output, where λ is a balancing coefficient.

Inference. The world model and distillation module are discarded. Only the action model runs:

$$a_{t:t+H-1} = \text{ActionModel}(o_t, l). \tag{4}$$

TABLE I

EVALUATION RESULTS ON THE LIBERO BENCHMARK. WE REPORT THE TASK SUCCESS RATE (%) FOR EACH SUITE AND THE AVERAGE ACROSS ALL TASKS. EACH TASK IS EVALUATED OVER 50 EPISODES (500 EPISODES PER SUITE).

Method	LIBERO				
	Spatial	Object	Goal	Long	Avg
LAPA [18]	73.8	74.6	58.8	55.4	65.7
UniVLA [19]	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT [2], [20]	97.6	98.4	97.9	94.5	97.1
π_0 [3]	96.8	98.8	95.8	85.2	94.2
π_0-Fast [21]	96.4	96.8	88.6	60.2	85.5
WorldVLA [22]	87.6	96.2	83.4	60.0	81.8
villa-X [23]	97.5	97.0	91.5	74.5	90.1
GR00T N1 [24]	94.4	97.6	93.0	90.6	93.9
$\pi_{0.5}$ [25]	98.8	98.2	98.0	92.4	96.9
DAWN	96.2	100.0	98.2	94.2	97.2
w/ WM Infer.	96.5	99.7	98.7	94.1	97.3
w/o WM Distill	96.0	99.4	97.4	93.6	96.6

This eliminates the need for video generation at inference while retaining the action-centric structural priors distilled during training.

III. EXPERIMENTS

A. Implementation Details

World model. The frozen world model is built on a 2B-parameter video DiT backbone (LTX-Video [14]) adapted to multi-view robot observations, providing intermediate structural features as distillation targets. Details are in Appendix B.

Appearance generation pathway. The auxiliary appearance branch uses InstructPix2Pix [16], a conditional latent diffusion model fine-tuned with LoRA [17]. It models the appearance content of the world model's future features, encouraging the action branch to specialize in dynamics through complementary learning.

Action prediction module. A Qwen3-VL [15] backbone with a lightweight MLP action head independently encodes observations and instructions at deployment, as the world model and distillation module are discarded. Details are in Appendix C.

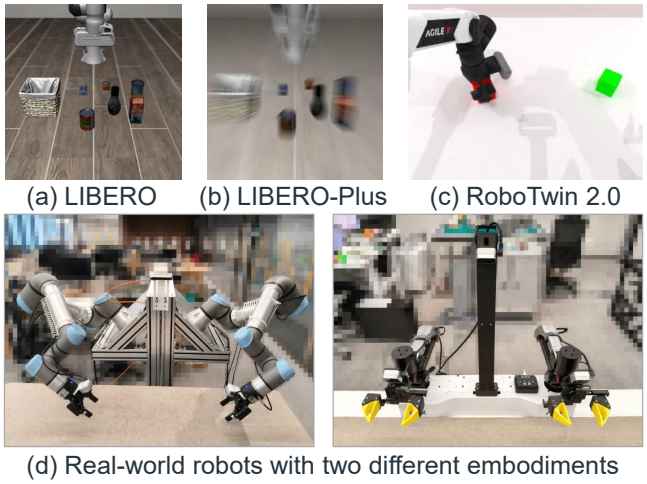


Fig. 4. **Evaluation benchmarks and robot platforms.** We evaluate DAWN on three simulation benchmarks—LIBERO, LIBERO-Plus, and RoboTwin 2.0—and on two real-world dual-arm embodiments used for physical manipulation experiments.

B. Simulation Benchmarks

We evaluate on three benchmarks: LIBERO [30] (single-arm tabletop), LIBERO-Plus [31] (robustness under domain perturbations), and RoboTwin 2.0 [32] (bimanual manipulation). We compare the full DAWN with two variants: DAWN w/ WM Infer., which retains the world model at inference to generate future video from the current observation for action prediction; and DAWN w/o WM Distill, which removes distillation entirely.

Benchmarks. *LIBERO* comprises four suites (Spatial, Object, Goal, Long) with 10 tasks each evaluated over 50 episodes. *LIBERO-Plus* extends LIBERO with perturbations along seven dimensions (Camera, Robot, Language, Light, Background, Noise, Layout); policies are trained on original LIBERO data without fine-tuning. *RoboTwin 2.0* evaluates

TABLE II

EVALUATION RESULTS ON THE LIBERO-PLUS BENCHMARK. WE REPORT THE AVERAGE SUCCESS RATE (%) ACROSS EACH PERTURBATION DIMENSION, WHERE EACH PERTURBATION INCLUDES THE FOUR TASK SUITES FROM THE ORIGINAL LIBERO BENCHMARK. AVG DENOTES THE AVERAGE SCORE OVER ALL TASKS.

Method	LIBERO-Plus							
	Camera	Robot	Language	Light	Background	Noise	Layout	Avg
UniVLA [19]	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
OpenVLA-OFT [2], [20]	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0 [3]	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-Fast [21]	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
WorldVLA [22]	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
Fast-WAM [12]	15.9	42.8	66.6	77.8	51.0	37.4	59.0	48.6
DAWN	53.0	68.4	88.2	96.5	97.1	76.3	82.4	78.7
w/ WM Infer.	50.2	70.4	86.3	95.9	97.9	76.1	84.5	78.5
w/o WM Distill	42.9	62.9	84.9	96.3	96.1	74.7	83.0	75.4

TABLE III

EVALUATION RESULTS ON ROBOTWIN 2.0. WE REPORT THE TASK SUCCESS RATE (%) UNDER THE CLEAN SETTING. ALL POLICIES ARE JOINTLY TRAINED ACROSS ALL 10 TASKS. BACKGROUND COLORS INDICATE HORIZON CATEGORY: SHORT (≤ 400), MEDIUM (400–800), AND LONG (≥ 800) STEPS.

Task	DP [26]	ACT [27]	π_0 [3]	RDT [28]	DP3 [29]	DAWN	w/ WM Infer.	w/o WM Distill
Pick Diverse Bottles	6	7	27	2	52	53	55	43
Place Mouse Pad	0	0	7	1	4	33	33	10
Place Object Scale	1	0	10	1	15	66	64	56
Place Object Stand	22	1	36	15	60	84	85	70
Open Laptop	49	56	85	59	82	84	85	75
Place Bread Basket	14	6	17	10	26	80	78	71
Place Object Basket	15	15	16	33	65	82	85	71
Handover Block	10	42	45	45	70	52	53	41
Place Cans Plasticbox	40	16	34	6	48	95	93	65
Stack Blocks Two	7	25	42	21	24	84	84	69
Average	16.4	16.8	31.9	19.3	44.6	71.3	71.5	57.1

coordinated dual-arm control on 10 jointly trained tasks categorized by horizon length into short (≤ 400), medium (400–800), and long (≥ 800) steps. The benchmark suite and real-world platforms are summarized in Figure 4, and qualitative simulation rollouts are in Appendix E-A.

LIBERO. As shown in Table I, DAWN achieves the highest average success rate among prior baselines, surpassing OpenVLA-OFT [2], [20] and $\pi_{0.5}$ [25]. The improvement over DAWN w/o WM Distill validates that distilled structural priors contribute meaningful future-predictive information. Moreover, DAWN closely matches DAWN w/ WM Infer., confirming that our distillation preserves WAM-level performance without inference-time video generation.

LIBERO-Plus. Table II evaluates robustness under domain perturbations. Fast-WAM [12], which uses video co-training but no explicit future generation at inference, obtains 15.9% on Camera and 37.4% on Noise. DAWN directly distills action-centric structural priors, achieving the highest average among baselines and surpassing OpenVLA-OFT by 9.1 points. DAWN is also comparable to or even outperforms DAWN w/ WM Infer., confirming that distilled priors preserve generalization benefits without video generation.

RoboTwin 2.0. Table III shows DAWN achieves the highest average among baselines, outperforming DP3 [29] by 26.7 points with advantages most pronounced on long-horizon tasks. DAWN remains on par with DAWN w/ WM Infer., confirming that the distilled priors largely preserve the world model's spatiotemporal reasoning capacity. The gap to DAWN w/o WM Distill further shows that distillation becomes increasingly important with task complexity.

C. Real-World Experiments

Experiment setup. We deploy DAWN on two real robot platforms: a dual-arm UR5e system and a dual-arm ARX ACONE system, each equipped with three RGB cameras (one head-mounted and two wrist-mounted). The hardware embodiments and camera layouts are detailed in Figure 6 in Appendix D. We design six manipulation tasks spanning rigid-object insertion, deformable-object folding, and multi-step sequential manipulation: *Insert the Drill*, *Hang the Towel on the Hanger*, *Place the Chain into the Box and Close the Box*, *Fold the Cloth*, *Fold the Towel*, and *Stack the Blocks*; their initial and goal configurations are shown in Figure 7 in Appendix D. For each task, we collect 300 teleoperated demonstrations for training. We compare against two baselines: GR00T N1.5 [24], a powerful VLA foundation model pretrained on large-scale cross-embodiment data, and Video Prediction Policy (VPP) [33], a lightweight yet effective world-action model that also leverages pretraining. All methods are trained on the same 300 demonstrations per task.

TABLE IV

REAL-WORLD COMPARISON. EACH METHOD USES 300 TRAINING DEMONSTRATIONS PER TASK AND IS EVALUATED OVER 20 TRIALS. BEST IN **BOLD**.

Task	GR00T N1.5	VPP	DAWN
Insert Drill	10/20	6/20	**11**/20
Hang Towel	13/20	8/20	**15**/20
Chain & Close Box	**16**/20	10/20	**16**/20
Fold Cloth	**14**/20	7/20	13/20
Fold Towel	11/20	5/20	**14**/20
Stack Blocks	**18**/20	15/20	**18**/20
Avg. Success (%)	68.3	42.5	**72.5**

Comparison with baselines. As shown in Table IV, each task is evaluated over 20 trials. DAWN achieves the highest average success rate (72.5%), outperforming GR00T N1.5 (68.3%) and VPP (42.5%). GR00T N1.5 is competitive on individual tasks but lacks explicit future-predictive reasoning; VPP, despite its pretrained world model, does not disentangle dynamics from appearance, limiting effectiveness on visually complex manipulation. These results confirm that distilled structural priors transfer effectively to real-world settings.

TABLE V

REAL-WORLD ABLATION ON *Stack the Blocks* (300 DEMOS, 20 TRIALS AT EXTREME POSITIONS).

w/o WM Distill	w/o Disentangle	DAWN
8/20	15/20	**18**/20

Real-world ablation. Table V ablates on the *Stack the Blocks* task, where all variants are trained on the same 300 demonstrations and blocks are deliberately placed at extreme positions to stress-test robustness. The results show a clear progression: the pure action model (w/o WM Distill) achieves only 8/20, adding entangled world model distillation (w/o Disentangle) improves to 15/20, and the full DAWN with disentangled distillation reaches 18/20, confirming that both world model distillation and dynamics-appearance disentanglement are essential for robust real-world performance under spatially challenging conditions. Representative successful and failed real-world rollouts are provided in Appendix E-B; see Figures 10 and 11.

D. Analysis

1) Ablation Studies: We ablate components of DAWN on RoboTwin 2.0, as shown in Figure 5, reporting success rates across horizon categories.

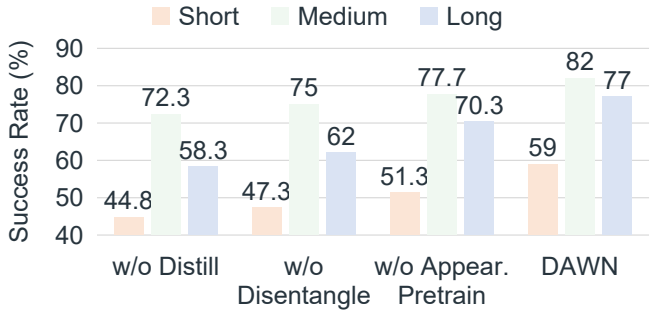


Fig. 5. **Ablation study on RoboTwin 2.0.** Success rates (%) across task horizon categories for four model variants.

Effect of dynamics-appearance disentanglement. Adding the appearance pathway (trained from scratch) yields gains over entangled distillation, especially on long-horizon tasks, supporting the role of a dedicated appearance branch in encouraging the action branch to specialize in dynamics through complementary learning.

Effect of appearance pathway pretraining. Initializing with pretrained InstructPix2Pix weights achieves the strongest performance, as pretrained visual generation priors enable more effective appearance absorption from early training, amplifying disentanglement.

2) Inference Efficiency: Table VI compares inference latency on a single NVIDIA RTX 4090, averaged over 10 runs. DAWN runs at 79.6 ms per step, 1.9× faster than DAWN w/ WM Infer. at 148.5 ms and 5.6× faster than Fast-WAM [12] at 442.0 ms, which still routes observations through the full video DiT backbone as a single-pass encoder. Combined with the competitive task performance in Tables I–III, the results confirm that discarding the world model at inference effectively eliminates generative overhead while preserving the benefits of world model distillation.

TABLE VI

INFERENCE EFFICIENCY. AVG. INFERENCE TIME (MS) PER STEP ON A SINGLE RTX 4090.

Method	Time (ms) ↓
Fast-WAM [12]	442.0
DAWN w/ WM Infer.	148.5
DAWN	**79.6**

IV. CONCLUSION

We present a training-inference asymmetric framework that bridges the gap between WAM effectiveness and real-time deployment efficiency. By distilling action-centric structural priors from a frozen world model through dynamics-appearance disentangled distillation, our approach transfers future-predictive advantages into compact action representations without requiring the world model at inference. Complementary learning with an auxiliary appearance generator encourages the distilled representations to focus on action-relevant dynamics rather than visual appearance, enabling video-free deployment that matches WAM-level performance while maintaining standard VLA inference efficiency, and offering a practical path toward deploying world-model-informed policies in real-time closed-loop control.

Limitations and future work. The distilled representations depend on the coverage and quality of the frozen world model; richer contact dynamics, deformable-object interactions, and highly novel scenes may benefit from stronger teachers or additional sensory supervision. Future work could extend this framework with uncertainty-aware distillation, tactile feedback, and continual closed-loop refinement, enabling the action model to adapt through real physical interaction.

REFERENCES

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

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "OpenVLA: An open-source vision-language-action model," in *Conference on Robot Learning (CoRL)*, 2024.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," in *Robotics: Science and Systems (RSS)*, 2025.

[4] Y. Guo, Y. Hu, J. Zhang, Y.-J. Wang, X. Chen, C. Lu, and J. Chen, "Prediction with action: Visual policy learning via joint denoising process," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[5] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," in *Robotics: Science and Systems (RSS)*, 2025.

[6] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, "Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets," in *Robotics: Science and Systems (RSS)*, 2025.

[7] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," in *International Conference on Learning Representations (ICLR)*, 2024.

[8] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, H. Zhang, and M. Zhu, "GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[9] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023.

[10] Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, L. Kaelbling, A. Zeng, and J. Tompson, "Video language planning," in *International Conference on Learning Representations (ICLR)*, 2024.

[11] Y. Shen, F. Wei, Z. Du, Y. Liang, Y. Lu, J. Yang, N. Zheng, and B. Guo, "VideoVLA: Video generators can be generalizable robot manipulators," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

[13] H. Yan, Z. Zhong, J. Zhu, J. He, W. Yuan, W. Song, X. Gong, Y. Cai, G. Zhao, X. Yan, B. Liu, Y.-C. Chen, and H. Li, "S-VAM: Shortcut video-action model by self-distilling geometric and semantic foresight," *arXiv preprint arXiv:2603.16195*, 2026.

[14] Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi, "LTX-Video: Realtime video latent diffusion," *arXiv preprint arXiv:2501.00103*, 2024.

[15] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, *et al.*, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[16] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[18] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, *et al.*, "Latent action pretraining from videos," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 28 213–28 239.

[19] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv preprint arXiv:2506.19850*, 2025.

[20] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," in *Robotics: Science and Systems (RSS)*, 2025.

[21] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "FAST: Efficient action tokenization for vision-language-action models," in *Robotics: Science and Systems (RSS)*, 2025.

[22] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang, D. Zhao, and H. Chen, "WorldVLA: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[23] X. Chen, H. Wei, P. Zhang, C. Zhang, K. Wang, Y. Guo, R. Yang, Y. Wang, X. Xiao, L. Zhao, J. Chen, and J. Bian, "villa-X: Enhancing latent action modeling in vision-language-action models," in *International Conference on Learning Representations (ICLR)*, 2026.

[24] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, *et al.*, "GR00T N1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[25] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," in *Conference on Robot Learning (CoRL)*, 2025.

[26] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

[27] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[28] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: A diffusion foundation model for bimanual manipulation," in *International Conference on Learning Representations (ICLR)*, 2025.

[29] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3d representations," in *Robotics: Science and Systems (RSS)*, 2024.

[30] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "LIBERO: Benchmarking knowledge transfer for lifelong robot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[31] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, J. Fu, J. Gong, and X. Qiu, "LIBERO-Plus: In-depth robustness analysis of vision-language-action models," *arXiv preprint arXiv:2510.13626*, 2025.

[32] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. 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.

[33] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," in *International Conference on Machine Learning (ICML)*, 2025.

[34] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems (RSS)*, 2024.

[35] Open X-Embodiment Collaboration, A. O'Neill, A. Rehman, A. Gupta, *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[36] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, *et al.*, "DROID: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems (RSS)*, 2024.

[37] AgiBot-World-Contributors, Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, *et al.*, "AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[38] Physical Intelligence, A. Amin, R. Aniceto, A. Balakrishna, K. Black, *et al.*, "$\pi^*_{0.6}$: A VLA that learns from experience," *arXiv preprint arXiv:2511.14759*, 2025.

[39] E. Yu, H. Lv, J. Sun, K. Lin, R. Zhang, Y. Shi, *et al.*, "DM0: An embodied-native vision-language-action model towards physical AI," *arXiv preprint arXiv:2602.14974*, 2026.

[40] R. Cai, J. Guo, X. He, P. Jin, J. Li, *et al.*, "Xiaomi-Robotics-0: An open-sourced vision-language-action model with real-time execution," *arXiv preprint arXiv:2602.12684*, 2026.

[41] Y. Yang, S. Zeng, T. Lin, X. Chang, D. Qi, J. Xiao, H. Liu, R. Chen, Y. Chen, D. Huo, *et al.*, "ABot-M0: VLA foundation model for robotic manipulation with action manifold learning," *arXiv preprint arXiv:2602.11236*, 2026.

[42] H. Yuan, Z. Liang, A. Chen, Y. Wang, H. Li, P. Lin, Y. Huang, Z. Lei, T. Zhang, J. Zhang, *et al.*, "Qwen-RobotManip Technical Report: Alignment Unlocks Scale for Robotic Manipulation Foundation Models," *arXiv preprint arXiv:2606.17846*, 2026. [Online]. Available: https://arxiv.org/abs/2606.17846

[43] Xiaomi Robotics Team, J. Guo, P. Jin, J. Li, P. Li, Y. Li, F. Liu, W. Peng, O. Qin, Y. Su, *et al.*, "Xiaomi-Robotics-1: Scaling Vision-Language-Action Models with over 100K Hours of Real-World Trajectories," *arXiv preprint arXiv:2607.15330*, 2026. [Online]. Available: https://arxiv.org/abs/2607.15330

[44] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang, *et al.*, "Magma: A foundation model for multimodal AI agents," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[45] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, Y. Peng, C. Shen, F. Feng, and Y. Xu, "ChatVLA: Unified multimodal understanding and robot control with vision-language-action model," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[46] Z. Zhou, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "ChatVLA-2: Vision-language-action model with open-world embodied reasoning from pretrained knowledge," *arXiv preprint arXiv:2505.21906*, 2025.

[47] J. Zhang, X. Chen, Q. Wang, M. Li, Y. Guo, Y. Hu, J. Zhang, S. Bai, J. Lin, and J. Chen, "VLM4VLA: Revisiting vision-language-models in vision-language-action models," *arXiv preprint arXiv:2601.03309*, 2026.

[48] A. Goyal, H. Hadfield, X. Yang, V. Blukis, and F. Ramos, "VLA-0: Building state-of-the-art VLAs with zero modification," *arXiv preprint arXiv:2510.13054*, 2025.

[49] C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia, *et al.*, "OpenHelix: A short survey, empirical analysis, and open-source dual-system VLA model for robotic manipulation," *arXiv preprint arXiv:2505.03912*, 2025.

[50] C. Tai, Z. Zheng, H. Long, H. Wu, H. Xiang, Z. Long, J. Xiong, R. Shi, S. Zhang, G. Qiu, *et al.*, "RealMirror: A comprehensive, open-source vision-language-action platform for embodied AI," *arXiv preprint arXiv:2509.14687*, 2025.

[51] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U.-X. Tan, N. Majumder, and S. Poria, "NORA: A small open-sourced generalist vision language action model for embodied tasks," *arXiv preprint arXiv:2504.19854*, 2025.

[52] S. Fan, K. Wu, Z. Che, X. Wang, D. Wu, F. Liao, N. Liu, Y. Zhang, Z. Zhao, Z. Xu, M. Li, Q. Liu, S. Zhang, M. Wan, and J. Tang, "XR-1: Towards versatile vision-language-action models via learning unified vision-motion representations," *arXiv preprint arXiv:2511.02776*, 2025.

[53] Y. Hu, J. Zhang, Y. Luo, Y. Guo, X. Chen, X. Sun, K. Feng, Q. Lu, S. Chen, Y. Zhang, *et al.*, "BagelVLA: Enhancing long-horizon manipulation via interleaved vision-language-action generation," *arXiv preprint arXiv:2602.09849*, 2026.

[54] H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg, and P. Abbeel, "OTTER: A vision-language-action model with text-aware visual feature extraction," *arXiv preprint arXiv:2503.03734*, 2025.

[55] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, 2022.

[56] G. Wang, C. Zhang, Q. Liu, J. Zhang, J. Cai, J. Liu, and X. Liu, "LIBERO-X: Robustness litmus for vision-language-action models," *arXiv preprint arXiv:2602.06556*, 2026.

[57] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "SIMPLER: Evaluating real-world robot manipulation policies in simulation," in *Conference on Robot Learning (CoRL)*, 2024.

[58] R. Cui, Z. Zhang, J. Pang, H. Chi, J. Guo, S. Zhang, S. Xie, X. Jin, Y. Mu, J. Yang, *et al.*, "LIBERO-Safety: A Comprehensive Benchmark for Physical and Semantic Safety in Vision-Language-Action Models," *arXiv preprint arXiv:2606.23686*, 2026, eCCV 2026. [Online]. Available: https://arxiv.org/abs/2606.23686

[59] W.-C. Tseng, G. Hussein, Y. Dong, A. Z. Ren, L. X. Shi, X. Wang, S. Levine, Z. Li, J. Gu, F. Shkurti, M.-Y. Liu, and Q. Vuong, "SC3-Eval: Evaluating Robot Foundation Models via Self-Consistent Video Generation," *arXiv preprint arXiv:2606.18610*, 2026. [Online]. Available: https://arxiv.org/abs/2606.18610

[60] Y. Liu, S. Zhang, Z. Dong, B. Ye, T. Yuan, X. Yu, L. Yin, C. Lu, J. Shi, L. J.-T. Yu, *et al.*, "FASTer: Toward powerful and efficient autoregressive vision–language–action models with learnable action tokenizer and block-wise decoding," in *International Conference on Learning Representations (ICLR)*, 2026.

[61] Y. Lu, Z. Liu, X. Fan, Z. Yang, J. Hou, J. Li, K. Ding, and H. Zhao, "FASTER: Rethinking real-time flow VLAs," *arXiv preprint arXiv:2603.19199*, 2026.

[62] Y. Wang, H. Zhu, M. Liu, J. Yang, H.-S. Fang, and T. He, "VQ-VLA: Improving vision-language-action models via scaling vector-quantized action tokenizers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[63] Z. Li, X. Gao, X. Wang, and J. Fu, "LatBot: Distilling universal latent actions for vision-language-action models," *arXiv preprint arXiv:2511.23034*, 2025.

[64] R. Xu, Y. Zhang, J. Wang, Y. Wang, and J. Yu, "Motion-Focused Latent Action Enables Cross-Embodiment VLA Training from Human EgoVideos," *arXiv preprint arXiv:2606.18955*, 2026, iROS 2026. [Online]. Available: https://arxiv.org/abs/2606.18955

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

[66] J. Wen, M. Zhu, Y. Zhu, Z. Tang, J. Li, Z. Zhou, C. Li, X. Liu, Y. Peng, C. Shen, and F. Feng, "DiffusionVLA: Scaling robot foundation models via unified diffusion and autoregression," in *International Conference on Machine Learning (ICML)*, 2025.

[67] M. Reuss, H. Zhou, M. Rühle, Ö. E. Yağmurlu, F. Otto, and R. Lioutikov, "FLOWER: Democratizing generalist robot policies with efficient vision-language-flow models," in *Conference on Robot Learning (CoRL)*, 2025.

[68] Y. Wen, H. Li, K. Gu, Y. Zhao, T. Wang, and X. Sun, "LLaDA-VLA: Vision language diffusion action models," *arXiv preprint arXiv:2509.06932*, 2025.

[69] Z. Liang, Y. Li, T. Yang, C. Wu, S. Mao, T. Nian, L. Pei, S. Zhou, X. Yang, J. Pang, Y. Mu, and P. Luo, "Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies," *arXiv preprint arXiv:2508.20072*, 2025.

[70] J. Chen, W. Song, P. Ding, Z. Zhou, H. Zhao, F. Tang, D. Wang, and H. Li, "Unified diffusion VLA: Vision-language-action model via joint discrete denoising diffusion process," *arXiv preprint arXiv:2511.01718*, 2025.

[71] W. Zhang, Y. Wang, H. Luo, H. Yuan, Y. Feng, S. Zheng, Q. Jin, and Z. Lu, "DiG-Flow: Discrepancy-guided flow matching for robust VLA models," *arXiv preprint arXiv:2512.01715*, 2025.

[72] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu, *et al.*, "HybridVLA: Collaborative diffusion and autoregression in a unified vision-language-action model," in *International Conference on Learning Representations (ICLR)*, 2026.

[73] J. Wen, M. Zhu, J. Liu, Z. Liu, Y. Yang, L. Zhang, S. Zhang, Y. Zhu, and Y. Xu, "dVLA: Diffusion vision-language-action model with multimodal chain-of-thought," *arXiv preprint arXiv:2509.25681*, 2025.

[74] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, *et al.*, "CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[75] L. Zhong, Y. Liu, Y. Wei, Z. Xiong, M. Yao, S. Liu, and G. Ren, "ACoT-VLA: Action chain-of-thought for vision-language-action models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[76] Z. Zhong, J. Li, J. He, H. Yan, X. Gong, G. Zhao, Y. Cai, J. Gao, X. Yan, B. Liu, Y. Chen, L. Yang, and H. Li, "DualCoT-VLA: Visual-linguistic chain of thought via parallel reasoning for vision-language-action models," *arXiv preprint arXiv:2603.22280*, 2026.

[77] Q. Shou, F. Zhu, S. Chen, P. Yan, Z. Yan, Y. Miao, X. Pang, Z. Hong, R. Shi, H. Huang, J. Zhang, and S. Guo, "HALO: A unified vision-language-action model for embodied multimodal chain-of-thought reasoning," *arXiv preprint arXiv:2602.21157*, 2026.

[78] C. Gu, J. Liu, H. Chen, R. Huang, Q. Wuwu, Z. Liu, X. Li, Y. Li, R. Zhang, P. Jia, P.-A. Heng, and S. Zhang, "ManualVLA: A unified VLA model for chain-of-thought manual generation and robotic manipulation," *arXiv preprint arXiv:2512.02013*, 2025.

[79] G. Yang, T. Zhang, H. Hao, W. Wang, Y. Liu, *et al.*, "Vlaser: Vision-language-action model with synergistic embodied reasoning," *arXiv preprint arXiv:2510.11027*, 2025.

[80] Z. Yang, Y. Chen, X. Zhou, J. Yan, D. Song, Y. Liu, Y. Li, Y. Zhang, P. Zhou, H. Chen, and L. Sun, "Agentic robot: A brain-inspired framework for vision-language-action models in embodied agents," *arXiv preprint arXiv:2505.23450*, 2025.

[81] C.-P. Huang, Y.-H. Wu, M.-H. Chen, Y.-C. F. Wang, and F.-E. Yang, "ThinkAct: Vision-language-action reasoning via reinforced visual latent planning," *arXiv preprint arXiv:2507.16815*, 2025.

[82] C.-P. Huang, Y. Man, Z. Yu, M.-H. Chen, J. Kautz, Y.-C. F. Wang, and F.-E. Yang, "Fast-ThinkAct: Efficient vision-language-action reasoning via verbalizable latent planning," *arXiv preprint arXiv:2601.09708*, 2026.

[83] S. Bai, J. Lyu, W. Zhou, Z. Li, D. Wang, L. Xing, X. Zhao, P. Wang, Z. Wang, C. Chi, B. Chen, and S. Zhang, "Latent reasoning VLA: Latent thinking and prediction for vision-language-action models," *arXiv preprint arXiv:2602.01166*, 2026.

[84] Z. Liu, J. Liu, H. Chen, J. Yu, Z. Guo, C. Hou, C. Gu, X. Mi, R. Zhang, K. Wu, Z. Che, J. Tang, P.-A. Heng, and S. Zhang, "LaST_0: Latent spatio-temporal chain-of-thought for robotic vision-language-action model," *arXiv preprint arXiv:2601.05248*, 2026.

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

[86] A. Ye, Z. Zhang, B. Wang, X. Wang, D. Zhang, and Z. Zhu, "VLA-

R1: Enhancing reasoning in vision-language-action models," *arXiv preprint arXiv:2510.01623*, 2025.

[87] Z. Fang, Z. Liu, J. Liu, H. Chen, Y. Zeng, S. Huang, Z. Chen, L. Chen, S. Zhang, and F. Zhao, "DualVLA: Building a generalizable embodied agent via partial decoupling of reasoning and action," *arXiv preprint arXiv:2511.22134*, 2025.

[88] D. Lei and L. Shan, "Think Less, Act Early: Reinforced Latent Reasoning with Early Exit in Vision-Language-Action Models," *arXiv preprint arXiv:2606.15099*, 2026, iCML 2026. [Online]. Available: https://arxiv.org/abs/2606.15099

[89] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, J. Gu, Z. Wang, Y. Ding, B. Zhao, D. Wang, and X. Li, "SpatialVLA: Exploring spatial representations for visual-language-action models," in *Robotics: Science and Systems (RSS)*, 2025.

[90] C. Li, J. Wen, Y. Peng, Y. Peng, and Y. Zhu, "PointVLA: Injecting the 3d world into vision-language-action models," *IEEE Robotics and Automation Letters*, 2026.

[91] L. Sun, B. Xie, Y. Liu, H. Shi, T. Wang, and J. Cao, "GeoVLA: Empowering 3d representations in vision-language-action models," *arXiv preprint arXiv:2508.09071*, 2025.

[92] V. Bhat, Y.-H. Lan, P. Krishnamurthy, R. Karri, and F. Khorrami, "3D CAVLA: Leveraging depth and 3d context to generalize vision language action models for unseen tasks," in *CVPR 2025 Workshop*, 2025.

[93] Z. Liu, Y. Gu, Y. Wang, X. Xue, and Y. Fu, "ActiveVLA: Injecting active perception into vision-language-action models for precise 3d robotic manipulation," *arXiv preprint arXiv:2601.08325*, 2026.

[94] M. Argus, J. Bratulic, H. Masnavi, M. Velikanov, N. Heppert, A. Valada, and T. Brox, "cVLA: Towards efficient camera-space VLAs," *arXiv preprint arXiv:2507.02190*, 2025.

[95] S. Deng, M. Yan, Y. Zheng, J. Su, W. Zhang, X. Zhao, H. Cui, Z. Zhang, and H. Wang, "StereoVLA: Enhancing vision-language-action models with stereo vision," *arXiv preprint arXiv:2512.21970*, 2025.

[96] J. Qian, B. Han, C. Shi, L. Xiao, L. Yang, S. Shi, and L. Jiang, "GeoPredict: Leveraging predictive kinematics and 3d Gaussian geometry for precise VLA manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[97] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang, L. Wang, T. Kong, and T. Tan, "BridgeVLA: Input-output alignment for efficient 3d manipulation learning with vision-language models," *arXiv preprint arXiv:2506.07961*, 2025.

[98] T. Zhang, H. Duan, H. Hao, Y. Qiao, J. Dai, and Z. Hou, "Grounding actions in camera space: Observation-centric vision-language-action policy," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[99] X. Wang, C. Wang, Y. Xu, M. Ye, F.-C. Zhang, J. Tian, X. Zhan, L. Zhu, C. Lu, and L. Yang, "LaMP: Learning vision-language-action policies with 3d scene flow as latent motion prior," *arXiv preprint arXiv:2603.25399*, 2026.

[100] F. Li, W. Song, H. Zhao, J. Wang, P. Ding, D. Wang, L. Zeng, and H. Li, "Spatial forcing: Implicit spatial representation alignment for vision-language-action model," *arXiv preprint arXiv:2510.12276*, 2025.

[101] Y. Feng, W. Zhang, Y. Wang, H. Luo, H. Yuan, S. Zheng, and Z. Lu, "Spatial-aware VLA pretraining through visual-physical alignment from human videos," *arXiv preprint arXiv:2512.13080*, 2025.

[102] F.-Y. Sun, S. Wu, C. Jacobsen, T. Yim, H. Zou, A. Zook, S. Li, Y.-H. Chou, E. Can, X. Wu, C. Eppner, V. Blukis, J. Tremblay, J. Wu, S. Birchfield, and N. Haber, "3D-GENERALIST: Vision-language-action models for crafting 3d worlds," in *International Conference on 3D Vision (3DV)*, 2026.

[103] J. Yang, J. Zhange, J. Li, H. Guo, W. Huang, J. Wang, P. Liu, and S.-T. Xia, "LAWM-3D: Learning 3D-Aware Latent Actions from Human Videos for Generalizable Robot World Models," *arXiv preprint arXiv:2608.05706*, 2026. [Online]. Available: https://arxiv.org/abs/2608.05706

[104] X. Sun, Y. Zhuang, M. S. L. Negrete, M.-V. Coldea, C. Liang, H. Zhang, C. Liu, Z. Zeng, S. Li, Q. Wang, F. Miao, and D. Rakita, "Artificial Foveated Perception for Mitigating Shortcut Learning in Robotic Foundation Models," *arXiv preprint arXiv:2607.10655*, 2026, coRL 2026. [Online]. Available: https://arxiv.org/abs/2607.10655

[105] R. Zhang, M. Dong, Y. Zhang, L. Heng, X. Chi, G. Dai, L. Du, D. Wang, Y. Du, and S. Zhang, "MoLe-VLA: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[106] C. Ni, C. Chen, X. Wang, Z. Zhu, W. Zheng, B. Wang, T. Chen, G. Zhao, H. Li, Z. Dong, *et al.*, "SwiftVLA: Unlocking spatiotemporal dynamics for lightweight VLA models at minimal overhead," *arXiv preprint arXiv:2512.00903*, 2025.

[107] J. Ye, N. Gao, S. Yang, J. Zheng, Z. Wang, Y. Chen, P. Chen, Y. Chen, S. Liu, and J. Jia, "StarVLA-α: Reducing complexity in vision-language-action systems," *arXiv preprint arXiv:2604.11757*, 2026.

[108] X. Tan, Y. Yang, P. Ye, J. Zheng, B. Bai, X. Wang, J. Hao, and T. Chen, "Think twice, act once: Token-aware compression and action reuse for efficient inference in vision-language-action models," *arXiv preprint arXiv:2505.21200*, 2025.

[109] W. Li, R. Zhang, R. Shao, J. He, and L. Nie, "CogVLA: Cognition-aligned vision-language-action models via instruction-driven routing & sparsification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[110] S. Dong, C. Fu, H. Gao, Y.-F. Zhang, C. Yan, C. Wu, X. Liu, Y. Shen, J. Huo, D. Jiang, H. Cao, Y. Gao, X. Sun, R. He, and C. Shan, "VITA-VLA: Efficiently teaching vision-language models to act via action expert distillation," *arXiv preprint arXiv:2510.09607*, 2025.

[111] W. Song, J. Chen, P. Ding, Y. Huang, H. Zhao, D. Wang, and H. Li, "CEED-VLA: Consistency vision-language-action model with early-exit decoding," *arXiv preprint arXiv:2506.13725*, 2025.

[112] W. Ye, P. Li, T. Yuan, Y. Xu, X. Wu, C. Zhao, J. Liu, N. Liu, Y. Huang, and L. Wang, "E-TTS: A New Embodied Test-Time Scaling Framework for Robotic Manipulation," *arXiv preprint arXiv:2606.27268*, 2026, eCCV 2026. [Online]. Available: https://arxiv.org/abs/2606.27268

[113] J. Sheng, H. Wang, and M. Liu, "SkeleWAM: Skeleton World-Action Modeling for Efficient Robotic Manipulation," *arXiv preprint arXiv:2610.02120*, 2026. [Online]. Available: https://arxiv.org/abs/2610.02120

[114] Rho Team, S. Bagaria, D. Chen, D. Fortier, J. Fu, M. Harrison, T. Hellebrekers, N. Joshi, A. Kolobov, D. Moore, *et al.*, "Rho: A Foundation for Efficiently Adaptable VLA Models," *arXiv preprint arXiv:2609.38164*, 2026. [Online]. Available: https://arxiv.org/abs/2609.38164

[115] Z. Yang, Q. Wang, Y. Wang, X. Guo, B. Yu, S. Liu, J. Xu, H. Dong, and M. Li, "Jetson-PI: Towards Onboard Real-Time Robot Control via Foresight-Aligned Asynchronous Inference," *arXiv preprint arXiv:2607.12659*, 2026, coRL 2026. [Online]. Available: https://arxiv.org/abs/2607.12659

[116] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, C. Lu, and W. Zhang, "ForceVLA: Enhancing VLA models with a force-aware MoE for contact-rich manipulation," *arXiv preprint arXiv:2505.22159*, 2025.

[117] J. Bi, K. Y. Ma, C. Hao, M. S. Zheng, and H. Soh, "VLA-Touch: Enhancing vision-language-action model with dual-level tactile feedback," *IEEE Robotics and Automation Letters*, 2026.

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

[119] H. Shi, B. Xie, Y. Liu, L. Sun, F. Liu, T. Wang, E. Zhou, H. Fan, X. Zhang, and G. Huang, "MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation," *arXiv preprint arXiv:2508.19236*, 2025.

[120] H. Li, F. Shen, D. Chen, L. Yang, X. Wang, J. Shi, Z. Bing, Z. Liu, and A. Knoll, "ReMem-VLA: Empowering vision-language-action model with memory via dual-level recurrent queries," *arXiv preprint arXiv:2603.12942*, 2026.

[121] H. Jang, S. Yu, H. Kwon, H. Jeon, Y. Seo, and J. Shin, "ContextVLA: Vision-language-action model with amortized multi-frame context," *arXiv preprint arXiv:2510.04246*, 2025.

[122] W. Song, Z. Zhou, H. Zhao, J. Chen, P. Ding, H. Yan, Y. Huang, F. Tang, D. Wang, and H. Li, "ReconVLA: Reconstructive vision-language-action model as effective robot perceiver," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[123] Z. Zheng, J. Yu, X. Peng, j. shi, M. Li, C. Zhang, W. Li, D. Wang, H. Lu, and X. Jia, "Mem-World: Memory-Augmented Action-Conditioned World Models for Persistent Robot Manipulation," *arXiv preprint arXiv:2606.18960*, 2026, coRL 2026. [Online]. Available: https://arxiv.org/abs/2606.18960

[124] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu, *et al.*, "EgoVLA: Learning vision-language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.

[125] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-H0: Vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[126] S. Kareer, K. Pertsch, J. Darpinian, J. Hoffman, D. Xu, S. Levine, C. Finn, and S. Nair, "Emergence of human to robot transfer in vision-language-action models," *arXiv preprint arXiv:2512.22414*, 2025.

[127] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, *et al.*, "Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos," *arXiv preprint arXiv:2510.21571*, 2025.

[128] J. Zheng, J. Li, Z. Wang, D. Liu, X. Kang, Y. Feng, Y. Zheng, J. Zou, Y. Chen, J. Zeng, *et al.*, "X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model," *arXiv preprint arXiv:2510.10274*, 2025.

[129] C. Li, Z. Yang, H. Zhang, F. Chen, C. Zhu, A. Bolimera, and M. Savvides, "MetaVLA: Unified meta co-training for efficient embodied adaption," *arXiv preprint arXiv:2510.05580*, 2025.

[130] Y. Chen, J. Yang, Y. Xu, Q. Ma, K. He, P. Li, K. Wang, Z. He, X. Wu, J. Liu, *et al.*, "XEWorld: Can Action-Conditioned World Models Generalize to Unseen Robot Embodiments?" *arXiv preprint arXiv:2608.05799*, 2026. [Online]. Available: https://arxiv.org/abs/2608.05799

[131] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng, and H. Li, "WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control," *arXiv preprint arXiv:2512.11047*, 2025.

[132] Z. Wu, Y. Zhou, X. Xu, Z. Wang, and H. Yan, "MoManipVLA: Transferring vision-language-action models for general mobile manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[133] H. Im, E. Jeong, A. Kolobov, J. Fu, and Y. Lee, "TwinVLA: Data-efficient bimanual manipulation with twin single-arm vision-language-action models," *arXiv preprint arXiv:2511.05275*, 2025.

[134] Z. Zhang, J. Pang, Z. Yang, K. Li, M. Liao, *et al.*, "Dexora: Open-source VLA for high-DoF bimanual dexterity," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[135] Y. Cui, Y. Zhang, L. Tao, Y. Li, X. Yi, and Z. Li, "End-to-end dexterous arm-hand VLA policies via shared autonomy: VR teleoperation augmented by autonomous hand VLA policy for efficient data collection," *arXiv preprint arXiv:2511.00139*, 2025.

[136] Z. Wang, Y. Chen, Y. Liu, J. Ye, P. Chen, C. Lu, S. Liu, B. Yu, and J. Jia, "VP-VLA: Visual prompting as an interface for vision-language-action models," *arXiv preprint arXiv:2603.22003*, 2026.

[137] X. Li, L. Xu, M. Zhang, J. Liu, Y. Shen, I. Ponomarenko, J. Xu, L. Heng, S. Huang, S. Zhang, and H. Dong, "Object-centric prompt-driven vision-language-action model for robotic manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[138] P. Li, Y. Wu, Z. Xi, W. Li, Y. Huang, Z. Zhang, Y. Chen, J. Wang, S.-C. Zhu, T. Liu, and S. Huang, "ControlVLA: Few-shot object-centric adaptation for pre-trained vision-language-action models," in *Conference on Robot Learning (CoRL)*, 2025.

[139] H. Xie, B. Wen, J. Zheng, Z. Chen, F. Hong, H. Diao, and Z. Liu, "DynamicVLA: A vision-language-action model for dynamic object manipulation," *arXiv preprint arXiv:2601.22153*, 2026.

[140] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, W. Zhang, H. Cui, Z. Zhang, and H. Wang, "GraspVLA: A grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv preprint arXiv:2505.03233*, 2025.

[141] M. Li, Z. Zhao, Z. Che, F. Liao, K. Wu, Z. Xu, P. Ren, Z. Jin, N. Liu, and J. Tang, "SwitchVLA: Execution-aware task switching for vision-language-action models," *arXiv preprint arXiv:2506.03574*, 2025.

[142] S. Yang, H. Li, B. Wang, Y. Chen, Y. Tian, T. Wang, H. Wang, F. Zhao, Y. Liao, and J. Pang, "InstructVLA: Vision-language-action instruction tuning from understanding to manipulation," *arXiv preprint arXiv:2507.17520*, 2025.

[143] K. Xu, Z. Zhu, A. Chen, S. Zhao, Q. Huang, Y. Yang, H. Lu, R. Xiong, M. Tomizuka, and Y. Wang, "Seeing to act, prompting to specify: A Bayesian factorization of vision language action policy," *arXiv preprint arXiv:2512.11218*, 2025.

[144] W. Zhao, G. Li, Z. Gong, P. Ding, H. Zhao, and D. Wang, "Unveiling the potential of vision-language-action models with open-ended multimodal instructions," *arXiv preprint arXiv:2505.11214*, 2025.

[145] T. Yang, G. Chen, Y. Chen, Z. Liang, Y. Liu, Z. Chen, C. Xu, H. Liang, J. Pang, Y. Mu, and P. Luo, "HiVLA: A visual-grounded-centric hierarchical embodied manipulation system," *arXiv preprint arXiv:2604.14125*, 2026.

[146] Y. Yang, J. Sun, S. Kou, Y. Wang, and Z. Deng, "LoHoVLA: A unified vision-language-action model for long-horizon embodied tasks," *arXiv preprint arXiv:2506.00411*, 2025.

[147] Z. Zhang, W. Luo, H. Wang, Y. Sheng, Y. Wang, H. Guo, H. Ren, X. Du, Y. Che, T. Cao, *et al.*, "Anticipation-VLA: Solving long-horizon embodied tasks via anticipation-based subgoal generation," *arXiv preprint arXiv:2605.01772*, 2026.

[148] P. Yuan, A. Ma, Y. Yao, H. Yao, M. Tomizuka, and M. Ding, "REMAC: Self-reflective and self-evolving multi-agent collaboration for long-horizon robot manipulation," *arXiv preprint arXiv:2503.22122*, 2025.

[149] X. Cai, Y. Cai, B. Chen, J. Chen, Z. Chen, S. Feng, T. Hou, J. Huang, H. Jiang, R. Ju, *et al.*, "
tau
$_0$-VLA: a Hierarchical Robot Foundation Model with World-Model-Guided Test-Time Computation," *arXiv preprint arXiv:2608.16885*, 2026. [Online]. Available: https://arxiv.org/abs/2608.16885

[150] S. Krishna, B. Eisner, H. Zhan, Y. Yuan, H. Zhen, C. Gan, S. Tulsiani, and D. Held, "GHOST: Hierarchical Sub-Goal Policies for Generalizing Robot Manipulation," *arXiv preprint arXiv:2606.10025*, 2026, rSS 2026. [Online]. Available: https://arxiv.org/abs/2606.10025

[151] J. Huang, Z. Li, Y. Hu, R. Qi, Y. Xiao, M. Coates, T. Cao, Z. Zhang, and Y. Zhang, "RoboHarness: Memory-Driven Orchestration of Heterogeneous Robot Policies for Long-Horizon Planning," *arXiv preprint arXiv:2607.18060*, 2026, eCCV Workshop Best Paper. [Online]. Available: https://arxiv.org/abs/2607.18060

[152] H. Li, P. Ding, R. Suo, Y. Wang, Z. Ge, D. Zang, K. Yu, M. Sun, H. Zhang, D. Wang, and W. Su, "VLA-RFT: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators," *arXiv preprint arXiv:2510.00406*, 2025.

[153] W. Xiao, H. Lin, A. Peng, H. Xue, T. He, Y. Xie, F. Hu, J. Wu, Z. Luo, L. Fan, G. Shi, and Y. Zhu, "Self-improving vision-language-action models with data generation via residual RL," *arXiv preprint arXiv:2511.00091*, 2025.

[154] Z. Bai, C. Gao, and M. Z. Shou, "EVOLVE-VLA: Test-time training from environment feedback for vision-language-action models," *arXiv preprint arXiv:2512.14666*, 2025.

[155] Z. Jiang, K. Liu, Y. Qin, S. Tian, Y. Zheng, M. Zhou, Z. Zhang, C. Yu, H. Li, and D. Zhao, "World4RL: Diffusion World Models for Policy Refinement With Reinforcement Learning for Robotic Manipulation," *IEEE Robotics and Automation Letters*, 2026. [Online]. Available: https://doi.org/10.1109/LRA.2026.3728345

[156] J. Yang, K. Lin, W. Zhang, J. Li, T. Lin, L. Wu, Y.-Q. Zhang, H. Zhao, P. Luo, Z. Su, *et al.*, "Self-Improving Robot Policy with Compositional World Model," *Robotics: Science and Systems XXII*, 2026. [Online]. Available: https://doi.org/10.15607/rss.2026.xxii.012

[157] K. Shi, J. Shi, P. Hebbar, Z. Zhao, T. Amarnath, Y. Su, S. Bahl, and D. Pathak, "FlowDPG: Deterministic Policy Gradient on Flow Matching Policies for Real-World Manipulation," *arXiv preprint arXiv:2606.22303*, 2026, coRL 2026. [Online]. Available: https://arxiv.org/abs/2606.22303

[158] J. Kwok, X. Zhang, M. Xu, Y. Liu, A. Mirhoseini, C. Finn, and M. Pavone, "Scaling verification can be more effective than scaling policy learning for vision-language-action alignment," *arXiv preprint arXiv:2602.12281*, 2026.

[159] J. Kwok, C. Agia, R. Sinha, M. Foutter, S. Li, I. Stoica, A. Mirhoseini, and M. Pavone, "RoboMonkey: Scaling test-time sampling and verification for vision-language-action models," in *Conference on Robot Learning (CoRL)*, 2025.

[160] Q. Gu, Y. Ju, S. Sun, I. Gilitschenski, H. Nishimura, M. Itkina, and F. Shkurti, "SAFE: Multitask failure detection for vision-language-action models," *arXiv preprint arXiv:2506.09937*, 2025.

[161] Y. Yang, Z. Duan, T. Xie, F. Cao, P. Shen, P. Song, C. Zhao, P. Jin, G. Sun, S. Xu, Y. You, and J. Liu, "FPC-VLA: A vision-language-action framework with a supervisor for failure prediction and correction," *Expert Systems with Applications*, 2026.

[162] C. Zhang, H. Zhao, H. Cheng, T. Pan, and L. Zeng, "WISE: World-model-guided Imagination Scheduling for Efficient Post-training of

Vision-Language-Action Models," *arXiv preprint arXiv:2609.03681*, 2026. [Online]. Available: https://arxiv.org/abs/2609.03681

[163] J. Liu, F. Gao, B. Wei, X. Chen, Q. Liao, Y. Wu, C. Yu, and Y. Wang, "What can rl bring to vla generalization? an empirical study," *Advances in Neural Information Processing Systems*, vol. 38, pp. 97 121–97 151, 2026.

[164] J. Zhou, K. Ye, J. Liu, T. Ma, Z. Wang, R. Qiu, K.-Y. Lin, Z. Zhao, and J. Liang, "Exploring the limits of vision-language-action manipulation in cross-task generalization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[165] I. Fang, J. Zhang, S. Tong, and C. Feng, "From intention to execution: Probing the generalization boundaries of vision-language-action models," *arXiv preprint arXiv:2506.09930*, 2025.

[166] S. Huang, J. Shao, K. Wang, Q. Chen, J. Sun, Y. Guo, M. Schwager, and J. Bohg, "Breaking lock-in: Preserving steerability under low-data VLA post-training," *arXiv preprint arXiv:2604.23121*, 2026.

[167] K. Lei, H. Li, D. Yu, Z. Wei, L. Guo, Z. Jiang, Z. Wang, S. Liang, and H. Xu, "Performant robotic manipulation with real-world reinforcement learning," *Science Robotics*, 2026. [Online]. Available: https://doi.org/10.1126/scirobotics.aed6267

[168] L. Wang, Z. Wang, X. Li, L. Fan, C. Lu, and N. Ye, "Scaling by Diversified Experience for Vision-Language-Action Models," *arXiv preprint arXiv:2606.09009*, 2026, iCML 2026. [Online]. Available: https://arxiv.org/abs/2606.09009

[169] Y. Zhang, C. Wang, O. Lu, Y. Zhao, Y. Ge, Z. Sun, X. Li, C. Zhang, C. Bai, and X. Li, "Align-then-steer: Adapting the vision-language action models through unified latent guidance," *arXiv preprint arXiv:2509.02055*, 2025.

[170] W. Li, S. Wang, Y. Chen, H. Zhuang, and Q. Wu, "AntiGrounding: Lifting robotic actions into VLM representation space for decision making," *arXiv preprint arXiv:2506.12374*, 2025.

[171] H. Zhao, J. Zhang, W. Song, P. Ding, and D. Wang, "VLA^2: Empowering vision-language-action models with an agentic framework for unseen concept manipulation," *arXiv preprint arXiv:2510.14902*, 2025.

[172] D. M. Nguyen, N. T. Diep, B. G. Nguyen, T.-B. Ho, D. Le, T. Q. Nguyen, T.-L. Ha, N. Tran, B. Thach, N. X. Tran, *et al.*, "FOCA: Future-Oriented Conditioning for Data-Efficient Vision-Language-Action Adaptation," *arXiv preprint arXiv:2606.20867*, 2026, iCML 2026. [Online]. Available: https://arxiv.org/abs/2606.20867

[173] J. Sun, W. Zhang, Z. Qi, S. Ren, Z. Liu, H. Zhu, G. Sun, X. Jin, and Z. Chen, "VLA-JEPA: Enhancing vision-language-action model with latent world model," *arXiv preprint arXiv:2602.10098*, 2026.

[174] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "UP-VLA: A unified understanding and prediction model for embodied agent," in *International Conference on Machine Learning (ICML)*, 2025.

[175] J. Won, K. Lee, H. Jang, D. Kim, and J. Shin, "Dual-stream diffusion for world-model augmented vision-language-action model," *arXiv preprint arXiv:2510.27607*, 2025.

[176] Y. Tian, Y. Jin, B. Yu, Y. Shi, H. Wu, C. H. Liu, K. Chen, and C. Huang, "STARRY: Spatial-temporal action-centric world modeling for robotic manipulation," *arXiv preprint arXiv:2604.26848*, 2026.

[177] Z. Liu, Y. Gu, S. Zheng, Y. Fu, X. Xue, and Y.-G. Jiang, "TriVLA: A triple-system-based unified vision-language-action model with episodic world modeling for general robot control," *arXiv preprint arXiv:2507.01424*, 2025.

[178] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, F. Lu, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin, "DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025.

[179] Z. Huang, J. Zhang, H. Liu, C. Zhang, R. Cheng, and L. Zhang, "Learning Transferable Dynamics Priors from Action to World Modeling," *arXiv preprint arXiv:2606.29501*, 2026, eCCV 2026. [Online]. Available: https://arxiv.org/abs/2606.29501

[180] L. Chen, H. Li, W. Yang, M. Zhao, and D. Jiang, "ViPSim: Collaborating Visual and Parameter Spaces for Consistent Long-Horizon Embodied World Models," *arXiv preprint arXiv:2606.28804*, 2026, rSS 2026. [Online]. Available: https://arxiv.org/abs/2606.28804

[181] Z. Shan, Z. Wu, X. Wang, Z. Zhu, and Z. Wang, "DVG-WM: Disentangled Video Generation Enables Efficient Embodied World Model for Robotic Manipulation," *arXiv preprint arXiv:2606.32028*, 2026. [Online]. Available: https://arxiv.org/abs/2606.32028

[182] X. Liu, Z. Bai, H. Ci, K. Y. Ma, and M. Z. Shou, "World-VLA-Loop: Closed-loop learning of video world model and VLA policy," *arXiv preprint arXiv:2602.06508*, 2026.

[183] Z. Zhang, H. Ren, Y. Sun, Y. Sheng, H. Wang, H. Lin, Z. Wu, P.-L. Bacon, and Y. Yu, "Towards practical world model-based reinforcement learning for vision-language-action models," *arXiv preprint arXiv:2603.20607*, 2026.

[184] A. Yu, Z. Chen, P. Song, Z. Hong, H. Wang, D. Zhang, T. He, Y. Ding, and D. Zhang, "WM-DAgger: Enabling efficient data aggregation for imitation learning with world models," *arXiv preprint arXiv:2604.11351*, 2026.

[185] GigaBrain Team, B. Wang, B. Li, C. Ni, G. Huang, *et al.*, "GigaBrain-0.5M*: A VLA that learns from world model-based reinforcement learning," *arXiv preprint arXiv:2602.12099*, 2026.

[186] L. Li, Q. Zhang, Y. Luo, S. Yang, R. Wang, F. Han, M. Yu, Z. Gao, N. Xue, X. Zhu, Y. Shen, and Y. Xu, "Causal World Modeling for Robot Control," *arXiv preprint arXiv:2601.21998*, 2026, rSS 2026. [Online]. Available: https://arxiv.org/abs/2601.21998

[187] Y. Chen, Y. Ge, H. Zhou, M. Ding, Y. Ge, and X. Liu, "DIAL: Decoupling intent and action via latent world modeling for end-to-end VLA," *arXiv preprint arXiv:2603.29844*, 2026.

[188] R. Wang, Q. Liu, Y. Deng, G. Liu, Z. Liu, and K. Jia, "EVA: Aligning video world models with executable robot actions via inverse dynamics rewards," *arXiv preprint arXiv:2603.17808*, 2026.

[189] S. Chen, J. Liu, J. Wu, Y. Guo, T. Wang, S. Qian, H. Chen, J. Cao, J. Tang, and S. Zhang, "Do Robotic World Models Really Follow Actions? Diagnosing and Aligning Action-Conditioned Generation for Policy Learning," *arXiv preprint arXiv:2608.24885*, 2026. [Online]. Available: https://arxiv.org/abs/2608.24885

[190] Q. Shen, S. Zhang, Y. Liao, Q. Li, Z. Tan, S. Wang, S. Yan, and X. Wang, "World Action Models: A Survey," *arXiv preprint arXiv:2606.20781*, 2026, survey. [Online]. Available: https://arxiv.org/abs/2606.20781

[191] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," in *International Conference on Learning Representations (ICLR)*, 2024.

[192] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation," in *Conference on Robot Learning (CoRL)*, 2025.

[193] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic manipulation by imitating generated videos without physical demonstrations," in *International Conference on Learning Representations (ICLR)*, 2026.

[194] P. Li, H. Wu, Y. Huang, C. Cheang, L. Wang, and T. Kong, "GR-MG: Leveraging partially annotated data via multi-modal goal-conditioned policy," *IEEE Robotics and Automation Letters*, 2025.

[195] M. J. Kim, Y. Gao, T.-Y. Lin, Y.-C. Lin, Y. Ge, G. Lam, P. Liang, S. Song, M.-Y. Liu, C. Finn, and J. Gu, "Cosmos policy: Fine-tuning video models for visuomotor control and planning," in *International Conference on Learning Representations (ICLR)*, 2026.

[196] Y. Huang, X. Lv, J. Xu, Z. Yu, J. Zhang, R. Hu, W. Feng, S. Zou, H. Xiao, Z. Zhou, *et al.*, "PAIWorld: A 3D-Consistent World Foundation Model for Robotic Manipulation," *arXiv preprint arXiv:2606.18375*, 2026. [Online]. Available: https://arxiv.org/abs/2606.18375

[197] J. Luo, H. Zhu, C. Gao, J. Wang, C. Wang, T. He, J. Li, and Z. Chen, "EmbodiedVAE: Disentangled Video VAE for Efficient and Controllable Embodied Manipulation," *arXiv preprint arXiv:2608.02990*, 2026, eCCV 2026. [Online]. Available: https://arxiv.org/abs/2608.02990

[198] L. Yang, Y. Bai, G. Eskandar, F. Shen, M. Altillawi, D. Chen, S. Majumder, Z. Liu, G. Kutyniok, and A. Valada, "RoboEnvision: A long-horizon video generation model for multi-task robot manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[199] X. Jia, B. Yang, Z. Ge, X. Nie, Y. Zhou, *et al.*, "GuidedVLA: Specifying task-relevant factors via plug-and-play action attention specialization," *arXiv preprint arXiv:2605.12369*, 2026.

[200] L. Yang, Y. Bai, G. Eskandar, F. Shen, M. Altillawi, D. Chen, Z. Liu, and A. Valada, "CoVAR: Co-generation of video and action for robotic manipulation via multi-modal diffusion," *arXiv preprint arXiv:2512.16023*, 2025.

[201] J. Pai, L. Achenbach, V. Montesinos, B. Forrai, O. Mees, and E. Nava, "mimic-video: Video-action models for generalizable robot control beyond VLAs," *arXiv preprint arXiv:2512.15692*, 2025.

[202] Q. Lv, W. Kong, H. Li, J. Zeng, Z. Qiu, D. Qu, H. Song, Q. Chen, X. Deng, and J. Pang, "F1: A vision-language-action model bridging understanding and generation to actions," *arXiv preprint arXiv:2509.06951*, 2025.

[203] T. Ma, J. Zheng, Z. Wang, C. Jiang, A. Cui, J. Liang, and S. Yang, "DiT4DiT: Jointly modeling video dynamics and actions for generalizable robot control," *arXiv preprint arXiv:2603.10448*, 2026.

[204] J. Chen, W. Song, J. Wang, S. Zhou, X. Gong, Z. Fan, Z. Zhou, J. E, H. Yan, F. Li, *et al.*, "UniWAM: Unified World-Action Model," *arXiv preprint arXiv:2610.02054*, 2026. [Online]. Available: https://arxiv.org/abs/2610.02054

[205] J. Zhu, J. Zhang, T. Su, T. Liu, Z. Wang, K. Xie, Z. Huang, C. Ma, Y. He, T. Wang, *et al.*, "DSWAM: A Dual-System World Action Foundation Model for Fine-Grained Robot Manipulation," *arXiv preprint arXiv:2607.04927*, 2026. [Online]. Available: https://arxiv.org/abs/2607.04927

[206] C. Gu, H. Yu, J. Zhang, H. Lin, W. Zhang, J. Wang, H. Jin, S. Xie, J. Jiang, and Z. Wang, "GeniWorld: A Generalizable Interactive World Model for Robotic Manipulation via Visual Actions," *arXiv preprint arXiv:2608.06332*, 2026. [Online]. Available: https://arxiv.org/abs/2608.06332

[207] X. Chen, Z. Yin, Y. Zhang, L. Zhang, H. Zheng, T. Mu, S. Zuo, D. Zhou, D. Wu, X. Zhou, *et al.*, "Magic-W0: A Structured World-Action Foundation Model for Physical Intelligence," *arXiv preprint arXiv:2609.39870*, 2026. [Online]. Available: https://arxiv.org/abs/2609.39870

[208] J. Cen, S. Huang, Y. Yuan, K. Li, H. Yuan, C. Yu, Y. Jiang, J. Guo, X. Li, H. Luo, F. Wang, D. Zhao, and H. Chen, "RynnVLA-002: A unified vision-language-action and world model," *arXiv preprint arXiv:2511.17502*, 2025.

[209] Y. Lin, J. He, S. Bao, C. Zhao, Y. Li, X. Wang, Y. Wang, C. Chi, and J. Zhang, "JEPA-WAM: Learning Vision-Language-Action Policies with Joint-Embedding World Modeling," *arXiv preprint arXiv:2608.09381*, 2026. [Online]. Available: https://arxiv.org/abs/2608.09381

[210] G. Yan, J. Liu, Y. Fan, L. Cai, M. Liao, J. Zhang, and D. Fox, "Flex-
pi: A Multi-Stream World-Action Model with Compute Flexibility," *arXiv preprint arXiv:2608.10860*, 2026. [Online]. Available: https://arxiv.org/abs/2608.10860

[211] A. Ye, B. Wang, C. Ni, G. Huang, G. Zhao, *et al.*, "GigaWorld-Policy: An efficient action-centered world-action model," *arXiv preprint arXiv:2603.17240*, 2026.

[212] Y. Zhu, R. Shao, J. He, W. Li, B. Zhao, Y. Wang, X. Yuan, T. Tan, M. Zhang, X. Peng, and Z. Yu, "ATI-VLA: Action-Centric Predictive Vision-Language-Action Models via Actionable Alignment Then Adaptive Injection," *arXiv preprint arXiv:2610.01741*, 2026. [Online]. Available: https://arxiv.org/abs/2610.01741

[213] R. Li, H. Zhang, J. Jin, Q. Zeng, Z. Zhuang, Y. Tang, S. Lyu, and D. Wang, "World-value-action model: Implicit planning for vision-language-action systems," *arXiv preprint arXiv:2604.14732*, 2026.

[214] Z. Wang, J. Li, Y. Cui, Y. Gao, X. Zhan, J. Yu, and X. Ma, "World Value Models for Robotic Manipulation," *arXiv preprint arXiv:2606.24742*, 2026. [Online]. Available: https://arxiv.org/abs/2606.24742

[215] H. Zhao, J. Wang, W. Song, S. Chen, Y. Liu, Y. Wang, H. Li, and D. Wang, "FRAPPE: Infusing world modeling into generalist policies via multiple future representation alignment," *arXiv preprint arXiv:2602.17259*, 2026.

[216] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, *et al.*, "FLARE: Robot learning with implicit world modeling," in *Conference on Robot Learning (CoRL)*, 2025.

[217] Y. Zhang, J. Zhao, C. Fan, J. Hu, X. Yuan, C. Bai, and X. Li, "V-JEPA Policy: Building Effective World-Action Models on Predictive Visual Latents," *arXiv preprint arXiv:2609.37250*, 2026. [Online]. Available: https://arxiv.org/abs/2609.37250

[218] Y. Zheng, S. Gu, W. Li, Y. Zheng, Y. Zang, S. Tian, X. Li, C. Hao, C. Gao, S. Liu, H. Li, Y. Chen, S. Yan, and W. Ding, "OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation," *arXiv preprint arXiv:2603.19201*, 2026.

[219] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, *et al.*, "World action models are zero-shot policies," *arXiv preprint arXiv:2602.15922*, 2026.

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

[221] C. Fu, X. Chi, Y. Rui, Y.-k. Wang, Z. Qian, X. Zhang, Y. Lou, K. Zhang, K. Ge, C. W. Mak, *et al.*, "Token-World: World Modeling in Vision-Language Model Token Space for Robot Manipulation," *arXiv preprint arXiv:2610.00575*, 2026. [Online]. Available: https://arxiv.org/abs/2610.00575

[222] S. Gao, S. Zhou, Y. Du, J. Zhang, and C. Gan, "AdaWorld: Learning adaptable world models with latent actions," in *International Conference on Machine Learning (ICML)*, 2025.

[223] H. Bi, H. Tan, S. Xie, Z. Wang, S. Huang, H. Liu, R. Zhao, Y. Feng, C. Xiang, Y. Rong, *et al.*, "Motus: A unified latent action world model," *arXiv preprint arXiv:2512.13030*, 2025.

[224] J. Lyu, K. Liu, X. Zhang, H. Liao, Y. Feng, *et al.*, "LDA-1B: Scaling latent dynamics action model via universal embodied data ingestion," *arXiv preprint arXiv:2602.12215*, 2026.

[225] A. D. Vuong, T. V. Vo, A. Sohail, H. Ding, L. Ma, X. Liang, A. Duan, I. Laptev, and I. Reid, "World2Act: Latent action post-training via skill-compositional world models," *arXiv preprint arXiv:2603.10422*, 2026.

[226] Y. Deng, G. Liu, and K. Jia, "DexWorldModel: Causal latent world modeling towards automated learning of embodied tasks," *arXiv preprint arXiv:2604.16484*, 2026.

[227] P. Zhou, L. Chen, S. Chen, D. Chen, W. Zhao, R. Jin, G. Ren, and J. Luo, "Act2Goal: From world model to general goal-conditioned policy," *arXiv preprint arXiv:2512.23541*, 2025.

[228] Z. Wang, H. Li, W. Wang, J. Cai, J. Zeng, Y. Chen, J. Pang, and S. Liu, "RoboInter1.5: A Holistic Intermediate Representation Suite for Embodied World Modeling and Robotic Manipulation," *arXiv preprint arXiv:2607.18709*, 2026. [Online]. Available: https://arxiv.org/abs/2607.18709

[229] M. K. Govind, D. Reilly, S. Patel, H. Le, and S. Das, "World Action Models Enable Continual Imitation Learning with Recurrent Generative Replays," *arXiv preprint arXiv:2606.27374*, 2026. [Online]. Available: https://arxiv.org/abs/2606.27374

[230] H. Zhen, Q. Sun, H. Zhang, J. Li, S. Zhou, Y. Du, and C. Gan, "Learning 4d embodied world models," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[231] Z. Liu, S. Li, E. Cousineau, S. Feng, B. Burchfiel, and S. Song, "Geometry-aware 4d video generation for robot manipulation," in *International Conference on Learning Representations (ICLR)*, 2026.

[232] J. Guo, Q. Li, P. Li, Z. Chen, N. Sun, Y. Su, H. Wang, Y. Zhang, X. Li, and H. Liu, "Unified 4d world action modeling from video priors with asynchronous denoising," *arXiv preprint arXiv:2604.26694*, 2026.

[233] W. Huang, H. Sun, Y. Guo, Y. Ma, H. Li, J. Long, Z. Mo, Z. Guan, Y. Guo, S. Di, and J. Xiong, "NoiseGate: Learning per-latent timestep schedules as information gating in world action models," *arXiv preprint arXiv:2605.07794*, 2026.

[234] R. Wang, Y. Zhang, J. Lin, K. Luo, J. Wang, Z. Wang, and X. Qi, "When to trust imagination: Adaptive action execution for world action models," *arXiv preprint arXiv:2605.06222*, 2026.

[235] C. Zhang, J. Tong, X. Li, Y. Wang, and H. Li, "Keep the Future, Drop the Rollout: RIFT for World Action Models," *arXiv preprint arXiv:2608.11521*, 2026. [Online]. Available: https://arxiv.org/abs/2608.11521

[236] W. Zhao, H. Jiang, X. Shi, L. Liu, F. Huang, Z. Su, W. Sui, and X. Wang, "Faster-WAM: Efficient Inference-Time Future Conditioning for Robust World Action Models," *arXiv preprint arXiv:2608.04404*, 2026. [Online]. Available: https://arxiv.org/abs/2608.04404

[237] Q. Feng, J. Yu, J. Liu, Y. Jia, Z. Wu, H. Chen, Z. Qian, S. Gu, P. Jia, S. Ma, and S. Zhang, "HarmoWAM: Harmonizing generalizable and precise manipulation via adaptive world action models," *arXiv preprint arXiv:2605.10942*, 2026.

[238] G. Yuan, Q. Qiao, J. Zhang, and D. Xu, "AdaWorldPolicy: World-model-driven diffusion policy with online adaptive learning for robotic manipulation," *arXiv preprint arXiv:2602.20057*, 2026.

[239] Y. Liu, P. Sun, S. Li, Y. Xie, L. Zhang, X. Chao, S. Dong, F. Chen, X.-P. Zhang, and W. Ding, "OA-WAM: Object-addressable world action model for robust robot manipulation," *arXiv preprint arXiv:2605.06481*, 2026.

[240] L. Fan, Z. Xu, C. Cao, W. Zhang, M. Yuan, and J. Chen, "AIM: Intent-aware unified world action modeling with spatial value maps," *arXiv preprint arXiv:2604.11135*, 2026.

[241] R. Wu, L. Ge, and X. Li, "ActiveWAM: Evidence-Aware Active Vision for World-Action Models," *arXiv preprint arXiv:2610.01698*, 2026. [Online]. Available: https://arxiv.org/abs/2610.01698

[242] Y. Jiang, Y. Guo, H. Yang, G. Lei, N. Chen, Y. Zhang, S. Yan, B. Lin, F. Gao, and B. Qi, "CKT-WAM: Parameter-efficient context knowledge transfer between world action models," *arXiv preprint arXiv:2605.06247*, 2026.

[243] MotuBrain Team, C. Xiang, F. Bao, H. Liu, H. Tan, *et al.*, "MotuBrain: An advanced world action model for robot control," *arXiv preprint arXiv:2604.27792*, 2026.

[244] B. Pan, F. Liu, H. Lu, J. Wang, and Y. Shi, "SelfWAM: A Self-Grounded Unified World Action Model for Fast Robot Control," *arXiv preprint arXiv:2608.00725*, 2026. [Online]. Available: https://arxiv.org/abs/2608.00725

[245] X. Yao, H. Zhou, O. Mees, Y. Meng, T. Xiao, Y. Bisk, J. Oh, E. Johns, M. Shridhar, D. Shah, *et al.*, "Bridging language and action: A survey of language-conditioned robot manipulation," *The International Journal of Robotics Research*, 2026. [Online]. Available: https://doi.org/10.1177/02783649261468360

[246] F. Wang, Z. Wang, G. Pei, M. Zhang, C. Liang, J. Hu, Z. Li, J. Wu, N. Han, Z. Zhang, *et al.*, "World Models for Robotic Manipulation: A Survey," *SmartBot*, 2026. [Online]. Available: https://doi.org/10.1002/smb2.70053

[247] T. Jiang, T. Yuan, Y. Liu, C. Lu, J. Cui, X. Liu, S. Cheng, J. Gao, H. Xu, and H. Zhao, "Galaxea open-world dataset and g0 dual-system vla model," *arXiv preprint arXiv:2509.00576*, 2025.

[248] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, *et al.*, "RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation," in *Robotics: Science and Systems (RSS)*, 2025.

[249] S. Wu, X. Liu, S. Xie, P. Wang, X. Li, B. Yang, Z. Li, K. Zhu, H. Wu, Y. Liu, *et al.*, "Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation," *arXiv preprint arXiv:2511.17441*, 2025.

[250] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[251] Z. Zhao, H. Jing, X. Liu, J. Mao, A. Jha, H. Yang, R. Xue, S. Zakharov, V. Guizilini, and Y. Wang, "Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation," *arXiv preprint arXiv:2510.08807*, 2025.

Appendix Table of Contents

Appendix A: Related Work 14
A-A Vision-Language-Action Models 14
A-B World Action Models 15
Appendix B: World Model Details 15
B-A Architecture 15
B-B Training Data and Configuration 15
Appendix C: DAWN Architecture and Training Details . 15
C-A Model Architecture 15
C-B Training Hyperparameters 16
Appendix D: Real-World Experimental Setup 16
Appendix E: Qualitative Visualization 16
E-A Simulation Rollout Visualization 16
E-B Real-World Rollout Visualization 16

APPENDIX A
RELATED WORK

A. Vision-Language-Action Models

The remarkable generalization achieved by large-scale vision-language pretraining has motivated researchers to repurpose such multimodal backbones for robotic control, giving rise to the Vision-Language-Action (VLA) paradigm—end-to-end models that map raw visual observations and language instructions directly to motor commands. RT-2 [1] pioneered this direction by demonstrating that a pretrained vision-language model can be fine-tuned to output tokenized robot actions, thereby transferring web-scale knowledge to manipulation tasks. This insight was rapidly amplified by open-source generalist policies such as OpenVLA [2] and Octo [34], which leveraged large cross-embodiment datasets [35]–[37], and by the π_0 family [3], [25], [38] that progressively scaled flow-matching VLAs toward open-world generalization and continual learning from online experience. In parallel, industrial foundation models [24], [39]–[43] and multimodal agent architectures [44]–[46] have driven the paradigm toward real-time deployment and unified perception–action capabilities, supported by a growing ecosystem of open designs [19], [28], [47]–[54] evaluated on standardized benchmarks [30], [31], [55]–[59].

A central design question in VLAs is how to represent continuous robot actions within a discrete language-model backbone. One strand of research develops compact discrete tokenizers—via DCT-based compression [21], [60], [61], vector quantization [62], or universal latent action distillation [63], [64]—that enable autoregressive generation of multi-step action chunks. A complementary strand replaces discrete decoding entirely with continuous diffusion or flow-matching action heads [65]–[71], allowing multimodal action distributions to be captured without discretization artifacts. Hybrid architectures that combine autoregressive reasoning with diffusion-based action generation [72], [73] further suggest that the two paradigms offer complementary strengths, and the optimal design trade-off remains an open question.

Reactive, single-step policies frequently struggle with compositional or long-horizon manipulation, motivating the integration of explicit or latent reasoning into VLAs. Explicit chain-of-thought methods produce visual or linguistic rationales prior to acting [74]–[80], whereas latent reasoning approaches encode deliberation into hidden representations for faster inference [81]–[88]. Orthogonally, monocular VLAs often misjudge depth and contact geometry, prompting the integration of point clouds, stereo cues, and explicit 3D representations [29], [89]–[104]. Because both reasoning and 3D perception incur additional computational cost, a parallel thread investigates making VLAs lightweight enough for real-time control through layer skipping, token pruning, early-exit decoding, and knowledge distillation [105]–[115].

Beyond core architecture design, extending VLA capabilities along multiple axes has attracted broad attention. On the input side, force/torque and tactile sensors supply contact-rich feedback that vision alone cannot capture [116]–[118], while recurrent and reconstructive memory mechanisms enable history-aware perception over extended manipulation sequences [119]–[123]. On the data side, learning from large-scale egocentric and third-person human videos bridges the embodiment gap between web-sourced data and robot demonstrations [64], [124]–[127]. On the embodiment side, cross-platform transfer [128]–[130] and morphology-specific architectures for whole-body locomotion [131], mobile manipulation [132], bimanual coordination [133], and dexterous hands [134], [135] extend VLAs well beyond single-arm table-top settings. Task-level generalization has been pursued through visual prompting interfaces [136]–[138], handling of dynamic objects and task switching [139]–[141], richer instruction following [142]–[144], and hierarchical planning for long-horizon goals [145]–[151].

Although imitation learning dominates VLA training, its performance ceiling is bounded by the quality and coverage of demonstrations. Reinforcement learning offers a principled route to surpass this ceiling through reward-verified fine-tuning in simulation [152]–[157], test-time verification and best-of-N sampling [158]–[162], and systematic empirical studies that characterize when and how RL enhances VLA generalization [20], [163]–[168]. Complementary adaptation strategies that steer pretrained VLAs toward novel scenarios without full retraining have also shown promise [169]–[172].

Most relevant to our work, a growing body of VLA methods enriches the action policy with a predictive world model. These approaches learn next-state or next-frame predictions as auxiliary objectives [173]–[181], close the loop between world-model imagination and policy improvement [156], [182]–[186], or decouple intent prediction from action execution [187], [188]. Recent diagnostics also test whether action-conditioned world models faithfully follow the supplied actions [189]. While these methods demonstrate that world knowledge benefits action prediction, they typically *jointly* train the predictive and control components, coupling their representational capacity and optimization dynamics. In contrast, our framework first trains a world model in isolation to master state-transition dynamics, then distills the learned representations into a lightweight action policy—allowing each module to specialize without mutual interference.

B. World Action Models

The finding that video generation models trained on internet-scale data implicitly capture physical dynamics has spurred a family of methods that employ video prediction as a planning substrate. A recent survey organizes this expanding world-action-model design space [190]. Pioneering work demonstrated that text-conditioned video synthesis can serve as a universal policy planner [9], [10], and subsequent methods extracted control signals from actionless video via dense correspondences [191] or leveraged generated human-hand videos to guide robot execution [192], [193]. Scaling this idea, large video foundation models have been pretrained on web-scale data and then adapted for manipulation [7], [8], [11], [194]–[197], with extensions to long-horizon video generation [180], [198], task-relevant attention specialization [199], and multi-modal video–action co-generation [200], [201].

A key limitation of cascaded video-then-action pipelines is that errors in the predicted video compound before reaching the action decoder. Joint world-action models (WAMs) address this by simultaneously denoising or generating video and actions within a shared architecture, enabling bidirectional information flow between predictive and control representations [4]–[6], [202]–[207]. This paradigm has been extended to unified autoregressive formulations [22], [208]–[210], action-centered architectures [211], [212], implicit value-based planning [149], [213], [214], future representation alignment [215]–[217], and multi-modal world modeling that incorporates tactile feedback [218]. Recent results explore zero-shot policies and cross-embodiment transfer [219], [220], while predictive visual representations can support generalist control [33].

Operating in pixel space, however, is computationally expensive and entangles visual appearance with the underlying dynamics. Latent action pretraining [18], [103], [221] addresses this by learning a compact action space from unlabeled video, enabling large-scale VLA training without paired action labels. This principle has been extended to adaptable latent world models [23], [130], [222], [223], billion-parameter latent dynamics modules [224], and skill-compositional post-training [225]–[229]. Concurrently, 4D spatiotemporal world models [230]–[232] capture geometric evolution over time. A growing line of work also addresses computational and deployment challenges: some investigate whether test-time imagination is always necessary or can be bypassed via distilled shortcuts [12], [13], [112], [162], [233]–[236], others explore online adaptation to changing conditions [172], [237], [238], and recent efforts enrich WAM representations through object-level structure [113], [239], intent-aware spatial value maps [240], [241], parameter-efficient cross-model knowledge transfer [242], and system-level integration of multiple predictive components [243], [244]. ActiveWAM further couples observation selection with world-action modeling [241].

Recent surveys summarize the broader language-conditioned manipulation and world-model landscape [245], [246]. Despite the rapid progress in both VLA and WAM communities, many approaches couple world modeling and action generation through shared representation learning or inference-time world-model computation. Our work challenges this premise. By first training a world model to learn rich, transferable state-transition dynamics and then distilling this knowledge into a dedicated action policy, we allow each module to specialize in its own objective, yielding a decoupled framework that combines the predictive power of world models with the deployment efficiency of lightweight VLA policies.

APPENDIX B
WORLD MODEL DETAILS

A. Architecture

The backbone is built on LTX-Video [14], a 2B-parameter Diffusion Transformer (DiT) pretrained for video generation. In our implementation, the video VAE provides $16\times$ spatial downsampling and $8\times$ temporal downsampling, substantially reducing the token count for efficient training and inference. Given a language instruction and the current multi-view observation, the backbone predicts a future clip at 256×256 resolution over a horizon of $K=24$ frames using a flow-matching objective in the VAE latent space.

B. Training Data and Configuration

The video diffusion backbone is adapted on a mixture of robot manipulation videos and egocentric human activity data:

- *Robot videos*: Open X-Embodiment [35], AgiBot-World [37], Galaxea [247], RoboMIND [248], and RoboCOIN [249], collectively spanning single-arm and bimanual platforms with diverse end-effectors, camera viewpoints, and task categories.
- *Human videos*: Ego4D [250] and Humanoid Everyday [251], providing rich egocentric interaction patterns including object manipulation, tool use, and dexterous hand activities.

The combined video corpus covers a broad spectrum of embodiments, interaction dynamics, and visual environments, enabling the backbone to internalize generalizable spatiotemporal priors. During DAWN training, the video backbone is kept entirely frozen. The Qwen3-VL backbone, the MLP action head, and the distillation module are jointly optimized, while the frozen teacher provides stable spatiotemporal targets.

For downstream evaluation, the action model is fine-tuned on benchmark-specific demonstration data: LIBERO training demonstrations for both LIBERO and LIBERO-Plus evaluation, and the RoboTwin 2.0 clean-setting demonstrations for RoboTwin 2.0 evaluation.

APPENDIX C
DAWN ARCHITECTURE AND TRAINING DETAILS

A. Model Architecture

Table VII summarizes the architecture of each component in DAWN. The configuration shown corresponds to the

RoboTwin 2.0 training setup.

TABLE VII
MODEL ARCHITECTURE OVERVIEW.

Component	Specification	Parameters
VLM backbone	Qwen3-VL-4B-Instruct	4B
Hidden dimension	2048	—
Observation resolution	224×224	—
Action head	MLP	—
Hidden size	1024	—
Action chunk length	50	—
Action dimension	14	—
World model backbone	LTX-Video DiT (frozen)	2B
Num layers	28	—
Input resolution	256×256	—
Future horizon	24 frames	—
Appearance pathway	InstructPix2Pix + LoRA	1B
Feature decoder	Conv2d + GroupNorm + Conv2d	—

B. Training Hyperparameters

Table VIII lists the training hyperparameters used for DAWN on the RoboTwin 2.0 benchmark.

TABLE VIII
TRAINING HYPERPARAMETERS.

Hyperparameter	Value
Optimizer	AdamW
Learning rate (VLM backbone)	1×10^{-5}
Learning rate (action head)	1×10^{-4}
Learning rate (appearance pathway)	5×10^{-6}
LR scheduler	Cosine (min LR 5×10^{-7})
Warmup steps	5,000
Batch size (per device)	16
Gradient accumulation steps	6
Distributed strategy	DeepSpeed ZeRO-2
GPUs	2× NVIDIA H200

APPENDIX D
REAL-WORLD EXPERIMENTAL SETUP

Figure 6 shows the two real-world robot embodiments used in our physical experiments, including the head-mounted and wrist-mounted RGB camera views. Figure 7 summarizes the initial and goal configurations for the six manipulation tasks evaluated in the real world.

APPENDIX E
QUALITATIVE VISUALIZATION

A. Simulation Rollout Visualization

Figure 8 presents representative successful rollouts of DAWN across all three simulation benchmarks: LIBERO, LIBERO-Plus, and RoboTwin 2.0. These qualitative examples illustrate the diversity of manipulation skills acquired by our policy, including precise spatial reasoning, robust generalization under domain perturbations, and coordinated bimanual control over varying task horizons. Figure 9 shows a representative failure case on RoboTwin 2.0, where the policy fails to coordinate simultaneous dual-arm bottle grasping, highlighting a remaining challenge in synchronized bimanual control.

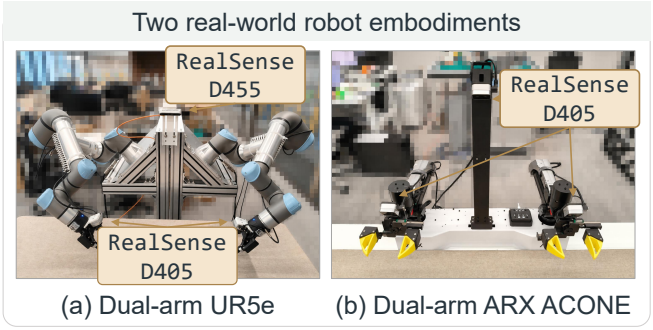


Fig. 6. **Real-world robot embodiments.** Left: the dual-arm UR5e setup with a head-mounted RealSense D455 and wrist-mounted RealSense D405 cameras. Right: the dual-arm ARX ACONE setup with the same multi-view camera configuration.

B. Real-World Rollout Visualization

Figure 10 presents representative successful real-world rollouts across the six physical tasks. Figure 11 shows a representative failure case on towel folding, where the policy makes partial progress but does not complete the final fold.

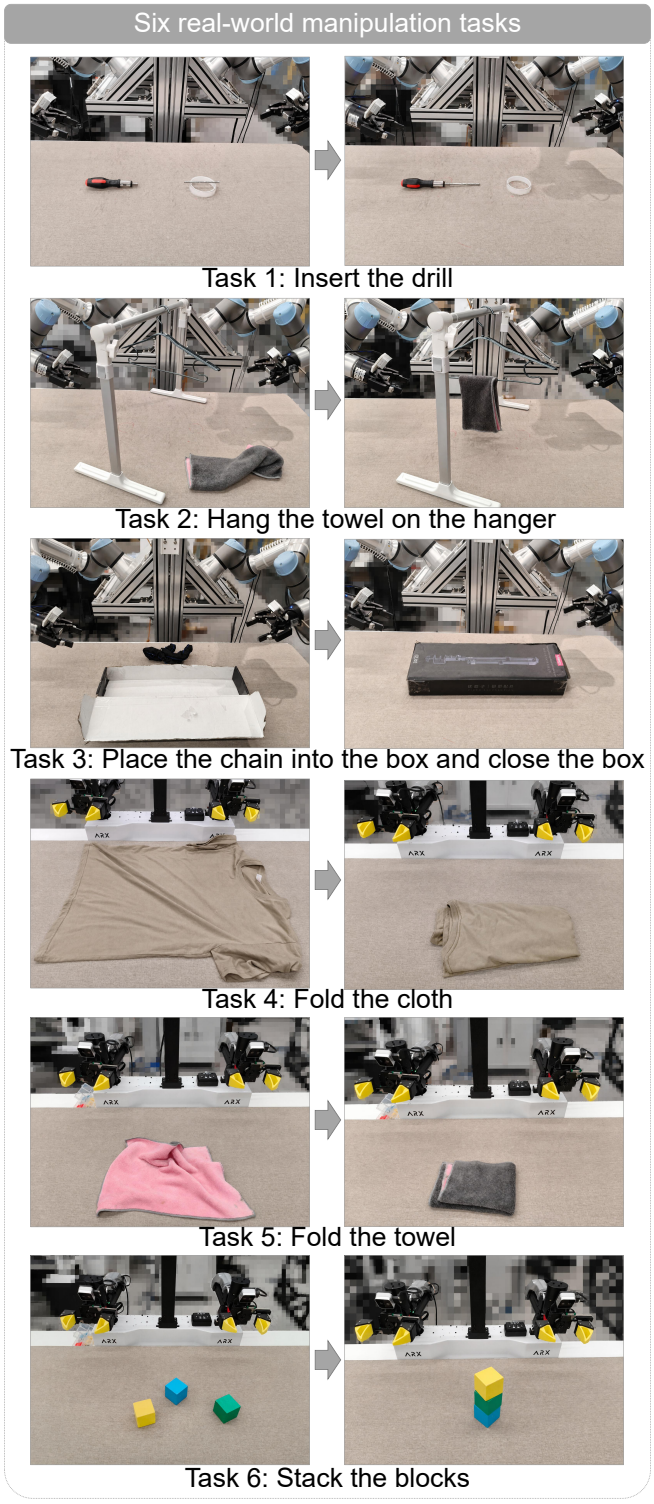


Fig. 7. **Real-world manipulation tasks.** Initial and goal configurations for six tasks: Insert the Drill, Hang the Towel on the Hanger, Place the Chain into the Box and Close the Box, Fold the Cloth, Fold the Towel, and Stack the Blocks.

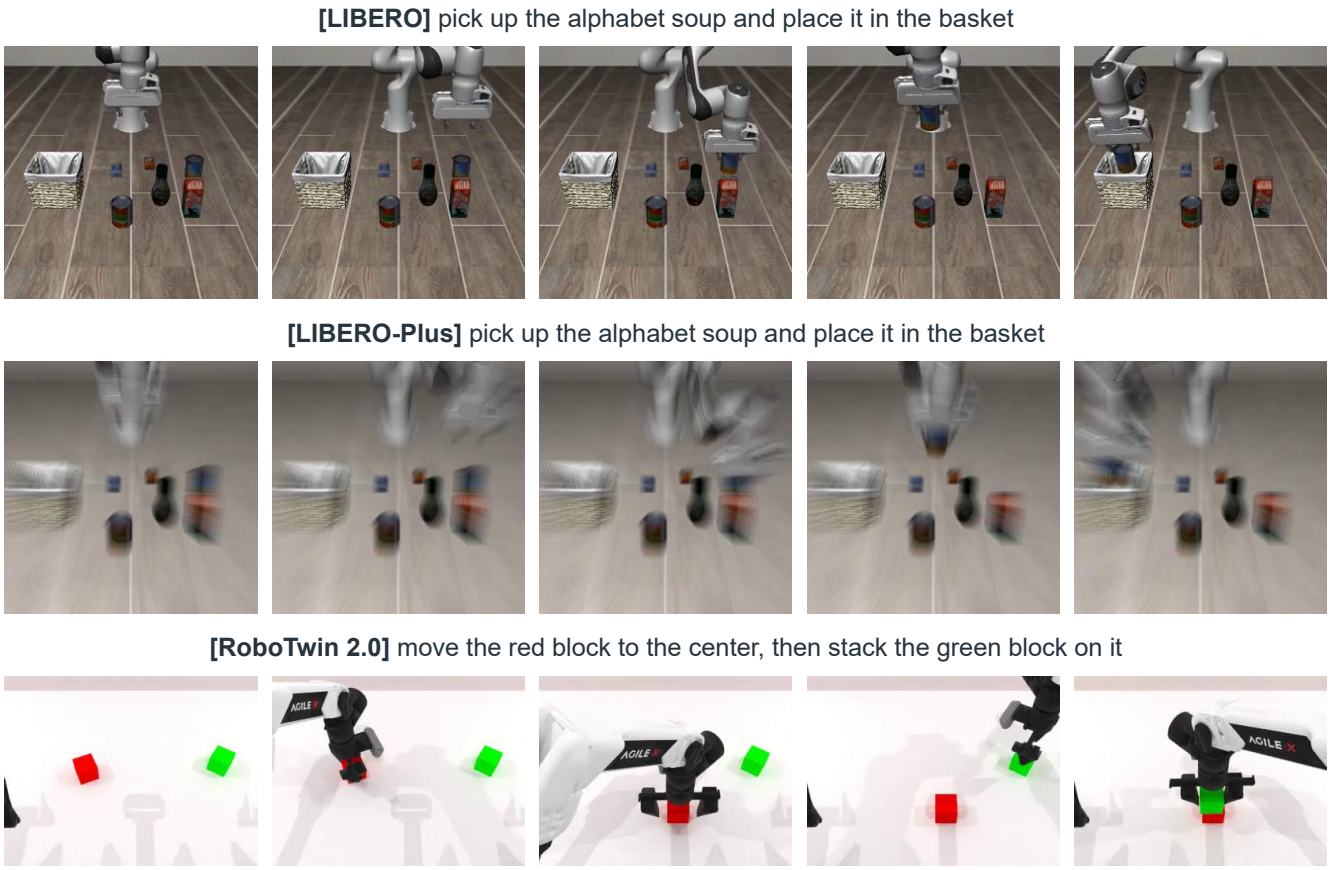


Fig. 8. **Qualitative visualization of successful rollouts on three benchmarks.** Each row shows keyframe sequences from representative tasks in LIBERO, LIBERO-Plus, and RoboTwin 2.0, demonstrating the diverse manipulation capabilities of DAWN.



Fig. 9. **Representative simulation failure case on RoboTwin 2.0.** A bimanual task in which the policy fails to coordinate simultaneous dual-arm grasping.



Fig. 10. **Representative successful real-world rollouts.** Keyframe sequences from six physical manipulation tasks demonstrate that DAWN transfers to insertion, hanging, box closing, deformable-object folding, and block stacking across the two real-world embodiments.



Fig. 11. **Representative real-world failure case.** A towel-folding rollout in which the policy grasps and moves the towel but does not complete the final fold, illustrating a remaining failure mode under deformable-object manipulation.