WSA_1: A 3D-CENTRIC WORLD-SPATIAL-ACTION MODEL FOR GENERALIZABLE ROBOT CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

World-Action Models (WAMs) have emerged as a promising paradigm for robot control by using predicted world evolution to guide action generation. However, existing WAMs largely follow a one-way world-to-action paradigm. Real-world interaction, in contrast, is inherently driven by bidirectional temporal causality with implicit physics: the evolving world informs what actions should be taken, while the executed actions, in turn, determine how the world evolves. Without explicitly modeling this action-to-world direction, WAMs cannot assess whether the consequences of their generated actions align with the world evolution they predict, preventing WAMs from capturing consistent interaction dynamics from video demonstrations and thereby constraining their generalization capabilities in real-world scenarios. In this paper, we introduce **WSA_1**, a new modeling paradigm for generalizable robot control. WSA_1 develops a World-Spatial-Action understanding by jointly modeling three interconnected processes: how actions transform the 3D world (world level), how spatial states evolve toward desired outcomes (spatial level), and how target 3D world guide appropriate robot behaviors (action level). By integrating these processes, WSA_1 enables more coherent and tightly coupled reasoning over embodied interactions. Compared with prior WAMs trained on tens of thousands of hours of real-robot data, WSA_1 leverages 6k hours of video demonstrations (only 1k from real robot), demonstrating strong data efficiency. Moreover, even our 3B-parameter WSA_1 surpasses larger WAMs with 6B–8B parameters, highlighting its parameter efficiency. Experiments show that WSA_1 achieves a 93.3% success rate on RoboTwin2.0, delivering an average 20% improvement over state-of-the-art models on real-world robot control tasks.

1 INTRODUCTION

The pursuit of generalist robot policies has driven rapid progress in embodied foundation models, which seek to equip robots with transferable capabilities for interacting with diverse physical environments. Among recent approaches, World-Action Models (WAMs) (Bi et al., 2026; Pai et al., 2025; Ye et al., 2026; Kim et al., 2026) have emerged as a promising paradigm for robot control. Rather than generating actions solely from the current observation and task instruction, WAMs explicitly predict how the world may evolve and use the predicted future to guide action generation. Such predictive modeling provides robots with an intermediate notion of "what should happen next" before deciding "what to do next", offering a potentially powerful basis for generalization beyond the training distribution.

However, existing WAMs largely formulate this relationship in only one direction: *world-to-action*. Predicted world evolution provides guidance for generating robot actions, yet the reverse direction—how those actions transform the world—is not explicitly modeled within the same interaction process. This creates an important disconnect. In real-world manipulation, world evolution and robot action are intrinsically coupled: the evolving world constrains which actions are appropriate, while the executed actions determine the subsequent world state. A model that captures only the former can generate actions according to a predicted future, but lacks an explicit mechanism for learning whether those actions would actually induce a world transition consistent with that future. As a result, world prediction and action generation are learned as only partially coupled components rather than two complementary views of the same physical interaction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

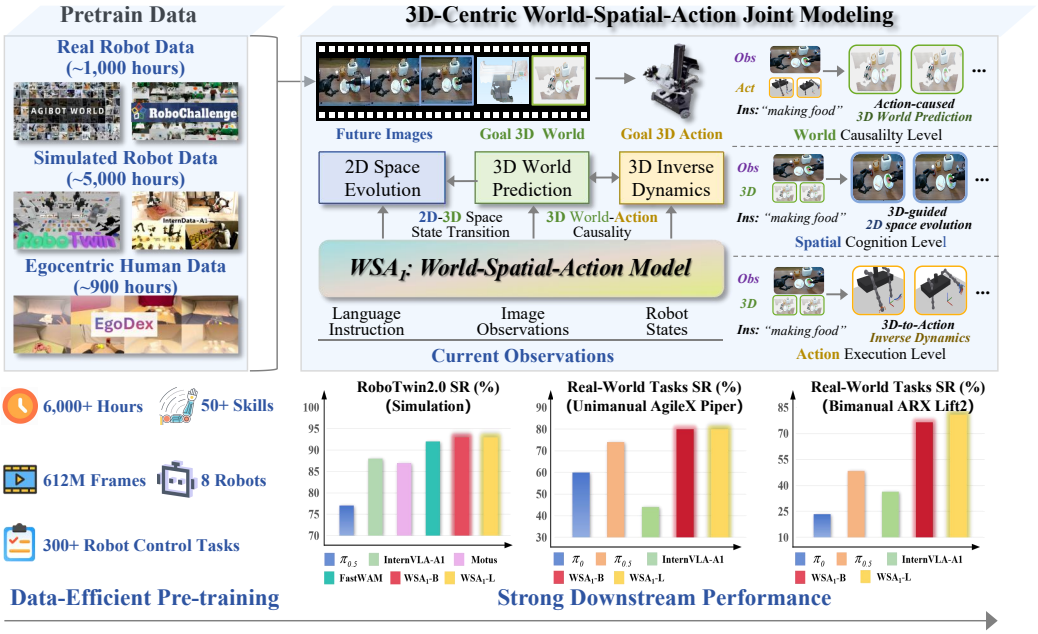

Figure 1: **WSA**$_1$: A generalizable robot foundation model based on 3D-centric world-spatial-action joint modeling, which learns bidirectional world-action physical commonsense from real-world interactions. It achieves strong performance across diverse simulated and real-robot benchmarks using 6K hours of pre-training data (only 1k from real robot), demonstrating strong data efficiency.

This observation motivates a different perspective on World-Action modeling: effective embodied learning should capture not only how a future world guides actions, but also how actions give rise to that future world. Importantly, connecting these two directions requires more than independently learning a forward world model and an inverse action model. Robot interaction is ultimately realized through changes in 3D spatial configurations: objects move, contacts change, and spatial relations evolve as actions are executed. Such spatial evolution provides a common representation that connects an action with the world transformation it produces. We therefore argue that *world evolution, spatial evolution, and action generation should be modeled jointly*, rather than treating world prediction merely as an auxiliary conditioning signal for control.

Based on this principle, we introduce **WSA**$_1$, a World-Spatial-Action modeling paradigm for generalizable robot control. WSA$_1$ jointly captures three interconnected processes. At the *world level*, it models how robot actions transform the 3D world, learning the target world state induced by an interaction. At the *spatial level*, it models how spatial states evolve from the current observation toward the target 3D world, providing a geometric connection between action-induced changes and their resulting outcomes. At the *action level*, it models how the target 3D world guides appropriate robot behaviors, learning the inverse mapping from desired world evolution to executable actions. Together, these processes connect action-to-world and world-to-action modeling through a shared spatial representation, allowing WSA$_1$ to learn more coherent world-action interaction dynamics from video demonstrations.

We instantiate WSA$_1$ as a unified embodied foundation model based on a Mixture-of-Transformers architecture, where complementary modeling experts jointly learn world evolution, spatial transformation, and action generation. Rather than relying heavily on large-scale real-robot supervision, WSA$_1$ is trained on 6,000 hours of heterogeneous video demonstrations, of which only 1,000 hours are collected from real robots. WSA$_1$ achieves a 93.3% success rate on RoboTwin2.0 and improves over state-of-the-art approaches by an average of 20% on real-world robot control tasks. Moreover, our 3B-parameter model surpasses existing WAMs with 6B–8B parameters, demonstrating strong data and parameter efficiency. In summary, our main contributions are three-fold:

- **World-Spatial-Action Modeling.** We introduce WSA_1, a 3D-centric world-action bidirectional learning paradigm that jointly models action-to-world evolution, spatial transformation, and world-to-action generation.
- **Unified Embodied Foundation Model.** We instantiate WSA_1 with a Mixture-of-Transformers architecture that jointly optimizes world, spatial, and action modeling across two model scales.
- **Efficient Generalization.** With 6,000 hours of demonstrations, including only 1,000 hours of real-robot data, WSA_1 achieves 93.3% success on RoboTwin2.0, improves real-world performance by 20% on average, and enables the 3B model to outperform larger 6B–8B WAMs.

2 Related Work

Embodied foundation models aim to transfer the broad visual and semantic knowledge of multi-modal foundation models (Chen et al., 2024; Bai et al., 2025b; Team Wan et al., 2025) to robot perception and control (Zhang et al., 2025a; Yu et al., 2025). Existing approaches can be broadly grouped into Vision-Language-Action (VLA) models and World-Action Models (WAMs).

Vision-Language-Action Models. VLA models (Wen et al., 2025; Black et al., 2025b;a; Yang et al., 2025; Zhao et al., 2025) leverage pretrained vision-language models and adapt their semantic priors to robot control through demonstration learning. Early methods such as RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2025b) formulate action generation as autoregressive token prediction, while recent approaches (Black et al., 2025b;a; Bjorck et al., 2025) adopt continuous generative objectives such as flow matching for more precise control. Several works further incorporate explicit 3D information, including depth and 3D representations, to enhance spatial reasoning (Zhen et al., 2024; Qu et al., 2025). Despite these advances, VLA models primarily focus on learning observation-to-action mappings, whereas explicitly modeling how robot actions transform the future 3D world remains less explored.

World-Action Models. WAMs (Pai et al., 2025; Ye et al., 2026; Kim et al., 2026; Yuan et al., 2026b; Lou et al., 2026; Bi et al., 2026) incorporate predictive world modeling into robot control by jointly learning future world evolution and action generation, often leveraging pretrained video models as world priors. Motus (Bi et al., 2026) demonstrates the benefit of future prediction for manipulation generalization, while Fast-WAM (Yuan et al., 2026b) shows that such gains can also arise from joint world-action learning without explicit video generation. More recent approaches extend world prediction to 3D geometry (Qian et al., 2026; Zhang et al., 2025b; Li et al., 2026b; Huang et al., 2026), enabling robot policies to exploit richer spatial information during manipulation. However, existing WAMs predominantly model how predicted world evolution guides action generation, while the reverse direction—how robot actions induce corresponding changes in the 3D world—remains less explicitly modeled.

Compared with prior VLA and WAM approaches, our WSA_1 goes beyond direct observation-to-action mapping and one-way world-to-action prediction by additionally modeling how robot actions transform the 3D world. It further connects these two directions through spatial-state evolution, jointly modeling action-to-world, spatial, and world-to-action dynamics within a unified framework.

3 Methodology

3.1 Problem Definition and Notation

In the context of robot control, the goal of embodied foundation model is to learn a robust policy that predicts next robot actions from current observations. Formally, at each timestep t, the observations include task instruction termed l, visual observation denoted as v_t, and robot proprioception termed s_t. In general, the robot action is a H-step chunk termed $A_t := (a_{t+1}, \ldots, a_{t+H})$, a short-term action trajectory with a sequence of end-effector pose or joints of a robot in next H timesteps. To train a robot policy π_θ parameterized by θ on diverse video demonstrations $\mathcal{D}_{\text{dem}}$, it is required to maximize the standard likelihood objective:

$$\max_{\theta} \mathbb{E}_{(v_t, s_t, A_t, l) \sim \mathcal{D}_{\text{dem}}} \log \pi_\theta(A_t \mid v_t, s_t, l). \tag{1}$$

Previous works formulate robot control problem as vision-language-action or world-action policy, which can be regarded as a joint distribution derived from the standard objective. Different from

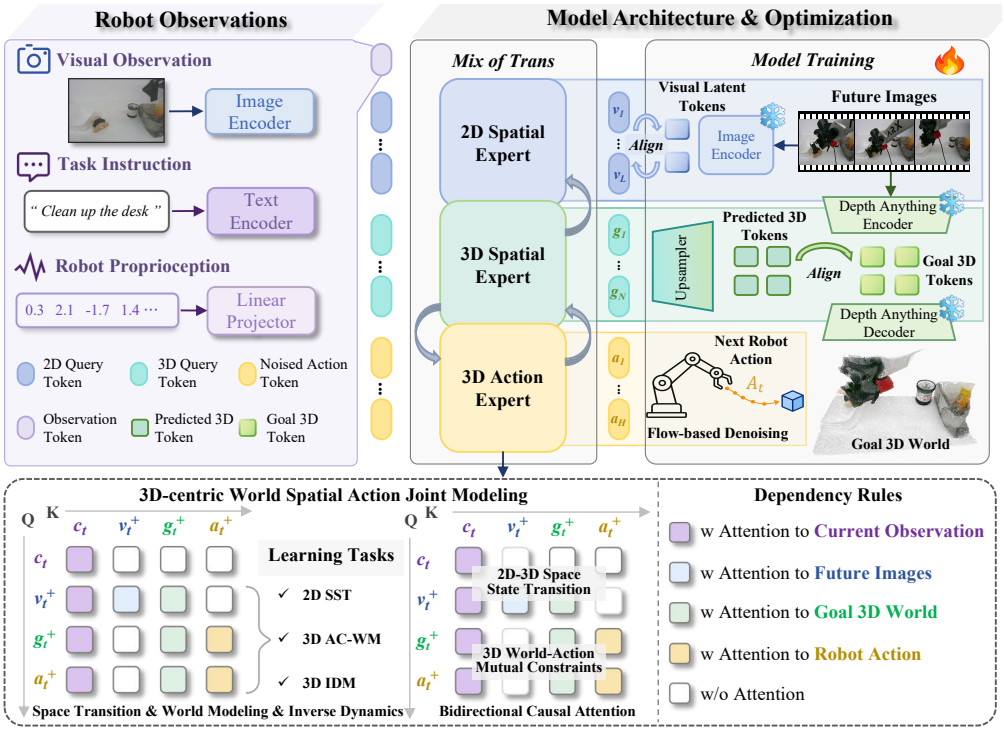

Figure 2: Overview of WSA_1. It adopts a Mixture-of-Transformers structure with three complementary experts: a 2D Spatial Expert for 2D-3D space state transition, a 3D Spatial Expert for action-caused 3D world prediction, and a 3D Action Expert for 3D inverse dynamics modeling. A bidirectional causal attention mechanism enforces dependency rules to integrate world–action mutual constraints within a shared latent space.

previous VLA and WAM paradigms that rely on unidirectional mappings, we formulate robot control as world-spatial-action joint learning that captures the interdependencies among 3D world dynamics, visual space cognition, and robot actions. Specifically, it jointly learns three complementary conditional distributions that impose mutual constraints among 3D world model, 2D space transition, and 3D inverse dynamics model:

$$\underbrace{p(g_{t+1:t+K} \mid a_{t+1:t+H}, \mathbf{O}_t)}_{\text{Action-caused 3D World Model}}, \quad \underbrace{p(v_{t+1:t+N} \mid g_{t+1:t+K}, \mathbf{O}_t)}_{\text{3D-guided 2D Space Transition}}, \quad \underbrace{p(a_{t+1:t+H} \mid g_{t+1:t+K}, \mathbf{O}_t)}_{\text{3D Inverse Dynamics Model}} \tag{2}$$

where $\mathbf{O}_t := \{v_t, s_t, l\}$ denotes robot current observation, $\{g_{t+1}, \ldots, g_{t+K}\}$ are target 3D world representations in near future, and $\{v_{t+1}, \ldots, v_{t+N}\}$ represent visual states of images in near future. These three distributions correspond to the three levels of the WSA framework: the world causality level (predicting how actions shape the 3D world), the spatial cognition level (grounding visual predictions in 3D geometry), and the action execution level (generating motor commands from 3D world states).

3.2 Model Architecture

Inspired by the recent works (Bi et al., 2026; Black et al., 2025a), WSA_1 is built upon a mixture-of-transformer (MoT) architecture to enable seamless integration of 2D-3D space transition, 3D world prediction and robot action generation, involving 2D spatial expert in parallel with 3D spatial expert and 3D action expert.

2D Spatial Expert (2D-SE). Given a human instruction, generating 3D goal-directed state transition in 2D space enhances understanding of world dynamics. The 2D-SE $f_{2D}(\cdot)$ serves as the 2D-3D state transition branch of WSA_1. It is initialized from a pre-trained vision-language model, thereby inheriting strong vision-language priors. The goal of 2D-SE is to predict future 2D visual dynamics

to reach the target state of the 3D world. Specifically, given goal states $G_t := (g_{t+1}, \ldots, g_{t+K})$ of 3D world, it predicts visual tokens $h_v = f_{2D}(G_t, O_t)$ of those images in near future, instead of RGB pixels. This design allows the model to focus on the essential visual dynamics rather than temporal variations of redundant pixels.

3D Spatial Expert (3D-SE). Given a sequence of executable robot actions, predicting target world in 3D space reflects causal effect of the robot behavior. Under this context, the 3D-SE $f_{3D}(\cdot)$ is designed as an internal mechanism for 3D causal world modeling. Instead of directly construct native 3D representations such as 3D Gaussian or 3D point cloud, the 3D-SE predicts latent representations $h_g = f_{3D}(O_t, A_t)$ of target 3D world, which are derived from 3D foundation models such as DepthAnything (Lin et al., 2025). We instantiate 3D-SE with a lightweight transformer model of the same depth as 2D-SE, thereby the architecture of 3D-SE can be tightly coupled with 2D-SE for space state transition.

3D Action Expert (3D-AE). Given target 3D world, predicting next robot action tells why and how robot to behave in 3D space. The 3D-AE $f_{act}(\cdot)$ aims to predict a full action chunk conditioned on both current observations and future 3D latent representations. Specifically, we instantiate 3D-AE as denoising diffusion transformer. Given noisy actions $\hat{A}_t = (1-\tau)A_t + \tau\omega$, where $\omega \sim \mathcal{N}(0, I)$ and $\tau \in [0, 1]$, it employs iterative cross-attention and denoising procedure to predict latent action representations $h_{act} = f_{act}(O_t, G_t, \hat{A}_t)$. This yields a 3D inverse dynamics formulation, where 3D-AE maps predicted 3D latent world representations back to executable action.

World-Spatial-Action Causal Attention. Three experts are coupled within the mixture-of-transformer architecture via a shared attention mechanism. Formally, let h_t denote tokens of current observation, which are extracted via 2D-SE. As illustrated in Figure 2, we adopt a 3D-centric causal attention mechanism, which defines the following dependency rules: (1) All predicted tokens (h_v, h_g, h_{act}) can attend to the current observation tokens h_t, ensuring all predictions are grounded in the current state of the robot and the environment. (2) The visual tokens of future subgoal images h_v can attend to the 3D scene tokens h_g, encouraging the model to learn 2D-3D visual consistency and ground semantic visual predictions in physically consistent 3D geometry. (3) The 3D scene tokens h_g and action tokens h_{act} have full bidirectional attention access to each other, enabling joint learning of action-caused 3D world prediction and 3D goal-directed action generation. This bidirectional attention is the key to modeling world-action interdependence of the WSA paradigm, as it reflects a causal inductive bias that drives the model to capture how actions shape world evolution and how world states inform action generation, rather than merely learning correlational mappings.

Based on these four components, we implement two variants of WSA_1, where their 2D-SE module is equipped with two different pretrained models and 3D-SE/AE is randomly initialized: WSA_1-B, a base model initialized from the pretrained Qwen3-VL-2B (Bai et al., 2025a) with a total of 3B parameters; and WSA_1-L, a large variant built on the pretrained Wan2.2-5B (Team Wan et al., 2025) with a total of 6B parameters.

3.3 World-Spatial-Action Joint Learning

Analogous to the internal model theory in cognitive neuroscience (Wolpert & Ghahramani, 2000; Clark, 2013), the WSA paradigm is designed to mirror three core computational principles of the human brain: 1) 3D spatial grounding of perception, 2) forward predictive modeling of action consequences, and 3) inverse motor control for goal-directed behavior. These three principles correspond to following three levels of the WSA learning framework, respectively.

2D Space State Transition (2D SST). Analogous to the brain's ability to form mental imagery of future scenes (Kosslyn, 1996), the 2D Spatial Expert learns to predict 3D-grounded visual state transition. To this end, we adopt a MSE loss over the predicted visual tokens of future images, to minimize the discrepancy between the predicted image tokens and the ground-truth tokens extracted via pre-trained tokenizer. Formally, this transition learning objective is defined as:

$$\mathcal{L}_{\text{2D}} = \mathbb{E}_{(O_t, G_t, V_t)\sim\mathcal{D}_{dem}} \left\| f_{2D}(O_t, G_t) - f_{enc}(V_t) \right\|_2^2, \tag{3}$$

where $f_{enc}(\cdot)$ denotes a pre-trained VAE-based tokenizer (e.g., COSMOS (Agarwal et al., 2025)).

Action-Caused 3D World Modeling (3D AC-WM). We adopt MSE loss over the predicted 3D world latent representations, to minimize the error between the predicted future 3D states and the

ground-truth 3D geometric states on the demonstration data. This 3D world model learning objective can be formalized as:

$$\mathcal{L}_{\text{3D}} = \mathbb{E}_{(O_t, A_t, G_t)\sim\mathcal{D}_{dem}} \| f_{3D}(O_t, A_t) - f_g(G_t) \|_2^2 , \tag{4}$$

where $f_g(\cdot)$ denotes a pre-trained 3D foundation model, e.g., Depth-Anything (Lin et al., 2025). This corresponds to the forward internal model in motor control theory (Wolpert & Ghahramani, 2000), which predicts how motor actions will alter the state of the environment.

3D Inverse Dynamics Modeling (3D IDM). Following previous works (Qian et al., 2026; Black et al., 2025a), we adopt flow matching loss (Lipman et al., 2022) for continuous action generation. This loss function aims to minimize the error between the predicted flow field and the ground-truth one, conditioned on the current observation and the goal 3D world. This learning process can be formulated as:

$$\mathcal{L}_{\text{ACT}} = \mathbb{E}_{(O_t, G_t, A_t)\sim\mathcal{D}_{dem}} \left\| \omega - A_t - f_{act}(O_t, G_t, \hat{A}_t) \right\|_2^2 . \tag{5}$$

This trains 3D-AE to predict flow field $(\omega - A_t)$, thereby facilitating coherent action generation. The overall learning objective is the sum of the three losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{ACT}}. \tag{6}$$

Table 1: Data mixture and training objectives used for pretraining. A checkmark indicates that the corresponding loss is enabled, while "–" indicates that the loss is masked out.

Data Type	**Num. frames**	**Num. tasks**	**Sampling weight**	$\mathcal{L}_{\text{2D}}$	$\mathcal{L}_{\text{3D}}$	$\mathcal{L}_{\text{ACT}}$
Simulation Data	424M	70	0.54	✓	✓	✓
Egocentric Human Data	89M	194	0.10	✓	✓	–
Real Robot Data	99M	36	0.36	✓	✓	✓

3.4 Pre-train Data

High-quality embodied data is the cornerstone of pre-training generalizable embodied foundation models that can acquire versatile manipulation skills for flexible real-world robot control. Existing robot foundation models (Black et al., 2025b; Li et al., 2026a) typically rely on tens of thousands of hours of robot demonstration data to achieve broad generalization, with real-robot robot data constituting a major portion of their pre-training corpora. However, as demonstrated in recent works (Bjorck et al., 2025; Yin et al., 2026), data diversity is the key to robot foundation model pre-training. Accordingly, we carefully collect a pre-train dataset with only 6,000 hours of demonstration videos from a mixture of heterogeneous data sources, including simulation data (Tian et al., 2025), egocentric human data (Hoque et al., 2025) and real robot data (Bu et al., 2025; Yakefu et al., 2025). As illustrated in Table1, the mixture of data sources covers 8 different robot embodiments and 300+ robot control tasks, including 50+ atomic manipulation skills such as grasping and placing. Since there exists embodiment gap between human and robot, we do not apply action generation loss $\mathcal{L}_{\text{ACT}}$ to WSA models when pre-training on human data.

4 Experiments

We evaluate WSA_1 in both simulation and real-world settings, focusing on generalization, the contribution of World-Spatial-Action joint modeling, and data efficiency. We first investigate the merits of WSA modeling in simulation, and then examine whether the merits observed in simulation are consistent to that of the real robot setting. Together, these experiments provide a comprehensive validation of whether 3D-centric bidirectional world-spatial-action modeling offers a practical and data-efficient pathway to generalizable robot control. More results and analysis are in Appendix.

4.1 Simulation Experiments

Benchmarks. We evaluate on RoboTwin2.0 (Chen et al., 2025), which contains 50 bimanual manipulation tasks under clean and randomized settings, and LIBERO (Liu et al., 2023), covering

Table 2: Comparison on RoboTwin2.0 and LIBERO benchmarks. We report the mean success rate (SR) and the best results are shown in **bold**.

Method	Size	RoboTwin2.0			LIBERO			
		Clean	Rand.	Avg.	Spatial	Object	Goal	Long
Vision–Language–Action (VLA) models								
OpenVLA (Kim et al., 2025b)	7B	–	–	–	84.7	88.4	79.2	53.7
OpenVLA-OFT (Kim et al., 2025a)	7B	–	–	–	97.6	98.4	97.9	94.5
π_0 (Black et al., 2025b)	3B	65.9	58.4	62.2	96.8	98.8	95.8	85.2
$\pi_{0.5}$ (Black et al., 2025a)	3B	82.7	76.8	79.8	98.8	98.2	98.0	92.4
X-VLA (Zheng et al., 2026)	1B	72.8	72.8	72.8	98.2	98.6	97.8	97.6
InternVLA-A1 (Cai et al., 2026)	3B	89.4	89.6	89.5	–	–	–	–
GR00T N1 (Bjorck et al., 2025)	2B	–	–	–	95.6	97.6	94.2	89.6
SpatialVLA (Qu et al., 2025)	4B	–	–	–	88.2	89.9	78.6	55.5
HoloBrain-0-QW (Lin et al., 2026b)	1B	91.9	92.3	92.1	97.2	99.6	97.6	95.2
Pose-VLA (RGB) (Lin et al., 2026a)	3B	79.9	79.1	79.5	96.5	98.0	97.1	92.4
ABot-M0 (Yang et al., 2026b)	4B	86.1	85.1	85.6	98.8	99.8	**99.0**	96.6
GeoPredict (Qian et al., 2026)	–	–	–	–	98.0	98.2	95.7	94.0
World–Action (WAM) models								
Motus (Bi et al., 2026)	8B	88.7	87.0	87.8	96.8	99.8	96.6	97.6
4D-WAM (Yang et al., 2026a)	6B	91.9	90.8	91.3	98.8	99.5	98.6	97.4
Fast-WAM (Yuan et al., 2026b)	6B	91.9	91.8	91.8	98.2	**100.0**	97.0	95.2
LingBot-VA (Li et al., 2026a)	5B	92.9	91.6	92.2	98.5	99.6	97.2	**98.5**
World–Spatial–Action (WSA) models								
WSA_1-B (Ours)	3B	92.2	92.7	92.4	98.6	99.6	97.2	94.2
WSA_1-L (Ours)	6B	**93.5**	**93.1**	**93.3**	**99.4**	99.8	98.0	95.6

spatial, object, goal-conditioned, and long-horizon manipulation. The randomized setting introduces scene-level distribution shifts and is used to evaluate out-of-distribution (OOD) generalization.

Evaluation Protocol. We adopt multi-task evaluation with success rate (SR) as the metric. Unless otherwise stated, models are fine-tuned on all 27,500 RoboTwin2.0 demonstrations. For scene generalization (OOD-C2R), models are trained only on clean demonstrations and tested in randomized scenes. For zero-shot cross-embodiment generalization (OOD-CE), the trained AgileX embodiment is replaced by an unseen ARX-X5 at test time. By default, WSA_1-B without pre-training is adopted to conduct ablation studies for fast investigation.

Main Results on Simulation Benchmarks. As shown in Table 2, WSA_1-L attains 93.5%, 93.1% and 93.3% scores on RoboTwin-2.0 in the clean mode, in the randomized mode and on average, which is the highest average among all entries and 1.1 points above the strongest baseline, LingBot-VA (92.2%). The gap between the clean and the randomized mode is 0.4 points for WSA_1-L and 0.5 points for WSA_1-B, while it amounts to 7.5 points for π_0 and 5.9 points for $\pi_{0.5}$. WSA_1-B makes the comparison favourable on the parameter-efficiency side as well, since at 3B parameters it surpasses the 8B Motus (87.8%), the 6B 4D-WAM (91.3%) and the 6B Fast-WAM (91.8%). On LIBERO, the comparison is less discriminative, because the strongest baselines already operate close to saturation. Averaged over the four suites, WSA_1-L reaches 98.2% and WSA_1-B reaches 97.4%, both are competitive with 8B Motus (97.7%), 6B 4D-WAM (98.6%), and the 6B Fast-WAM (97.6%). WSA_1 therefore obtains the best accuracy–efficiency trade-off.

Ablation Study of World-Action Bidirectional Modeling. As illustrated in Table 3, bidirectional world-action modeling improves the average success rate from 73.5% to 80.2%. Removing either direction degrades performance: removing world-to-action reduces the score to 71.0%, while removing action-to-world yields 73.5%. Notably, the latter matches the fully unidirectional setting, suggesting that the gain comes from explicitly coupling action-conditioned world prediction with world-guided action generation, rather than simply combining two prediction targets. Under the causal ordering implemented by the unidirectional variant, the predicted world is not conditioned on the actions, which is why that variant behaves like the model whose action-to-world direction has been removed.

Ablation Study of Learning Objectives. As reported in Table 4, optimizing the action objective alone yields 59.7% average SR. Adding 2D SST or 3D AC-WM improves it to 75.8% and 74.1%, respectively, while jointly optimizing all three objectives further reaches 80.2%. These results show that the spatial and world objectives provide complementary supervision for action learning, with consistent gains also observed in the randomized setting.

Table 3: Ablation study of bidirectional world-action modeling on RoboTwin2.0.

Setting	Clean	Rand	Avg.	Drop.
unidirectional WSA	72.5	74.5	73.5	–
bidirectional WSA	**80.7**	**79.6**	**80.2**	–
w/o action-to-world	72.6	74.4	73.5	↓ 6.7
w/o world-to-action	70.4	71.6	71.0	↓ 9.2

Table 4: Ablation study of learning objectives on RoboTwin2.0.

Setting	$\mathcal{L}_{\text{ACT}}$	$\mathcal{L}_{\text{2D}}$	$\mathcal{L}_{\text{3D}}$	Clean	Rand	Avg.
Action only	✓	–	–	59.5	59.9	59.7
w SST	✓	✓	–	75.8	75.8	75.8
w AC-WM	✓	–	✓	73.4	74.7	74.1
Full WSA	✓	✓	✓	**80.7**	**79.6**	**80.2**

Investigation of WSA Pre-train/Post-train. As shown in Table 5, we compare three ways of applying the WSA objectives across pre-training and post-training. Using WSA only during post-training yields poor OOD transfer, with 5.4% under randomized OOD-C2R and 6.6% under randomized OOD-CE, indicating that the three-level objectives alone are insufficient to handle substantial distribution shifts. Introducing WSA during pre-training substantially improves generalization, raising randomized OOD-C2R performance to 75.9% and clean performance from 61.1% to 83.4%. Retaining WSA objectives during post-training further achieves the best overall results (83.8%, 77.8%, 47.1%, and 47.6%). These results indicate that WSA-based pre-training learns transferable physical interaction priors, while preserving the objectives during post-training further strengthens their transfer to unseen scenes and embodiments.

Table 5: Investigation of WSA joint modeling for pre-training and post-training.

Pre-train	Post-train	OOD-C2R		OOD-CE	
		Clean	Rand	Clean	Rand
None	WSA	61.1	5.4	8.2	6.6
WSA	Action only	83.4	75.9	46.7	46.4
WSA	WSA	**83.8**	**77.8**	**47.1**	**47.6**

Table 6: Investigation of scene generalization. All models are fine-tuned with clean data only.

Method	Clean	Rand	Avg.
$\pi_{0.5}$ (Black et al., 2025a)	70.7	46.0	58.4
ABot-M0 (Yang et al., 2026b)	57.4	30.4	43.9
RobotManip (Yuan et al., 2026a)	73.2	62.6	67.9
Fast-WAM (Yuan et al., 2026b)	77.8	1.90	39.9
InternVLA-A1 (Cai et al., 2026)	**86.6**	74.6	80.6
WSA_1	83.8	**77.8**	**80.8**

Table 7: Investigation of zero-shot cross embodiment generalization. Unseen ARX-X5 is adopted for testing.

Method	Clean	Rand	Avg.
$\pi_{0.5}$ (Black et al., 2025a)	33.5	19.0	26.3
LingBot-VA (Li et al., 2026a)	17.2	16.6	16.9
RobotManip (Yuan et al., 2026a)	–	42.9	42.9
Fast-WAM (Yuan et al., 2026b)	25.9	22.4	24.2
InternVLA-A1 (Cai et al., 2026)	18.1	17.5	17.8
WSA_1	**47.1**	**47.6**	**47.4**

Scene Generalization. As shown in Table 6, all models are fine-tuned on clean demonstrations and tested in randomized scenes to evaluate scene-level generalization. WSA_1-B achieves the highest randomized success rate (77.8%) and the highest average (80.8%), outperforming InternVLA-A1 by 3.2 points in the randomized setting. More importantly, WSA_1-B shows only a 6.0-point drop from clean to randomized scenes, substantially smaller than those of the competing methods. For example, InternVLA-A1 drops by 12.0 points, while Fast-WAM decreases from 77.8% to 1.9%. These results suggest that WSA_1 is more robust to scene-level distribution shifts and less dependent on appearance-specific correlations.

Cross-embodiment Generalization. As shown in Table 7, all models are fine-tuned on the source embodiment and evaluated on an unseen ARX-X5 robot without target-platform fine-tuning. WSA_1-B achieves 47.1% and 47.6% in the clean and randomized modes, respectively, averaging 47.4%. It outperforms $\pi_{0.5}$ by 13.6 and 28.6 points in the two settings and exceeds the best remaining average, RobotManip (42.9%), by 4.5 points. Nevertheless, the absolute performance remains well below the in-distribution result in Table 2, indicating that zero-shot cross-embodiment transfer remains challenging despite the clear relative gains.

Summary of Exploration Study. Overall, the experiments provide two main findings: 1) *Generalization.* WSA_1 consistently performs strongly across in-distribution, scene generalization (OOD-C2R), and zero-shot cross-embodiment (OOD-CE) settings, while remaining parameter-efficient; 2) *Modeling effectiveness.* What makes 3D world modeling useful for robot control is the bidirectional joint modeling, not the objective itself. This bidirectional world-action modeling is where generalization comes from, and it must be acquired during pre-training.

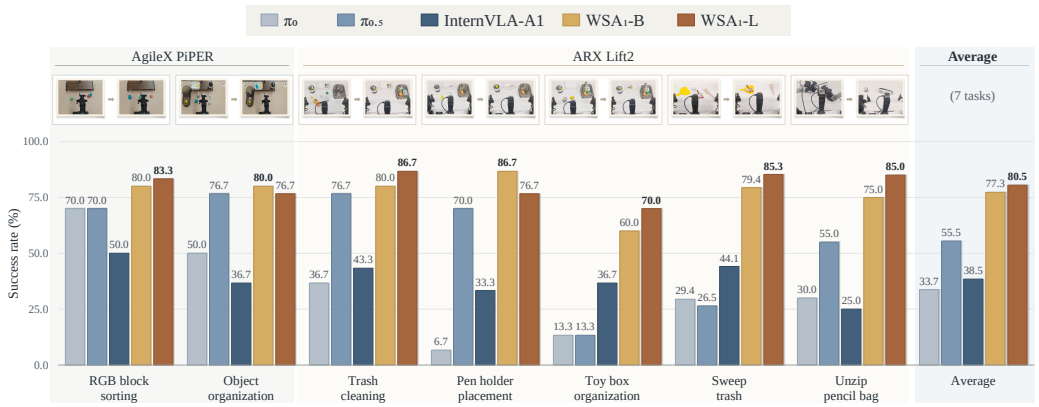

Figure 3: **Real-robot success rates across seven tasks.** Bars compare five policies on each task and their mean success rate over all seven tasks. Thumbnails show initial and final states. The first two tasks use AgileX PiPER, and the remaining five use ARX Lift2.

4.2 Real Robot Experiments

Experimental Setup. Real-robot experiments are conducted on two platforms: (1) a unimanual AgileX PiPER with three cameras and a 7-DoF arm, and (2) a bimanual ARX Lift2 with three cameras and two 7-DoF arms. We design seven tabletop manipulation tasks and evaluate all policies with objects initialized in predefined positions, orientations, and scene layouts to evaluate generalization capability. We report two metrics: Success Rate (SR), the proportion of successfully completed rollouts; and Completeness (C), the proportion of completed sub-tasks.

Baselines. We compare WSA_1 with three representative open-source embodied foundation models: π_0 (Black et al., 2025b) for continuous action prediction, $\pi_{0.5}$ (Black et al., 2025a) for open-world robot control, and InternVLA-A1 (Cai et al., 2026), which unifies visual understanding, video generation, and action generation.

Main Results on Real-world Manipulation Tasks. As shown in Figure 3, WSA_1 achieves leading performance over baselines in the real-world setting. Across 7 tabletop manipulation tasks, WSA_1-L achieves an average success rate of 80.5% and an average completeness score of 85.2%, while WSA_1-B reaches 77.3% and 82.3%, substantially outperforming the second-best model $\pi_{0.5}$ with 55.5% and 63.3%, respectively.

The per-task results further show that the advantage becomes more pronounced on challenging bimanual tasks. On the two unimanual PiPER tasks, the baselines remain competitive, with $\pi_{0.5}$ reaching 70.0% on *RGB Block Sorting* and 76.7% on *Object Organization*. On the five bimanual ARX Lift2 tasks, however, both WSA_1 models achieve at least 60.0% success on every task, while π_0 and InternVLA-A1 remain below 50.0% throughout. The largest gaps appear on tasks requiring long-horizon execution and precise spatial reasoning: WSA_1-B reaches 86.7% on *Pen Holder Placement* versus 6.7% for π_0, while WSA_1-L achieves 70.0% on *Toy Box Organization*, 85.0% on *Unzip Pencil Bag*, and 85.3% on *Sweep Trash*, substantially exceeding the baselines.

These results highlight the effectiveness of the proposed WSA architecture. By coupling 3D-aligned visual state transition, 3D world dynamics, and 3D-conditioned action generation, WSA_1 enables more reliable execution in spatially complex and long-horizon tasks.

5 Conclusion

This work proposes WSA, a 3D-centric world-spatial-action paradigm that jointly models 3D space state transition, action-caused 3D world prediction, and 3D inverse dynamics within a Mixture-of-Transformers architecture. We instantiate WSA at two scales, i.e., WSA_1-B (3B) and WSA_1-L (6B). With only 6,000 hours of demonstrations (1,000 hours from real robots), WSA achieves strong performance across simulation and real-world benchmarks, demonstrating a data-efficient path toward generalizable robot foundation models.

AI USE STATEMENT

In the preparation of this paper, we employed generative AI solely as a writing assistance tool for limited text polishing and language refinement. These AI tools were not involved in any aspects of research ideation, conceptual development, technical analysis, algorithm design, experimental execution, or result analyses. All scientific contributions, methodological innovations, and intellectual content remain entirely our own.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *CoRL*, volume 305, pp. 17–40, 2025a.

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

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

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. URL https://arxiv.org/abs/2506.18088.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://doi.org/10.15607/RSS.2023.XIX.026.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. In *CVPR*, 2026.

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

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

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

Stephen M Kosslyn. *Image and brain: The resolution of the imagery debate*. MIT press, 1996.

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

Peiyan Li, Yixiang Chen, Yuan Xu, Jiabing Yang, Xiangnan Wu, Jun Guo, Nan Sun, Long Qian, Xinghang Li, Xin Xiao, Jing Liu, Nianfeng Liu, Tao Kong, Yan Huang, Liang Wang, and Tieniu Tan. Multi-view video diffusion policy: A 3d spatio-temporal-aware video action model. *arXiv preprint arXiv:2604.03181*, 2026b.

Haitao Lin, Hanyang Yu, Jingshun Huang, He Zhang, Yonggen Ling, Ping Tan, Xiangyang Xue, and Yanwei Fu. Universal pose pretraining for generalizable vision-language-action policies. In *Proceedings of Robotics: Science and Systems*, 2026a. URL https://arxiv.org/abs/2602.19710.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Xuewu Lin, Tianwei Lin, Yun Du, Hongyu Xie, Yiwei Jin, Jiawei Li, Shijie Wu, Qingze Wang, Mengdi Li, Mengao Zhao, Ziang Li, Chaodong Huang, Hongzhe Bi, Lichao Huang, and Zhizhong Su. Holobrain-0 technical report. *arXiv preprint arXiv:2602.12062*, 2026b. URL https://arxiv.org/abs/2602.12062.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: benchmarking knowledge transfer for lifelong robot learning. In *NeurIPS*, 2023.

Yunfan Lou, Xiaowei Chi, Xiaojie Zhang, Zezhong Qian, Chengxuan Li, Rongyu Zhang, Yaoxu Lyu, Guoyu Song, Chuyao Fu, Haoxuan Xu, Pengwei Wang, and Shanghang Zhang. Mask world model: Predicting what matters for robust robot policy learning. *arXiv preprint arXiv:2604.19683*, 2026.

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

Yunchao Ma, Ze Chen, Erjin Zhou, Ziming Liu, Haowei Zhang, Kai Liu, and Haoqiang Fan. Table30 V2: Evaluating generalized models by real robots at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4461–4467, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/GigaBrainChallenge/html/Ma_Table30_V2_Evaluating_Generalized_Models_by_Real_Robots_at_Scale_CVPRW_2026_paper.html.

NVIDIA GEAR. GR00T N1.5: An improved open foundation model for generalist humanoid robots. Research project page, 2025. URL https://research.nvidia.com/labs/gear/gr00t-n1_5/.

NVIDIA GEAR Team. NVIDIA Isaac GR00T N1.6: An improved open foundation model for generalist humanoid robots. NVIDIA Research project page, 2025. URL https://research.nvidia.com/labs/gear/gr00t-n1_6/.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint 2512.15692*, 2025.

Jingjing Qian, Boyao Han, Chen Shi, Lei Xiao, Long Yang, Shaoshuai Shi, and Li Jiang. Geopredict: Leveraging predictive kinematics and 3d gaussian geometry for precise vla manipulation. In *CVPR*, pp. 13529–13539, 2026.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model. In *Robotics: Science and Systems (RSS)*, 2025.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing, 2026. URL https://arxiv.org/abs/2604.05014.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, Xuhong Huang, Pei Lin, Junyang Lin, Dayiheng Liu, Shuai Bai, Jingren Zhou, Jiazhao Zhang, Haoqi Yuan, Gengze Zhou, Hang Yin, Ye Wang, Yiyang Huang, Zixing Lei, Wujian Peng, Delin Chen, Yingming Zheng, Jingyang Fan, Xianwei Zhuang, Xin Zhou, Haoyang Li, Anzhe Chen, Tong Zhang, Xuejing Liu, Yuchong Sun, Ruizhe Chen, Zhaohai Li, Chenxu Lü, Zhibo Yang, Tao Yu, and Xionghui Chen. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026. URL https://arxiv.org/abs/2605.30280.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. In *CoRL*, 2025.

Daniel M Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature neuroscience*, 3(11):1212–1217, 2000.

Adina Yakefu, Bin Xie, Chongyang Xu, Enwen Zhang, Erjin Zhou, Fan Jia, Haitao Yang, Haoqiang Fan, Haowei Zhang, Hongyang Peng, et al. Robochallenge: Large-scale real-robot evaluation of embodied policies. *arXiv preprint arXiv:2510.17950*, 2025.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026a. URL https://arxiv.org/abs/2608.08023.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026b.

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

Zhenhan Yin, Xuanhan Wang, Jiahao Jiang, Kaiyuan Deng, Pengqi Chen, Shuangle Li, Chong Liu, Xing Xu, Jingkuan Song, Lianli Gao, and Heng Tao Shen. Mivla: Towards generalizable vision-language-action model with human-robot mutual imitation pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1535–1545, June 2026.

Zhaoshu Yu, Bo Wang, Pengpeng Zeng, Haonan Zhang, Ji Zhang, Zheng Wang, Lianli Gao, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. A survey on efficient vision-language-action models. *arXiv preprint arXiv:2510.24795*, 2025.

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, Jie Zhang, Jingyang Fan, Gengze Zhou, Qihang Peng, Chenxu Lv, Xiaoyue Chen, An Yang, Fei Huang, Junyang Lin, Dayiheng Liu, Jingren Zhou, Chenfei Wu, and Xiong-Hui Chen. Qwen-RobotManip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026a. URL https://arxiv.org/abs/2606.17846.

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

Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097*, 2025a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. In *NeurIPS*, 2025b.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, pp. 1702–1713, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016. URL https://doi.org/10.15607/RSS.2023.XIX.016.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. In *International Conference on Machine Learning (ICML)*, pp. 61229–61245, 2024.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.10274.

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

This appendix records the configuration behind every experiment illustrated in the main paper and adds more evidence to support core claims in this paper. The first part specifies how the two WSA models were pre-trained and fine-tuned, how the world-modeling objectives were supervised, how large the deployed models are and what they cost at inference time, and the protocol under which every evaluation was run. The second part reports the full benchmark comparisons, the two distribution-shift studies, the directional ablation together with its qualitative evidence, and the real-robot task suite with its per-task scores.

A Implementation Details

A.1 Pre-training configuration

WSA_1-B and WSA_1-L share an optimisation schedule of 300,000 training steps at an effective batch size of 96 on eight GPUs, with cosine decay to a final learning rate of 1×10^{-5}, bfloat16 precision and random seed 42. Three things differ between the scales, and Table 8 makes them explicit. The 2D Spatial Expert is initialized from a different backbone, Qwen3-VL-2B-Instruct for WSA_1-B against Wan2.2-TI2V-5B for WSA_1-L, which also changes the frozen image tokenizer each model carries. The visual and textual context is larger for WSA_1-L, which is pre-trained at 384×320 with a 128-token text context while WSA_1-B uses 224×224 and 48 tokens, so the larger model runs a higher peak learning rate (1×10^{-4} against 5×10^{-5}) with a longer warmup (5,000 against 2,000 steps). The loss weighting is scale-specific as well: (λ_{2D}, λ_{3D}, λ_{ACT}) is (0.01, 0.01, 1.0) for WSA_1-B and (1.0, 0.1, 1.0) for WSA_1-L, so the 2D state-transition objective carries two orders of magnitude more weight at the larger scale, while the action objective stays at 1.0 at both. Actions are represented as deltas with a horizon of 50 for WSA_1-B and 32 for WSA_1-L, and inference uses 10 flow-matching steps.

Table 8: Pretraining hyperparameters for WSA_1-B and WSA_1-L. Loss weights are ordered as $(\lambda_{\text{2D}}, \lambda_{\text{3D}}, \lambda_{\text{ACT}})$.

Configuration	**WSA_1-B (3B)**	**WSA_1-L (6B)**
Backbone initialization	Qwen3-VL-2B-Instruct	Wan2.2-TI2V-5B
Optimizer	AdamW	AdamW
AdamW $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.95, 10^{-8})$	$(0.9, 0.95, 10^{-8})$
Weight decay	0.01	0.01
Learning-rate schedule	Warmup + cosine decay	Warmup + cosine decay
Peak $\rightarrow$ final learning rate	$5\times10^{-5} \rightarrow 1\times10^{-5}$	$1\times10^{-4} \rightarrow 1\times10^{-5}$
Warmup steps	2,000	5,000
Training steps	300,000	300,000
Per-device batch	12	12
Number of GPUs	8	8
Gradient accumulation	1	1
Effective global batch	96	96
Numerical precision	bfloat16	bfloat16
Random seed	42	42
Text context length	48	128
Visual input resolution	224×224	384×320
Action representation	Delta	Delta
Action horizon	50	32
Flow-matching inference steps	10	10
Loss weights (2D/3D/ACT)	0.01/0.01/1.0	1.0/0.1/1.0

A.2 Supervision of the 3D world modeling objective

The 3D world modeling objective is supervised by a frozen DA3-Large-1.1 teacher (Lin et al., 2025). Each of the three camera views is processed at 504×504, and target features are read from four hidden layers of the teacher, 11, 15, 19 and 23, with indices starting at zero. Four layers of the WSA_1 backbone are aligned to them through a single cross-attention block followed by a SiLU feed-forward network, layers 13, 19, 23 and 27 for WSA_1-B and 14, 19, 24 and 29 for WSA_1-L. Every view

contributes 144 scene tokens and 1,296 target features arranged on a 36×36 grid, all projected to a feature dimension of 2,048. Because the teacher is frozen, these objectives inherit only the structure that its representations contain, and no part of it is loaded at inference. Table 9 summarizes the 3D supervision configuration.

Table 9: 3D supervision configuration during pretraining. Layer indices start at zero.

Configuration	**WSA$_1$-B**	**WSA$_1$-L**
DA3 specification	DA3-Large-1.1	
input resolution per view	504×504	
selected hidden layers in DA3	$11, 15, 19, 23$	
selected hidden layers in WSA$_1$	$13, 19, 23, 27$	$14, 19, 24, 29$
Feature dimension	2,048	
Camera views	3	
Scene tokens per view	144	
Target features per view	1,296 (36×36 grid)	
Feature alignment module	One cross-attention block + SiLU FFN	

A.3 Downstream fine-tuning

Fine-tuning uses one recipe across every benchmark and platform: AdamW with (β_1, β_2, ϵ) = (0.9, 0.95, 10^{-8}), weight decay 0.01, cosine learning-rate decay, bfloat16 precision, seed 42 and eight GPUs, with 10 flow-matching steps at inference. Table 10 lists what that recipe is instantiated with, namely the platform a policy is trained on, the schedule length, the effective batch size, the peak and final learning rates, the horizon H of predicted actions together with the number E of actions executed before replanning, the action representation and the loss weights. Three of these depend on the benchmark rather than on the model scale. The action representation is delta everywhere except LIBERO with WSA$_1$-L, which uses absolute actions. The execution horizon E is 24 on RoboTwin2.0, 10 on LIBERO and 50 or 48 on the real robots. And the 2D and 3D loss weights are re-tuned per benchmark, from (0.005, 0.01, 1.0) on LIBERO with WSA$_1$-B to (1.0, 0.1, 1.0) on RoboTwin2.0 with WSA$_1$-L, while λ_{ACT} remains 1.0 in every row.

Table 10: Hyperparameters for downstream adaptation. Batch denotes the effective global batch size. Loss weights are ordered as $(\lambda_{\text{2D}}, \lambda_{\text{3D}}, \lambda_{\text{ACT}})$.

Model	**Platform**	**Duration**	**Batch**	**Peak LR**	**Final LR**	**Warmup**	**Action** (H/E)	**Loss weights**
RoboTwin2.0								
WSA$_1$-B	Aloha	200k steps	128	5×10^{-5}	5×10^{-6}	2,000	Delta (50/24)	0.01/0.01/1.0
WSA$_1$-L	Aloha	5 epochs	96	1×10^{-4}	1×10^{-6}	5,000	Delta (32/24)	1.0/0.1/1.0
LIBERO								
WSA$_1$-B	Franka	128k steps	128	7.5×10^{-5}	5×10^{-6}	2,000	Delta (10/10)	0.005/0.01/1.0
WSA$_1$-L	Franka	110k steps	128	7.5×10^{-5}	5×10^{-6}	2,000	Abs. (12/10)	0.1/0.1/1.0
Real robot								
WSA$_1$-B	PiPER	30k steps	128	5×10^{-5}	5×10^{-6}	600	Delta (50/50)	0.01/0.01/1.0
WSA$_1$-B	Lift2	40k steps	128	5×10^{-5}	5×10^{-6}	600	Delta (50/50)	0.01/0.01/1.0
WSA$_1$-L	PiPER	40k steps	96	6×10^{-5}	1×10^{-5}	600	Delta (48/48)	0.1/0.1/1.0
WSA$_1$-L	Lift2	50k steps	96	6×10^{-5}	1×10^{-5}	600	Delta (48/48)	0.1/0.1/1.0

A.4 Parameter budget and inference cost

Table 11 splits the trained parameters by expert. WSA$_1$-B holds 3.156B: 2.128B in the 2D Spatial Expert, 0.440B in the 3D Spatial Expert, 0.440B in the 3D Action Expert and about 0.148B in the remainder. WSA$_1$-L holds 6.457B, with 5.000B, 0.436B and 1.021B in the three experts. The growth from the smaller to the larger scale is therefore concentrated in the 2D Spatial Expert, which supplies 2.872B of the 3.301B added, or 87% of the increase. Every one of these parameters is updated during

training, and the frozen components, namely the image tokenizer and the frozen 3D teacher backbone, are listed separately and excluded from the totals.

Table 11: The statistics of model parameters

Model	2D-SE	3D-SE	3D-AE	Other	Total
WSA_1-B	2.128B	0.440B	0.440B	∼0.148B	3.156B
WSA_1-L	5.000B	0.436B	1.021B	∼0.1M	6.457B
Frozen components excluded from policy totals					
WSA_1-B	Cosmos image tokenizer (76.8M); DA3-Large backbone (304.4M)				
WSA_1-L	Wan VAE (704.7M); DA3-Large backbone (304.4M)				

Table 12 measures a single action-inference call on one A800-SXM4-80GB with the frozen DA3 disabled. WSA_1-B peaks at 6.18 GiB and returns a chunk in 812.6 ms at the median and 870.1 ms at the 95th percentile, sustaining 1.210 chunks per second; WSA_1-L peaks at 14.01 GiB and returns a chunk in 1060.1 ms, sustaining 0.945 chunks per second. Doubling the trained parameter count therefore costs 2.3 times the peak memory and reduces chunk throughput by 22%. The parameter count resident at inference is slightly above the trained total, 3.233B against 3.156B for WSA_1-B and 7.161B against 6.457B for WSA_1-L, and every row is measured with batch size 1 and 10 flow-matching steps.

Table 12: Single-GPU action inference on an A800-SXM4-80GB with BF16, batch size 1, 10 denoising steps, the DA3 teacher disabled, 10 warmup runs, and 50 measured runs.

Model	Params (B) policy / inference	Peak VRAM (GiB)	Latency p50/p95 (ms)	Chunk/s
WSA_1-B	3.156 / 3.233	6.18	812.6 / 870.1	1.210
WSA_1-L	6.457 / 7.161	14.01	1060.1 / 1088.5	0.945

B Additional Results and Analyses

B.1 Full simulation results

Table 13 reproduces the complete LIBERO comparison behind the summary in Table 2, with 24 baselines drawn from both the VLA and the WAM families. The benchmark is close to saturation: 14 of the 26 entries average at least 97.0% and six average at least 98.0%, and the remaining headroom sits in the long-horizon suite, the only column in which the field is not compressed into the last two points. WSA_1-L averages 98.2%, fifth among the 26 entries and 1.0 point below the best entry (Being-H0.7 at 99.2%), while WSA_1-B averages 97.4%; both sit inside the cluster of 6B and 8B models rather than above it. That is why Section 4.1 treats LIBERO as a consistency check and separates the two paradigms on RoboTwin2.0 and on the real robots.

Table 14 reports the per-task results on all 50 RoboTwin2.0 tasks.

B.2 Robustness under scene-level shift

Table 15 isolates the effect of state and action normalization on transfer to the unseen ARX-5 robot. One WSA_1-B checkpoint is evaluated twice, first with the ALOHA means and standard deviations supplied with the checkpoint and then with 14-dimensional statistics recomputed from the official RoboTwin ARX-5 demonstrations on Hugging Face. These statistics standardize the input state and convert normalized action predictions back to their original scale, while the policy parameters, the embodiment mapping and the evaluation settings are held fixed. Switching the statistics raises success from 17.0% to 47.1% in the clean setting and from 17.2% to 47.6% under randomization, differences of 30.1 and 30.4 points. The absolute level of a zero-shot cross-embodiment result is therefore a property of the normalization statistics as much as of the policy, and the WSA_1-B entry of Table 7 is the one obtained with ARX-5 statistics.

Table 13: Full comparison with prior methods on LIBERO. Dashes indicate unreported results.

Method	Spatial	Object	Goal	Long	Avg.
ACT (Zhao et al., 2023)	96.3	99.5	97.0	90.8	95.9
Diffusion Policy (Chi et al., 2023)	78.3	92.5	68.3	50.5	72.4
Octo (Team et al., 2024)	78.9	85.7	84.6	51.1	75.1
OpenVLA (Kim et al., 2025b)	84.7	88.4	79.2	53.7	76.5
π_0 (Black et al., 2025b)	96.8	98.8	95.8	85.2	94.2
$\pi_{0.5}$ (Black et al., 2025a)	98.8	98.2	98.0	92.4	96.9
OpenVLA-OFT (Kim et al., 2025a)	97.6	98.4	97.9	94.5	97.1
X-VLA (Zheng et al., 2026)	98.2	98.6	97.8	97.6	98.1
GR00T N1 (Bjorck et al., 2025)	95.6	97.6	94.2	89.6	94.3
GR00T N1.5 (Bjorck et al., 2025; NVIDIA GEAR, 2025)	98.2	99.4	97.2	87.8	95.7
GR00T N1.6 (Bjorck et al., 2025; NVIDIA GEAR Team, 2025)	97.7	98.5	97.5	94.4	97.0
SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1
StarVLA-OFT (4B, spec.) (StarVLA Community, 2026)	99.0	99.8	98.5	94.1	97.9
Qwen-VLA-Instruct (Wang et al., 2026)	–	–	–	–	97.9
HoloBrain-0-QW (Lin et al., 2026b)	97.2	99.6	97.6	95.2	97.4
Pose-VLA (RGB) (Lin et al., 2026a)	96.5	98.0	97.1	92.4	96.0
GeoPredict (Qian et al., 2026)	98.0	98.2	95.7	94.0	96.5
LingBot-VA (Li et al., 2026a)	98.5	99.6	97.2	**98.5**	98.5
ABot-M0 (Yang et al., 2026b)	98.8	99.8	**99.0**	96.6	98.6
Being-H0.7 (Luo et al., 2026)	–	–	–	–	**99.2**
DreamVLA (Zhang et al., 2025b)	97.5	94.0	89.5	89.5	92.6
Motus (Bi et al., 2026)	96.8	99.8	96.6	97.6	97.7
4D-WAM (Yang et al., 2026a)	98.8	99.5	98.6	97.4	98.6
Fast-WAM (Yuan et al., 2026b)	98.2	**100.0**	97.0	95.2	97.6
WSA$_1$-B	98.6	99.6	97.2	94.2	97.4
WSA$_1$-L	**99.4**	99.8	98.0	95.6	98.2

B.3 Cross-embodiment transfer and normalization statistics

Table 16 examines how state and action normalization affects transfer to ARX-X5. We compare the ALOHA means and standard deviations supplied with the WSA$_1$-B checkpoint with 14-dimensional statistics estimated from the official RoboTwin ARX-X5 demonstrations on Hugging Face. These statistics standardize the input state and convert normalized action predictions back to their original scale. Policy parameters, embodiment mapping, and evaluation settings remain fixed. Using ARX-X5 statistics improves success by 30.1 percentage points in Clean and 30.4 points in Randomized, showing that normalization substantially affects transfer with a fixed policy. The WSA$_1$-B results in Table 7 use ARX-X5 statistics.

B.4 More details of WSA modeling

World–action modeling directions. Table 17 records which information paths each setting of the directional ablation retains, and corresponds row by row to the four settings compared in Table 3. Removing G→A blocks scene information from both the state and the action representations, including the indirect scene→state→action path, and costs 9.2 points of average success; the fully unidirectional baseline removes G→V as well and uses triangular attention, yet its success rate matches the variant that removes action-to-world alone, which is why Section 4.1 attributes the gain to the coupling between the two directions rather than to the number of prediction targets. Figure 4 shows the same ablation qualitatively, plotting future point clouds decoded with frozen DA3 heads for four WSA$_1$-B variants alongside the teacher reference. The bidirectional model follows the reference geometry most closely and the ablated variants depart from it, so the ordering of the columns follows the ordering of success rates in Table 3, which is consistent with the reading that a more accurate learned world accompanies stronger manipulation generalization.

Table 14: Per-task success rates (%) on all 50 RoboTwin2.0 tasks. Each method is evaluated with 100 rollouts per task and domain. Fast-WAM and $\pi_{0.5}$ results are from Yuan et al. (2026b), Appendix Table 3. Bold marks the best score in each domain, including ties; averages are over all 50 tasks.

Task	$\pi_{0.5}$		Fast-WAM		**WSA$_1$-B**		**WSA$_1$-L**	
	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.
Adjust Bottle	**100**	99	**100**	**100**	**100**	99	**100**	**100**
Beat Block Hammer	96	93	99	97	98	95	**100**	**100**
Blocks Ranking RGB	92	85	**100**	**100**	95	95	**100**	99
Blocks Ranking Size	49	26	**94**	**98**	92	92	82	86
Click Alarmclock	98	89	**100**	**100**	**100**	**100**	95	98
Click Bell	99	66	**100**	**100**	**100**	**100**	99	99
Dump Bin Bigbin	92	97	**97**	96	**97**	**99**	**97**	98
Grab Roller	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Handover Block	66	57	95	81	91	84	**96**	**85**
Handover Mic	98	97	99	**100**	**100**	**100**	**100**	**100**
Hanging Mug	18	17	**58**	**62**	50	50	40	42
Lift Pot	96	85	**100**	**100**	**100**	99	**100**	**100**
Move Can Pot	51	55	90	88	86	95	**99**	**96**
Move Pillbottle Pad	84	61	**100**	99	99	**100**	**100**	99
Move Playingcard Away	96	84	**100**	**100**	**100**	**100**	**100**	**100**
Move Stapler Pad	56	42	77	64	65	75	**82**	**81**
Open Laptop	90	96	98	**100**	**100**	**100**	98	**100**
Open Microwave	34	77	62	45	**100**	**100**	90	77
Pick Diverse Bottles	81	71	80	**85**	86	76	**90**	79
Pick Dual Bottles	93	63	**100**	**96**	94	89	98	**96**
Place A2B Left	87	82	95	93	92	**97**	**99**	**97**
Place A2B Right	87	84	93	**99**	94	94	**98**	95
Place Bread Basket	77	64	91	93	**99**	94	98	**97**
Place Bread Skillet	85	66	90	**93**	90	89	**94**	92
Place Burger Fries	94	87	96	99	**100**	**100**	**100**	**100**
Place Can Basket	62	62	71	69	80	79	**81**	**80**
Place Cans Plasticbox	94	84	99	96	**100**	99	**100**	**100**
Place Container Plate	99	95	96	**100**	99	**100**	**100**	**100**
Place Dual Shoes	75	75	94	88	**95**	**94**	91	89
Place Empty Cup	**100**	99	**100**	**100**	**100**	99	**100**	**100**
Place Fan	87	85	96	96	98	**99**	**100**	98
Place Mouse Pad	60	39	83	**89**	**86**	86	83	84
Place Object Basket	80	76	89	88	84	87	**91**	**90**
Place Object Scale	86	80	90	**97**	85	88	**96**	93
Place Object Stand	91	85	90	94	95	95	**96**	**97**
Place Phone Stand	81	81	97	99	**99**	97	98	**100**
Place Shoe	92	93	96	99	99	99	**100**	**100**
Press Stapler	87	83	**90**	**97**	83	78	87	82
Put Bottles Dustbin	84	79	**95**	90	**95**	92	92	**98**
Put Object Cabinet	80	79	**94**	89	90	**93**	93	91
Rotate QRcode	89	87	**93**	89	87	**92**	88	90
Scan Object	72	65	89	92	85	88	**95**	**96**
Shake Bottle	99	97	**100**	**100**	**100**	**100**	**100**	**100**
Shake Bottle Horizontally	99	99	**100**	**100**	**100**	**100**	**100**	**100**
Stack Blocks Three	91	76	95	97	98	98	**99**	**100**
Stack Blocks Two	97	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Stack Bowls Three	77	71	80	81	**92**	**93**	91	91
Stack Bowls Two	95	96	92	98	97	**100**	**99**	**100**
Stamp Seal	79	55	**90**	**94**	73	82	75	89
Turn Switch	62	54	61	59	61	**75**	**66**	73
Average	82.74	76.76	91.88	91.78	92.18	92.70	**93.52**	**93.14**

Table 15: Performance drop under OOD-C2R setting. Δ is Randomized$-$Clean in percentage points.

Task	Clean	Rand.	Δ
Open Microwave	71	32	-39
Click Bell	95	65	-30
Handover Block	74	46	-28
Put Object Cabinet	56	36	-20
Beat Block Hammer	94	75	-19
Place Can Basket	52	64	$+12$
Turn Switch	49	53	$+4$
Pick Diverse Bottles	75	79	$+4$
Hanging Mug	29	33	$+4$

Table 16: Effect of normalization statistics on WSA_1-B transfer to ARX-X5. Success rates (%) are measured over 45 tasks with 100 episodes per task and setting, using the same checkpoint.

Statistics source	Clean	Randomized
ALOHA (checkpoint)	17.0	17.2
ARX-X5 (recomputed)	47.1	47.6

Table 17: Information-flow paths in the directional ablation.

Setting	G$\rightarrow$V	A$\rightarrow$G	G$\rightarrow$A
Fully unidirectional	No	No	Yes
Fully bidirectional	Yes	Yes	Yes
w/o action-to-world	Yes	No	Yes
w/o world-to-action	Yes	Yes	No

B.5 Visualization of predicted 3D world

We compare the predicted point clouds across WSA_1-B variants. As shown in Figure 4, bidirectional WSA modeling yields the best performance, indicating that the better the 3D world representation the model learns, the stronger manipulation generalization it achieves.

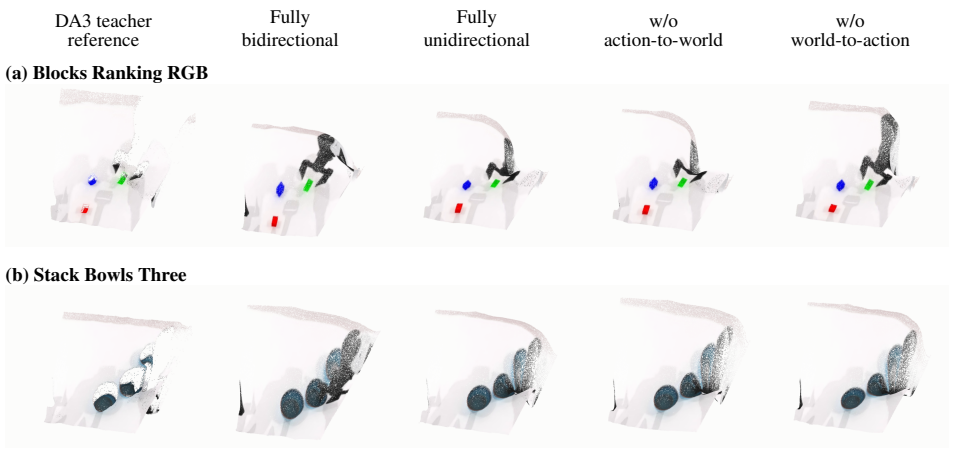

Figure 4: Decoded future point clouds for the directional attention ablation on (a) Blocks Ranking RGB and (b) Stack Bowls Three. The first column shows the DA3 teacher reference from the future observation. The remaining columns show point clouds obtained by decoding the predicted 3D features with frozen DA3 heads.

B.6 Real-Robot Tasks and Per-Task Results

We use two platforms: a 7-DoF single-arm AgileX PiPER and a 14-DoF bimanual ARX Lift2, each observed by three cameras. Trials begin from predefined object positions, orientations, and layouts.

Success requires completion of the full task. Completeness is the fraction of predefined task stages completed before termination. Trial counts and stage-progress rules are specific to each task, and task-level scores are averaged with equal weight. We train and evaluate all five policies in Table 18 using the same downstream demonstrations and evaluation settings across the seven tasks. Table 19 summarizes the tasks and demonstration data.

The evaluation sets contain 30 trials for each of the first five tasks, 34 trials for Sweep Trash, and 20 trials for Unzip Pencil Bag. In the order listed in the table, WSA_1-L completes the full task in 25/30, 23/30, 26/30, 23/30, 21/30, 29/34, and 17/20 trials. Averaging the seven task success rates with equal weight gives 80.52%.

Table 18: Per-task real-robot results (%). SR denotes full-task success and C denotes stage completeness. Each policy uses the same number of trials per task. Averages are unweighted means over the seven tasks, computed separately for SR and C.

Task	**Robot**	π_0		$\pi_{0.5}$		InternVLA-A1		WSA_1-B		WSA_1-L	
		SR	C	SR	C	SR	C	SR	C	SR	C
RGB Block Sorting	PiPER	70.00	70.00	70.00	70.00	50.00	50.00	80.00	80.00	**83.33**	**83.33**
Object Organization	PiPER	50.00	50.00	76.67	76.67	36.67	36.67	**80.00**	**80.00**	76.67	76.67
Trash Cleaning	Lift2	36.67	53.33	76.67	86.67	43.33	58.33	80.00	90.00	**86.67**	**91.67**
Pen Holder Placement	Lift2	6.67	28.33	70.00	76.67	33.33	38.33	**86.67**	**90.00**	76.67	76.67
Toy Box Organization	Lift2	13.33	18.33	13.33	26.67	36.67	48.33	60.00	71.67	**70.00**	**85.00**
Sweep Trash	Lift2	29.41	40.19	26.47	44.12	44.12	78.41	79.41	89.21	**85.29**	**93.12**
Unzip Pencil Bag	Lift2	30.00	42.50	55.00	62.50	25.00	42.50	75.00	75.00	**85.00**	**90.00**
Average		33.73	43.24	55.45	63.33	38.45	50.37	77.30	82.27	**80.52**	**85.21**

Table 19: Real-robot tasks and demonstration data. Traj. / Time denotes the trajectory count and total collection time.

Initial / Final	**Robot**	**Traj. / Time**	**Task**
	AgileX PiPER 7-DoF	30 / 21 min	**RGB Block Sorting**: place red, green, and blue blocks from left to right in the specified order.
	AgileX PiPER 7-DoF	50 / 39 min	**Object Organization**: sort toys, stationery, and trash into their designated containers.
	ARX Lift2 14-DoF	30 / 5 min	**Trash Cleaning**: grasp trash, hand it between grippers, and place it in the bin.
	ARX Lift2 14-DoF	30 / 4 min	**Pen Holder Placement**: pick up stationery and place it in the pen holder.
	ARX Lift2 14-DoF	30 / 4 min	**Toy Box Organization**: place tabletop toys into the toy box and arrange them.
	ARX Lift2 14-DoF	90 / 31 min	**Sweep Trash**: use a broom to sweep trash into a dustpan.
	ARX Lift2 14-DoF	30 / 8 min	**Unzip Pencil Bag**: stabilize the bag, unzip it, and return it to the tabletop.

B.7 World and Action Predictions

To reveal how WSA_1 performs robot control, we present qualitative visualizations of the model's predictions on various tasks. As illustrated in Figure 5, the visualization covers four views: visual observation with text instruction, predicted future images, goal 3D world prediction with 3D point cloud and depth map, and trajectory of executed actions. From the results, we find that predicted visual plans well capture the task objective and well align the generated 3D world. Furthermore, driven by explicit 3D world modeling, the action trajectories are well-aligned with the predicted 3D goal states. These visual results demonstrate that WSA_1 unifies visual thinking, 3D world dynamics, and physical action generation within a shared latent space. It jointly models how actions alter the 3D world and how 3D world states guide action generation, thus leading to robust and generalizable manipulation behaviors.

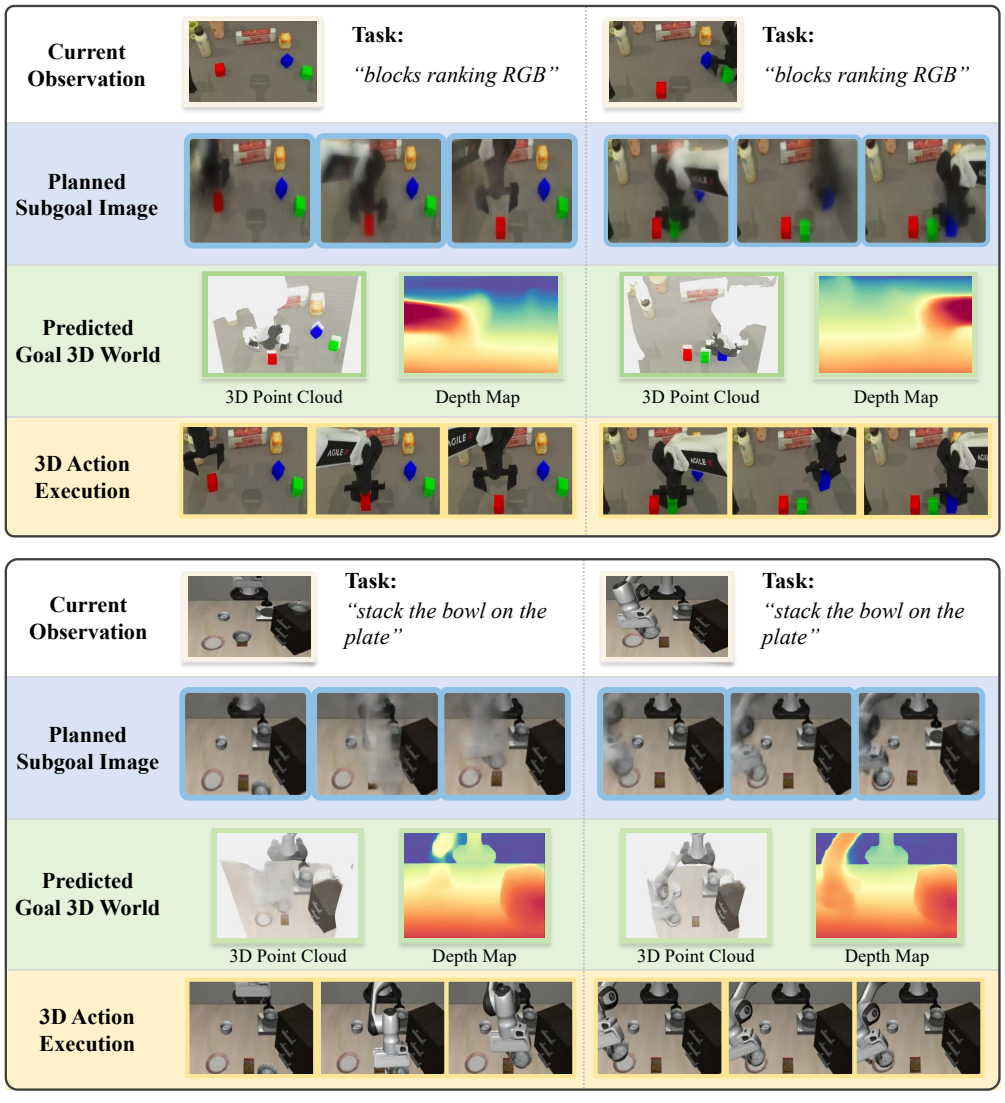

Figure 5: Representative WSA_1 predictions. Rows show the current observation and instruction, predicted subgoal image, depth and point-cloud visualizations obtained by decoding predicted 3D features with frozen DA3 heads, and action execution.

B.8 Cross-Embodiment Execution

Figure 6 shows execution sequences for three RoboTwin tasks on the source ALOHA embodiment and on ARX-X5 under zero-shot transfer.

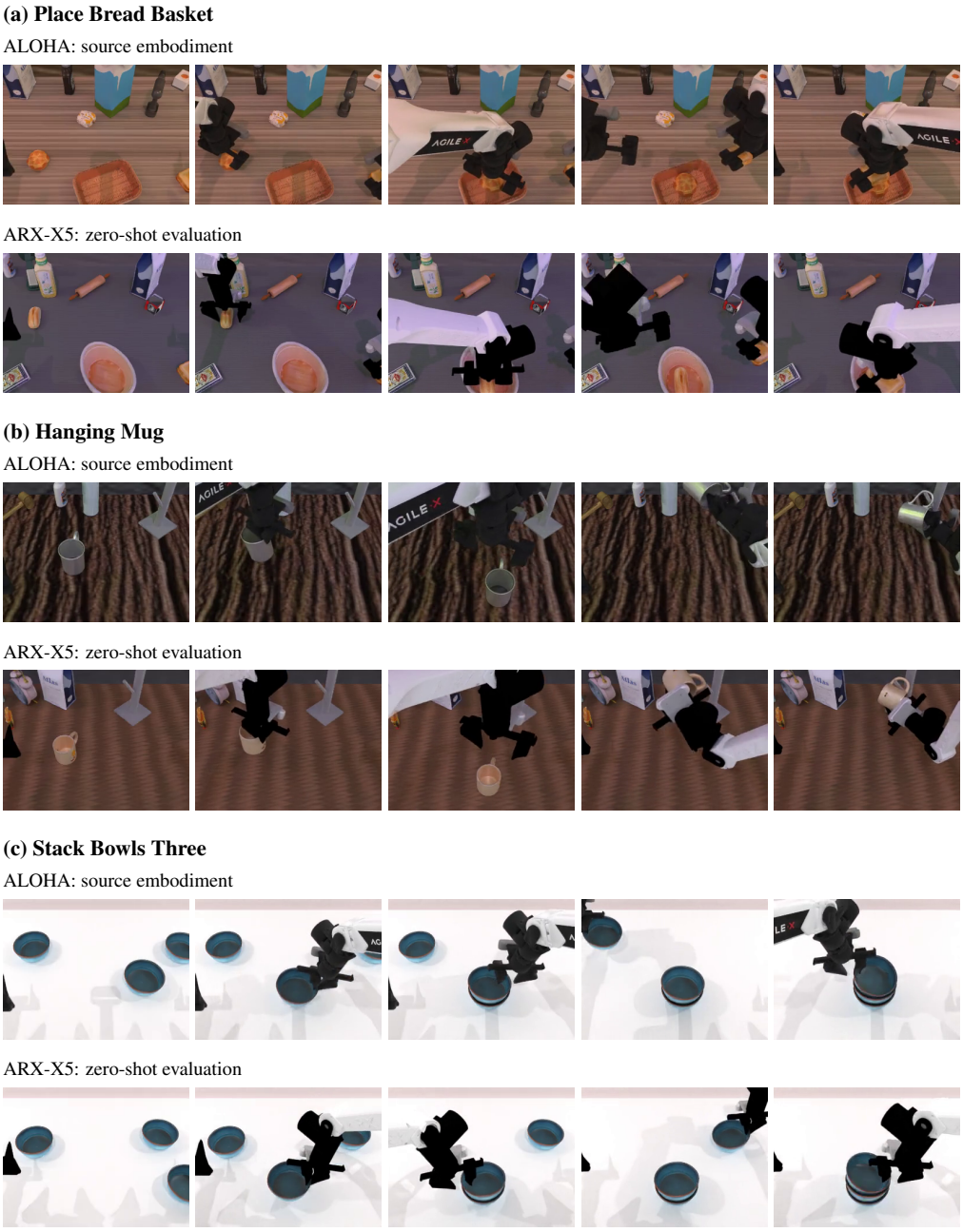

Figure 6: Cross-embodiment execution on three RoboTwin tasks: (a) *Place Bread Basket*, (b) *Hanging Mug*, and (c) *Stack Bowls Three*. Each task shows an ALOHA rollout and an ARX-X5 zero-shot rollout, with five frames progressing from left to right within each row. The ARX-X5 rollouts use a policy adapted on ALOHA demonstrations without target-embodiment policy fine-tuning.

B.9 Long-Horizon Real-Robot Execution

Figure 7 shows WSA_1-B performing *Pick Out the Green Blocks* on RoboChallenge 2.0 (Table30 V2) (Ma et al., 2026).

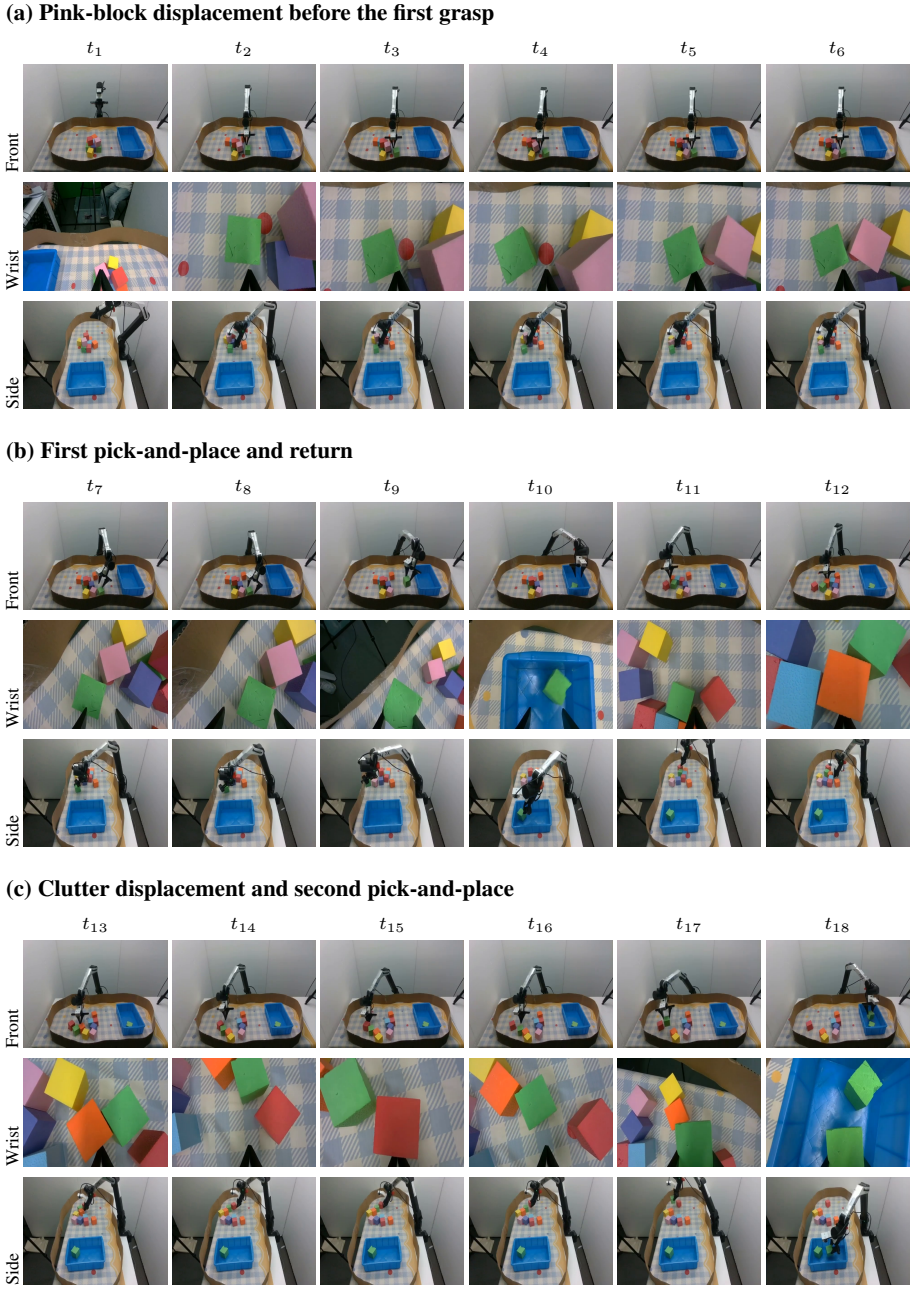

Figure 7: Long-horizon WSA_1-B execution of *Pick Out the Green Blocks* on RoboChallenge 2.0. Eighteen selected time points (t_1–t_{18}) progress from left to right within panels (a)–(c). Each time point is shown from synchronized front, wrist, and side views. The sequence captures the pink block being pushed aside before the first grasp, further displacement of non-target blocks, and two successive transfers of green blocks from the cluttered pile to the blue container.