BRIDGE-WA: LEARNING ACTION-RELEVANT WORLD DYNAMICS FOR ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

General-purpose vision-language-action (VLA) models leverage large-scale vision-language priors to understand scenes and instructions, but primarily generate actions directly from the current observation. World-action models (WAMs) further model future scene states, offering a broader view of how the environment may evolve. However, for robotic manipulation, predicting the entire future scene can introduce information beyond what is necessary for action generation; more importantly, effective actions require knowing not only **what** the scene may become, but also **where** and **how** relevant changes unfold. To bridge action generation with these action-relevant aspects of the world, we present **BRIDGE-WA**, a general world-action framework that learns complementary representations of future states, spatial changes, and local motion. Specifically, BRIDGE-WA consists of a Latent World Dynamics Module (LWDM) and WorldBridge. LWDM predicts future states, spatial changes, and local motion from VLM outputs, supervised by corresponding world targets. The WorldBridge are embedded into the action transformer and inject layer-specific combinations of these world priors through multi-source attention, spatiotemporal biases, and reliability-gated feature modulation. This design grounds action generation in action-relevant future dynamics while adaptively regulating world guidance, enabling robust generalization to visual disturbances and viewpoint shifts. We evaluate BRIDGE-WA on LIBERO-Plus, LIBERO-Dynamic, RoboTwin 2.0, VLABench, and real-world robots, where it achieves maximum success-rate improvements of 11.1%, 42.0%, 3.7%, 23.4%, and 11.1%, respectively, and achieves state-of-the-art average success rates on LIBERO-Dynamic, RoboTwin 2.0 and real-world. In particular, BRIDGE-WA demonstrates strong generalization to visual variations and viewpoint shifts in both simulation and real-world settings.

1 INTRODUCTION

General-purpose manipulation requires more than recognizing objects and parsing instructions. A robot must also reason about intervention: after it pushes, grasps, pours, or inserts an object, which regions should change, how should they move, and which visual details should remain irrelevant? Recent embodied foundation models and generalist vision-language-action (VLA) models have advanced through large-scale, cross-embodiment robot training and the transfer of web-scale vision-language priors to robot control (Brohan et al., 2023; O'Neill et al., 2024; Team et al., 2024; Bjorck et al., 2025; Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; Li et al., 2024; Black et al., 2025; Team et al., 2025; Cai et al., 2026; Ma et al., 2026a; Cheang et al., 2025; Bu et al., 2025; Wang et al., 2026a; Yuan et al., 2026a; Zheng et al., 2026; Wu et al., 2026). Action chunking and diffusion- or flow-based models further improve temporal consistency and continuous control (Zhao et al., 2023; Chi et al., 2025; Liu et al., 2025; Black et al., 2024; Kim et al., 2025; Black et al., 2026). Yet most direct VLA models still learn a reactive mapping from current observation and language to actions, leaving future consequences weakly represented and making policies vulnerable to shortcut correlations in backgrounds, colors, and camera-specific appearance, thereby leading to action errors and task failures, as shown in the Fig 1(a).

World models offer a natural way to close this gap: predict how the environment will evolve, then use that prediction to guide action. Latent dynamics, imagination-based agents, visual foresight, and robotic video prediction have shown that anticipated futures can support planning and manipu-

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.

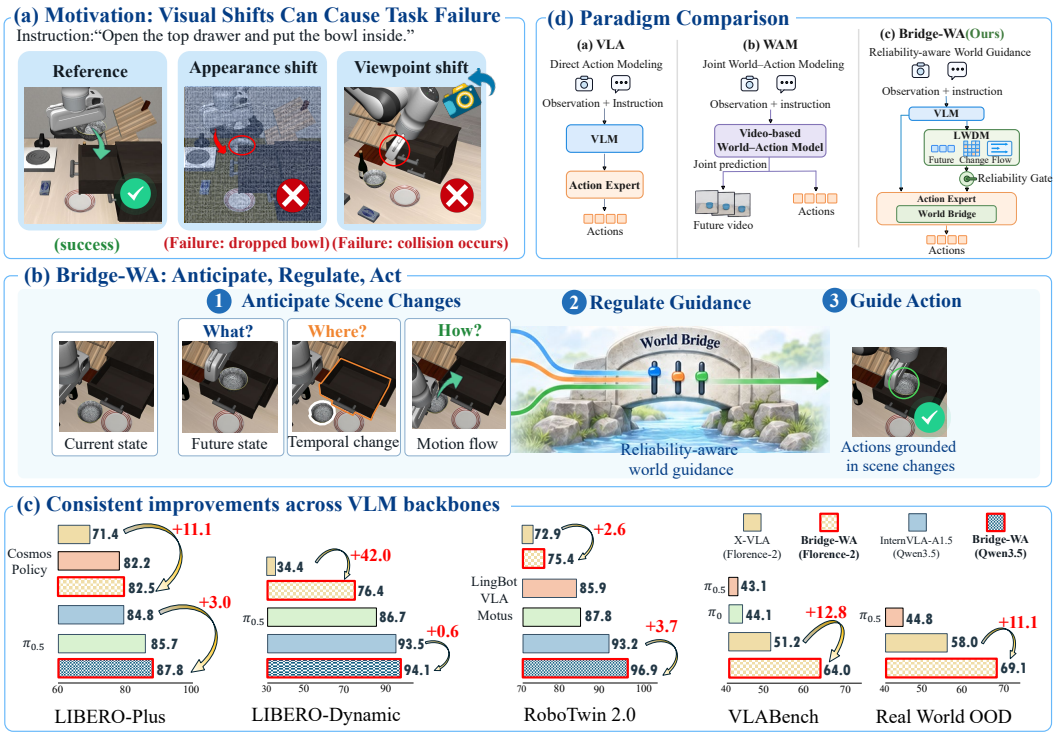


Figure 1: **(a)** Motivation: Appearance and viewpoint shifts can cause manipulation failures despite unchanged task semantics. **(b)** In contrast, Bridge-WA anticipates **what**, **where**, and **how** the scene will change, regulates the reliability of these world priors, and grounds actions in the predicted changes, enabling successful task completion under visual shifts. **(c)** BRIDGE-WA consistently improves different VLM backbones across four simulation and real-world benchmarks. **(d)** Unlike standard VLAs that directly predict actions and world-action models that jointly generate future video, BRIDGE-WA injects reliability-aware world priors into the action expert.

lation (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023; Schrittwieser et al., 2020; Finn & Levine, 2017; Ebert et al., 2018; Babaeizadeh et al., 2021; Du et al., 2023; Zhou et al., 2024). Recent world-action models extend this idea by jointly modeling future states and actions (Hou et al., 2026; Wang et al., 2026b; Luo et al., 2026; Liu et al., 2026; Bi et al., 2026; Yuan et al., 2026b; Kim et al., 2026; Zhao et al., 2026b; Ma et al., 2026b). The challenge is practical: running or decoding a large generative world model can be expensive, and dense image/video prediction allocates capacity to many pixels that do not affect control. For manipulation, the useful future is often smaller than a photorealistic rollout: it is a structured **change field** that tells the policy where the intended interaction takes effect and how the affected regions move.

To address these challenges, We propose BRIDGE-WA, a general world-action framework that conditions action generation on complementary world priors, as shown in the Fig 1(b). During training, we construct supervision directly from observed futures in robot demonstrations. Frozen WAN VAE (Wan et al., 2025) and DINOv3 (Siméoni et al., 2025) encoders provide future tokens T_f that capture complementary latent and visual structure, while differences between current and future features yield change maps M_c that highlight regions of scene change. Image-space optical flow provides motion maps M_o describing local displacement direction and magnitude. A Latent World Dynamics Module (LWDM) learn to recover the three types of priors from the current multimodal context. WorldBridge injects layer-specific combinations of these representations into the action expert through multi-source attention, change- and motion-aware attention biases, and reliability-gated feature modulation. This design grounds action generation in action-relevant future dynamics while adaptively regulating world guidance, enabling robust generalization to visual disturbances and viewpoint shifts, with its generalization performance illustrated in Fig.1(c).

Compared with direct VLA models (Black et al., 2025; Zheng et al., 2026; Ma et al., 2026a), it provides explicit future-aware structure. Compared with WAMs that require large future-model components at training or deployment (Yuan et al., 2026b; Bi et al., 2026; Ye et al., 2026; Li et al., 2026), it

keeps the online path close to a direct action transformer. By emphasizing change and motion rather than raw appearance, the policy is encouraged to ignore visually salient but causally irrelevant scene factors, a property aligned with prior work on robot visual representations and robustness under distribution shift (Nair et al., 2023; Ma et al., 2023; Karamcheti et al., 2023; Hendrycks & Dietterich, 2019; Hendrycks et al., 2021). Fig.1(d) presents a comparison of three model architectures.

Our contributions are summarized as follows:

- We propose BRIDGE-WA, a general world-action framework that learns complementary future-state, spatial-change, and local-motion priors from supervision extracted from observed robot demonstrations. By transferring these priors into compact current-context LWDM, our framework requires neither additional world-model pre-training nor video diffusion transformer execution during policy deployment.
- We introduce WorldBridge, which conditions the action transformer through layer-specific multi-source attention, change- and motion-aware attention biases, and reliability-gated feature modulation. Together with paired clean/augmented training and action-visual-world consistency, this design encourages action-relevant world guidance and stable action generation under observation perturbations.
- We demonstrate the generality of BRIDGE-WA across three VLM backbones, four simulation benchmarks and two real-world robots, achieving consistent out-of-distribution(OOD) gains under appearance and viewpoint shifts and state-of-the-art average success rates on LIBERO-Dynamic, RoboTwin 2.0, and real-world evaluations.

2 RELATED WORK

Vision-Language-Action Models. Vision-language-action (VLA) models connect visual perception and language understanding with robot action generation. Representative models such as RT-2, OpenVLA, π_0, $\pi_{0.5}$, and InternVLA, among others, advance generalist robot control through pretrained vision-language representations, large-scale robot training, and expressive action architectures (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; 2025; Bjorck et al., 2025; Li et al., 2024; Team et al., 2025; Cai et al., 2026; Ma et al., 2026a; Cheang et al., 2025; Bu et al., 2025; Wang et al., 2026a; Yuan et al., 2026a; Zheng et al., 2026; Wu et al., 2026). Action chunking and diffusion- or flow-based action modeling further improve the expressiveness and temporal structure of predicted actions (Zhao et al., 2023; Chi et al., 2025; Liu et al., 2025; Black et al., 2024; Kim et al., 2025; Black et al., 2026). Despite these advances, many direct VLA models are trained primarily to predict demonstrated actions, with limited explicit supervision for the resulting scene evolution. This can leave models reliant on appearance correlations: changes in background, lighting, or camera viewpoint may disrupt action prediction even when the task and required interaction remain unchanged. Stronger semantic representations and more expressive action decoders do not, by themselves, ensure that models distinguish task-relevant changes from incidental visual variation. BRIDGE-WA addresses this limitation by guiding action generation with complementary future-state, spatial-change, and local-motion priors, encouraging the model to attend to anticipated task-relevant changes rather than incidental appearance.

World Models for Robotics. World models provide predictive representations for planning, policy learning, evaluation, and synthetic data generation in robotics (Hou et al., 2026). Visual foresight, video prediction, and future-generation methods show that anticipated observations can guide manipulation and improve generalization (Finn & Levine, 2017; Ebert et al., 2018; Babaeizadeh et al., 2021; Du et al., 2023; Zhou et al., 2024; Wu et al., 2024; Zhou et al., 2026). Recent world-action models further couple future-state modeling with action generation through latent future reasoning or object-addressable states (Wang et al., 2026b; Luo et al., 2026; Liu et al., 2026; Yuan et al., 2026b). However, predicting the visual future is not the same as extracting what matters for control: full image/video rollouts can devote substantial computation to reconstructing appearance rather than representing action-relevant changes. Executing a large world model online further burdens every control decision with predictive computation. BRIDGE-WA instead distills world supervision into compact outcome, change, and motion priors that directly guide action generation, without reconstructing future images or running a large world model at deployment.

Learning Action-Relevant Representations. A complementary line of work learns explicit spatial and actionable representations for manipulation. Transporter Networks, CLIPort, and

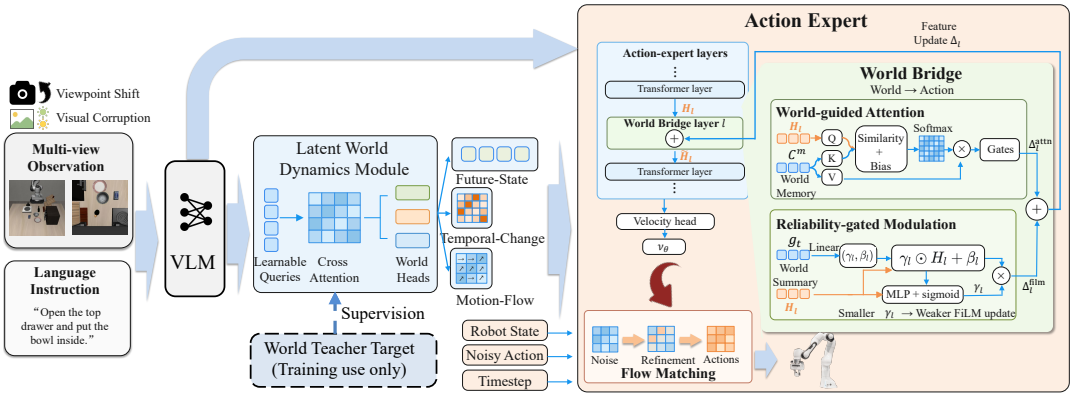


Figure 2: **The BRIDGE-WA framework.** A LWDM learns future, change, and flow representations from training-only supervision. WorldBridge, the world-to-action interface, couples these priors to the action transformer through world-guided attention and reliability-gated FiLM. Flow matching produces the action chunk. Complementary conditioning paths are detailed in Appendix B.

Where2Act encode spatial correspondence, separate semantic "what" from spatial "where", or identify potential interaction regions (Zeng et al., 2021; Shridhar et al., 2022; Mo et al., 2021). PerAct, RVT, and Act3D further exploit 3D structure and multi-view representations for spatially precise action prediction (Shridhar et al., 2023; Goyal et al., 2023; Gervet et al., 2023). Representation-learning methods such as R3M, VIP, and Voltron investigate visual features that support downstream robot control (Nair et al., 2023; Ma et al., 2023; Karamcheti et al., 2023). Recent world-action models further learn predictive representations through object-addressable dynamics, geometry-aware policy spaces, or compressed world-generator features (Liu et al., 2026; Zhao et al., 2026a; Zeng et al., 2026). Unlike these approaches, BRIDGE-WA explicitly represents complementary future outcomes, spatial changes, and local motion, providing structured world guidance for action generation.

3 BRIDGE-WA

3.1 OVERVIEW

Problem Formulation We formulate robotic manipulation as conditional action-sequence generation. Given current multi-view images $I_t = \{I_t^{(v)}\}_{v=1}^{V}$, a language instruction ℓ, and robot state s_t, a policy π_θ generates a chunk of T_a future actions:

$$A_t = (a_t, \ldots, a_{t+T_a-1}) = \pi_\theta(I_t, \ell, s_t), \qquad a_t \in \mathbb{R}^{d_a}. \tag{1}$$

Here t is the observation timestep, v indexes the V camera views, and d_a is the robot's action dimension; actions control end-effector or joint motion and the gripper. The policy executes a prefix of the chunk before replanning from new observations.

Overview BRIDGE-WA is a general world-action framework that unifies predictive world representation, world-conditioned action generation, and robust policy learning (Fig. 2). A vision-language backbone first encodes the current context. A latent world predictor then estimates **what** the scene may become, **where** it changes, and **how** local content moves. WorldBridge, the framework's world-to-action interface, selectively integrates these complementary priors into a flow-matching action transformer. Matched clean/augmented training promotes consistent actions under observation perturbations. In the following sections, we detail the design of: 1) **Latent World Dynamics Module**, 2) **WorldBridge**, and 3) **Policy Learning**.

3.2 LATENT WORLD DYNAMICS MODULE

Future-World Supervision We distill complementary future-world information from robot demonstrations into a world predictor. Frozen WAN-VAE and DINOv3 encoders (Wan et al., 2025; Siméoni et al., 2025) extract future tokens T_f from observed future frames. Normalized current-to-future feature differences yield change maps M_c, while optical flow provides motion maps M_o. Together, these targets describe future outcomes, spatial changes, and local displacements. Targets

are cached offline and fused across valid prediction horizons (Appendix B), providing training-only supervision.

Predictive Representation Distillation Conditioned on the current multimodal context, five learned-query heads predict WAN/DINO future tokens, WAN/DINO change maps, and motion flow. Regression and cosine-alignment losses form the distillation objective $\mathcal{L}_W$, transferring future-world information into representations predictable from current inputs. The predictions are organized into three memories: C^F combines predicted WAN/DINO features, C^C fuses the two change predictions, and C^M encodes motion. At inference, the policy uses these current-context representations without future observations or teacher encoders.

3.3 WorldBridge: Selective World-to-Action Conditioning

Layer-Specific Attention Within BRIDGE-WA, WorldBridge connects the predicted world representation to action generation. It updates the action-token stream while preserving the multimodal context pathway. At transformer layer l, action tokens H_l query each active world memory:

$$R_l^m = \mathrm{softmax}\left(\frac{(H_l W_Q^m)(C^m W_K^m)^\top}{\sqrt{d_h}} + \mathbf{1}(b^m)^\top\right)(C^m W_V^m)W_O^m, \quad m \in \{F, C, M\}. \tag{2}$$

Here $H_l \in \mathbb{R}^{N_a \times d}$ contains the hidden representations of the N_a action tokens entering action-transformer layer l, and d is the hidden width. The index m denotes future (F), change (C), or motion (M) memory C^m. The matrices $W_Q^m, W_K^m, W_V^m, W_O^m$ are attention projections, d_h is the head width, and b^m is broadcast over action queries. Softmax normalizes over memory tokens, yielding the ungated update R_l^m. Future attention uses no extra bias; change scores and motion magnitudes provide key biases for the other two memories. These biases prioritize anticipated changes without imposing explicit temporal positional encodings. Active memories are then gated and applied sequentially:

$$H_l^{(0)} = H_l, \qquad H_l^{(j)} = H_l^{(j-1)} + \Delta_l^{m_j}\left(H_l^{(j-1)}, C^{m_j}\right), \qquad \Delta_l^{\mathrm{attn}} = H_l^{(J)} - H_l. \tag{3}$$

Here $\mathcal{M}_l = (m_1, \ldots, m_J)$ is the ordered sequence of the J memories active at layer l, and j denotes its position in that sequence. $H_l^{(j)}$ is the action-token stream after the first j memories, initialized by $H_l^{(0)} = H_l$. Each $\Delta_l^{m_j}$ applies modality, layer, channel, and reliability gates to $R_l^{m_j}$; their accumulated change is Δ_l^{attn}. A layer schedule emphasizes future information earlier and change/motion refinement later. Modalities are applied sequentially, with shared attention weights but distinct layer gates. Each residual is scaled by modality, layer, and channel gates, together with a learned reliability score that compares pooled action and world features. Appendix B specifies the gates and layer schedule.

Reliability-Gated Modulation We complement token-level attention with reliability-gated FiLM conditioning (Perez et al., 2018):

$$\Delta_l^{\mathrm{film}} = \eta r_l(\gamma_l \odot H_l + \beta_l), \qquad \Delta_l = \Delta_l^{\mathrm{attn}} + \Delta_l^{\mathrm{film}}, \qquad \widetilde{H}_l = H_l + \Delta_l + \mathcal{A}(\Delta_l), \tag{4}$$

Here g_t is the fused global world summary(Fig.2), from which layer-specific channel-wise scale and shift vectors (γ_l, β_l) are derived. The scalar η sets residual strength, and $r_l \in (0, 1)$ gates the FiLM residual Δ_l^{film} using action–modulation compatibility. The total world update is $\Delta_l = \Delta_l^{\mathrm{attn}} + \Delta_l^{\mathrm{film}}$. The learned adapter $\mathcal{A}$ further processes this update, producing the conditioned action representation $\widetilde{H}_l$. Reliability depends on action and modulation features, allowing weaker updates when world guidance is less compatible with the action stream; it is not a calibrated uncertainty estimate. Zero-initialized residual output projections provide a stable initialization for training. Language-conditioned, address-keyed visual slots supply current scene evidence, while bounded state-dependent scalars adjust guidance strength (Appendix B).

3.4 Robust Policy Learning

Matched Clean and Augmented Training We train clean and augmented observations of the same demonstration with shared action labels, flow noise ϵ, and interpolation time τ. Mild robot-state, language, and appearance perturbations preserve task supervision. State corruption is aligned between numeric and tokenized channels; appearance perturbations do not relocate objects or modify action labels. A detached, action-gradient-directed visual perturbation additionally regularizes

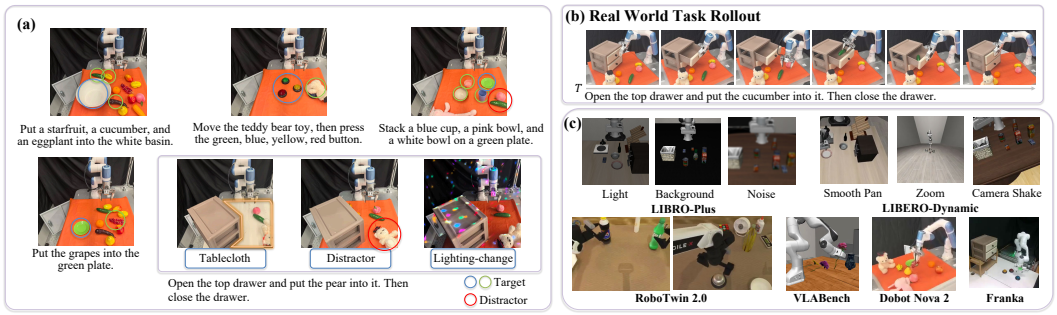


Figure 3: **Evaluation overview.** (a) Real-world tasks with diverse unseen objects and visual disturbances. (b) A real-world rollout of the "Put Item in Drawer" task. (c) Simulation benchmarks and real-world robots used in our evaluation.

feature sensitivity. Both branches learn the native flow-matching velocity (Lipman et al., 2023):

$$X_\tau = (1-\tau)A_t + \tau\epsilon, \qquad \mathcal{L}_{\text{act}} = \mathbb{E}\|v_\theta(X_\tau, \tau; I_t, \ell, s_t) - (\epsilon - A_t)\|^2. \tag{5}$$

Here ϵ is sampled Gaussian noise with the shape of A_t, and $\tau \in [0,1]$ is interpolation time, distinct from observation time t. X_τ interpolates from actions to noise; v_θ predicts its target velocity $\epsilon - A_t$. The expectation averages squared error over demonstrations, noise, and τ. We also align augmented outputs with detached clean targets: SmoothL1 on action velocities and cosine losses on pooled visual features, fused world memories, and visual slots. Denoting their weighted sum by $\mathcal{L}_{\text{con}}$, the objective is

$$\mathcal{L} = \lambda_a\left[(1-\alpha_n)\mathcal{L}^c_{\text{act}} + \alpha_n\mathcal{L}^a_{\text{act}}\right] + \tfrac{1}{2}(\mathcal{L}^c_W + \mathcal{L}^a_W) + \rho_n\mathcal{L}_{\text{con}}, \tag{6}$$

Here c/a denote clean/augmented branches, λ_a weights action supervision, and $\mathcal{L}_W$ is world-prior distillation. At optimizer update n, α_n balances the two action losses and ρ_n ramps consistency strength; the two world losses are equally weighted. Demonstration reweighting, feature perturbation, and the consistency terms are detailed in Appendix C.

4 Experiments

4.1 Simulation Benchmarks and Real-World Setup

Simulation Benchmarks We evaluate BRIDGE-WA on **LIBERO-Plus**, our **LIBERO-Dynamic**, **RoboTwin 2.0**, and **VLABench** (Fig. 3(c)). Together, these benchmarks test static visual shifts, time-varying viewpoints, bimanual domain randomization, and systematic single-arm generalization. We follow each benchmark's official training and evaluation protocol and report success rate, together with intention and progress scores where defined. Exact datasets, observation and action spaces, action horizons, perturbations, and episode counts are preserved in Appendix D.1.

Real-World We evaluate on the **Dobot Nova 2** and **Franka** robots with an Intel RealSense third-person camera and a wrist camera. Demonstrations are stored with RGB images, cartesian robot states, and absolute cartesian actions, with RGB standardized to 640×480. The suite contains seven tasks, as shown in the Fig.3(a)(b). Easy evaluation uses 10 in-distribution trials per task; Hard evaluation uses 15 trials per task across tablecloth, distractor-object, and lighting shifts, with five trials per factor. We report per-task and average success rates for each track. More real-world robot experimental setups are provided in the Appendix D.2.

Experimental Details Experimental details, including hardware, VLM backbones, model sizes, and training configurations, are provided in Appendix E.

4.2 Main Results

On LIBERO-Plus, LIBERO-Dynamic As shown in Tables 1 and 2, BRIDGE-WA consistently improves three heterogeneous VLA backbones—X-VLA, $\pi_{0.5}$, and InternVLA-A1.5—built on Florence-2, PaliGemma, and Qwen3.5, respectively. On LIBERO-Plus, it raises their weighted success rates by 11.1, 3.0, and 3.2 points, achieving a best overall score of 88.7 without using any LIBERO-Plus data. The largest gains occur under camera, robot-state, and visual-noise shifts, where policies must distinguish task-relevant changes from nuisance observations. On LIBERO-Dynamic, BRIDGE-WA improves X-VLA by 42.01 points and further raises the strong InternVLA-A1.5 baseline across every camera-motion profile. These consistent gains across different VLM scales and policy architectures demonstrate that BRIDGE-WA provides a general world-to-action interface rather than a backbone-specific modification.

Table 1: **Performance comparison on LIBERO-Plus.** Success rates (%) across LIBERO-Plus variants and their weighted average. Δ denotes the gain over the corresponding backbone. $*$ and § indicate Florence-2 and PaliGemma VLMs, respectively.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
World Action Models								
WorldVLA(Cen et al., 2025)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
Fast-WAM(Yuan et al., 2026b)	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
SG-WAM(Zhao et al., 2026a)	58.6	48.9	81.4	89.8	86.1	80.7	74.2	73.0
Faster-WAM(Zhao et al., 2026b)	53.8	71.6	**94.7**	96.3	61.3	63.6	79.1	73.6
Faster-WAM(Ma et al., 2026b)	67.9	49.0	92.1	94.3	57.0	82.3	82.7	75.0
VLA-JEPA(Sun et al., 2026)	64.2	67.7	88.1	91.8	93.4	65.8	83.9	79.5
GE-Act(Liao et al., 2026)	60.7	77.0	77.4	95.8	86.0	90.9	80.2	80.3
Cosmos-Policy(Kim et al., 2026)	75.8	63.3	81.7	96.5	88.9	92.7	82.2	82.2
ImageWAM(Zhang et al., 2026)	80.8	50.3	91.4	98.1	85.5	93.8	80.5	83.1
OA-WAM(Liu et al., 2026)	80.5	89.6	85.3	96.5	95.9	75.6	82.8	83.9
Vision-Language-Action Models								
OpenVLA(Kim et al., 2024)	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
NORA(Hung et al., 2025)	2.2	37.0	65.1	45.7	58.6	12.8	62.1	39.0
UniVLA(Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0Black et al. (2024)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-Fast(Pertsch et al., 2025)	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA(Tan et al., 2025)	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT(Kim et al., 2025)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
AVA-VLA(Xiao et al., 2026)	55.5	25.9	85.6	95.5	88.9	78.0	74.1	70.1
HoloBrain-0(Lin et al., 2026)	65.5	58.2	78.7	88.1	90.3	66.9	79.5	74.0
StarVLA(Community, 2026)	52.5	49.8	88.5	95.7	95.7	73.0	76.9	74.1
ABot-M0(Yang et al., 2026)	60.4	67.9	86.4	96.2	91.6	86.4	82.6	80.5
X-VLA(Zheng et al., 2026)	23.4	89.7	75.7	88.2	96.0	62.7	71.8	71.4
BRIDGE-WA* (Ours)	60.7	**90.6**	81.0	98.6	95.5	86.9	72.5	82.5
Δ	+37.3	+0.9	+5.3	+10.4	−0.5	+24.2	+0.7	+11.1
$\pi_{0.5}$(Black et al., 2025)	75.4	77.5	85.6	96.9	94.6	89.7	85.7	85.7
BRIDGE-WA§ (Ours)	77.2	85.6	86.7	98.3	97.9	92.5	**88.5**	**88.7**
Δ	+1.8	+8.1	+1.1	+1.3	+3.3	+2.8	+2.8	+3.0
InternVLA-A1.5(Ma et al., 2026a)	83.1	55.1	86.9	96.4	98.2	95.6	85.2	84.8
BRIDGE-WA (Ours)	**89.8**	67.0	85.7	**98.7**	**98.5**	**98.5**	83.6	88.0
Δ	+6.7	+11.9	−1.2	+2.3	+0.3	+2.9	−1.6	+3.2

Table 2: **Performance comparison on LIBERO-Dynamic.** Success rates (%) across LIBERO-Dynamic camera-motion profiles and their average.

Method	Smooth Pan	Orbit	Zoom	View Jump	Camera Shake	Total
X-VLA(Zheng et al., 2026)	37.65	38.40	27.45	21.30	46.95	34.35
BRIDGE-WA* (Ours)	81.65	77.80	70.25	72.90	79.20	76.36
Δ	+44.00	+39.40	+42.80	+51.60	+32.25	+42.01
InternVLA-A1.5(Ma et al., 2026a)	93.10	93.30	92.35	93.35	95.15	93.45
BRIDGE-WA (Ours)	**93.20**	**93.85**	**93.90**	**93.50**	**96.05**	**94.10**
Δ	+0.10	+0.55	+1.55	+0.15	+0.90	+0.65

On RoboTwin 2.0 Table 3 shows that BRIDGE-WA consistently improves language-conditioned bimanual manipulation. It improves X-VLA by 2.6 points and InternVLA-A1.5 by 3.7 points, reaching $97.1/96.6$ success rates on the Easy/Hard settings and a state-of-the-art average of 96.9. Importantly, the gain is preserved under the Hard setting, which jointly varies distractors, textures, lighting, tabletop height, and instructions. The improvement on a bimanual platform with 14-dimensional actions and a 50-step action horizon indicates that structured world priors extend beyond single-arm, short-horizon control.

Table 3: **Performance on RoboTwin 2.0.** Success rates (%) on the Easy and Hard settings of RoboTwin 2.0 and their average.

Method	Easy	Hard	Average
World Action Models			
Motus(Bi et al., 2026)	88.7	87.0	87.8
Faster-WAM(Ma et al., 2026b)	89.7	88.6	89.2
Being-H0.7(Luo et al., 2026)	90.2	89.6	89.9
Fast-WAM(Yuan et al., 2026b)	91.9	91.8	91.8
LingBot-VA(Li et al., 2026)	92.9	91.5	92.2
Enfold(Zeng et al., 2026)	91.6	91.9	91.8
Vision-Language-Action Models			
π_0(Black et al., 2024)	65.9	58.4	62.2
StarVLA(Community, 2026)	88.7	87.8	88.3
$\pi_{0.5}$(Black et al., 2025)	82.7	76.8	79.8
LingBot-VLA(Wu et al., 2026)	86.5	85.3	85.9
InternVLA-A1(Cai et al., 2026)	89.4	89.6	89.5
X-VLA(Zheng et al., 2026)	72.9	72.8	72.9
BRIDGE-WA* (Ours)	75.2	75.7	75.4
Δ	+2.3	+2.9	+2.6
InternVLA-A1.5(Ma et al., 2026a)	93.3	93.0	93.2
BRIDGE-WA (Ours)	**97.1**	**96.6**	**96.9**
Δ	+3.8	+3.6	+3.7

On VLABench As reported in Table 4, BRIDGE-WA substantially improves both task completion and execution progress across all five VLABench tracks. Average SR increases from 29.4 to 52.8 and PS from 51.2 to 64.0, with particularly large SR gains on In-Distribution (+38.0), Semantic Instruction (+28.0), and Unseen Texture (+21.0). In contrast, the average intention score changes only marginally. This suggests that the improvement does not primarily arise from better

Table 4: **Performance comparison on VLABench.**

Method	In-Dist.			Cross-Cat.			Common Sense			Semantic Instr.			Unseen Texture			Avg.		
	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS
$\pi_{0.5}$ (Black et al., 2025)	40.6	75.0	60.8	22.6	49.6	35.3	18.0	57.5	41.6	16.1	57.1	30.3	25.6	62.0	47.4	24.6	60.2	43.1
π_0(Black et al., 2024)	47.0	67.8	62.7	21.2	44.0	33.6	29.1	54.9	43.0	17.3	58.0	38.7	32.2	50.6	42.5	29.4	55.0	44.1
X-VLA(Zheng et al., 2026)	40.0	88.0	66.5	11.0	42.0	24.0	33.0	66.0	51.7	39.0	82.0	62.5	24.0	73.0	51.2	29.4	70.2	51.2
BRIDGE-WA* (Ours)	78.0	85.0	85.8	23.0	39.0	28.8	51.1	74.2	64.4	67.0	82.0	80.3	45.0	76.0	60.3	52.8	71.2	64.0
Δ	+38.0	−3.0	+19.3	+12.0	−3.0	+4.8	+18.1	+8.2	+12.7	+28.0	+0.0	+17.8	+21.0	+3.0	+9.1	+23.4	+1.0	+12.8

Table 5: **Component masking ablation on VLABench.** F, C, and O denote future, change, and flow priors, respectively.

World Prior			In-Dist.		Cross-Cat.		Common Sense		Semantic Instr.		Unseen Texture		Avg.	
F	C	O	SR	PS	SR	PS	SR	PS	SR	PS	SR	PS	SR	PS
			17.0	51.0	12.0	28.2	14.0	47.3	15.0	48.7	15.0	45.2	14.6	44.1
✓			73.0	82.5	18.0	24.3	43.0	57.0	60.0	74.3	42.0	57.7	47.2	59.2
	✓		34.0	63.3	19.0	31.2	27.0	53.3	34.0	61.8	32.0	53.7	29.2	52.7
		✓	20.0	31.7	4.0	10.8	8.0	18.7	22.0	30.0	7.0	18.5	12.2	21.9
✓	✓		76.0	85.0	**25.0**	**31.5**	46.0	61.7	65.0	78.7	43.0	61.2	51.0	63.6
	✓	✓	45.0	66.3	6.0	16.8	24.0	44.5	38.0	58.8	26.0	44.3	27.8	46.2
✓		✓	77.0	**86.3**	21.0	28.0	47.0	60.8	66.0	79.8	42.0	59.3	50.6	62.9
✓	✓	✓	**78.0**	85.8	23.0	28.8	**51.1**	**64.4**	**67.0**	**80.3**	**45.0**	**60.3**	**52.8**	**64.0**

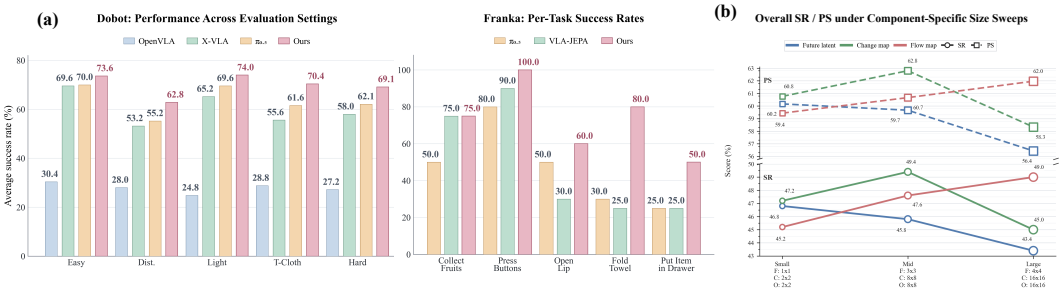


Figure 4: **Real-world evaluation and component analysis.** (a) Performance across visual disturbances on Dobot and per-task success rates on Franka. The detailed success rates on Dobot robot are reported in the Appendix F. (b) Overall SR and PS under different resolutions of the future, change, and flow representations on VLABench.

instruction recognition, but from translating semantic intent into more effective actions. By jointly representing the intended future, spatial support of change, and local motion, BRIDGE-WA improves execution under semantic, categorical, and appearance shifts.

On Real-World Robots Fig.4(a) show that the simulation gains transfer to both Dobot Nova 2 and Franka robots. On Dobot, BRIDGE-WA improves the average Easy success rate from 69.6 to 73.6 and the aggregated Hard result from 58.0 to 69.1, with the larger OOD gain indicating improved robustness to distractors, lighting, and tablecloth changes. On Franka, it achieves a 73.0 average success rate, compared with 49.0 for the strongest baseline. These gains are obtained with only 50 demonstrations per task, supporting the claim that action-relevant world priors suppress nuisance appearance while preserving the scene dynamics required for robust execution.

4.3 ABLATION STUDIES

World Component Analysis Table 5 reveals complementary but asymmetric roles of the three world priors. Future representations provide the strongest individual contribution, improving average SR/PS from 14.6/44.1 to 47.2/59.2, confirming that anticipated outcomes are the primary guidance for action generation. Change maps also yield clear gains by localizing where interactions should occur, whereas flow alone is insufficient because local motion is ambiguous without task outcome and spatial context. When combined with future and change priors, however, flow provides additional motion guidance. The full model achieves the best overall SR/PS of 52.8/64.0 and the strongest results on Common Sense, Semantic Instruction, and Unseen Texture. These results support our "what–where–how" decomposition: future priors define the goal, change priors localize its spatial support, and flow priors refine how the change should unfold.

WorldBridge Architecture Analysis Table 6 shows that WorldBridge benefits primarily from reliability-aware integration rather than unregulated feature injection.

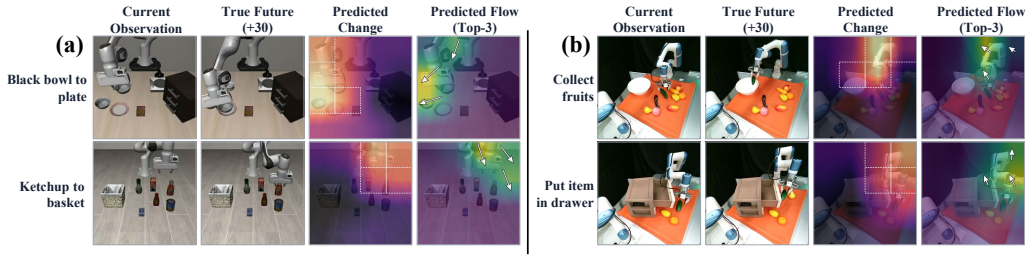


Figure 5: **Spatial interpretation.** Given the current observation, BRIDGE-WA predicts action-relevant change regions and the dominant 2-D motion vectors. The predicted changes concentrate on the robot and manipulated objects, while the flow directions are consistent with the observed future motion.

Table 6: **WorldBridge Ablation results on LIBERO-Plus.**

Config	Total
Full Model	**82.45**
w/o Multi-Source Attention	81.28
w/o Spatiotemporal Attention Bias	82.13
w/o Modulation Reliability Gate	75.87
w/o Visual Invariance Adapter	82.24
w/o All Reliability Gates	71.34

Removing the modulation reliability gate reduces performance by 6.58 points, while removing reliability gating from both guided attention and feature modulation causes the largest drop of 11.11 points. This confirms that predicted world priors have input-dependent reliability and must be adaptively regulated before influencing action generation. Multi-source attention contributes 1.17 points by allowing the action transformer to access complementary world memories separately. The smaller drops from removing the spatiotemporal attention bias or visual invariance adapter suggest that their functions are partially supported by content-based attention and robust training, while their inclusion still improves the complete model. Overall, reliability gating is the central mechanism that converts predictive world features into stable action guidance.

Future-Prediction Transfer Analysis WorldBridge captures predictive information beyond current VLM features. Diagnostic heads trained on LIBERO and evaluated zero-shot on LIBERO-Plus show consistently lower changed-region prediction error across future horizons, including a 2.21% relative reduction at horizon +30. Positive improvements across all perturbation families demonstrate transferable future-state modeling under distribution shifts; full results are provided in Appendix F.2.

World-prior Size Sweep Analysis The size sweep in Fig. 4(b) shows that each prior needs a different spatial resolution budget. Future latents work best as compact global summaries: 1×1 reaches 46.8% SR and 60.2% PS, while 4×4 drops to 43.4% SR and 56.4% PS. Change maps peak at the intermediate 8×8 scale (49.4% SR / 62.8% PS), balancing localization and stability. Flow maps benefit from finer resolution, improving from 45.2% SR at 2×2 to 49.0% SR at 16×16. Overall, BRIDGE-WA benefits from asymmetric capacity allocation: compact future tokens, medium-resolution change maps, and higher-resolution flow maps.

4.4 VISUALIZATION ANALYSIS

We visualize the world priors learned by BRIDGE-WA in both simulation and real-world environments. As shown in Fig. 5(a)(b), the change maps identify **where** action-relevant changes occur, while suppressing backgrounds and distractors; the flow vectors capture **how** manipulated regions move and align with the observed future motion. Together with future-state representations of **what** the scene may become, these priors focus action generation on relevant dynamics, improving robustness to visual and viewpoint shifts.

5 CONCLUSION AND LIMITATIONS

We presented BRIDGE-WA, a general world-action framework that replaces dense future-scene generation with action-relevant representations of **what** may happen, **where** changes occur, and **how** they unfold. LWDM predicts these priors from the current context, while WorldBridge selectively injects them into the action transformer through reliability-aware conditioning. Across three VLA backbones, four simulation benchmarks, and two real-robot platforms, BRIDGE-WA consistently improves manipulation robustness, particularly under visual and viewpoint shifts. However, the ablations further confirm the complementarity of the world priors and the importance of reliability gating. The learned priors depend on demonstration coverage and may not fully capture occluded 3D dynamics. Future work will explore geometry-aware priors and broader cross-embodiment training.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation; to edit and polish the manuscript for grammar, clarity, and readability; and to modify and debug software code. We did not use generative AI tools to generate synthetic data, develop theoretical or conceptual frameworks, formulate mathematical claims, assist with proofs, propose or refine research hypotheses, design research methods or experiments, clean or reformat datasets, conduct qualitative or thematic analysis, or interpret experimental results. All AI-assisted text was carefully reviewed and revised by the authors to ensure accuracy and preserve the intended meaning. All AI-assisted code modifications were manually reviewed and tested by the authors for correctness. We take full responsibility for the final content of this work, including all text, code, claims, and artifacts produced or revised with the assistance of generative AI.

REFERENCES

Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. Pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. Pi0.5: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

Kevin Black, Manuel Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *Advances in Neural Information Processing Systems*, 38:33383–33407, 2026.

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

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

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.

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

Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.

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

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

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

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.

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

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3949–3965. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/gervet23a.html.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pp. 694–710. PMLR, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, et al. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *Robotics: Science and Systems*, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

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

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

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 88446–88463, 2026.

Xuewu Lin, Tianwei Lin, Yun Du, Hongyu Xie, Yiwei Jin, Jiawei Li, Shijie Wu, Qingze Wang, Mengdi Li, Mengao Zhao, et al. Holobrain-0 technical report. *arXiv preprint arXiv:2602.12062*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

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

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025.

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

Jingxian Lu, Wenke Xia, Yuxuan Wu, Zhiwu Lu, and Di Hu. When would vision-proprioception policies fail in robotic manipulation? *arXiv preprint arXiv:2602.12032*, 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.

Haoxiang Ma, Junhao Cai, Xiaoxu Xu, Hao Li, Yuyin Yang, Yang Tian, Jiafei Cao, Hongrui Zhu, Zherui Qiu, Yuqiang Yang, et al. Internvla-a1. 5: Unifying understanding, latent foresight, and action for compositional generalization. *arXiv preprint arXiv:2607.04988*, 2026a.

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

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6813–6823, 2021.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pp. 894–906. PMLR, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. In *European Conference on Computer Vision*, pp. 478–497. Springer, 2026.

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

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

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

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026a.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, et al. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026b.

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

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024.

Lei Xiao, Jifeng Li, Juntao Gao, Feiyang Ye, Yan Jin, Jingjing Qian, Jing Zhang, Yong Wu, and Xiaoyuan Yu. Ava-vla: improving vision-language-action models with active visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13453–13463, 2026.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pp. 2825–2834. PMLR, 2018.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

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

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026a.

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.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pp. 726–747. PMLR, 2021.

Weili Zeng, Yitong Xing, Fulong Liu, Chengqun Yang, Antao Xiang, Feng Tian, Jingnan Gao, Jisong Cai, Xin Wang, Xiaomin Wu, et al. Enfold: Folding world model imagination into predictive representations for ultra-efficient embodied control. *arXiv preprint arXiv:2607.26657*, 2026.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks, 2024a. *URL https://arxiv.org/abs/2412.18194*.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. Imagewam: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026.

Juntu Zhao, Wenbo Lu, Di Zhang, Yufeng Liu, Yushen Liang, Tianluo Zhang, Yifeng Cao, Junyuan Xie, Yingdong Hu, Shengjie Wang, et al. Do you need proprioceptive states in visuomotor policies? *arXiv preprint arXiv:2509.18644*, 2025.

Ruiteng Zhao, Zhengshen Zhang, Yue Su, Wenshuo Wang, Jiahui Li, Zhiyuan Yang, Francis EH Tay, Marcelo H Ang Jr, and Haiyue Zhu. Sg-wam: Self-guided world modeling in geometry-aware policy space. *arXiv preprint arXiv:2608.01397*, 2026a.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

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

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

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. Gem-4d: Geometry-enhanced video world models for robot manipulation. *arXiv preprint arXiv:2605.22882*, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, pp. 61885–61896. PMLR, 2024.

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

Appendix

A Notation and Symbol Definitions

Table 7 lists additional variables introduced in the appendices.

Table 7: Additional notation used in the appendices.

Symbol	Meaning
World targets and distillation	
h, h'	Future-frame offset and alternative offset in a horizon sum.
k; FW, FD, CW, CD, M	Prediction-head index; WAN future, DINO future, WAN change, DINO change, and motion flow, respectively.
W, D in Z^W, Z^D	WAN-VAE and DINO feature sources, respectively.
Z_t^W, Z_t^D	Encoded current WAN and DINO features.
$Z_{t,h}^W, Z_{t,h}^D$	Last WAN temporal latent of a clip ending at $t+h$, and DINO features of that future frame.
$Y_{t,h}^k, \bar{Y}_t^k, \widehat{Y}_t^k$	Target at one horizon, horizon-fused target, and current-context prediction for field k.
$m_{t,h}$	Binary indicator of an available within-episode future. Distinct from memory-type index m.
$\xi_h^k, \pi_{t,h}^k$	Learned horizon logit and its normalized, validity-masked mixture weight.
$\pi^{k,0}, \pi^k$	Initial and learned horizon distributions used by the mixture regularizer; distinct from policy π_θ.
P_t, Q_k, h_k	Valid current multimodal prefix, learned queries, and output mapping of prediction head k. The mapping h_k is distinct from horizon h.
$\lambda_2^k, \lambda_{\cos}^k$	Nonnegative coefficients for field-wise mean squared error and cosine distance.
$\mathcal{L}_{\text{bal}}, \lambda_{\text{bal}}$	Horizon-prior KL regularizer and its coefficient.
Memory formation and attention	
$\mathcal{S}_m, U_k$	Set of sources fused into memory m, and representation of source k. Here k ranges over fusion sources, including foresight, not just prediction heads.
W_k, e_k, N_m	Source projection, source identity embedding, and output token count for memory m.
ζ^m, α_k^m	Source-fusion logits and normalized mixture weight of source k.
$\mathcal{C}_t^x, \phi_x$	Token source indexed by $x \in \{P, F, C, M\}$ and its projection into the shared global-summary space.
$\boldsymbol{\zeta}^g, \alpha_x^g$	Trainable global-summary fusion logits and normalized weight of source x.
Reliability, residuals, and state guidance	
$\bar{H}_l, \bar{C}_a^m$	Pooled action features and pooled memory features projected to a compatible action-feature space. Subscript a here means action-aligned, not augmented.
f_m, b_m^r, r_l^m	Learned compatibility mapping, scalar reliability bias, and attention reliability gate for memory m at layer l.
$u_m, u_{l,m}, c_m$	Modality gate logit, layer-specific modality gate logit, and channel-wise gate vector.
$\mathcal{I}_m$	Set of transformer layers receiving memory m.
Δ_l^m	Gated attention residual for memory m at layer l.
$q(s_t), q_1, \ldots, q_4$	Bounded state-conditioned scalars that rescale future, change, motion, and global context, respectively.
p_t	Demonstration-derived transition score used to scale gradients through the state-guidance path.
Robust learning and optimization	
$v, \widetilde{v}$	Clean and augmented predicted action velocities in the consistency loss; shorthand for evaluations of v_θ.
$V, \bar{V}, \widetilde{V}$	Visual feature tensor, pooled clean features, and pooled augmented features in consistency/perturbation equations. Here V is not the number of cameras.
$S_i, \widetilde{S}_i, K_s, i$	Clean slot, augmented slot, number of slots, and slot index.

Continued on next page

Table 7 (continued)

Symbol	Meaning
$\lambda_{\mathrm{con},a}, \lambda_{\mathrm{con},v}$	Action and visual consistency weights.
$\lambda_{\mathrm{con},w}, \lambda_{\mathrm{con},s}$	World-memory and slot consistency weights.
δ_V, G, ϵ_V	Visual perturbation, action-loss gradient with respect to visual features, and relative perturbation magnitude.
j, η_j, λ_j	Optimizer parameter-group index, learning rate, and source-anchor strength.
$\theta_j^{\mathrm{src}}, \widetilde{\theta}_j, \theta_j^+$	Fixed source parameters, parameters after the optimizer update but before anchoring, and parameters after anchoring. Here the tilde does not denote augmentation.

B World-Prior Distillation and Conditioning

B.1 Demonstration-derived supervision

Let $Z_{t,h}^W$ denote the last temporal WAN-VAE latent of a causal clip ending at $t+h$, and $Z_{t,h}^D$ the DINOv3 patch features of that future frame. With corresponding current features Z_t^W, Z_t^D, the five supervision fields are

$$\begin{aligned} Y_{t,h}^{FW} &= \mathrm{Pool}_W(Z_{t,h}^W), & Y_{t,h}^{FD} &= \mathrm{Pool}_D(Z_{t,h}^D), \\ Y_{t,h}^{CW} &= \mathrm{Pool}_2\big(\mathcal{N}(\|Z_{t,h}^W - Z_t^W\|_2)\big), & Y_{t,h}^{CD} &= \mathrm{Pool}_2\big(\mathcal{N}(\|Z_{t,h}^D - Z_t^D\|_2)\big), \\ Y_{t,h}^{M} &= \mathrm{Pool}_2(\mathrm{Flow}(I_t, I_{t+h})). \end{aligned} \tag{7}$$

Here $\mathcal{N}$ normalizes each change map by its spatial maximum. Pool_2 denotes spatial pooling, whereas future-token pooling operates on the feature sequence. Optical flow (Farnebäck, 2003) is normalized by image dimensions. These targets encode observed temporal changes rather than identified causal effects.

To aggregate complementary horizons, we learn a masked mixture for each field:

$$\pi_{t,h}^k = \frac{m_{t,h} \exp \xi_h^k}{\sum_{h'} m_{t,h'} \exp \xi_{h'}^k}, \qquad \bar{Y}_t^k = \sum_h \pi_{t,h}^k (Y_{t,h}^k). \tag{8}$$

The validity mask m excludes futures outside the episode; samples without a valid future receive no distillation loss. The teacher encoders are frozen, while mixture weights are learned. With $d_{\cos}(x,y) = 1 - \cos(\mathrm{vec}\, x, \mathrm{vec}\, y)$, the objective is

$$\mathcal{L}_W = \sum_k \Big[\lambda_2^k \,\mathrm{MSE}(\widehat{Y}_t^k, \bar{Y}_t^k) + \lambda_{\cos}^k d_{\cos}(\widehat{Y}_t^k, \bar{Y}_t^k)\Big] + \lambda_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}}, \tag{9}$$

where $\mathcal{L}_{\mathrm{bal}}$ averages $D_{\mathrm{KL}}(\pi^{k,0} \| \pi^k)$ over fields and encourages retention of the initial horizon prior.

B.2 Prediction and memory formation

For a valid current-context prefix P_t, excluding padding and teacher-forced actions, each predictor uses learned queries Q_k:

$$\widehat{Y}_t^k = h_k(\mathrm{CrossAttn}(Q_k, (P_t), (P_t))), \quad k \in \{FW, FD, CW, CD, M\}. \tag{10}$$

The operator h_k contains residual processing and a field-specific projection; $\,$ denotes stop-gradient. Projected source representations U_k are normalized, pooled, and fused into memories:

$$C^m = \sum_{k \in \mathcal{S}_m} \alpha_k^m [(\mathrm{Pool}_{N_m}(U_k W_k)) + e_k], \qquad \alpha^m = \mathrm{softmax}(\zeta^m), \quad m \in \{F, C\}. \tag{11}$$

Future sources comprise predicted DINO features and the backbone's trainable foresight representation; change sources comprise WAN and DINO change predictions. Motion uses a projection of the predicted flow. Foresight tokens remain active without video-denoising supervision.

To provide global FiLM conditioning, we pool and project the current prefix and the three world memories, then combine them with learned source weights:

$$g_t = \sum_{x\in\{P,F,C,M\}} \alpha_x^g \phi_x(\mathrm{Pool}(\mathcal{C}_t^x)), \qquad \boldsymbol{\alpha}^g = \mathrm{softmax}(\boldsymbol{\zeta}^g), \quad \mathcal{C}_t^P = P_t, \quad \mathcal{C}_t^m = C^m. \tag{12}$$

Here x indexes the current prefix (P) or a future, change, or motion source. $\mathcal{C}_t^x$ denotes the corresponding token sequence, ϕ_x is a source-specific projection into a shared feature space, and α_x^g is the normalized weight obtained from trainable logits $\boldsymbol{\zeta}^g$. Consequently, g_t is a global summary computed from current-context predictions rather than an additional supervision target or cached future input.

B.3 Selective world-to-action conditioning

The action transformer allocates overlapping layer intervals to future, change, and motion guidance, progressing from outcome information to local refinement. Within a layer, active memories update the action stream sequentially. Pooled action and projected memory features define a compatibility score:

$$r_l^m = \sigma\big(f_m[\bar{H}_l, \bar{C}_a^m, \bar{C}_a^m - \bar{H}_l, \bar{H}_l \odot \bar{C}_a^m] + b_m^r\big). \tag{13}$$

With channel gate c_m, modality logit u_m, and layer logit $u_{l,m}$, the attention contribution is

$$\Delta_l^m = \mathbf{1}_{\{l\in\mathcal{I}_m\}}\, \eta\sigma(u_m)\sigma(u_{l,m}) r_l^m (c_m \odot R_l^m). \tag{14}$$

Reliability is learned compatibility, not calibrated uncertainty. Change biases use predicted WAN change scores, while motion biases use displacement magnitudes. Applying Eq. equation 3 to these gated contributions gives Δ_l^{attn}. FiLM has a separate reliability gate computed from action and modulation features. Its affine transformation acts on the original layer input; the adapter processes $\Delta_l = \Delta_l^{\mathrm{attn}} + \Delta_l^{\mathrm{film}}$. Residual output projections are zero-initialized.

Current evidence and state guidance Language-conditioned visual slots complement predicted dynamics. Each slot contains a learned address and visual content retrieved with camera-aware queries. Actions query slot addresses as keys and full slots as values. In the LIBERO setting, slots and world targets use the main camera, while the VLM retains both main and wrist observations. A trainable visual adapter supports adaptation of the frozen visual pathway.

Normalized robot state modulates guidance through bounded scalars:

$$q(s_t) = 0.1\tanh(\mathrm{MLP}((s_t))) \in \mathbb{R}^4. \tag{15}$$

These rescale (C^F, C^C, C^M, g_t) by $(1+q_1, \ldots, 1+q_4)$. To limit reliance on proprioceptive shortcuts (Zhao et al., 2025; Lu et al., 2026), a demonstration-derived transition score p_t scales gradients through this guidance path by $1 - 0.5p_t$, without changing its forward computation or requiring transition labels at inference.

C Robust Learning and Experimental Protocol

C.1 Paired consistency and regularization

Clean and augmented observations share action labels, flow noise, and interpolation time. State perturbations are aligned across numeric and tokenized inputs, preserving gripper information. Task-preserving language rewrites and photometric perturbations alter observations without relocating objects or changing action labels. The consistency objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{con}} = {} & \lambda_{\mathrm{con},a}\,\mathrm{SmoothL1}(\widetilde{v}, (v)) + \lambda_{\mathrm{con},v} d_{\cos}(\widetilde{\bar{V}}, (\bar{V})) \\ & + \frac{\lambda_{\mathrm{con},w}}{3}\sum_m d_{\cos}(\mathrm{Pool}(\widetilde{C}^m), (\mathrm{Pool}(C^m))) + \frac{\lambda_{\mathrm{con},s}}{K_s}\sum_i d_{\cos}(\widetilde{S}_i, (S_i)).\end{aligned} \tag{16}$$

Here tildes denote augmented features, $\bar{V}$ denotes pooled visual features, and S_i denotes a slot. Clean targets are detached for consistency only; clean action and world supervision remain differentiable. To regularize feature sensitivity, we additionally perturb augmented visual features along a detached action-loss gradient direction:

$$\delta_V = \epsilon_V \|V\|_2 (G/\max(\|G\|_2, \varepsilon)), \qquad G = \nabla_V \mathcal{L}_{\mathrm{act}}. \tag{17}$$

Norms are computed per example. Demonstration-derived start, state-rarity, and motion-transition scores mildly reweight supervision toward underrepresented or changing states. Such reweighting preserves demonstrated action labels rather than synthesizing recovery trajectories.

To retain pretrained behavior during adaptation, we use a proximal parameter anchor (Xuhong et al., 2018). Following an optimizer update to $\widetilde{\theta}_j$, parameter group j is updated as

$$\theta_j^+ = \frac{\widetilde{\theta}_j + \eta_j \lambda_j \theta_j^{\mathrm{src}}}{1 + \eta_j \lambda_j}, \tag{18}$$

where θ^{src} is the fixed initialization and η_j is the learning rate. This regularizes parameter drift without an additional teacher-policy forward pass.

C.2 Training Configuration

World targets use horizons of 16, 24, and 30 frames; these horizons describe scene evolution beyond the execution chunk. The predictor uses 16 queries per field, and world attention operates at width 256. In the 24-layer action transformer, zero-indexed future/change/motion intervals are $[0, 16)$, $[8, 24)$, and $[16, 24)$.

D Simulation Benchmarks and Real-World Setup

D.1 Simulation Benchmarks

We evaluate BRIDGE-WA on LIBERO-Plus, our LIBERO-Dynamic, RoboTwin 2.0, and VLABench, as shown in Fig. 3(c). We train exclusively on the official LIBERO demonstrations (Liu et al., 2023), using language instructions, third-person and wrist-camera RGB observations, and robot proprioception. Each action is represented as $\mathbf{a}_t = [\Delta\mathbf{p}_t, \Delta\mathbf{r}_t, g_t] \in \mathbb{R}^7$, where $\Delta\mathbf{p}_t, \Delta\mathbf{r}_t \in \mathbb{R}^3$ are end-effector translation and rotation commands and g_t is the gripper command. The policy predicts an $T_a = 8$ action chunk. **LIBERO-Plus** (Fei et al., 2025) evaluates the same LIBERO-only trained policy on seven controlled shifts: camera viewpoint, robot initial state, language instruction, lighting, background, visual noise, and scene layout. We report success rates for each LIBERO-Plus perturbation. Unlike the static shifts in LIBERO-Plus, **LIBERO-Dynamic** evaluates closed-loop robustness to time-varying camera viewpoints. It extends the 40 canonical LIBERO tasks with five predefined camera-motion profiles: *smooth panning*, *orbital motion*, *continuous zoom*, *abrupt viewpoint jumps*, and *camera shake*. Only the third-person camera moves; the wrist camera and success criteria remain unchanged. We evaluate 50 initial states per task over five balanced difficulty levels, yielding 10,000 episodes with policy-independent camera viewpoint trajectories. No LIBERO-Dynamic data are used for training, and success rates are reported by camera-motion profile. **RoboTwin 2.0** (Chen et al., 2025) evaluates language-conditioned bimanual manipulation under *Easy* and *Hard* settings. Easy contains clean scenes, while Hard randomizes distractor objects, background textures, lighting, tabletop height, and language instructions. We train on the official 50 tasks using 50 clean and 500 randomized trajectories per task. Inputs comprise the instruction, head-camera and bilateral wrist-camera RGB observations, and proprioception. Actions are $\mathbf{a}_t = [\mathbf{q}_t^L, g_t^L, \mathbf{q}_t^R, g_t^R] \in \mathbb{R}^{14}$, where $\mathbf{q}_t^L, \mathbf{q}_t^R \in \mathbb{R}^6$ are left- and right-arm joint commands and g_t^L, g_t^R are gripper commands. The action-chunk horizon is $T_a = 50$. We evaluate 100 episodes per task with unseen instruction variants and report mean success rates separately for Easy and Hard. **VLABench** (Zhang et al.) evaluates single-arm manipulation under in-distribution and systematic generalization settings. We train on its official primitive-task dataset containing 10 tasks and 500 demonstrations per task. Inputs include the instruction, front-, secondary-, and wrist-camera RGB observations, and proprioception. Each action is $\mathbf{a}_t = [\mathbf{p}_t, \mathbf{r}_t^{6\mathrm{D}}, g_t] \in \mathbb{R}^{10}$, where $\mathbf{p}_t \in \mathbb{R}^3$ is the end-effector position and $\mathbf{r}_t^{6\mathrm{D}} \in \mathbb{R}^6$ is its continuous rotation representation; the action-chunk horizon is $T_a = 30$. We evaluate policy on five tracks: *In-Distribution*, *Cross-Category*, *Common Sense*, *Semantic Instruction*, and *Unseen Texture*. The latter four test zero-shot transfer to unseen categories, implicit or knowledge-dependent instructions, and appearance shifts. We report success rate (SR), intention score (IS), and progress score (PS), averaged across tracks.

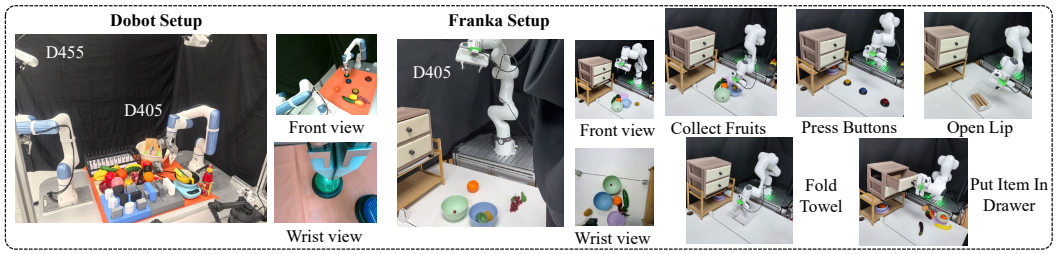


Figure 6: Dobot and Franka real-world setup.

D.2 Real-Robot Experimental Setup, Task Definitions, and Evaluation Protocol

Overview We evaluate the real-robot task execution capability and OOD generalization capability of BRIDGE-WA on two single-arm platforms: a Dobot Nova2 robot and a Franka FR3 robot. Both platforms use LeRobot-style data with synchronized third-person RGB, wrist RGB, Cartesian robot state, Cartesian end-effector action, timestamps, and task instructions. The action vector is seven-dimensional, $\mathbf{a}_t = [x_t, y_t, z_t, \mathrm{roll}_t, \mathrm{pitch}_t, \mathrm{yaw}_t, g_t] \in \mathbb{R}^7$, where $g_t \in [0, 1]$ denotes the gripper command. Demonstrations are collected at 30 Hz. The Dobot setup uses an Intel RealSense D455 as the third-person camera and an Intel RealSense D405 as the wrist camera, with 480×480 videos for both views. The Franka setup uses an ORBBEC Femto Bolt as the third-person camera at 720×720 resolution and an Intel RealSense D405 as the wrist camera at 480×480 resolution. At training time, images are resized to the policy resolution and the same action-chunk interface is used across embodiments. The robot is reset to a fixed home pose before each rollout. Object poses are randomly sampled within task-specific regions, while preserving reachability and avoiding initial collisions. The dataset contains 50 demonstrations per task.

D.2.1 Real-Robot Task Definitions

Pick Grape On the table, there is a bunch of grapes, other fruits, and a plate. The robot needs to pick up and place the grapes. Language instruction: Put the grapes into the [COLOR] plate.

Collect Fruits There are various fruits on the table. The robot needs to collect the specified fruits into the specific target. Language instruction: Put a [FRUIT], a [FRUIT], and an [FRUIT] into the [COLOR] [TARGET].

Push Buttons/Press Buttons There are multiple buttons of different colors on the table. The robot needs to press the buttons in a specific sequence. Language instruction: Press the [COLOR] button, then the [COLOR] button, followed by the [COLOR] button, and finally the [COLOR] button.

Stack Bowls There are a cup, some bowls and a plate on the table. The robot needs to stack a cup, the bowls and plate in a specific order. Language instruction: Stack a [COLOR] cup,a [COLOR] bowl and a [COLOR] bowl on a [COLOR] plate.

Put Item In Drawer There is a cabinet on the table. The robot needs to open the specified drawer, place the target object inside, and then close the drawer. Language instruction: Open the [Top/Middle/Bottom] drawer and put the [OBJECT] into it.

Fold Towel There is a towel on the table. The robot needs to pick up one corner of the towel and then fold it. Language instruction: Fold the towel.

Open Lip There is a box on the table. The robot needs to pick up the lid and open it. Language instruction: Open the plastic lid of the biscuit case.

Hard-track perturbations For Hard-track evaluation, we introduce controlled visual and scene perturbations while keeping the task goal and success criterion unchanged. In the distractor setting, additional irrelevant objects are placed on the tabletop or in the surrounding scene, such as extra fruit items, plush bears, and other non-target objects. In the tablecloth setting, we change the tablecloth color and texture to alter the workspace appearance. In the lighting setting, we turn on dynamic colored lights to create non-stationary illumination changes and color shifts. These perturbations

are designed to test whether the policy can focus on the task-relevant object, contact region, and intended scene change rather than relying on background appearance.

D.2.2 Real-Robot Evaluation Protocol

Evaluation contains one Easy track and three Hard tracks. Easy uses 10 trials per task with in-distribution object poses and lighting. Each Hard track uses 5 trials per task and isolates one visual shift: distractor objects, lighting generalization, or tablecloth generalization. All methods are evaluated from matched initial states with the same language instruction sequence and the same manual reset procedure. On the Dobot robot, for binary tasks, a trial receives 1 for success and 0 otherwise. `PushButtons` receives 0, 0.5, or 1 according to the number of correctly pressed targets. `CollectFruits` and `StackBowls` receive 0, 0.3, 0.6, or 1 according to task-specific completion stages. The Hard score is the sum over the three hard tracks, normalized as an average success rate. On the Franka robot, during evaluation, each trial is scored by stage, with a value between 0 and 1 assigned according to the successfully completed stages.

E Experimental Details

E.1 Implementation Details

We train our models using 8 NVIDIA A100 80GB GPUs and 4 NVIDIA RTX PRO 6000 GPUs. We instantiate the VLM backbones with Florence-2 (Xiao et al., 2024), PaliGemma (Beyer et al., 2024), and Qwen3.5 (Team, 2026), following X-VLA (Zheng et al., 2026), $\pi_{0.5}$ (Black et al., 2025), and InternVLA-A1.5 (Ma et al., 2026a), respectively. The corresponding VLM sizes are 0.9B, 3B, and 2B parameters. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and weight decay 0.01. We clip the gradient norm to 1.0 and train with BF16 mixed precision.

E.2 Baselines

E.2.1 Vision-Language-Action Models

X-VLA (Zheng et al., 2026) is a scalable flow-matching VLA built from standard Transformer encoders. It assigns each data source an embodiment-specific learned soft prompt, allowing a shared policy to exploit heterogeneous cross-robot data with only a small number of added parameters.

$\pi_{0.5}$ (Black et al., 2025) extends π_0 toward open-world manipulation through heterogeneous co-training. It combines multi-robot actions with web data, object detections, high-level semantic prediction, and hybrid multimodal examples to improve long-horizon generalization in unseen environments.

LingBot-VLA (Wu et al., 2026) is a dual-arm foundation policy trained on roughly 20K hours of real-world data from nine robot configurations. Its emphasis is practical cross-platform generalization, efficient post-training, and a high-throughput open training and deployment stack.

InternVLA-A1.5 (Ma et al., 2026a) preserves VLM semantics through continued VQA and subtask-prediction training while attaching a lightweight expert for continuous actions. Learnable foresight tokens distill task-relevant future information from a frozen video generator, whose prediction branch is removed at deployment.

E.2.2 World Action Models

Fast-WAM (Yuan et al., 2026b) retains video-prediction co-training but removes explicit future imagination at test time. It isolates the representational benefit of world-model supervision from the cost of iterative video generation, enabling real-time action prediction without discarding future-aware training.

VLA-JEPA (Sun et al., 2026) pretrains an action policy through leakage-free latent future prediction: the student observes only the current frame, while a target encoder converts future frames into supervision. Predicting action-relevant state transitions in latent space reduces sensitivity to camera motion, background changes, and other pixel-level nuisance variation.

Cosmos Policy (Kim et al., 2026) adapts a pretrained video diffusion model to control in a single post-training stage without adding a separate action architecture. Actions, future observations, and values are represented as latent frames, enabling the same model to generate trajectories and perform value-guided test-time planning.

Motus (Bi et al., 2026) unifies understanding, video generation, and action experts in a Mixture-of-Transformers architecture with a shared diffusion scheduler. Optical flow supplies embodiment-

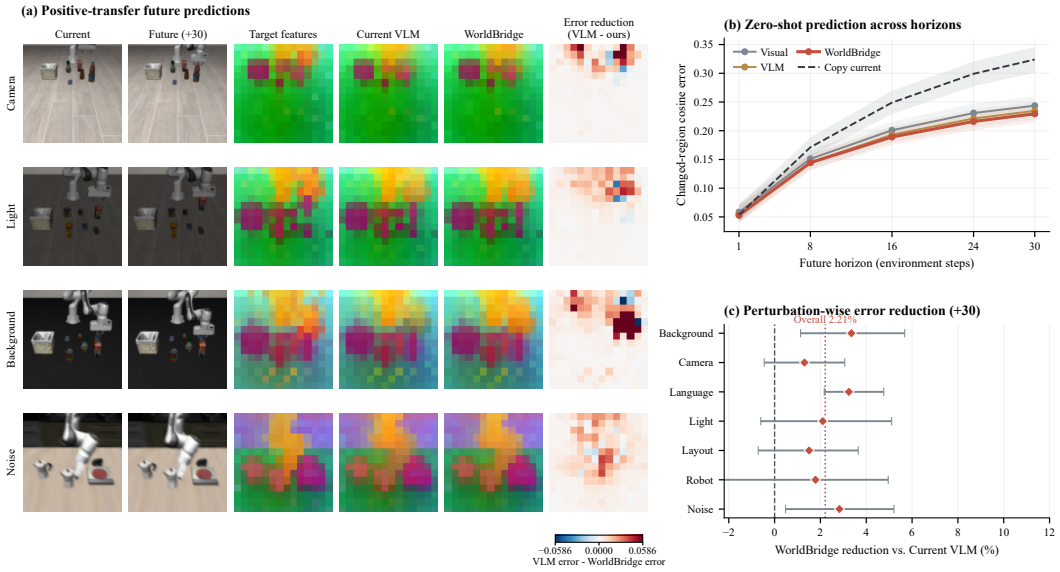


Figure 7: **Future-feature prediction under LIBERO-Plus perturbations.** (a) Positive-transfer examples for Camera, Light, Background, and Sensor Noise. Red regions indicate lower future-feature prediction error for WorldBridge than for Current VLM. (b) Changed-region cosine error across future horizons. (c) WorldBridge error reduction relative to Current VLM at horizon $+30$, with task/episode-clustered 95% confidence intervals.

agnostic latent-action supervision, allowing the model to switch among world modeling, inverse dynamics, VLA control, and joint video–action prediction.

F Additional Experimental Results and Analysis

F.1 Detailed Real-Robot Results

For the Franka robot, we evaluate the Easy track. The per-task success rates for BRIDGE-WA, $\pi_{0.5}$ (Black et al., 2025), and VLA-JEPA (Sun et al., 2026) are reported in Tab. 8. Tab.9 details the success rate of different models for each task and each track on the Dobot robot.

Table 8: Per-task success rates (%) on the Franka real-robot evaluation.

Method	Collect Fruits	Press Buttons	Open Lip	Fold Towel	Put Item In Drawer	Average
$\pi_{0.5}$ (Black et al., 2025)	50.0	80.0	50.0	30.0	25.0	47.0
VLA-JEPA (Sun et al., 2026)	**75.0**	90.0	30.0	25.0	25.0	49.0
BRIDGE-WA (Ours)	**75.0**	**100.0**	**60.0**	**80.0**	**50.0**	**73.0**

F.2 Ablation Study

Future-Prediction Transfer Analysis We investigate whether WorldBridge provides predictive information beyond the current VLM representation. As illustrated in Fig. 7(a), WorldBridge more closely recovers the target future features, with improvements concentrated around regions undergoing future robot or object changes. These examples provide a spatial interpretation of the learned future priors, while the following quantitative results use the complete evaluation cohort. Equal-capacity diagnostic heads are trained only on original LIBERO and evaluated without adaptation on LIBERO-Plus. Figure 7(b) shows that WorldBridge maintains lower changed-region prediction error than Current VLM across future horizons. At horizon $+30$, the error decreases from 0.2347 to 0.2295, corresponding to a relative reduction of 2.21%. Figure 7(c) further reports positive point estimates across all seven perturbation families. The confidence intervals are entirely positive for Background, Language, and Sensor Noise. Overall, these results indicate that WorldBridge contributes transferable future-state information beyond the current VLM context.

Table 9: Dobot real-robot evaluation results (%). Easy uses 10 trials per task; Hard contains three tracks with 5 trials each: distractors, lighting generalization, and tablecloth generalization. Each PushButtons test is scored as 0, 0.5, or 1; each CollectFruits and StackBowls test is scored as 0, 0.3, 0.6, or 1; all other tasks are scored as 0 or 1. Hard aggregates the results over the three hard tracks.

Method	PickGrape					StackBowls					PushButtons				
	Easy	Dist.	Light	T-Cloth	Hard	Easy	Dist.	Light	T-Cloth	Hard	Easy	Dist.	Light	T-Cloth	Hard
OpenVLA (Kim et al., 2024)	30.0	20.0	40.0	20.0	26.7	12.0	18.0	6.0	12.0	12.0	80.0	70.0	40.0	**80.0**	63.3
X-VLA (Zheng et al., 2026)	80.0	60.0	80.0	60.0	66.7	34.0	24.0	46.0	32.0	34.0	80.0	70.0	40.0	**80.0**	63.3
$\pi_{0.5}$ (Black et al., 2025)	80.0	60.0	80.0	**100.0**	80.0	**40.0**	**30.0**	**58.0**	**42.0**	**43.3**	**90.0**	80.0	70.0	60.0	70.0
BRIDGE-WA (Ours)	**90.0**	**80.0**	**100.0**	**100.0**	**93.3**	30.0	18.0	40.0	**42.0**	33.3	**90.0**	**90.0**	**90.0**	70.0	**83.3**

Method	CollectFruits					PutItemInDrawer					Avg. SR				
	Easy	Dist.	Light	T-Cloth	Hard	Easy	Dist.	Light	T-Cloth	Hard	Easy	Dist.	Light	T-Cloth	Hard
OpenVLA (Kim et al., 2024)	20.0	12.0	18.0	12.0	14.0	10.0	20.0	20.0	20.0	20.0	30.4	28.0	24.8	28.8	27.2
X-VLA (Zheng et al., 2026)	84.0	72.0	**100.0**	66.0	79.3	**70.0**	**40.0**	**60.0**	40.0	**46.7**	69.6	53.2	65.2	55.6	58.0
$\pi_{0.5}$ (Black et al., 2025)	80.0	66.0	80.0	66.0	70.7	60.0	**40.0**	**60.0**	40.0	**46.7**	70.0	55.2	69.6	61.6	62.1
BRIDGE-WA (Ours)	**88.0**	**86.0**	**100.0**	**80.0**	**88.7**	**70.0**	**40.0**	40.0	**60.0**	**46.7**	**73.6**	**62.8**	**74.0**	**70.4**	**69.1**

G LIBERO-DYNAMIC BENCHMARK

G.1 MOTIVATION AND BENCHMARK DEFINITION

Existing LIBERO evaluations mainly alter the initial environment state or apply a fixed visual perturbation throughout an episode. Such protocols measure robustness to static distribution shifts, but do not fully evaluate whether a policy can maintain temporally consistent perception and control when the observation viewpoint changes during task execution.

We introduce **LIBERO-Dynamic**, as shown in the Fig8, a zero-shot evaluation benchmark for temporally varying camera perturbations. Unlike static camera benchmarks, LIBERO-Dynamic moves the external agent-view camera as the robot executes the task. The manipulation task, object dynamics, robot state, language instruction, and success condition remain unchanged. Therefore, performance degradation can be attributed primarily to the policy's robustness to dynamic viewpoint changes.

Let $\mathcal{T}$ denote a LIBERO task, s_i its i-th initial environment state, and $H_{\mathcal{T}}$ its maximum episode horizon. For camera variant v, LIBERO-Dynamic assigns a deterministic camera trajectory

$$\Gamma_{\mathcal{T},i,v} = \{\mathbf{C}_{\mathcal{T},i,v}(t)\}_{t=0}^{H_{\mathcal{T}}-1}, \tag{19}$$

where $\mathbf{C}(t) = (\mathbf{p}(t), \mathbf{R}(t))$ contains the camera position and orientation at simulation step t. We parameterize time using

$$u_t = \frac{t}{H_{\mathcal{T}} - 1}, \tag{20}$$

so that the camera motion spans the complete task horizon and remains comparable across tasks with different maximum lengths.

Each tuple $(\mathcal{T}, s_i, v)$ maps to one unique but deterministic trajectory. The mapping is fixed in a released evaluation manifest rather than sampled online. Consequently, different models observe exactly the same initial state and camera motion, enabling paired and reproducible comparison.

Difference from static camera evaluation. LIBERO-Dynamic evaluates three capabilities that are weakly measured by a fixed viewpoint: temporal visual consistency, recovery from changing spatial projections, and action prediction under continuously changing camera–workspace geometry. It is an evaluation-only benchmark: models are trained on the standard LIBERO data and evaluated on LIBERO-Dynamic without dynamic-camera training data or benchmark-specific fine-tuning.

G.2 DYNAMIC CAMERA VARIANTS

LIBERO-Dynamic contains five complementary camera-motion variants. They cover continuous, discontinuous, low-frequency, and high-frequency viewpoint changes.

Table 10: Dynamic camera variants in LIBERO-Dynamic.

Variant	Description
Smooth Pan	Applies smooth low-frequency changes in azimuth and elevation. A subset of trajectories follows an arc around the workspace while continuously looking toward its center.
Orbit	Moves the camera along a continuous orbit around the workspace center. The look-at direction is updated at every simulation step to preserve the manipulation region in view.
Zoom	Changes the camera–workspace distance along the viewing direction, producing temporally varying object scale and perspective while retaining the task-relevant workspace.
View Jump	Switches between several separated, valid viewpoints during task execution. Some jumps are sampled from center-looking orbital poses. This variant evaluates recovery from abrupt changes in visual correspondence.
Camera Shake	Adds band-limited translational and rotational perturbations at a higher temporal frequency, simulating unstable camera mounting or external vibration.

A camera trajectory is accepted only when it satisfies workspace-visibility constraints. In particular, the camera height, viewing direction, distance, and angular range are bounded, and center-looking trajectories keep the workspace center inside a valid image region. Invalid poses that point toward the floor, ceiling, or outside the workspace are rejected or projected back into the valid camera region. These constraints prevent degenerate samples in which the task becomes unobservable rather than perceptually challenging.

G.3 Difficulty Levels

Each variant is divided into five ordered difficulty levels. Difficulty is increased by jointly scaling the translational displacement, angular excursion, motion speed, active duration, jump distance or frequency, and shake magnitude, depending on the variant. The levels are defined such that

$$A_1 < A_2 < A_3 < A_4 < A_5, \tag{21}$$

where A_d denotes the effective perturbation magnitude at difficulty level d. Exact trajectory parameters are stored in the frozen benchmark manifest.

G.4 Evaluation Protocol

We evaluate all 40 tasks from four standard LIBERO suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-10. Each suite contains ten tasks. Every task is evaluated from 50 initial states under each of the five camera variants. Thus, one variant contains 2000 rollouts, and the complete benchmark contains 10000 rollouts. We report the success rate for each camera variant and the overall average.

H Additional Visualizations

We provide additional qualitative visualizations to contextualize the benchmark tasks, the learned world-prior representations, and the real-robot execution behavior. The benchmark-scene overview in Fig. 9 visualizes the ten VLABench tasks and the 15 RoboTwin 2.0 tasks. The world-prior visualization in Fig. 10 illustrates how BRIDGE-WA decomposes predicted scene evolution into future, change, and motion priors on simulation benchmarks. Finally, the Dobot rollout sequence in Fig. 11 presents representative executions for the five real-robot tasks, showing the temporal progression from initial observation to task completion.

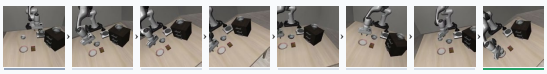
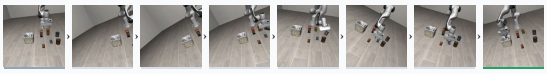
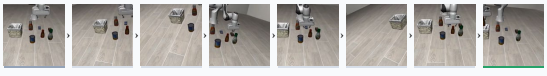
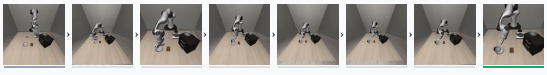
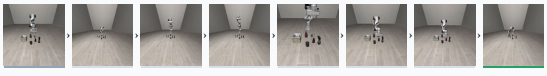
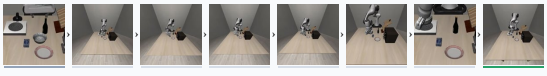
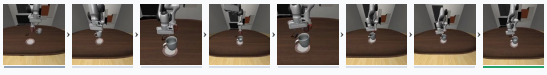
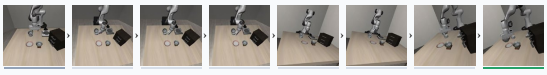
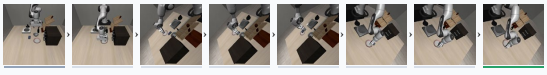
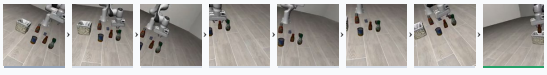
Camera	Suite	Task	Rollout [t_0 → Perturbation Response → Completion]
Smooth Pan	Spatial	black bowl: stove → plate	
	Object	cream cheese → basket	
	Goal	push plate → stove front	
	Long	mug → microwave; close	
Orbit	Spatial	black bowl: top drawer → plate	
	Object	BBQ sauce → basket	
	Goal	push plate → stove front	
	Long	bowl → bottom drawer; close	
Zoom	Spatial	black bowl: stove → plate	
	Object	BBQ sauce → basket	
	Goal	bowl → stove	
	Long	mug → plate; pudding → right	
View Jump	Spatial	black bowl: ramekin → plate	
	Object	cream cheese → basket	
	Goal	push plate → stove front	
	Long	turn on stove; moka pot → stove	
Camera Shake	Spatial	black bowl: between objects → plate	
	Object	BBQ sauce → basket	
	Goal	bowl → cabinet top	
	Long	soup + sauce → basket	

Figure 8: **Qualitative examples from LIBERO-Dynamic.** Representative rollout sequences under five dynamic camera variants: Smooth Pan, Orbit, Zoom, View Jump, and Camera Shake. For each variant, we show tasks from the Spatial, Object, Goal, and Long suites. Frames are ordered from left to right, beginning with the initial observation at t_0, followed by the time-varying camera perturbation and the policy response, and ending with task completion.

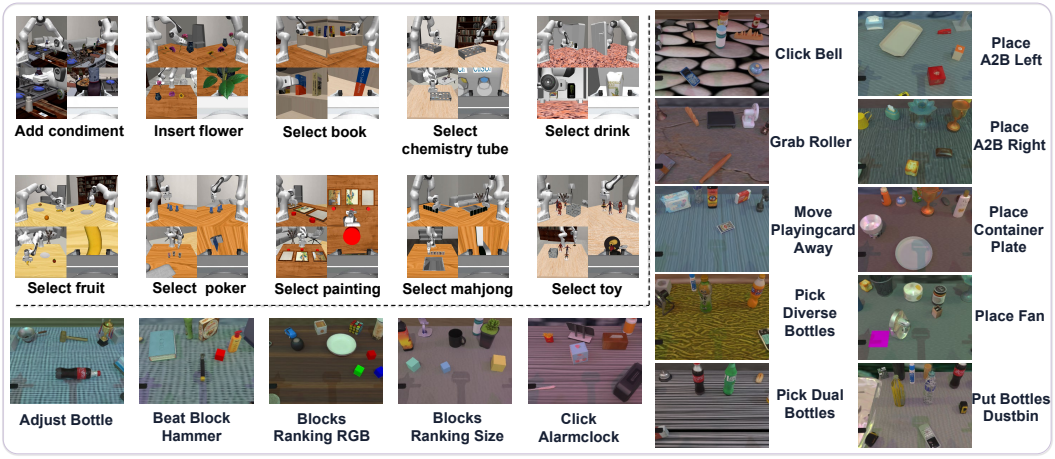


Figure 9: Task and scene visualizations for the simulation benchmarks on VLABench and RoboTwin 2.0.

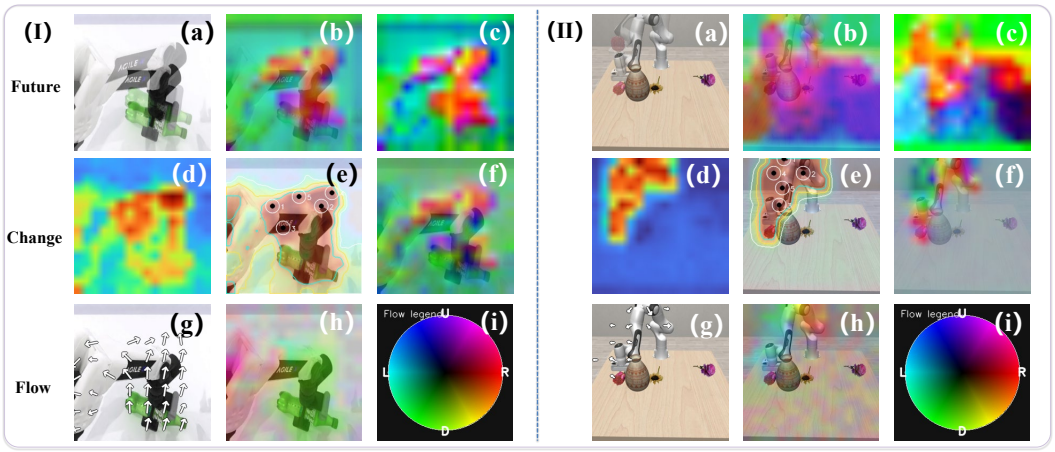


Figure 10: Qualitative world-prior visualizations on RoboTwin 2.0 (I) and VLABench (II). We show the factorized predictive signals used by BRIDGE-WA: future representations summarize the intended outcome, change maps identify action-relevant regions expected to change, and motion-flow maps indicate the local direction of scene evolution. (a) shows the future RGB target overlaid on the current observation. (b) overlays the predicted future latent representation on the scene, while (c) visualizes the future latent map. (d) presents the change map, and (e) overlays the enhanced change response on the observation; the marked points indicate high-confidence regions where action-relevant state changes are expected. (f) highlights feature-level differences between the current and future states. (g) visualizes the dense flow field on the scene, and (h) illustrates the temporal motion direction implied by the flow. (i) gives the color-wheel legend, where hue denotes motion direction.

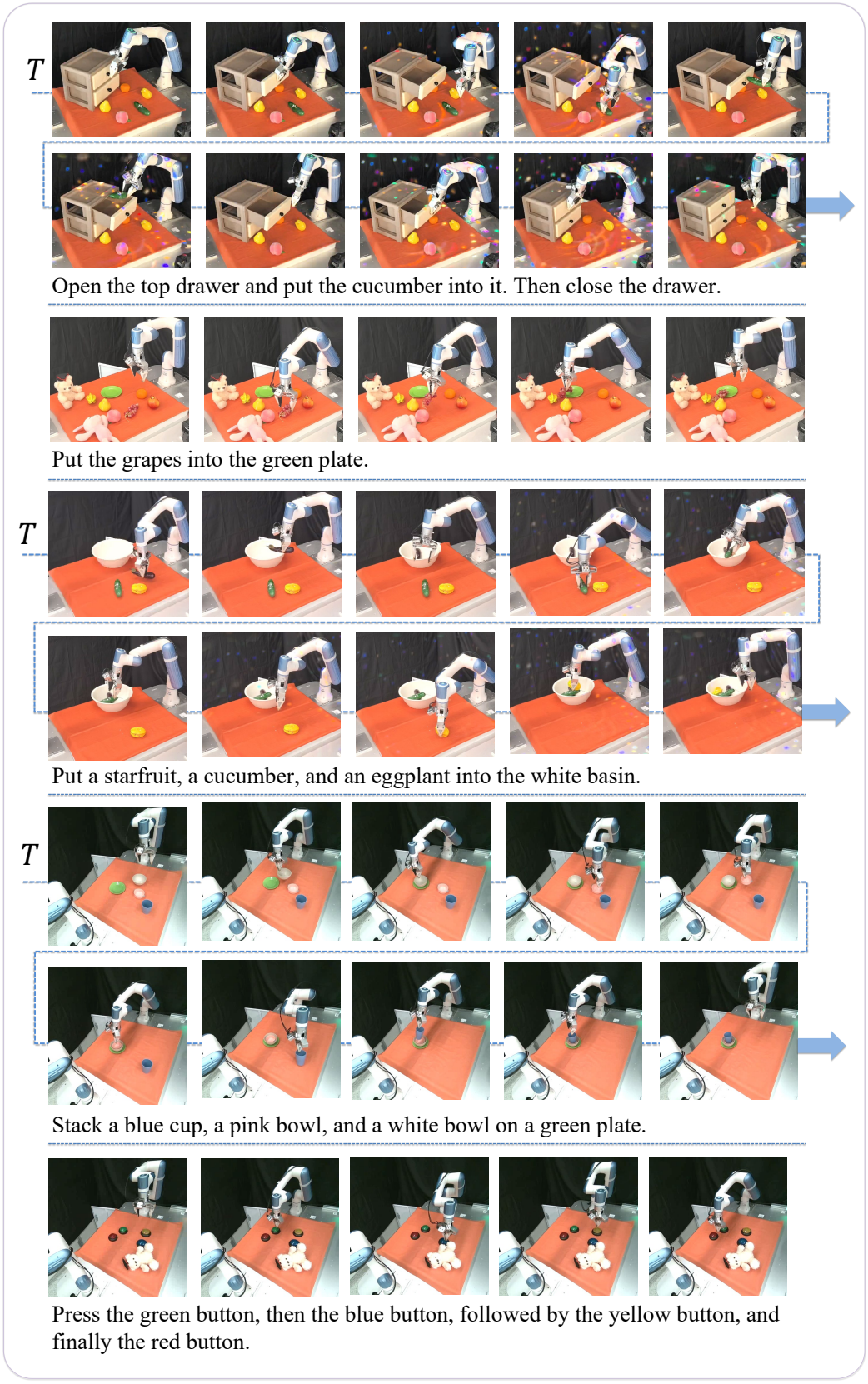


Figure 11: Dobot real-robot execution sequences. The figure shows temporal rollouts for the five real-robot tasks, covering object selection, placement, pushing, stacking, and drawer manipulation.