# BRIDGEWAM: DOES SHARED ATTENTION ALIGN PHYSICAL WORLD REPRESENTATIONS WITH ACTIONS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

World Action Models (WAMs) commonly couple video and action experts through dense, layer-wise Mixture-of-Transformers (MoT) attention, exposing action prediction to irrelevant visual features while enforcing matched expert depths. This coupling hinders selective visual-feature use and introduces redundant computation. We introduce Bridge-of-Experts (BoE), instantiated as BRIDGEWAM, which aggregates visual, language, and proprioceptive priors through 32 Latent Bridge Queries (LBQs) into a shared, limited latent space. Joint video–action supervision lets action gradients guide the Video Expert toward representations better suited for action prediction, while the Action Expert learns to selectively retrieve task-relevant information through cross-attention. Decoupling expert depths and reusing LBQs during action denoising yield a $5.06$B model with a 30-layer Video Expert and a two-layer Action Expert, achieving $98.25\%$ success on LIBERO and $72.24\%$ pooled success on LIBERO-Plus, demonstrating effective task execution and strong generalization with an ultralight action head. Ablations support joint representation adaptation and unified LBQ conditioning. Relative to Fast-WAM, BRIDGEWAM reduces single-sample inference FLOPs by $28.72\%$ in simulation profiling and real-robot latency to $77.1$ ms ($2.47\times$ speedup). Qualitative experimental results further indicate preferential attention to target objects and task destinations across perturbations. Our project page is available at `https://bridgewam.github.io/Bridge-Wam-show/`.

## 1 INTRODUCTION

World Action Models (WAMs) aim to predict robot actions from visual observations, language instructions, and proprioceptive state while retaining the broad visual priors of video-generation models. Useful WAMs must therefore simultaneously predict precise action chunks, remain robust when scenes and objects change, and run with low enough latency for closed-loop control. Recent WAMs have improved these properties by coupling a Video Expert and an Action Expert with Mixture-of-Transformers (MoT) attention Ye et al. (2026); Bi et al. (2025); Li et al. (2026); Yuan et al. (2026). However, information transfer between the two modalities remains largely unchanged.

WAM designs progressively simplify inference from joint video–action generation to lighter visual conditioning. Joint models process current and future video, action, and state tokens together Ye et al. (2026). IDM-style models use asymmetric attention: the video branch operates independently while the action branch reads the visual sequence, but future-video generation remains necessary Li et al. (2026); Zhang et al. (2026). Fast-WAM then removes future-video generation at inference to reduce latency, retaining MoT's direct video K/V interface to action attention Yuan et al. (2026).

However, these layer-wise MoT designs retain central limitations. First, each Video Expert layer directly exposes its dense visual features to the corresponding action layer by concatenating video and action keys/values, placing task-relevant information and redundant details in competition for attention. When the Action Expert cannot adequately distinguish them, redundant features may receive attention comparable to useful features, limiting effective visual-feature utilization and constraining action accuracy and generalization. Second, one-to-one layer coupling forces both experts to share the same depth, introducing redundant parameters and inference overhead. Excessive action-side



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.

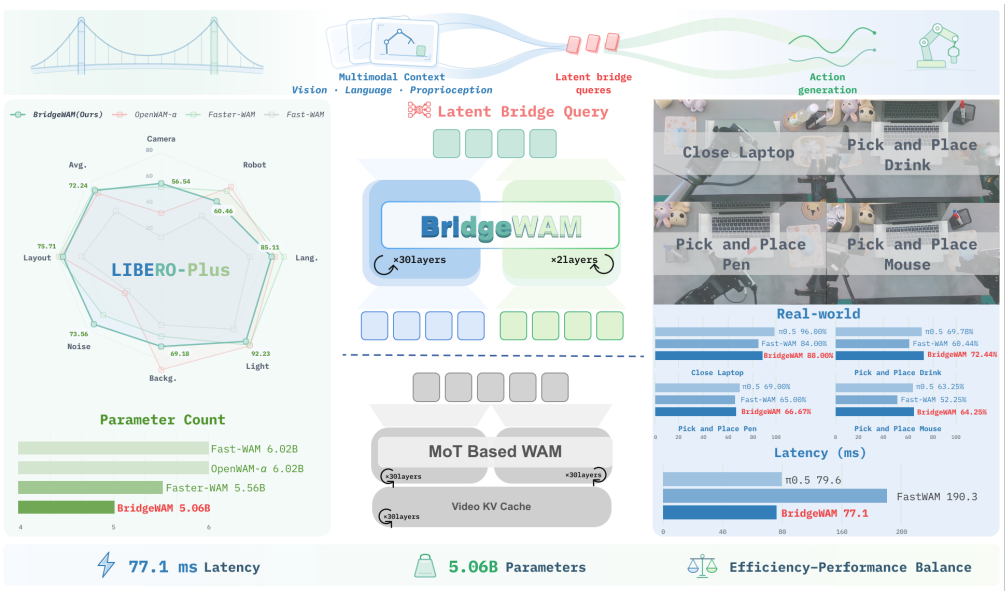


Figure 1: **Overview of BRIDGEWAM.** Left: LIBERO-Plus success rates and parameters. Center: LBQs replace dense MoT video K/V, linking 30-layer Video and two-layer Action Experts. Right: real-world manipulation tasks, success rates, and measured inference latency in evaluated settings.

capacity may favor memorizing specific scene patterns over learning generalizable environmental regularities, weakening generalization. WAMs therefore need a learnable interface that operates in a compact latent space and enables the Action Expert to allocate more attention to task-relevant visual features while allowing action supervision to shape visual representations. This interface should support independent expert depths to reduce redundant parameters and improve inference efficiency.

We introduce Bridge-of-Experts (BoE), a new architecture that overcomes these limitations, and instantiate it as BRIDGEWAM, with an overview shown in Fig. 1. As illustrated in Fig. 2, BoE preserves modality-specific computation and connects the two experts through a compact set of Latent Bridge Queries (LBQs), rather than directly mixing video and action tokens. Within the Video Expert, 32 LBQs aggregate the visual, language, and proprioceptive priors needed for action generation into a shared, limited latent space, providing the Action Expert with latent features optimized for action generation through iterative learning. During end-to-end optimization, action-loss gradients flow through the LBQs into the Video Expert, guiding it to produce representations better suited to the Action Expert's learning and prediction. BoE decouples the depths of the two experts. BRIDGEWAM combines a 30-layer Video Expert with a two-layer alternating cross–self-attention Action Expert: cross-attention retrieves information from the LBQs, while self-attention coordinates the action chunk. By retaining the backbone's representational capacity with a lightweight action head, this design balances efficiency and performance. Our contributions are:

- **We introduce BoE, a video–action interface built on a limited latent space.** BoE uses 32 LBQs to construct a limited latent space that aggregates visual, language, and proprioceptive priors into a unified latent representation tailored to the Action Expert. Joint video–action optimization sends action gradients through the LBQs to shape action-relevant video representations and trains the Action Expert to prioritize task-relevant visual information.
- **We develop BRIDGEWAM with decoupled expert depths to balance efficiency and performance.** BRIDGEWAM pairs a 30-layer Video Expert with a two-layer alternating cross–self-attention Action Expert for LBQ readout and action-chunk coordination. The lightweight action head and reusable LBQ interface reduce parameter redundancy and inference cost, yielding a 5.06B-parameter BRIDGEWAM that retains the video backbone's representational capacity while balancing performance and efficiency.
- **We demonstrate that BRIDGEWAM overcomes the above limitations of MoT** through simulation and real-world task evaluations and precise ablation studies.

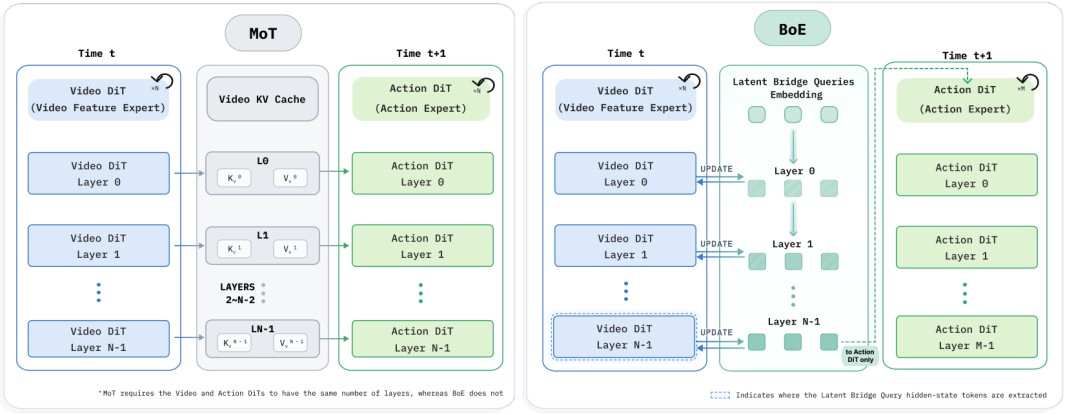


Figure 2: **Comparison of the BoE and MoT architectures.** BoE uses Latent Bridge Queries (LBQs) to transfer action-relevant visual features between experts. Unlike MoT's matched-depth coupling, BoE decouples expert depths, enabling BRIDGEWAM to use only two Action DiT layers while retaining the full Video Expert for rich visual representation learning.

## 2 RELATED WORK

**Generalist Vision–Language–Action Policies.** Generalist vision–language–action (VLA) policies map visual observations and language instructions directly to robot actions. RT-1 and RT-2 scale behavior cloning and vision–language pretraining for cross-task generalization (Brohan et al., 2023; Zitkovich et al., 2023). OpenVLA, $\pi_0$, $\pi_{0.5}$, and GR00T N1 extend generalist control using large pretrained backbones with autoregressive, diffusion, or flow-matching action decoders (Kim et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025; Bjorck et al., 2025). Despite their strong ability to generalize semantically, these action-centric models learn physical dynamics only implicitly from demonstrations annotated with robot actions. This limitation motivates the development of World Action Models, which explicitly model how visual scenes evolve over time.

**World Action Models and Video–Action Coupling.** World Action Models (WAMs) jointly model visual dynamics and robot actions, providing dense supervision for scene evolution and physical interactions (Gao et al., 2026; Cen et al., 2025; Bi et al., 2025; Kim et al., 2026). DreamZero jointly denoises future-video and action latents, enabling cross-modal interaction throughout inference (Ye et al., 2026). LingBot-VA and LingBot-VA 2.0 instead infer actions from predicted visual transitions through causal inverse dynamics (Li et al., 2026; Zhang et al., 2026). Fast-WAM reduces visual-imagination cost by training with future-video prediction but omitting future generation at inference, directly conditioning actions on latent Video DiT representations (Yuan et al., 2026). AHA-WAM further decouples their temporal roles through a low-frequency video planner and a high-frequency action executor (Cai et al., 2026). These WAMs primarily couple modalities through shared attention or by directly injecting dense video K/V representations. However, the way in which information is communicated between the heterogeneous representation spaces of video and action remains insufficiently characterized in existing action-generation pipelines.

**Cross-Modal Representation Transfer with Learnable Queries.** MetaQuery uses learnable queries to extract generation conditions from a frozen multimodal language model for transfer to a diffusion model (Pan et al., 2025). GenCeption inserts learnable tokens into a pretrained Video DiT to extract sparse structured predictions from its generative representations (Wang et al., 2026a). These studies establish compact, learnable query interfaces for task-relevant information extraction, primarily through one-way condition transfer or task readout. Building on this insight, our Latent Bridge Queries (LBQs) use a shared set of learnable queries to construct a latent space for video–action associations. The Video Expert learns to write action-relevant physical representations into this shared latent space, while the Action Expert learns to selectively retrieve from the same space the information required for control and action generation under changing physical conditions.

# 3 METHODOLOGY

## 3.1 PRELIMINARIES

**Embodied Policies.** Following Motus Bi et al. (2025), we describe embodied modeling paradigms through their conditional distributions. Let $o_t$, $l$, and $s_t$ denote the current visual observation, language instruction, and robot proprioceptive state, respectively. We use $H$ and $K$ for the action and video horizons, which cover the same prediction interval at potentially different sampling rates:

| Paradigm | Conditional distribution |
|---|---|
| VLA | $p(a_{t+1:t+H} \mid o_t, l)$ |
| WM | $p(o_{t+1:t+K} \mid o_t, a_{t+1:t+H})$ |
| Joint WAM | $p(o_{t+1:t+K}, a_{t+1:t+H} \mid o_t, l, s_t)$ |
| IDM WAM | $p(a_{t+1:t+H} \mid o_{t:t+K}, l, s_t)$ |
| Fast-WAM | $p(a_{t+1:t+H} \mid o_t, l, s_t)$ |
| BRIDGEWAM (ours) | $p(a_{t+1:t+H} \mid \mathbf{b}_t)$ |

We show canonical VLA and WM formulations and make state conditioning explicit for the WAM policies considered here. VLA predicts actions; WM predicts their visual consequences. Joint WAMs generate future observations and actions together, whereas IDM WAMs condition actions on visual transitions. Fast-WAM uses a VLA-like inference interface while retaining video supervision during training Yuan et al. (2026). The MoT-based WAM designs considered here couple their experts through shared attention over video and action representations Bi et al. (2025); Yuan et al. (2026). With one-to-one layer coupling, the Action Expert must match the Video Expert's depth.

**Bridge-of-Experts.** We propose Bridge-of-Experts (BoE) to replace this coupling with a compact latent interface. Within the Video Expert, Latent Bridge Queries (LBQs) integrate the current observation, language instruction, and robot state into a set of final hidden states:

$$\mathbf{b}_t = f_V(o_t, l, s_t; \theta_V, \mathbf{\Phi}) \in \mathbb{R}^{M \times d_V}, \qquad M = 32, \tag{1}$$

where $o_t$, $l$, and $s_t$ denote the current visual observation, language instruction, and robot state, respectively. $\theta_V$ denotes the Video Expert parameters, and $\mathbf{\Phi}$ contains the learnable initial LBQ embeddings. The encoding function $f_V$ maps these inputs and initial queries to the final, input-conditioned LBQ representations $\mathbf{b}_t$. Here, $M$ is the number of queries, and $d_V$ is the feature dimension of each token in the Video Expert. The Action Expert receives all external conditioning through $\mathbf{b}_t$, projected into its own token space. This fixed-size interface aligns visual, semantic, and proprioceptive information through joint video and action supervision without passing dense video tokens directly to the action branch during conditioning and denoising.

During training, LBQs read current-frame features and text–state context, while future-video queries read the LBQs to predict scene evolution. The generation loss over all valid future frames also updates the LBQs embeddings, encouraging predictive information that supports future-video generation be injected to them. Action-loss gradients flow through the bridge into the Video Expert, providing implicit feedback that encourages visual representations useful for control. At inference, LBQs are computed once from the current context and reused throughout action denoising. BoE therefore permits independent expert depths while keeping the interface compact: BRIDGEWAM combines a 30-layer Video Expert with a two-layer alternating cross–self-attention Action Expert for compact action conditioning at every denoising step.

## 3.2 LEARNING AND READING THE LATENT BRIDGE

In Fig. 3, we initialize LBQs with $\mathbf{B}^0 = \mathbf{\Phi}$ and propagate them through the Video DiT. At layer $\ell$, $\mathbf{V}^\ell = [\mathbf{V}_0^\ell; \mathbf{V}_f^\ell]$ contains current-frame and future-video tokens, and $\mathbf{B}^\ell$ denotes LBQ states. Future-video queries read video and LBQ keys/values, whereas LBQ queries read only current-frame and LBQ keys/values, excluding future-video keys/values from LBQ updates at every Video DiT layer:

$$\begin{aligned} \mathbf{R}_{V_f}^\ell &= \mathrm{Attn}\Big(\mathbf{Q}_{V_f}^\ell, [\mathbf{K}_V^\ell; \mathbf{K}_B^\ell], [\mathbf{U}_V^\ell; \mathbf{U}_B^\ell]; \mathbf{M}_f\Big), \\ \mathbf{R}_B^\ell &= \mathrm{Attn}\big(\mathbf{Q}_B^\ell, [\mathbf{K}_{V_0}^\ell; \mathbf{K}_B^\ell], [\mathbf{U}_{V_0}^\ell; \mathbf{U}_B^\ell]\big). \end{aligned} \tag{2}$$

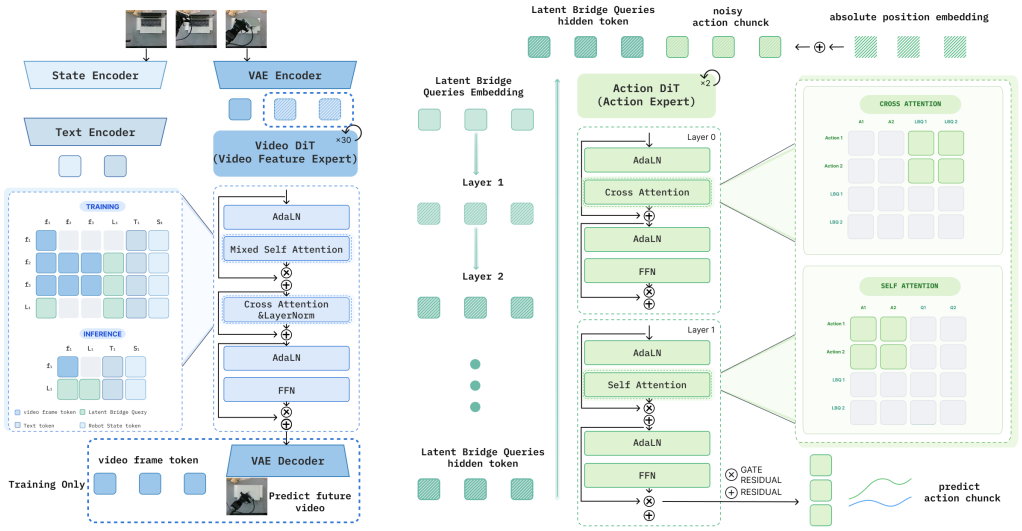


Figure 3: **Overview of the BridgeWAM architecture.** Latent Bridge Queries (LBQs) first aggregate action-relevant representations from the Video DiT and are read by the Action DiT to guide action prediction, forming a compact bridge for the video-to-action information flow.

Here, $\mathbf{Q}, \mathbf{K}, \mathbf{U}$ denote queries, keys, and values computed with shared Video DiT projections; subscripts identify source tokens, and $\cdot;\cdot$ denotes concatenation. $\mathbf{M}_f$ preserves video masking while allowing future queries to read all LBQs, whose states are updated through residual readouts, text–state cross-attention, and an FFN. Video tokens use 3D RoPE and LBQs use identity RoPE; the K/V restriction blocks LBQ access to future-video tokens. The final bridge $\mathbf{b}_t = \mathbf{B}^{L_V}$ is projected into the Action DiT space as $\widehat{\mathbf{B}} = \Pi(\mathbf{b}_t)$; action queries read these LBQ keys/values without directly receiving text–state tokens or dense video K/V. We jointly fine-tune both experts and the learnable LBQ embeddings $\mathbf{\Phi}$ with equally weighted video and action flow-matching losses. Video loss updates the Video Expert and $\mathbf{\Phi}$, while action loss updates the Action Expert, $\Pi$, $\mathbf{\Phi}$, and the Video Expert through the differentiable LBQ pathway. Thus, each LBQ update combines contributions from both losses, enabling action supervision to guide the Video Expert toward visual representations that support task-relevant perception, action prediction, and closed-loop control.

### 3.3 Alternating Cross–Self Action Head

Following the alternating design in GR00T N1 Bjorck et al. (2025), we use a two-layer Action DiT: cross-attention retrieves information from LBQs, then self-attention coordinates actions within the chunk. For noisy actions $\{\mathbf{a}_{\sigma,i}\}_{i=0}^{H-1}$, we form $\mathbf{x} = [\mathbf{x}_0; \ldots; \mathbf{x}_{H-1}]$ with learned absolute positions that distinguish temporal slots within the action chunk before cross-attention is applied:

$$\mathbf{x}_i = \mathrm{Linear}_{\mathrm{in}}(\mathbf{a}_{\sigma,i}) + \mathbf{p}_i, \qquad \mathbf{p}_i = \mathrm{Emb}_{\mathrm{pos}}(i). \tag{3}$$

Each action receives a learned embedding of its position within the chunk to distinguish temporal slots before cross-attention, while 1D RoPE in self-attention encodes relative positions between actions. The full mapping from action tokens to predicted action flow is

$$\begin{aligned} \mathbf{z} &= \mathrm{SelfBlock}_{t_A}\left(\mathrm{CrossBlock}_{t_A}(\mathbf{x}, \widehat{\mathbf{B}})\right), \\ \widehat{\mathbf{u}}_A &= \mathrm{Linear}_{\mathrm{out}}(\mathbf{z}). \end{aligned} \tag{4}$$

Each block comprises attention followed by an FFN, with residual connections and AdaLN conditioned on the diffusion timestep $t_A$. Self-attention and FFN residual branches use timestep-conditioned gates; cross-attention uses an ungated residual. The cross block uses action queries and LBQ keys/values; the self block uses only action tokens. The output $\widehat{\mathbf{u}}_A$ predicts action flow, with the same LBQs reused as the action-conditioning context at every denoising step.

Table 1: Results on LIBERO and LIBERO-Pro. We report task success rates (%) on the Spatial, Object, Goal, and Long-Horizon suites, together with their average.

| Method | Embodied PT. | LIBERO | | | | | LIBERO-Pro (zero-shot) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ |
| **OpenVLA** Kim et al. (2024) | ✓ | 84.7 | 88.4 | 79.2 | 53.7 | 76.5 | 40.5 | 32.0 | 25.0 | 20.5 | 29.5 |
| $\pi_0$ Black et al. (2024) | ✓ | 96.8 | 98.8 | 95.8 | 85.2 | 94.1 | 52.0 | 33.2 | 54.6 | 33.2 | 43.2* |
| $\pi_{0.5}$ Physical Intelligence et al. (2025) | ✓ | 98.8 | 98.2 | 98.0 | 92.4 | 96.9 | 55.6 | 52.2 | 48.0 | **57.0** | 53.0* |
| **LingBot-VA** Li et al. (2026) | ✓ | 98.5 | 99.6 | 97.2 | **98.5** | 98.5 | – | – | – | – | – |
| **Motus** Bi et al. (2025) | ✓ | 96.8 | 99.8 | 96.6 | 97.6 | 97.7 | – | – | – | – | – |
| *World Action Model* | | | | | | | | | | | |
| **Fast-WAM** Yuan et al. (2026) | ✗ | 97.8 | 99.4 | 95.4 | 94.6 | 96.8 | 57.6 | 49.2 | 58.1 | 40.8 | 51.4 |
| **Faster-WAM** Zhao et al. (2026) | ✗ | **99.6** | 99.8 | **98.2** | 98.2 | **99.0** | **59.6** | 52.8 | **59.6** | 45.8 | **54.5** |
| **BRIDGEWAM (Ours)** | ✗ | 99.0 | **100.0** | 97.4 | 96.6 | 98.3 | 57.8 | **55.1** | 58.4 | 43.5 | 53.7 |

* Reproduced using the same hyperparameter settings as the official LIBERO-Pro repository.

Table 2: Zero-shot results on LIBERO-Plus. We report task success rates (%) under seven types of perturbations and their average. "Embodied PT." indicates whether embodied pre-training is used.

| Method | Embodied PT. | LIBERO-Plus (zero-shot) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Camera ↑ | Robot ↑ | Lang. ↑ | Light ↑ | Backg. ↑ | Noise ↑ | Layout ↑ | Avg. ↑ |
| *Vision Language Model* | | | | | | | | | |
| **OpenVLA** Kim et al. (2024) | ✓ | 0.6 | 3.9 | 38.7 | 14.2 | 38.4 | 5.4 | 42.5 | 19.7 |
| **OpenVLA-OFT** Kim et al. (2025) | ✓ | 56.4 | 31.9 | 79.5 | 88.7 | 93.3 | 75.8 | 74.2 | 69.6 |
| $\pi_0$ Black et al. (2024) | ✓ | 13.8 | 6.0 | 58.8 | 85.0 | 81.4 | **79.0** | 68.9 | 53.6 |
| *World Action Model* | | | | | | | | | |
| **Fast-WAM** Yuan et al. (2026) | ✗ | 15.4 | 44.4 | 68.5 | 79.0 | 52.6 | 37.3 | 61.5 | 49.5 |
| **OpenWAM-$\alpha$** Wang et al. (2026b) | ✗ | 33.8 | **76.1** | 88.0 | **97.0** | **87.1** | 39.8 | 77.5 | 69.2 |
| **Faster-WAM** Zhao et al. (2026) | ✗ | 53.8 | 71.6 | **94.7** | 96.3 | 61.3 | 63.6 | **79.1** | **73.6** |
| **BRIDGEWAM (Ours)** | ✗ | **56.5** | 60.5 | 85.1 | 92.2 | 69.2 | 73.6 | 75.7 | 72.2 |

# 4 EXPERIMENTS

This section systematically evaluates BRIDGEWAM in terms of effective visual representation use and decoupled expert depths, examining the properties of BoE described in the Introduction. First, quantitative experiments in simulation and on real robots (Secs. 4.2 and 4.3) test whether an LBQ interface built a limited latent space and a lightweight Action Expert can maintain effective task execution and generalize under environmental perturbations and distribution shifts. Meanwhile, ablation studies (Sec. 4.4) examine the joint adaptation of visual representations and action prediction, the aggregation of multimodal priors in a limited latent space, and the effects of the Action Expert's readout mechanism on task performance. Additionally, efficiency experiments (Sec. 4.4) compare parameter counts, computational cost, and inference latency across simulation-input profiling and real-robot deployment settings, testing whether decoupled expert depths and LBQ conditioning reuse translate into practical computational savings and, together with the quantitative results, assessing the balance between performance, generalization, and inference efficiency. Finally, qualitative analysis (Sec. 4.5) illustrates how BoE guides the Action Expert to assign greater attention to visual features of target objects and task destinations during action prediction, and examines the stability of this guidance under different perturbations.

## 4.1 EXPERIMENTAL SETUP

We evaluate BRIDGEWAM on LIBERO Liu et al. (2023), LIBERO-Plus Fei et al. (2025), LIBERO-Pro Zhou et al. (2025), RoboTwin 2.0 Chen et al. (2025), and real-world manipulation tasks. Unless otherwise specified, BRIDGEWAM initializes its 30-layer Video Expert from Wan2.2- TI2V-5B Wan et al. (2025) and employs 32 Latent Bridge Queries together with a lightweight two-layer alternating cross–self-attention Action Expert. The video and action objectives are jointly optimized with equal weights. The model is trained for 10 epochs on LIBERO and 5 epochs on RoboTwin 2.0. LIBERO experiments use two RGB camera views, whereas RoboTwin 2.0 uses three views. During evaluation, we use 10 action-denoising steps and report task success rates following the evaluation protocol and corresponding task definitions of each benchmark.

## 4.2 QUANTITATIVE EVALUATION IN SIMULATION

**LIBERO Benchmarks.** BRIDGEWAM averages 98.3% on LIBERO (1.45 points above Fast-WAM), 100% on LIBERO-Object, and 53.7% on LIBERO-Pro (Table 1). On LIBERO-Plus, it raises Fast-WAM's average from 49.5% to 72.2%, with the best Camera score (56.5%; Table 2).

Table 3: Fine-grained task comparison on RoboTwin 2.0 Simulation (Clean vs. Randomized).

| Task | $\pi_0$ Black et al. 2024 | | $\pi_{0.5}$ Physical Intelligence et al. 2025 | | Motus Bi et al. 2025 | | Motus from WAN2.2 Bi et al. 2025 | | LingBot-VA Li et al. 2026 | | Fast-WAM Yuan et al. 2026 | | Faster-WAM Zhao et al. 2026 | | BridgeWAM (Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C | R | C | R | C | R | C | R | C | R | C | R | C | R | C | R |
| Adjust Bottle | 99 | 95 | **100** | 99 | 89 | 93 | 99 | 97 | 90 | 94 | **100** | **100** | **100** | 99 | **100** | 99 |
| Blocks Ranking RGB | 80 | 63 | 92 | 85 | 99 | 97 | 92 | 88 | 99 | 98 | **100** | **100** | **100** | **100** | 99 | 99 |
| Click Bell | 71 | 48 | 99 | 66 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | 99 | 99 |
| Grab Roller | 98 | 94 | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |
| Handover Mic | 97 | 97 | 98 | 97 | 78 | 63 | 98 | 95 | 94 | 96 | 99 | **100** | **100** | 99 | **100** | **100** |
| Lift Pot | 80 | 72 | 96 | 85 | 96 | 99 | 90 | 87 | **100** | 99 | **100** | **100** | **100** | **100** | **100** | **100** |
| Shake Bottle Horizontally | 98 | 92 | 99 | 99 | **100** | 98 | **100** | 97 | **100** | 99 | **100** | **100** | **100** | **100** | **100** | **100** |
| Shake Bottle | 94 | 91 | 99 | 97 | **100** | 97 | **100** | 96 | **100** | 97 | **100** | **100** | **100** | **100** | **100** | **100** |
| Move Pillbottle Pad | 67 | 46 | 84 | 61 | 93 | 96 | 83 | 83 | 99 | 99 | **100** | 99 | 98 | 99 | 99 | **100** |
| Move Playingcard Away | 74 | 65 | 96 | 84 | **100** | 96 | 50 | 47 | **100** | 99 | **100** | **100** | **100** | **100** | 98 | **100** |
| .....(50 tasks) | | | | | | | | | | | | | | | | |
| Click Alarmclock | 77 | 68 | 98 | 89 | **100** | **100** | **100** | 99 | 99 | **100** | **100** | **100** | 99 | **100** | 97 | 99 |
| Dump Bin Bigbin | 88 | 83 | 92 | 97 | 95 | 91 | 94 | 96 | 89 | 96 | 97 | 96 | 97 | **98** | **98** | 97 |
| Open Laptop | 71 | 81 | 90 | 96 | 95 | 91 | 89 | 89 | 92 | 94 | 98 | **100** | 95 | 99 | **99** | **100** |
| Place Container Plate | 97 | 92 | **99** | 95 | 98 | 99 | 97 | **100** | **99** | 97 | 96 | **100** | 98 | 97 | **99** | **100** |
| Place Burger Fries | 81 | 76 | 94 | 87 | **98** | 98 | 95 | 90 | 97 | 95 | 96 | **99** | **98** | 97 | **98** | 98 |
| Place Object Stand | 82 | 68 | 91 | 85 | 98 | 97 | 91 | 93 | **99** | 96 | 90 | 94 | 95 | 96 | 96 | **98** |
| Place Phone Stand | 49 | 53 | 81 | 81 | 87 | 86 | 80 | 78 | 97 | 97 | 97 | **99** | **100** | 96 | 98 | 95 |
| Place Shoe | 76 | 76 | 92 | 93 | **99** | 97 | 95 | 92 | 98 | 98 | 96 | **99** | 97 | 98 | 97 | 96 |
| Stack Bowls Two | 94 | 95 | 95 | 96 | 98 | **98** | 98 | **98** | 94 | **98** | 92 | **98** | 98 | 96 | **99** | 97 |
| Stack Blocks Two | 93 | 79 | 97 | **100** | **100** | 98 | 96 | 94 | **100** | 98 | **100** | **100** | **100** | **100** | **100** | 99 |
| Stack Bowls Three | 77 | 75 | 77 | 71 | 79 | 87 | **90** | 74 | 86 | 83 | 80 | 81 | 83 | 78 | **88** | **90** |
| **Average (%)** | 65.92 | 58.40 | 82.74 | 76.76 | 88.66 | 87.02 | 77.56 | 77.00 | **92.90** | 91.50 | 91.88 | 91.78 | 92.78 | **92.26** | 89.84 | 90.22 |

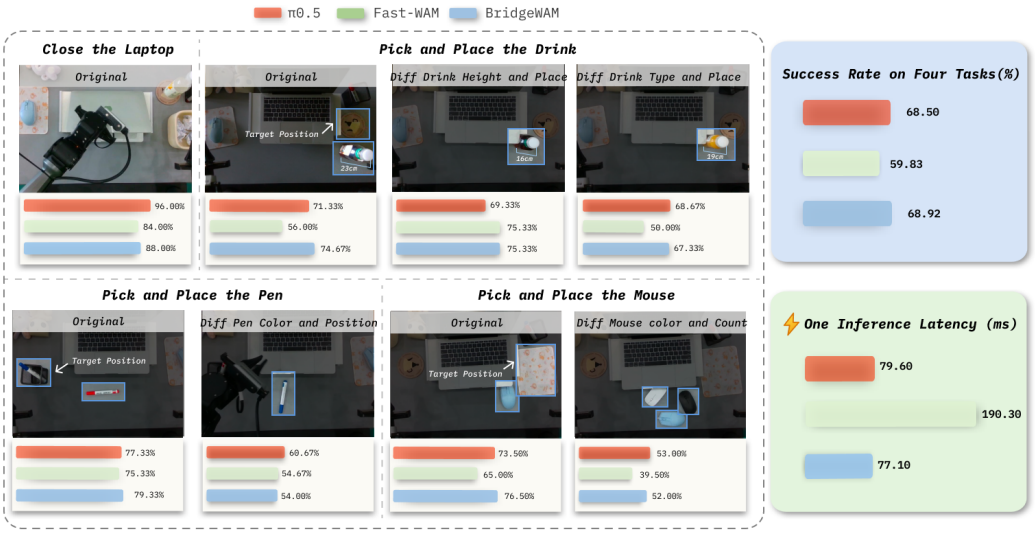


Figure 4: **Real-world manipulation tasks and results.** We compare $\pi_{0.5}$, Fast-WAM, and BRIDGEWAM on closing a laptop and picking and placing a drink, a pen, and a mouse. The task panels show the original settings and variations in object height, type, color, position, or count, together with success rates (%). The right panels summarize success rates and inference latency (ms).

**RoboTwin 2.0.** We train for five epochs (29,355 steps) on Fast-WAM's 27,500 demonstrations (2,500 clean; 25,000 randomized), evaluating 50 tasks with 100 trials per task and setting (Table 3). Under randomization, BRIDGEWAM scores 90% on *Stack Bowls Three* (Faster-WAM: 78%; LingBot-VA: 83%), 98% on *Place Object Stand*, and 100% on *Move Pillbottle Pad*. Success on *Handover Mic*, *Lift Pot*, and *Open Laptop* is near-perfect in both settings.

**Analysis.** Across LIBERO, LIBERO-Plus, LIBERO-Pro, and RoboTwin 2.0, BRIDGEWAM supports effective task execution and strong generalization with an ultralight two-layer Action Expert. Faster-WAM achieves higher aggregate success rates in these comparisons, potentially benefiting from future-aware conditioning beyond current-observation features (Zhao et al., 2026), although the present comparison does not isolate this effect. Together with the efficiency analysis in Sec. 4.4, these results highlight BoE's performance–efficiency balance: BRIDGEWAM uses 5.06B parameters versus Faster-WAM's 5.56B, approximately 9.0% fewer (Fig. 1), and requires 54.67% fewer profiled FLOPs under the respective native configurations (Table 6). The recorded H20Z latency is 88.05 ms versus 321.16 ms, approximately $1/3.65$ of Faster-WAM's. These results demonstrate that BoE balances effective task execution with reduced computational cost and inference latency.

Table 4: Action-head depth ablation on LIBERO and LIBERO-Plus. We report task success rates (%); the default uses two action layers.

| Method | LIBERO | | | | | LIBERO-Plus (Zero-shot) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ |
| **Fast-WAM** | 97.8 | 99.4 | 95.4 | 94.6 | 96.80 | 52.21 | 69.06 | 37.32 | 40.06 | 49.54 |
| **BRIDGEWAM (Ours)** | **99.0** | **100.0** | 97.4 | 96.6 | **98.25** | **77.64** | 84.47 | **60.29** | **67.17** | **72.24** |
| **4 layers** | 97.2 | **100.0** | 97.2 | 96.6 | 97.75 | 67.44 | **84.63** | 54.30 | 60.26 | 66.56 |
| **8 layers** | 97.4 | 99.8 | **97.6** | **96.8** | 97.90 | 70.69 | 83.76 | 49.83 | 58.44 | 65.50 |
| **16 layers** | 98.8 | 99.4 | 96.2 | 93.4 | 96.95 | 67.94 | 80.70 | 52.37 | 62.13 | 65.66 |
| **30 layers (self-cross)** | 97.8 | **100.0** | **97.6** | 94.8 | 97.55 | 69.40 | 78.04 | 60.09 | 64.31 | 67.89 |

Table 5: Interface and video–action coupling ablations on LIBERO and LIBERO-Plus. We report task success rates (%);

| Method | LIBERO | | | | | LIBERO-Plus (Zero-shot) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ | Spatial ↑ | Object ↑ | Goal ↑ | Long ↑ | Avg. ↑ |
| **Fast-WAM** | 97.8 | 99.4 | 95.4 | 94.6 | 96.80 | 52.21 | 69.06 | 37.32 | 40.06 | 49.54 |
| **BRIDGEWAM (Ours)** | **99.0** | **100.0** | 97.4 | 96.6 | **98.25** | **77.64** | **84.47** | **60.29** | **67.17** | **72.24** |
| **Frozen-Video-DiT** | 87.8 | 97.0 | 84.8 | 72.4 | 85.50 | 32.22 | 38.52 | 15.44 | 16.36 | 25.48 |
| **MoT-readout** | 98.0 | 99.4 | 97.4 | 96.4 | 97.80 | 71.57 | 80.70 | 36.09 | 54.78 | 60.48 |
| **Text-state-LBQ-input** | 96.4 | 99.2 | 97.0 | 96.2 | 97.20 | 64.70 | 79.98 | 44.11 | 49.07 | 59.29 |
| **BRIDGEWAM Joint** | 97.4 | 99.6 | 96.6 | **96.8** | 97.60 | 76.52 | 72.12 | 46.51 | 59.59 | 63.41 |
| **BRIDGEWAM IDM** | 98.8 | 97.8 | **99.0** | 96.6 | 98.05 | 76.23 | 71.49 | 58.55 | 64.19 | 67.45 |

### 4.3 QUANTITATIVE EVALUATION ON REAL ROBOTS

Figure 4 compares task success and inference latency across the three methods. We deploy BRIDGEWAM on the *PiPER X from SongLing* real-robot platform and evaluate four manipulation tasks with their OOD settings. We use task-specific staged scoring, with final scores computed over completed operation stages; for example, within a single trial, successfully grasping an object and achieving the target each earn credit for their respective stages. For drink pick-and-place, grasping the drink, placing it within the target region, and leaving it upright each earn one point. Scoring criteria are detailed in Appendix A.4. These results show that the LBQ-mediated video–action interface transfers beyond simulation and maintains stable, action-relevant representations under visual variation, observation noise, and execution disturbances during closed-loop manipulation.

### 4.4 ABLATION STUDIES AND EFFICIENCY

Table 4 shows that increasing the Action Expert depth from two to eight layers reduces the LIBERO-Plus average from 72.24% to 65.50%, while scaling the action head does not improve the standard LIBERO score. This pattern supports our hypothesis that excessive action-side capacity can encourage the model to absorb task-irrelevant visual details, weakening its robustness to distribution shifts rather than improving action prediction. The four-, sixteen-, and thirty-layer variants produce LIBERO-Plus averages of 66.56%, 65.66%, and 67.89%, respectively, consistent with the same capacity–generalization trade-off across the tested action-head depths.

Table 5 shows that the Frozen-Video-DiT and MoT-readout variants probe two information-transfer directions of the LBQ interface. In the Frozen-Video-DiT variant, the Action Expert receives visual weights learned by Fast-WAM on LIBERO, but the frozen representation cannot be read effectively for action prediction: the LIBERO-Plus average falls to 25.48%. This result indicates that video features need to be learned in a form connected to action learning, rather than treated as a fixed representation. The MoT-readout variant reaches 60.48%, below the full model's 72.24%, suggesting that LBQs learned through action cross-attention provide a more effective readout than the conventional MoT mechanism. The text-state-LBQ-input variant reaches 59.29%, which further suggests that the prior information used by the Action Expert can be compressed into a finite LBQ latent space, self-aligned across modalities, and jointly read and integrated by action tokens. Under the same backbone and action-head configuration, the Joint and IDM variants reach 63.41% and 67.45%, respectively, comparing BoE with existing video–action coupling paradigms under con-

Table 6: Inference efficiency on LIBERO with 30-layer Video DiTs, H20Z GPUs, `bfloat16`, an action horizon of 32, and 10 denoising steps. Appendix A.8 details the profiling scope.

| Method | Layers | | FLOPs (T) | | | Acceleration | | Peak Mem. (MiB) |
|---|---|---|---|---|---|---|---|---|
| | Video | Action | Video ↓ | Action ↓ | Total ↓ | Latency (ms) ↓ | Speedup ↑ | |
| **Fast-WAM** Yuan et al. (2026) | 30 | 30 | **1.005** | 1.043 | 2.048 | 405.91 | 1.00× | 23,773.86 |
| **Faster-WAM** Zhao et al. (2026) | 30 | 30 | 2.745 | 0.476 | 3.221 | 321.16 | 1.26× | 22,962.00 |
| **BRIDGEWAM** | 30 | 2 | 1.432 | **0.028** | **1.460** | **88.05** | **4.61**× | **21,912.21** |

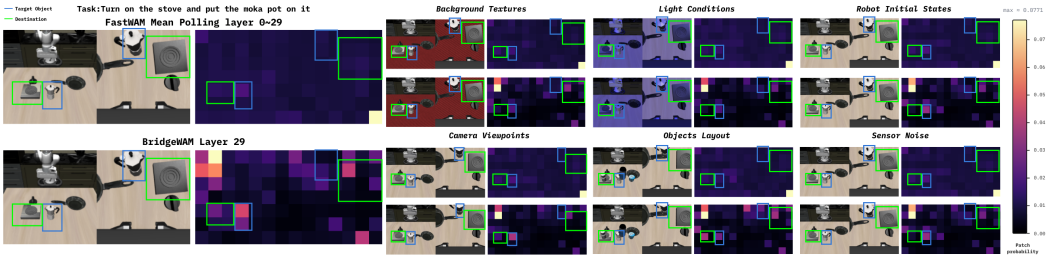


Figure 5: **Action-conditioned visual routing maps.** The top row shows Fast-WAM attention averaged across Video-DiT layers, while the bottom row shows the LBQ-mediated routing at the final Video-DiT layer of BRIDGEWAM. The columns cover variations in background texture, lighting, camera viewpoint, robot initial state, object layout, and sensor noise.

trolled conditions. Detailed designs for all ablations and the model pipelines of the two variants are provided in the supplementary material for reproducibility and detailed comparison.

Table 6 also quantifies our architecture's efficiency via Inference cost, inference latency, and FLOPs. All three models use H20Z GPUs, `bfloat16`, an identical cached LIBERO input, a 32-step action horizon, and 10 denoising steps. Latency averages 10 CUDA-event timings after three warm-ups, from just before Video DiT layer 0 to the final clean action latent, excluding observation preprocessing, VAE/text encoding, and postprocessing. Runtime-shape FLOPs count linear and attention matrix multiplications and linear biases, counting each multiply–add as two FLOPs; action FLOPs include conditioning and all denoising steps. Appendix A.8 details the measurement protocol.

### 4.5 QUALITATIVE ANALYSIS

**Action-Conditioned Visual Routing.** To visualize which visual information is preferentially accessed by the Action Expert, we construct an action-conditioned visual-routing map. For BRIDGEWAM, we use the Action Expert's attention scores over the latent bridge queries as query-importance weights and apply them to the corresponding LBQ-to-video-patch attention maps from the Video Expert. For Fast-WAM, we directly aggregate the action-to-visual-token attention across all Video-DiT layers. The resulting patch-space maps reveal the visual information preferentially routed toward action prediction; they should be interpreted as routing visualizations rather than strict causal attributions. Figure 5 shows representative routing maps for the two models under changes in background texture, lighting, camera viewpoint, robot initial state, object layout, and sensor noise. Experimental results show that the BoE not only enables the Action Expert to focus its attention on visual features that are more relevant to the current task, but also improves robustness to perturbations and enhances the model's generalization performance across the evaluated visual changes.

## 5 CONCLUSION

We presented BoE, a differentiable LBQ interface that aligns video representations with action prediction and enables selective retrieval by the Action Expert. Reusing the compact bridge supports independent expert depths and efficient conditioning, while experiments validate its promise for robust, longer-horizon control across diverse settings. Additional real-robot tests on long-horizon tasks, not included in the reported evaluation, also indicate limited performance.

### AI USE STATEMENT

We used large language models (LLMs) to polish the language of author-written text for clarity, fluency, and readability. For example, when preparing the Related Work section, the authors first read and organized relevant papers and manually added their citations to the bibliography. Based on their understanding of the field's technical development, the authors then drafted the discussion of how prior work relates to the core technical ideas and how these ideas evolved. The resulting author-written paragraphs were subsequently provided to LLMs for language polishing. The authors take responsibility for the final manuscript, including its claims, citations, and reported results.

### REPRODUCIBILITY STATEMENT

We provide a link to the project homepage at the end of the abstract and a link to our anonymous Git repository at the beginning of the appendix. To support reproducibility, we describe the model architecture and training objectives in Section 3, and the experimental setup and simulation evaluation protocols in Sections 4.1 and 4.2. Appendix A.1 describes action-head initialization, and Appendix A.2 specifies the ablation variants. The staged scoring approach for real-robot experiments is described in Appendix A.4. Appendix A.8 documents the hardware, numerical precision, timing boundaries, and FLOPs accounting used for efficiency measurements.

## REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025. doi: 10.48550/arXiv.2512.13030. URL https://arxiv.org/abs/2512.13030.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Jim Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. doi: 10.48550/arXiv.2503.14734. URL https://arxiv.org/abs/2503.14734.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. doi: 10.48550/arXiv.2410.24164. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2023. doi: 10.48550/arXiv.2212.06817. URL https://arxiv.org/abs/2212.06817.

Jisong Cai, Long Ling, Shiwei Chu, Zhongshan Liu, Jiayue Kang, Zhixuan Liang, Wenjie Xu, Yinan Mao, Weinan Zhang, Xiaokang Yang, Ru Ying, Ran Zheng, and Yao Mu. AHA-WAM: Asynchronous horizon-adaptive world-action modeling with observation-guided context routing.

*arXiv preprint arXiv:2606.09811*, 2026. doi: 10.48550/arXiv.2606.09811. URL https://arxiv.org/abs/2606.09811.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025. doi: 10.48550/arXiv.2506.21539. URL https://arxiv.org/abs/2506.21539.

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. doi: 10.48550/arXiv.2506.18088. URL https://arxiv.org/abs/2506.18088.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. doi: 10.48550/arXiv.2510.13626. URL https://arxiv.org/abs/2510.13626.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi Jim Fan. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026. doi: 10.48550/arXiv.2602.06949. URL https://arxiv.org/abs/2602.06949.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. doi: 10.48550/arXiv.2406.09246. URL https://arxiv.org/abs/2406.09246.

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

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026. doi: 10.48550/arXiv.2601.16163. URL https://arxiv.org/abs/2601.16163.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. doi: 10.48550/arXiv.2601.21998. URL https://arxiv.org/abs/2601.21998.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. doi: 10.48550/arXiv.2306.03310. URL https://arxiv.org/abs/2306.03310.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025. doi: 10.48550/arXiv.2504.06256. URL https://arxiv.org/abs/2504.06256.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A

vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. doi: 10.48550/arXiv.2504.16054. URL https://arxiv.org/abs/2504.16054.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Video generation models are general-purpose vision learners. *arXiv preprint arXiv:2607.09024*, 2026a. doi: 10.48550/arXiv.2607.09024. URL https://arxiv.org/abs/2607.09024.

Y. Wang et al. OpenWAM: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, September 2026b. doi: 10.48550/arXiv.2609.07398. URL https://arxiv.org/abs/2609.07398.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, February 2026. doi: 10.48550/arXiv.2602.15922. URL https://arxiv.org/abs/2602.15922.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, March 2026. doi: 10.48550/arXiv.2603.16666. URL https://arxiv.org/abs/2603.16666.

Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, Jiaming Zhou, Yishu Shen, Yudong Jin, Fangyi Xu, Shuailei Ma, Jiaqi Liao, Guanxing Lu, Zifan Shi, Yongkun Wen, Yujie Zhao, Weixuan Tang, Xinyang Wang, Chaojian Li, Jiapeng Zhu, Ka Leong Cheng, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026. doi: 10.48550/arXiv.2607.08639. URL https://arxiv.org/abs/2607.08639.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-wam: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026. doi: 10.48550/arXiv.2608.04404. URL https://arxiv.org/abs/2608.04404.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *[arXiv preprint arXiv:2510.03827*, 2025. doi: 10.48550/arXiv.2510.03827. URL https://arxiv.org/abs/2510.03827.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag Sanketi, Grecia Salazar, Michael Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Tsang-Wei Edward Lee, Lisa Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Kehang Han, Keerthana Gopalakrishnan, Montse Gonzalez Arenas, Chuyuan

Fu, Pete Florence, Chelsea Finn, Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, and Anthony Brohan. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. doi: 10.48550/arXiv.2307.15818. URL https://arxiv.org/abs/2307.15818.