LIBERO-VPro: Benchmarking Closed-Loop Visual Robustness of Robotic Foundation Models

Huiqiong Li[1], Zhiting Mei[2], Anirudha Majumdar[2], Jingjing Chen[3], Yu-Gang Jiang[3], Bin Zhu[1†]
[1]Singapore Management University [2]Princeton University [3]Fudan University
Correspondence: binzhu@smu.edu.sg
https://huiqiongli.github.io/LIBERO-VPro/

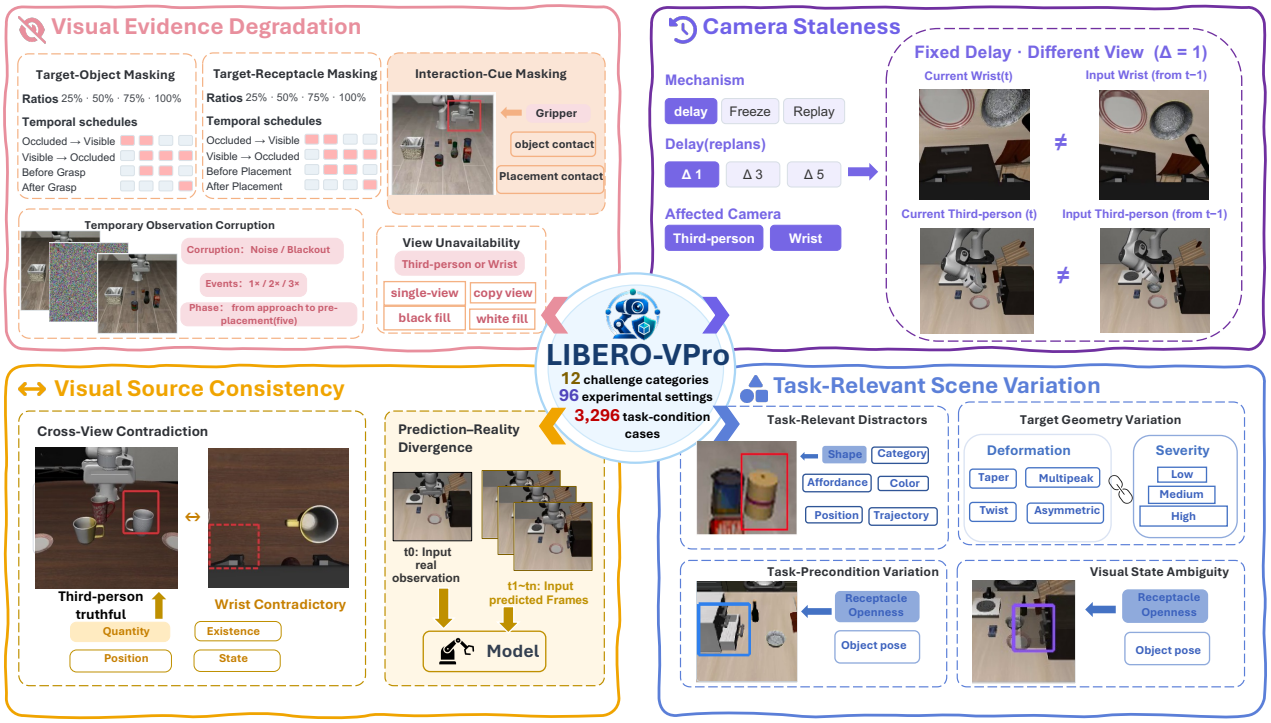

Fig. 1. Overview of LIBERO-VPro. The benchmark organizes 12 challenge categories into four complementary dimensions of closed-loop visual robustness: Visual Evidence Degradation, Camera Staleness, Visual Source Consistency, and Task-Relevant Scene Variation. LIBERO-VPro comprises 96 experimental settings and 3,296 task-condition cases.

***Abstract*— Robotic foundation models achieve impressive performance on standard manipulation benchmarks, yet these evaluations typically assume clean, timely, and consistent visual observations throughout execution. We introduce LIBERO-VPro, a benchmark for systematically evaluating the closed-loop visual robustness of robotic foundation models by perturbing the visual evidence available during execution. LIBERO-VPro covers four complementary dimensions, including Visual Evidence Degradation, Camera Staleness, Visual Source Consistency, and Task-Relevant Scene Variation, spanning 12 challenge categories, 96 experimental settings, and 3,296 task-condition cases. We evaluate three vision-language-action models and three world-action models over approximately 196,000 simulated episodes, complemented by 200 real-world rollouts on a Franka Research 3. Our results reveal that strong nominal performance can mask substantial weaknesses in visual grounding and adaptation. Models often remain successful despite severe object-level occlusion, yet degrade sharply when local interaction cues are disrupted or familiar spatial priors are violated. They are also highly sensitive to stale or missing observations and struggle when changed task preconditions require behavioral adaptation. Finally, VLAs and WAMs exhibit distinct robustness profiles, showing that visual robustness is multi-dimensional and architecture-dependent. LIBERO-VPro provides a systematic diagnostic framework for developing robotic foundation models that can more reliably ground and adapt their actions under challenging visual conditions.**

†Corresponding author and project lead.

I. Introduction

Robotic foundation models are rapidly advancing toward general-purpose manipulation by integrating visual observations, language instructions, and action generation. Vision-language-action models (VLAs) connect visual and language representations to robot control [11, 12, 13, 14], while world-action models (WAMs) incorporate future world states prediction into policy learning [15, 16, 17]. Benchmarks such as LIBERO [1] and RoboTwin [2] have enabled standardized evaluation across diverse manipulation tasks. However, high task success under nominal conditions does not necessarily imply that a policy can robustly use visual feedback during closed-loop execution.

TABLE I

COMPARISON WITH REPRESENTATIVE ROBOTIC MANIPULATION BENCHMARKS IN TERMS OF VISUAL ROBUSTNESS COVERAGE. ✓ INDICATES EXPLICIT COVERAGE, △ PARTIAL CORRESPONDENCE, AND ✗ NO COVERAGE.

Benchmark	Visual Evidence Degradation					Visual Source Consistency		Task-Relevant Scene Variation				Camera Staleness
	Target-Obj. Mask.	Target-Rec. Mask.	Interact. Cue Mask.	Temp. Obs. Corr.	View Unavail.	Cross-View Contrad.	Pred.–Reality Divergence	Task-Rel. Distr.	Task Precond. Var.	Visual State Ambig.	Target Geom. Var.	
LIBERO [1]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
RoboTwin 2.0 [2]	✗	✗	✗	✗	✗	✗	✗	△	✗	✗	✗	✗
RoboCasa [3]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
COLOSSEUM [4]	✗	✗	✗	✗	✗	✗	✗	△	✗	✗	△	✗
VLATest [5]	✗	✗	✗	✗	✗	✗	✗	△	✗	✗	✗	✗
RobustVLA [6]	✗	✗	✗	△	✗	✗	✗	△	✗	✗	✗	✗
LIBERO-PRO [7]	✗	✗	✗	✗	✗	✗	✗	✗	△	✗	△	✗
LIBERO-Plus [8]	✗	✗	✗	△	✗	✗	✗	△	△	✗	✗	✗
VLA-Arena [9]	✗	✗	✗	△	✗	✗	✗	△	✗	✗	✗	✗
RoboTrustBench [10]	△	△	✗	✗	✗	✗	✗	△	△	△	✗	✗
LIBERO-VPro(Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Standard manipulation benchmarks typically assume that observations remain complete, synchronized, and mutually consistent throughout a rollout. For physical deployment, nevertheless, the visual evidence available during execution may not always be reliable. Task-relevant objects can become occluded by the robot or surrounding clutter; camera streams can be temporarily corrupted, delayed, or unavailable; different viewpoints can provide conflicting evidence; and the physical scene can change in ways that invalidate previously successful action patterns. These failures occur inside the perception–action loop, where the policy must continuously interpret imperfect observations and adjust its behavior accordingly. Existing robustness benchmarks, such as LIBERO-Plus [8], LIBERO-PRO [7] and VLA-Arena [9], have substantially expanded evaluation beyond nominal settings by varying object appearance, lighting, viewpoints, initial states, instructions, and environments. Nevertheless, as shown in Table I, these evaluations primarily perturb the scene or task configuration, while the visual observations available to the policy during closed-loop execution generally remain temporally coherent and internally consistent. Consequently, an important question remains underexplored: *How robust are robotic foundation models when the visual evidence they rely on becomes incomplete, stale, inconsistent, or behaviorally misleading during closed-loop execution?*

To address this question, we introduce **LIBERO-VPro**, a benchmark for closed-loop visual robustness of robotic foundation models built upon LIBERO. As shown in Fig. 1, LIBERO-VPro consists of four complementary dimensions. These dimensions are designed to isolate four fundamental requirements for reliable closed-loop visual control, whether task-relevant evidence is available, temporally current, mutually consistent across sources, and sufficient to support behavioral adaptation when the scene changes. *Visual Evidence Degradation* removes or corrupts task-relevant visual information at different spatial and temporal scales. *Camera Staleness* introduces temporal mismatch between observations and the current physical state. *Visual Source Consistency* introduces contradictory evidence across views or between predicted and realized observations. *Task-Relevant Scene Variation* modifies visual properties that should require the policy to re-ground its behavior in the current scene. Together, these dimensions span 12 challenge categories, 96 experimental settings, and 3,296 task-condition cases, enabling controlled diagnosis of different sources of closed-loop visual failure.

We conduct a comprehensive evaluation using six representative robotic foundation models on LIBERO-VPro, including three VLAs, OpenVLA-OFT [18], π_0 [19], $\pi_{0.5}$ [14], and three WAMs, FastWAM [17], LingBot-VA [16], LaWAM [20], over approximately 196,000 simulated episodes. The evaluation reveals several failure modes that are largely hidden by nominal task success. First, policies often remain successful when the target object or receptacle is occluded, but degrade sharply when interaction cues such as the gripper is removed. Second, models handle appearance-based distractors well but fail when distractors violate familiar spatial layouts, revealing strong reliance on learned spatial priors. Third, changing task preconditions is far more disruptive than introducing visual ambiguity, highlighting limited behavioral adaptation. Finally, VLAs and WAMs exhibit distinct robustness profiles, showing that visual robustness is inherently multi-dimensional. To examine whether these failures extend beyond simulation, we conduct 200 real-world rollouts on a Franka Research 3 using $\pi_{0.5}$ and LaWAM across two manipulation tasks. Representative perturbations from all four LIBERO-VPro dimensions reveal broadly consistent trends. Both models are highly sensitive to missing views, cross-view inconsistencies, and violations of familiar spatial layouts, while robustness to stale observations varies by task and model. These results indicate that LIBERO-VPro captures failure modes that also arise in physical closed-loop manipulation. Our main contributions are as follows:

- We introduce LIBERO-VPro, a benchmark that systematically perturbs visual observations *during* closed-loop execution along four complementary dimensions, covering 12 challenge categories, 96 experimental settings, and 3,296 task-condition cases.
- We conduct a comprehensive evaluation of three VLA

and three WAM models, producing approximately 196,000 simulated episodes, complemented by 200 physical rollouts on a Franka Research 3, providing controlled analysis of visual robustness across both simulation and real-world manipulation.
- We reveal strong reliance on spatial priors, vulnerability to disrupted interaction cues and scene changes, and distinct robustness profiles across VLAs and WAMs.

II. Related Work

A. Robotic Foundation Models

Vision-language-action (VLA) models extend vision-language pretraining to continuous robot control, from RT-1/ RT-2 [11, 12] to OpenVLA [13] and embodied multimodal reasoning [21]. Our VLA baselines include OpenVLA-OFT [18], π_0 [19], and $\pi_{0.5}$ [14]. World-action models (WAMs) augment control with future-state prediction or latent dynamics: GR-2 [22], FastWAM [17], LingBot-VA [16], and LaWAM [20] represent video-generation pretraining, test-time imagination, causal world modeling, and latent-dynamics policy generation, respectively. VLAs map observations directly to actions, whereas WAMs explicitly model future dynamics, motivating their joint robustness evaluation.

B. Robotic Manipulation Benchmarks

LIBERO [1] and nominal simulators [23, 24, 25, 26, 27] provide standardized manipulation tasks, while MimicGen [28], GenAug [29], and RoboCasa [3] scale data and scene diversity. Recent benchmarks evaluate appearance, lighting, background, and viewpoint changes [4], progressive difficulty [8], object, initial-state, instruction, and environment shifts [7], domain randomization [2, 30], confounding objects [5], multimodal perturbations [6], or offline generated-video quality [10]. Existing methods primarily modify scene, task, or generic visual conditions without testing our stage-specific transient corruption, cross-view contradiction, or recursive predicted-feedback settings. LIBERO-VPro instead perturbs visual availability, temporal freshness, cross-source consistency, and task-relevant scene content during execution. Table I compares their coverage.

III. Benchmark Construction

LIBERO-VPro evaluates closed-loop visual robustness by systematically modifying the visual evidence available to a manipulation policy during execution. As shown in Fig. 1, LIBERO-VPro organizes visual challenges into four complementary dimensions: Visual Evidence Degradation, Camera Staleness, Visual Source Consistency, and Task-Relevant Scene Variation. The benchmark spans 12 challenge categories, 96 experimental settings, and 3,296 task-condition cases. Fig. 2 summarizes the complete hierarchy.

A. Visual Evidence Degradation

Reliable manipulation should not depend on every visual cue remaining continuously available. In realistic execution, task-relevant entities may be occluded by the robot, manipulated objects, or surrounding clutter, while camera observations may be temporarily corrupted or lost. We therefore

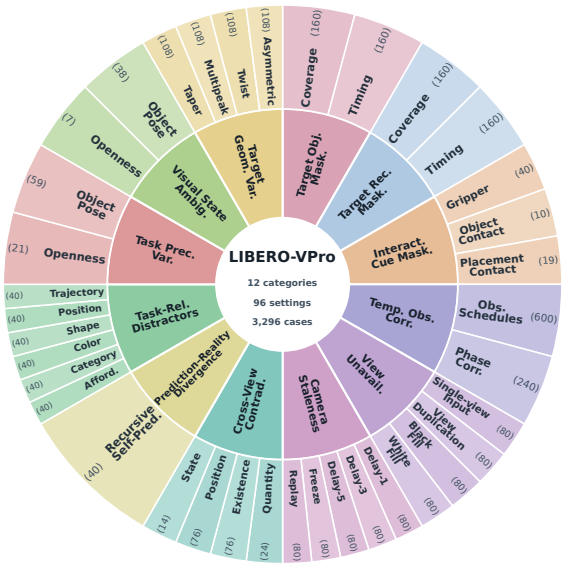

Fig. 2. LIBERO-VPro benchmark hierarchy: 12 challenge categories decomposed into 96 experimental settings totaling 3,296 cases.

evaluate information loss at multiple spatial scales, from individual objects and interaction regions to complete frames and camera views.

1) Target-Object and Target-Receptacle Masking: We separately occlude the target object and target receptacle to test whether policies depend on their online visual appearance during execution. Each category includes four masking ratios (25%, 50%, 75% and 100%) and four temporal schedules that control when the relevant entity becomes visible or occluded, including transitions around grasping or placement.

2) Interaction-Cue Masking: We selectively remove local interaction cues while preserving the surrounding scene. Perturbations target the robot gripper, gripper-object contact region, and placement contact region, enabling us to isolate the importance of fine-grained visual feedback during manipulation.

3) Temporary Observation Corruption: We introduce transient full-frame corruption using either blackout or visual noise. Corruption is controlled by both occurrence frequency and manipulation phase from approach to pre-placement.

4) View Unavailability: We selectively remove either the third-person or wrist-camera observation. We consider single-view input as well as replacements using constant images or duplicated observations from the remaining camera.

B. Camera Staleness

Closed-loop control requires observations to reflect the robot's current physical state. In practice, camera latency, frame drops, and asynchronous sensor pipelines can cause policies to act on outdated observations, potentially producing actions that are individually plausible but inappropriate for the present state. We simulate such temporal mismatch through three fixed-delay levels, observation freezing, and

historical replay. Each perturbation is independently applied to either the third-person or wrist camera while the other view remains current. This design allows us to measure both overall sensitivity to stale observations and the relative dependence of each policy on individual camera streams.

C. Visual Source Consistency

Robotic foundation models increasingly integrate information from multiple visual sources, including external camera views (third-person and wrist) and, for WAMs, internally predicted future observations. Reliable control requires these sources to remain consistent about the task-relevant state of the environment. Unlike Camera Staleness, which tests whether an observation is temporally current, Visual Source Consistency examines failures caused by disagreement between different sources of visual evidence. We consider two settings: conflicts across physical camera views and divergence between predicted and realized visual futures.

1) Cross-View Contradiction: Multi-view observations may provide inconsistent evidence due to occlusion, viewpoint-dependent visibility, or perception errors. We introduce controlled contradictions between the third-person and wrist cameras along four task-relevant attributes: quantity, existence, position, and state. One view is manipulated while the other remains truthful, with both kept temporally aligned. This setting tests whether policies can reconcile conflicting visual evidence rather than simply rely on a preferred camera view.

2) Prediction–Reality Divergence: WAMs may additionally rely on internally predicted future observations for action generation, making the consistency between imagined and realized scene evolution critical for closed-loop control. We therefore provide a ground-truth observation only at initialization and recursively feed the model's predicted frames back as subsequent visual input. This setting examines whether prediction errors accumulate into control failures when imagined futures increasingly replace observations of the realized environment. Because explicit future-frame generation is required, this evaluation is applied only to compatible WAMs.

D. Task-Relevant Scene Variation

The previous three dimensions perturb the quality, timeliness, or consistency of visual observations while largely preserving the underlying task configuration. In contrast, Task-Relevant Scene Variation changes the scene itself while keeping observations complete, current, and mutually consistent. It therefore tests whether a policy can re-ground its behavior in the current scene rather than rely on visual or action patterns memorized from the training distribution.

1) Task-Relevant Distractors: A robust manipulation policy should identify the instructed target from current visual evidence rather than rely on superficial similarity or familiar spatial layouts. We introduce distractors that share the target's category, color, shape, or affordance, place a competing object at the target's original position, or obstruct its typical manipulation trajectory respectively. These conditions probe robustness to appearance-based confusion as well as reliance on learned spatial and motion priors.

2) Task-Precondition Variation: Manipulation strategies often depend on the current state of the environment. When that state changes, successful execution may require additional or altered actions. We vary two task-relevant preconditions, including receptacle openness and object pose, while keeping the language instruction unchanged. For example, placing an object into a closed receptacle requires first opening it. This scenario tests whether policies adapt their action sequence to the observed scene rather than replay a nominal manipulation routine.

3) Visual State Ambiguity: A related but distinct challenge arises when the physical state is unchanged but its visual evidence becomes difficult to interpret. We therefore preserve the underlying task state while weakening the cues associated with openness and object pose, so that multiple interpretations become visually plausible. We apply the same two attributes as Task-Precondition Variation but render multiple states simultaneously in the observation, such as a bowl appearing both upright and inverted.

4) Target Geometry Variation: Objects within the same semantic category can exhibit substantial geometric variation in the real world, requiring manipulation policies to adjust grasp locations and end-effector poses accordingly. We introduce four target-object deformations, including Taper, Multipeak, Twist, and Asymmetric, each at three severity levels while preserving object category. This scenario evaluates whether policies adapt manipulation behavior to the observed geometry rather than reuse shape-specific grasp strategies learned during training.

IV. Experiments

A. Experimental Setup

Evaluated Models. We evaluate six representative robotic foundation models: three VLAs, including OpenVLA-OFT [18], π_0 [19], and $\pi_{0.5}$ [14], as well as three WAMs, including FastWAM [17], LingBot-VA [16], and LaWAM [20]. All models are trained on the standard LIBERO demonstrations [1]. We use the officially released checkpoints and inference pipelines whenever available. Since LingBot-VA only releases a LIBERO-Long checkpoint, we additionally train it on LIBERO-Object, LIBERO-Spatial, and LIBERO-Goal using the official training recipe and default hyperparameters.

Evaluation Protocol. Experiments cover 40 manipulation tasks from four LIBERO suites: LIBERO-Long, LIBERO-Object, LIBERO-Spatial, and LIBERO-Goal. The 12 challenge categories in LIBERO-VPro yield 96 experimental settings and 3,296 task-condition cases. Each case is evaluated over ten trials with different initial configurations, resulting in approximately 196,000 episodes in total. Unless otherwise specified, perturbations affect only the visual observations.

Metric and Interpretation. We report task Success Rate (SR) to evaluate the performance. For conventional robustness conditions, such as camera staleness or view unavailability, higher SR indicates stronger resilience. However, several scenarios, including object masking, interaction-cue

TABLE II

SUCCESS RATE (SR, %) ACROSS THE ELEVEN LIBERO-VPRO CHALLENGE CATEGORIES SHARED BY ALL SIX MODELS. NORMAL DENOTES CLEAN-CONDITION PERFORMANCE, WHILE GRAY ↓ VALUES REPORT THE ABSOLUTE SR DROP RELATIVE TO NORMAL. CATEGORIES MARKED ↑ ARE ROBUSTNESS INDICATORS, WHERE HIGHER SR INDICATES GREATER RESILIENCE. CATEGORIES MARKED ◇ ARE DIAGNOSTIC INDICATORS, WHERE HIGH SR MAY REFLECT RELIANCE ON NON-VISUAL CUES OR LEARNED PRIORS RATHER THAN GENUINE VISUAL GROUNDING.

Model		Normal	Visual Evidence Degradation					Visual Source Consistency	Task-Relevant Scene Variation				Camera ↑ Staleness	Overall Avg.
			Target Obj. ◇ Mask.	Target Rec. ◇ Mask.	Interact. ◇ Cue Mask.	Temp. Obs. ↑ Corr.	View ↑ Unavail.	Cross-View ◇ Contrad.	Task-Rel. ↑ Distractors	Task Prec. ↑ Var.	Visual State ◇ Ambig.	Target ↑ Geom. Var.		
VLA	OpenVLA-OFT	97.0	91.28 ↓ 5.7	54.34 ↓42.7	16.23 ↓80.8	76.39 ↓20.6	**29.25** ↓67.8	54.47 ↓42.5	65.08 ↓31.9	39.38 ↓57.6	88.22 ↓ 8.8	81.74 ↓15.3	24.02 ↓73.0	56.40 ↓40.6
	π_0	92.5	81.00 ↓11.5	87.50 ↓ 5.0	**71.59** ↓20.9	70.94 ↓21.6	28.19 ↓64.3	51.53 ↓41.0	61.25 ↓31.2	37.13 ↓55.4	70.67 ↓21.8	79.31 ↓13.2	27.00 ↓65.5	60.55 ↓31.9
	$\pi_{0.5}$	97.2	93.41 ↓ 3.8	95.66 ↓ 1.5	47.54 ↓49.7	82.80 ↓14.4	22.72 ↓74.5	59.21 ↓38.0	**70.92** ↓26.3	**45.50** ↓51.7	85.78 ↓11.4	85.35 ↓11.9	20.52 ↓76.7	64.49 ↓32.7
WAM	FastWAM	98.8	93.38 ↓ 5.4	95.09 ↓ 3.7	69.42 ↓29.4	81.34 ↓17.5	10.12 ↓88.7	63.63 ↓35.2	69.50 ↓29.3	43.63 ↓55.2	**94.67** ↓ 4.1	**88.24** ↓10.6	17.00 ↓81.8	66.00 ↓32.8
	LingBot-VA	97.0	84.69 ↓12.3	**96.78** ↓ 0.2	70.43 ↓26.6	67.89 ↓29.1	21.00 ↓76.0	**70.11** ↓26.9	69.75 ↓27.2	**45.50** ↓51.5	93.33 ↓ 3.7	86.39 ↓10.6	**34.50** ↓62.5	**67.31** ↓29.7
	LaWAM	98.5	**93.44** ↓ 5.1	91.06 ↓ 7.4	69.13 ↓29.4	**85.97** ↓12.5	17.97 ↓80.5	58.37 ↓40.1	69.50 ↓29.0	41.38 ↓57.1	91.33 ↓ 7.2	87.75 ↓10.8	16.60 ↓81.9	65.68 ↓32.8

masking, cross-view contradiction, and visual state ambiguity, are primarily diagnostic. High success may indicate robustness, but can also arise from reliance on memorized spatial priors, alternative camera views, or non-visual cues. We therefore interpret these conditions together with their controlled comparisons rather than treating all SR values as uniformly higher-is-better.

B. Overall Results

Table II summarizes performance across the 11 challenge categories shared by all six models. First, visual robustness is highly scenario-dependent. No model consistently dominates across all categories, showing that robustness is multi-dimensional rather than a single capability that scales uniformly with model performance. For example, $\pi_{0.5}$ performs strongly under semantic distractors and task precondition variation but is highly vulnerable to camera staleness and missing views. Second, all models are surprisingly tolerant to removal of object-level evidence like target-object and target-receptacle masking, which often preserve high success rate. This tolerance likely reflects reliance on learned spatial priors and remaining scene cues. In contrast, masking local interaction cues causes substantially larger degradation, indicating that fine manipulation depends more strongly on visual feedback around the robot-object interaction. Third, View Unavailability and Camera Staleness are among the most disruptive conditions, showing that both the availability and temporal freshness of visual observations are critical for closed-loop control. Notably, all three VLAs outperform all three WAMs under View Unavailability, suggesting that the evaluated WAMs are particularly sensitive to changes in their expected multi-view input structure.

C. What Visual Evidence Do Policies Rely On?

Interaction Cues Matter More Than Static Object Appearance. We first examine how performance changes when visual evidence is removed at different spatial scales. Target-object and target-receptacle masking preserve relatively high success for most models, even under severe occlusion. In contrast, Interaction-Cue Masking causes substantially larger degradation. Fig. 3 further decomposes the interaction region. Across all six models, masking the gripper is the most disruptive intervention, whereas masking the target object

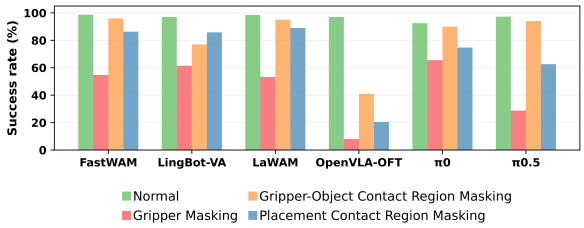

Fig. 3. Success rate (%) under Interaction-Cue Masking.

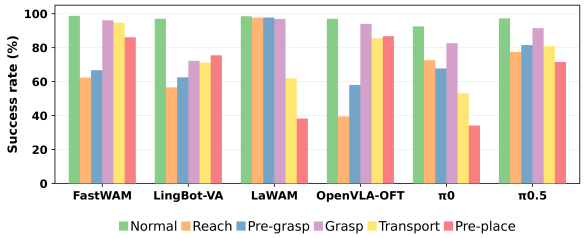

Fig. 4. Success rate (%) under phase-triggered Observation Corruption.

or receptacle is considerably less harmful. This contrast indicates that policies can often compensate for missing object evidence using learned task and spatial priors, but remain strongly dependent on local visuomotor feedback around the end effector.

Sensitivity to Visual Loss Depends on Execution Phase. Transient corruption provides a complementary temporal view of visual dependence. Fig. 4 shows that corruption during the Grasp phase has relatively limited impact, whereas corruption during reaching, transport, and pre-placement produces larger and strongly model-dependent degradation. The temporal profiles also differ across policies. OpenVLA-OFT is more sensitive to early-stage corruption, while π_0 and LaWAM degrade more strongly after grasping. These results show that visual robustness depends not only on how much information is lost, but also on when it becomes unavailable.

D. How Do Policies Use Multi-View Visual Evidence?

Stale and Missing Views Reveal Strong Camera Dependence. Table III reveals substantial asymmetry in how policies use the third-person and wrist cameras. Five of the six models are more sensitive to perturbations of the wrist view,

TABLE III

PER-VIEW SUCCESS RATE (%) UNDER CAMERA STALENESS AND VIEW UNAVAILABILITY. FOR STALENESS, ONLY THE INDICATED CAMERA STREAM IS MADE STALE WHILE THE OTHER REMAINS CURRENT. FOR AVAILABILITY, ONLY THE INDICATED VIEW IS RETAINED AND THE OTHER IS REMOVED.

	Model	Normal	Staleness		Availability	
			Third	Wrist	Third	Wrist
VLA	OpenVLA-OFT	97.0	37.15 ↓59.9	10.90 ↓86.1	15.31 ↓81.7	43.19 ↓53.8
VLA	π_0	92.5	32.85 ↓59.6	21.15 ↓71.3	7.62 ↓84.9	**48.75** ↓43.8
VLA	$\pi_{0.5}$	97.2	21.35 ↓75.8	19.70 ↓77.5	2.19 ↓95.0	43.25 ↓54.0
WAM	FastWAM	98.8	21.65 ↓77.2	12.35 ↓86.5	0.56 ↓98.2	19.69 ↓79.1
WAM	LingBot-VA	97.0	**55.75** ↓41.2	13.25 ↓83.8	1.38 ↓95.6	40.62 ↓56.4
WAM	LaWAM	98.5	9.95 ↓88.5	**23.25** ↓75.2	**35.00** ↓63.5	0.94 ↓97.6

TABLE IV

SUCCESS RATE (%) UNDER CROSS-VIEW CONTRADICTION.

	Model	Normal	Quantity	Existence	Position	State
VLA	OpenVLA-OFT	97.0	76.7 ↓20.3	53.3 ↓43.7	46.5 ↓50.5	66.4 ↓30.6
VLA	π_0	92.5	69.6 ↓22.9	47.1 ↓45.4	47.2 ↓45.3	67.9 ↓24.6
VLA	$\pi_{0.5}$	97.2	79.2 ↓18.0	56.8 ↓40.4	52.5 ↓44.7	74.3 ↓22.9
WAM	FastWAM	98.8	79.6 ↓19.2	59.6 ↓39.2	63.0 ↓35.8	61.4 ↓37.4
WAM	LingBot-VA	97.0	**85.0** ↓12.0	**68.8** ↓28.2	**65.8** ↓31.2	**75.0** ↓22.0
WAM	LaWAM	98.5	81.7 ↓16.8	61.3 ↓37.2	46.3 ↓52.2	67.9 ↓30.6

whereas LaWAM shows stronger dependence on the third-person camera. Camera Staleness and View Unavailability expose complementary vulnerabilities. Staleness provides visually valid but outdated observations, whereas view loss removes a source entirely. Strong degradation under both conditions indicates that models rely heavily on the expected temporal and multi-view structure of their inputs rather than flexibly reallocating attention to the remaining evidence.

Conflicting Views Are Hardest When They Disagree Spatially. We next examine Cross-View Contradiction, where temporally aligned views provide incompatible evidence about object quantity, existence, position, or state. Table IV shows that quantity contradiction is generally the least disruptive, while existence and position contradictions cause substantially larger degradation. Importantly, high success under contradiction does not necessarily imply successful conflict resolution. Quantity or state inconsistencies may preserve the target location, allowing policies to execute the task without explicitly resolving the disagreement. By contrast, position contradiction provides competing positive evidence at different locations, directly interfering with the spatial representation needed for action generation.

E. Can Policies Re-Ground When the Scene Changes?

Spatial Distractors Expose Reliance on Learned Positional Priors. Table V lists the results of six distractors in the scene. Distractors sharing the target's category, color, shape, or affordance cause relatively limited degradation. In contrast, placing a distractor at the target's familiar position causes success to collapse across models, while trajectory obstruction causes the second-largest degradation. This pattern indicates that semantic similarity alone is not the dominant challenge. Instead, policies rely strongly on learned spatial priors. When a competitor occupies the target's familiar location,

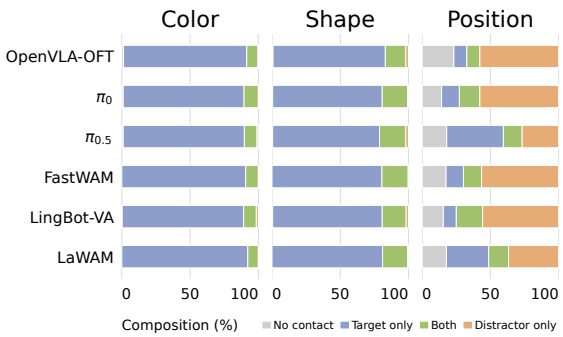

Fig. 5. Robot-object contact composition under three Task-Relevant distractors. Each bar partitions trials by whether the gripper contacts the target, the distractor, both, or neither.

TABLE V

SUCCESS RATE (%) UNDER TASK-RELEVANT DISTRACTORS. EACH COLUMN CORRESPONDS TO A DIFFERENT DISTRACTOR TYPE TARGETING THE TARGET OBJECT'S CATEGORY, COLOR, SHAPE, AFFORDANCE, SPATIAL POSITION, OR MANIPULATION TRAJECTORY.

	Model	Normal	Category	Color	Shape	Afford.	Position	Trajectory
VLA	OpenVLA-OFT	97.0	80.00 ↓17.0	92.75 ↓ 4.2	88.50 ↓ 8.5	93.25 ↓ 3.8	4.25 ↓92.8	31.75 ↓65.2
VLA	π_0	92.5	80.25 ↓12.2	86.25 ↓ 6.2	84.25 ↓ 8.2	89.00 ↓ 3.5	2.50 ↓90.0	25.25 ↓67.2
VLA	$\pi_{0.5}$	97.2	88.25 ↓ 9.0	93.00 ↓ 4.2	91.75 ↓ 5.5	**94.25** ↓ 3.0	**22.25** ↓75.0	36.00 ↓61.2
WAM	FastWAM	98.8	**92.25** ↓ 6.5	**94.50** ↓ 4.3	**92.75** ↓ 6.0	**94.25** ↓ 4.5	4.75 ↓94.0	38.50 ↓60.3
WAM	LingBot-VA	97.0	**92.25** ↓ 4.8	91.50 ↓ 5.5	92.50 ↓ 4.5	86.75 ↓10.2	8.00 ↓89.0	**47.50** ↓49.5
WAM	LaWAM	98.5	88.75 ↓ 9.8	94.00 ↓ 4.5	90.75 ↓ 7.8	**94.25** ↓ 4.2	14.50 ↓84.0	34.75 ↓63.8

models frequently act toward the memorized position rather than re-grounding the target from the current observation. Fig. 5 shows the corresponding behavioral shift, with position displacement substantially increasing distractor-only contact, whereas semantic distractors (e.g., color and shape) more often preserve target-directed interaction.

Task-Precondition Variation Is Harder Than Visual Ambiguity. Task-Precondition Variation and Visual State Ambiguity provide a controlled comparison over the same two attributes: receptacle openness and object pose. The former changes the physical scene, whereas the latter preserves the underlying state but weakens its visual evidence. Table VI shows that actual precondition changes cause severe degradation, while visual ambiguity preserves substantially higher performance for both openness and object pose. The results suggest that the primary difficulty is not simply uncertainty in visual interpretation, but adapting the manipulation strategy when the scene requires a different action sequence. For example, placing an object into a closed receptacle requires an additional opening action before the nominal placement behavior. Current policies often fail to make such scene-conditioned adjustments, revealing limited behavioral adaptation beyond familiar training trajectories.

Geometry Variation Is Comparatively Well Tolerated. Table VII evaluates four target-geometry transformations. Twist consistently produces the largest degradation, while asymmetric deformation is generally the least disruptive. Nevertheless, performance remains relatively strong compared with positional distractors and task-precondition changes. The results show that current robotic foundation models

TABLE VI

SUCCESS RATE (%) COMPARING TASK-PRECONDITION VARIATION WITH VISUAL STATE AMBIGUITY OVER THE SHARED ATTRIBUTES OF RECEPTACLE OPENNESS AND OBJECT POSE.

	Model	Normal	Task-Precondition Variation		Visual State Ambiguity	
			Openness	Object Pose	Openness	Object Pose
VLA	OpenVLA-OFT	97.0	58.6 ↓38.4	25.3 ↓71.7	95.7 ↓ 1.3	86.8 ↓10.2
	π_0	92.5	51.9 ↓40.6	24.7 ↓67.8	81.4 ↓11.1	68.7 ↓23.8
	$\pi_{0.5}$	97.2	64.8 ↓32.4	**30.3** ↓66.9	97.1 ↓ 0.1	83.7 ↓13.5
WAM	FastWAM	98.8	59.5 ↓39.3	29.7 ↓69.1	**100.0**	**93.7** ↓ 5.1
	LingBot-VA	97.0	**67.6** ↓29.4	29.3 ↓67.7	94.3 ↓ 2.7	93.2 ↓ 3.8
	LaWAM	98.5	61.9 ↓36.6	26.6 ↓71.9	97.1 ↓ 1.4	90.3 ↓ 8.2

TABLE VII

SUCCESS RATE (%) UNDER TARGET GEOMETRY VARIATION.

	Model	Normal	Taper	Multipeak	Twist	Asymmetric	Overall Avg.
VLA	OpenVLA-OFT	97.2	80.6 ↓16.6	79.6 ↓17.6	75.9 ↓21.3	85.2 ↓12.0	80.3 ↓16.9
	π_0	93.1	83.3 ↓ 9.8	84.3 ↓ 8.8	70.4 ↓22.7	81.5 ↓11.6	79.9 ↓13.2
	$\pi_{0.5}$	97.5	90.7 ↓ 6.8	88.0 ↓ 9.5	79.6 ↓17.9	88.0 ↓ 9.5	86.6 ↓10.9
WAM	FastWAM	98.6	**89.5** ↓ 9.1	**89.7** ↓ 8.9	82.1 ↓16.5	**91.6** ↓ 7.0	**88.2** ↓10.4
	LingBot-VA	96.7	86.5 ↓10.2	85.9 ↓10.8	**85.5** ↓11.2	87.7 ↓ 9.0	86.4 ↓10.3
	LaWAM	98.6	89.5 ↓ 9.1	89.3 ↓ 9.3	82.9 ↓15.7	89.4 ↓ 9.2	87.8 ↓10.8

generalize reasonably well to moderate geometric variation, while remaining more vulnerable to changes that invalidate learned spatial priors.

F. Can a WAM Rely on Its Own Predicted Futures?

The previous experiments evaluate challenges shared by VLAs and WAMs. We finally investigate a capability specific to WAMs that explicitly generate future observations: whether their predicted futures remain sufficiently grounded to support closed-loop control. Starting from a real observation, we recursively replace subsequent visual inputs with the model's own predictions. Among the evaluated models, this experiment applies to LingBot-VA only.

As shown in Table VIII, recursive self-prediction causes substantial degradation across all four LIBERO suites, reducing overall success from 97.0% to 53.8%. These results show that predicted futures can retain sufficient task-relevant information for partial closed-loop execution, but prediction errors accumulate when generated observations are recursively fed back into the policy. The particularly large degradation on LIBERO-Long is consistent with the greater difficulty of maintaining reliable visual dynamics over longer manipulation sequences.

G. Real-World Evaluation

Experimental Setup. We further examine whether the failure modes identified in simulation transfer to physical manipulation. Experiments are conducted on a Franka Research 3 equipped with third-person and wrist RGBD cameras, both of which are Intel RealSense 435IF. We evaluate $\pi_{0.5}$ and LaWAM as representative VLA and WAM policies on two tasks with different temporal structures. The *Pick-and-Place* task requires a sequence of grasping, transport, and placement actions, whereas the *Move* task relocates a target object through fewer manipulation stages. For each task, we evaluate the Normal condition and four perturbations

TABLE VIII

SUITE-WISE SUCCESS RATE (%) OF LINGBOT-VA UNDER NORMAL OBSERVATION AND RECURSIVE PREDICTED-FRAME FEEDBACK.

Condition	Long	Goal	Object	Spatial	Overall Avg.
Normal	98.0	97.0	98.0	95.0	97.0
Recursive Self-Pred.	22.0 ↓76.0	81.0 ↓16.0	52.0 ↓46.0	60.0 ↓35.0	53.8 ↓43.2

TABLE IX

SUCCESS RATE (%) IN REAL-WORLD EXPERIMENTS UNDER NORMAL AND VISUALLY PERTURBED CONDITIONS.

Task	Model	Normal	View Unavail. (Third)	Cross-View Contrad. (Third)	Same-Pos. Distractor	Camera Staleness (Third)	Overall Avg.
Pick-and-Place	$\pi_{0.5}$	100	0 ↓ 100	30 ↓ 70	10 ↓ 90	20 ↓ 80	15.0 ↓85.0
	LaWAM	80	0 ↓ 80	0 ↓ 80	0 ↓ 80	30 ↓ 50	7.5 ↓72.5
Move	$\pi_{0.5}$	100	10 ↓ 90	0 ↓ 100	0 ↓ 100	100 ↓ 0	27.5 ↓72.5
	LaWAM	80	0 ↓ 80	0 ↓ 80	0 ↓ 80	40 ↓ 40	10.0 ↓70.0

representative of LIBERO-VPro's four dimensions: *View Unavailability*, *Cross-View Contradiction*, *Same-Position Distractor*, and *Camera Staleness*. For *View Unavailability*, the third-person observation is replaced with a black frame. For *Cross-View Contradiction*, the target is hidden from the third-person view while remaining visible in the wrist view. For *Same-Position Distractor*, the target object is moved nearby and a distractor is placed at its original position. For *Camera Staleness*, the third-person stream is delayed by one replanning cycle while the wrist observation remains current. The task instruction remains unchanged across conditions. Each model is evaluated for ten trials under every task-condition pair, yielding 200 physical rollouts in total.

Results. As shown in Table IX, both policies perform well under Normal observations but degrade substantially under most visual perturbations. View Unavailability causes near-total failure, confirming the strong dependence on complete multi-view observations observed in simulation. Cross-View Contradiction is similarly disruptive: even when the target remains visible from the wrist camera, inconsistent evidence across views substantially reduces task success. The Same-Position Distractor also causes severe degradation, supporting the simulation finding that policies often rely on familiar spatial layouts rather than consistently re-grounding the target from current visual evidence. Nevertheless, Camera Staleness exhibits stronger task dependence. For Pick-and-Place, delaying the third-person stream sharply reduces the performance of both models, consistent with the need for temporally aligned feedback across multiple manipulation stages. In contrast, the shorter Move task is considerably more tolerant: $\pi_{0.5}$ retains its Normal performance at 100%, whereas LaWAM drops from 80% to 40%. This contrast indicates that the effect of stale observations depends not only on delay magnitude, but also on the temporal structure of the task and the policy's reliance on each camera stream.

Fig. 6 illustrates representative behaviors. Under the Same-Position Distractor, $\pi_{0.5}$ successfully re-localizes the displaced target in one rollout, whereas LaWAM misses the object. Under Third-Person Camera Staleness, $\pi_{0.5}$ completes the Move task despite the delayed observation, while

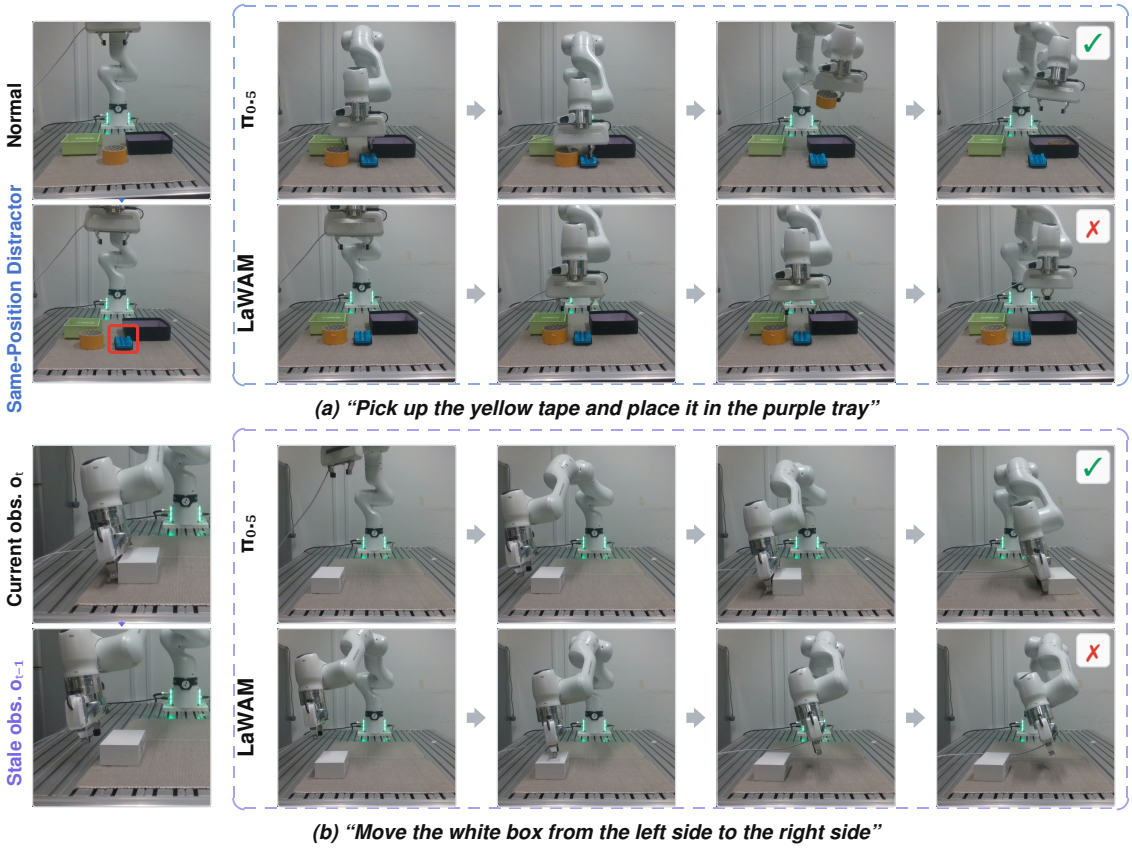

Fig. 6. Representative real-world rollouts under two LIBERO-VPro perturbations. (a) Same-Position Distractor on the Pick-and-Place task where the target object is moved nearby while a distractor occupies its original location. (b) Third-Person Camera Staleness on the Move task where the third-person observation is delayed by one replanning cycle while the wrist view remains current. The red box marks the distractor. ✓ = success, ✗ = failure.

LaWAM follows a pushing trajectory without establishing contact. Overall, these physical experiments are consistent with the main simulation trends, particularly the sensitivity to missing views, cross-view inconsistency, and violations of learned spatial priors, while also showing that temporal robustness can depend on task structure and model design.

V. Conclusion

We have presented LIBERO-VPro, a benchmark for evaluating the closed-loop visual robustness of robotic foundation models. Our results reveal substantial reliance on learned visuospatial priors and vulnerability to disrupted interaction cues, stale observations, and scene changes requiring behavioral adaptation. VLAs and WAMs further exhibit distinct robustness profiles, showing that visual robustness is multi-dimensional and architecture-dependent. We hope LIBERO-VPro provides a systematic diagnostic platform for developing more robust and reliable robotic foundation models.

References

[1] B. Liu *et al.*, "LIBERO: Benchmarking knowledge transfer for lifelong robot learning," *arXiv:2306.03310*, 2023.

[2] T. Chen *et al.*, "RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv:2506.18088*, 2025.

[3] S. Nasiriany *et al.*, "RoboCasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv:2406.02523*, 2024.

[4] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The Colosseum: A benchmark for evaluating generalization for robotic manipulation," *arXiv:2402.08191*, 2024.

[5] Z. Wang, Z. Zhou, J. Song, Y. Huang, Z. Shu, and L. Ma, "VLATest: Testing and evaluating vision-language-action models for robotic manipulation," *Proceedings of the ACM on Software Engineering*, 2025.

[6] J. Guo *et al.*, "On robustness of vision-language-action model against multi-modal perturbations," in *International Conference on Learning Representations*, 2026.

[7] X. Zhou *et al.*, "LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization," *arXiv:2510.03827*, 2025.

[8] S. Fei *et al.*, "LIBERO-Plus: A progressive robustness benchmark for visual-language-action models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[9] B. Zhang *et al.*, "VLA-Arena: An open-source framework for benchmarking vision-language-action models," in *Forty-third International Conference on Machine Learning*, 2026.

[10] H. Li, J. Wang, Z. Mei, A. Majumdar, J. Chen, and B. Zhu, "Robotrustbench: Benchmarking the trustworthiness of video world models for robotic manipulation," in *Findings of the Association for Computational Linguistics: EMNLP*, 2026.

[11] A. Brohan *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv:2212.06817*, 2022.

[12] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv:2307.15818*, 2023.

[13] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv:2406.09246*, 2024.

[14] Physical Intelligence *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," *arXiv:2504.16054*, 2025.

[15] S. Ye *et al.*, "World action models are zero-shot policies," *arXiv:2602.15922*, 2026.

[16] L. Li *et al.*, "Causal world modeling for robot control," *arXiv:2601.21998*, 2026.
[17] T. Yuan, Z. Dong, Y. Liu, and H. Zhao, "Fast-WAM: Do world action models need test-time future imagination?" *arXiv:2603.16666*, 2026.
[18] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv:2502.19645*, 2025.
[19] K. Black *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv:2410.24164*, 2024.
[20] J. Chen *et al.*, "LaWAM: Latent world action models for efficient dynamics-aware robot policies," *arXiv:2606.15768*, 2026.
[21] D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," in *International Conference on Machine Learning*, 2023.
[22] C.-L. Cheang *et al.*, "GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv:2410.06158*, 2024.
[23] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark & learning environment," *arXiv:1909.12271*, 2019.
[24] J. Gu *et al.*, "ManiSkill2: A unified benchmark for generalizable manipulation skills," in *International Conference on Learning Representations*, 2023.
[25] Y. Zhu *et al.*, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv:2009.12293*, 2020.
[26] T. Yu *et al.*, "Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on Robot Learning*, 2020.
[27] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.
[28] A. Mandlekar *et al.*, "MimicGen: A data generation system for scalable robot learning using human demonstrations," in *Conference on Robot Learning*, 2023.
[29] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "GenAug: Retargeting behaviors to unseen situations via generative augmentation," in *Robotics: Science and Systems*, 2023.
[30] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.