# IN-CONTEXT WORLD MODELING FOR ROBOT CONTROL

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

## ABSTRACT

Modern Vision-Language-Action (VLA) models often fail to generalize to novel setups, such as altered camera viewpoints or robot morphologies, because they are typically conditioned only on current observations and language instructions. By ignoring the underlying system configuration as a variable, these models implicitly assume a fixed execution context encountered during training, necessitating data-intensive fine-tuning for any new environment. In this work, we introduce In-Context World Modeling (ICWM), a framework that treats system identification as an in-context adaptation problem. ICWM enables robot policies to autonomously infer essential system variables from a short history of self-generated, task-agnostic interactions. Unlike traditional In-Context Learning that uses demonstrations to specify what task to perform, ICWM leverages the context window to understand how the system operates. By processing these interactions before task execution, the model implicitly captures the world dynamics of the current system, enabling adaptation to novel configurations without parameter updates. Extensive experiments in simulation and on real-world robot platforms demonstrate that ICWM significantly outperforms standard VLA baselines on novel camera viewpoints.

## 1 INTRODUCTION

Consider a human operator handed a joystick to control a robot, but with no prior knowledge of the control mapping. Does pushing forward move the robot left, right, or forward? The operator's first instinct is not to attempt the task directly, but to explore: move the joystick randomly, observe the robot's response, and infer the input-output relationship. Within seconds, this brief calibration phase yields an internal model of the system dynamics, enabling the transition from stochastic exploration to goal-directed control. Crucially, this calibration is entirely self-generated—the operator needs no prior knowledge of the task, no external guidance, and no task-specific experience under the new control mapping.

Modern Vision-Language-Action (VLA) models Zitkovich et al. (2023); Kim et al. (2024); Black et al., however, lack this calibration ability entirely. The standard formulation $\pi_\theta(a_t \mid o_t, l)$ conditions only on the current observation and language instruction, treating system configuration $\psi$ (like camera viewpoints, morphology) as a fixed constant absorbed into model parameters during training. When deployment conditions deviate from training, the model has no mechanism to recover the correct action-observation correspondence, and performance degrades Fei et al. (2025). Scene-specific fine-tuning remains the prevailing remedy, yet it requires human intervention for every new setup—contradicting the goal of generalist deployment.

We argue this failure is a *system identification* problem: the policy lacks knowledge of $\psi$ at test time. This naturally motivates repurposing the context window to recover $\psi$ through targeted interaction. However, existing In-Context Learning (ICL) methods Fu et al. (2024); Shah et al. (2025) treat context as *behavior specification* (what to do) rather than *system identification* (how the system operates), and still require human-provided demonstrations at test time.

In this work, we propose In-Context World Modeling (ICWM) (Fig. 1), which repurposes the context window for system identification rather than behavior specification. Before task execution, the robot performs a short sequence of random exploratory movements and records the resulting visual transitions; these self-probed clips are prepended as context, from which the model implicitly recovers



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.

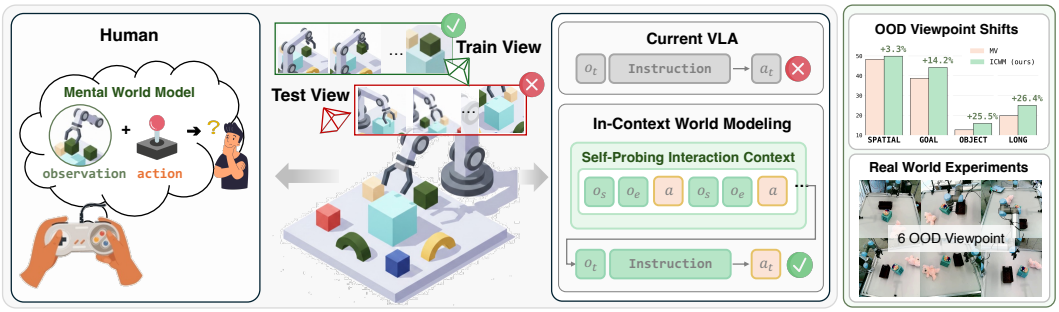



Figure 1: In-Context World Modeling (ICWM). Standard VLA models often fail in novel system configurations due to fixed observation-action assumptions. Similar to how humans explore unfamiliar controls to build a mental world model, ICWM enables robots to autonomously infer system dynamics from self-probing interaction context.

the current system configuration and adapts its actions accordingly. Surprisingly, the task-agnostic random movements alone provide sufficient context for implicit system identification without any task-specific demonstrations and we show this simple strategy yields consistent improvements over standard VLA training on novel viewpoints across both simulation and real-robot platforms.

In summary, this work makes the following contributions:

- We reframe VLA generalization as a test-time system identification problem, identifying the absence of explicit $\psi$ conditioning as an underexplored failure mode that persists across standard training strategies.
- We propose In-Context World Modeling (ICWM), which implicitly identifies the system configuration from self-generated, task-agnostic exploratory transitions prepended as context, achieving test-time adaptation without parameter updates or task-specific demonstrations.
- We validate ICWM on both simulation benchmarks and real-robot experiments, demonstrating significant improvements over standard VLA training on novel camera viewpoints, with further generalization to semantic scene variations and robot morphological changes.

## 2 Related Work

### 2.1 In-Context Adaptation for Robotics

The emergence of large-scale pre-trained models Radford et al. (2019); Touvron et al. (2023); OpenAI (2022) has established In-Context Learning (ICL) as an emergent capability for on-the-fly skill acquisition Brown et al. (2020); Dong et al. (2022). In robotics, ICL has been explored via in-context imitation: models are prompted with expert trajectories, provided by teleoperation or retrieved from offline buffers, to perform next-token prediction over observation-action sequences Fu et al. (2024); An et al. (2025); Sridhar et al. (2025); complementary approaches leverage human play videos for cross-embodiment transfer Shah et al. (2025); Jain et al. (2024). Despite differences in data modality, all such methods treat context as behavior specification ("what to do"), and critically require human-provided demonstrations at test time. A related line of work pursues adaptation through meta-learning. Gradient-based methods Finn et al. (2017); Rajeswaran et al. (2019) meta-optimize an initialization for rapid test-time fine-tuning, while recurrent meta-RL Duan et al. (2016); Zintgraf et al. (2019) encodes interaction history to infer *which task* to perform—relying on reward signals unavailable prior to task execution. In contrast, ICWM uses task-agnostic self-generated interactions to understand *how the system operates*, requiring no demonstrations, no reward signal, and no parameter updates at test time–a harder and more practical deployment setting.

### 2.2 World Modeling for Robotic Control

World modeling captures environment dynamics to ground an agent's actions in causal state transitions Ha & Schmidhuber (2018); LeCun & Courant (2022); Ding et al. (2024). Existing approaches

fall into two paradigms: forward models predict future observations either in pixel space Wu et al. (2023); Li et al. (2024); Zhao et al. (2025); Cen et al. (2025) or latent space Hafner et al. (2019; 2020); Wu et al. (2022); Hafner et al. (2023); Zheng et al. (2025), while inverse dynamics models abduce actions from observed visual changes Du et al. (2023); Jang et al. (2025); Tian et al. (2024); Baker et al. (2022); some works unify both objectives Zhu et al. (2025); Li et al. (2025). All of these introduce dedicated parameters and training objectives for world modeling. ICWM instead realizes world modeling *implicitly*: the model leverages standard sequence modeling to extract the time-invariant causal structure (e.g., control mappings and camera viewpoints) directly from task-agnostic interaction histories. This implicit paradigm introduces zero additional parameters, treating world modeling as an emergent inference capability at test time.

# 3 Preliminary and Motivation

## 3.1 Standard VLA Formulation and Its Limitation

We consider a robot manipulation task where a policy $\pi$ maps multimodal observations and language instructions to actions. Let $\mathcal{O}$ be the observation space and $\mathcal{I}$ be the space of natural language instructions. A standard VLA policy $\pi_\theta(a_t \mid o_t, l)$ processes the current observation $o_t \in \mathcal{O}$ and instruction $l \in \mathcal{I}$ to predict an action $a_t \in \mathcal{A}$. The control task is thus formulated as:

$$\pi_\theta(a_t \mid o_t, l). \tag{1}$$

where the model parameters $\theta$ are optimized on a large-scale dataset $\mathcal{D}$ collected under a set of specific system setups. In practice, $a_t$ is often structured as an action chunk to ensure temporal smoothness and control stability during execution.

This formulation embeds an implicit assumption: the system configuration $\psi$—encompassing camera viewpoints, mounting offsets, and the robot's morphological properties—is fixed and known. An ideal policy should instead condition on $\psi$ explicitly:

$$\pi^*_\theta(a_t \mid o_t, l, \psi). \tag{2}$$

Without $\psi$, training forces the model to marginalize over all configurations in the dataset:

$$\pi_\theta(a_t \mid o_t, l) \approx \int \pi^*_\theta(a_t \mid o_t, l, \psi) p(\psi) d\psi. \tag{3}$$

At deployment, when a specific $\psi'$ is realized, this averaged policy lacks the context to correctly interpret the observation-action correspondence, leading to degraded performance on novel configurations. This motivates explicit recovery of $\psi$ at test time.

## 3.2 Interaction Context Enriches Configuration Information

A single observation often provides insufficient information about the underlying system configuration. We formalize this intuition by considering the information carried by an interaction trajectory. Let $\psi$ denote the task-relevant system configuration, including camera viewpoints, mounting offsets, and robot morphological properties. Given an initial observation $o_0$, a probing sequence $a_{1:t}$ produces additional observations $o_{1:t}$. We define the interaction context as $\mathcal{T}_t = (o_{0:t}, a_{1:t})$.

We make three assumptions. (A1) The initial observation is partially informative, but does not uniquely determine the configuration: $I(\psi; o_0) < H(\psi)$. (A2) The probing actions are exogenous with respect to the unknown configuration: $I(\psi; a_{1:t} \mid o_0) = 0$. (A3) The probing sequence is configuration-sensitive, meaning that the resulting observations contain additional information about the configuration: $I(\psi; o_{1:t} \mid o_0, a_{1:t}) > 0$.

**Proposition 1.** *Under assumptions A1–A3, the interaction context provides strictly more information about the system configuration than the initial observation:*

$$I(\psi; \mathcal{T}_t) > I(\psi; o_0). \tag{4}$$

The proposition formalizes the role of exploratory interaction: actions alone do not reveal the hidden configuration, but their induced visual changes can provide additional evidence about the observation–action correspondence. The result applies only to configuration-sensitive probing and does not claim that arbitrary movements reveal task-relevant system properties.

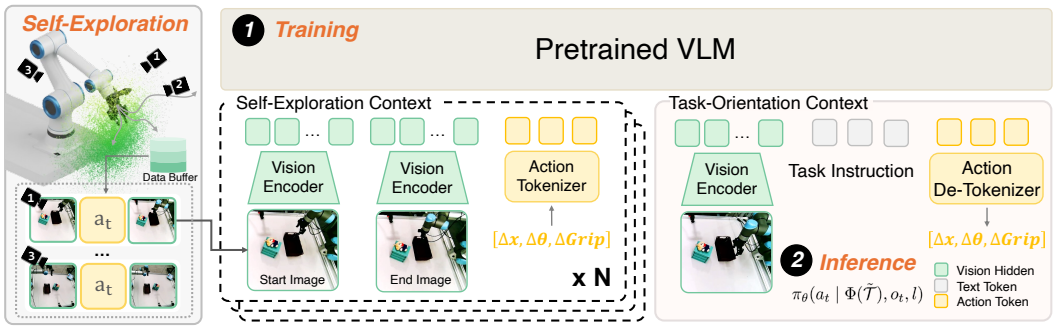



Figure 2: Overview of the ICWM Training and Inference Pipeline. (1) Training: The model is trained on data collected across diverse system configurations, where task-agnostic interaction clips are prepended to each training sample as context. (2) Inference: At test time, the robot first performs stochastic exploration to collect system context, which then guides the policy to generate precise actions via in-context inference.

## 4 In-Context World Modeling

To resolve the information incompleteness identified in Sec. 3, we introduce In-Context World Modeling (ICWM) (Fig. 2). The core philosophy of ICWM is to transform the VLA from a static mapping into an adaptive inference mechanism that recovers the latent configuration $\psi$ through environmental interactions.

Specifically, given a short history of task-agnostic interaction clips $\mathcal{T} = \{(o_i^s, a_i, o_i^e)\}_{i=1}^N$, the configuration can be inferred implicitly with a function $\Psi(\mathcal{T})$. By conditioning the policy on interactions collected under the current configuration $\psi$, the policy is reformulated as:

$$a_t \sim \pi_\theta(a_t \mid o_t, l, \Psi(\mathcal{T})). \tag{5}$$

The $\Psi(\mathcal{T})$ term serves as a representation of the world dynamics under $\psi$, enabling the model to interpret $o_t$ in the context. This interaction-centric conditioning allows the policy to adapt its action selection to the specific visual and physical setup encountered at test time, thereby overcoming the generalization collapse of standard VLA models.

### 4.1 In-Context Training

To implement $\Psi(\mathcal{T})$, we adopt a parameter-efficient design where $\Psi$ shares its parameters with the VLA backbone $\pi_\theta$, motivated by the structural symmetry between action prediction and configuration inference—both require understanding the correspondence between observations and actions.

Each training sample is constructed by prepending $N$ task-agnostic interaction clips to the task query. These clips are randomly sampled from a pool of interaction segments collected across all training trajectories and viewpoints, ensuring diversity in the interaction context. The model is trained on data collected across diverse system configurations, where the interaction context $\mathcal{T}$ naturally varies with the underlying $\psi$. This variation provides an implicit training signal: to accurately predict task actions across configurations, the model must learn to extract and utilize the dynamics information carried in $\mathcal{T}$, implicitly modeling the action-to-observation mapping under each specific system setup. The model is trained with the following loss:

$$\mathcal{L} = -\log \pi_\theta(a_t \mid \Psi(\mathcal{T}), o_t, l), \tag{6}$$

where $\Psi(\mathcal{T})$ denotes the hidden states induced by the interaction context.

### 4.2 Test-Time Active Probing and Inference

At deployment, ICWM enables task-specific demonstration-free adaptation to novel system configurations through a two-phase inference protocol that requires no gradient updates, prior calibration, or knowledge of the target environment.

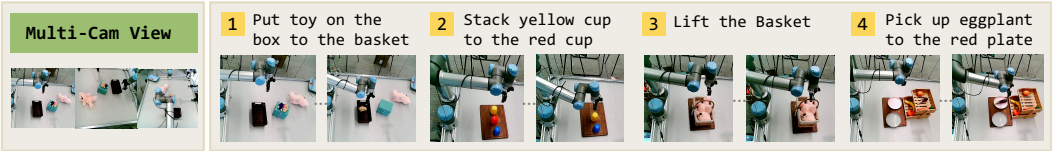



Figure 3: Real-World Task Suite with four manipulation tasks, including stacking, lifting, and pick-and-place, evaluated across six distinct camera viewpoints.

**Active Probing Phase.** Before task execution, the robot collects the interaction context $\mathcal{T} = \{(o_i^s, a_i, o_i^e)\}_{i=1}^N$ by performing $N$ task-agnostic probing actions. For each step $i \in \{1, \ldots, N\}$, a random target pose is sampled within the robot's safe workspace, and the robot executes an action $a_i$ to reach it. The resulting transition $(o_i^s, a_i, o_i^e)$ is recorded. These probing movements are designed to be spatially diverse, sampling multiple directions relative to the end-effector's pose to provide sufficient coverage of the local dynamics manifold. The probing workspace is defined to avoid contact with task-relevant objects, ensuring that the task initial state remains undisturbed throughout this phase.

**In-Context Execution Phase.** Once $\mathcal{T}$ is collected, the configuration inference function $\Psi$ processes the interaction context to implicitly recover the latent system configuration. Conditioned on the inferred hidden representation $\Psi(\mathcal{T})$, the current observation $o_t$, and the language instruction $l$, the policy generates the task action $a_t \sim \pi_\theta(a_t \mid o_t, l, \Psi(\mathcal{T}))$. Since $\Psi$ shares parameters with the VLA backbone, this is implemented as a single forward pass where the Transformer first attends to $\mathcal{T}$, building configuration-aware hidden states, before processing the task query $(o_t, l)$ to produce actions aligned with the deployed physical setup.

## 5 Experiments

In this section, we evaluate the effectiveness of In-Context World Modeling (ICWM) in enabling vision-language-action models to adapt to novel system configurations without parameter updates. We focus on answering three questions:

(1) *Generalization*: Does ICWM improve generalization to unseen viewpoints?

(2) *System Identification*: Does the model truly perform implicit world modeling rather than pattern matching?

(3) *Versatility*: Does ICWM generalize beyond geometric shifts to semantic and morphological perturbations?

### 5.1 Experimental Setup

**Simulated Benchmark.** We build a viewpoint-shifted evaluation protocol on top of LIBERO (Liu et al., 2023), termed **LIBERO-View**, to assess model robustness under camera viewpoint changes. It comprises four task suites (Spatial, Object, Goal, Long) that assess spatial reasoning, object understanding, goal conditioning, and long-horizon execution, and adopts a cross-view protocol: training on 8 azimuthal angles ($\psi \in \{30°, 60°, \ldots, 330°\}$) and evaluating on both in-domain and 6 unseen OOD viewpoints ($\psi' \in \{45°, 135°, \ldots, 315°\}$), yielding $500 \times 15 \times 4$ total episodes. Details are in App. C.

**Real-Robot Setup.** Our platform is a UR5e manipulator with a 12-camera multi-view system, split into 6 training and 6 held-out test viewpoints. We evaluate four manipulation tasks (stacking, lifting, pick-and-place) with 25 trials per task per novel viewpoint (600 total), reporting the average success rate. Details are in App. D.

**Baselines.** We use two controlled comparisons. (1) **Multi-View BC (MV)** follows the standard multi-view behavior cloning paradigm used in current VLA training, sharing the same architecture and multi-view training data with ICWM but omitting the interaction context. (2) **Explicit Config-**

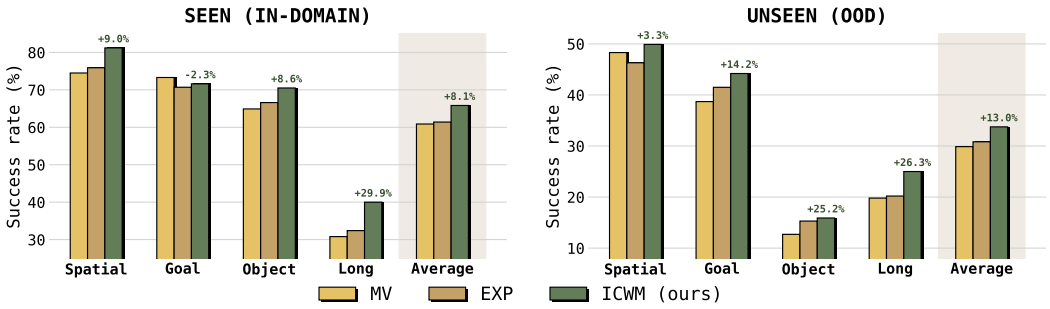



Figure 4: Success Rates (%) on LIBERO-View for Seen and Unseen Viewpoints.

**uration (EXP)** augments MV with the ground-truth camera angle provided as a text input. These baselines isolate the contribution of configuration inference from the interaction prefix.

**Implementation Details.** Our model utilizes Qwen2.5-VL-3B (Bai et al., 2025) as the backbone and FAST (Pertsch et al., 2025) as action tokenizer, with action chunk size 5 and $N{=}5$ context clips. Implemented in PyTorch and trained on 8 NVIDIA A100 GPUs, the system is optimized via AdamW with a weight decay of $10^{-4}$ and a learning rate peaking at $5 \times 10^{-5}$ after a 50k-step warmup, followed by cosine decay. Training clips are trajectory slices bucketed by viewpoint ($\sim$10k each) without any extra collection; each step resamples from the same angle bucket, so each sample sees many prefixes, spanning tasks and without temporal order. The implementation details of self-exploration is detailed in App. E.

### 5.2 Simulation Results

As shown in Fig. 4, ICWM consistently outperforms all baselines in both seen and unseen viewpoints, leading to several key observations: **(1) Viewpoint generalization is a fundamental challenge.** While a seen-to-unseen performance drop is observed globally, ICWM demonstrates significantly stronger resilience, improving the OOD success rate by 13.0% over the Multi-View BC baseline. This confirms that while multi-view training expands spatial data, geometric extrapolation remains an open challenge that standard imitation learning struggles to resolve without test-time adaptation. **(2) Implicit identification outperforms explicit specification.** ICWM improves OOD success rate by 9.5% over Explicit Configuration. Even with ground-truth camera angles, Exp lacks OOD generalization and an understanding of system dynamics, strengths that ICWM gains from interaction context. **(3) ICWM yields the greatest gains on long-horizon tasks.** On LIBERO-Long, ICWM surpasses MV by 29.9% (seen) and 26.3% (unseen), the largest relative margins across all suites. This is because long-horizon tasks amplify small spatial errors from viewpoint shift, causing cascading failures in baselines, while ICWM mitigates this by continuously grounding actions in system dynamics.

### 5.3 Real-world Results

As shown in Fig. 5, we evaluate ICWM on the UR5e under novel camera viewpoints without any parameter updates or task-specific demonstrations. ICWM effectively transfers the viewpoint adaptation capability from simulation to real-world manipulation. The MV baseline achieves 68% success under the training-aligned viewpoint but drops to 17% under viewpoint shifts, whereas ICWM substantially recovers the performance gap by leveraging interaction context for test-time calibration.

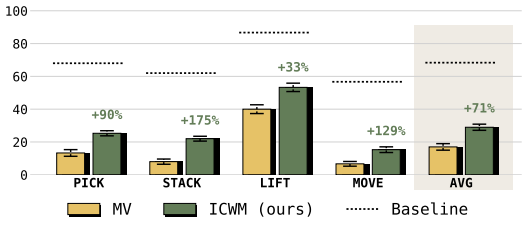



Figure 5: Real-world evaluation on the UR5e platform. ICWM substantially outperforms the Multi-View (MV) baseline under novel viewpoints.

As shown in Fig. 6, viewpoint shifts lead to systematic errors such as end-effector position offsets and premature gripper closure, rather than random execution failures, indicating a mismatch between visual observations and the learned action policy. ICWM resolves these failures through dynamic

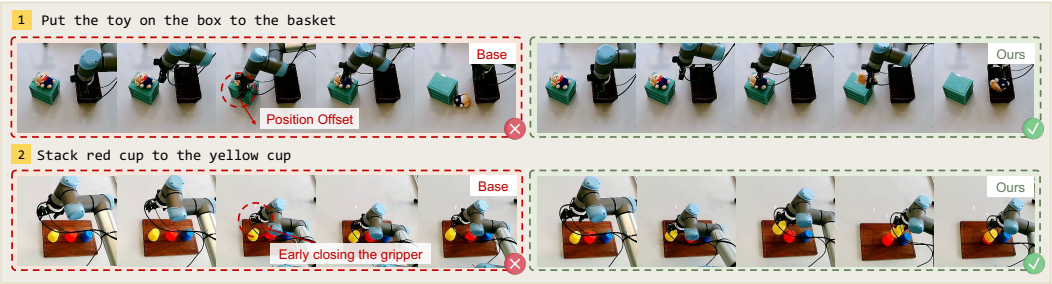



Figure 6: Qualitative Comparison. Without ICWM, standard policies exhibit (a) position offsets or (b) premature gripper closure due to viewpoint shifts; ICWM resolves these ambiguities through dynamic calibration.

Table 1: Ablation on Interaction Context. Removing any context component degrades performance, confirming that joint observation of actions and outcomes is essential for system dynamics.

| | ICWM | w/o act. | w/o img. | w/o ctx. | false ctx. |
|---|---|---|---|---|---|
| 45° | 36.6 | 30.8 ↓ 15.8% | 17.6 ↓ 51.9% | 32.4 ↓ 11.5% | 27.8 ↓ 24.0% |
| 135° | 2.2 | 2.0 ↓ 9.1% | 0.8 ↓ 63.6% | 1.6 ↓ 27.3% | 1.2 ↓ 45.5% |
| 225° | 8.8 | 8.2 ↓ 6.8% | 3.0 ↓ 65.9% | 7.8 ↓ 11.4% | 4.2 ↓ 52.3% |
| 255° | 28.4 | 28.0 ↓ 1.4% | 13.2 ↓ 53.5% | 27.8 ↓ 2.1% | 24.4 ↓ 14.1% |
| 285° | 36.6 | 25.6 ↓ 30.1% | 14.6 ↓ 60.1% | 27.4 ↓ 25.1% | 23.6 ↓ 35.5% |
| 315° | 37.6 | 35.0 ↓ 6.9% | 16.2 ↓ 56.9% | 35.0 ↓ 6.9% | 32.4 ↓ 13.8% |
| avg | 25.0 | 21.6 ↓ 13.6% | 10.9 ↓ 56.4% | 22.0 ↓ 12.0% | 18.9 ↓ 24.4% |

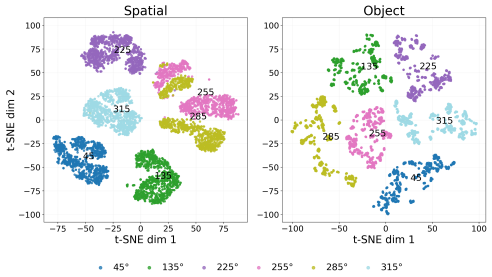



Figure 7: t-SNE of $\Psi(\mathcal{T})$ across OOD viewpoints (perplexity=30).

context-based calibration. By inferring the current configuration from robot-environment interactions, ICWM maintains accurate object localization and grasp execution across different viewpoints, achieving reliable manipulation without additional training.

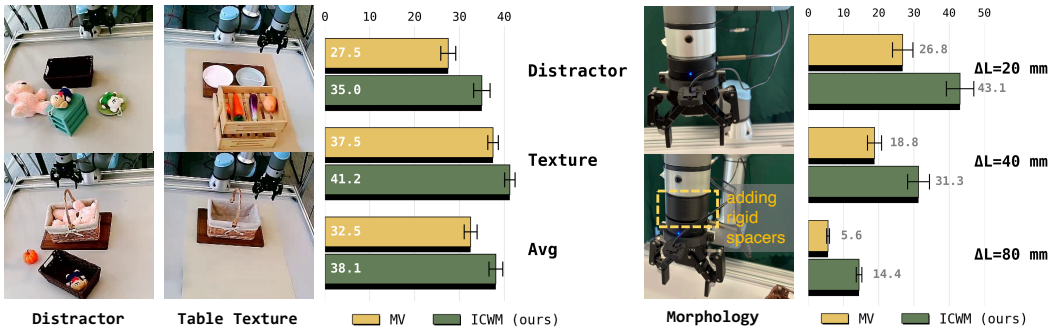
## 6 Analysis

### 6.1 What Does the Interaction Context Contribute?

**Ablation on context components.** We ablate interaction context across five settings: (1) full-context (ICWM), (2) *w/o actions* (action tokens omitted), (3) *w/o images* (image tokens omitted), (4) *w/o context* (all interaction tokens excluded), and (5) *false context* (clips from a 180°-offset viewpoint). The asymmetric drops in Tab. 1 reveal several key findings: *w/o images* causes the largest collapse (avg. $-56.4\%$)—without visual outcomes, the model mimics exploratory actions as task demonstrations, actively hurting performance; *w/o actions* degrades more moderately, confirming that visual flow provides a coarse spatial anchor but full calibration requires the paired $(o_i^s, a_i, o_i^e)$ tuple. Critically, *false context* performs worse than no context at all (18.9 vs. 22.0), indicating that misaligned context actively misleads the policy's world model rather than being passively ignored. This negative transfer, symmetric in magnitude to the gains from correct context ($+13.6\%$), confirms that the model genuinely conditions on context content for configuration inference.

**Is in-context capability an emergent property of specialized training?** To determine whether in-context adaptation arises from standard sequence modeling, we evaluate a BC policy trained without in-context supervision under the same interaction context. Performance collapses to near-zero ($< 1\%$) when interaction tokens are prepended, confirming that this capability must be explicitly incentivized during training rather than emerging naturally from imitation learning.

**Are the Learned Implicit Representations Identifiable?** We examine whether $\Psi(\mathcal{T})$ forms a structured representation with respect to system configuration. We visualize the hidden representations $\Psi(\mathcal{T})$ via t-SNE across the six OOD viewpoints (1,024 points per viewpoint). As shown in Fig. 7, the representations exhibit tight within-viewpoint clustering, demonstrating stability, and clear between-viewpoint separation, demonstrating identifiability.




(a) Semantic perturbations

(b) Morphological generalization

Figure 9: Robustness to semantic scene variations and robot morphological changes. ICWM maintains a performance margin over the baseline when facing novel out-of-distribution perturbations.

### 6.2 How Many Interaction Clips Are Needed?

While additional interaction clips provide more evidence about the relationship between viewpoints and robot motion, they also increase the context length processed by the policy. We therefore study the effect of the number of interaction clips available at inference time by varying $N \in \{1, 3, 5, 8\}$. As shown in Fig. 8, success improves consistently as more interaction clips are provided across the evaluated viewpoints, suggesting that multiple transitions offer complementary information for inferring the current visuomotor configuration. However, the marginal improvement becomes smaller with longer contexts. Based on this accuracy–efficiency trade-off, we use $N = 5$ for the main experiments.

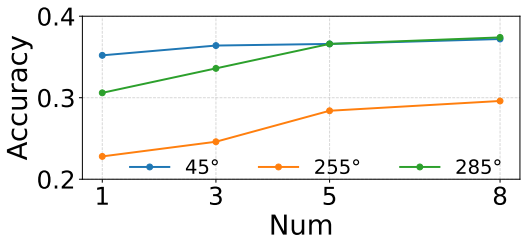



Figure 8: LIBERO-Long success as a function of the number of interaction clips.

### 6.3 What Probing Strategy is Effective?

A practical question for deploying ICWM is whether the choice of probing strategy matters. We evaluate four strategies: (a) *Random*, which samples target poses uniformly across all spatial directions; (b) *XY-only*, which restricts movements to the horizontal plane; (c) *Z-only*, which moves exclusively along the vertical axis; and (d) *R-only*, which varies only the end-effector orientation. As shown in Tab. 2, all four strategies consistently outperform the multi-view baseline by 15–27%, confirming that the benefit of ICWM stems from the interaction format itself rather than any particular movement pattern. Performance differences across strategies suggest that different axes expose different aspects of the local dynamics manifold, with no single strategy dominating across all viewpoints. Future work could explore adaptive probing policies that select interaction contexts to maximize task-relevant information gain.

Table 2: Success rates (%) under different probing strategies on OOD viewpoints.

| | MV | Random | R-only | Z-only | XY-only |
|---|---|---|---|---|---|
| 45° | 30.4 | **36.6** | 35.8 | 34.6 | 34.6 |
| 135° | 1.2 | 2.2 | 1.0 | **3.0** | 1.0 |
| 225° | 2.8 | 8.8 | 6.0 | 13.0 | **14.8** |
| 255° | 24.0 | 28.4 | 30.4 | 25.2 | **30.6** |
| 285° | 32.8 | 36.6 | **38.8** | 31.2 | 33.6 |
| 315° | 27.6 | **37.6** | 29.2 | 30.0 | 35.0 |
| Avg | 19.8 | **25.0** | 23.4 | 22.8 | 24.9 |

### 6.4 Does ICWM Generalize Beyond Camera Shifts?

To evaluate whether ICWM extends beyond viewpoint changes, we study generalization across three increasingly challenging settings. First, we test zero-shot robustness to semantic scene variations, where the trained policy encounters unseen visual appearances. Second, we evaluate zero-shot embodiment generalization by modifying the robot morphology at test time without additional training.

Finally, we examine whether ICWM can interpolate across unseen kinematic configurations when trained with boundary morphology variations. Experimental details are provided in Sec. D.2.

**Semantic perturbations generalization.** We evaluate the same trained model under unseen scene variations, including distractor objects and novel table textures. As shown in (Fig. 9a), ICWM maintains a consistent margin over MV under both distractor objects (35.0 vs. 27.5) and novel table textures (41.2 vs. 37.5). The more moderate gains relative to viewpoint generalization likely reflect the scarcity of diverse scene-configuration data in current datasets rather than a fundamental limitation of the mechanism.

**Morphological generalization.** We next evaluate embodiment changes without retraining. Attaching rigid spacers ($\Delta L \in \{20, 40, 80\}$ mm) to the gripper flange alters forward kinematics at test time. ICWM consistently outperforms MV across all offsets with a stable margin of +60% (Fig. 9b); at $\Delta L = 80$ mm, where MV fails largely (5.6), ICWM retains non-trivial success (14.4) by inferring the altered kinematics from probing context.

**Interpolation across unseen morphologies.** We further validate this on a WindowX platform with systematically varied link lengths ($\{100\%, 90\%, 80\%, 70\%\}$ of the original), training on two boundary configurations (100% and 70% link length) and evaluating zero-shot on two interpolated OOD configurations (90% and 80%). As the link-length offset increases from 10% to 20%, MV's average success rate collapses by more than half (57% $\rightarrow$ 28%), while ICWM degrades far more gracefully (77% $\rightarrow$ 62%), widening its margin over MV from 20 to 34 points (Fig. 10). This mirrors the pattern observed with spacer attachments: as kinematic uncertainty grows, ICWM's advantage over MV grows with it, consistent with the benefit stemming from explicit kinematic inference via the probing context.

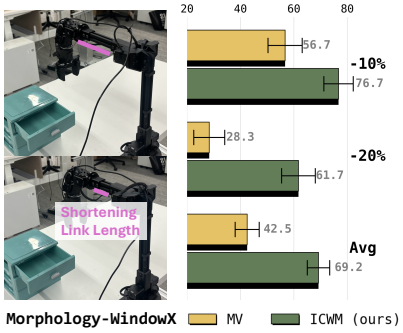



Figure 10: Morphological generalization on WindowX.

### 6.5 Inference Latency Analysis

We evaluate the computational overhead of ICWM on a single NVIDIA RTX 4090 GPU. The extra latency is mainly due to the $2N$ image and $N$ action tokens. While the baseline VLA requires 0.112s per inference step, our ICWM-enhanced model with $N = 3$ and $N = 5$ context clips incurs a latency of 0.165s and 0.185s respectively, without compromising control loop stability (Fig. 11). Furthermore, since the interaction context $\tilde{\mathcal{T}}_\psi$ is static under a fixed configuration, its hidden states can be precomputed and reused via KV caching, effectively reducing per-step inference cost back to near-baseline levels.

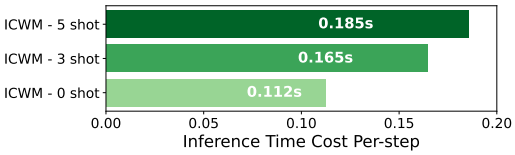



Figure 11: Comparison of inference time across different setting.

## 7 Conclusion

In this work, we identified the absence of explicit system configuration conditioning in modern VLA models as an underexplored limitation to generalization. Inspired by human motor adaptation, we proposed In-Context World Modeling (ICWM), which transforms the VLA policy from a static mapping into an adaptive inference engine. By conditioning the policy on self-generated forward interaction clips, ICWM implicitly captures the underlying sensory-motor relationship at test time without any parameter updates. Extensive experiments on simulation and real-world platforms show that ICWM improves viewpoint generalization, especially on long-horizon tasks. Ablations and representation analyses confirm that ICWM exploits action–observation correspondence rather than static visual cues. Evaluations under semantic and morphological shifts further demonstrate its ability to adapt VLA policies beyond camera changes.

### AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript, as well as for assisting the generation of code used for visualization and statistical plotting. We have not used generative AI tools for developing the core research ideas, designing the methodology, conducting experiments, analyzing results, or generating scientific claims. Other recommended disclosure tasks are not applicable to this work.

All AI-assisted outputs were reviewed and verified by the authors. In particular, AI-assisted text revisions were manually checked for correctness, clarity, and consistency with the authors' intended meaning, and AI-generated code for visualization was tested and validated by the authors before use. The authors take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work studies algorithmic improvements for vision-language-action models in robotic manipulation. Our experiments are conducted in controlled simulation environments and laboratory settings. We do not collect human data, process personal information, or deploy the system in safety-critical applications. We acknowledge that future deployment of autonomous robotic systems requires careful consideration of safety, reliability, and human-robot interaction.

### REPRODUCIBILITY STATEMENT

We have made efforts to facilitate reproducibility of our work. The main paper describes the proposed In-Context World Modeling (ICWM) framework and the overall experimental methodology. The appendix provides additional details on the model architecture, training procedure, implementation choices, evaluation protocols, and ablation settings. We also document the simulation environments, benchmark tasks, and data processing procedures used in our experiments. The source code, pretrained models, and evaluation scripts will be made publicly available upon publication to support independent verification and future research.

## REFERENCES

Vuong Dinh An, Minh Nhat Vu, Dong An, and Ian D. Reid. Action tokenizer matters in in-context imitation learning. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13490–13496, 2025. URL https://api.semanticscholar.org/CorpusID:276742267.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL https://api.semanticscholar.org/CorpusID:276449796.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *ArXiv*, abs/2206.11795, 2022. URL https://api.semanticscholar.org/CorpusID:249953673.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler,

Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

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*, abs/2506.21539, 2025. URL https://api.semanticscholar.org/CorpusID:280010695.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, J. Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58:1 – 38, 2024. URL https://api.semanticscholar.org/CorpusID:274192171.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:255372865.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *ArXiv*, abs/2302.00111, 2023. URL https://api.semanticscholar.org/CorpusID:256459809.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. $Rl^2$: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Senyu Fei, Siyin Wang, Junhao Shi, Z. G. 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*, abs/2510.13626, 2025. URL https://api.semanticscholar.org/CorpusID:282102298.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Letian Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, Will Panitch, Fangchen Liu, Hui Li, and Kenneth Y. Goldberg. Icrt: In-context imitation learning via next-token prediction. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5937–5944, 2024. URL https://api.semanticscholar.org/CorpusID:271974730.

David R Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:52171619.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ArXiv*, abs/1912.01603, 2019. URL https://api.semanticscholar.org/CorpusID:208547755.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *ArXiv*, abs/2010.02193, 2020. URL https://api.semanticscholar.org/CorpusID:222133157.

Danijar Hafner, J. Pašukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse domains through world models. *ArXiv*, abs/2301.04104, 2023. URL https://api.semanticscholar.org/CorpusID:255569874.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment, 2021. URL https://arxiv.org/abs/2007.04309.

Vidhi Jain, Maria Attarian, Nikhil J. Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R. Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *ArXiv*, abs/2403.12943, 2024. URL https://api.semanticscholar.org/CorpusID:268532100.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xi Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke S. Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models. 2025. URL https://api.semanticscholar.org/CorpusID:279447300.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *ArXiv*, abs/2406.09246, 2024. URL https://api.semanticscholar.org/CorpusID:270440391.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots, 2021. URL https://arxiv.org/abs/2107.04034.

Yann LeCun and Courant. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. 2022. URL https://api.semanticscholar.org/CorpusID:251881108.

Peiyan Li, Hongtao Wu, Yan Huang, Chi-Hou Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, 10:1912–1919, 2024. URL https://api.semanticscholar.org/CorpusID:271957548.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *ArXiv*, abs/2503.00200, 2025. URL https://api.semanticscholar.org/CorpusID:276741531.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: efficient action tokenization for vision-language-action models. *CoRR*, abs/2501.09747, 2025. doi: 10.48550/ARXIV.2501.09747. URL https://doi.org/10.48550/arXiv.2501.09747.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Rutav Shah, Shuijing Liu, Qi Wang, Zhenyu Jiang, Sateesh Kumar, Mingyo Seo, Roberto Mart'in-Mart'in, and Yuke Zhu. Mimicdroid: In-context learning for humanoid robot manipulation from human play videos. *ArXiv*, abs/2509.09769, 2025. URL https://api.semanticscholar.org/CorpusID:281309736.

Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. Ricl: Adding in-context adaptability to pre-trained vision-language-action models. *ArXiv*, abs/2508.02062, 2025. URL https://api.semanticscholar.org/CorpusID:280422322.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *ArXiv*, abs/2412.15109, 2024. URL https://api.semanticscholar.org/CorpusID:274859727.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Hongtao Wu, Ya Jing, Chi-Hou Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *ArXiv*, abs/2312.13139, 2023. URL https://api.semanticscholar.org/CorpusID:266374724.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and P. Abbeel. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:250088882.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025. URL https://api.semanticscholar.org/CorpusID:277435005.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. Flare: Robot learning with implicit world modeling. *ArXiv*, abs/2505.15659, 2025. URL https://api.semanticscholar.org/CorpusID:278782778.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *ArXiv*, abs/2504.02792, 2025. URL https://api.semanticscholar.org/CorpusID:277510147.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

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 T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

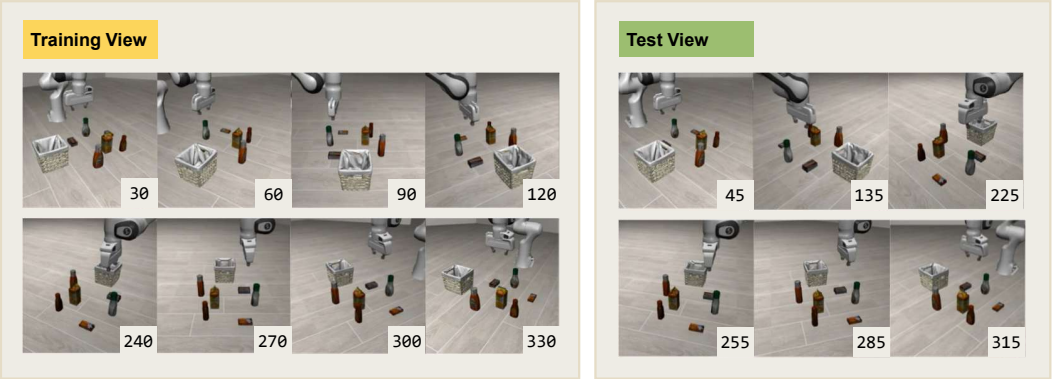



Figure 12: Viewpoint Distribution for Training and Evaluation. We illustrate the 8 in-domain training angles and 6 out-of-domain testing angles used to benchmark viewpoint generalization.

## A Proof of Proposition 1

We prove that an interaction context $\mathcal{T}_t = (o_{0:t}, a_{1:t})$ contains strictly more information about the system configuration $\psi$ than the initial observation $o_0$ under assumptions A1–A3.

*Proof.* By the chain rule of mutual information,

$$\begin{aligned} I(\psi; \mathcal{T}_t) &= I(\psi; o_{0:t}, a_{1:t}) \\ &= I(\psi; o_0, a_{1:t}) + I(\psi; o_{1:t} \mid o_0, a_{1:t}). \end{aligned} \tag{7}$$

We first consider the action sequence term. By assumption A2, the probing actions are independent of the unknown configuration conditioned on the initial observation:

$$I(\psi; a_{1:t} \mid o_0) = 0. \tag{8}$$

Therefore,

$$I(\psi; o_0, a_{1:t}) = I(\psi; o_0). \tag{9}$$

Substituting this equality gives

$$I(\psi; \mathcal{T}_t) = I(\psi; o_0) + I(\psi; o_{1:t} \mid o_0, a_{1:t}). \tag{10}$$

By assumption A3, the interaction-induced observations are informative about the hidden configuration:

$$I(\psi; o_{1:t} \mid o_0, a_{1:t}) > 0. \tag{11}$$

Hence,

$$I(\psi; \mathcal{T}_t) > I(\psi; o_0), \tag{12}$$

which proves the proposition. □

## B Additional Adaptation Baselines

We compare ICWM with additional adaptation strategies that use different forms of context or test-time adaptation. All methods use the same Qwen2.5-VL backbone, task data, and evaluation protocol as the controlled baselines. (1) RMA-style adaptation (Kumar et al., 2021) provides a learned configuration embedding, (2) PAD-style adaptation (Hansen et al., 2021) updates the policy from the available transitions, (3) Memory-style adaptation uses execution history, and Oracle demonstration provides task-relevant trajectories as an upper-bound reference. The results are reported on LIBERO-Long across the six unseen viewpoints.

Table 3: Additional adaptation baselines on LIBERO-Long unseen viewpoints.

| Method | 45° | 255° | 285° | 315° | 135° | 225° | Avg |
|---|---|---|---|---|---|---|---|
| MV | 30.4 | 24.0 | 32.8 | 27.6 | 1.2 | 2.8 | 19.8 |
| RMA-style | 31.2 | 24.2 | 28.6 | 30.8 | 1.0 | 6.2 | 20.3 |
| PAD-style | 28.2 | 22.8 | 34.2 | 28.0 | 1.0 | 2.4 | 19.4 |
| Memory-style | 35.6 | 24.8 | 33.8 | 32.4 | 1.4 | 3.6 | 21.9 |
| Oracle demonstration | 37.8 | 32.6 | 40.2 | 38.0 | 3.4 | 13.0 | 27.5 |
| **ICWM** | **36.6** | **28.4** | **36.6** | **37.6** | **2.2** | **8.8** | **25.0** |

The additional comparisons show that context design determines the quality of configuration inference. ICWM improves over MV on the average unseen-view success rate and remains competitive with the oracle demonstration setting while using task-agnostic probing transitions. The per-view results also preserve the same pattern observed in the main experiments: the gap-adjacent views are substantially more difficult than the bracketed interpolation views.

# C Details of Simulated Experiment Setup

## C.1 Experimental Configuration

**System Identification Benchmark Design.** Unlike the standard LIBERO setup, which typically evaluates policies under a fixed, single camera viewpoint, our experimental protocol is specifically designed to stress-test OOD viewpoint generalization. We distribute 14 discrete azimuthal angles around the workspace center. We designate 8 In-Domain (ID) angles for training: $\psi_{train} \in \{30°, 60°, 90°, 120°, 240°, 270°, 300°, 330°\}$, while withholding 6 Out-of-Domain (OOD) angles: $\psi_{test} \in \{45°, 135°, 225°, 255°, 285°, 315°\}$ exclusively for evaluation. Fig. 12 illustrates the drastic visual shifts between these angles, which introduce spatial ambiguities that standard VLA policies struggle to resolve without explicit system identification.

**LIBERO Suite Specifications.** To evaluate the versatile capabilities of our model, we utilize the LIBERO benchmark Liu et al. (2023), categorized into four specialized suites:

- LIBERO-Spatial: Benchmarks spatial reasoning by presenting identical objects in randomized initial layouts, requiring the agent to disambiguate relative geometric cues (e.g., "pick up the black bowl next to the plate and place it on the plate").
- LIBERO-Goal: Evaluates goal-conditioning by assigning multiple distinct objectives to the same scene configuration, assessing whether the policy can map different instructions to varied behavioral distributions (e.g., "open the top drawer and put the bowl inside").
- LIBERO-Object: Focuses on semantic grounding by introducing a diverse set of object categories within fixed layouts, testing the model's ability to generalize to novel visual appearances (e.g., "pick up the bbq sauce and place it in the basket").
- LIBERO-Long: Challenges the model with 10 complex, multi-stage manipulation sequences that require sustained temporal consistency and the ability to recover from small execution errors over extended horizons (e.g., "turn on the stove and put the moka pot on it").

**Data Synthesis and Preprocessing.** The training data is generated by replaying expert demonstration trajectories from the original benchmark in the simulator and re-rendering them from the 8 ID camera poses. Following the preprocessing protocol of OpenVLA Kim et al. (2024), we filter out unsuccessful episodes and remove redundant frames where action norms are near-zero and the gripper state is static, ensuring a high-density learning signal.

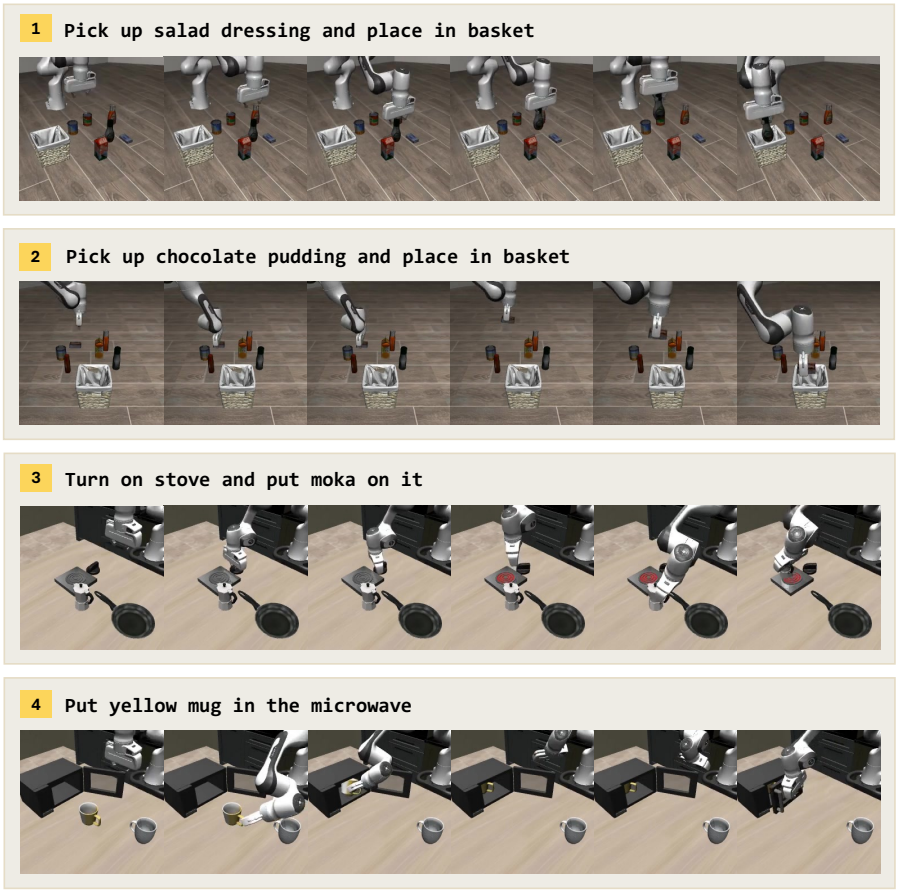



Figure 13: Visualization of Successful Task Rollouts in Simulation. By utilizing the interaction prefix to resolve viewpoint-induced ambiguities, the agent achieves precise grasping and multi-stage execution (e.g., turning on a stove) without any environment-specific fine-tuning.

### C.2 Detailed Results

Tab. 4 and Tab. 5 provide the exhaustive breakdown of success rates for ID and OOD viewpoints. Notably, ICWM maintains robust performance across the majority of evaluated conditions, outperforming multi-view behavior cloning and explicit configuration approaches in most settings. In particular, for LIBERO-Long, ICWM achieves a $25.0\%$ average OOD success rate, a significant margin over the $19.8\%$ achieved by the Multi-View (MV) baseline, demonstrating the power of in-context calibration for complex tasks.

We also observe that certain OOD viewpoints, particularly 135°, pose challenges for all methods, including ICWM. We attribute this to viewpoint-specific geometric constraints: at 135°, the camera angle may introduce a certain degree of object occlusion and reduce the effective workspace visible to the model, which can cause manipulation targets to occasionally exit the field of view during execution. This suggests a perceptual limitation shared across methods, rather than a failure specific to ICWM alone.

To further investigate the model's behavior, we visualize the rollout trajectories in Figure 13. ICWM demonstrates remarkable execution stability; even when the initial viewpoint causes a spatial offset, the model utilizes the in-context world model to "re-align" its end-effector during the first few steps of the task, ensuring a successful grasp that baselines frequently miss due to viewpoint-induced depth and coordinate errors.

Table 4: Success Rates (%) on In-Domain (Seen) Viewpoints.

| | | 30° | 60° | 90° | 120° | 240° | 270° | 300° | 330° | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Spatial | $\pi$-FAST | 19.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.0 | 4.1 |
| | $\pi_{0.5}$ | 39.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 20.6 | 7.6 |
| | NORA | 11.4 | 0.6 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 17.8 | 3.8 |
| | MV | 79.4 | 77.2 | 69.4 | 65.4 | 72.8 | 71.4 | 81.8 | 78.4 | 74.5 |
| | EXP | 79.8 | 81.8 | 68.2 | 69.2 | 70.6 | 75.6 | 79.4 | 82.6 | 75.9 |
| | ICWM (ours) | **84.6** | **84.0** | **75.6** | **73.8** | **83.0** | **77.0** | **83.6** | **88.0** | **81.2** |
| Goal | $\pi$-FAST | 7.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 | 2.2 |
| | $\pi_{0.5}$ | 40.2 | 4.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 28.0 | 9.2 |
| | NORA | 5.6 | 0.0 | 0.4 | 0.0 | 0.2 | 0.0 | 0.0 | 4.8 | 1.4 |
| | MV | 78.4 | **77.2** | 68.4 | **74.0** | **72.8** | **69.0** | 71.4 | 75.4 | **73.3** |
| | EXP | 78.2 | 71.0 | 66.4 | 68.4 | 68.0 | 62.8 | **72.8** | **77.6** | 70.7 |
| | ICWM (ours) | **82.2** | 75.4 | **71.0** | 69.0 | 62.2 | 64.6 | 72.2 | 76.0 | 71.6 |
| Object | $\pi$-FAST | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 11.8 | 1.5 |
| | $\pi_{0.5}$ | 13.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.6 | 43.4 | 8.1 |
| | NORA | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.6 | 0.6 |
| | MV | 66.0 | 67.0 | **72.4** | 60.2 | 54.6 | 52.6 | 73.8 | 72.8 | 64.9 |
| | EXP | 62.0 | 68.2 | **72.4** | **63.4** | 62.4 | 55.6 | 73.6 | 75.6 | 66.6 |
| | ICWM (ours) | **74.0** | **70.6** | 71.6 | 60.4 | **69.8** | **62.0** | **76.6** | **79.0** | **70.5** |
| Long | $\pi$-FAST | 2.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.6 | 0.7 |
| | $\pi_{0.5}$ | 13.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 | 2.9 |
| | NORA | 1.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 6.2 | 1.0 |
| | MV | 36.2 | 32.0 | 29.6 | 24.8 | 28.0 | 28.2 | 33.4 | 34.0 | 30.8 |
| | EXP | 39.8 | 32.6 | 26.4 | 31.0 | 30.4 | 26.8 | 33.2 | 39.0 | 32.4 |
| | ICWM (ours) | **45.2** | **41.2** | **37.2** | **39.4** | **38.4** | **34.2** | **39.8** | **44.6** | **40.0** |




Figure 14: Visibility across the gap-adjacent viewpoints. The highlighted 135° and 225° views show the workspace and gripper–target relationship under the most challenging viewing directions; intermediate angles illustrate the change in line of sight.

### C.3 Visibility at Gap-Adjacent Viewpoints

The 135° and 225° viewpoints lie adjacent to the large angular coverage gap in the training distribution. At these viewpoints, the camera line of sight exposes partial occlusion around the arm, gripper, and target workspace. The intermediate views in Fig. 14 illustrate the gradual change in visibility across this region and provide a qualitative explanation for the low success rates shared by all methods.

Table 5: Success Rates (%) on Out-of-Domain (Unseen) Viewpoints.

| | | 45° | 135° | 225° | 255° | 285° | 315° | Avg |
|---|---|---|---|---|---|---|---|---|
| Spatial | $\pi$-FAST | 2.6 | 0.0 | 0.0 | 0.0 | 0.4 | 3.4 | 1.1 |
| | $\pi_{0.5}$ | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.8 | 1.8 |
| | NORA | 6.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.6 | 1.6 |
| | MV | 73.2 | **3.2** | **18.2** | 51.4 | 68.2 | **75.6** | 48.3 |
| | EXP | 71.2 | 1.8 | 11.0 | 53.6 | 65.2 | 75.2 | 46.3 |
| | ICWM (ours) | **78.2** | 2.4 | 17.0 | **55.4** | **71.2** | 75.0 | **49.9** |
| Goal | $\pi$-FAST | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.2 |
| | $\pi_{0.5}$ | 29.2 | 0.0 | 0.0 | 0.0 | 0.0 | 6.2 | 5.9 |
| | NORA | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| | MV | 60.4 | 13.2 | 13.8 | 34.2 | **53.2** | 57.6 | 38.7 |
| | EXP | 62.6 | **21.8** | **18.0** | 40.8 | 51.2 | 54.6 | 41.5 |
| | ICWM (ours) | **74.8** | 18.6 | **18.0** | **41.0** | 50.0 | **62.8** | **44.2** |
| Object | $\pi$-FAST | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.2 | 0.7 |
| | $\pi_{0.5}$ | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 6.6 | 1.2 |
| | NORA | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | MV | 25.4 | **0.2** | 0.4 | 11.4 | 13.8 | 24.8 | 12.7 |
| | EXP | 23.4 | **0.2** | 0.0 | 15.2 | 14.8 | **38.2** | 15.3 |
| | ICWM (ours) | **28.6** | 0.0 | **1.0** | **24.2** | **20.8** | 20.8 | **15.9** |
| Long | $\pi$-FAST | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | $\pi_{0.5}$ | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 0.1 |
| | NORA | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | MV | 30.4 | 1.2 | 2.8 | 24.0 | 32.8 | 27.6 | 19.8 |
| | EXP | 29.8 | 1.0 | 6.4 | 24.0 | 27.8 | 32.2 | 20.2 |
| | ICWM (ours) | **36.6** | **2.2** | **8.8** | **28.4** | **36.6** | **37.6** | **25.0** |

# D Details of Real-world Experiments

## D.1 Experimental Configuration

**Physical Platform and Perception.** As illustrated in Fig. 15, our real-robot workstation is centered around a 6-DoF UR5e manipulator equipped with a Robotiq 2F-85 parallel gripper, providing a versatile platform for diverse manipulation tasks. To create a challenging multi-view sensing environment, we deploy an array of 12 cameras strategically positioned at varying elevations and azimuthal angles. This setup ensures comprehensive coverage of the workspace while introducing significant perspective-induced spatial distortions.

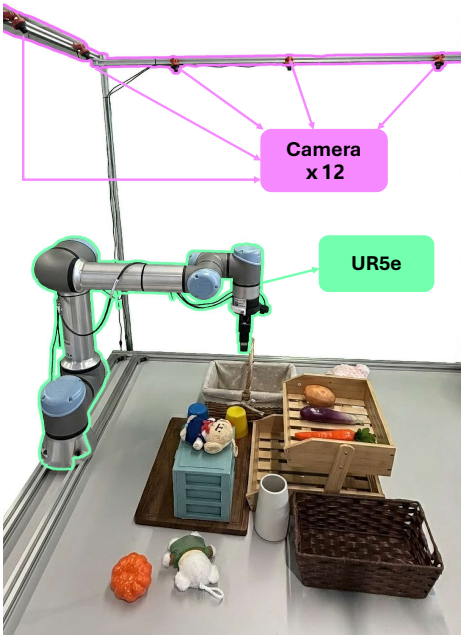



Figure 15: Real-World Experimental Platform. Our physical setup comprises a 6-DoF UR5e manipulator equipped with a parallel gripper and a comprehensive 12-camera perception array distributed at diverse elevations and azimuthal angles.

**Generalization Protocol and Data Collection.** To rigorously evaluate zero-shot adaptation, we partition the 12-camera system into two distinct subsets: 6 cameras are designated for training, while the remaining 6 cameras are withheld exclusively for testing. This balanced split ensures that the model is evaluated on a wide range of previously unseen perspectives, requiring success to depend entirely on functional system identification rather than the memorization of specific camera-to-robot geometries. For task-specific knowledge, we collect approximately 100–150 human demonstrations per task via teleoperation, covering four representative manipulation tasks:

1. Spatial Reasoning & Disambiguation: "Put the toy on the box into the basket." This task requires the agent to understand the vertical spatial relationship between the toy and the box, necessitating precise end-effector positioning to take the toy without disturbing the support surface—a process highly sensitive to viewpoint-induced depth errors

2. Fine-grained Alignment: "Stack the yellow cup onto the red cup." This serves as a benchmark for high-precision motor control, where the agent must align the principal axes of two objects under novel perspective projections.

3. Structural Manipulation: "Lift the basket." This task focuses on handle-centric grasping of large-scale empty containers, testing the model's ability to ground actions on specific structural affordances of an object.

4. Multi-Object Semantic Grounding: "Pick up the eggplant and place it onto the red plate." Conducted in a cluttered scene with multiple distractor objects, this task assesses the model's ability to maintain correct object-instruction alignment when viewed from unfamiliar angles that may cause occlusion or visual overlap.

For task-specific knowledge, we collect approximately 100–150 high-quality human demonstrations per task via teleoperation. These demonstrations capture the necessary precision and coordination for successful execution across diverse initial object layouts.

## D.2 Details of Generalization Experiments

**Semantic Perturbations.** Semantic perturbation experiments are evaluated across 4 tasks × 4 in-domain viewpoints × 10 trials per condition (160 trials per bar). We evaluate two types of scene-level variation: (1) Distractor objects: 10 task-irrelevant objects are placed within the workspace; (2) Novel table textures: the table surface is replaced with 4 unseen textures absent during training.

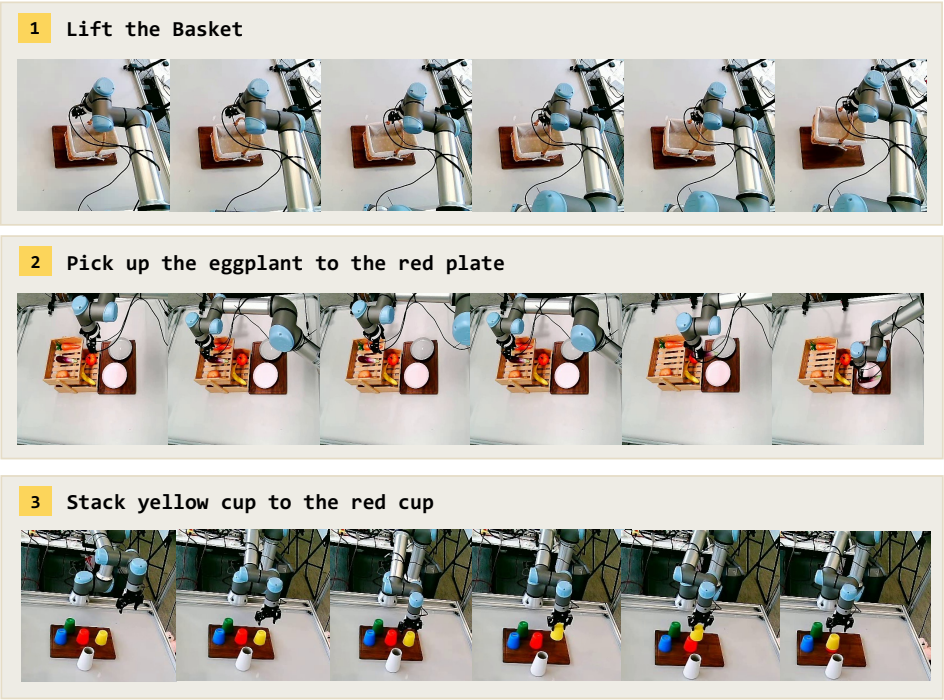



Figure 16: The supplementary successful trajectories from real-robot experiments. These trajectories illustrate the ICWM model's capability in completing lifting and pick-and-place maneuvers, as well as its robustness against object disturbances.

Table 6: Success rates (%) on OOD link-length configurations (WindowX platform). The model is trained on 100% and 70% link lengths, and evaluated on interpolated OOD configurations at 90% and 80%.

| Task | 90% length (10% reduction) | | 80% length (20% reduction) | |
|---|---|---|---|---|
| | MV | ICWM (ours) | MV | ICWM (ours) |
| Pick Cube | 90 | **95** | 40 | **80** |
| Stack Cube | 15 | **65** | 10 | **60** |
| Place Ball | 65 | **70** | 35 | **45** |
| Average | 57 | **77** | 28 | **62** |

**Morphological Generalization.** Morphological generalization experiments are evaluated across 4 tasks × 4 in-domain viewpoints × 10 trials x 3 morphological settings, with rigid spacers of $\Delta L \in \{20, 40, 80\}$ mm attached to the gripper flange to systematically alter forward kinematics at test time.

# E Details of Self-Exploration Probing

## E.1 Implementation Details

To enable test-time adaptation, we implement a stochastic self-exploration phase that allows the agent to synchronize its internal coordinate system with the current camera viewpoint. This phase is designed to be task-agnostic and prioritizes safety during real-robot deployment.

The exploration process involves moving the robot's end-effector toward a randomly sampled target coordinate $g_{probe} \in \mathbb{R}^3$. The control is executed via a fixed-step movement policy rather than a goal-reaching controller. It is important to note that the robot is not required to reach the target

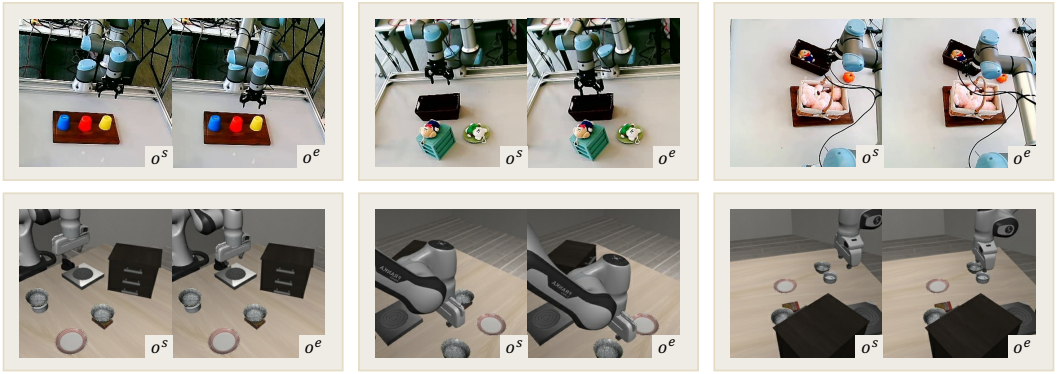



Figure 17: Visualization of start ($o^s$) and end frames ($o^e$) for random interaction clips in the real world and simulation.

point; the primary objective is to generate diverse visual-motor transitions $(o_t, a_t, o_{t+1})$ that expose the underlying action-observation mapping under the current configuration $\psi$. These interaction fragments are then collected and organized as the in-context prefix $\mathcal{T}$.

**Workspace-Constrained Random Exploration (Real-Robot).** For real-world experiments, probing targets $g_{probe}$ are sampled uniformly at random within the robot's reachable workspace. Rather than requiring manual specification, this workspace bounding box is derived automatically from the robot's forward kinematics model—given the joint limits of the UR5e, the reachable Cartesian space in the robot base frame can be computed analytically, requiring no human annotation or scene-specific calibration. The robot moves toward each sampled target via fixed-step increments, ensuring all motions remain within the operable joint limits. The resulting transitions cover diverse spatial regions of the workspace, forming the in-context prefix $\mathcal{T}$.

Crucially, since the bounding box is defined entirely in the robot's base frame, it remains invariant across arbitrary camera viewpoint shifts or semantic scene changes, requiring only a one-time specification for a fixed physical workstation setup. The random movements are task-agnostic and configuration-agnostic, carrying no task-relevant information. Our zero-shot claim refers specifically to the absence of task-specific demonstrations or parameter updates for any novel configuration; in simulation, this step is unnecessary entirely.

**Timing Analysis.** The probing phase introduces minimal overhead. Before task execution, the robot performs 20 probing actions spanning the workspace in approximately 5–6 seconds total. During task execution, $N$=5 triplets are randomly sampled from these 20 recorded transitions as the in-context prefix at each inference step. Crucially, this 20-step probing is performed *once per deployment*: the same context pool is reused throughout the entire task execution with no additional probing required. Given the multi-step nature of VLA manipulation tasks, this one-time overhead constitutes a small fraction of total deployment time, and is further amortized to negligible levels when the same configuration is reused across tasks.

### E.2 Case Visualization

We present the start ($o^s$) and end frames ($o^e$) of random action interaction clips from both the real world and simulation in Fig. 17.