Reflective VLA: In-Context Action Consequences Make VLAs Generalize

Qing Lian[1], Kent Yu[1,2,3], and Lei Zhang[1,2,3]

[1] Futian Laboratory
[2] International Digital Economy Academy (IDEA)
[3] Visincept

lianqing1997@gmail.com

Abstract

Most vision-language-action (VLA) models are reactive: they predict the next action from the current instruction and observation, implicitly assuming that the current observation fully specifies the action-relevant state. In embodied control, however, embodiment-specific factors such as camera-to-robot geometry, robot calibration, or systematic actuation bias are often hard to identify from a single observation. As a result, reactive policies cannot reliably disambiguate these factors in general, overfitting to training environments and generalizing poorly at deployment. We propose Reflective VLA, which conditions each decision on a context of observation–action–consequence triplets. Each triplet records not only what the robot observed and executed, but also how the scene changed afterward, exposing the deployment-specific mapping from actions to observed effects. Architecturally, Reflective VLA routes all observation modalities through the VLM under shared attention, so the action expert reasons directly over past triplets and the current observation. A block-causal mask enables parallel multi-frame training without leakage and supports KV-cached real-time inference. On standard LIBERO and SimplerEnv-Bridge, Reflective VLA preserves strong in-distribution performance. Under distribution shift on LIBERO-Plus and the harder LIBERO-Plus-Hard, it improves average success rate by 5.0 and 4.2 percentage points over a matched reactive baseline. Ablations with a matched history-only baseline further show that action consequences—rather than additional context length alone—are the key to cross-environment generalization.

1 Introduction

Vision-language-action (VLA) systems build on pretrained vision-language backbones Bai et al. [2025a,b], Beyer et al. [2024] and fine-tune on large-scale robot demonstrations Open X-Embodiment Collaboration et al. [2024], AgiBot-World-Contributors et al. [2025] to produce language-conditioned control policies Zitkovich et al. [2023], Kim et al. [2024], Black et al. [2025], Li et al. [2024a], NVIDIA et al. [2025]. By unifying scene understanding, instruction following, and low-level control in one model, they have substantially broadened the range of manipulation tasks a generalist policy can perform. Yet despite training on large and diverse datasets, current VLAs still generalize poorly to deployment environments unseen during training Fei et al. [2025], Zhang et al. [2025a].

Cross-environment generalization in embodied control often requires inferring embodiment-specific factors that are hard to identify from single observations, such as camera geometry, robot calibration, and systematic actuation bias. In contrast, the *interaction context*—an observation, an executed action, and the resulting observation—exposes these factors through how actions translate into observed changes: a commanded motion reveals camera extrinsics through its pixel displacement and calibration offsets through its pose residual. Existing VLAs condition on a single frame, omitting

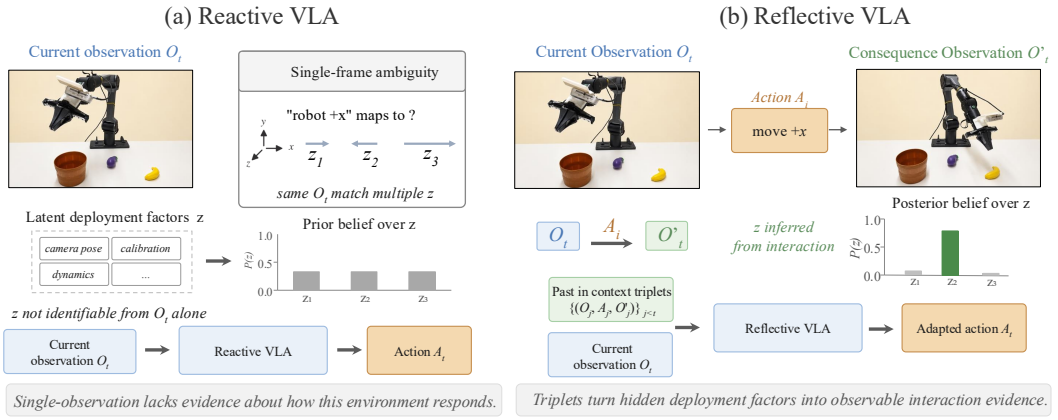


Figure 1: **From reactive to reflective control. (a) Reactive VLA.** Identifying embodiment-specific latent factors z—camera pose, calibration, etc.—from a single observation O_t is ill-posed: many z are consistent with the same frame, so A_t overfits training embodiments. **(b) Reflective VLA.** Conditioning on past causal triplets $\{(O_j, A_j, O'_j)\}_{j<t}$, where O'_j is the action-aligned observation after A_j, rules out deployments inconsistent with the evidence, sharpening $P(z \mid \text{context})$ and yielding an A_t adapted to the current embodiment.

this evidence, so the policy must instead memorize the embodiments seen during training. Recent temporal-context and memory-based VLAs Liu et al. [2025], Shi et al. [2026] improve state tracking, but lack explicit action–consequence binding, which is critical for identifying deployment-specific sensing and control factors.

We therefore cast cross-environment generalization in VLAs as an in-context learning (ICL) problem over causal interaction triplets. Given a small prompt of such triplets, the policy can infer the current deployment online from interaction feedback, analogous to how large language models infer task structure from few-shot demonstrations Brown et al. [2020], Xie et al. [2022]. Instantiating this in a dual-system VLA raises two challenges. First, the prompt interleaves heterogeneous modalities—images, proprioception, and continuous action chunks—which must be packed into a causal sequence without bottlenecking the action decoder. Second, dual-system VLAs must propagate context through both the VLM prefix and the action expert; naive training repeats forward passes across target frames, while naive inference recomputes the history prefix at every step, making real-time control costly.

We address these challenges with **Reflective VLA**, an ICL framework for dual-system VLAs. The current observation is augmented with a small set of past triplets, each consisting of an observation, an executed action chunk, and the resulting observation. To pack heterogeneous modalities into a single causal sequence without bottlenecking the action decoder, all modalities share the VLM token space and a continuous action expert attends densely to the full prompt under shared attention. To make ICL training tractable, a block-causal mask supervises all K context frames in a single forward pass instead of K separate ones. The same causal structure supports KV-cached inference, so the extended context does not compromise real-time control at deployment.

We evaluate Reflective VLA on standard LIBERO and SimplerEnv, as well as on LIBERO-Plus Fei et al. [2025] and our extension, LIBERO-Plus-Hard, which target deployment shifts in sensing, embodiment, and layout. Reflective VLA preserves strong standard-benchmark performance, improving over a matched reactive baseline by 1.1 and 5.3 points on LIBERO and SimplerEnv, respectively. Under deployment shifts, it improves held-out generalization by 5.0 and 4.2 points on LIBERO-Plus and LIBERO-Plus-Hard, respectively, without test-time fine-tuning. Ablations show that these gains come from observation–action–consequence evidence rather than longer context alone.

In summary, our main contributions are:

1. We identify observation–action–consequence interaction context as a key signal for cross-environment VLA generalization, and formulate it as in-context learning over causal triplets.
2. We introduce **Reflective VLA**, a dual-system VLA that places multimodal observations and historical action consequences in a shared VLM token space, with block-causal training and KV-cached real-time inference.

3. Across LIBERO, SimplerEnv, LIBERO-Plus, and LIBERO-Plus-Hard, Reflective VLA improves standard-benchmark performance and held-out deployment generalization over matched reactive baselines; history-only ablations confirm that aligned consequence observations are the critical ingredient.

2 Related Work

Vision-language-action models. Recent generalist policies build on pretrained vision-language backbones and train on large-scale robot datasets to produce language-conditioned control Reed et al. [2022], Brohan et al. [2023], Zitkovich et al. [2023], Kim et al. [2024], Team et al. [2024]. Architecturally, recent VLAs follow two main designs. One line casts control as autoregressive prediction over discretized or tokenized actions Zitkovich et al. [2023], Kim et al. [2024, 2025], Pertsch et al. [2025]. A more recent line decouples high-level reasoning from low-level control by pairing the VLM backbone with a dedicated continuous action expert based on diffusion or flow matching—a *dual-system* design that preserves action fidelity at high control rates Black et al. [2025], Li et al. [2024a], Wu et al. [2026], Zheng et al. [2026], NVIDIA et al. [2025]. We adopt this dual-system design and take an orthogonal direction: rather than scaling data or refining the action representation, we equip a fixed VLA with in-context interaction evidence as a new axis for cross-environment generalization, without any test-time updates.

Temporal context and memory for robot control. Temporal context is a natural response to partial observability in robot control. Imitation policies such as ACT, Diffusion Policy, and RDT use short observation histories and action chunks to stabilize visuomotor control Zhao et al. [2023], Chi et al. [2023], Liu et al. [2025]; VLA systems such as RoboFlamingo, 4D-VLA, and MemoryVLA further incorporate visual history, spatiotemporal representations, or explicit memory for language-conditioned manipulation Li et al. [2024b], Zhang et al. [2025b], Shi et al. [2026]. These methods improve state estimation, task progress tracking, and robustness to transient perceptual ambiguity. However, temporal history alone is not equivalent to interaction feedback: existing histories do not explicitly bind an executed action chunk to its aligned consequence. Reflective VLA therefore distinguishes temporal context from action-conditioned consequence context, and tests this distinction with history-only ablations that remove consequence observations from the same historical milestones.

In-context adaptation for sequential decision making. In-context learning provides a complementary view of adaptation: a sequence model can infer the relevant task or environment from examples in its context Brown et al. [2020], Xie et al. [2022]. This idea connects to meta-RL Duan et al. [2016], Finn et al. [2017], Rakelly et al. [2019] and to transformer policies that adapt behavior from contextual experience at test time, including Decision Transformer, Prompt-DT, and algorithm distillation Chen et al. [2021], Xu et al. [2022], Laskin et al. [2023]. Reflective VLA brings this viewpoint to embodied VLA control, but uses a different adaptation signal: structured multimodal observation–action–consequence triplets, without rewards, textual self-reflection, a separate system-identification module, or test-time parameter updates.

3 Method

Reflective VLA extends a reactive dual-system VLA with an in-context interface over observation–action–consequence triplets. We first define the reactive formulation, then describe how triplets are constructed and packed, how the model is trained on them in parallel using a block-causal mask, and how they are reused during online inference.

3.1 Preliminary: Standard Reactive VLA Formulation

At control step t, let $\mathcal{L}$ denote the language instruction, $\mathcal{O}_t$ the current multimodal observation, and $A_t = [a_t, \ldots, a_{t+C-1}]$ an action chunk of horizon C. The observation may include third-person images, wrist images, and proprioceptive states. A standard reactive VLA predicts the next action chunk from only the current instruction and observation:

$$A_t \sim \pi_\theta(\cdot \mid \mathcal{L}, \mathcal{O}_t). \tag{1}$$

This interface covers recent dual-system VLAs that pair a VLM prefix with a continuous diffusion or flow-matching action expert. Although their action generation objectives differ, they share the same

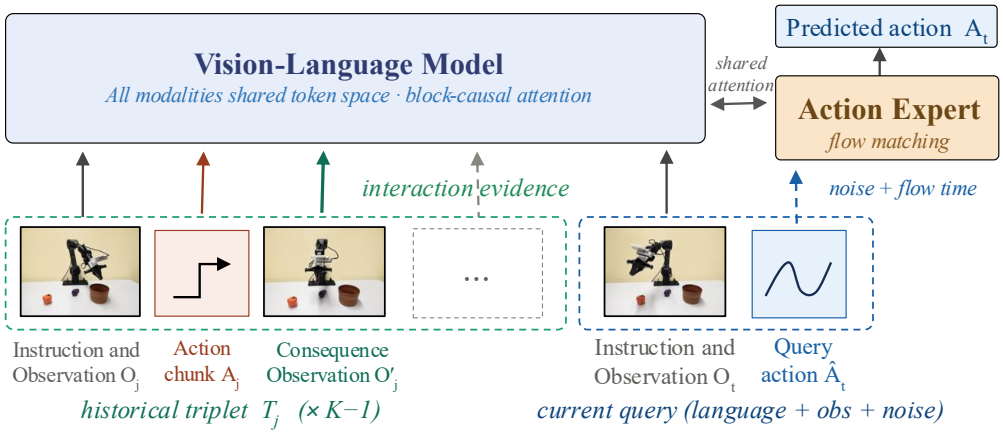


Figure 2: **Reflective VLA: observation–action–consequence in-context learning.** Past triplets $\{(O_j, A_j, O'_j)\}_{j=1}^{K-1}$ and the current observation O_t share a single VLM token sequence; a flow-matching action expert attends to this prefix and denoises $\hat{A}_t$ into the predicted chunk A_t. The aligned consequence O'_j carries the *interaction evidence* that exposes environment factors.

limitation for our purpose: the action expert conditions on the current observation, but not on past executed actions and their observed consequences.

3.2 From Reactive to In-Context Generalization

A reactive VLA learns $\pi(A_t \mid \mathcal{L}, \mathcal{O}_t)$, predicting the next action from the current instruction and observation. Different deployments may share task semantics but differ in camera-to-robot geometry, robot calibration, or systematic actuation bias. We summarize these embodiment-specific factors as an environmental latent variable z, on which the appropriate action depends. Inferring z from a single observation $\mathcal{O}_t$ is ill-posed: the mapping from z to $\mathcal{O}_t$ is many-to-one, leaving the policy unable to disambiguate the current sensing and control conditions.

This missing evidence can be recovered by interaction feedback: when the robot executes an action and observes the resulting change, the observation–action–consequence relation reveals how the current deployment responds to commands. Formally, an adaptive policy can be written as a marginal over the posterior on z:

$$\pi(A_t \mid \mathcal{L}, \mathcal{O}_t, \mathcal{H}) = \int \pi(A_t \mid \mathcal{L}, \mathcal{O}_t, z)\, P(z \mid \mathcal{H})\, dz, \tag{2}$$

where $\mathcal{H}$ is the interaction context accumulated so far. We do not explicitly estimate z or compute this integral; the formulation serves only as a motivating abstraction, since transformers can implicitly approximate such posterior updates via in-context inference Xie et al. [2022].

The key requirement is that $\mathcal{H}$ is diagnostic of z. Under an idealized Markov view, if $\mathcal{H}$ consists of interaction triplets, the posterior factorizes as

$$P(z \mid \mathcal{H}) \propto P(z) \prod_i P_{\text{env}}(\mathcal{O}'_i \mid \mathcal{O}_i, A_i, z), \tag{3}$$

where $\mathcal{O}'_i$ is the observation after executing action chunk A_i. The likelihood depends explicitly on how the environment responds to executed actions, so a context containing only $(\mathcal{O}_i, A_i)$ pairs cannot expose the response term and provides limited evidence for identifying deployment-specific factors.

Reflective VLA therefore conditions each decision on interaction context with explicit consequences:

$$A_t \sim \pi_\theta\big(A_t \mid \mathcal{L}, \mathcal{H}, \mathcal{O}_t\big), \quad \mathcal{H} = \big\{(\mathcal{O}_i, A_i, \mathcal{O}'_i)\big\}_{i=1}^{K-1}, \tag{4}$$

where $\mathcal{H}$ is the observation–action–consequence context.

3.3 Reflective VLA

Observation–action–consequence context. Reflective VLA represents recent interaction history as structured observation–action–consequence triplets. Each triplet stores the observation before an

action chunk, the action chunk executed from that observation, and the resulting observation after the chunk completes. Since the policy predicts C-step action chunks, we define the consequence observation as $\mathcal{O}'_i = \mathcal{O}_{i+C}$ rather than the immediate next frame $\mathcal{O}_{i+1}$. This action-aligned consequence better captures the visible effect of the executed action, such as end-effector displacement, residual control error and etc.

At training time, A_i is the demonstration action chunk and $\mathcal{O}'_i$ the observation after the chunk horizon; at deployment, A_i is the chunk actually executed and $\mathcal{O}'_i$ the subsequent observation collected from the environment. Thus, each stored triplet represents a realized interaction rather than only a planned transition. We embed each previous action chunk into eight tokens in the VLM token space via a learned projection g_A, and write the resulting prefix tokens simply as A_i. The i-th context element is then $T_i = (\mathcal{L}, \tau_i, \mathcal{O}_i, A_i, \mathcal{O}'_i)$, where the language instruction $\mathcal{L}$ is re-inserted at the start of every triplet so that each unit is a self-contained $(\mathcal{L}, \mathcal{O}, A, \mathcal{O}')$ block, matching the per-triplet structure shown in Figure 3. We refer to these as causal triplets because the action token links a pre-action observation to its observed post-action consequence.

Multimodal prompt construction. Given historical milestones $m_1, \ldots, m_{K-1}$ and the current query timestep t, Reflective VLA packs past triplets and the current observation—each prefixed by the language instruction $\mathcal{L}$—into a single multimodal sequence:

$$X_{\text{in}} = \Big[T_{m_1}, \ldots, T_{m_{K-1}},\ (\mathcal{L}, \tau_t, \mathcal{O}_t)\Big]. \tag{5}$$

The historical triplets provide evidence about how the current deployment responds to actions, while the final $(\mathcal{L}, \tau_t, \mathcal{O}_t)$ block serves as the query for predicting the next action chunk. The current target action is not inserted into the prefix; it is generated by the continuous action expert.

Shared-attention dual-system architecture. For the action decoder to use interaction context, all observation modalities must be visible to it. Some dual-system VLAs route only part of the observation through the VLM backbone or condition the action module on a compressed prefix NVIDIA et al. [2025], Zheng et al. [2026], which can bottleneck reasoning over (O, A, O') structure. Reflective VLA therefore routes all observation modalities—third-person images, wrist images, and proprioception—into a shared VLM token sequence. Images are encoded by the visual backbone, while proprioceptive states and previous action chunks are projected into the token space with a learned non-linear head. A continuous flow-matching action expert is attached as a suffix that shares attention with the VLM prefix at every layer, following recent MoT-style VLA designs Liang et al. [2025], Black et al. [2025], Wu et al. [2026]. The action expert can thus attend directly to prior observations, prior actions, observed consequences, temporal indices, and the current observation when generating A_t.

Block-causal training. In-context conditioning lengthens each training sequence by a factor of K, so supervising each context position with an independent forward pass would scale training cost as $\mathcal{O}(K)$. We instead borrow the packed-sequence idea from LM training and supervise all K sampled frames jointly under a *block-causal* attention mask, which lets every sampled frame act as both context for later targets and as a prediction target itself within a single forward pass. Because the same mask supervises positions with $0, 1, \ldots, K-1$ preceding triplets, training naturally covers the full range of context lengths the model will encounter at deployment.

Sampling and packing. We sample K ordered frames $t_1 < \cdots < t_K$ from a trajectory and pack them into a single sequence: the first $K-1$ form the historical triplets and the K-th is the current query, but the block-causal mask supervises all K frames jointly. As reflected in T_i, $\mathcal{L}$ is re-inserted at the start of every triplet (and the query block) so each unit is a self-contained $(\mathcal{L}, \mathcal{O}, A, \mathcal{O}')$ structure, and special tokens demarcate triplet boundaries so the action expert can unambiguously parse where each unit begins and ends. Because adjacent action chunks are temporally smooth, a fixed inter-frame stride invites the policy to extrapolate A_{t_k} from past actions rather than learn from observed consequences; we therefore randomize the stride within a bounded range during training to break this shortcut.

Mask and objective. Each target frame t_k is realized by a query slot $\hat{A}_{t_k}$ in the suffix flow-matching expert; Figure 3 visualizes the resulting attention pattern as a row in the mask. The query attends only to

$$\mathcal{V}_{t_k} = \{T_{t_j} : j < k\} \cup \{(\mathcal{L}, \tau_{t_k}, \mathcal{O}_{t_k})\}, \tag{6}$$

Table 1: **In-distribution evaluation on LIBERO and SimplerEnv-Bridge.** Success rates (%) across the four standard LIBERO task suites and the SimplerEnv-Bridge benchmark. † denotes our reproduced reactive baseline $\pi_{0.5}$. "–" indicates the result is not provided.

	LIBERO					SimplerEnv-Bridge				
Method	Spatial	Object	Goal	Long	Avg	Spoon	Carrot	Cube	Eggplant	Avg
OpenVLA Kim et al. [2024]	84.7	88.4	79.2	53.7	75.9	4.2	0.0	8.3	45.8	14.6
CoT-VLA Zhao et al. [2025]	87.5	91.6	87.6	69.0	81.1	–	–	–	–	–
4D-VLA Zhang et al. [2025b]	93.8	92.8	95.6	86.5	92.2	–	–	–	–	–
ThinkAct Huang et al. [2025]	–	–	–	–	–	37.5	8.7	58.3	70.8	43.8
CogACT Li et al. [2024a]	97.2	98.0	90.2	88.8	93.2	71.7	50.8	15.0	67.5	51.3
InternVLA-M1 Intern Robotics [2025]	98.0	**99.0**	93.8	92.6	95.9	87.5	67.9	31.3	**100.0**	71.7
π_0 Black et al. [2025]	96.8	98.8	95.8	85.2	94.2	–	–	–	–	–
MemoryVLA Shi et al. [2026]	98.4	98.4	96.4	93.4	96.5	75.0	75.0	37.5	**100.0**	71.9
GR00T N1.5 NVIDIA et al. [2025]	–	–	–	–	–	82.0	72.0	54.0	63.0	67.8
$\pi_{0.5}$ Physical Intelligence [2025]	**98.8**	98.2	98.0	92.4	96.9	–	–	–	–	–
Reactive baseline $\pi^{\dagger}_{0.5}$	97.5	98.2	97.8	94.0	96.9	91.7	79.2	70.8	50.0	72.9
Reflective VLA (ours)	98.4	**99.0**	**98.2**	**94.6**	**98.0**	**95.8**	**83.3**	79.2	54.2	**78.2**

i.e., all completed prior triplets (each carrying its own copy of $\mathcal{L}$) together with the current language-prefixed observation. It is masked from the prefix action token A_{t_k} and consequence $\mathcal{O}'_{t_k}$ within its own triplet—both observed only after A_{t_k} is executed—and from every subsequent triplet $\{T_{t_j} : j > k\}$; sibling query slots $\{\hat{A}_{t_j}\}_{j \neq k}$ are also mutually masked, so each prediction depends solely on prefix evidence. The training objective sums the flow-matching action loss over all valid targets:

$$\mathcal{L}_{\text{train}} = \sum_{k=1}^{K} \mathcal{L}_{\text{act}}\big(A_{t_k}; \mathcal{V}_{t_k}\big), \tag{7}$$

where $\mathcal{L}_{\text{act}}$ is the same flow-matching objective used by the base VLA. This yields dense multi-frame supervision in one forward pass while ensuring that each prediction is conditioned only on past completed interactions and the current observation. The first target t_1 has no preceding triplets and provides reactive supervision, while $t_2, \ldots, t_K$ provide ICL supervision with progressively longer context, as the staircase pattern along the diagonal of Figure 3 shows.

At deployment, Reflective VLA keeps a rolling buffer of the most recent $K-1$ triplets. Unlike training, where ground-truth action chunks populate the historical context, at inference each triplet stores the policy's own predicted chunk $\hat{A}_t$ together with the observation $\mathcal{O}_{t+C}$ actually reached after executing it—so the in-context prefix reflects the realized rollout rather than a teacher trajectory. At each step, only the new observation (or triplet, after execution) is encoded; VLM-side keys and values for past triplets are cached once and reused, keeping the per-step inference cost roughly constant.

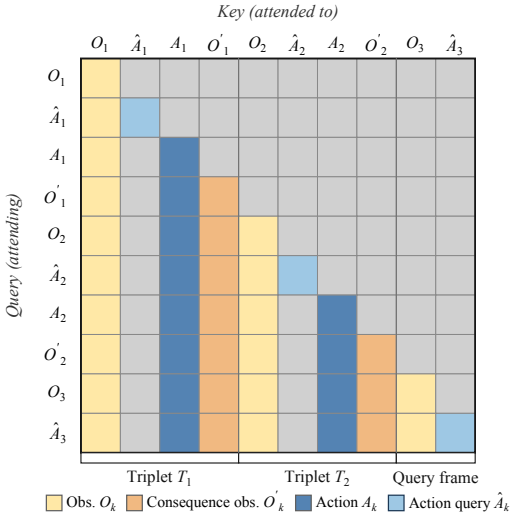


Figure 3: **Block-causal mask.** Each query $\hat{A}_k$ attends to $\mathcal{L}$, prior triplets $T_{<k}$, and $\mathcal{O}_k$, while its own A_k, $\mathcal{O}'_k$, future triplets and queries are masked, supervising all targets in one forward.

4 Experiments

We evaluate Reflective VLA along three axes. First, we test whether adding interaction context preserves strong in-distribution performance on standard manipulation benchmarks. Second, we evaluate robustness under perturbations that change visual appearance, sensing, or embodiment-specific properties. Third, we ablate the designed components to illustrate the contribution of the observation–action–consequence structure rather than from longer context alone.

Table 2: **Robustness under perturbations on LIBERO-Plus and LIBERO-Plus-Hard.** Success rates (%) on LIBERO-Plus (7 standard perturbation categories) and LIBERO-Plus-Hard (2 additional harder shifts: Multi-camera shift (Camera†) and Robot calibration shift (Rob. Calib†)).

	LIBERO-Plus								LIBERO-Plus-Hard		
Method	Camera	Robot	Lang	Light	Bg	Noise	Layout	Avg	Camera†	Rob. Calib†	Avg
UniVLA Bu et al. [2025]	1.8	46.2	69.6	69.0	81.0	21.2	31.9	45.8	–	–	–
π_0 Black et al. [2025]	13.8	6.0	58.8	85.0	81.4	79.0	68.9	56.3	–	–	–
OpenVLA-OFT Fei et al. [2025]	92.8	30.3	85.8	94.9	93.9	89.3	**77.6**	80.7	72.2	43.1	57.7
MemoryVLA Shi et al. [2026]	93.1	42.1	84.2	95.1	92.7	89.1	77.5	82.0	72.1	49.2	60.7
Reactive baseline $\pi^\dagger_{0.5}$	90.0	50.0	94.9	92.0	85.8	90.2	75.0	82.6	74.0	55.2	64.6
Reflective VLA (ours)	**95.7**	**72.9**	92.2	92.1	92.0	**95.4**	73.0	**87.6**	**76.3**	**61.3**	**68.8**

Table 3: **Ablations on the interaction context.** Success rates (%) on Camera, Camera†, and Rob. Calib†; Avg denotes the mean. (a) varies what each context element contains, with K fixed; (b) varies the number of context elements K with the full (O, A, O') structure. K=1 corresponds to the reactive baseline without historical triplets.

(a) Context composition (fixed K).

Context	Camera	Camera†	Rob. Calib†	Avg
Reactive (no history)	90.0	74.0	55.2	73.1
O	88.8	73.2	55.0	72.3
O, A	89.1	74.9	57.4	73.8
O, A, O'	**95.7**	**76.3**	**61.3**	**77.8**

(b) Context length (full (O, A, O')).

K	Camera	Camera†	Rob. Calib†	Avg
1	90.0	74.0	55.2	73.1
2	93.8	74.6	57.4	75.3
4	96.1	75.7	58.4	76.7
8	**95.7**	**76.3**	**61.3**	**77.8**

4.1 Experimental Setup

Simulation benchmarks. We evaluate on four simulation settings: *(i)* LIBERO Liu et al. [2023], the standard suite of 40 language-conditioned manipulation tasks across *Spatial*, *Object*, *Goal*, and *Long*, with a fixed camera and embodiment, used as our nominal in-distribution testbed; *(ii)* SimplerEnv-Bridge Li et al. [2024c], which evaluates VLAs trained on the BridgeDataV2 Walke et al. [2023] in the ManiSkill2 Gu et al. [2023] simulator, providing a real-to-sim transfer setting; *(iii)* LIBERO-Plus Fei et al. [2025], which extends LIBERO with seven perturbation categories spanning sensing, language, and robot–environment configuration; and *(iv)* LIBERO-Plus-Hard, our diagnostic extension targeting shifts in the action-to-observation mapping.

A context-diagnostic benchmark. While LIBERO-Plus provides broad perturbation coverage, not all of its categories are equally diagnostic of interaction-conditioned adaptation: shifts in language, or background can largely be absorbed by invariances in the pretrained VLM backbones. We therefore introduce LIBERO-Plus-Hard with two perturbations designed so that single-frame evidence is insufficient and the action-to-observation mapping must be inferred from interaction:

- **Multi-camera shift** (Camera† in Table 2). We jointly perturb the extrinsics of all camera views (third-person and wrist), so no single view preserves the nominal calibration; the camera-to-robot geometry must be recovered from the pixel-space displacement induced by past action chunks.
- **Robot calibration shift** (Rob. Calib† in Table 2). We inject an episode-level systematic offset between commanded and achieved end-effector motion, simulating calibration error, actuation bias, or mechanical backlash; this offset is unobservable from a static frame but recoverable from the residual between commanded actions and their consequences.

Baselines. Our primary baseline, denoted $\pi^\dagger_{0.5}$, is a reproduced *reactive* $\pi_{0.5}$ that uses the same backbone, training data, and network parameters as Reflective VLA but with context length K=1, conditioning each prediction only on the current instruction and observation. Where available and protocol-compatible, we additionally report published results for OpenVLA Kim et al. [2024], OpenVLA-OFT Kim et al. [2025], UniVLA Bu et al. [2025], CoT-VLA Zhao et al. [2025], 4D-VLA Zhang et al. [2025b], ThinkAct Huang et al. [2025], CogACT Li et al. [2024a], InternVLA-M1 Intern Robotics [2025], π_0 Black et al. [2025], $\pi_{0.5}$ Physical Intelligence [2025], MemoryVLA Shi et al. [2026], and GR00T N1.5 NVIDIA et al. [2025].

Implementation details. Following $\pi_{0.5}$ Physical Intelligence [2025], Reflective VLA adopts a Mixture-of-Transformers Liang et al. [2025] dual-system design, pairing a pretrained VLM prefix with a flow-matching continuous action expert as the suffix under shared attention. Unless otherwise stated, we use action chunk size $C{=}10$, context length $K{=}8$, corresponding to seven historical triplets and one query observation, and bounded randomized milestone sampling. Additional architectural details, hyperparameters, and training settings are provided in Sections A and C.1.

4.2 Main Results

In-distribution performance. Table 1 shows that Reflective VLA preserves strong in-distribution performance on both LIBERO and SimplerEnv-Bridge. On LIBERO it reaches 98.0% average success, achieving state-of-the-art—ahead of the published $\pi_{0.5}$ (96.9%) and memory-based MemoryVLA (96.5%)—and outperforming the matched reactive $\pi_{0.5}^{\dagger}$ baseline by 1.1 points. On SimplerEnv-Bridge, it reaches 78.2% average success, again achieving state-of-the-art and improving over the reactive baseline by 5.3 points; since BridgeData V2 Walke et al. [2023] spans diverse scenes, embodiments, and viewpoints, this gain reflects the model's ability to exploit in-context interaction evidence under embodiment-level variability already present in the training distribution. Performance on the *Long* suite is also slightly improved (94.0% → 94.6%), indicating that interaction context does not hurt temporally extended manipulation. Together, these results show that adding structured interaction context does not compromise the base policy's nominal task-solving ability.

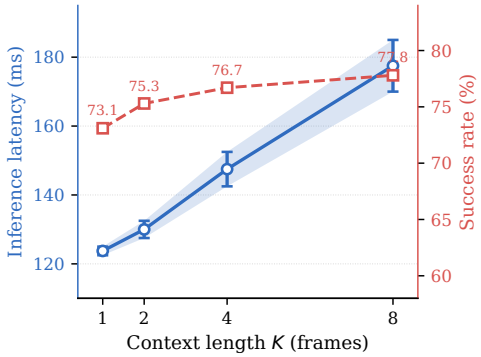
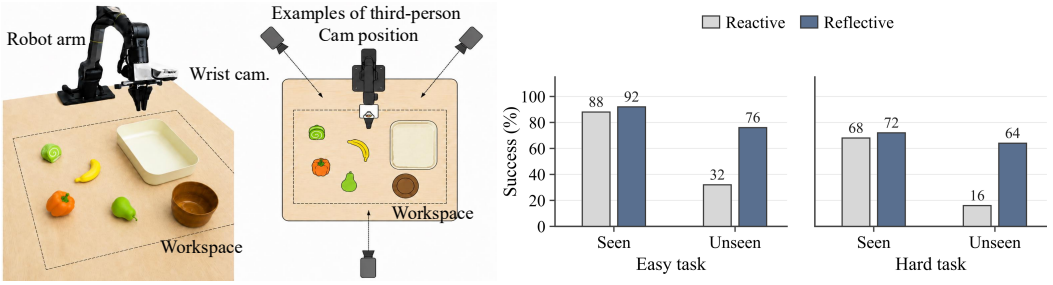


Figure 4: Latency–accuracy trade-off across context length K on the perturbation subset.

Robustness under perturbations. Table 2 evaluates robustness on the seven standard LIBERO-Plus perturbation categories and two harder shifts in LIBERO-Plus-Hard; MemoryVLA is reproduced on the same training data as ours for a fair comparison. On LIBERO-Plus, Reflective VLA achieves 87.6% average success, outperforming the matched reactive baseline (82.6%) and strong prior methods such as OpenVLA-OFT (80.7%). Reflective VLA improves on five of the seven categories, with the largest gains under *Robot* (+22.9 pp), *Background* (+6.2 pp), and *Noise* (+5.2 pp); both methods remain strong on language perturbations, where the reactive baseline slightly leads (94.9% vs. 92.2%). The categories with the largest gains directly affect either the sensing interface or the robot's spatial configuration, both of which are difficult to identify from a single frame but exposed through how past actions translate into observed consequences.

Discussion. The two LIBERO-Plus-Hard shifts further stress-test this hypothesis. Under *Multi-camera shift* (Camera†), where the extrinsics of both third-person and wrist views are perturbed, Reflective VLA improves from 74.0% to 76.3%. Under *Robot calibration shift* (Rob. Calib†), where the achieved end-effector motion deviates systematically from the commanded action, it improves from 55.2% to 61.3%. Together, these two shifts raise the average success rate from 64.6% to 68.8%, a 4.2-point gain over the matched reactive baseline. Unlike language or appearance perturbations, both shifts change the relationship between actions and their observed effects, making them difficult to identify from any single frame. The largest gain appears on Rob. Calib†, where the only diagnostic signal is the residual between commanded actions and their consequences—directly supporting the central design of Reflective VLA.

Context composition. Table 3a isolates the effect of context composition with K fixed. Adding observation history alone (O) yields no improvement over the reactive baseline (72.3% vs. 73.1%), and adding observation–action pairs without the resulting consequence (O,A) provides only a marginal gain (73.8%). In contrast, the full observation–action–consequence context (O,A,O′) improves the average from 73.1% to 77.8%, a 4.7-point gain that is most pronounced on Rob. Calib† (55.2% → 61.3%). This supports our hypothesis that the action-aligned consequence $\mathcal{O}'$ provides the key adaptation signal: temporal context alone, even when paired with the executed action, leaves the action-to-observation mapping unidentified.



(a) Real robot setup (b) Camera placement protocol (c) Results on seen and unseen camera setups

Figure 5: **Real-world setup.** (a) An Agilex Piper arm with RealSense D435i cameras over a tabletop workspace, with two tasks (place-into-box, place-into-bowl). (b) Third-person camera-placement protocol: ten placements span the left, front, and right of the workspace; demonstrations cover all ten, while evaluation uses five seen and five held-out placements drawn from the same regions.

Context length. Table 3b studies the effect of context length. Increasing K consistently improves robustness, from 73.1% at $K{=}1$ to 75.3% at $K{=}2$, 76.7% at $K{=}4$, and 77.8% at $K{=}8$. Most of the gain is captured by $K{=}4$ (+3.6 points over the reactive baseline), with diminishing returns thereafter, suggesting that a small number of recent interaction triplets already captures most of the useful deployment-specific evidence.

Latency–accuracy trade-off. Figure 4 reports per-step latency against success rate during inference. Since past triplets are encoded once and reused, latency grows sub-linearly in K: $K{=}8$ is only $1.43\times$ slower than $K{=}1$ (178 vs. 124 ms) despite an $8\times$ longer prompt, while $K{=}4$ already recovers most of the accuracy gain at $1.19\times$ baseline latency—a practical operating point when tighter control rates are required. We cap K at 8 because longer contexts exceed our GPU memory budget during block-causal training; KV-cached inference itself can scale further.

4.3 Real-World Experiments

We further evaluate Reflective VLA under real-world cross-camera generalization. Following the protocol in Figure 5, demonstrations cover ten third-person camera placements spanning the left, front, and right sides of the workspace; at test time we evaluate on five seen and five held-out placements ($N = 25$ trials each) without test-time fine-tuning. We consider two tabletop tasks—placing an object into a large square box (primarily grasping) and into a small bowl (additionally requires camera-dependent placement). Full protocol details are in Section D.

Reflective VLA preserves performance on seen viewpoints while substantially improving generalization to unseen ones. On seen placements, success improves only slightly over the reactive baseline (88%→92% on box, 68%→72% on bowl), indicating that interaction context does not compromise nominal performance when camera geometry is covered by training. Under unseen placements, the reactive baseline degrades sharply (32% box, 16% bowl) while Reflective VLA reaches 76% and 64%. Although the bowl task is harder overall due to its tighter placement tolerance, the gain on unseen viewpoints (+48 pp) matches the box task (+44 pp), suggesting that reflective context benefits both grasping robustness and camera-dependent spatial alignment.

5 Conclusion

We cast cross-environment generalization in VLAs as in-context inference over causal triplets $(\mathcal{O}, A, \mathcal{O}')$, enabling adaptation from interaction feedback without test-time fine-tuning. Two findings support this view: a matched history-only ablation that omits the action aligned consequence $\mathcal{O}'$ recovers little of the gain, showing that context length alone is insufficient; and Reflective VLA closes a substantial portion of the OOD gap on the simulation and real world generalization environments. These gains come at constant model size—only a change in how the policy uses its context window. Scaling context length, extending reflection to longer horizons, and applying this framework to diverse embodiments are natural next steps.

References

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *RSS*, 2025. doi: 10.15607/RSS.2025.XXI.010.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.025.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, volume 33, 2020.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. In *RSS*, 2025. doi: 10.15607/RSS.2025.XXI.014.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, 2021.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.026.

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, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhaoyuan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yuzhe Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. ThinkAct: Vision-language-action reasoning via reinforced visual latent planning. In *NeurIPS*, 2025.

Intern Robotics. InternVLA-M1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *CoRL*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *RSS*, 2025. doi: 10.15607/RSS.2025.XXI.017.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *ICLR*, 2023.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *ICLR*, 2024b.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024c.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025.

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

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: A diffusion foundation model for bimanual manipulation. In *ICLR*, 2025.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *ICRA*, 2024.

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. *arXiv preprint arXiv:2501.09747*, 2025.

Physical Intelligence. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *ICML*, 2019.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *ICLR*, 2026.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *RSS*, 2024. doi: 10.15607/RSS.2024.XX.090.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning*, volume 229, pages 1723–1736, 2023.

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

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *ICLR*, 2022.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua B. Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *ICML*, 2022.

Hongyin Zhang, Shuo Zhang, Junxi Jin, Qixin Zeng, Runze Li, and Donglin Wang. Robustvla: Robustness-aware reinforcement post-training for vision-language-action models. *arXiv preprint arXiv:2511.01331*, 2025a.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. In *NeurIPS*, 2025b.

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

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.016.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *ICLR*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.

Supplementary Material

This supplementary material provides the implementation and experimental details needed to reproduce Reflective VLA. Section A describes the architecture components, triplet construction, context-buffer behavior, block-causal masking, and flow-matching inference procedure. Section C details the training data construction, optimization settings, and evaluation protocol for LIBERO, LIBERO-Plus, LIBERO-Plus-Hard, and SimplerEnv-Bridge. Section B specifies the two diagnostic perturbation families used in LIBERO-Plus-Hard. Section D documents the real-world robot setup, control interface, tasks, and camera-placement protocol. Section E summarizes reproducibility resources, external assets, and current limitations.

A Implementation Details

A.1 Architecture components

Reflective VLA uses the same model family and training budget as the matched reactive baseline, and changes only the conditioning interface. We implement it with a dual-system Mixture-of-Transformers (MoT) architecture: a pretrained VLM encodes the language and multimodal observation prefix, and a continuous flow-matching transformer action expert is attached as a suffix with shared self-attention to the prefix. We instantiate the VLM with either PaliGemma-3B Beyer et al. [2024] or Qwen3-VL-2B Bai et al. [2025a], and use an action expert with hidden dimension 1024. Third-person and wrist images are processed through the native VLM visual pathway, while proprioceptive states and historical action chunks are projected into the token space by lightweight two-layer nonlinear fully connected projectors. Each historical action chunk is represented by eight learned tokens. The matched reactive baseline uses the same backbone, projectors, action expert, optimizer, and data, but sets $K{=}1$ so that no historical triplets are provided.

A.2 Triplet construction and context buffer

For each control step t, the query input contains the instruction, the current multimodal observation, and a bounded history of completed interaction triplets. The current observation comprises the third-person view, left/right wrist views (when available), and the proprioceptive state. For a chunk horizon C, each historical triplet is stored as

$$T_i = \big(\tau_i, \mathcal{O}_i, A_i, \mathcal{O}_{i+C}\big), \tag{8}$$

where $A_i = [a_i, \ldots, a_{i+C-1}]$ is the executed action chunk, projected by g_A into eight learned action tokens. We use $C{=}10$ for LIBERO and SimplerEnv-Bridge, and $C{=}30$ for the real-world experiments.

During training, triplets are sampled only from completed past interactions of the same episode, so each consequence $\mathcal{O}_{i+C}$ strictly precedes the query. To prevent the model from extrapolating the target action from a fixed temporal pattern of neighboring actions, we add a randomized stride of $[0, 15]$ environment steps to the backward spacing.

At inference, the policy maintains a rolling FIFO buffer of completed triplets from the current episode: after each executed chunk, the newly formed triplet is appended and the oldest is evicted once the buffer is full. Triplets are drawn from the buffer with a fixed backward stride for deterministic input layout, and at the start of an episode the model simply conditions on whatever context is available.

A.3 Prompt packing details

VLM input template

```
<|im_start|>user {instruction}
# historical triplets
<|frame_start|> <Third-Person OBS_i> <Wrist OBS_i> <PROPRIO_i>
<|action_start|><ACTION_i><|action_end|>
<Third-Person OBS_{i+C}> <Wrist OBS_{i+C}> <PROPRIO_{i+C}> <|frame_end|>
<|frame_sep|>   ···   <|frame_sep|>
# current query, no target action or consequence
<|frame_start|> <Third-Person OBS_t> <Wrist OBS_t> <PROPRIO_t>
```

B LIBERO-Plus-Hard Specification

B.1 Perturbation overview

LIBERO-Plus-Hard extends LIBERO-Plus with two diagnostic perturbation families targeting factors that are hard to identify from a single frame but directly govern the action-to-observation mapping: camera geometry and robot calibration. We denote them Camera† and Rob. Calib† in the result tables. Task semantics, objects, and language instructions are unchanged.

For each rollout, the perturbation parameters are sampled once at the start of the episode and held fixed throughout. All methods share the same task initializations, seeds, and sampled perturbation instances, and the success criterion follows the original LIBERO task-completion signal. The latent perturbation is not exposed to the model but can be inferred from observation–action–consequence triplets collected during the current episode.

B.2 Multi-camera shift

The multi-camera shift perturbs the visual sensing interface while leaving task, object layout, and robot dynamics unchanged. For the third-person view, we sample an episode-level camera transform with azimuth in $[-75°, 75°]$, elevation in $[0°, 15°]$, distance scale in $[1.0, 2.0]$, and endpoint rotations in $[-10°, 10°]$. For the wrist view, we sample a field of view in $[60°, 90°]$ and one of four image transforms (identity, horizontal flip, vertical flip, $180°$ rotation), applied consistently throughout the rollout. Both views remain visually valid but their geometry differs from the nominal LIBERO setting, so the action-to-pixel mapping must be re-identified online.

B.3 Robot calibration shift

The robot calibration shift perturbs the mapping from commanded actions to the motion executed by the simulator, simulating systematic calibration error, actuation bias, or mechanical offset. At the start of each rollout, we sample a fixed calibration bias and apply it to every absolute end-effector command before stepping the environment. Under the default moderate setting, the translational bias is sampled independently along each Cartesian axis from $[-10, 10]$ mm, and the rotational bias is a fixed axis-angle offset with magnitude up to $3°$; the gripper command is unchanged. We only collect the replayed trajectories that successfully complete the task.

The bias is unobservable from any single frame but recoverable from completed triplets, since the residual between the commanded action stored in context and the realized next observation directly reveals it.

C Experimental Protocol

C.1 Training details

Training data construction. For the standard LIBERO and SimplerEnv-Bridge experiments, we use the official LIBERO demonstrations and BridgeData V2 trajectories, with action targets converted to absolute end-effector control and a chunk horizon of $C{=}10$.

Table 4: Training hyperparameters. Batch size is reported per GPU; all reported runs use 8 H20 GPUs.

Setting	Steps	Batch/GPU	K
LIBERO reactive baseline	90k	32	1
LIBERO Reflective VLA	50k	4	8
LIBERO-Plus / Hard reactive	50k	32	1
LIBERO-Plus / Hard Reflective VLA	90k	4	8
SimplerEnv-Bridge reactive	80k	32	1
SimplerEnv-Bridge Reflective VLA	160k	6	4

For LIBERO-Plus, we use the released LIBERO-Plus training set and convert the trajectories to absolute end-effector control via replay. For LIBERO-Plus-Hard, we follow the same replay-based generation pipeline on the original LIBERO demonstrations, additionally injecting our two diagnostic perturbations during replay: camera-pose shifts and commanded-action biases. We merge the two sets and train a single model on the union, yielding 43,546 trajectories in total.

Optimization. All models are trained with AdamW (β_1=0.9, β_2=0.95, weight decay 0.01, gradient clipping at norm 1.0) under bf16 mixed precision with DeepSpeed ZeRO-2. The action expert and projectors use a peak learning rate of 10^{-4}; the VLM uses a $0.1\times$ multiplier and is held frozen for the initial period in Table 4, after which it follows the same linear warm-up and cosine decay schedule as the rest of the model.

For Reflective VLA on LIBERO/LIBERO-Plus we use K=8 context frames, and K=5 on SimplerEnv-Bridge. Adjacent context frames are separated by one action-chunk stride plus a random gap, sampled from $[0, 15]$ environment steps for LIBERO-family experiments and $[0, 5]$ for SimplerEnv-Bridge; at evaluation the gap is set to zero for deterministic fixed-stride selection.

C.2 Evaluation protocol

All models are evaluated from a fixed checkpoint without test-time fine-tuning, using the same pipeline for the reactive baseline and Reflective VLA. For Reflective VLA, the context buffer is reset per episode and populated online from the policy's own executed chunks and reached observations.

LIBERO. We evaluate the four official suites (*Spatial*, *Object*, *Goal*, *Long*) under absolute end-effector control, with 50 rollouts per task (seed 42) and a horizon of 800 steps (900 for *Long*). Success follows the environment's task-completion signal.

LIBERO-Plus and LIBERO-Plus-Hard. For LIBERO-Plus we evaluate the seven standard perturbation categories with one rollout per task; for LIBERO-Plus-Hard we additionally evaluate the multi-camera shift and the robot-calibration shift (default moderate level) on the LIBERO base suites. Perturbations are generated deterministically from task and rollout ids, so all methods see identical instances.

SimplerEnv-Bridge. We evaluate the four WidowX tasks (spoon on towel, carrot on plate, cube stacking, eggplant in basket) with 24 rollouts per task, reporting per-task success and the unweighted average.

All tables report success rates in percent, averaged over the reported suites or perturbation categories.

C.3 More qualitative results

In Figure 6, we show the qualitative results of Reflective VLA on LIBERO-Plus-Hard dataset.

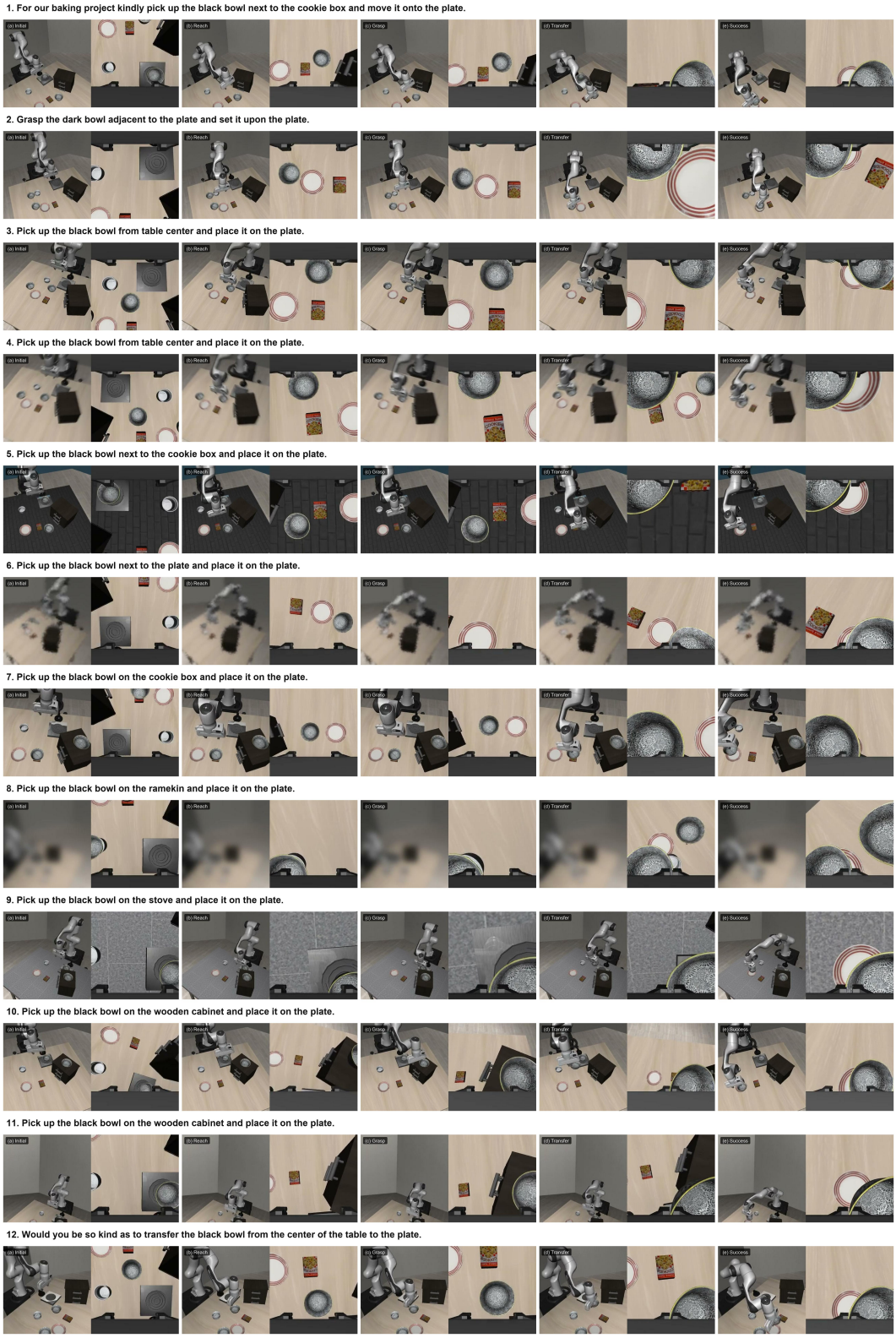


Figure 6: Qualitative results of reflective VLA on LIBERO-Plus-Hard dataset.

D Real-World Protocol

D.1 Hardware and control interface

We use a tabletop setup with an Agilex Piper arm and Intel RealSense D435i cameras. Each policy receives the language instruction, the current third-person view, and the proprioceptive state, and predicts chunked delta end-effector pose and gripper commands with horizon $C{=}30$, executed through the same interface for the reactive baseline and Reflective VLA. For Reflective VLA, the context buffer is reset per trial and populated online from executed chunks and their resulting observations; the reactive baseline conditions only on the current observation and instruction. All hardware, controller, camera streams, and task initializations are identical across methods, with no test-time fine-tuning or camera-specific calibration.

D.2 Tasks and camera placements

We use two language-conditioned pick-and-place tasks: a *box* task (placement into a large square box) and a *bowl* task (placement into a smaller bowl, requiring tighter spatial alignment). We collect 500 demonstrations across ten third-person camera placements spanning the left, front, and right sides of the workspace. Evaluation uses both seen and held-out placements from the same regions, preserving task semantics and hardware while changing camera-to-robot geometry. For each task and condition, we run five trials per placement ($N = 25$ total) with matched initial object configurations; success is judged by a human (target object inside the container at rollout end).

Statistical significance. We report Wilson 95% confidence intervals in Section 4.3. On the *box* task, the reactive baseline reaches 88% [70.0, 95.8] (seen) and 32% [17.2, 51.6] (held-out), versus 92% [75.0, 97.8] and 76% [56.6, 88.5] for Reflective VLA. On the *bowl* task, the corresponding numbers are 68% [48.4, 82.8] / 16% [6.4, 34.7] for the baseline and 72% [52.4, 85.7] / 64% [44.5, 79.8] for Reflective VLA. On both tasks, the held-out intervals do not overlap, indicating that the cross-camera generalization gain is significant at the 95% level; seen-placement intervals overlap, consistent with the small in-distribution gap.

E Reproducibility, Assets, and Limitations

Reproducibility. We provide the codebase for training, evaluation, and perturbation generation in the supplementary material. Refer to the README for setup instructions, dependency versions, and dataset preparation steps; benchmarks depending on external assets are linked to their official downloads rather than redistributed.

Assets. Our simulated experiments build on public assets from LIBERO, LIBERO-Plus, BridgeData V2, and SimplerEnv. Pretrained vision-language backbones are obtained from their official releases under their respective licenses. The LIBERO-Plus-Hard perturbations are specified procedurally in our evaluation scripts (Section B). Real-world experiments use the hardware described in Section D.

Limitations and Future work. Reflective VLA predicts the first chunk reactively, and assumes consequences are observable within the chunk horizon—delayed effects or contact-rich dynamics may need longer contexts. Three further constraints stem from our compute and data budget: (i) we cap context at $K{=}8$ frames due to training memory; (ii) providing history can induce a shortcut where the policy extrapolates from past action chunks instead of reasoning from the current observation, partially mitigated by frame-boundary tokens and stride randomization; (iii) in-context generalization benefits from data diversity, so scaling training data should yield further gains. Our real-world study is also limited to tabletop cross-camera generalization with few trials per condition.