BEYOND THE CURRENT OBSERVATION: HISTORY-INFORMED ADAPTATION FOR VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In robotic manipulation, the task-relevant scene state is often shaped by prior interactions. A robot must infer prior events, retain relevant evidence, and use it to guide subsequent actions. Most existing Vision-Language-Action (VLA) models, however, implicitly assume that the current observation alone provides sufficient information for action prediction. We introduce HI-Adapter, a lightweight and transferable history-informed adaptation framework that equips pretrained VLA policies to understand and utilize recent interaction history without robot-video pretraining. HI Adaptation comprises three components: a spatiotemporal vision adapter that aggregates multi-view observations across time to contextualize visual evidence; a reconstruction-guided memory adapter that distills task-relevant historical information into compact memory tokens; and an action expert adapter that selectively incorporates relevant memory representations into the action-generation process. Together, they improve temporal grounding and memory-guided manipulation in a 1B-parameter VLA policy while training only 2.2% of its parameters. Our method achieves success rates of 37.6% on RoboMME, outperforming the baselines by 18.7%. It further reaches 19.16% on real-world tasks requiring recent historical context. Its consistent improvements when integrated into $\pi_{0.5}$ demonstrate transferability across VLA architectures. These results establish history-informed adaptation as an effective and parameter-efficient approach for overcoming the reliance of existing VLA policies on the current observation alone.

1 INTRODUCTION

In embodied AI, Vision-Language-Action (VLA) models have emerged as a scalable paradigm for generalist robot learning, unifying visual perception, language grounding, and action generation within a single policy while demonstrating broad transfer across tasks, environments, and embodiments (RT-1, RT-2, OpenVLA, and $\pi_{0.5}$)(Brohan et al., 2023; Zitkovich et al., 2023; Black et al., 2025a). Yet, despite these advances, many VLA policies still derive their decisions primarily from immediate perceptual context. This formulation is effective when the present scene uniquely specifies the task state, but becomes insufficient in the common presence of temporal ambiguity. Visually similar observations may require different actions depending on what occurred earlier. Humans resolve such ambiguity by grounding instructions in recent events rather than in a single snapshot. Likewise, correct robot actions may require interpreting the current scene in relation to recent interaction history.

Although expanding the temporal context is an intuitive response, it does not resolve the underlying inference problem. Under temporal partial observability, task-relevant evidence may be sparse and transient, while naive context expansion can dilute the information that matters. Robotic manipulation therefore requires more than processing additional frames. A policy must determine which past event changed the latent task state, consolidate the relevant evidence into a reusable internal representation, and use that representation to condition action generation. This requirement goes beyond mainstream video understanding, which primarily aims to recognize, localize, or describe events, rather than determine which historical evidence should influence robot action control. More importantly, learning these capabilities from scratch is costly, while large-scale robot trajectories

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.

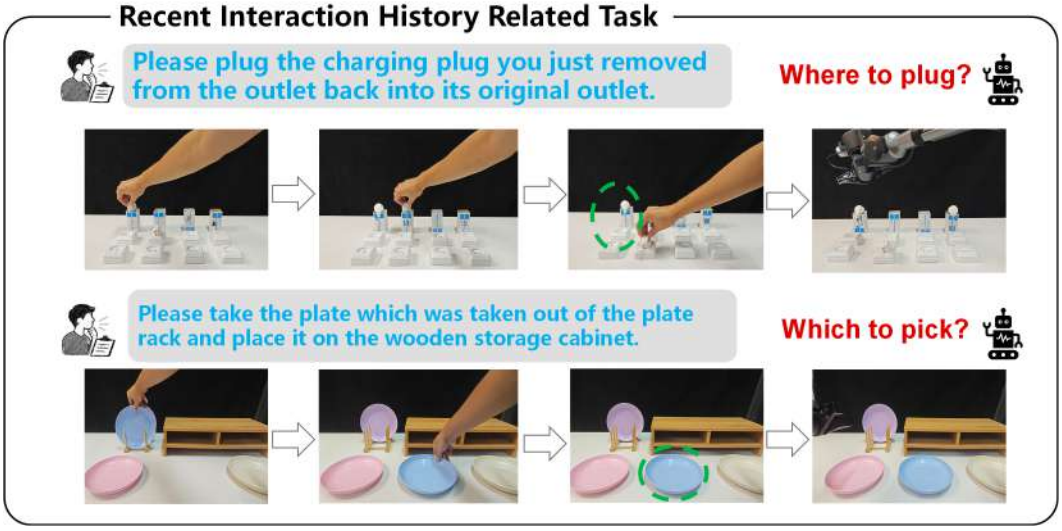


Figure 1: **Temporal grounding through recent interaction history.** Two manipulation tasks illustrate the need to ground instructions in past interactions. Top: returning a charging plug to its original outlet. Bottom: moving the plate previously removed from the rack onto the wooden cabinet. Although the current scene reveals the available objects and locations, it does not resolve these history-dependent references. Green dashed circles highlight the target outlet and plate, respectively.

with explicit history dependence remain scarce. These considerations lead to a concrete research question: ***How can pretrained VLA policies be adapted in a lightweight manner to understand, retain, and utilize recent interaction history for action generation?***

Inspired by adapter-based parameter-efficient adaptation in computer vision (Yang et al., 2023), we propose **History-Informed Adaptation (HI Adaptation)**, a parameter-efficient framework that equips pretrained VLA models with temporal perception and short-term memory for action generation. HI-Adaptation framework comprises two components: **(a) Spatiotemporal Adapter.** We extend a frozen single frame Transformer by reusing its pretrained self-attention along the temporal dimension. The proposed lightweight query-based adapters modulate both spatial and temporal attention outputs, complemented by parallel adapter branches alongside the frozen MLPs. This design adapts pretrained visual representations without training a separate temporal backbone from scratch. **(b) Short-Term Interaction Memory Module.** This module recurrently aggregates the resulting spatiotemporal representations. The feature reconstruction and temporal-order supervision encourage the memory to preserve interaction content and relative chronology. Lightweight cross-attention allows the action policy to access this memory, while policy-side query-based adapters modulate its attention and feed-forward computations to support history-conditioned action generation. Together, these components integrate temporal interaction modeling, memory retention, and history-conditioned action generation.

We evaluate the effectiveness and efficiency of HI Adaptation in simulation and on real robots. Our evaluation includes six manipulation tasks that depend on recent interaction history and closely reflect real-world workflows in biological laboratories, households, and industrial settings. We further evaluate HI Adaptation on the RoboMME benchmarks (Dai et al., 2026) and assess its applicability across different VLA backbones. The results demonstrate improved performance on history-dependent tasks with parameter-efficient updates across the evaluated backbones, supporting the applicability of the framework beyond a single policy architecture. Our main contributions are summarized as follows:

- We propose HI Adaptation, a parameter-efficient framework that equips pretrained frame-based VLAs with spatiotemporal modeling and history-conditioned action generation through query-based adaptation, while keeping the pretrained backbones frozen.

- We introduce recurrent interaction memory with task-oriented and action-oriented slots, using feature reconstruction and temporal-order supervision to preserve relevant interaction evidence while reserving capacity for action-driven learning.

- We evaluate HI Adaptation on RoboMME and six real-world manipulation tasks, demonstrating improved average performance across RDT-1B and $\pi_{0.5}$ and supporting its applicability beyond a single backbone or simulation setting.

2 Related Work

Vision-Language-Action Models. Vision-language-action (VLA) models have emerged as a scalable paradigm for generalist robot learning by mapping visual observations and language instructions to robot actions (Zhang et al., 2025). Early work demonstrated the potential of scaling robot policies with increasingly diverse data and pretrained multimodal representations. RT-1 demonstrated multi-task generalization from large-scale real-robot demonstrations, while PaLM-E and RT-2 further incorporated pretrained vision-language representations into embodied reasoning and control (Brohan et al., 2023; Driess et al., 2023; Zitkovich et al., 2023). Open X-Embodiment and RT-X extended this scaling paradigm across heterogeneous robot embodiments, and OpenVLA provided an open-source generalist VLA trained on large-scale cross-embodiment data (Open X-Embodiment Collaboration et al., 2024; Kim et al., 2025). In parallel, action generation has evolved from autoregressive prediction of discretized action tokens toward continuous generative modeling. Octo employs diffusion-based action decoding to model multimodal continuous action distributions (Ghosh et al., 2024), while RDT-1B scales this paradigm with a diffusion Transformer trained across heterogeneous robotic embodiments (Liu et al., 2025). More recently, π_0 and $\pi_{0.5}$ augment pretrained vision-language backbones with flow-matching action experts, enabling the generation of continuous action chunks for dexterous and generalist manipulation (Black et al., 2025b;a). However, many existing VLA policies typically condition only on the current observation or a short, fixed frame stack, leaving task-relevant past events unmodeled and action prediction ambiguous.

Parameter-Efficient Adaptation. Despite the strong generalization capabilities of foundation models (FMs), adapting them to downstream scenarios through full-parameter fine-tuning requires substantial computation and may compromise previously acquired knowledge(Li et al., 2025). Parameter-efficient fine-tuning (PEFT) addresses this issue by updating only a small subset of parameters, while keeping most pretrained weights frozen(Han et al., 2024). Representative approaches include lightweight adapters that insert trainable modules into frozen networks (Houlsby et al., 2019), low-rank adaptation (LoRA) that parameterizes weight updates with low-rank matrices (Hu et al., 2022), and prompt-based methods that introduce compact trainable representations (Li & Liang, 2021; Lester et al., 2021). These techniques have increasingly been adopted and further specialized for VLA models, enabling efficient adaptation to downstream tasks and robot embodiments while reducing the cost(Wang et al., 2026; Römer et al., 2026; Jiang et al., 2026). However, most existing approaches focus on specializing what a pretrained VLA can accomplish under its original observation structure. This motivates adapting a frame-based VLA to extend its temporal modeling capability without extensively modifying the original model.

Memory-Augmented Policies. To address the limited temporal context of frame-based VLAs, recent methods have explored different ways of incorporating historical observations into action prediction. ContextVLA compresses multi-frame history into a single context token, while HAMLET summarizes each observation into compact moment tokens and aggregates them through a temporal memory module (Jang et al., 2025; Koo et al., 2026). Although computationally efficient, compressing historical observations into a small number of tokens may lose detailed information from earlier observations. MemoryVLA instead maintains an explicit perceptual-cognitive memory bank with retrieval and consolidation mechanisms (Shi et al., 2026), enabling longer-term information storage but making temporal reasoning dependent on memory selection. ReMem-VLA propagates historical information through frame- and chunk-level recurrent queries(Li et al., 2026), although earlier information must be continually preserved within compact recurrent states. In contrast, our work treats history integration as a lightweight adaptation problem and introduces functional adapters across visual perception, temporal memory, and action generation.

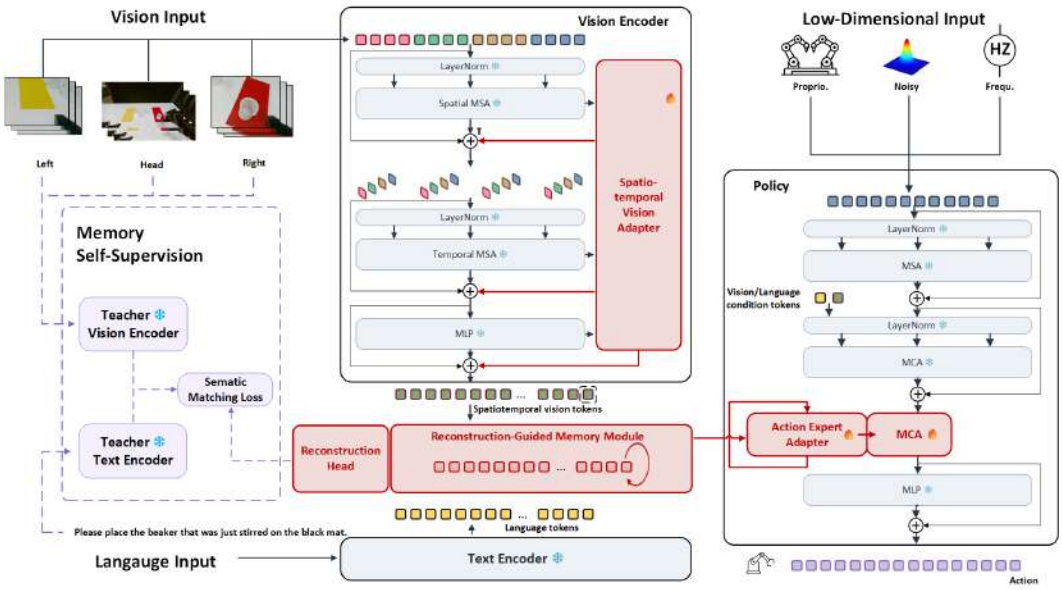


Figure 2: Overview of HI-Adaptation. Recent multi-view observations are processed by a frozen vision encoder augmented with trainable spatiotemporal adapters to produce history-aware visual tokens. A reconstruction-guided memory module recurrently consolidates these representations into a compact memory of past interactions. During training, a reconstruction head supervises the memory through a semantic matching loss guided by frozen vision and text teachers.

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

We propose **HI Adaptation**, a parameter-efficient framework that extends pretrained frame-based VLA models with temporal modeling and history-conditioned action generation. At time t, the robot has observed a sequence of multi-view images $\mathcal{O}_{1:t} = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$ and receives a task instruction c. The visual backbone with temporal modeling encodes these observations into spatiotemporal hidden representations $\mathbf{Z}_{1:t}$. The current-time spatiotemporal tokens $\mathbf{Z}_t$ and instruction embedding $\mathbf{e}_c$ form the conditioning representation $\mathbf{C}_t = (\mathbf{Z}_t, \mathbf{e}_c)$, while historical spatiotemporal features are aggregated into an interaction-memory representation $\mathbf{M}_t$. Together with the robot's proprioceptive state $\mathbf{s}_t$, these representations condition the policy to generate an H-step action chunk:

$$\widehat{\mathbf{A}}_t = (\hat{\mathbf{a}}_t, \ldots, \hat{\mathbf{a}}_{t+H-1}) \sim \pi_{\theta,\phi}(\cdot \mid \mathbf{C}_t, \mathbf{M}_t, \mathbf{s}_t), \tag{1}$$

where θ and ϕ denote pretrained and adaptation parameters, respectively.

In this non-Markovian observation setting, accurate manipulation requires recovering task-relevant state information from past interactions. Our goal is therefore to equip observation embeddings with spatiotemporal modeling capabilities and strengthen historical interaction representations, enabling the policy to use both temporally contextualized observations and retained interaction evidence. To this end, HI-Adaptation comprises two components: **(a) Spatiotemporal Perception Adaptation**, which models inter-frame dependencies within the pretrained visual backbone, and **(b) Interaction Memory and Policy Adaptation**, which retains historical interaction information and integrates it into action generation. Figure 2 illustrates the overall workflow.

To assess the transferability of the adaptation design across policy architectures, we instantiate HI-Adaptation with RDT-1B (Liu et al., 2025) and $\pi_{0.5}$ (Black et al., 2025a). We use RDT-1B as the illustrative backbone below.

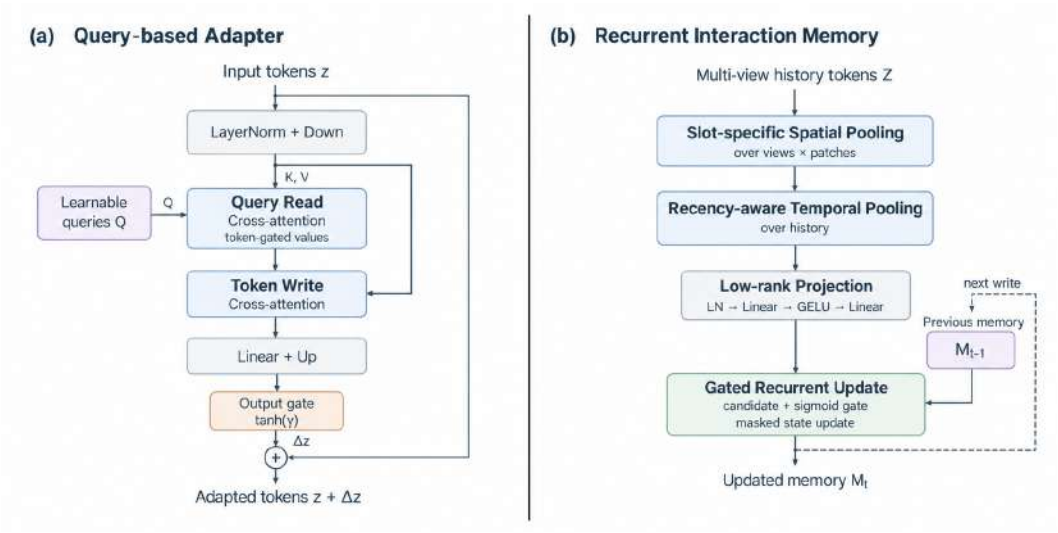


Figure 3: Overview of the two key modules. (a) The query-based adapter refines input tokens by using a small set of learnable queries to read task-relevant information through cross-attention, write the aggregated information back to the token sequence, and produce an adapted residual update. (b) The recurrent interaction memory compresses multi-view history tokens through slot-specific spatial pooling and recency-aware temporal pooling, projects them into compact update tokens, and recurrently updates the memory state via a gated mechanism conditioned on the previous memory.

3.2 Spatiotemporal Perception Adaptation

To move beyond snapshot-based perception, we first consider how to equip frame-based VLAs with temporal modeling. Directly concatenating historical visual tokens increases policy-side computation (Koo et al., 2026), while indiscriminate temporal pooling can obscure interaction-specific cues. Learning a dedicated video backbone introduces further training costs (Yang et al., 2023). Inspired by (Yang et al., 2023), we instead introduce lightweight query-based adapters that extend a frozen image encoder to spatiotemporal modeling. The key is to integrate history into visual representations before action generation, rather than merely lengthen the policy's input sequence.

Query-based adapter. Our adapter uses a query bottleneck to aggregate context and redistribute it as token-specific corrections, following the latent read–write principle (Jaegle et al., 2022). As illustrated in Figure 3 (a), input tokens $\mathbf{X}$ are projected to a lower-dimensional space, summarized by learned queries $\mathbf{Q}$, and updated through a gated residual:

$$\begin{aligned} \mathbf{S} &= \mathrm{Attn}_{\mathrm{read}}(\mathbf{Q}; P_{\downarrow}(\mathbf{X})), \\ \mathcal{A}(\mathbf{X}) &= \tanh(\gamma)\, P_{\uparrow}(\mathrm{Attn}_{\mathrm{write}}(P_{\downarrow}(\mathbf{X}); \mathbf{S}))\,. \end{aligned} \tag{2}$$

Here, $\mathrm{Attn}(Q; C)$ denotes cross-attention from queries Q to context C, and $P_{\downarrow}$ and $P_{\uparrow}$ reduce and restore feature width; normalization and value gating are implicit. The layer-local summaries $\mathbf{S}$ mediate cross-token interactions, while residual injection retains the pretrained pathway and token layout.

Insertion and temporal aggregation. Within a causal context of T frames, each block first applies spatial attention to $\mathbf{X}_\ell$ within individual views, then reuses the same frozen attention weights across time at corresponding view–patch positions. Query-based adapters modulate both attention outputs, while a parallel MLP-side adapter jointly aggregates context across time, views, and patches. This progression couples spatial appearance with inter-frame relations without updating the pretrained image backbone. After the final block, we retain only the current-time spatiotemporal tokens $\mathbf{Z}_t \in \mathbb{R}^{V \times N \times D_v}$, flatten their view and patch dimensions, and project them into the policy's visual-conditioning pathway.

3.3 Interaction Memory and Policy Adaptation

This component combines a Short-Term Interaction Memory Module with policy adaptation to retain and exploit recent key interactions. Memory injection alone does not ensure useful retention, since action loss provides only indirect, sparse supervision of which historical information should be preserved. Hence, we combine explicit supervision of memory content and temporal order with lightweight policy adaptation.

Recurrent Interaction Memory. As illustrated in Figure 3 (b), learned spatial and temporal pooling aggregates recent spatiotemporal representations into slot-wise updates, which are integrated with the previous memory through a gated recurrent update. The resulting fixed-size memory $\mathbf{M}_t \in \mathbb{R}^{J \times D_r}$ can retain interaction evidence beyond the visual context window.

We partition $\mathbf{M}_t = [\mathbf{M}_t^{\text{task}}; \mathbf{M}_t^{\text{act}}]$ into *task-oriented* and *action-oriented* slots, allocating eight slots to each group. Task-oriented slots receive explicit supervision for interaction content and chronology: a shared reconstruction trunk with view-specific heads predicts frozen SigLIP features of past observations selected by image–instruction similarity (Zhai et al., 2023), while a scalar order head learns their relative temporal order. In contrast, action-oriented slots receive no direct reconstruction or order targets, reserving capacity for control-relevant information not captured by the instruction-based selection. Both groups contribute to action generation and receive action supervision. This division encourages retention of instruction-relevant evidence without constraining the entire memory to reconstruct teacher-selected observations.

Memory-Conditioned Policy Adaptation. We use a two-stage attention pathway to integrate memory into the frozen policy. Layer-specific learned queries $\mathbf{Q}_\ell^\tau$, conditioned on diffusion timestep τ, first read from a low-dimensional projection $\mathbf{H}_t$ of the memory. The response $\mathbf{R}_\ell$ is combined with these queries through a gated residual, refined by a lightweight feed-forward block, and projected to the policy width to form $\mathbf{B}_\ell$. Policy tokens $\mathbf{X}_\ell$ then retrieve information from this readout:

$$\mathbf{R}_\ell = \mathrm{CA}(\mathbf{Q}_\ell^\tau, \mathbf{H}_t, \mathbf{H}_t), \qquad \mathbf{X}'_\ell = \mathbf{X}_\ell + \tanh(\eta_\ell)\,\mathrm{CA}(\mathbf{X}_\ell, \mathbf{B}_\ell, \mathbf{B}_\ell). \tag{3}$$

Here, $\mathrm{CA}(Q, K, V)$ denotes cross-attention, with normalization omitted for clarity. The first operation summarizes memory for the current denoising stage; the second retrieves historical evidence according to the policy's current hidden state. The memory-conditioned residual enters before the native MLP, preserving the existing visual and language conditioning. Finally, the query-based adapters from Section 3.2 modulate the policy's self-attention and native cross-attention outputs, together with a parallel MLP-side branch, for parameter-efficient adaptation.

3.4 Training Objectives

To encourage memory to retain key interaction evidence, we introduce SigLIP-guided auxiliary supervision through feature reconstruction and temporal-order prediction. These objectives explicitly constrain both the content of retained interactions and their temporal relationships.

Teacher and causal target selection. A frozen SigLIP model scores observations preceding time t by cosine similarity between instruction and anchor-view image embeddings. A bounded set of selected observations provides normalized multi-view features $\mathbf{y}_{k,v}$ and observation indices u_k to supervise memory content and chronology. Instruction similarity serves as a proxy for task relevance, directing memory toward potentially useful interaction evidence without additional event annotations.

Loss functions and overall objective. Feature reconstruction encourages task-oriented memory slots to preserve the content of selected interactions. We match their normalized predictions $\widehat{\mathbf{y}}_{j,v}$ to teacher targets using Hungarian assignment (Kuhn, 1955) based on multi-view cosine costs. Denoting the slot assigned to target k by $\pi(k)$, we minimize

$$\mathcal{L}_{\text{rec}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_v m_{k,v} w_v \left(1 - \langle \widehat{\mathbf{y}}_{\pi(k),v}, \mathbf{y}_{k,v} \rangle\right)}{\max(1, \sum_v m_{k,v} w_v)}, \tag{4}$$

where K is the number of matched targets, $m_{k,v}$ indicates view validity, and w_v weights each view. Content-based matching allows slots to retain different interactions without prescribing their identities.

To preserve chronology rather than content alone, an order head predicts a scalar score s_j from each task-oriented slot. Reusing the content-based assignment, we encourage later observations to receive larger scores:

$$\mathcal{L}_{\text{ord}} = \frac{1}{|\mathcal{P}|} \sum_{(k,k') \in \mathcal{P}} \max\big(0, \delta + s_{\pi(k)} - s_{\pi(k')}\big), \tag{5}$$

where $\mathcal{P} = \{(k, k') : u_k < u_{k'}\}$ contains chronologically ordered pairs of matched targets and δ is a positive margin. Thus, temporal relationships are supervised by observation order rather than slot position.

Action supervision complements these objectives by learning how retained evidence should influence control. We denote the backbone-specific action objective by $\mathcal{L}_{\text{act}}$, using DDPM-based denoising for RDT-1B and flow matching for $\pi_{0.5}$ (Liu et al., 2025; Black et al., 2025a). The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}}, \tag{6}$$

where λ_{rec} and λ_{ord} balance the auxiliary objectives. Reconstruction and order losses directly supervise task-oriented slots, while action supervision shapes both slot groups, leaving capacity for control-relevant information beyond the teacher-selected targets.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate HI-Adaptation on real-world tasks and simulation benchmark requiring temporal grounding in recent interaction history. Our real-world evaluation comprises six tasks across three application domains. In biological laboratory scenarios, *Contamination Trace* requires tracing direct and secondary contamination among visually identical test tubes and isolating those with the requested contamination status; *Stir Memory* requires selecting a beaker by its stirring frequency or order after rearrangement. In household scenarios, *Occluded Object Identity* requires identifying a plate by its prior occlusion or initial position after the plates exchange locations; *Tray Return by Event* requires retrieving and relocating the plate previously removed from a rack among identical distractors. In industrial scenarios, *Inspected Part Packaging* requires packaging the previously inspected part after it has been returned to a tray; *Plug Restoration* requires reinserting a previously removed plug into the designated socket. Across these tasks, the correct action depends on prior interactions rather than the final scene alone. In simulation, we evaluate on RoboMME (Dai et al., 2026).

Baselines. We develop HI Adaptation on RDT-1B (Liu et al., 2025) and extend it to $\pi_{0.5}$ (Black et al., 2025a) to assess cross-backbone applicability. Real-world experiments compare each backbone with and without HI-Adaptation, yielding four policy variants. We use the same two end-to-end VLA backbones as baselines in simulation and conduct ablations to examine the contributions of memory-related components and policy adaptation.

Implementation details. For the full RDT-based model, we insert query-based adapters into every visual Transformer block and all 28 policy blocks. The adapters use a bottleneck width of 256 and eight attention heads. The visual spatial-attention, temporal-attention, and MLP-side insertion points use 4, 8, and 8 queries, respectively, while each policy-side adapter uses 8 queries. The visual history buffer retains up to 16 sampled observations, and recurrent memory contains 16 slots with shape $\mathbf{M}_t \in \mathbb{R}^{16\times 2048}$, excluding the batch dimension. Memory readout uses 64 layer-specific learned queries at width 512, with a shared readout core across policy blocks. The pretrained visual, language, and policy backbones remain frozen in the full parameter-efficient configuration.

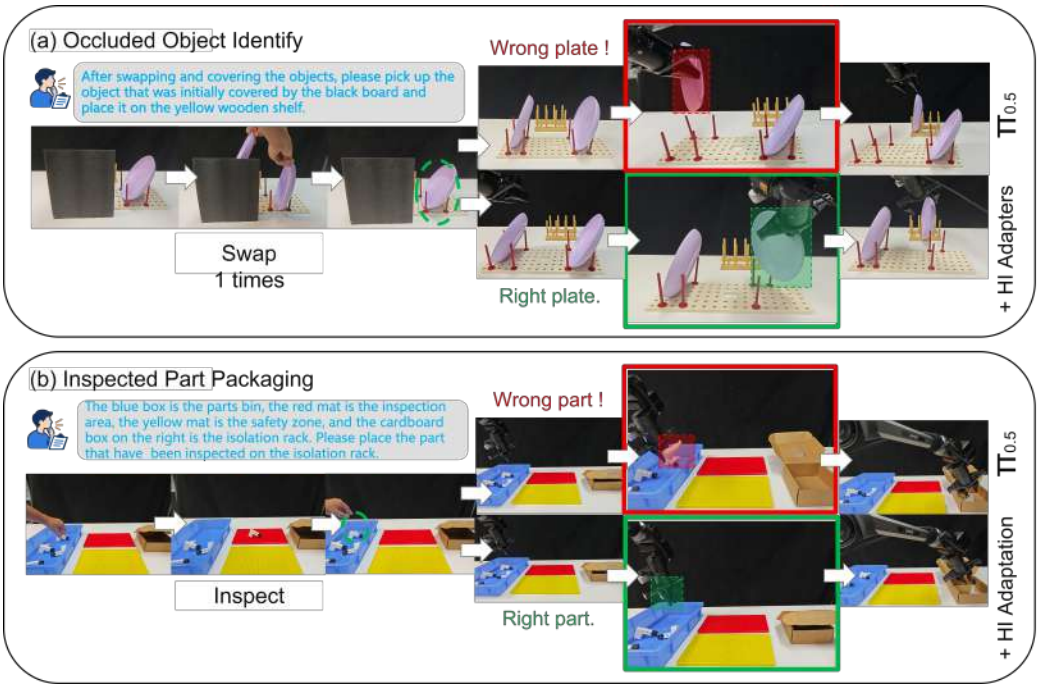


Figure 4: **Example rollouts of real-world tasks.** We provide example rollouts of $\pi_{0.5}$ with H.I. Adaptation and $\pi_{0.5}$ without H.I. Adaptation. With HI Adaptation, the model can locate the correct object; without it, the model remains uncertain which one to grab.

Table 1: **Comparison experiments on RoboMME.**

Model/Task	Counting	Permanence	Reference	Imitation	Avg.
RDT-1B	0.0%	2.5%	1.0%	1.5%	1.3%
+ Spatiotemporal Adaptation	4.0% (+4.0%)	13.0% (+11.5%)	13.0% (+8.5%)	5.5% (+4.0%)	8.9%
+ HI Adaptation	6.5% (+6.5%)	20.0% (+17.5%)	20.5% (+19.0%)	12.0% (+8.5%)	14.8%
$\pi_{0.5}$	25.0%	19.5%	19.0%	12.0%	18.9%
+ HI Adaptation	32.0%(+7.0%)	48.5%(+19.0%)	47.0%(+18.0%)	23.0%(+11.0%)	37.6%

4.2 Main Results

Evaluation on simulation benchmark. We evaluate HI Adaptation on RoboMME to examine its effectiveness on memory-dependent manipulation. Table 1 reports success rates across its four task suites: Counting, Permanence, Reference, and Imitation (Dai et al., 2026). Native RDT-1B rarely completes these tasks, whereas $\pi_{0.5}$ provides a stronger baseline. This gap may reflect differences in both semantic grounding and manipulation priors: RDT-1B conditions its policy on separately encoded visual and language features, while $\pi_{0.5}$ incorporates a pretrained VLM and heterogeneous robot-data training (Liu et al., 2025; Black et al., 2025a). Nevertheless, stronger pretrained priors do not replace access to past interactions when the current observation is insufficient to determine the correct action.

HI Adaptation improves RDT-1B's average success rate by 14.5 percentage points, with particularly pronounced gains of 17.5 and 19.0 percentage points on Permanence and Reference, respectively. Permanence requires maintaining object locations through occlusion and scene changes; for example, VideoUnmaskSwap requires locating a hidden cube after its covering container changes position. Reference requires grounding instructions in earlier observations or interactions: VideoRepick identifies a previously manipulated cube, while VideoPlaceButton resolves a destination relative to an earlier button press (Dai et al., 2026). These demands align with our emphasis on re-

taining interaction content and temporal relationships. The variation across suites is also consistent with RoboMME's finding that different memory designs offer task-dependent advantages rather than uniform improvements (Dai et al., 2026). Applying HI-Adaptation to $\pi_{0.5}$ yields a further average gain of **[X] percentage points**, supporting its applicability beyond the backbone on which it was developed.

Real-World Evaluation. To assess applicability beyond simulation, we evaluate each model variant using a single policy shared across all six real-world tasks, rather than task-specific policies. Each policy is tested in 20 trials per task, and we report task-wise success rates and their average across tasks. Table 2 compares native RDT-1B and $\pi_{0.5}$ with their HI Adaptation counterparts, with representative executions shown in Figure 4.

Both native backbones struggle, achieving average success rates of only **5.83%** and **11.66%**, respectively. These tasks combine history-dependent target selection with precise physical manipulation, and common observed failures include dropped objects and incorrect target selection. Successful execution therefore requires both identifying the appropriate object from interaction history and reliably manipulating it. HI Adaptation improves the average success rates of RDT-1B and $\pi_{0.5}$ by **14.16 percentage points** and **19.16 percentage points**, respectively. These gains across two backbones support the practical value of historical interaction conditioning in workflow-motivated manipulation, extending its benefits beyond standardized simulation tasks.

Table 2: **Real-world evaluation results.**

Model / Task	Contam. Trace	Stir Mem.	Occl. Identity	Tray Return	Insp. Pack.	Plug Rest.	Avg.
RDT-1B	0%	10%	0%	15%	10%	0%	5.83%
+ HI Adaptation	5% (+5%)	25% (+15%)	10% (+10%)	25% (+10%)	20% (10%)	0% (0%)	14.16% (+8.33%)
$\pi_{0.5}$	5%	25%	5%	20%	10%	5%	11.66%
+ HI Adaptation	15% (+10%)	25% (0%)	20% (+15%)	40% (+20%)	15% (+5%)	0% (-5%)	19.16% (+7.5%)

Task abbreviations: Contam. Trace: Contamination Trace; Stir Mem.: Stir Memory; Occl. Identity: Occluded Object Identity; Tray Return: Tray Return by Event; Insp. Pack.: Inspected Part Packaging; Plug Rest.: Plug Restoration.

4.3 More Analysis

Ablation Study on RoboMME. We conduct ablation experiments on two representative robotic VLA backbones, RDT-1B and $\boldsymbol{\pi_{0.5}}$, over four cognitive memory tasks in RoboMME: Counting, Permanence, Reference, and Imitation. For the RDT-1B baseline, the vanilla model achieves only 1.3% average performance across all four tasks, showing poor inherent memory reasoning ability. Adding the spatiotemporal adaptation module brings gain, lifting the average score to 8.9%. This validates that explicit spatiotemporal modeling can partially remedy the weak memory capacity of the base policy. Further incorporating the proposed HI Adaptation yields another performance boost, pushing the average metric to 14.8%. Across all subtasks, we observe consistent absolute improvements: +6.5% on Counting, +17.5% on Permanence, +19.0% on Reference, and +8.5% on Imitation.

5 Conclusion

We presented HI Adaptation, a parameter-efficient framework that enables pretrained frame-based VLAs to model, retain, and utilize recent interaction history. By combining spatiotemporal perception adaptation with recurrent memory and policy adaptation, the method connects historical evidence to action generation beyond the current observation. Reconstruction and temporal-order supervision guide memory toward relevant interaction content and chronology, while preserving capacity for action-driven learning. Experiments on RoboMME and six real-world tasks show improved average performance across RDT-1B and $\pi_{0.5}$, supporting cross-backbone applicability. Fine-grained manipulation remains challenging, highlighting the need to improve historical grounding and physical execution together.

REFERENCES

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert 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 *Proceedings of Robotics: Science and Systems*, 2025b. 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 *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.025.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. RoboMME: Benchmarking and understanding memory for robotic generalist policies. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8m30ogkPk2.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *The Tenth International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.14795.

Huiwon Jang, Sihyun Yu, Heeseung Kwon, Hojin Jeon, Younggyo Seo, and Jinwoo Shin. Contextvla: Vision-language-action model with amortized multi-frame context. *arXiv preprint arXiv:2510.04246*, 2025.

Yuhua Jiang, Junjie Lu, Xinyao Qin, Xiaoyu Chen, Kaixin Wang, Feifei Gao, and Li Zhao. Vla-gse: Boosting parameter-efficient fine-tuning in vla with generalized and specialized experts. *arXiv preprint arXiv:2605.06175*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, 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. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025.

Myungkyu Koo, Daewon Choi, Taeyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. HAMLET: Switch your vision-language-action model into a history-aware policy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KcJ9U0x6kO.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. URL https://aclanthology.org/2021.emnlp-main.243/.

Hang Li, Fengyi Shen, Dong Chen, Liudi Yang, Xudong Wang, Jinkui Shi, Zhenshan Bing, Ziyuan Liu, and Alois Knoll. Remem-vla: Empowering vision-language-action model with memory via dual-level recurrent queries. *arXiv preprint arXiv:2603.12942*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. URL https://aclanthology.org/2021.acl-long.353/.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Ralf Römer, Yi Zhang, Yuming Li, and Angela P. Schoellig. Clare: Continual learning for vision-language-action models via autonomous adapter routing and expansion. *IEEE Robotics and Automation Letters*, 11(7):8331–8338, 2026. doi: 10.1109/LRA.2026.3693992.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=54U3XHf7qq.

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 18638–18646, 2026.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.03024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.15343.

Dapeng Zhang, Jing Sun, Chenghui Hu, Xiaoyan Wu, Zhenlong Yuan, Rui Zhou, Fei Shen, and Qingguo Zhou. Pure vision language action (vla) models: A comprehensive survey. *arXiv preprint arXiv:2509.19012*, 2025.

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 *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023.

A Experimental Details

A.1 Model Details

In our experiments, we evaluate diffusion-based and flow-matching based VLAs: $\pi_{0.5}$(Black et al., 2025a)[1], RDT-1B(Liu et al., 2025)[2]. All checkpoints are obtained from their official repositories.

A.2 Evaluation details

Our evaluations cover both real-world and simulated environments, as illustrated in Figure 5.

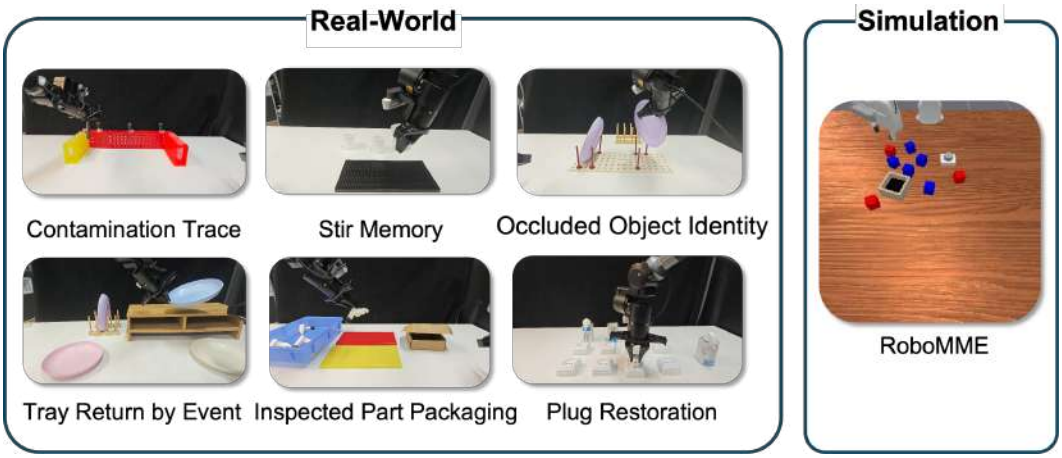


Figure 5: **Evaluation environments.** Left: The real-world evaluation contains six tabletop manipulation tasks that demand reasoning over historical context. On the right, simulated benchmarks RoboMME(Dai et al., 2026), featuring a broad set of manipulation challenges.

Real-robot environment. As visualized in Figure 6, our experimental platform is built on a dual-arm setup with two PiPER X manipulators. The perception system adopts three calibrated Intel RealSense cameras for multi-view scene sensing: each manipulator is fitted with an Intel RealSense D435i camera to capture high-resolution wrist-view observations stored at 640×480, while an Intel RealSense D455 camera is centrally mounted between the two robotic arms to provide a full-scene global view of the tabletop workspace with global RGB streams recorded at 1280×720. The robot control loop operates at 30 Hz. Leveraging this dual-robot physical testbed, we define six practical-oriented tabletop manipulation tasks as depicted in Figure 5: (i) Contamination Trace, where the robot selects the proper container according to past sample-loading records; (ii) Stir Memory, where the robot picks the target beaker leveraging previous stirring operation history; (iii) Occluded Object Identity, where the robot identifies the desired object based on historical occlusion and position observations; (iv) Tray Return by Event, where the robot fetches the correct tableware by referring to past event logs; (v) Inspected Part Packaging by Inspection History, where the robot chooses the qualified part relying on prior inspection history; and (vi) Plug Restoration by Removal History, where the robot completes peg insertion guided by earlier peg-removal records. We collect 70

[1] `gs://openpi-assets/checkpoints/pi05_base`

[2] `https://huggingface.co/robotics-diffusion-transformer/rdt-1b`

Table 3: **Frame statistics of the evaluated real-world tasks.** We compute the mean, maximum, and standard deviation of per-episode frame counts for all real-world datasets. Such metrics help determine an appropriate upper bound on timesteps in evaluation.

Task	Mean #frames	Max #frames	Std #frames
Contamination Trace	431.5	659.0	118.2
Stir Memory	402.5	675.0	63.7
Occluded Object Identity	432.5	579.0	62.5
Tray Return by Event	439.7	785.0	104.0
Inspected Part Packaging	502.6	687.0	60.2
Plug Restoration	358.7	479.0	38.8

demonstration trajectories for each task to train our model, with detailed frame statistics summarized in Table 3. Given that the average trajectory length across all tasks is approximately 428 frames, we set a maximum evaluation horizon of 800 timesteps for each trial. Any episode that cannot finish the task within this timestep limit is categorized as a failure.

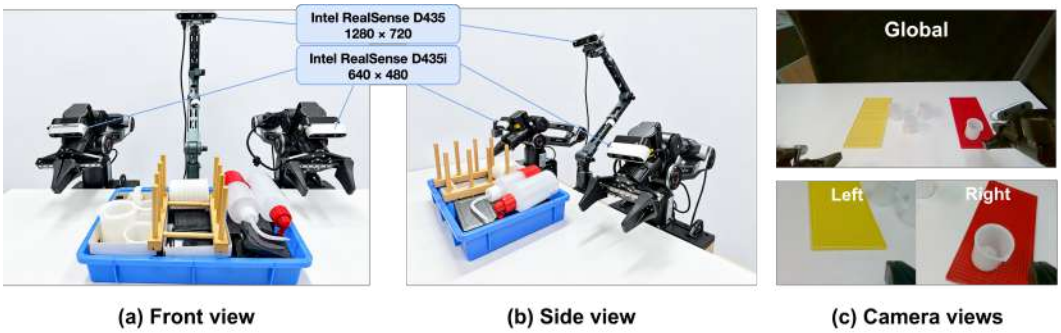


Figure 6: **Real-robot platform.** (a) Front view and (b) side view of the dual-arm setup with Intel RealSense D435/D435i cameras. (c) Multi-view observations: global overhead view and left/right wrist-mounted eye-in-hand views.

Simulation environment. We evaluate our HI Adaptation, which are separately integrated into the $\pi_{0.5}$(Black et al., 2025a) and RDT-1B(Liu et al., 2025) policy backbones, on one standard simulation benchmark, RoboMME (Dai et al., 2026)[3]. We conduct 50 evaluation episodes for all tasks in it. All training and evaluation experiments are implemented on the basis of the officially released codebases of the benchmark. Representative tasks from RoboMME are visualized in Figure 7.

A.3 Training Details

Based on the official training recipes (Liu et al., 2025; Black et al., 2025a), we fine-tune RDT-1B and $\pi_{0.5}$.

- **RDT-1B**: We perform HI Adaptation fine-tuning for 60k iterations with a batch size of 32 and a learning rate of 1e-5.
- $\pi_{0.5}$: HI Adaptation fine-tuning runs for 20k iterations under a batch size of 32 and a learning rate of 2e-5.

[3] https://github.com/RoboMME/robomme_benchmark

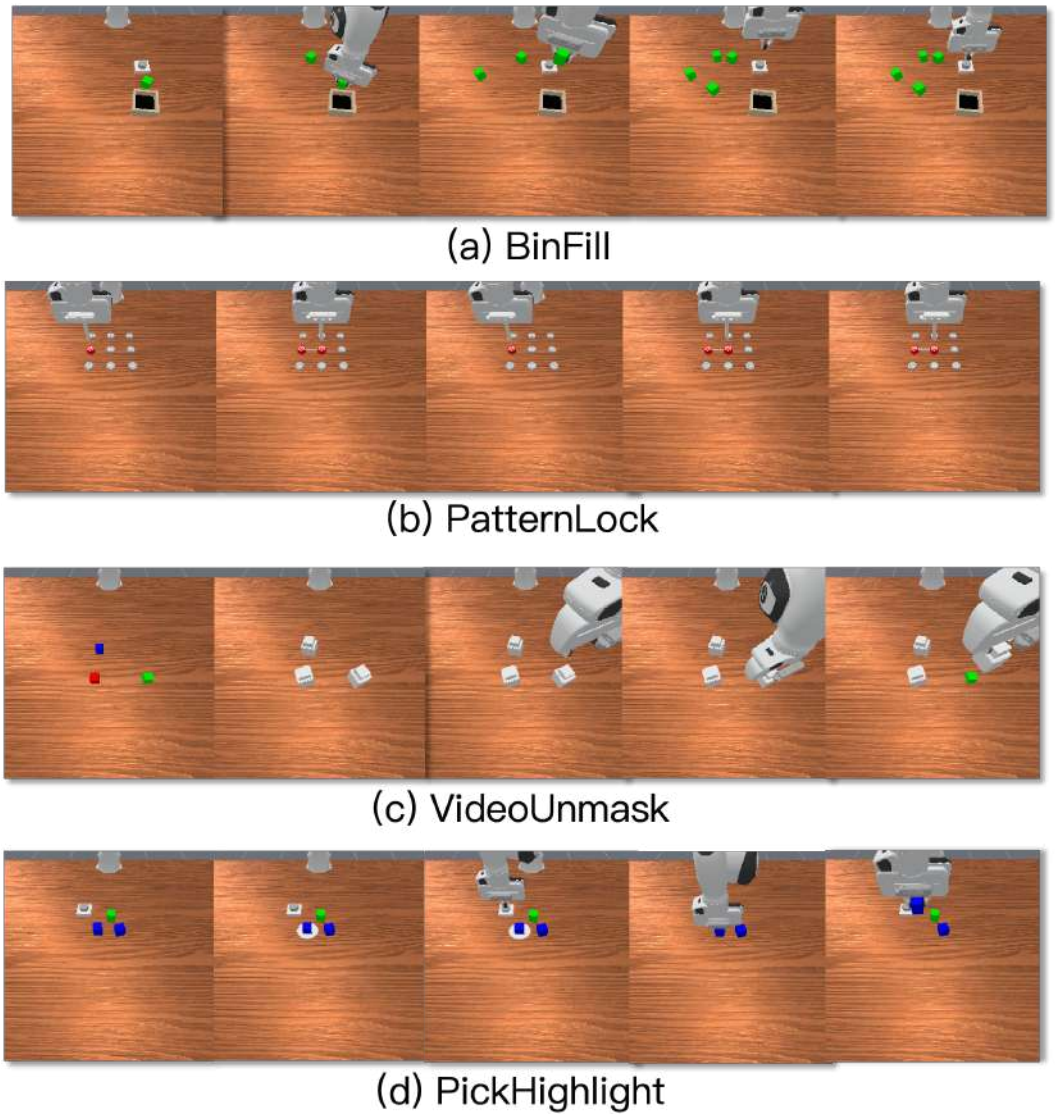


Figure 7: **Representative RoboMME tasks.**

B More Experimental Results

Further example rollouts of real-world tasks. We additionally present rollout examples of the remaining four real-world task in Figures8. As can be readily seen, $\pi_{0.5}$ without HI Adaptation lacks historical perception and therefore exhibits poor performance under memory-centric scenarios; it frequently cannot recover from occlusions or ambiguous intermediate states. By contrast, $\pi_{0.5}$ with HI Adaptation exploits its memory module to accurately accomplish target tasks.

C The use of large language models

This manuscript incorporates the use of large language models during writing and revision stages. The model serves purely as a linguistic optimization tool, helping refine sentence structure, polish academic expressions, and eliminate language errors. All research frameworks, analytical contents, empirical results and final conclusions of this study are entirely original and independently completed by the authors, with no AI-generated research viewpoints or analytical findings adopted.

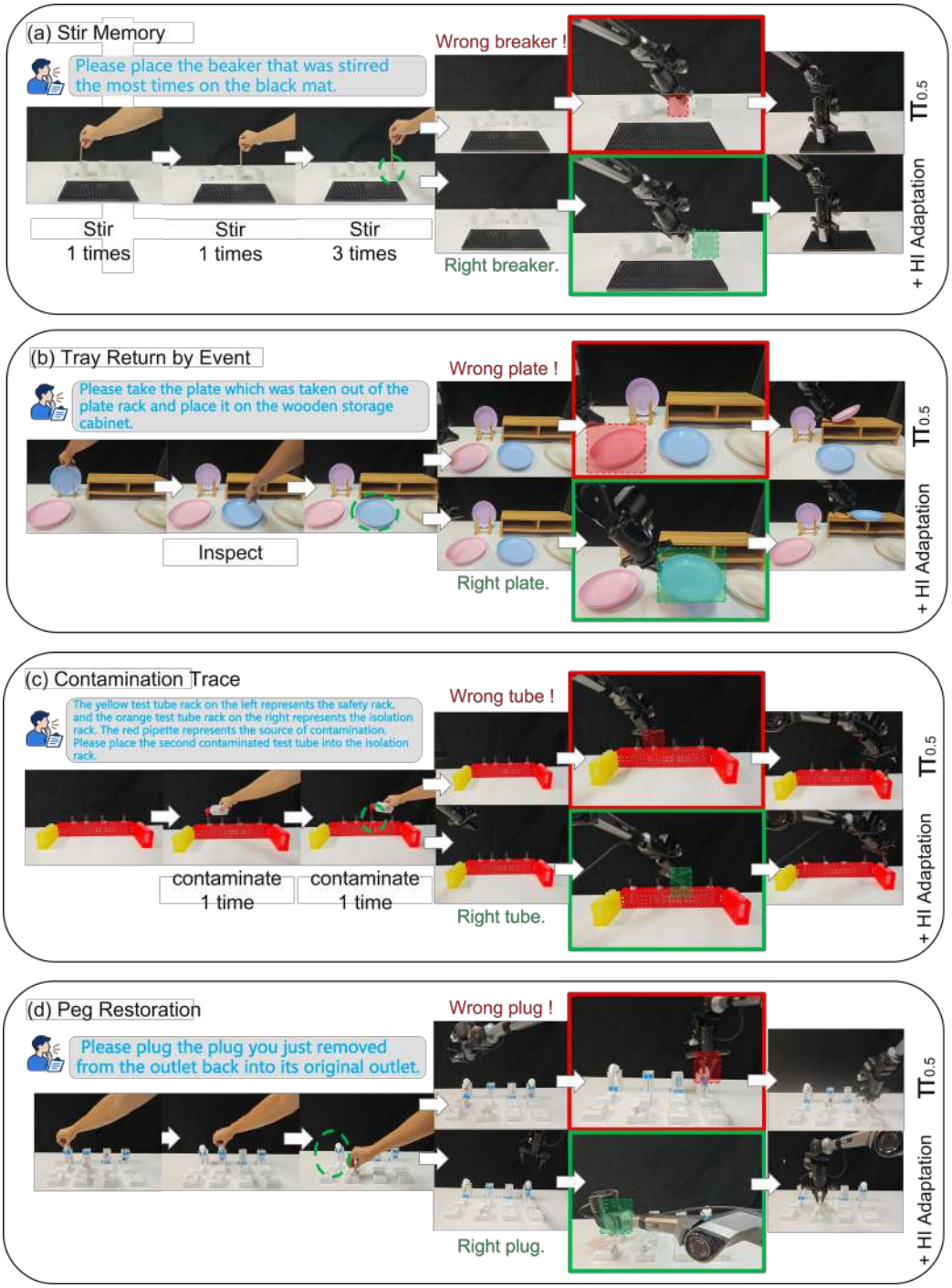


Figure 8: **Example rollouts of the rest four real tasks.**