HYBRID MEMORY-AUGMENTED VISION-LANGUAGE-ACTION MODELS FOR LONG-HORIZON MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models have emerged as a powerful paradigm for closed-loop robot control, yet their limited memory remains a critical bottleneck for long-horizon embodied manipulation. Existing memory-augmented VLAs typically rely on short temporal windows, recurrent memory, or explicit visual memories. These approaches either emphasize recent context, compress away long-range evidence, or retain redundant historical observations, making it difficult to jointly maintain temporal continuity and preserve task-relevant evidence over long horizons. To address these limitations, we propose **HERM-VLA**, a Hybrid Episodic–Recurrent Memory Vision-Language-Action Model that combines recurrent state propagation with explicit episodic evidence retention. The **Recurrent Working Memory** maintains a compact latent state to track task evolution, while the **Episodic Evidence Memory** preserves informative historical context as compact snapshots. An action-aware evidence consolidation mechanism selects action-relevant visual features together with cognitive and spatial context, reducing redundant historical information. The resulting episodic memory is summarized and integrated with the recurrent state and current context for action prediction, enabling effective long-horizon control. Experiments on long-horizon manipulation tasks demonstrate the effectiveness of HERM-VLA and the complementary roles of its two memory components.

1 INTRODUCTION

Vision-Language-Action (VLA) models have emerged as a powerful paradigm for closed-loop robot control by integrating visual perception, language understanding, and action generation (Kim et al., 2025). Recent advances (Brohan et al., 2023; Kim et al., 2025; Black et al., 2024) have broadened the capabilities of VLAs from semantic transfer and generalist manipulation to long-horizon control in complex environments (Nasiriany et al., 2026; Li et al., 2023). For long-horizon control, VLAs may function either as end-to-end control policies or as low-level action policies within embodied agent frameworks (Xu et al., 2026; Hu et al., 2026; Zhang et al., 2026). Regardless of their role, long-horizon tasks often exhibit non-Markovian dependencies (Shi et al., 2026), as appropriate actions may depend on task progression and prior execution outcomes, as well as previously observed objects or visual context that is no longer available in the current observation. Effective action generation therefore requires memory to continuously track the evolving execution state while retaining decision-relevant historical information over extended horizons.

Recent memory-augmented VLAs (Shi et al., 2026; Torne et al., 2026; Jang et al., 2025) have explored different mechanisms for representing and utilizing historical context. Short-horizon visual memory (Torne et al., 2026; Sridhar et al., 2026) retains or compresses a bounded sequence of recent observations to capture short-term temporal context, while information outside the observation window is no longer directly accessible. Recurrent-memory VLAs (Li et al., 2026a; Cherepanov et al., 2026) propagate compact latent states across control steps to accumulate historical context, but information from distant interactions may be progressively attenuated or overwritten as the state is continually updated. Explicit memory buffers (Yang et al., 2026a) instead preserve historical observations, keyframes, or perceptual and semantic representations beyond the recent context window. However, retaining complete observations or keyframes can introduce substantial redundancy,

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.

limiting the ability of explicit memory to compactly preserve decision-relevant information across extended interaction horizons. Consequently, existing memory designs still struggle to combine continuous task-state tracking with selective retention of informative historical context over long-horizon interactions.

Our key insight is that continuous task-state tracking and selective preservation of historical information should be supported by complementary memory mechanisms rather than a single form of memory. To this end, we propose **HERM-VLA**, a Hybrid Episodic–Recurrent Memory Vision-Language-Action model built around two complementary components: **Recurrent Working Memory** and **Episodic Evidence Memory**. Specifically, the Recurrent Working Memory summarizes recent interaction history into a compact latent state that is propagated across decisions. Learned gated updates balance information retained from previous interactions with the current decision context, allowing the state to track task evolution while maintaining temporal continuity.

Complementing the recurrent state, the Episodic Evidence Memory preserves selected historical evidence as compact snapshots in a bounded memory bank. An action-aware evidence consolidation mechanism leverages attention from the action model to select action-relevant visual features, which are stored together with cognitive context and coarse spatial information to form each episodic snapshot. A temporal Qformer summarizer aggregates the stored snapshots into a fixed-size set of memory tokens. Through *Hybrid Memory Integration*, these tokens are processed jointly with the recurrent representation and current visual-language context to guide action prediction. Extensive experiments on long-horizon manipulation benchmarks, including RoboCasa365, demonstrate the effectiveness of HERM-VLA in improving long-horizon task execution. Our main contributions are summarized as follows:

- We propose HERM-VLA, a hybrid episodic–recurrent VLA that couples continuous working-state propagation with explicit historical evidence retention and integrates both forms of memory into action generation.
- We develop an episodic memory mechanism that combines action-guided visual evidence selection with cognitive and spatial context retention, together with temporal aggregation to provide compact historical representations for policy conditioning.
- We conduct extensive experiments on long-horizon manipulation benchmarks, including RoboCasa365, demonstrating the effectiveness of HERM-VLA for long-horizon control.

2 Related Works

2.1 Vision-Language-Action Models

Vision-language-action (VLA) models (Kim et al., 2025; Black et al., 2024; Community, 2026; Li et al., 2026b) extend pretrained vision-language models (Achiam et al., 2023; Ouyang et al., 2022; Liu et al., 2023; Yang et al., 2025; Zhou et al., 2025) to robotic control by leveraging large-scale vision-language knowledge and robot demonstrations (O'Neill et al., 2024). Representative approaches such as RT-2 (Brohan et al., 2023), OpenVLA (Kim et al., 2025), and π_0 (Black et al., 2024) have demonstrated strong generalization across diverse manipulation tasks. More recent approaches (Yuan et al., 2026; Ye et al., 2026b; Tian et al., 2025) further incorporate predictive world modeling into robot policies, coupling future observation prediction with action generation. Despite these advances, long-horizon robotic manipulation is inherently non-Markovian (Chen et al., 2026b; Shi et al., 2026), as effective decision-making depends on temporal dependencies extending beyond the current observation, while existing VLA models remain limited in capturing historical context.

2.2 Memory-Augmented Vision-Language-Action Models

Recently, memory-augmented VLAs (Shi et al., 2026; Li et al., 2026a; Torne et al., 2026; Chen et al., 2026b; Zhou et al., 2026) have emerged to incorporate historical context into long-horizon manipulation. Within this line of work, approaches using short-horizon visual memory (Torne et al., 2026) condition the policy on a limited window of recent observations, while retrieval-based methods (Sridhar et al., 2026; Ji et al., 2026; Lin et al., 2025) selectively retrieve task-relevant observations, keyframes, or prior experiences from interaction history to inform subsequent decisions. Recurrent-memory VLAs (Li et al., 2024; 2026a; Xiao et al., 2026) compress interaction history

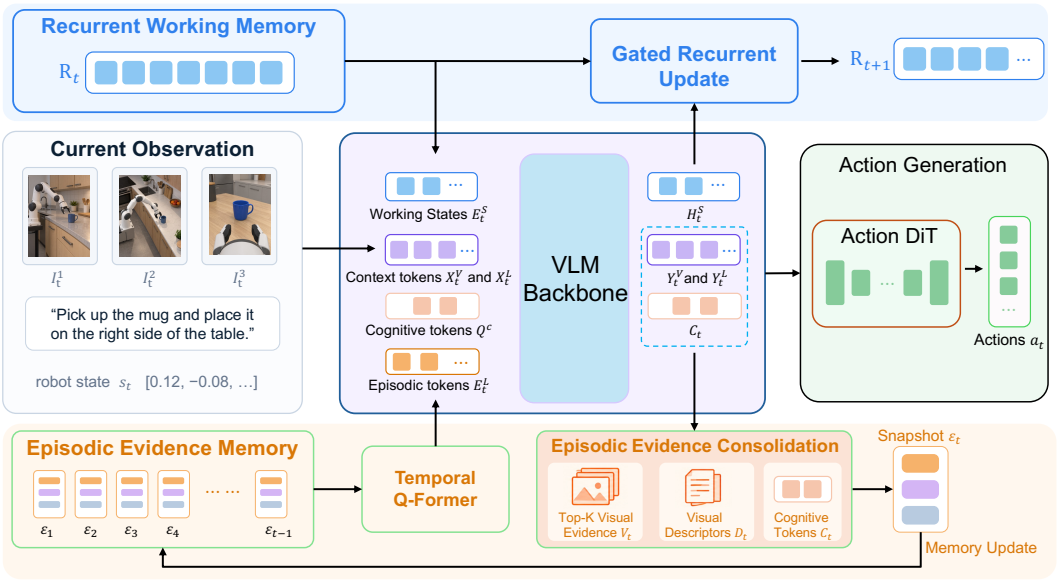

Figure 1: Overview of **HERM-VLA**. The framework combines Recurrent Working Memory and Episodic Evidence Memory to model complementary temporal information for long-horizon manipulation. The recurrent memory propagates a compact latent state across action chunks to maintain recent interaction, while the episodic memory preserves selected historical visual and cognitive tokens. At each decision step, the recurrent state, aggregated episodic evidence, and current observation are jointly incorporated into the vision-language backbone to condition action generation.

into compact latent states that are propagated across control steps. Furthermore, VLAs with explicit memory buffers (Shi et al., 2026; Yang et al., 2026a) preserve selected observations, keyframes, or compact perceptual and semantic representations as task-relevant evidence for subsequent decision-making. Despite these advances, existing memory designs still face a tension between maintaining temporal continuity and retaining task-relevant evidence over long horizons (Yang et al., 2026b). Recurrent memories compress interaction history into compact states, potentially creating an information bottleneck that discards fine-grained historical information (Yang et al., 2026a), whereas explicit memories require selective retention to avoid redundancy and increasing computational overhead (Shi et al., 2026; Yang et al., 2026b). HERM-VLA addresses this tension by coupling recurrent state propagation with selectively maintained episodic evidence, preserving temporal continuity while retaining task-relevant context beyond the effective horizon of recurrent memory.

3 Method

We propose **HERM-VLA**, a Hybrid Episodic–Recurrent Memory Vision-Language-Action model that integrates recurrent working memory with selectively maintained episodic evidence. Specifically, the *Recurrent Working Memory* maintains a compact latent representation for tracking recent task context, while the *Episodic Evidence Memory* preserves task-relevant visual observations and cognitive features from earlier decisions. Through *Hybrid Memory Integration*, the continuously updated working context and retrieved episodic evidence are jointly integrated with the current observation to guide action prediction.

3.1 Problem Formulation

Following prior work (Yang et al., 2026a; Shi et al., 2026), we formulate long-horizon robotic manipulation as a non-Markovian sequential decision-making problem, where action prediction depends on both the current observation and historical task context. At decision step t, the policy receives multi-view camera observations $I_t^{1:n_v}$, a language instruction ℓ, and the robot state s_t, and predicts an action chunk $a_t \in \mathbb{R}^{H \times d_a}$. Here, n_v is the number of camera views, H is the action horizon, and d_a is the action dimension. To capture historical context, HERM-VLA also maintains a recurrent

working state R_t and an episodic evidence bank $\mathcal{B}_t$, representing the *Recurrent Working Memory* and *Episodic Evidence Memory*, respectively. We denote the full policy by π, which consists of a vision-language backbone $f\mathrm{VLM}$ and an Action DiT for action generation. The policy is formulated as

$$a_t = \pi\left(I_t^{1:n_v}, \ell, s_t, R_t, \mathcal{B}_t\right), \tag{1}$$

where R_t provides a continuously evolving representation of recent task context, while $\mathcal{B}_t$ preserves selected historical visual evidence and cognitive context.

At each decision step, HERM-VLA jointly uses the recurrent working state R_t, episodic evidence in $\mathcal{B}_t$, and the current input for action prediction. The recurrent state is then updated with the current decision context, while action-relevant visual features and cognitive features from the current decision are stored in the episodic bank for subsequent decisions.

3.2 Hybrid Memory Representation

Recurrent working memory. To retain recent interaction context and capture task evolution, we introduce *Recurrent Working Memory*. Unlike a short-term frame window (Torne et al., 2026) that explicitly stores past observations, it summarizes recent interaction history into a compact latent state $R_t \in \mathbb{R}^{m\times d_s}$, where m denotes the number of memory tokens and d_s is the dimension of each token. This state is propagated across decisions and continuously updated with the current context. At the beginning of each episode, it is initialized from a set of learnable memory queries R_{init}.

Episodic evidence memory. To complement the compact recurrent state, we introduce *Episodic Evidence Memory*, which preserves action-relevant visual evidence together with cognitive context from earlier decisions. At each decision step i, we construct an *episodic evidence snapshot* $\mathcal{E}_i$, which consists of cognitive tokens C_i, selected visual evidence tokens V_i, and coarse visual descriptors D_i:

$$\mathcal{E}_i = (C_i, V_i, D_i). \tag{2}$$

Here, C_i denotes the cognitive tokens produced by the learned cognitive queries $Q^c \in \mathbb{R}^{n_c\times d}$, V_i contains the action-relevant visual tokens selected from each camera view and D_i denotes coarse visual descriptors that provide complementary global spatial information beyond the selected local visual tokens. The episodic evidence memory $\mathcal{B}_t$ at decision step t is then formed by the retained snapshots from preceding decisions:

$$\mathcal{B}_t = \{\mathcal{E}_i \mid i \in \mathcal{I}_t\}, \qquad \mathcal{I}_t \subseteq \{0, \ldots, t-1\}, \qquad |\mathcal{I}_t| \le K, \tag{3}$$

where $\mathcal{I}_t$ denotes the indices of the retained historical snapshots and K is the maximum number of snapshots stored in the bank.

3.3 Hybrid Memory Integration

In this section, we present the joint process of memory-guided action prediction, recurrent state updating, and episodic evidence consolidation. At each decision, the policy uses R_t and $\mathcal{B}_t$ together with the current input to predict a_t. The current decision also produces an updated working state R_{t+1} and an episodic evidence snapshot $\mathcal{E}_t$, which is inserted into the bank to form $\mathcal{B}_{t+1}$.

Hybrid memory encoding. For the recurrent working memory, we map the recurrent state $R_t \in \mathbb{R}^{m\times d_s}$ to the VLM hidden dimension d using a learned projector:

$$E_t^S = \phi_S(R_t), \tag{4}$$

where ϕ_S denotes MLP projector and $E_t^S \in \mathbb{R}^{m\times d}$ is the resulting recurrent memory representation.

For the episodic evidence memory, we first encode each snapshot in $\mathcal{B}_t$ independently and subsequently aggregate them into a compact representation. For each snapshot $\mathcal{E}_i = (C_i, V_i, D_i)$, we map its three components into a shared embedding space using separate projectors and apply joint self-attention to integrate their complementary information:

$$\widetilde{T}_i = f_{\mathrm{enc}}\left([\phi_C(C_i); \phi_V(V_i); \phi_D(D_i)]\right), \tag{5}$$

where ϕ_C, ϕ_V, and ϕ_D are independent MLP projectors with output dimension d, $[\cdot;\cdot]$ denotes concatenation of the projected tokens, and f_{enc} performs self-attention over the concatenated tokens

within each snapshot. To aggregate the encoded snapshots, we first concatenate those retained in $\mathcal{B}_t$ to form a historical token sequence:

$$M_t^L = \mathrm{Concat}_{i \in \mathcal{I}_t} \left(\widetilde{T}_i \right) \in \mathbb{R}^{N_t^L \times d}, \tag{6}$$

where $\mathcal{I}_t$ denotes the indices of the retained snapshots and N_t^L is the total number of historical tokens. Subsequently, the temporal Q-Former f_{TQ} uses learned queries $Q^L \in \mathbb{R}^{n_L \times d}$ to attend to M_t^L and summarize the historical evidence into n_L episodic memory tokens. To encode the temporal distance of each snapshot from the current decision, temporal rotary position embeddings (RoPE) are applied to the cross-attention keys. The resulting episodic memory representation is

$$E_t^L = f_{\mathrm{TQ}} \left(Q^L, \mathcal{R}_{\mathrm{temp}}(M_t^L) \right) \in \mathbb{R}^{n_L \times d}, \tag{7}$$

where $\mathcal{R}_{\mathrm{temp}}(\cdot)$ denotes temporal RoPE applied to the cross-attention keys according to each snapshot's temporal distance from the current decision. When $\mathcal{B}_t$ is empty, E_t^L is omitted from the subsequent memory integration.

Memory-guided action prediction. Let $X_t^V \in \mathbb{R}^{L_t^V \times d}$ and $X_t^T \in \mathbb{R}^{L_t^T \times d}$ denote the visual and textual representations of the current observations and language instruction, respectively. Together with the recurrent and episodic memory representations E_t^S and E_t^L, we introduce learned cognitive queries $Q^c \in \mathbb{R}^{n_c \times d}$ to extract a compact representation of the current decision context. These representations are concatenated and jointly processed by the VLM:

$$H_t = [H_t^S; H_t^L; Y_t^V; Y_t^T; C_t] = f_{\mathrm{VLM}} \left([E_t^S; E_t^L; X_t^V; X_t^T; Q^c] \right), \tag{8}$$

where H_t^S and H_t^L denote the output representations corresponding to the recurrent and episodic memory tokens, respectively. Y_t^V and Y_t^T denote the contextualized visual and textual representations. C_t denotes the cognitive representations extracted by the learned queries Q^c.

The resulting VLM hidden states H_t serve as the context for the Action Expert. Conditioned on H_t and the robot state s_t, the Action Expert, implemented as a DiT, predicts the action chunk a_t.

Recurrent working memory update. To track task evolution while retaining recent interaction context, we update the working state R_t using the recurrent-token hidden states H_t^S produced by the VLM. We first transform H_t^S using a learned projector. The previous memory tokens in R_t then serve as queries to attend to the projected features, extracting relevant information for each token to form a candidate state $\widetilde{R}_t$. A learned token-wise gate controls how much of the candidate representation is incorporated into the previous state:

$$\begin{aligned} g_{t,j} &= \sigma \left(f_{\mathrm{gate}} \left([R_{t,j}; \widetilde{R}_{t,j}] \right) \right), \\ R_{t+1,j} &= (1 - g_{t,j}) R_{t,j} + g_{t,j} \widetilde{R}_{t,j}, \end{aligned} \tag{9}$$

where j indexes the recurrent memory tokens, f_{gate} is a learned scalar projection, and σ denotes the sigmoid function.

Episodic evidence consolidation. To retain action-relevant visual evidence while reducing memory and computation overhead, we selectively preserve the most informative visual tokens from the current decision. Specifically, we use the cross-attention weights of the Action Expert as a relevance signal for visual evidence selection. Let $V_t^v \in \mathbb{R}^{N \times d}$ denote the visual tokens of camera view v within Y_t^V. We average the cross-attention weights over the action-query positions to obtain a relevance score $\alpha_{t,j}^v$ for each visual token and select the top-k tokens from each view:

$$J_t^v = \mathrm{TopK} \left(\{\alpha_{t,j}^v\}_{j=1}^N, k \right). \tag{10}$$

The selected visual evidence is formed by concatenating the top-k tokens from all camera views:

$$V_t = \mathrm{Concat}_{v=1}^{n_v} \left(V_t^v[J_t^v] \right). \tag{11}$$

In addition to the selected local visual evidence, we retain coarse visual descriptors D_t to provide complementary global spatial information. we extract a coarse visual descriptor grid $D_t^v \in \mathbb{R}^{g^2 \times d_v}$, where g denotes the spatial grid resolution and is set to 2 in our implementation. The descriptors from all views are aggregated as

$$D_t = \mathrm{Stack}_{v=1}^{n_v} \left(D_t^v \right). \tag{12}$$

We also retain the cognitive tokens C_t to preserve the high-level context of the current decision. Together with the selected visual evidence V_t and coarse visual descriptors D_t, they form the episodic evidence snapshot

$$\mathcal{E}_t = (C_t, V_t, D_t). \tag{13}$$

The resulting snapshot $\mathcal{E}_t$ is added to the episodic bank for subsequent decisions. With a capacity of K snapshots, the bank retains episodic evidence over a substantially longer temporal horizon, providing persistent historical context for long-horizon manipulation

3.4 Model training

We train HERM-VLA end-to-end with a flow-matching objective for action generation, together with auxiliary supervision on the recurrent and episodic memory representations.

Flow-matching action objective. Following standard flow-matching VLA formulations, we optimize the Action Expert with a conditional flow-matching objective:

$$\mathcal{L}_{\text{flow}} = \|\widehat{v}_\theta(\cdot \mid H_t, s_t) - v_t^\star\|_2^2, \tag{14}$$

where $\widehat{v}_\theta(\cdot \mid H_t, s_t)$ denotes the predicted velocity field conditioned on the VLM hidden states H_t and robot state s_t, and $v_t^\star$ denotes the flow-matching target. **Auxiliary memory supervision.** To encourage the recurrent working memory to retain information relevant to subsequent action evolution, we aggregate the recurrent hidden states H_t^S into a compact representation r_t using learned attention and predict the final action of the current chunk:

$$\widehat{a}_{t,H} = f_{\text{future}}(r_t). \tag{15}$$

The target $a_{t,H}$ provides auxiliary supervision for learning action-relevant recurrent representations.

For the episodic memory, we introduce an auxiliary progress prediction objective to encourage the episodic representation E_t^L to capture task-stage information.

$$\widehat{p}_t = \sigma\left(f_{\text{prog}}(E_t^L)\right), \qquad p_t = \frac{q_t}{\max(T_e - 1, 1)}, \tag{16}$$

where q_t denotes the within-episode decision index and T_e is the episode length. We optimize the two auxiliary predictions with smooth-L_1 losses, denoted by $\mathcal{L}_{\text{future}}$ and $\mathcal{L}_{\text{prog}}$. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{future}}\mathcal{L}_{\text{future}} + \lambda_{\text{prog}}\mathcal{L}_{\text{prog}}, \tag{17}$$

where λ_{future} and λ_{prog} balance the two auxiliary objectives.

4 Exeperments

In this section, we conduct comprehensive experiments on long-horizon manipulation tasks to evaluate the effectiveness of HERM-VLA. Extensive ablation studies are further conducted to examine the contributions of the Recurrent Working Memory, Episodic Evidence Memory, and action-aware evidence consolidation.

4.1 Experimental Settings

Benchmarks. We conduct experiments on two challenging benchmarks: RoboCasa365 (Nasiriany et al., 2026) for evaluating long-horizon manipulation performance and RMBench (Chen et al., 2026a) for evaluating memory-dependent manipulation capabilities. **RoboCasa365.** This large-scale household manipulation benchmark contains 365 kitchen tasks, including 65 atomic and 300 composite tasks with complex multi-step task structures. It uses a PandaOmron robot with a 12-dimensional action space. We train on the pretraining split and evaluate on the target split across Atomic, Composite-Seen, and Composite-Unseen categories. **RMBench.** This benchmark evaluates memory-dependent robotic manipulation through nine tasks with controlled memory requirements. It includes five tasks requiring single-step historical recall and four tasks requiring multi-step temporal integration, instantiated on the ALOHA-AgileX robot in RoboTwin 2.0. We conduct experiments on all nine tasks to assess memory-dependent manipulation performance.

Table 1: Performance comparison on the RoboCasa365 target split. We report task success rate (%) on Atomic Seen, Composite Seen, and Composite Unseen tasks, together with the average performance across the three evaluation settings.

Model	Atomic Seen	Composite Seen	Composite Unseen	Average
Diffusion Policy (Chi et al., 2025)	15.7%	0.2%	1.3%	6.1%
GigaWorld-Policy 0.1 (Ye et al., 2026a)	44.4%	11.8%	2.9%	20.7%
WorldDreamer (Wang et al., 2024)	66.3%	26.7%	9.0%	35.3%
$\pi_{0.5}$ (Black et al., 2024)	39.6%	7.1%	1.2%	16.9%
GR00T N1.5 (Bjorck et al., 2025)	50.7%	14.8%	2.7%	23.9%
GR00T N1.6 (Bjorck et al., 2025)	51.1%	9.4%	1.7%	21.9%
RLDX-1 (Kim et al., 2026)	67.6%	27.9%	8.5%	36.0%
HERM-VLA-GR00T	**68.6%**	**29.0%**	**11.2%**	**37.6%**

Baselines. We compare our approach with a broad range of competitive baselines spanning standard policies and memory-augmented VLA methods. The standard policy group includes Diffusion Policy (Chi et al., 2025), $\pi_{0.5}$ (Black et al., 2024), GR00T N1.5, GR00T N1.6 (Bjorck et al., 2025), and GigaWorld-Policy (Ye et al., 2026a). For memory-augmented methods, we consider several representative memory paradigms. RLDX-1 (Kim et al., 2026) incorporates multi-frame temporal context together with a history memory module, while MemER (Sridhar et al., 2026) retrieves task-relevant historical keyframes. Mem-0 (Chen et al., 2026a) combines sliding-window, and keyframe memories in a hierarchical policy. MemoryVLA (Shi et al., 2026) maintains an explicit perceptual-cognitive memory bank, whereas EventVLA (Yang et al., 2026a) dynamically preserves sparse keyframes.

Implementation details. We instantiate HERM-VLA with different backbone policies for the two benchmarks. On RoboCasa365, we build **HERM-VLA-GR00T** on top of GR00T N1.5, while on RMBench, we build **HERM-VLA-OFT** on top of Qwen-OFT. All models are implemented in PyTorch and trained on 8 NVIDIA H100 GPUs. For RoboCasa365, we use a batch size of 32 per GPU and train for 200,000 steps. For RMBench, we use a batch size of 8 per GPU and train for 80,000 steps. For both variants, the Recurrent Working Memory contains $m = 16$ memory slots with hidden dimension $d_s = 1536$ and is updated once per action chunk. The Episodic Evidence Memory maintains a bank of $K = 16$ snapshots. Each snapshot stores $n_c = 4$ cognitive tokens, the top-$k = 32$ action-relevant visual tokens from each camera view, and a 2×2 coarse descriptor grid for each view. The temporal Q-Former uses $n_L = 32$ learned queries, 6 transformer blocks, and 16 attention heads, with the relative temporal distance clipped at $R_{\max} = 1024$. We set the auxiliary loss weights to $\lambda_{\text{future}} = 0.02$ and $\lambda_{\text{prog}} = 0.02$.

4.2 Main Results

Robocasa365. As shown in Table 1, HERM-VLA-GR00T achieves the best overall performance on RoboCasa365, reaching an average success rate of **37.6%**. It consistently outperforms prior methods across all three evaluation settings, achieving **68.6%** on Atomic Seen, **29.0%** on Composite Seen, and **11.2%** on Composite Unseen. HERM-VLA also surpasses the strongest baseline, RLDX-1, across all evaluation categories, demonstrating competitive performance on both seen and unseen task compositions. More importantly, compared with its backbone GR00T N1.5, HERM-VLA improves the average success rate from 23.9% to 37.6%, with substantial gains on both Composite Seen (+14.2%) and Composite Unseen (+8.5 %). These results demonstrate that incorporating recurrent working memory and episodic evidence enables the policy to more effectively exploit historical context, particularly for complex long-horizon mobile manipulation tasks.

Table 2: Performance comparison on RMBench. We report the average task success rate (%).

Model	Success Rate
DP	5.8%
$\pi_{0.5}$	10.4%
QwenOFT	5.6%
MemER	8.7%
Mem-0	42.0%
MemoryVLA (QwenOFT)	41.7%
EventVLA	67.8%
HERM-VLA-OFT	**68.6%**

Table 3: We evaluate the contribution of Recurrent Working Memory and Episodic Evidence Memory using task success rate (%).

Recurrent Memory	Episodic Memory	Success Rate
×	×	59.11%
✓	×	66.22%
✓	✓	**68.56%**

Table 4: Recurrent memory update ablation on the Atomic target split. We compare update intervals of one and two action chunks using task success rate (%).

Update Interval	Success Rate
2-chunk interval	64.36%
1-chunk interval	**66.22%**

Table 5: Ablation of episodic snapshot composition on the RoboCasa365 Atomic target split. We compare different combinations of visual tokens, cognitive tokens, and coarse visual descriptors retained in each episodic snapshot. Performance is measured by task success rate (%).

Visual Tokens	Cognitive Tokens	Coarse Descriptors	Success Rate
✓	×	×	64.55%
✓	✓	×	65.78%
✓	✓	✓	**68.56%**

RMBench. As shown in Table 2, HERM-VLA-OFT achieves the best overall performance on RM-Bench among the compared methods. The consistent performance across memory-dependent tasks indicates that the proposed memory design can effectively preserve and utilize task-relevant historical information during manipulation. In particular, the combination of recurrent working memory and episodic evidence memory provides complementary temporal context, enabling the policy to better handle tasks that require reasoning over observations from earlier interaction steps. This demonstrate the effectiveness of HERM-VLA across both long-horizon manipulation and explicitly memory-dependent settings.

4.3 Ablation Studies

We conduct ablation studies on the Atomic target split of RoboCasa365 to examine the contribution of each memory component and the key design choices in HERM-VLA.

Effect of memory components. We first evaluate the contribution of the Recurrent Working Memory (RWM) and Episodic Evidence Memory (EEM). As shown in Table 4.3, introducing recurrent memory substantially improves the success rate from 59.11% to 66.22%, demonstrating the importance of maintaining a continuously propagated representation of interaction history. Incorporating episodic memory further improves the performance to 68.56%. These results suggest that the two memory mechanisms provide complementary benefits. Recurrent memory captures continuously evolving task context, while episodic memory preserves informative historical evidence beyond the recurrent state.

Recurrent memory update frequency. We further investigate the effect of recurrent memory update frequency. As shown in Table 4, updating the recurrent state after every action chunk achieves a success rate of 66.22%, compared with 64.36% when the state is updated every two action chunks. More frequent updates allow the memory to incorporate newly observed interaction context with less temporal delay, keeping the recurrent representation better synchronized with changes in the environment and the robot's execution state. In contrast, a longer update interval may cause the latent memory to rely on increasingly stale context, particularly when intermediate actions induce meaningful changes in object configurations. These results suggest that updating the recurrent memory at each action chunk provides a more responsive representation of recent interaction history, and we therefore adopt this setting in the final model.

Table 6: Effect of the number of temporal Q-Former queries on the RoboCasa365 Atomic target split. We compare 16 and 32 queries using task success rate (%).

Temporal Queries	Success Rate
16	68.33%
32	**68.56%**

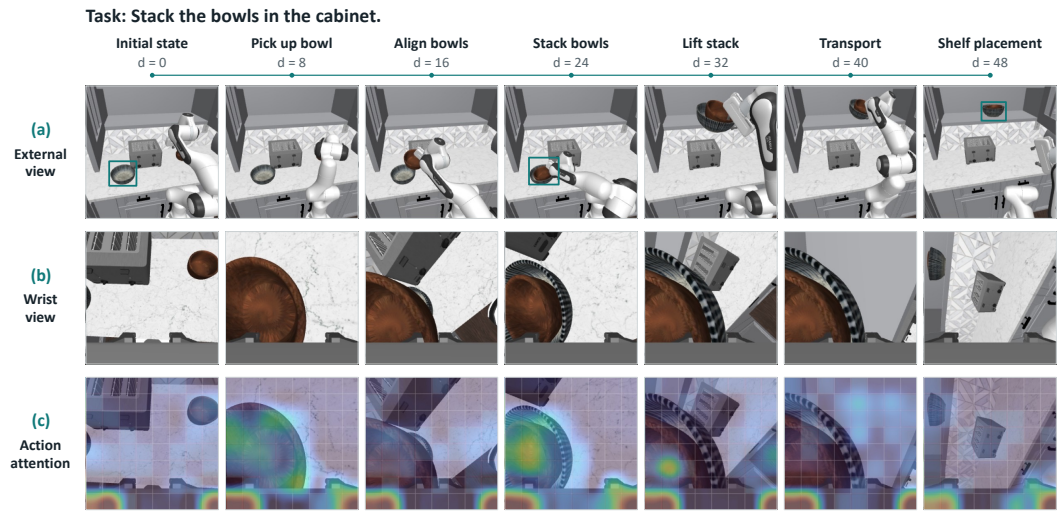

Figure 2: Qualitative case study of HERM-VLA on StackBowlsCabinet. From top to bottom: external views, wrist views, and action-attention maps with selected visual tokens during task progress.

Episodic snapshot composition. We further investigate the information retained in each episodic snapshot. As shown in Table 5, storing visual evidence together with cognitive features achieves a success rate of 65.78%, while further incorporating coarse spatial descriptors improves performance to 68.56%. The three components provide complementary information about historical observations. Visual tokens preserve fine-grained appearance cues, cognitive tokens summarize task-relevant semantics, and coarse descriptors retain global spatial structure that may be lost during selective visual token storage. Their combination therefore yields a richer episodic representation for conditioning subsequent action prediction.

Number of temporal queries. We further study the effect of the number of learned queries in the temporal Q-Former. As shown in Table 6, using 32 queries achieves a slightly higher success rate than using 16 queries (68.56% vs. 68.33%). The small performance gap indicates that the temporal aggregation module is not highly sensitive to the query number within this range. We adopt 32 queries in the final model to retain slightly higher aggregation capacity.

4.4 Case Study

Figure 2 shows HERM-VLA stacking bowls and placing them in the cabinet. As the wrist-camera viewpoint changes, the action-attention maps visualize the evidence selected for episodic consolidation, retaining 32 of 64 visual tokens per view. This selective retention complements Recurrent Working Memory, which propagates execution context across decisions. The rollout qualitatively illustrates coordinated multi-stage execution and action-aware evidence selection, consistent with the complementary roles of continuous task-state tracking and historical evidence preservation.

5 Limitations and Conclusion

We presented HERM-VLA, a Hybrid Episodic–Recurrent Memory Vision-Language-Action model for long-horizon embodied manipulation. By coupling Recurrent Working Memory with Episodic Evidence Memory, HERM-VLA maintains evolving interaction context while preserving task-relevant historical evidence beyond the effective horizon of recurrent state propagation. Content-aware memory consolidation further retains complementary visual, cognitive, and spatial information for subsequent action prediction. Experiments on RoboCasa365 and RMBench demonstrate the effectiveness of the proposed memory design, with ablations further validating the contribution of its individual components. Our current evaluation is primarily limited to simulated environments, and the episodic memory relies on a fixed-capacity bank and simple replacement strategy. Future work will investigate more adaptive memory management and evaluate the approach in real-world long-horizon manipulation settings.

AI USE STATEMENT

We used generative AI tools primarily for language polishing to improve the clarity, fluency, and presentation of the manuscript; all AI-assisted revisions were carefully reviewed by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or other ethical concerns outlined in the Code of Ethics.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the model architecture, training procedure, experimental settings, and evaluation protocols in the main paper and appendix. Upon publication, we will release the source code, model configurations, and relevant data and evaluation resources.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tianxing Chen, Yuran Wang, Mingleyang Li, Yan Qin, Hao Shi, Zixuan Li, Yifan Hu, Yingsheng Zhang, Kaixuan Wang, Yue Chen, et al. Rmbench: Memory-dependent robotic manipulation benchmark with insights into policy design. *arXiv preprint arXiv:2603.01229*, 2026a.

Yipeng Chen, Wentao Tan, Lei Zhu, Fengling Li, Jingjing Li, Guoli Yang, and Heng Tao Shen. Non-markovian long-horizon robot manipulation via keyframe chaining. *arXiv preprint arXiv:2603.01465*, 2026b.

Egor Cherepanov, Nikita Kachaev, Daniil Zelezetsky, Aydar Bulatov, Artem Pshenitsyn, Yuri Kuratov, Alexey Skrynnik, Aleksandr I Panov, and Alexey K Kovalev. μ vla: On recurrent memory for partially observable manipulation in vla models. *arXiv preprint arXiv:2606.12497*, 2026.

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. *The International Journal of Robotics Research*, 2025.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Jiaheng Hu, Mohit Shridhar, Caden Lu, Dhruv Shah, Hao-Tien Lewis Chiang, Jie Tan, and Annie Xie. What matters in orchestrating robot policies: A systematic study of hierarchical vla agents. *arXiv preprint arXiv:2606.10267*, 2026.

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.

Li Ji, Siyin Wang, Pengfang Qian, Xiaopeng Yu, Yihai Tian, Zhaoye Fei, Jingjing Gong, and Xipeng Qiu. Hime: Hierarchical embodied memory for long-horizon vision-language-action control. In *ICML*, 2026. URL https://openreview.net/forum?id=vVVbGj9cMC.

Dongyoung Kim, Huiwon Jang, Myungkyu Koo, Suhyeok Jang, Taeyoung Kim, Beomjun Kim, Byungjun Yoon, Changsung Jang, Daewon Choi, Dongsu Han, et al. Rldx-1 technical report. *arXiv preprint arXiv:2605.03269*, 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 *CoRL*, 2025.

Chengshu Li, Ruohan Zhang, Josiah Wong, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Proceedings of The 6th Conference on Robot Learning*, 2023.

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*, 2026a.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *ICLR*, volume 2024, pp. 26703–26721, 2024.

Xinghang Li, Peiyan Li, Long Qian, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Xinlong Wang, Di Guo, et al. What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, 2026b.

Min Lin, Xiwen Liang, Bingqian Lin, Liu Jingzhi, Zijian Jiao, Kehan Li, Yu Sun, Weijia Liufu, Yuhan Ma, Yuecheng Liu, et al. Echovla: Synergistic declarative memory for vla-driven mobile manipulation. *arXiv preprint arXiv:2511.18112*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tQJYKwc3n4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, pp. 6892–6903. IEEE, 2024.

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.

Ajay Sridhar, Jennifer Pan, Satvik Sharma, and Chelsea Finn. Scaling up memory for robotic control via experience retrieval. In *ICLR*, 2026.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *ICLR*, 2025.

Marcel Torne, Karl Pertsch, Homer Walke, Kyle Vedder, Suraj Nair, Brian Ichter, Allen Z Ren, Haohuan Wang, Jiaming Tang, Kyle Stachowicz, et al. Mem: Multi-scale embodied memory for vision language action models. *arXiv preprint arXiv:2603.03596*, 2026.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worldwarper: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.

Lei Xiao, Jifeng Li, Juntao Gao, Feiyang Ye, Yan Jin, Jingjing Qian, Jing Zhang, Yong Wu, and Xiaoyuan Yu. Ava-vla: Improving vision-language-action models with active visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13453–13463, June 2026.

Peiran Xu, Jiaqi Zheng, and Yadong Mu. Roboagent: Chaining basic capabilities for embodied task planning. In *CVPR*, June 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ganlin Yang, Zhangzheng Tu, Yuqiang Yang, Sitong Mao, Junyi Dong, Tianxing Chen, Jiaqi Peng, Jing Xiong, Jiafei Cao, Jifeng Dai, et al. Eventvla: Event-driven visual evidence memory for long-horizon vision-language-action policies. *arXiv preprint arXiv:2606.20092*, 2026a.

Sizhe Yang, Juncheng Mu, Tianming Wei, Chenhao Lu, Xiaofan Li, Linning Xu, Zhengrong Xue, Zhecheng Yuan, Dahua Lin, Jiangmiao Pang, et al. Memorywam: Efficient world action modeling with persistent memory. *arXiv preprint arXiv:2606.20562*, 2026b.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, et al. Harness vla: Steering frozen vlas into reliable manipulation primitives via memory-guided agents. *arXiv preprint arXiv:2607.08448*, 2026.

Jiacong Zhou, Jiaxu Miao, xianyun wang, and Jun Yu. Enhancing llm planning for robotics manipulation through hierarchical procedural knowledge graphs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025.

Jiacong Zhou, Jiaxu Miao, Yourun Lin, Xianyun Wang, Jun Xiao, and Jun Yu. Memory-augmented scene understanding and exploration for open-world aerial object-goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21616–21626, 2026.