

SkillMemo: Expert-guided Skill Memory Framework for Compositional Embodied Manipulation

Changyuan Wang¹ Chubin Zhang¹ Zhenyu Wu² Runhao Li³ Angyuan Ma² Ke Chao⁴
Yinan Liang² Xiuwei Xu² Ziwei Wang³ Yansong Tang^{1*} Jiwen Lu²

¹Shenzhen International Graduate School, Tsinghua University

²Department of Automation, Tsinghua University

³Nanyang Technological University

⁴Beijing Normal University

Project Page: <https://changyuanwang17.github.io/SkillMemo/>

Abstract: Embodied visuomotor models, including Diffusion Policy (DP) and Vision-Language-Action (VLA) models, have demonstrated promising performance on robotic manipulation benchmarks. However, their potential remains fundamentally constrained by the scarcity of large-scale embodied trajectory datasets, leading to insufficient compositional generalization in out-of-distribution (OOD) scenarios with limited capability to capture reusable skill structures. To address this limitation, we propose Skill-Based Memory (SkillMemo) framework that implicitly decomposes long-horizon demonstrations into latent atomic skills and integrates skill-level features into a dynamic episodic memory bank for solving compositional tasks. Specifically, we first introduce an expert-guided trajectory segmentation module built upon a Mixture-of-Experts (MoE) architecture, which implicitly partitions trajectories into distinct skill primitives represented by learned gating coefficients. We further design a skill-level episodic memory architecture that stores compact skill representations as retrievable key-value pairs. During inference, the memory bank retrieves the most relevant skill primitives which are subsequently fused with the model’s current gating distribution, providing a robust contextual prior to refine action predictions. Extensive experiments on the simulation benchmark and real-world manipulation tasks demonstrate that SkillMemo consistently enhances both DP and VLA backbones, achieving state-of-the-art performance and outperforming $\pi_{0.5}$, while exhibiting strong compositional generalization to unseen task configurations.

Keywords: Skill Learning, Memory-Augmented Policy, Robot Manipulation

1 Introduction

Embodied visuomotor models, including Diffusion Policy (DP) [1, 2, 3] and Vision-Language-Action (VLA) [4, 5, 6, 7, 8] models, have achieved remarkable success in robotic manipulation, primarily driven by large-scale vision-language pretraining followed by embodied trajectory finetuning [9, 10, 11]. Nonetheless, their real-world deployment remains fundamentally limited by the high cost of acquiring large-scale robot datasets. Training on insufficient demonstrations often results in monolithic policies that fail to capture reusable behavioral structures, leading to degraded performance when facing out-of-distribution (OOD) scenarios wherein novel compositions of known objects, receptacles, and instructions.

To address these challenges, prior efforts have investigated robotic reinforcement learning [12, 13, 14], motion primitive libraries [15, 16, 17], and memory-augmented policies [18, 19]. Among these, memory-augmented methods have emerged as a promising paradigm for long-horizon generalization.

*Corresponding author.

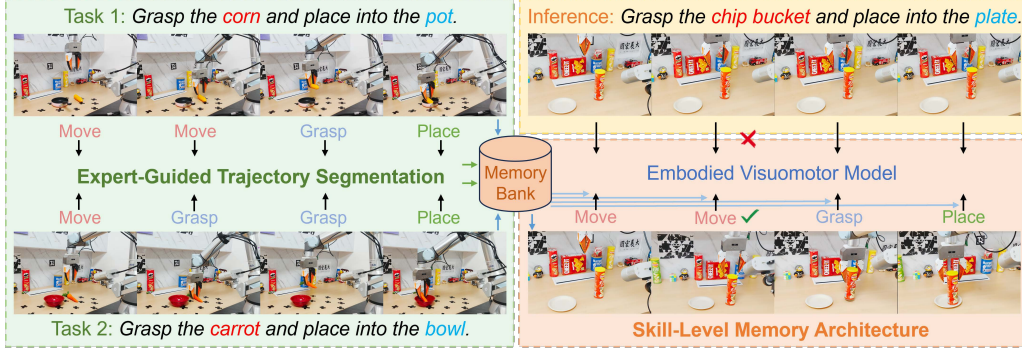


Figure 1: Unlike conventional models that struggle to capture reusable skill structures, our method implicitly decomposes trajectories into latent atomic primitives for memory storage, which are dynamically retrieved and fused to facilitate compositional generalization to unseen task configurations.

MemoryVLA [18] introduces a dual-stream Perceptual-Cognitive Memory Bank for context-aware decision making. HAMLET [19] addresses non-Markovian dependencies by encoding interaction history into temporal moment tokens. However, existing methods lack an explicit mechanism for memorizing behaviorally salient skill-level features. Consequently, they tend to exhibit inefficient reuse of prior experience and diminished capability for compositional generalization in unseen tasks.

In this paper, we propose a Skill-Based Memory (SkillMemo) framework to enhance robust compositional generalization for embodied manipulation. Unlike conventional visuomotor models that encode entire behavioral trajectories with a single monolithic network, SkillMemo implicitly decomposes long-horizon demonstrations into latent atomic skills and integrates skill-level features into a dynamic episodic memory bank for solving compositional tasks. Specifically, we first introduce an expert-guided trajectory segmentation module built on a Mixture-of-Experts (MoE) architecture, which implicitly partitions trajectories into distinct skill primitives represented by learned gating coefficients. Building upon this segmentation, we design a skill-level episodic memory architecture that stores compact skill representations as retrievable key-value pairs. During inference, the memory bank retrieves the most relevant skill primitives which are subsequently fused with the model’s current gating distribution, providing a robust contextual prior to refine action predictions. Extensive experiments on the simulation benchmark and real-world UR5e manipulation tasks demonstrate that SkillMemo consistently improves both DP and VLA backbones, achieving state-of-the-art performance and outperforming $\pi_{0.5}$ by 1.2% (98.0% vs. 96.8%), while exhibiting strong compositional generalization to unseen task configurations.

2 Related Works

2.1 Embodied Visuomotor Model

Recent advances in embodied AI have been largely driven by Diffusion Policy (DP) [1] and Vision-Language-Action (VLA) models [20, 8, 21]. While DP formulates control as iterative denoising and has been extended via 3D representations [22, 23] and hierarchical structures [24, 2], VLA models couple large-scale vision-language pretraining with action generation using extensive robot datasets [25, 9, 11, 7, 26]. Despite their impressive zero-shot capabilities, both paradigms struggle with compositional generalization when limited demonstrations are available, motivating our work on skill-level memory augmentation.

2.2 Skill Learning

To address the temporal complexity of manipulation, recent approaches explore skill decomposition. Explicit frameworks [17, 27, 16] decouple high-level planning from execution by constructing dynamic libraries of motor primitives or prompts for task composition. Alternatively, implicit methods discover skills within the policy architecture using phase-aware masking [28] or Mixture-

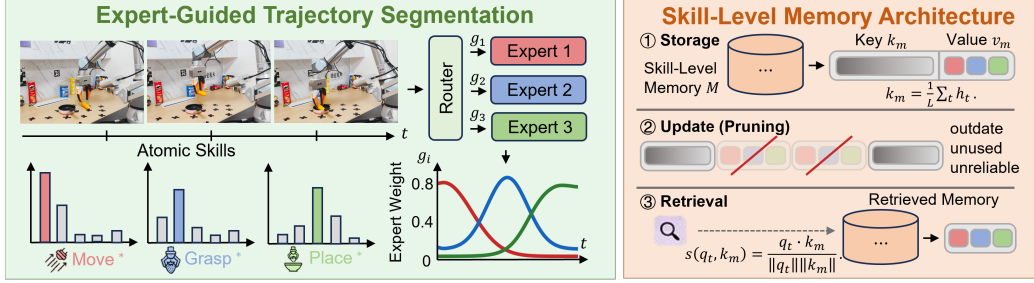


Figure 2: Overview of the SkillMemo framework. The Expert-Guided Trajectory Segmentation module implicitly* extracts latent atomic skills from demonstrations via MoE routing. The Skill-Level Memory Architecture structurally stores and dynamically prunes these learned gating profiles, enabling context-aware retrieval for robust compositional generalization on unseen task.

of-Experts (MoE) [2, 29]. Distinct from methods relying on static skill libraries or purely implicit partitioning, our approach uniquely couples MoE-based skill discovery with a dynamic episodic memory bank, enabling structured storage and compositional retrieval of learned skills.

2.3 Memory-based Framework

To mitigate the limitations of purely reactive policies, recent research integrates explicit memory mechanisms into visuomotor architectures. A prominent example is MemoryVLA [18], which utilizes a dual-stream Perceptual-Cognitive Memory Bank to store visual details and semantic summaries for context-aware retrieval. While other works explore temporal tokens or learnable prompts [30, 31, 19, 32, 33], these methods typically store unstructured or holistic visual-semantic representations. They lack the capacity to capture and reuse fine-grained, skill-level behavioral patterns. In contrast, SkillMemo structurally stores decomposed skill primitives and retrieves them based on behavioral relevance, providing robust contextual priors for compositional generalization.

3 Methods

3.1 Expert-Guided Trajectory Segmentation

Conventional VLA models typically learn entire behavioral trajectories using a single monolithic network, which leads to limited compositionality and poor generalization over heterogeneous, high-dimensional action sequences. To address this limitation, we introduce an expert-guided trajectory segmentation module that structurally decomposes entire trajectories into a sequence of latent atomic skills. This decomposition enables fine-grained specialization of network capacity and facilitates structured skill reuse. Formally, given a full trajectory $\tau = \{(x_t, a_t)\}_{t=1}^T$, where x_t and a_t denote the observation and action at time t , the explicit segmentation process partitions τ into a set of non-overlapping skill segments:

$$\tau \rightarrow \{s_k\}_{k=1}^K, \quad s_k = \{(x_t, a_t) \mid t \in [t^{\text{start}}, t^{\text{start}+L}]\}, \quad (1)$$

where each segment s_k corresponds to a semantically coherent skill with the chunk size L . However, explicit segmentation schemes often fail to capture all potential skill boundaries which leads to inaccurate skill partitioning. On the contrary, we introduce an implicit skill partitioning scheme, where the decomposition is dynamically learned from the model’s feature representations. Specifically, we implement this skill decomposition using a Mixture-of-Experts (MoE) architecture, which consists of N expert networks $\{E_1, \dots, E_N\}$ and a gating network G . Given an intermediate feature representation h_t at time t , the MoE layer computes:

$$y_t = \sum_{i=1}^N g_i(h_t) \cdot E_i(h_t), \quad (2)$$

where $E_i(\cdot)$ denotes the i -th expert and $g_i(h_t)$ represents its corresponding gating weight. Intuitively, each expert captures a recurring local motion primitive, while the gating network g implicitly defines

the temporal segmentation boundaries by modulating expert activations over time. This implicit partitioning mechanism enables adaptive and data-driven skill discovery, ensuring that the identified skills align with emergent task-level structures.

To further encourage distinct skill specialization among experts and promote emergent skill composition, we regularize the model using a synergy-based information loss derived from the theory of *Partial Information Decomposition (PID)*. For any pair of experts (A, B) , we define their synergistic information with respect to the task target G as:

$$\mathcal{L}_{\text{PID}} = -[I(\{A, B\}; G) - I(A; G) - I(B; G)], \quad (3)$$

where $I(\cdot; \cdot)$ denotes mutual information. Minimizing $\mathcal{L}_{\text{PID}}$ encourages the joint representation of the experts to contain more task-relevant information than the sum of individual expert representations. This drives each expert to specialize in complementary skills while enabling higher-order coordination, thereby facilitating the learning of distinct, non-overlapping skill primitives. Complementarily, we employ expert load-balancing losses and top-k routing strategies to prevent expert collapse.

3.2 Skill-Level Memory Architecture

While the MoE-based trajectory segmentation in Section 3.1 discovers a set of reusable atomic skills, a critical challenge remains: how to effectively leverage these skills for solving compositional novel tasks where previously seen objects are recombined in ways not seen during training. Conventional VLA policies are purely reactive, relying only on the current observation and their static network weights, thereby discarding the rich historical context embedded in past successful executions. To endow our agent with a form of experiential reasoning, we introduce a Skill-Level Episodic Memory Architecture that stores, retrieves, and composes previously learned skill primitives during inference.

Memory Storage. We first establish the foundational principle of the memory module: explicitly retaining previously acquired skills to facilitate task completion through adaptive skill composition during inference. To achieve this, we organize the memory bank $\mathcal{M}$ as a collection of Key-Value pairs, comprising two core components:

(1) *Memory Key.* For each atomic skill segment $s_k = \{(x_t, a_t)\}_{t=t_{\text{start}}}^{t_{\text{start}}+L}$ discovered by the segmentation module, we construct a compact memory entry. To enable efficient retrieval, we utilize the latent trajectory features h_t to determine the similarity between the inference observation and the stored skills. However, directly storing frame-wise features for all timesteps results in an unscalable memory bank. To address this, we employ temporal aggregation to extract a compact centroid feature for each skill segment:

$$\mathbf{k}_m = \frac{1}{L} \sum_{t \in \tau_m} h_t, \quad (4)$$

where τ_m denotes the m -th skill segment and h_t represents the latent feature at time t . By storing this compact centroid, we preserve the most representative characteristics of the skill trajectory, ensuring the efficiency of the retrieval process.

(2) *Memory Value.* The value component stores the functional knowledge required to reproduce the skill. Since the gating weights $g(h_t)$ govern expert activation and implicitly define the category of the skill, we designate the gating distribution as the core knowledge content. To ensure the preservation of detailed temporal dynamics and information validity, we store the complete sequence of gating weights across all time steps within the segment. The final memory bank $\mathcal{M}$ is thus organized as:

$$\mathcal{M} = \{(\mathbf{k}_m, \mathbf{v}_m) \mid \mathbf{v}_m = \{g(h_t)\}_{t \in \tau_m}\}_{m=1}^M, \quad (5)$$

where M is the total number of stored skill primitives, and $\mathbf{v}_m$ represents the m -th stored gating weights. This memory structure enables the model to efficiently retrieve the most relevant skills and significantly enhances the inference process through expert guidance.

Memory Retrieval. Following the construction of the memory bank during training, we leverage the stored procedural knowledge to enhance policy execution during inference. Standard VLA models

often struggle with generalization when facing out-of-distribution (OOD) scenarios due to the scarcity of diverse training samples. Our skill-based memory addresses this by providing a structured bank of reusable behaviors.

First, we identify the most relevant historical skills by comparing the current state representation with the stored memory keys. Let q_t denote the latent feature of the current observation at inference time. We compute the cosine similarity between q_t and every key $\mathbf{k}_m$ in the memory bank $\mathcal{M}$:

$$s(q_t, \mathbf{k}_m) = \frac{q_t \cdot \mathbf{k}_m}{\|q_t\| \|\mathbf{k}_m\|}, \quad (6)$$

where $s(\cdot, \cdot)$ denotes the similarity score. Based on these scores, we retrieve the top- N memory entries $\mathcal{M}_{\text{top}} = \{(\mathbf{k}_n, \mathbf{v}_n)\}_{n=1}^N$ that exhibit the highest affinity with the current state.

To ensure robustness, we introduce a memory reliability check before integrating the retrieved information. We calculate the average similarity score of the retrieved set and compare it against a predefined confidence threshold δ . If the retrieval is deemed reliable, we proceed to augment the current policy by fusing the retrieved expert activation profiles with the model’s current predicted gating distribution. Specifically, we refine the gating weights $g(h_t)$ by averaging them with the retrieved gating profiles:

$$g'(q_t) = \lambda g(q_t) + (1 - \lambda) \frac{1}{N} \sum_{n=1}^N \mathbf{v}_n[t'], \quad (7)$$

where λ is a hyperparameter controlling the strength of the memory intervention, and $\mathbf{v}_n[t']$ corresponds to the aligned timestep within the retrieved skill sequence. By explicitly incorporating these retrieved priors, the model can synthesize novel behaviors through the composition of known primitives, significantly improving compositional generalization on unseen task configurations.

Memory Update. To adapt to lifelong learning scenarios where the agent continuously encounters new tasks, we design a dynamic memory update mechanism for finite-capacity memory bank. Since the memory bank has a finite capacity M_{max} , it is crucial to prune obsolete information to accommodate new experiences. We implement a usage-tracking method that monitors the retrieval history and validation status of each memory entry.

When the memory bank reaches its capacity limit, we employ a prioritized pruning strategy to discard low-utility information. Specifically, we target memory entries for removal based on three criteria: (1) memories with the most outdated timestamps, mimicking the human cognitive process of forgetting distant, non-reinforced details; (2) memories that have remained unused, i.e., not successfully retrieved; (3) memories that consistently fail to meet the reliability threshold δ during the retrieval phase, indicating low-quality or irrelevant entries. This dynamic pruning strategy balances the trade-off between memory diversity and memory quality, ensuring that the memory bank remains compact, relevant, and conducive to effective skill composition.

4 Experiments

4.1 Implementation Details

We evaluate our proposed SkillMemo framework on multimodal embodied manipulation tasks by instantiating both diffusion policy (DP) and vision-language-action (VLA) frameworks as backbones. Specifically, we adopt Diffusion Policy [1] for low-level visuomotor control and UniACT-0.5B [15], UniVLA-7B [6] as representative large-scale VLA models. Further experimental setups, including hyperparameter configurations and hardware specifics, are detailed in Appendix A.

4.2 Ablation Studies

Component-wise Analysis. To isolate the contributions of each component, we conduct ablation studies under both the DP and VLA frameworks, evaluating task success rate on the LIBERO

Table 1: Effect of different methods we proposed, reporting accuracy of DP and UniACT model.

Model	EGTS	SLMA	Goal	Spatial	Object
UniAct	✓	✓	68.7	72.1	75.7
			71.1	78.6	79.3
			73.4	80.2	79.8
Model	EGTS	SLMA	Push-T	BlockPush	Kitchen
DP	✓	✓	52.9	73.6	57.1
			53.6	76.5	59.1
			55.2	78.1	60.6

Table 2: Effect of expert number on UniAct success rates and inference time.

Expert Number (N)	Goal	Spatial	Object	Inference Time (s)
1	68.7	72.1	75.7	3.20
2	69.7	75.6	77.9	2.78
3	70.3	76.3	78.5	3.60
5	71.1	78.6	79.3	3.93
8	72.3	78.9	79.5	5.43
10	72.7	79.5	80.1	7.94

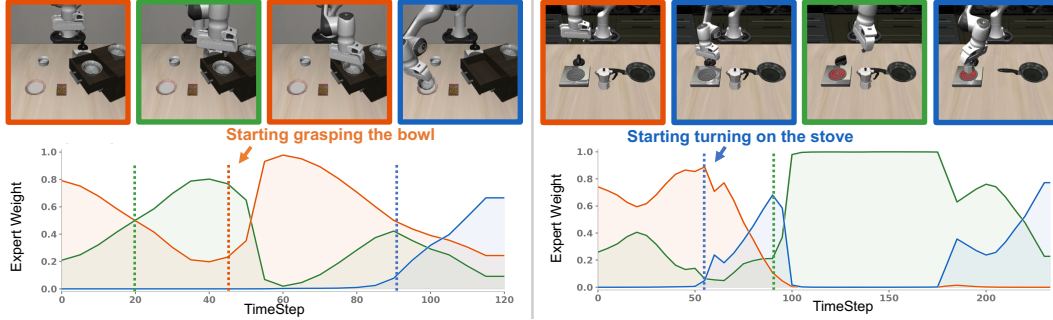


Figure 3: Visualization of expert activation weights over time. The gating coefficients dynamically shift and exhibit pronounced peaks during critical action phases such as *grasping a bowl* or *turning on a stove*, demonstrating that the EGTS module successfully drives individual experts to implicitly specialize in distinct, semantically meaningful motion primitives.

benchmark [34]. Starting from the backbone model, we progressively integrate the proposed *Expert-Guided Trajectory Segmentation* (EGTS) and *Skill-Level Memory Architecture* (SLMA). Table 1 reports the success rates across different configurations. Compared to the vanilla backbone, adding EGTS alone yields consistent improvements across all task suites, confirming the effectiveness of implicit skill decomposition. Further incorporating SLMA provides additional gains, particularly on tasks requiring compositional reasoning.

Impact of Experts. We further investigate the sensitivity of our framework to the number of experts ($N \in \{1, 2, 3, 5, 8, 10\}$), evaluating both task accuracy and computational efficiency in Table 2. As the number of experts increases, the model’s ability to represent diverse skill primitives improves, leading to a steady rise in success rates. However, we observe diminishing returns beyond $N=5$; specifically, increasing the expert count to 10 provides only marginal accuracy improvement while substantially increasing inference time. Therefore, we adopt $N=5$ as the default configuration.

Expert Activation Analysis. To further validate the effectiveness of our implicit skill partitioning, we visualize the temporal evolution of expert routing weights during task execution. As illustrated in Figure 3, the gating coefficients exhibit distinct, dynamic shifts that tightly align with critical motion transitions. Notably, when the robotic arm initiates significant semantic actions such as *grasping a bowl* or *turning on a stove*, the activation weights of specific experts demonstrate pronounced peaks. This temporal correlation confirms that the MoE routing mechanism successfully extracts and specializes in distinct latent atomic skills, even in the complete absence of explicit action labels.

4.3 Simulated Evaluation

To comprehensively evaluate the effectiveness of the proposed *Skill-Based Memory* (SkillMemo) framework, we conduct experiments under both DP and VLA backbones across various widely used benchmarks: LIBERO, Push-T, UR3 Block Push, and Franka Kitchen.

Benchmarks with DP backbone. Follow DP [1], we evaluate our framework on three standard continuous-control environments: Push-T, UR3 Block Push, and Franka Kitchen. Table 3 presents a comparative analysis of SkillMemo against various state-of-the-art baselines, including Diffusion

Table 3: Success rates compared to state-of-the-art Diffusion Policy(DP) baselines across various simulated benchmarks.

Model	Push-T	BlockPush	Kitchen	Avg.
DP [1]	52.9	73.6	57.1	61.2
SDP [2]	53.5	72.9	56.3	60.9
CP [35]	53.2	73.2	56.0	60.8
IMLE Policy [3]	53.7	77.2	57.5	62.8
STEP [36]	49.7	76.9	58.1	61.6
SkillMemo	55.2	78.1	60.6	64.6

Table 4: Zero-shot generalization to unseen task configurations. Models trained on specific LIBERO suites are evaluated cross-suite.

Dataset	Method	Goal	Spatial	Object
LIBERO-Goal	w/o memory	68.7	71.3	72.5
	w memory	73.4	75.6	79.1
LIBERO-Spatial	w/o memory	70.3	72.1	73.9
	w memory	72.0	80.2	78.2
LIBERO-Object	w/o memory	69.5	70.8	75.7
	w memory	71.6	73.4	79.8

Table 5: Success rates on the LIBERO benchmark. Our proposed SkillMemo framework consistently enhances the performance of various state-of-the-art VLA backbones across different scales.

Method	Params	LIBERO Benchmark (%) $\uparrow$				Average $\uparrow$
		Goal	Spatial	Object	Long	
OpenVLA [8]	7.0B	78.0	85.0	86.8	54.0	76.0
TriVLA [37]	3.4B	89.8	91.2	93.8	73.2	87.0
CogACT [38]	7.6B	90.2	97.2	98.0	88.8	93.2
π_0 [21]	3.3B	95.8	96.8	98.8	85.2	94.2
MemoryVLA [18]	7.3B	96.4	98.4	98.4	93.4	96.5
UniAct [15]	0.5B	68.7	72.1	75.7	51.4	67.2
UniAct-SkillMemo (Ours)	0.6B	73.4	80.2	79.8	57.2	72.7
UniVLA [6]	8.5B	91.8	96.5	95.6	92.0	93.9
UniVLA-SkillMemo (Ours)	8.9B	94.5	98.1	97.8	93.2	95.9
$\pi_{0.5}$ [7]	3.3B	98.0	98.8	98.2	92.4	96.8
$\pi_{0.5}$-SkillMemo (Ours)	3.6B	99.0	99.4	98.2	95.4	98.0

Policy (DP) [1], Sparse Diffusion Policy (SDP) [2], Consistency Policy (CP) [35], IMLE Policy [3], and STEP [36]. While existing methods fundamentally lack explicit temporal memory and struggle with fine-grained skill extraction, severely limiting their high-level task compositionality. In contrast, by structurally storing and dynamically retrieving latent atomic skills, SkillMemo consistently outperforms all baselines across every benchmark, achieving the highest average success rate of 64.6%, where surpassing DP by 3.4% and IMLE Policy by 1.8%.

Benchmarks with VLA backbone. Following OpenVLA [8], we evaluate our framework on the LIBERO benchmark [34] utilizing a 7-DoF Franka Panda robot across different suites: *Goal*, *Spatial*, *Object*, and *Long*. Each suite consists of 10 distinct manipulation tasks, while LIBERO-Long focuses on long-horizon tasks. We follow $\pi_{0.5}$ [7] which training single unified policy across all tasks for 30k steps. Validation reports the average success rate over 50 rollouts per task.

Table 5 compares SkillMemo against cutting-edge VLA architectures on the LIBERO benchmark. While recent approaches like MemoryVLA [18] leverage holistic visual-semantic memory, they lack fine-grained behavioral priors for continuous action execution. Furthermore, even highly optimized foundation models such as π_0 [21] and $\pi_{0.5}$ [7] lack dedicated mechanisms to dynamically retrieve and compose historical skills. By structurally integrating latent atomic skills into a dynamic memory bank, SkillMemo consistently enhances these strong backbones. Notably, $\pi_{0.5}$ -SkillMemo establishes average success rate of 98.0%, yielding a 1.2% (98.0% vs. 96.8%) improvement over $\pi_{0.5}$ and decisively surpassing MemoryVLA by 1.5% (98.0% vs. 96.5%), demonstrating that skill-level memory is crucial for maximizing VLA performance.

Unseen task generalization. To further evaluate the generalization ability of the learned skills, we conduct a zero-shot cross-suite evaluation where models trained on the *Goal*, *Spatial*, and *Object* subsets respectively are directly evaluated on the other subsets. As shown in Table 4, the SkillMemo framework trained on *Goal* achieves 75.6% on unseen *Spatial* suite, which even outperforms the in-distribution baseline with 3.5% (75.6% vs. 72.1%) trained explicitly on *Spatial*. This result strongly

Table 6: Single-task performance in real world. Success rates (%) are reported over 40 trials per task.

Model	Strawberry in Bowl	Lemon on Plate	Corn in Pot	Butter in Pot	Chip Bucket on Table
Diffusion Policy	77.5	77.5	82.5	62.5	67.5
SkillMemo	82.5	80.0	90.0	75.0	77.5

Table 7: Real-world generalization to unseen tasks. Marks with ✓ indicates the source tasks used for training. We evaluate the models on unseen compositional tasks to test skill transferability. Success rates (%) are reported over 40 trials.

Unseen Tasks	Strawberry in Bowl	Lemon on Plate	Butter in Pot	Diffusion Policy	SkillMemo
Strawberry in Bowl	✓	✓		77.5	82.5
Lemon on Plate	✓	✓		77.5	80.0
Strawberry on Plate	✓	✓		65.0	75.0
Lemon in Bowl	✓	✓		62.5	75.0
Strawberry in Pot	✓	✓	✓	60.0	70.0
Lemon in Pot	✓	✓	✓	57.5	72.5
Butter in Bowl	✓	✓	✓	65.0	70.0
Butter on Plate	✓	✓	✓	60.0	67.5

supports our hypothesis that the memory bank enables compositional reuse of skill primitives, allowing the model to generalize beyond the training distribution by recombining known skills.

4.4 Real-World Evaluation

Single-Task Evaluation. Following the hardware setup described in the Implementation Details, we deploy the policies on a real-world UR5e robotic manipulator. We curate a suite of five contact-rich manipulation tasks involving diverse objects and semantic goals: *Put the Strawberry into the Bowl*, *Place the Lemon on the Plate*, *Put the Corn into the Pot*, *Grasp the Butter in the Pot*, and *Place the Chip Bucket on the Table*. For each task, we collect a dataset of 50 expert demonstrations and evaluate using 40 independent rollouts. As shown in Table 6, SkillMemo consistently outperforms DP across all five tasks, with particularly notable improvements on more challenging tasks such as *Corn in Pot* (+7.5%) and *Butter in Pot* (+12.5%).

Unseen Task Generalization. To investigate the capability of SkillMemo to generalize to unseen object-receptacle combinations, we design a compositional evaluation protocol. Models are trained on a limited subset of source tasks and subsequently evaluated on novel pairings. As detailed in Table 7, we establish two generalization settings. In the first scenario, models are co-trained on *Strawberry in Bowl* and *Lemon on Plate*, and then tested on the unseen crossover combinations. In the second scenario, we extend the training set with *Butter in Pot* and evaluate generalization to more complex unseen targets. The results indicate that SkillMemo significantly outperforms DP in these compositional scenarios, improving *Strawberry on Plate* from 65.0% to 75.0%. This confirms that our framework effectively retrieves and fuses atomic skills from disparate training experiences to facilitate robust compositional generalization on unseen task configurations.

5 Conclusion

We presented SkillMemo, a skill memory framework for compositional embodied manipulation that addresses the limitations of existing visuomotor models in capturing and reusing structured behavioral patterns. By coupling expert-guided trajectory segmentation with a skill-level episodic memory architecture, SkillMemo enables the explicit decomposition, storage, retrieval, and composition of atomic skill primitives. Extensive experiments on both simulation benchmarks and real-world robotic tasks demonstrate that SkillMemo consistently enhances the performance of diverse backbone architectures while exhibiting zero-shot compositional generalization to unseen task compositions.

References

- [1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [2] Y. Wang, Y. Zhang, M. Huo, R. Tian, X. Zhang, Y. Xie, C. Xu, P. Ji, W. Zhan, M. Ding, et al. Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning. *arXiv preprint arXiv:2407.01531*, 2024.
- [3] K. Rana, R. Lee, D. Pershouse, and N. Suenderhauf. Imle policy: Fast and sample efficient visuomotor policy learning via implicit maximum likelihood estimation. *arXiv preprint arXiv:2502.12371*, 2025.
- [4] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.
- [5] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, H. Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.
- [6] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.
- [7] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- [8] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [9] A. O’Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.
- [10] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.
- [11] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, et al. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.
- [12] J. Luo, C. Xu, J. Wu, and S. Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105):eads5033, 2025.
- [13] A. Wagenmaker, M. Nakamoto, Y. Zhang, S. Park, W. Yagoub, A. Nagabandi, A. Gupta, and S. Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025.
- [14] L. Ankile, Z. Jiang, R. Duan, G. Shi, P. Abbeel, and A. Nagabandi. Residual off-policy rl for finetuning behavior cloning policies. *arXiv preprint arXiv:2509.19301*, 2025.
- [15] J. Zheng, J. Li, D. Liu, Y. Zheng, Z. Wang, Z. Ou, Y. Liu, J. Liu, Y.-Q. Zhang, and X. Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22508–22519, 2025.

- [16] Y. Yao, S. Liu, H. Song, D. Qu, Q. Chen, Y. Ding, B. Zhao, Z. Wang, X. Li, and D. Wang. Think small, act big: Primitive prompt learning for lifelong robot manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22573–22583, 2025.
- [17] W. Mao, W. Zhong, Z. Jiang, D. Fang, Z. Zhang, Z. Lan, H. Li, F. Jia, T. Wang, H. Fan, et al. Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. *arXiv preprint arXiv:2412.00171*, 2024.
- [18] H. Shi, B. Xie, Y. Liu, L. Sun, F. Liu, T. Wang, E. Zhou, H. Fan, X. Zhang, and G. Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025.
- [19] M. Koo, D. Choi, T. Kim, K. Lee, C. Kim, Y. Seo, and J. Shin. Hamlet: Switch your vision-language-action model into a history-aware policy. *arXiv preprint arXiv:2510.00695*, 2025.
- [20] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.
- [21] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [22] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.
- [23] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu. Generalizable humanoid manipulation with 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.
- [24] Y. Lu, Y. Tian, Z. Yuan, X. Wang, P. Hua, Z. Xue, and H. Xu. H^3 DP: Triply-hierarchical diffusion policy for visuomotor learning. *arXiv preprint arXiv:2505.07819*, 2025.
- [25] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- [26] P. Li, Y. Wu, Z. Xi, W. Li, Y. Huang, Z. Zhang, Y. Chen, J. Wang, S.-C. Zhu, T. Liu, et al. Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. *arXiv preprint arXiv:2506.16211*, 2025.
- [27] D. Li, B. Peng, C. Li, N. Qiao, Q. Zheng, L. Sun, Y. Qin, B. Li, Y. Luan, B. Wu, et al. An atomic skill library construction method for data-efficient embodied manipulation. *arXiv preprint arXiv:2501.15068*, 2025.
- [28] Y. Fan, P. Ding, S. Bai, X. Tong, Y. Zhu, H. Lu, F. Dai, W. Zhao, Y. Liu, S. Huang, et al. Longvla: Unleashing long-horizon capability of vision language action model for robot manipulation. *arXiv preprint arXiv:2508.19958*, 2025.
- [29] S. Huang, Z. Zhang, T. Liang, Y. Xu, Z. Kou, C. Lu, G. Xu, Z. Xue, and H. Xu. Mentor: Mixture-of-experts network with task-oriented perturbation for visual reinforcement learning. *arXiv preprint arXiv:2410.14972*, 2024.
- [30] M. Lei, H. Cai, Z. Cui, L. Tan, J. Hong, G. Hu, S. Zhu, Y. Wu, S. Jiang, G. Wang, et al. Robomemory: A brain-inspired multi-memory agentic framework for interactive environmental learning in physical embodied systems. *arXiv preprint arXiv:2508.01415*, 2025.
- [31] R. Li, W. Guo, Z. Wu, C. Wang, H. Deng, Z. Weng, Y.-P. Tan, and Z. Wang. Map-vla: Memory-augmented prompting for vision-language-action model in robotic manipulation. *arXiv preprint arXiv:2511.09516*, 2025.

- [32] G. Lan, K. Qu, R. Zurbrügg, C. Chen, C. E. Mower, H. Bou-Ammar, and M. Hutter. Experience is the best teacher: Grounding vlms for robotics through self-generated memory. *arXiv preprint arXiv:2507.16713*, 2025.
- [33] Y. Lei, S. Mao, S. Zhou, H. Zhang, X. Li, and P. Luo. Dynamic mixture of progressive parameter-efficient expert library for lifelong robot learning. *arXiv preprint arXiv:2506.05985*, 2025.
- [34] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.
- [35] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*, 2024.
- [36] J. Li, Y. Cong, Y. Wang, H. Xia, S. Huang, Y. Zhang, N. Xu, and G. Dai. Step: Warm-started visuomotor policies with spatiotemporal consistency prediction. *arXiv preprint arXiv:2602.08245*, 2026.
- [37] Z. Liu, Y. Gu, S. Zheng, Y. Fu, X. Xue, and Y.-G. Jiang. Trivla: A triple-system-based unified vision-language-action model with episodic world modeling for general robot control. *arXiv preprint arXiv:2507.01424*, 2025.
- [38] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

A Extended Implementation Details

A.1 Simulation Setup

For simulation experiments, we conduct evaluations on the widely utilized LIBERO benchmark [34]. LIBERO comprises a collection of long-horizon and language-conditioned robotic manipulation tasks, designed to test generalization to compositional and unseen task instructions. All models are trained on 8 NVIDIA A6000 GPUs with PyTorch. We use 32 samples per GPU for a global batch size of 256 and adopt AdamW with a learning rate of 2×10^{-5} . The policy takes as input a single RGB image observation and a language instruction, and outputs a 7-dimensional continuous action (6-DoF end-effector pose plus gripper command). The MoE module consists of $N=5$ experts by default. The memory bank stores up to $M_{\max}=1000$ skill entries.

A.2 Real-World Setup

For real-world validation, we deploy the DP model on a single 7-DoF UR5e robotic manipulator integrated with a Weiss WSG-50 parallel-jaw gripper, as illustrated in Figure 4. An ORBBEC Femto Bolt RGB-D camera is mounted on a flexible boom arm in a side-view configuration to capture visual observations. The robotic arm is anchored to a height-adjustable lifting table with a mobile base, which facilitates the evaluation of policy robustness across varying viewpoints and mounting heights. The action space is defined as a 6-dimensional vector, comprising 3D end-effector translation, 3D rotation (Euler angles), and a binary gripper command. We collect 50 expert demonstrations per task through kinesthetic teaching.

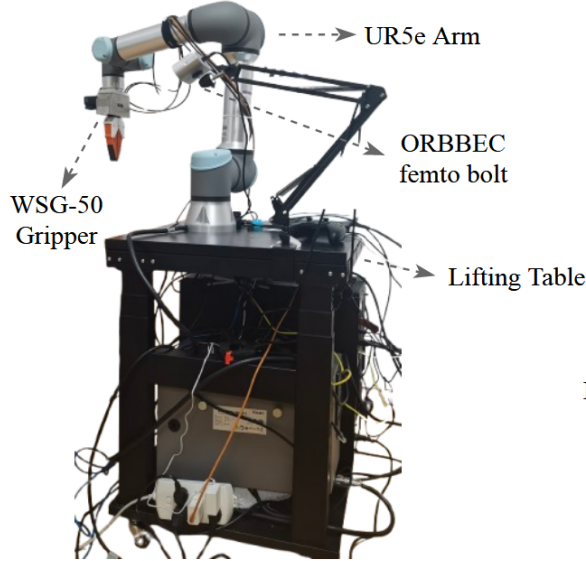


Figure 4: Hardware setup for real-world experiments, featuring a UR5e manipulator, WSG-50 gripper, ORBBEC Femto Bolt camera, and a mobile lifting table.

A.3 Qualitative Results in the Real World

To further demonstrate the effectiveness and robustness of our proposed SkillMemo framework in physical environments, we provide qualitative visualizations of the real-world task executions. As shown in Figure 5, we visualize the continuous execution frames of four distinct compositional manipulation tasks: *Place the Strawberry into the Bowl*, *Place the Corn into the Pot*, *Put the Butter into the Pot*, and *Put the Chip Bucket onto the Table*.

The visualizations confirm that by retrieving and composing latent atomic skills from the memory bank, the UR5e manipulator can smoothly and accurately handle varied objects and receptacles, exhibiting precise grasping and placing behaviors even under complex real-world dynamics.

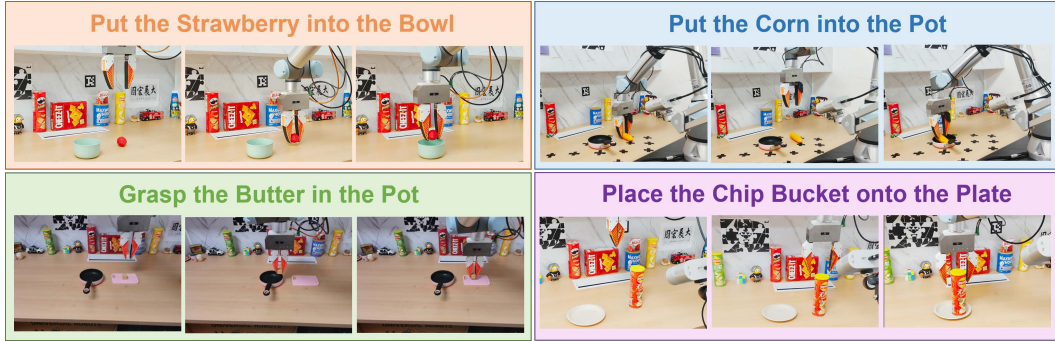


Figure 5: Visualization of successful rollouts for real-world manipulation tasks using the SkillMemo framework. The continuous frames illustrate the UR5e robot effectively executing diverse compositional tasks, demonstrating accurate skill retrieval, robust grasping, and precise placement across different object-receptacle combinations.