Learning What to Share and What to Personalize: Hierarchical Strategy Co-Evolution for Agent Memory

Yupeng Han Shuochen Liu Kai Zhang* Ze Liu Zhihong Pan Xianquan Wang

State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China
yupenghan@mail.ustc.edu.cn kkzhang08@ustc.edu.cn

Abstract

Memory-augmented agents maintain compact user profiles throughout extended conversations, enabling personalized and consistent responses without the need to process the entire dialogue history. The quality of these user profiles relies on the underlying memory management strategy: at each step, the agent must determine what to retain, compress, or discard. However, existing methods typically employ a static, one-size-fits-all strategy established before training. In practice, the optimal memory decision is inherently user-specific and dynamically evolves alongside policy optimization. To address this, we propose **HiPS** (**Hi**erarchical **P**ersonalized **S**trategy), a framework that decouples memory management into a globally shared foundation and a user-specific adaptive tier. Specifically, HiPS employs **Universal Strategy** to extract shared principles from cross-persona trajectories, alongside **Persona Delta Distillation** to generate tailored rules for users whose behaviors diverge from general patterns. **Cross-Level Rule Flow** dynamically calibrates their boundary by promoting broadly validated personal rules and demoting contradicted global ones. The architecture establishes a co-evolution loop where a mechanism guarantees that all strategy refinements are anchored to task outcomes. Extensive experiments demonstrate consistent improvements over memory-augmented baselines[1].

1 Introduction

Large language models (LLMs) are increasingly deployed as personalized conversational agents, which require maintaining coherent and tailored interactions for individual users (Jiang et al., 2025; Zhao et al., 2025). However, the inherent constraints of finite context windows prohibit LLMs from retaining unbounded dialogue histories (Xu et al., 2026b; Gao et al., 2025). Moreover, naive storage and retrieval of raw dialogue fail to capture the dynamic, evolving user preferences (Zhang et al., 2026b). This challenge motivates the integration of external memory systems, requiring the agent to dynamically evaluate at each turn whether to retain, compress, or discard incoming context.

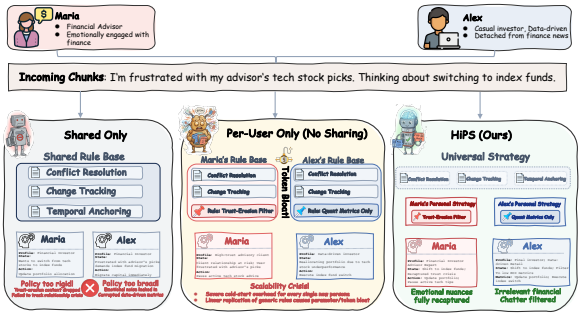

Figure 1: Same conversation, different users, different needs. A shared strategy (left) applies identical rules to all users, losing Maria's emotional context while over-preserving Alex's trivial details. A hypothetical per-user independent approach (center) would redundantly rediscover universal rules and fail on cold-start users. HiPS (right) decomposes the strategy into a shared level and per-user adaptive level.

Along this line, most existing memory systems employ static workflows, converting raw dialogue into external memory via predefined extraction and compression rules (Chhikara et al., 2025; Xu et al., 2026b; Fang et al., 2026). However, they lack the capacity to learn from interaction feedback or adapt to diverse user behaviors. Recent approaches formalize memory operations as learnable actions, optimizing update policies through reinforcement learning (RL) (Ouyang et al., 2026; Yu et al., 2026b; Zhou et al., 2026) or natural-language strategy distillation (Xu et al., 2026a). While more adaptive, they impose a strictly user-agnostic management paradigm. When optimizing for population-averaged rewards, the learning signals for niche behaviors are overwhelmed by the majority (Poddar et al., 2024), resulting in a one-

*Corresponding author.
[1]https://github.com/Hyp26cs/HiPS

size-fits-all compromise. Moreover, methods utilizing explicit strategies (Xu et al., 2026a) keep these rules frozen during training. Blind to online feedback, the prescribed strategy misaligns with actual rollout trajectories as the policy evolves, making it imperative that strategies be both *personalized* and *adaptive* (Hu et al., 2026b). Yet, individualizing every operation is redundant, as foundational rules are universally beneficial while niche behaviors require tailored handling. Since the optimal boundary is unknown a priori, this naturally raises the question: *Can we dynamically discover and co-evolve this partition using on-policy evidence?*

To this end, we introduce **HiPS** (**Hi**erarchical **P**ersonalized **S**trategy), a framework that decomposes the management strategy into a shared universal level and a per-user adaptive level, dynamically learning the boundary between them from on-policy evidence. As illustrated in Fig. 1, HiPS contains three core mechanisms: (1) **Universal Strategy Distillation (USD)** evolves shared rules using cross-persona trajectories. By contrasting high- and low-reward episodes, USD generates structured revisions that validate or refine rules. (2) **Persona Delta Distillation (PDD)** formulates adaptive, per-user rules specifically for individuals whose behaviors diverge from the population norm. A divergence criterion selectively identifies users who need personalization, thereby preventing overfitting and noise for those adequately served by universal rules. Finally, (3) **Cross-Level Rule Flow** dynamically calibrates the boundary between these tiers by promoting broadly validated personal rules to the universal level and substituting contradicted global rules with targeted PDD revisions. Unlike previous approaches that freeze strategies before training (Xu et al., 2026a), HiPS interleaves strategy discovery and policy optimization within a unified, continuous loop. In this paradigm, the active strategy directs model rollouts to generate trajectories that subsequently serve as empirical feedback for further strategy refinement. To prevent self-reinforcing confirmation bias (Tan et al., 2025a), strategy evolution is anchored strictly to objective task outcomes, reserving rule adherence exclusively for policy optimization. Our contributions are summarized as below:

- We formulate the **strategy personalization problem**, demonstrating that the boundary between shared and user-specific memory rules must be discovered empirically rather than predefined.
- We propose **HiPS**, an RL framework that co-evolves a universal strategy and user-specific adaptive rules alongside a policy via evidence-based distillation and cross-level rule flow.
- Extensive experiments demonstrate consistent performance gains. We uncover a **component importance flip** where universal rules dominate in-domain tasks and adaptive mechanisms drive out-of-domain generalization.

2 Related Work

Memory architectures. LLM agent memory systems manage how historical information is stored, organized, and retrieved within a bounded context window. Memory bank approaches apply segmentation, summarization, and selective forgetting to maintain long-term quality (Chhikara et al., 2025; Xu et al., 2026b; Fang et al., 2026). Structured indices such as tree- and graph-based retrieval improve access efficiency (Packer et al., 2024; Li et al., 2025). Personalization-oriented systems extract and maintain user profiles for downstream conditioning (Jiang et al., 2025). Despite their diversity, these systems share two fundamental limitations: their management strategies are **fixed**, rendering them incapable of learning from interaction feedback, and **shared**, imposing identical rules across all users regardless of behavioral differences.

RL-trained and strategy-based memory policies. To overcome the rigidity of fixed pipelines, recent studies formulate memory operations as learnable actions via reinforcement learning (RL) (Yan et al., 2025; Wang et al., 2026; Yu et al., 2026b; Zhou et al., 2026; Ouyang et al., 2026). While these methods effectively learn parametric and procedural *policies*, they fail to develop declarative and interpretable *strategies*: management behaviors are implicitly encoded within model parameters, rendering them neither inspectable nor editable. A parallel line of research formulates strategies as natural-language rules. For instance, MemCoE (Xu et al., 2026a) optimizes management guidelines using TextGrad before RL training, subsequently injecting them as system prompts for GRPO fine-tuning. EverMemOS (Hu et al., 2026a) introduces skill retirement through experience clustering and distillation, while MemSkill (Zhang et al., 2026a) reframes traditional static memory operations into learnable and evolvable "memory skills", dynamically optimizing both skill selection and

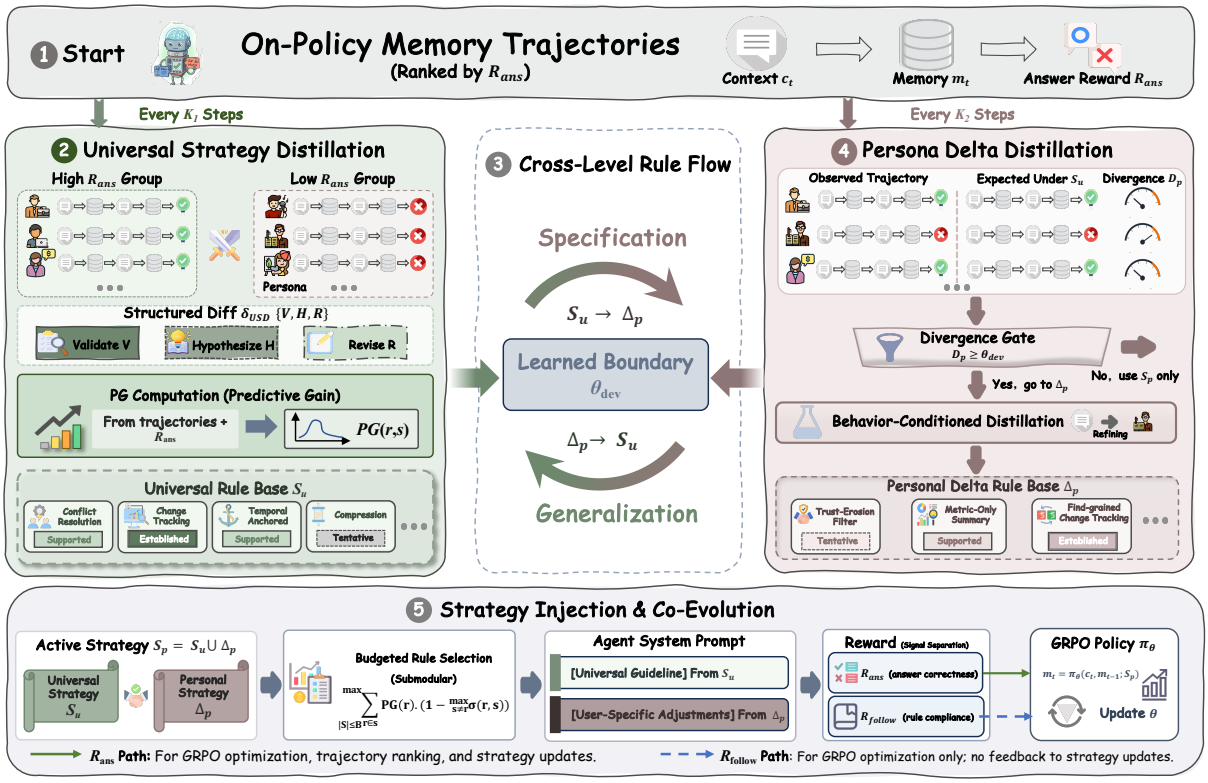

Figure 2: HiPS framework overview. Universal Strategy Distillation (USD) evolves shared strategies S_u from cross-persona trajectories. Persona Delta Distillation (PDD) evolves per-user adaptations Δ_p for users whose divergence exceeds θ_{div}. Cross-Level Flow promotes widely validated deltas to S_u. The active strategy $S_p = S_u \cup \Delta_p$ is injected into the agent prompt via submodular selection.

refinement. Although these approaches yield interpretable strategies, the resulting rules are inherently **globally shared and frozen prior to training**: a single guideline serves all users and remains static once optimized. HiPS differs by employing *user-indexed* strategies that dynamically *co-evolve* with the policy driven by on-policy evidence, rather than relying on static, globally shared rules.

3 Preliminary

We consider a conversational setting in which a memory-augmented agent interacts with a user p over multiple sessions. The dialogue history is segmented into contextual chunks $\{c_1, \ldots, c_K\}$, processed sequentially by the agent π_θ. To facilitate long-term personalization, the agent maintains a compact memory state m_t that is updated as new information arrives. The memory update mechanism is governed by a set of management rules, denoted as S_p, which is injected into the agent's system prompt to regulate its actions at step t:

$$m_t = \pi_\theta(c_t, m_{t-1};\, S_p). \tag{1}$$

Prior work treats the management strategy as a single, globally shared entity that is fixed prior to training. This introduces two primary limitations: the strategy lacks user-specific adaptability, and it remains static as the policy evolves. To overcome these limitations and enable dynamic adaptation, we decompose S_p into shared and personalized components, allowing this partition to co-evolve with the policy during training. Formally,

$$S_p = S_u \cup \Delta_p, \tag{2}$$

where S_u comprises universally beneficial rules applicable to the entire population, and Δ_p contains adaptive rules tailored to the behavioral patterns of user p. After processing all K dialogue chunks, the agent generates a final response to a query q conditioned on the terminal memory state m_K.

4 Method

In this section, we propose HiPS, a framework that decouples the memory management strategy into a shared universal tier and a user-specific adaptive tier. Specifically, Universal Strategy Distillation (USD) (Section 4.1) abstracts broadly applicable

principles from cross-persona trajectories. Subsequently, Persona Delta Distillation (PDD) (Section 4.2) captures idiosyncratic rules specifically for users exhibiting divergent behaviors. Finally, a Cross-Level Rule Flow mechanism (Section 4.3) calibrates the partition by migrating rules between the two tiers as evidence accumulates. The overall architecture of HiPS is illustrated in Fig. 2.

4.1 Universal Strategy Distillation (USD)

Existing strategies are either manually crafted or learned from data, yet both are subsequently frozen, thereby failing to incorporate dynamic evidence accumulated as the policy evolves. To break this static paradigm, S_u should be continuously refined using live trajectories. While contrastive analysis of trajectories can identify promising patterns, distilling these insights into reliable rule updates presents two challenges: (1) Conventional free-form feedback treats the guideline as a monolithic text, lacking per-rule granularity. (2) Without granular evidence tracking, a rule proposed in one cycle may be arbitrarily overwritten in the next. USD resolves these issues by utilizing *structured diffs* to decompose updates into rule-level operations, coupled with a rigorous evidence tracking mechanism to maintain a persistent validation history.

Contrastive feedback as structured diff. At regular training intervals, USD samples a persona-balanced set of the top-k and bottom-k trajectories to prevent dominant users from skewing the optimization signal. An LLM meta-optimizer is prompted to compare these contrastive sets. Instead of yielding a free-form summary, the optimizer outputs a structured diff, a discrete set of operations applied to the current rule set (see Appendix F for prompt). Formally,

$$\delta_{\text{USD}} = \{\text{V} : [\cdot],\ \text{H} : [\cdot],\ \text{R} : [\cdot]\}, \tag{3}$$

where V signifies validating an existing rule, H represents hypothesizing a new rule, and R denotes revising or retiring a rule. Crucially, new hypotheses must strictly adhere to the "[Label]: When [condition], [action]" format and prescribe explicit management behaviors.

Evidence tracking. Since a single distillation cycle may yield spurious hypotheses, USD incorporates strict evidence tracking to filter out noise. Each rule's evidence level, including `supported`, `established`, and `tentative` (see Appendix E for details) is iteratively updated based on its validation signals: consistent validation operations promote the rule, whereas revision or contradiction operations demote it. To complement the LLM judgment with a data-driven safeguard, rules are automatically promoted if their predictive gain (Eq. 4) exceeds θ_{val}, and demoted if it falls below θ_{rev}.

4.2 Persona Delta Distillation (PDD)

Under globally shared strategies, rules benefiting a minority receive diluted signals while locally harmful ones persist. Although per-user adaptation resolves this, naively generating Δ_p for everyone wastes computational budget and introduces noise for those adequately served by S_u. To mitigate this, PDD employs a *divergence criterion* to gate personalization, selectively identifying users who genuinely require adaptive rules.

Divergence-gated personalization. To quantify behavioral deviation, we first evaluate the extent to which each universal rule correlates with task success. We define the *predictive gain* of a rule r as the absolute difference in the expected success rate (denoted by $Y \in \{0, 1\}$) between trajectories that comply with r and those that do not:

$$\text{PG}(r) = \left| P(Y{=}1 \mid r^{+}) - P(Y{=}1 \mid r^{-}) \right|, \tag{4}$$

where r^{+} and r^{-} indicate rule compliance and non-compliance, respectively. PG provides a proxy for rule importance: rules with high PG strongly predict task outcomes, while near-zero PG signals irrelevance. Crucially, PG is used only for relative ranking and threshold-based gating, not as an absolute measure of rule quality. Building upon this, we define the divergence of user p as the aggregate discrepancy in rule efficacy between the specific user and the broader population:

$$D_p = \sum_{r \in S_u} \left| \text{PG}(r \mid p) - \text{PG}(r) \right|. \tag{5}$$

Users exhibiting a divergence $\mathcal{D}(p) \geq \theta_{\text{div}}$ trigger the PDD module. Conversely, users below this threshold are deemed sufficiently covered by S_u and bypass the personalization step. This gating mechanism resolves the cold-start dilemma by defaulting low-divergence users to the universal baseline, optimizing resource allocation.

Behavior-conditioned distillation. For users identified as divergent, PDD prompts an LLM to

formulate management behaviors that strictly differ from S_u, using the universal rules as an anchoring context. To prevent Δ_p from degenerating into mere factual knowledge, the generated rules are constrained to be *management-oriented* to prescribe actions rather than asserting facts and *behavior-conditioned*. This ensures robust generalization to unseen users at inference time.

4.3 Cross-Level Rule Flow

If S_u and Δ_p evolve in strict isolation, the boundary between them becomes brittle. Without a dynamic transfer mechanism, a personalized rule that ultimately proves globally beneficial remains permanently trapped in Δ_p, while a universal rule contradicted by a subset of users persists erroneously in S_u. To resolve this structural rigidity, the Cross-Level Rule Flow dynamically calibrates the partition by migrating rules between tiers based on accumulated cross-persona evidence.

Generalization ($\Delta_p \to S_u$). When a Δ_p rule at "supported" level or above appears in $\geq \theta_{\text{flow}}$ fraction of personas that have Δ_p, it is promoted to S_u and pruned from individual deltas. Intuitively, a management pattern emerging across diverse users signifies a universal principle rather than a localized adaptation. To ensure robust aggregation, we match rules via semantic similarity, consolidating equivalent behaviors despite syntactic variations.

Specialization ($S_u \to \Delta_p$). When USD revises or demotes a rule within S_u, users who previously benefited from the original rule risk losing critical guidance. To mitigate this, PDD hypothesizes persona-specific replacements, effectively restoring the management behavior in a localized form. This mechanism prevents a one-size-fits-all revision from deteriorating the performance of minority users who relied on the deprecated rule.

Together, generalization and specialization establish a cycle. Local discoveries that prove widely beneficial are assimilated into S_u, whereas universal rules that are contradicted by specific subpopulations are replaced by tailored adaptations in Δ_p.

4.4 Strategy Injection and Co-Evolution

In this co-evolutionary framework, the active strategy directs policy rollouts, while the resulting trajectories provide empirical feedback for subsequent strategy refinement. We link these phases using budgeted rule injection and a combined reward that decouples strategy evolution from adherence.

Submodular selection. The active strategy $S_p = S_u \cup \Delta_p$ is constrained by a maximum token budget B. To select the most representative and diverse set of rules, we greedily maximize a submodular coverage objective penalized by redundancy:

$$\max_{\substack{S \subseteq S_p \\ \sum_{r \in S} |r| \leq B}} \sum_{r \in S} \text{PG}(r) \cdot \big(1 - \max_{s \neq r} \sigma(r, s)\big), \quad (6)$$

where $\sigma(\cdot, \cdot)$ denotes semantic similarity. The token budget is allocated between the universal and personalized levels proportionally to their aggregate predictive gains, subject to a minimum floor for each level as detailed in Appendix C.

Guideline-aligned reward. To align the policy with the discovered strategies, we define a persona-aware adherence reward:

$$R_{\text{follow}}(\tau, p) = \frac{\sum_{r \in S_p} \text{PG}(r) \cdot h(\tau, r)}{\sum_{r \in S_p} \text{PG}(r)}, \quad (7)$$

where $h(\tau, r) \in [0, 1]$ is a compliance check. The combined reward for Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is $R(\tau, p) = R_{\text{ans}}(\tau) + \lambda \cdot R_{\text{follow}}(\tau, p)$. Consequently, the identical trajectory naturally yields varying adherence scores depending on the user's specific Δ_p.

Anti-circularity and co-evolution. A rule could validate itself if compliance inflates the signal used to select which trajectories inform strategy updates. We disrupt this direct path by ranking the distillation buffer exclusively by R_{ans}, while restricting R_{follow} to GRPO advantage computation. An indirect path remains where R_{follow} shapes the policy, which alters rollout behavior and subsequently affects future R_{ans}. However, its indirect path is benign when rule compliance genuinely improves task success, which is precisely the condition under which strategy refinement should occur. A vulnerability arises only when compliance is rewarded without corresponding task improvement. Separating R_{ans} for distillation from R_{follow} for policy optimization prevents this reinforcement loop. The detailed training procedure, detailed in Appendix F, integrates these components into a unified loop.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate on four personalized memory benchmarks: **PersonaMem** (Jiang et al.,

Method	PersonaMem		PrefEval		PersonaBench				PERMA			
	32K	128K	Expl.	Impl.	0	0.3	0.5	0.7	C-S	C-M	N-S	N-M
Long Context	42.17	20.74	32.40	25.90	27.77	22.38	19.45	12.59	24.11	22.88	24.82	22.81
RAG	52.41	38.02	47.40	32.40	32.21	29.13	25.28	23.36	51.63	36.84	51.77	35.09
Mem0	48.53	39.67	57.60	46.40	17.43	18.95	19.31	16.49	52.76	51.67	51.49	51.15
A-Mem	55.42	39.88	62.30	52.80	30.09	28.52	25.81	24.10	53.05	53.73	52.91	49.87
LightMem	52.41	36.74	64.20	54.80	19.14	17.83	19.61	17.51	59.43	55.53	59.43	55.53
MemAgent	59.93	50.86	77.40	64.40	30.74	28.34	24.47	23.53	53.98	53.90	55.74	48.84
Mem-α	58.43	48.18	75.70	63.10	19.31	18.01	17.35	16.64	52.19	51.67	54.46	46.53
MemSkill	64.45	55.37	82.60	68.40	31.19	**29.89**	24.85	23.36	62.83	51.92	60.14	53.21
HiPS (Ours)	**73.49**	**62.01**	**89.20**	**69.40**	**32.27**	29.76	**25.99**	**25.09**	**66.95**	**56.56**	**63.83**	**56.30**

Table 1: **Overall comparison across twelve evaluation settings.** PersonaMem (32K/128K) is in-domain; PrefEval (Explicit/Implicit), PersonaBench (4 noise levels), and PERMA (C-S/C-M/N-S/N-M denote Clean/Noise $\times$ Single/Multi-Domain) are out-of-domain. Higher is better. Best results are in **bold**.

2025) for preference evolution over varying context scales; **PrefEval** (Zhao et al., 2025) for explicit and implicit queries with 50 distraction turns; **PersonaBench** (Tan et al., 2025b) for personalized retrieval and QA over corpora; and **PERMA** (Liu et al., 2026) combining two noise levels with two task scopes. We report accuracy for PersonaMem, PrefEval, and PERMA, and the F1 score for PersonaBench. See Appendix A for details.

Baselines. We compare our framework against a diverse selection of baselines. **LongContext** feeds raw histories directly into the context window, whereas **RAG** retrieves the top-k relevant segments from a vector store. We also include three retrieval-based methods that dynamically maintain an external memory bank: **Mem0** (Chhikara et al., 2025), **A-Mem** (Xu et al., 2026b), and **LightMem** (Fang et al., 2026). Furthermore, we compare against three RL-based memory agents that learn memory evolution actions: **MemAgent** (Yu et al., 2026a), **MEM-α** (Wang et al., 2026), and **MemSkill** (Zhang et al., 2026a). For a fair comparison, all baselines use the same configurations.

Implementation. We employ Qwen2.5-7B-Instruct (Qwen et al., 2025) as the backbone language model for most methods, whereas MEM-α utilizes Qwen3-4B and MemSkill uses Qwen3-Next-80B-A3B-Instruct. For retrieval, all-MiniLM-L6-v2 (Wang et al., 2020) extracts the top ten candidates. Our training set consists of 423 instances sampled from 70% of the PersonaMem 32k subset. To reduce computational overhead, training uses retrieved dialogues as context, whereas inference utilizes the full history, processing context in 4K-token chunks. Baselines are implemented using their public repositories; for a fair comparison, MemAgent, MEM-α and MemSkill are initialized from public checkpoints and fine-tuned on our 423 samples. All experiments run on four NVIDIA H800 GPUs, with shared hyperparameters detailed in Appendix C.

5.2 Main Results

Overall Comparison with Baselines. As shown in Tab. 1, HiPS consistently demonstrates superior performance across all twelve evaluation settings. This performance gap indicates that learning an explicit, dynamically evolving memory-management strategy is significantly more effective than relying on fixed context inclusion or manually designed retrieval heuristics. Specifically, the performance of the Long Context baseline degrades severely under noisy, long-horizon histories, falling to 20.74 on PersonaMem 128K. In contrast, HiPS maintains stable personalized performance by utilizing its distilled strategies to filter out irrelevant information during memory updates. Compared to explicit memory-bank baselines such as Mem0, A-Mem, and LightMem, HiPS delivers substantial and consistent improvements across both in-domain and out-of-domain tasks. Although reinforcement learning-based memory agents, specifically MemAgent, MEM-α, and MemSkill, are competitive, they still lag behind the holistic performance of our framework. This consistent advantage highlights the efficacy of our hierarchical co-evolution design, where universal strategy dis-

Configuration	PersonaMem		PERMA			
	32K	128K	C-S	C-M	N-S	N-M
HiPS (Full)	**73.49**	**62.01**	**66.95**	**56.56**	**63.83**	**56.30**
w/o Flow	71.08	60.49	45.39	40.87	43.12	36.50
w/o Gate	69.28	59.63	51.63	44.99	52.62	46.79
w/o USD	66.27	53.70	61.28	51.93	60.14	49.10
w/o PG	67.47	52.22	60.70	51.41	59.43	53.98
w/o PDD	69.88	56.05	60.99	53.98	60.71	51.41

Table 2: Ablation results on PersonaMem and PERMA. Cell color intensity indicates drop from Full (<3 to ≥15 pp). Rows ordered by average drop.

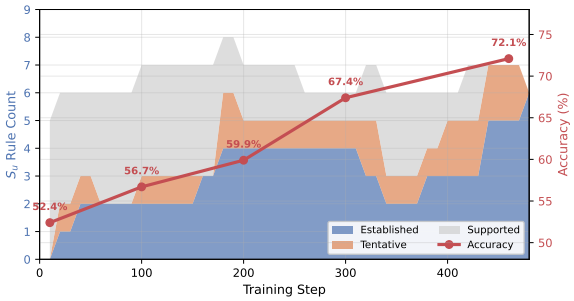

Figure 3: Evolution of S_u evidence distribution and accuracy throughout training on PersonaMem.

tillation induces a transferable global guideline, and persona delta distillation dynamically adapts to user-specific behavioral patterns.

Generalization Across Settings. Across the performance detailed in Tab. 1, HiPS exhibits robust generalizability. It consistently outperforms baselines on both in-domain tasks and out-of-domain benchmarks. This includes the explicit and implicit preference queries of PrefEval, the increasingly noisy contexts of PersonaBench, and the diverse domain crossings of PERMA. This strong generalization capability is driven by our hierarchical strategy decomposition, where universal strategy distillation establishes stable, globally shared memory organizations, and persona delta distillation handles targeted adaptations for divergent users. Detailed performance comparisons across different user categories are provided in Appendix B.

5.3 Ablation Study

Tab. 2 ablation results reveal a striking in-domain versus out-of-domain importance flip between PersonaMem and PERMA. In-domain, removing **USD** or **PG** causes the largest declines, reducing PersonaMem 128K accuracy from 62.01 to 53.70 and 52.22, respectively, confirming that universal rule quality is critical when distributions align. Under this setting, **PDD** yields a moderate drop to 56.05, while **Gate** and **Flow** reduce accuracy only to 59.63 and 60.49, as training personas are well-served by S_u alone. Conversely, this ranking reverses out-of-domain. Ablating **Flow** triggers the most severe degradation, reducing C-S accuracy from 66.95 to 45.39, an average PERMA decrease of 19.4 points, while removing **Gate** drops C-S accuracy to 51.63, averaging an 11.9-point loss. These drops show that cross-level rule migration and divergence-gated personalization are essential when universal rules do not transfer, especially in multi-domain settings such as C-M and N-M, where cross-domain interests stress-test persona diversity. In contrast, USD and PG matter less out-of-domain, dropping performance by only 5.3 and 4.5 points on average. Finally, ablating **PDD** causes consistent, moderate drops of 6.0 on PersonaMem 128K and 4.1 on average across PERMA.

5.4 Strategy Quality Analysis

Evolutionary Dynamics of Universal Strategies. Fig. 3 traces S_u evidence distribution and accuracy throughout training on PersonaMem 32K validation set. Starting from 5 seed rules (all supported), the system hypothesizes new rules (appearing as tentative) and validates successful ones. The stacked area shows evidence of evolution: rules progressively mature from tentative to supported to established, with accuracy rising from 52.4% to 72.1% by convergence. The rule count oscillates before stabilizing at 6 established rules at step 480+, reflecting the create-validate-prune lifecycle.

Qualitative Analysis of Personalized Deltas. Fig. 4 illustrates how the impact of Δ_p varies across user profiles. For P15, the finance enthusiast, a specialized financial logging rule yields an 18.6% point improvement by preventing granular data compression. Similarly, P2, the programmer, gains 17.9% points by preserving conflicting portfolio entries instead of overwriting them, while P0, the musician, gains 15.8% points through custom milestone tracking. Conversely, P6, the marketing specialist, experiences a 3.3% point drop. Because P6's professional activities are already well-managed by S_u, additional rules over-constrain model generation, validating our divergence-gating design to selectively apply personalization.

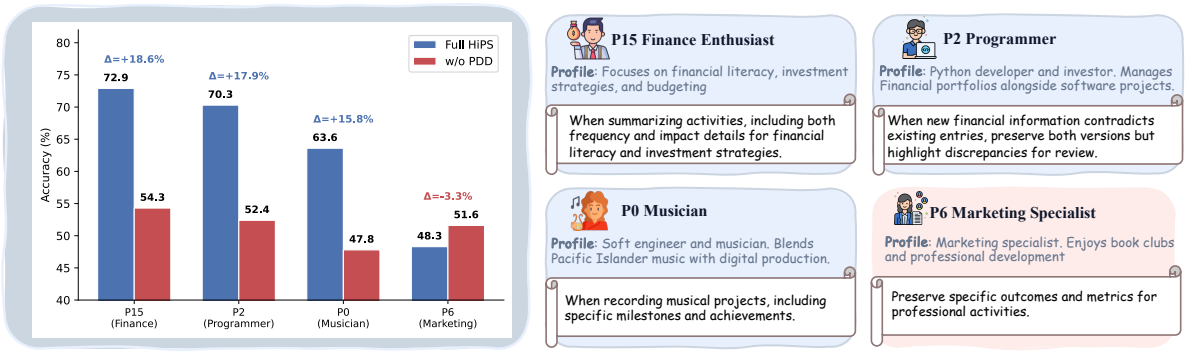

Figure 4: Per-persona impact of persona-specific strategies (Δ_p). We compare Full HiPS (with Δ_p) against w/o PDD (without Δ_p) on PersonaMem 128K. PDD yields an average improvement of +6.0%, benefiting 20 personas. However, the gains are highly heterogeneous: P15 (finance enthusiast) gains +18.6%, while P6 (marketing specialist) suffers -3.3%. We highlight four representative personas to illustrate why.

Method	Qwen2.5-7B Instruct	GPT-4o -mini	Gemini 2.5 flash	GPT-5
RAG	52.41	51.58	60.83	63.74
A-Mem	55.42	55.91	62.74	65.93
▼ *Optimized w/ Qwen2.5-7B-Instruct*				
HiPS (S_u only)	57.83	56.74	64.25	67.18
HiPS (S_u+Δ_p)	**61.24**	59.83	67.45	70.16
▼ *Optimized w/ GPT-4o-mini*				
HiPS (S_u only)	56.47	57.35	65.18	68.42
HiPS (S_u+Δ_p)	60.07	**61.35**	**68.12**	**71.48**

Table 3: Cross-model transferability of strategies (without RL policy). Strategies are distilled with one LLM and injected into others' prompts for inference.

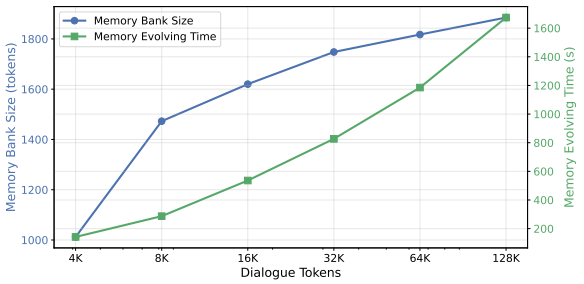

Figure 5: Scaling analysis on PersonaMem (128K). We increase the total dialogue tokens (each evolution round processes 4K tokens) and report the resulting memory bank size (left) and memory evolving time (right).

5.5 Cross-model Transfer

Tab. 3 presents the transferability of the distilled strategies across various backbone LLMs. Both HiPS variants consistently outperform baselines such as RAG and A-Mem across all models, demonstrating that our strategy captures model-agnostic memory management principles rather than overfitting to a specific LLM. Moreover, the S_u+Δ_p configuration surpasses the S_u-only baseline, confirming that hierarchical personalization provides complementary benefits across backbones. Notably, optimization via `GPT-4o-mini` generalizes well, achieving strong results on three backbones, including `GPT-5` at 71.48 and `Gemini 2.5 flash` at 68.12. These results confirm that HiPS produces portable strategies, enabling deployment across diverse backbone models.

5.6 Scaling Analysis

Fig. 5 analyzes how our framework scales with longer dialogue context on PersonaMem (128K). As the dialogue tokens grow from 4K to 128K, the memory bank size increases from 1,000 to around 1,900 tokens, and the curve exhibits a clear trend: it rises steadily in the short-context regime and gradually flattens as the dialogue lengthens, which demonstrates that HiPS effectively consolidates preference-relevant information while filtering out redundant content to keep memory growth bounded. Meanwhile, the memory evolving time scales approximately linearly with dialogue length, ranging from roughly 140 to 1,700 seconds across the full spectrum. This indicates that the computational overhead remains predictable and well-controlled, confirming that HiPS can efficiently handle ultra-long dialogues without incurring prohibitive memory or time costs.

6 Conclusion

In this paper, we introduce HiPS, a framework that decouples memory management strategies into a universal baseline and a user-specific adaptive delta, with the partition updated using online evidence. Our empirical results verify that HiPS secures consistent performance gains over retrieval and RL-

trained counterparts, with its advantages becoming pronounced as the dialogue context scales. By establishing a continuous co-evolution loop between multi-tiered strategies and active policies, this work paves the way for developing robust and truly personalized long-term memory systems for agents.

Limitations

While HiPS demonstrates robust improvements across diverse memory benchmarks, we identify several gentle limitations that offer promising avenues for future research. First, our evaluation is primarily focused on English-language conversational benchmarks. While the core algorithmic principles of strategy distillation and policy co-evolution are model-agnostic, memory management strategies in other typological languages may exhibit different syntactic structures or require localized prompt adjustments. Testing our framework in multilingual and cross-lingual settings remains an open area for future work. Second, our experiments evaluate interaction histories spanning up to 1M tokens, which represents a medium-to-long timeline of conversational sessions. Over extremely long-term lifecycles, such as years of continuous daily interaction, a user's fundamental baseline personality traits may undergo gradual, paradigm-level shifts. Handling such slow-moving, long-term personal baseline drift would require additional meta-distillation mechanisms to update the universal seed rules themselves occasionally.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (Grants No.62406303).

References

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *Preprint*, arXiv:2504.19413.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2026. Lightmem: Lightweight and efficient memory-augmented generation. *Preprint*, arXiv:2510.18866.

Pengyu Gao, Jinming Zhao, Xinyue Chen, and Long Yilin. 2025. An efficient context-dependent memory framework for LLM-centric agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 1055–1069, Albuquerque, New Mexico. Association for Computational Linguistics.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. 2026a. Evermemos: A self-organizing memory operating system for structured long-horizon reasoning. *Preprint*, arXiv:2601.02163.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2026b. Memory in the age of ai agents. *Preprint*, arXiv:2512.13564.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *Preprint*, arXiv:2504.14225.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, and 3 others. 2025. Memos: An operating system for memory-augmented generation (mag) in large language models. *Preprint*, arXiv:2505.22101.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. 2026. Perma: Benchmarking personalized memory agents via event-driven preference and realistic task environments. *Preprint*, arXiv:2603.23231.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. 2026. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *The Fourteenth International Conference on Learning Representations*.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Chuyi Tan, Peiwen Yuan, Xinglin Wang, Yiwei Li, Shaoxiong Feng, Yueqi Zhang, Jiayi Shi, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. 2025a. Diagnosing and mitigating system bias in self-rewarding rl. *Preprint*, arXiv:2510.08977.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R N, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025b. PersonaBench: Evaluating AI models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 878–893, Vienna, Austria. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. 2026. MEM-α: LEARNING MEMORY CONSTRUCTION VIA REINFORCEMENT LEARNING.

Derong Xu, Shuochen Liu, Pengfei Luo, Pengyue Jia, Yingyi Zhang, Yi Wen, Yimin Deng, Wenlin Zhang, Enhong Chen, Xiangyu Zhao, and Tong Xu. 2026a. Learning how and what to memorize: Cognition-inspired two-stage optimization for evolving memory. *Preprint*, arXiv:2605.00702.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026b. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, and 1 others. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2026a. Memagent: Reshaping long-context LLM with multi-conv RL-based memory agent. In *The Fourteenth International Conference on Learning Representations*.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. 2026b. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. *Preprint*, arXiv:2601.01885.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026a. Memskill: Learning and evolving memory skills for self-evolving agents. *Preprint*, arXiv:2602.02474.

Yingyi Zhang, Junyi Li, Wenlin Zhang, Pengyue Jia, Xianneng Li, Yichao Wang, Derong Xu, Yi Wen, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2026b. Evoking user memory: Personalizing LLM via recollection-familiarity adaptive retrieval. In *The Fourteenth International Conference on Learning Representations*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low, and Paul Pu Liang. 2026. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. In *The Fourteenth International Conference on Learning Representations*.

A Datasets

A.1 PersonaMem Benchmark

PersonaMem (Jiang et al., 2025) is a large-scale benchmark designed to evaluate long-term personalization in conversational LLMs. It features interaction histories for **20 simulated personas**, with each persona defined by rich static attributes, such as demographics, alongside dynamic traits and preferences that evolve across **15 diverse real-world task domains**, including food recommendation, travel planning, and therapy consultation. For every persona, multi-session conversations are constructed where the user engages with a chatbot via **7 types of in-situ queries** that probe distinct personalization capabilities, such as recalling user facts, tracking preference evolution, and providing preference-aligned suggestions. Each session consists of 15 to 30 user-assistant turns. These

Statistic	PersonaMem		
Tokens per history	~32k	~128k	~1M
# QA pairs	589	2727	2674
# Sessions per history	10	20	60
Avg. # utterances	167.1	758.3	3607.9

Table 4: Statistics of the PersonaMem dataset at different context lengths. Token counts denote the approximate total context length per interaction history; utterance counts are averaged over histories.

Statistic	Value
Explicit queries	1,000
Implicit queries	1,000
Maximum inserted conversations	24
Maximum inserted turns	326
Avg. turns / conversation	13.58
Total tokens	108,102
Avg. tokens / conversation	4,504.25

Table 5: Dataset statistics for PrefEval multiple-choice classification.

histories are instantiated at three context scales by concatenating 10, 20, or 60 sessions, yielding approximate context lengths of 32k, 128k, and 1M tokens, respectively. At evaluation time, models must select appropriate responses to user queries conditioned on the interaction history, testing their ability to adapt to dynamic user profiles. Tab. 4 summarizes the main statistics of PersonaMem.

A.2 PrefEval Benchmark

PrefEval (Zhao et al., 2025) is a long-context, multi-session benchmark designed to evaluate whether LLMs can infer, retrieve, and act on user preferences in realistic conversational settings. The benchmark emphasizes four core aspects: preference inference, long-context retrieval, preference following, and personalization proactiveness. The dataset comprises 1,000 unique preference-query pairs and spans 20 everyday topics grouped into seven domains: *Entertainment*, comprising shows, music and books, sports, and games; *Travel*, spanning activities, restaurants, hotels, and transportation; *Lifestyle*, including diet, beauty, fitness, and health; *Shopping*, encompassing home, fashion, motors, and technology; *Education*, covering educational resources and learning styles; *Professional Ownership*; and *Professional Work Style*. PrefEval supports two evaluation formats, namely

User	Queries	Corpus	Conv.	AI	E-com.
1	48	110	84	23	3
2	43	90	78	8	4
3	42	64	51	12	1
4	46	85	71	14	0
5	44	84	59	21	4
6	40	94	79	14	1
Sum	263	527	422	92	13

Table 6: Statistics of the PersonaBench subset across six users. *Corpus* is the sum of *Conv.*, *AI*, and *E-com.*.

a free-form generation setting and a 4-way multiple-choice classification setting where exactly one option is consistent with the stated preference. To challenge long-range personalization capabilities, the benchmark inserts unrelated multi-session dialogue turns between the preference revelation and the final query. In our experiments, we employ 1,000 explicit and 1,000 implicit instances under the multiple-choice classification setting, with 50 intervening turns inserted as distractor context. Tab. 5 reports the summary statistics of this subset.

A.3 PersonaBench Benchmark

PersonaBench (Tan et al., 2025b) evaluates personalized retrieval and question answering grounded in user-specific contexts. For each user, the dataset provides a heterogeneous personal corpus consisting of conversations with friends, denoted as *Conv.*, dialogues with AI assistants, denoted as *AI*, and e-commerce purchase histories, denoted as *E-com.*. The evaluation queries are typically short and underspecified, requiring models to resolve implicit intent by grounding responses in evidence distributed across historical interactions and behaviors. This setup tests a model's ability to align with diverse, user-dependent semantics under realistic contextual ambiguity. Tab. 6 summarizes the per-user query counts and corpus statistics for the six-user subset used in our experiments.

A.4 PERMA Benchmark

PERMA (Liu et al., 2026) is an event-driven, long-context benchmark for evaluating whether personalized memory agents can maintain, update, and synthesize dynamic persona states in realistic conversational environments, with an emphasis on three aspects: task completion, preference consistency, and informational confidence. The dataset comprises 10 representative user profiles containing 2,166 fine-grained preference details across 10 countries, and spans 20 distinct domains,

Metric	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Total
User Demographics											
Age	35-44	25-34	65+	55-64	35-44	18-24	25-34	35-44	25-34	25-34	–
Gender	M	M	F	M	M	M	M	M	F	M	–
Education	Grad.	Univ.	–	Univ.	–	Grad.	Univ.	–	Grad.	Univ.	–
Country	CA	MX	FI	US	AU	UK	CH	IL	RU	BE	–
Interaction Statistics (Base Dataset: Clean / Noisy)											
# Interests	17	16	15	15	17	13	13	17	16	12	151
# Queries	63	62	60	59	63	51	50	63	60	49	580
# Events	81	81	85	79	82	79	75	81	81	84	808
# Dialogs	356/364	340/358	373/379	347/352	348/370	349/357	321/329	350/364	359/362	365/375	3.5k/3.6k
# Tokens (k)	34/34	32/33	33/36	31/31	33/33	32/32	31/31	32/36	34/39	33/34	324/331
Style-aligned Long-Context Dataset											
# Events	156	141	156	139	137	107	176	94	155	127	1,388
# Dialogs	1009	848	1084	876	888	802	1015	798	969	869	9,158
# Tokens (k)	139.0	61.4	174.0	88.9	197.2	69.0	151.1	62.7	143.7	78.4	1,165.4

Table 7: Comprehensive statistics of the dataset across multiple dimensions. The table summarizes user demographics for profiles, followed by statistics for the base (comparing Clean and Noise) and the style-aligned long-context dataset. Note: For User Demographics Profile (P), Gender: M (Male), F (Female); Education: Grad. (Graduate Degree), Univ. (University); Country codes follow ISO 3166-1 alpha-2. In Interaction Statistics, # denotes the count, and (k) represents thousands of tokens.

including *Travel*, *Finance*, *Shopping*, *Entertainment*, *Messaging*, and *Calendar*. PERMA supports two evaluation formats: an 8-option multiple-choice question (MCQ) probing setting in which options are systematically ablated, and a multi-turn interactive setting driven by an LLM-based user simulator that terminates upon successful task completion. To stress realistic, long-horizon personalization, the benchmark structures interactions temporally through event-driven timelines and injects controlled within-session noise to simulate real-world user erraticism. In our experiments, we evaluate memory agents across four key scenarios–*Clean-Single*, *Clean-Multi*, *Noise-Single*, and *Noise-Multi*–which span single- and multi-domain tasks under both standard dialogue histories and those perturbed with text-variability noise; comprehensive statistics of the dataset are reported in Tab. 7.

B Comparison in Different Categories

Comparison in Different Categories of PersonaMem. Tab. 8 reports category-wise results on PersonaMem under 32K and 128K interaction histories. Across both scales, **HiPS** achieves the best overall performance (73.49 at 32K; 62.01 at 128K), and the gains are concentrated on memory-dependent personalization abilities. On 32K histories, **HiPS** leads in Suggest ideas (51.85), Prefs evolve (84.44), Update reasons (89.66), Aligned recs (73.33), and New Scenarios (53.33), which jointly drives a clear margin over the strongest baselines (e.g., 73.49 vs. 64.45 for MemSkill). When scaling to 128K, *Long Context* degrades sharply overall (20.74), while memory-based methods remain substantially stronger; within them, **HiPS** stays best-performing and ranks first on Recall facts (68.42), Suggest ideas (44.02), Latest prefs (71.82), Prefs evolve (71.55), Update reasons (72.49), Aligned recs (54.73), and New Scenarios (44.13). In contrast, Recall facts is not a comparative strength for **HiPS** under the 32K scale (71.43 vs. 72.87 for MemAgent), indicating that while simpler fact-retrieval mechanisms can be effective in shorter horizons, **HiPS**'s architectural advantage lies in tracking and applying evolving, structured preferences as the context scales up. Overall, the category-wise gains suggest that explicitly *structuring* memory operations and then *selecting* what to keep is most effective for preference-heavy queries, and the advantage becomes more pronounced as the interaction history grows longer.

Comparison in Different Categories of PrefEval. Tab. 9 breaks down PrefEval performance by domain under **Explicit** and **Implicit** preference settings. Under Explicit Memory, **HiPS** attains the best overall accuracy (89.20) and shows consistently strong gains on preference-heavy domains, ranking first across all evaluated topics, including Travel (90.20), Entertain (94.20), Lifestyle

(91.70), Shop (84.10), Education (86.30), Professional (85.20), and Pet (78.10). This comprehensive lead demonstrates that **HiPS** can robustly extract and apply directly stated constraints. Under Implicit Memory, the task becomes more challenging for all methods, yet **HiPS** again leads overall (69.40) and improves most clearly on domains that require inferring latent preferences from context, such as Travel (71.90), Lifestyle (71.60), Shop (64.50), and Professional (63.40). A notable exception is Education, where **HiPS** (65.10) slightly trails MemSkill (65.80), suggesting that certain implicit reasoning contexts may occasionally benefit from alternative skill-driven memory-update pipelines. Compared with memory-bank baselines (Mem0/A-Mem/LightMem), the advantage of **HiPS** is broad across domains in both settings, indicating that it better resists long-range distractors and preserves preference-relevant signals. Overall, these domain-wise results reflect PrefEval's construction: the inserted unrelated turns make long-range preference retrieval and faithful preference following the main bottlenecks, and **HiPS** improves most on domains where precise preference identification and consistent application are essential.

C Hyperparameters

Tab. 10 lists all HiPS hyperparameters and their default values.

Hyperparameter Sensitivity Analysis To evaluate how key hyperparameters in HiPS influence personalization accuracy, we conduct a parameter sensitivity analysis on the PersonaMem 128K benchmark. As shown in the sensitivity curves, the trends demonstrate clear trade-offs across all key configurations: The distillation frequencies, K_1 and K_2, govern the update speed of universal and personalized rules. For the USD frequency K_1, the optimal performance is achieved at 20 steps. Too frequent distillation, such as every 5 steps, introduces short-term interaction noise, while too infrequent updates, such as every 80 steps, fail to adapt in a timely manner. For the PDD frequency K_2, the peak occurs at 10 steps, indicating that personalized strategies require more frequent updates than universal rules to capture rapid user-specific behavioral shifts. Regarding the strategy token budget B, performance improves sharply as the budget increases from 80 tokens, peaking around 200 to 300 tokens. Although 300 tokens yields a marginally higher accuracy, we select 200 as the default value to balance strategy expressiveness with prompt token efficiency. The divergence threshold θ_{div} controls personalization gating. Low values like 0.1 trigger excessive, unnecessary personalization that introduces content-level noise, whereas high values like 0.5 over-filter users, preventing divergent individuals from receiving tailored rules. The optimal threshold is 0.3. The flow threshold θ_{flow} regulates rule migration and exhibits robust performance across a wide range of values, reaching its peak at 0.6. The validation and revision thresholds, θ_{val} and θ_{rev}, peak at 0.1 and 0.02, respectively. These values indicate that a moderately conservative update policy is optimal to prevent the premature promotion of spurious rules while ensuring the timely pruning of contradicted guidelines. Finally, the quality weight λ controls the balance between task reward and strategy adherence, peaking at 0.3. Removing the adherence signal entirely, where λ is 0.0, reduces performance, while overweighting adherence at 0.8 degrades task success by forcing compliance over objective outcomes.

D Seed Rules for S_u

The five seed rules used to initialize S_u. All seeds start at the Supported evidence level and must earn Established status through training; they can be revised or pruned if contradicted by evidence.

1. *Structure: When processing chunks, organize memory into labeled sections such as Identity, Preferences, and Activities.*
2. *Conflict resolution: When new information contradicts existing memory, mark the old entry as deprecated with a reason and keep both.*
3. *Relevance filtering: When encountering one-off details like transient locations or momentary moods, omit them unless explicitly stated as long-term.*
4. *Compression: When duplicate or redundant entries appear across sections, merge them into a single consolidated entry.*
5. *Evidence-bound: When updating memory, only store information directly supported by the current conversation chunk.*

E Evidence Level Lifecycle

Each rule carries an evidence level that reflects its validation history across distillation cycles:

Method	Recall facts	Suggest ideas	Latest prefs	Prefs evolve	Update reasons	Aligned recs	New Scenarios	Overall
▼ *32K memory corpus data*								
Long Context	41.38	22.22	-	42.22	72.41	33.33	26.67	42.17
RAG	37.93	22.22	-	64.44	86.21	60.00	33.33	52.41
Mem0	47.93	19.41	-	46.61	79.58	57.04	42.75	48.53
A-Mem	45.71	18.52	-	73.33	79.31	66.67	33.33	55.42
LightMem	47.06	12.90	-	66.19	75.76	38.18	24.56	52.41
MemAgent	**72.87**	23.66	-	69.78	83.84	50.91	38.60	59.93
Mem-α	45.71	29.63	-	71.11	86.21	66.67	40.00	58.43
MemSkill	57.14	40.74	-	73.33	**89.66**	66.67	46.67	64.45
HiPS	71.43	**51.85**	-	**84.44**	**89.66**	**73.33**	**53.33**	**73.49**
▼ *128K memory corpus data*								
Long Context	9.26	22.15	15.08	28.87	36.49	22.02	16.67	20.74
RAG	53.70	20.25	33.73	52.58	59.46	39.45	36.36	38.02
Mem0	56.25	20.49	37.77	55.09	57.20	41.64	29.13	39.67
A-Mem	64.81	18.99	35.71	54.64	62.16	40.37	37.88	39.88
LightMem	30.99	14.67	35.33	61.29	64.31	35.82	28.17	36.74
MemAgent	56.14	27.80	61.66	64.22	66.17	40.69	34.74	50.86
Mem-α	59.26	30.38	50.40	60.82	64.86	45.87	39.39	48.18
MemSkill	64.81	37.34	61.51	67.01	68.92	50.46	42.42	55.37
HiPS	**68.42**	**44.02**	**71.82**	**71.55**	**72.49**	**54.73**	**44.13**	**62.01**

Table 8: Category-wise accuracy (%) on PersonaMem under 32K and 128K interaction histories. "–" indicates the category is not available in the dataset.

Method	Travel	Entertain	Lifestyle	Shop	Education	Professional	Pet	Overall
▼ *Explicit Memory*								
Long Context	33.90	42.30	37.80	24.10	24.70	23.70	8.00	32.40
RAG	48.50	55.90	52.40	40.30	40.60	38.00	25.90	47.40
Mem0	60.10	64.00	62.30	49.70	53.20	50.20	38.70	57.60
A-Mem	63.90	69.20	65.60	56.70	56.80	55.50	43.30	62.30
LightMem	66.10	70.90	67.20	59.30	57.50	56.30	47.00	64.20
MemAgent	78.60	81.10	82.40	71.60	73.20	73.10	65.20	77.40
Mem-α	78.10	80.20	79.40	69.80	71.60	70.70	60.60	75.70
MemSkill	84.10	86.50	85.00	79.10	79.30	78.50	68.80	82.60
HiPS	**90.20**	**94.20**	**91.70**	**84.10**	**86.30**	**85.20**	**78.10**	**89.20**
▼ *Implicit Memory*								
Long Context	28.00	32.50	31.10	19.30	20.30	15.60	8.00	25.90
RAG	33.50	42.40	36.20	24.20	27.60	24.10	9.40	32.40
Mem0	48.30	53.20	51.50	39.20	40.30	39.50	26.40	46.40
A-Mem	54.80	58.70	56.60	46.60	49.20	45.00	34.80	52.80
LightMem	57.50	60.40	59.60	48.60	47.50	47.50	37.30	54.80
MemAgent	67.10	69.00	67.50	58.40	60.70	59.00	50.30	64.40
Mem-α	64.60	71.10	65.30	56.60	59.60	55.40	45.60	63.10
MemSkill	68.40	73.00	71.50	63.60	**65.80**	62.70	54.30	68.40
HiPS	**71.90**	**74.20**	**71.60**	**64.50**	65.10	**63.40**	**55.30**	**69.40**

Table 9: Domain-wise accuracy (%) on PrefEval multiple-choice classification under Explicit vs. Implicit preference.

- **Tentative:** Newly hypothesized, awaiting confirmation. Not injected into the prompt.
- **Supported:** Validated by at least one distillation cycle. Eligible for prompt injection.
- **Established:** Consistently validated across multiple cycles. Highest priority for injection.

Rules are promoted when trajectory data shows they predict task success (predictive gain exceeds θ_{val}), and demoted when contradicted by evidence (predictive gain falls below θ_{rev}). Stale Tentative rules that receive no validation within T_{stale} consecutive cycles are pruned.

F USD and PDD Prompts

USD prompt. The USD prompt instructs the LLM to compare high-reward and low-reward trajectories across users and identify management pat-

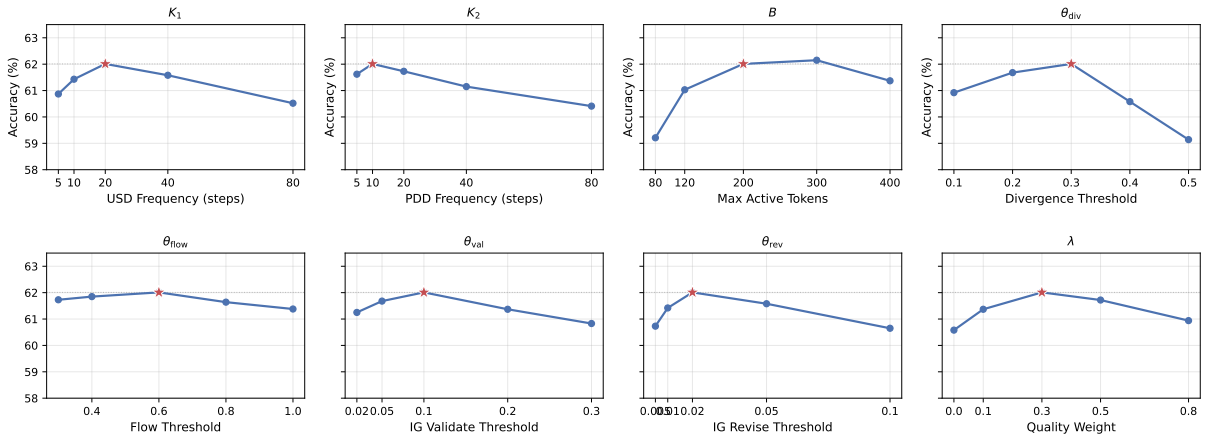

Figure 6: Parameter sensitivity analysis of HiPS on the PersonaMem 128K benchmark. The red stars indicate the selected default values. From top-left to bottom-right, the subplots demonstrate the impact on prediction accuracy of the USD frequency K_1, PDD frequency K_2, strategy token budget B, divergence threshold θ_{div}, flow threshold θ_{flow}, validation threshold θ_{val}, revision threshold θ_{rev}, and quality weight λ.

Hyperparameter	**Symbol**	**Value**
Universal Strategy Distillation (USD)		
USD distillation frequency	K_1	20
Distillation trajectories per cycle	k	5
Trajectory buffer size	—	50
Promotion threshold	θ_{val}	0.1
Demotion threshold	θ_{rev}	0.02
Stale pruning cycles	T_{stale}	3
Persona Delta Distillation (PDD)		
PDD distillation frequency	K_2	10
Gating threshold	θ_{div}	0.3
Rule Budget & Flow		
Total token budget	B	200
Minimum budget floor	—	$0.2B$
Migration threshold	θ_{flow}	0.6
Policy Optimization		
Reward weight (adherence)	λ	0.3

Table 10: Hyperparameter settings used for HiPS strategy co-evolution.

terns that generalize. To focus the LLM on management patterns rather than content, the prompt explicitly instructs: "IGNORE user-specific content (names, hobbies, topics). Focus only on how the memory was ORGANIZED and MANAGED." The prompt is enriched with behavioral feature statistics (verbosity, structural organization, change tracking) and their contrastive attribution across high- and low-reward groups. Key output constraints: (1) structured diffs only (validate/hypothesize/revise), (2) rules follow "[Label]: When [condition], [action]" format, (3) rules describe management actions, not topic-level content preferences. When a hypothesized rule overlaps an existing entry (word overlap > 0.7), it is treated as a validation rather than a new hypothesis, preventing rule explosion.

PDD prompt. The PDD prompt presents S_u as anchoring context and instructs the LLM to identify management behaviors for a specific user that differ from the universal rules. Key constraints beyond USD: (1) rules must describe management actions, not content topics, (2) rule conditions must reference observable behavioral patterns, not user identities or domain names. Concretely, a rule like "preserve full context for financial discussions rather than summarizing" is valid because the action is about *how* to handle information; "the user is interested in finance" is not, because it states *what* the user cares about without prescribing a management decision. During training, rules are persona-specific; behavior-conditioning enables generalization to unseen users at inference time.

G Compliance Estimation

Both the predictive gain (PG, Eq. 4) and the adherence reward (R_{follow}, Eq. 7) require estimating whether a trajectory complies with a given rule. We use a lightweight keyword-based heuristic: for each rule, we define a set of trigger keywords derived from the rule's condition and action clauses; a trajectory segment is marked as compliant (r^+) if it contains at least one keyword from the action set and the corresponding condition keywords appear in the preceding context, and as non-compliant (r^-) otherwise. This heuristic is intentionally simple to avoid adding LLM calls during training. While imperfect, PG is used for relative ranking (submodular

Universal Strategy Distillation Template

```
You are analyzing a memory management agent that maintains a structured user memory profile across
multiple conversation sessions.

At each episode, the agent reads conversation chunks one by one, updates a running user memory, and
then answers a question about the user based solely on this memory.

## Current universal management rules:
{current_rules}

## New evidence from recent episodes across DIFFERENT users:

### High-reward traces, where the memory led to correct answers:
{top_traces}

### Low-reward traces, where the memory led to wrong answers:
{bottom_traces}

## Instructions

Compare the traces. IGNORE user-specific content, such as names, hobbies, or topics. Focus ONLY on how
the memory was ORGANIZED and MANAGED.

Good rules describe management ACTIONS, for example:
- "Use labeled sections for different information types"
- "Mark changed preferences as deprecated instead of deleting"

Bad rules describe topics, for example:
- "Prioritize cooking preferences", which represents content
  rather than management

Output ONLY a JSON object with no other text:
- "validate": list of rule indices as integers supported by evidence
- "hypothesize": list of 1 to 2 NEW management rules, requiring fewer than 20 words per rule
- "revise": list of rule indices as integers that conflict with evidence

Example:
{
 "validate": [0, 2],
 "hypothesize": ["Use deprecated markers when preferences change"],
 "revise": [1]
}

If no changes needed:
{"validate": [], "hypothesize": [], "revise": []}
```

Figure 7: The complete prompt template used for Universal Strategy Distillation.

selection, budget allocation) and threshold-based gating (divergence, auto-promotion), both of which are robust to systematic bias in the estimate: a uniform upward shift in all PG values preserves their relative ordering and does not change which users exceed the divergence threshold. The adherence reward R_{follow} serves as a dense auxiliary signal alongside the sparse task reward R_{ans}; its weight $\lambda = 0.3$ is small, limiting the influence of any single compliance misjudgment.

H Algorithm

Algorithm 1 outlines the complete co-evolutionary training procedure for our framework. At each step, the system dynamically constructs an active strategy by merging the universal baseline S_u and the personalized delta Δ_p through budgeted submodular selection. This active strategy guides the policy rollouts to generate interaction trajectories. After computing both the task correctness and strategy adherence rewards, the system strictly separates their downstream application. It updates the trajectory distillation buffer using exclusively the task reward to prevent self-validation loops, while uti-

Persona Delta Distillation Template

```
You are analyzing a memory management agent serving ONE specific user.

The agent already follows these universal management rules:
{universal_rules}

Current personalized rules for this user, where indices refer to this list:
{current_delta}

## New evidence from this user's recent episodes:

### High-reward traces, where the memory worked well:
{top_traces}

### Low-reward traces, where the memory was less useful:
{bottom_traces}

## Instructions

Identify management behaviors specific to THIS user that DIFFER from the universal rules.
Do NOT repeat any universal rule.

Each rule must describe a MANAGEMENT ACTION, not a topic:
WRONG: "Prioritize financial interests"
RIGHT: "Preserve full preference change history instead of keeping only the latest"

Output ONLY a JSON object with no other text:
- "validate": list of rule indices as integers
  from personalized rules that are confirmed
- "hypothesize": list of 1 to 2 NEW user-specific
  management rules, requiring fewer than 20 words per rule
- "revise": list of rule indices as integers
  that conflict with evidence

Example:
{"validate": [0],
 "hypothesize": ["Keep full preference change history with reasons"],
 "revise": []}

If no changes needed:
{"validate": [], "hypothesize": [], "revise": []}
```

Figure 8: The complete prompt template used for Persona Delta Distillation.

lizing the combined reward to execute the GRPO policy update. To ensure continuous strategy refinement, the framework periodically triggers its dual distillation modules. The system executes Persona Delta Distillation every K_2 steps to generate adaptive rules for highly divergent users, and it invokes Universal Strategy Distillation every K_1 steps to abstract shared principles and elevate widely successful personalized rules into the global standard.

Algorithm 1 HiPS: Hierarchical Strategy-Policy Co-Evolution

Require: Policy π_θ, seed rules S_u, empty $\{\Delta_p\}$, buffer $\mathcal{B}$

for each training step t **do**
  Select $S_p = S_u \cup \Delta_p$ via submodular optimization (Eq. 6)
  Inject S_p into rollout prompts
  Collect trajectories $\{\tau\}$; compute R_{ans}, R_{follow}
  Update buffer $\mathcal{B}$ with (m_K, R_{ans}, t) *// task reward only*
  GRPO update on π_θ using $R_{\text{ans}} + \lambda \cdot R_{\text{follow}}$
  if $t \bmod K_2 = 0$ **then**
    Recompute PG for all rules; auto-promote / demote by thresholds
    PDD: evolve Δ_p for users with $D(p) \geq \theta_{\text{div}}$ (Eq. 5)
  end if
  if $t \bmod K_1 = 0$ **then**
    USD: evolve S_u from cross-persona trajectories
    Generalize frequent Δ_p rules $\rightarrow S_u$; consolidate duplicates
  end if
end for