BENCHMARKING MULTIMODAL SELF-EVOLVING MEMORY FOR PERSONALIZED VISION-LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized AI assistants need long-term memory that evolves as a user's interactions grow over time and across domains. Yet evaluations centered on text and static memory snapshots provide limited insight into how multimodal memory supports lifelong personalization. We introduce MEMORYVCD, a benchmark that separates three memory design dimensions: *how memory is represented*, *how it is updated*, and *how it grows*. Grounded in real users' time-stamped multimodal histories, it combines five input formats with two growth axes: a within-domain time axis with advancing prediction targets and a cross-domain axis with fixed targets. We evaluate six personalization tasks across up to twelve vision-language backbones from three providers and four main memory-update methods. Visual signals broadly improve personalization, and model scaling trends remain visible alongside differences in task and format strengths across families. No update method dominates: accumulated history is especially useful for ranking and choice, while compression and recency retain advantages for quantitative prediction and local detail. More history does not always yield the best score, and cross-domain memory brings substantial benefits but can induce persistent negative transfer in visual prediction. These findings motivate memory systems that select, retain, and transfer evidence as memory evolves. MEMORYVCD provides a shared testbed for systematically studying multimodal lifelong personal AI systems.

1 INTRODUCTION

AI assistants increasingly support people across work, health, and everyday life, where useful assistance depends on understanding the individual over months and years. The same request can have different right answers for different users, making personalization central to sustained assistance (Zhang et al., 2025; Salemi et al., 2024). Real-world interactions are inherently multimodal: language, visual signals, feedback, and temporal context jointly reveal what a person values. These signals also evolve as new experiences arrive, interests change, and activities move across domains, even while some preferences persist. Personal memory must therefore be both *multimodal* and *self-evolving*: its state must incorporate new evidence while retaining information that remains useful over time and across contexts (Huang et al., 2026a). Understanding how memory supports this evolving dynamics is essential to building reliable lifelong personalized agents.

Existing benchmarks fall short on three fronts, and no prior benchmark closes all three at once. **(i) They are text-only.** LoCoMo (Maharana et al., 2024), MemBench (Tan et al., 2025), LongMemEval (Wu et al., 2025), LaMP (Salemi et al., 2024), and MSC (Xu et al., 2022) treat history as a token sequence and discard the images that dominate real user footprints. **(ii) They are simulated or short.** Most use LLM-generated dialogues with no real user feedback; the few real-user sets capture only a handful of interactions per person, too few to study memory over a long horizon. **(iii) They read memory at a single frozen snapshot**, so any claim about *how memory evolves* as history grows is out of reach, and, as we show, single-snapshot evaluation actively misranks methods. The closest prior effort, Zhang et al. (2026b) add real cross-domain users but is still text-only, single-snapshot, and does not separate storage format from update policy. Table 1 contrasts MEMORYVCD with these prior benchmarks across eight realism axes a lifelong personal agent needs (an expanded discussion is in Appendix H).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

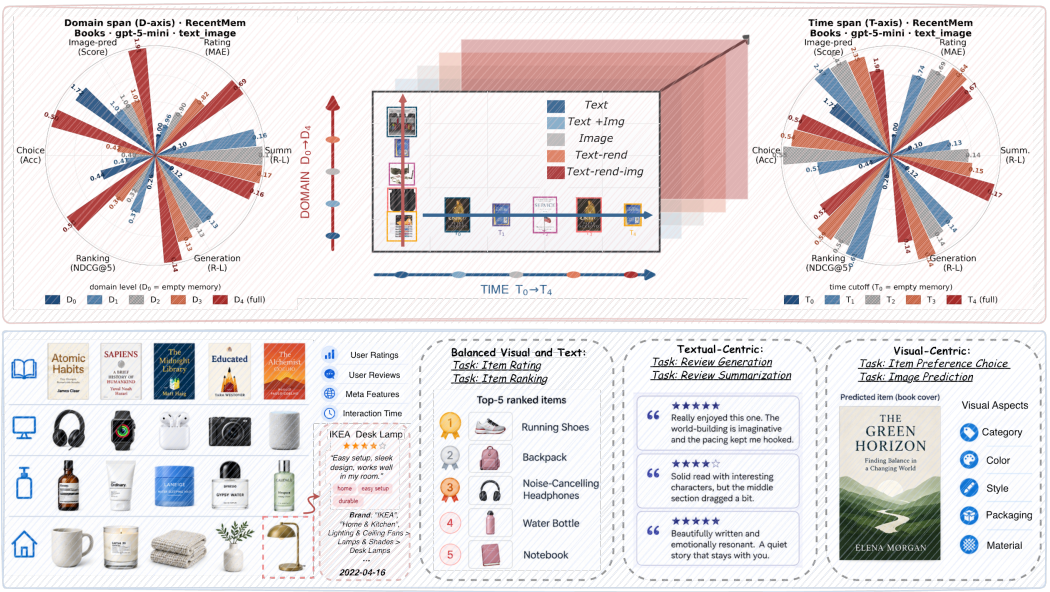


Figure 1: **Overview of MEMORYVCD.** A real user's multi-year, multimodal history (images, reviews, ratings, timestamps across four domains) is stored in *five input formats*; from memory alone the agent answers *six* tasks in three groups by which modality carries the signal (Balanced, Textual, Visual). Each task is scored as memory *evolves* along two axes, within-domain time (T_0 to T_4) and cross-domain (D_0 to D_4), from empty to full, so we measure memory as it grows rather than at one frozen snapshot.

We introduce MEMORYVCD (Figure 1), a benchmark and harness that studies memory as three separable design choices. Built on Amazon Reviews 2023 (Hou et al., 2024), it provides real, time-stamped, image-paired histories of hundreds to thousands of items for lifelong users in each of Books, Personal Care, Electronics, and Home & Kitchen, and evaluates six user-facing tasks organized into three categories by modality emphasis, balanced visual & text, textual-centric, and visual-centric (Table 2). Two design choices make it a memory testbed rather than another QA set. First, *five input formats* spanning text, rendered-text pixels, and visual inputs let us ask how visual signals support personalization. Second, a *self-evolving* protocol grows the memory pool along two parallel growth axes, a within-domain time axis and a cross-domain axis, each with five cutoffs from empty to full; to stay close to reality the time axis is scored on a dynamic test that advances with the growing history, while Appendix M additionally reports a fixed-test variant in which only the memory grows. On this grid we evaluate four widely-used update policies (a recency window, BM25 retrieval as in LoCoMo, LLM-summarized gists as in ReadAgent (Lee et al., 2024), and an evolving persona as in PersonaAgent) across up to twelve backbones from three providers.

MEMORYVCD serves as shared research infrastructure and a testbed for multimodal, self-evolving personal memory. By connecting real user histories with a common evaluation protocol, it enables systematic study of how memory supports personalization as evidence changes over time and across domains. The resulting comparisons establish the foundation infrastructure and testbed for identifying current limitations and measuring progress toward lifelong personalized agents.

Our contributions.

- **A real, multimodal, lifelong benchmark** that grounds every memory item in a real user's time-stamped, image-paired history and separates three memory design axes prior work conflates: *what modality* to store (five formats), *what update rule* to apply, and *along which axis* memory grows. It comprises **314,116** real interactions (median **773** per user) spanning roughly a decade, each carrying visual evidence, a review (title and body), an integer rating, and a timestamp.
- **A self-evolving evaluation protocol** with two parallel longitudinal axes, a within-domain time axis (T_0 to T_4) and a cross-domain axis (D_0 to D_4), each grown from empty to full; the time axis

Table 1: **Head-to-head on eight realism axes for lifelong personal memory** (✓ yes, ● partial, ✗ no). **Real**=real users; **Impl.**=implicit (behavior-learned) preference; **X-Dom/X-Mod**=cross-domain/-modality; **Effic.**=efficiency measured; **Self-Ev.**=memory read as it evolves. MEMORYVCD is the only one satisfying all eight.

Benchmark	Ctx	Span	Real	Impl.	X-Dom	X-Mod	Effic.	Self-Ev.
Text-domain memory benchmarks								
LoCoMo (Maharana et al., 2024)	9K	days to weeks	✗	✗	✗	●	✗	✗
LongMemEval (Wu et al., 2025)	100K+	weeks	✗	✗	✗	✗	✗	✗
MemoryAgentBench (Hu et al., 2026)	1M	varies	●	●	✗	✗	●	●
MemoryCD (Zhang et al., 2026b)	400K	months to years	✓	✓	✓	✗	✗	✗
Vision-domain memory benchmarks								
Mem-Gallery (Bei et al., 2026)	∼80K	days to months	✗	●	●	●	●	●
MemLens (Ren et al., 2026)	32K–256K	weeks to months	✗	●	●	✓	✗	✗
MemEye (Guo et al., 2026)	∼60K	hours to weeks	✗	●	●	✓	✗	●
MEMORYVCD (ours)	**670K+**	**months to years**	✓	✓	✓	✓	✓	✓

is scored on a dynamic test that advances with the growing history to stay close to reality, and Appendix M additionally reports a fixed-test variant in which only the memory grows.

- **A large, reproducible study** that evaluates **4** widely-used update policies over **5** storage modalities (plain text, rendered-text pixels, visual inputs, and their two hybrids), **2** orthogonal growth axes with **5** evolution cutoffs each (time T_0 to T_4 and domain D_0 to D_4), up to **12** vision-language backbones from three providers, **12** domains, and **6** user-facing tasks; Appendix D adds a controlled comparison of **8** further memory-update methods on both targets and both axes.

2 THE MEMORYVCD BENCHMARK

Task Formulation. MEMORYVCD treats each user u as a long-horizon behavioral process $\mathcal{M}_u = \{(i, m_i, v_i, r_{u,i}, t_{u,i}, x_{u,i})\}_{i \in \mathcal{I}_u}$, where i is an item, m_i its text metadata, v_i its product image, $r_{u,i} \in \{1, \ldots, 5\}$ the user's rating, $t_{u,i}$ the timestamp, and $x_{u,i}$ the review text. Interactions are sorted by t; the last three are held out as the test set and earlier interactions are memory candidates. We evaluate six user-facing tasks, grouped into three categories by which modality carries the preference signal, **Balanced Visual & Text** (Rating, Ranking), **Textual-Centric** (Review Generation, Summarization), and **Visual-Centric** (Preference Choice, Image Prediction) (Table 2).

Five input formats. The same memory item is rendered as (1) `text`, (2) `text_image` (text + visual input), (3) `image_only`, (4) `text_rendered` (the review words painted into an image), and (5) `text_rendered_image`. This spectrum lets us test how visual encoding contributes to personalization across tasks and model families.

Memory methods. We implement **twelve** memory-update methods spanning the current design space, organized into **four update-mechanism categories**. **Recency** keeps recent items and drops old ones (*RecentMem* (Tan et al., 2025), *Zep* (Rasmussen et al., 2025)); **Accumulation** keeps the raw history and grows the pool (*LoCoMo* (Maharana et al., 2024), *Stack-All*); **LLM compression** summarizes or distills the history (*ReadAgent* (Lee et al., 2024), *LightMem* (Fang et al., 2026), *SimpleMem* (Liu et al., 2026), *MemOS* (Li et al., 2025)); and **LLM update** evolves an explicit state from feedback (*PersonaAgent* (Zhang et al., 2025), *Mem0* (Chhikara et al., 2025), *A-Mem* (Xu et al., 2025), *MemSkill* (Zhang et al., 2026a)). For clarity in the main text, we foreground one representative per category, *RecentMem*, *LoCoMo*, *ReadAgent*, and *PersonaAgent*; all twelve methods are evaluated under the identical protocol. Appendix D gives the full twelve-method comparison, including build/inference efficiency, and per-method descriptions.

Self-evolving protocol. We sort each user's interactions by timestamp and grow memory along two orthogonal axes, re-evaluating the same six tasks at five cutoffs on each. Memory is *cumulative*: the state at cutoff k contains everything available up to k and strictly extends the state at $k-1$, i.e. $M_{u,k} = U(M_{u,k-1}, \Delta H_{u,k})$ for an update rule U (formal definition in Appendix A).

Table 2: The six MEMORYVCD tasks, grouped by which modality carries the preference signal (Balanced, Textual, Visual). Each probes a different facet of preference and needs a different memory ability. Image Prediction is judge-scored 1 to 5 (1 = category-unrelated floor, 5 = perfect visual match); its rubric, scale, prompt, and full per-task definitions are in Appendix A.

Task	What it probes (preference angle)	Input	Metrics
Balanced Visual & Text, signal is carried by both modalities			
Rating Prediction	Absolute intensity: recover the user's rating scale	$\mathcal{M}_u$, target title	MAE, RMSE
Item Ranking	Relative order over a 20-candidate pool	$\mathcal{M}_u$, candidates	NDCG@{K}, Recall@{K}
Textual-Centric, signal is carried mainly by review and item meta text			
Review Generation	The user's own voice: tone and length	$\mathcal{M}_u$, target title	ROUGE-1/L, BLEU-1/4
Review Summarization	Which aspects matter: recurring opinions	$\mathcal{M}_u$, title, full review	ROUGE-1/L, BLEU-1/4
Visual-Centric, signal is carried mainly by product images			
Item Pref. Choice	Visual taste among near-identical titles	$\mathcal{M}_u$, target + 4 hard negs	Accuracy
Image Prediction	Future visual taste: color, form, material	$\mathcal{M}_u$ only	LLM judge

Time axis (T_0 to T_4): a within-domain, genuinely *dynamic* setting. Cutoffs are taken *per user* by *interaction count* (not calendar time): at cutoff k the model predicts a three-interaction window at that point in the user's own timeline from all earlier interactions, so the memory *and* the prediction target advance together, T_0 predicts the user's first three interactions from empty memory, T_4 their last three from full history. Unlike single-snapshot benchmarks that score one frozen query set, this tracks online personalization as it actually unfolds.

Domain axis (D_0 to D_4): a cross-domain setting with a *fixed* test. The three test interactions are always the target domain's last three; D_0 is empty memory and D_1 to D_4 add source domains one at a time in a fixed order, ending with the target domain itself at D_4 (the full-lifelong upper bound). Source-domain memory is time-filtered to precede the earliest target test timestamp, so cross-domain history never leaks future information. Because the test set is held fixed here, differences across D-cutoffs isolate the effect of memory *composition*.

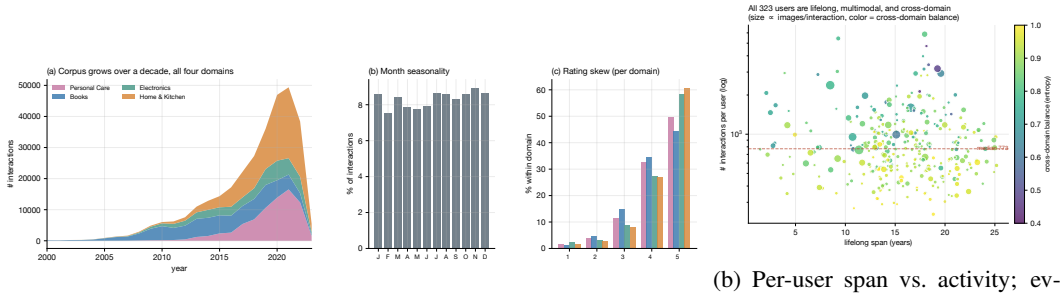
All update methods receive the *same* accessible history at each cutoff and differ only in how they select, compress, or organize it, so the comparison is fair by construction; only their implementations differ (Appendix D).

2.1 CORPUS AND STATISTICS

MEMORYVCD is built on Amazon Reviews 2023 (Hou et al., 2024) and spans **twelve** Amazon domains in total: **four primary domains** used as evaluation targets (Personal Care, Books, Electronics, Home & Kitchen) plus eight additional domains that populate the cross-domain growth axis (Appendix A). We select **323 lifelong users** by taking those simultaneously active in all four primary domains (their overlap) and keeping the ones with the most interactions there, yielding **314,116 real interactions** (median **773** per user) each carrying visual evidence, review title and body, an integer rating, and a timestamp. Full dataset construction, the user-selection criterion, per-task definitions, and the image-prediction judge rubric are in Appendix A. Two properties make this corpus a genuine *self-evolving* testbed rather than a static snapshot. Figure 2 shows it spans a decade and is multimodal end-to-end; Figure 3 then places a single *user* at the center, in both modalities and along both benchmark axes.

Change and stability coexist in individual user histories. Figure 3 puts the user at the center of the analysis. Panels (a,b) follow each user's mean review-text and product-image embeddings over time, with calendar year as height. Some users show large shifts, while others remain relatively stable. The histories therefore change differently across users. These plots motivate testing how memory methods use new evidence, but do not show the memory states constructed by those methods.

Panels (c,d) follow users across twelve domains. Their embeddings shift with domain content, yet the roughly parallel trajectories suggest that some relative relationships between users persist across domains. This pattern motivates the possibility of domain-invariant user structure alongside domain-specific variation. The text and image views show these relationships from different perspectives; the plots motivate, rather than establish, transferability, which Section 3.3 tests behaviorally.

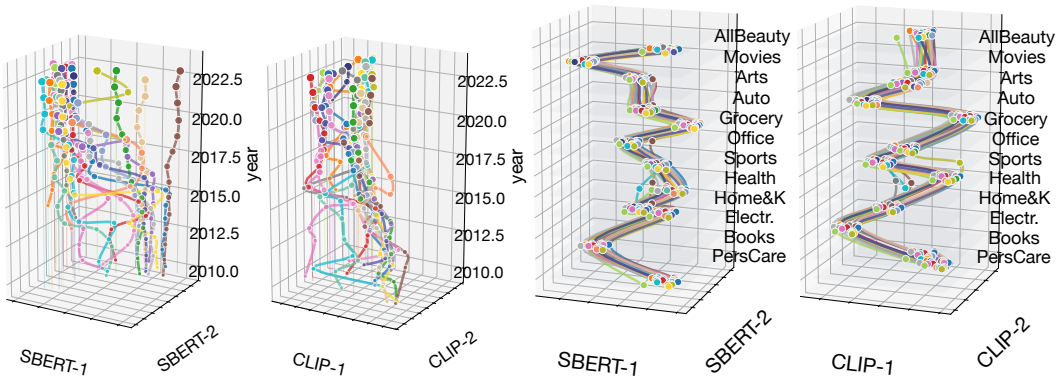


(a) Interaction volume by year (stacked over domains), month-of-year seasonality, and the per-domain rating distribution.

(b) Per-user span vs. activity; every user is lifelong, image-rich, and cross-domain balanced.

Figure 2: **The corpus spans roughly a decade and is multimodal end-to-end.** Statistics here are for the four primary evaluation domains (Books, Personal Care, Electronics, Home & Kitchen); the full dataset spans twelve domains in total, with the complete breakdown in Appendix A. (a) Interaction volume rises across all four primary domains through the 2010s, activity is roughly aseasonal, and ratings concentrate at 4 to 5⋆, so rating prediction must recover each user's personal scale. (b) Every one of the 323 users spans years, carries review images, and is spread across domains; color is a user's *cross-domain balance* (normalized entropy of their per-domain split; 1.0 = perfectly even), and nearly all sit near the top, i.e. genuinely multi-domain.

These plots describe variation across time and domains. They motivate measuring several memory states, but do not establish how much of that variation is predictable from personal history. The prediction experiments test that question directly. Appendix F provides further descriptive analyses.



(a) Time · text **(b)** Time · image **(c)** Domain · text **(d)** Domain · image

Figure 3: **One user at the center, across time, domain, and modality** (20 most active users). (a,b) Per-user 3D *time landscapes* (height = year) in review-text (SBERT) and product-image (CLIP) space: taste drifts along coherent per-user paths around a stable baseline, and differently in the two modalities. (c,d) All 12 domains stacked as z-layers, one ribbon per user through their per-domain mean embedding: despite large domain-to-domain shifts, the near-vertical, roughly parallel ribbons *suggest* a per-user signature that persists across domains (an exploratory cue tested behaviorally in Section 3.3). The four panels vary exactly the benchmark's three design axes.

3 Experiments and Findings

We follow the evaluation protocol and task settings defined in Section 2, with task details in Appendix A. The format sweep uses twelve backbones, five formats, RecentMem, and the T_4 setting. The trajectory sweeps use `gpt-5-mini` and `gpt-5-nano`, four methods, and two formats on Books and Home & Kitchen. Decoding settings are held fixed within each comparison.

Table 3: **Storage axis: input-format sweep across 12 backbones from three providers** (OpenAI, Anthropic, Google; H&K target, T_4/RecentMem). Rows are the five memory formats grouped by task, columns are the backbones. Per (task, model) the winning format is **bold** and shaded, with intensity tracking within-column rank (each task block reads as a heatmap). Rating is MAE (lower better), all others higher-better. Books and per-domain breakdowns in Appendix I.

		OpenAI						Anthropic			Google		
Task	**Fmt**	5.5	5	5-mini	5-nano	4o	4o-mini	opus-4.7	sonnet-4.6	haiku-4.5	3.1-pro	3-flash	2.5-flash
Rating	*MAE (lower)*												
	T	0.640	0.667	0.687	0.763	0.753	0.877	0.727	0.830	0.893	0.653	0.613	0.653
	Tr	0.667	0.660	0.703	0.753	0.747	0.833	0.783	0.790	0.877	0.633	0.600	0.637
	I	0.677	0.677	0.663	0.847	0.807	0.970	**0.693**	0.853	0.940	0.617	0.617	0.713
	T+I	**0.627**	**0.610**	**0.647**	**0.737**	0.767	**0.827**	0.710	0.837	**0.853**	0.620	0.613	**0.607**
	Tr+I	0.627	0.627	0.653	0.780	**0.717**	0.850	0.760	**0.780**	0.856	**0.607**	**0.567**	0.627
Summ	*R-L (higher)*												
	T	0.220	0.185	0.172	0.157	0.142	0.155	0.238	0.209	**0.194**	0.229	0.214	0.208
	Tr	**0.221**	0.195	**0.184**	0.156	**0.148**	**0.159**	0.230	**0.210**	0.161	0.223	0.218	0.203
	I	0.193	0.145	0.158	0.144	0.121	0.135	0.198	0.166	0.147	0.197	0.168	0.160
	T+I	0.221	0.187	0.133	0.158	0.137	0.143	**0.245**	0.205	0.174	**0.234**	0.213	**0.210**
	Tr+I	0.221	**0.200**	0.181	**0.159**	0.136	0.142	0.227	0.199	0.168	0.226	**0.219**	0.179
Gen	*R-L (higher)*												
	T	0.159	**0.136**	0.140	0.122	0.131	0.127	0.157	0.149	0.143	**0.177**	0.167	0.159
	Tr	0.166	0.128	0.145	0.111	**0.143**	**0.134**	**0.165**	0.161	**0.146**	0.173	0.170	0.156
	I	0.142	0.109	0.119	0.104	0.120	0.119	0.145	0.121	0.113	0.133	0.138	0.128
	T+I	0.162	0.134	0.144	**0.128**	0.141	0.133	0.158	0.148	0.142	0.177	0.169	**0.160**
	Tr+I	**0.171**	0.133	**0.148**	0.111	0.139	0.134	0.165	**0.163**	0.142	0.173	**0.172**	0.158
Rank	*NDCG (higher)*												
	T	0.435	0.365	0.366	0.288	0.268	0.153	0.407	0.353	0.229	0.387	0.378	0.393
	Tr	0.330	0.259	0.316	0.191	0.165	0.131	0.387	0.313	0.202	0.398	0.389	0.341
	I	0.408	0.370	0.352	0.248	**0.284**	0.155	0.418	0.327	0.293	0.390	0.383	0.348
	T+I	**0.451**	**0.399**	**0.396**	**0.290**	0.272	**0.177**	**0.432**	0.373	**0.327**	0.401	**0.394**	**0.417**
	Tr+I	0.382	0.285	0.350	0.213	0.152	0.151	0.383	**0.386**	0.193	**0.402**	0.382	0.352
Choice	*Acc (higher)*												
	T	0.353	0.333	0.353	0.283	0.317	0.300	0.287	0.287	0.297	**0.337**	0.273	**0.283**
	Tr	0.320	0.337	0.297	0.230	0.267	0.273	0.287	0.303	0.280	0.320	0.283	0.250
	I	0.277	0.330	0.253	0.303	**0.320**	0.307	0.267	0.310	**0.323**	0.277	0.273	0.283
	T+I	**0.367**	**0.360**	**0.360**	**0.310**	0.310	**0.317**	**0.340**	0.307	0.273	0.330	0.260	0.260
	Tr+I	0.343	0.300	0.323	0.270	0.287	0.293	0.323	**0.313**	0.250	0.337	**0.303**	0.247
ImgPred	*1-5 (higher)*												
	T	1.750	1.210	1.177	1.193	2.020	1.633	1.760	1.900	1.920	1.647	1.623	1.970
	Tr	1.743	1.193	1.183	1.197	2.163	1.697	**1.783**	**1.958**	**1.987**	1.653	1.605	1.963
	I	1.757	**1.230**	1.183	1.210	2.040	**1.757**	1.783	1.733	1.797	**1.827**	**1.767**	1.960
	T+I	1.757	1.173	1.187	1.214	2.130	1.627	1.760	1.650	1.787	1.727	1.713	**2.030**
	Tr+I	**1.770**	1.223	**1.210**	**1.217**	**2.170**	1.630	1.770	1.701	1.693	1.737	1.623	2.010

3.1 Experiment 1: How should inputs be represented?

Fixing the cutoff (T_4) and update rule (RecentMem), we compare input formats for the memory and the task-provided target or candidate items. Table 3 reports the full input-format sweep across **12 vision-language backbones from three providers** (OpenAI, Anthropic, Google) on the Home & Kitchen target, all six tasks and five input formats. Every (model, task) block is shaded as a within-task heatmap and the bold entry is the winning format.

Finding 1: visual signals broadly improve personalization. Visual information provides useful personalization cues beyond plain text. Across the storage sweep (Table 3), visually encoded formats achieve the best reported score in most model–task settings, with the text-plus-visual hybrid the most frequent winner. The benefits extend beyond visual prediction to rating, ranking, and choice: for example, adding visual signals improves GPT-5 ranking from 0.365 to 0.399 and choice from 0.333 to 0.360. Visual appearance can complement textual descriptions, while rendered text offers another route for models to access the same semantic content. Together, these results support visual encoding as a useful channel for personalized prediction.

Finding 2: model scaling benefits personalization, with family-specific task and format strengths. The benefits associated with model scaling remain visible in personalization. Within a family, higher-capability models generally perform better on several tasks: with text-plus-visual inputs, GPT-5

Table 4: **Memory trajectories along both axes** (`gpt-5-mini`, `text_image`). *Left:* time axis (T_0 to T_4, H&K target). *Right:* domain axis (D_0 to D_4, Books target). T_0/D_0 = empty-memory baseline; **Avg** = mean over the four non-empty cutoffs; Δ = its relative change vs. that baseline (green = memory helps the metric, red = hurts). Green rows mark the per-task winner.

		Time axis H&K target							**Domain axis** Books target						
Task	**Method**	T_0	T_1	T_2	T_3	T_4	**Avg**	Δ	D_0	D_1	D_2	D_3	D_4	**Avg**	Δ
Rating MAE, lower is better															
	RecentMem	0.867	0.653	0.560	0.610	0.647	0.617	−29%	1.000	0.963	0.903	0.823	0.687	0.844	−16%
	LoCoMo	0.867	0.667	0.543	0.593	0.660	0.616	−29%	1.000	0.853	0.757	0.740	0.710	0.765	−24%
	ReadAgent	0.867	0.627	**0.453**	0.527	0.580	0.547	−37%	1.000	0.833	0.710	0.700	**0.643**	0.722	−28%
	PersonaAgent	0.867	0.607	0.560	0.623	0.693	0.621	−28%	1.000	0.850	0.693	0.680	0.663	0.722	−28%
Summarization ROUGE-L, higher is better															
	RecentMem	0.144	0.183	0.183	**0.191**	0.133	0.172	+20%	0.101	0.162	0.166	**0.166**	0.163	0.164	+63%
	LoCoMo	0.144	0.179	0.178	0.177	0.178	0.178	+24%	0.101	0.151	0.153	0.152	0.157	0.153	+52%
	ReadAgent	0.144	0.179	0.169	0.186	0.160	0.174	+20%	0.101	0.152	0.139	0.133	0.151	0.144	+42%
	PersonaAgent	0.144	0.163	0.164	0.167	0.167	0.165	+15%	0.101	0.144	0.139	0.134	0.140	0.139	+38%
Generation ROUGE-L, higher is better															
	RecentMem	0.127	0.146	**0.151**	0.140	0.144	0.145	+14%	0.117	0.131	0.131	0.133	**0.137**	0.133	+14%
	LoCoMo	0.127	0.143	0.146	0.140	0.139	0.142	+12%	0.117	0.126	0.125	0.124	0.129	0.126	+8%
	ReadAgent	0.127	0.136	0.135	0.128	0.124	0.131	+3%	0.117	0.121	0.115	0.119	0.122	0.119	+2%
	PersonaAgent	0.127	0.131	0.129	0.122	0.119	0.125	−1%	0.117	0.117	0.117	0.116	0.118	0.117	+0%
Ranking NDCG@5, higher is better															
	RecentMem	0.281	0.438	0.428	0.394	0.396	0.414	+47%	0.263	0.369	0.318	0.362	0.504	0.388	+48%
	LoCoMo	0.281	0.626	0.719	0.798	**0.811**	0.738	+163%	0.263	0.359	0.445	0.489	**0.738**	0.508	+93%
	ReadAgent	0.281	0.464	0.469	0.462	0.434	0.457	+63%	0.263	0.367	0.365	0.415	0.572	0.430	+63%
	PersonaAgent	0.281	0.506	0.583	0.587	0.585	0.565	+101%	0.263	0.403	0.528	0.566	0.683	0.545	+107%
Choice Accuracy, higher is better; winner differs by axis															
	RecentMem	0.371	0.390	0.397	0.320	0.360	0.367	−1%	0.440	0.407	0.397	0.417	0.503	0.431	−2%
	LoCoMo	0.371	0.400	**0.440**	0.387	0.360	0.397	+7%	0.440	0.417	0.457	0.473	0.547	0.474	+8%
	ReadAgent	0.371	0.380	0.373	0.357	0.343	0.363	−2%	0.440	0.437	0.433	0.480	0.503	0.463	+5%
	PersonaAgent	0.371	0.420	0.437	0.410	0.360	0.407	+10%	0.440	0.397	0.450	0.473	**0.620**	0.485	+10%
Image Prediction GPT-5 judge 1 to 5; note the Books D-axis U-shape															
	RecentMem	1.157	1.185	1.129	1.177	1.187	1.169	+1%	**1.717**	1.007	1.004	1.019	**1.907**	1.234	−28%
	LoCoMo	1.157	1.193	1.137	1.163	1.103	1.149	−1%	1.717	1.007	1.007	1.007	1.011	1.008	−41%
	ReadAgent	1.157	1.170	1.157	**1.227**	1.213	1.192	+3%	1.717	1.007	1.011	1.015	1.791	1.206	−30%
	PersonaAgent	1.157	1.183	1.147	1.153	1.150	1.158	+0%	1.717	1.004	1.004	1.011	1.030	1.012	−41%

improves over GPT-5-nano on rating (0.610 vs. 0.737 MAE), ranking (0.399 vs. 0.290), and choice (0.360 vs. 0.310); Claude Opus similarly exceeds Haiku on summarization and ranking (Table 3). This is an empirical scaling trend across model tiers, rather than a fitted parameter-count law, and it need not hold for every task.

Scaling also interacts with model-family specialization. Claude Opus is strong on summarization, Gemini Pro on review generation, and GPT models on ranking and choice. The formats each family exploits best differ as well: GPT ranking favors text-plus-visual inputs, Claude visual prediction benefits from rendered text, and Gemini Pro attains its best visual-prediction score with visual-only inputs. Thus, effective personalization depends jointly on model capability, the task, and the storage format.

3.2 Experiment 2: Evaluating Memory Evolution Over Time

Fixing the format to `text_image` and varying the update policy, we track a representative setting (Home & Kitchen, `gpt-5-mini`) across T_0 to T_4; the left block of Table 4 reports the full grid.

Finding 3: no memory-update method dominates; accumulation and detail preservation serve different tasks. Accumulating evidence across history is broadly useful for tasks that require a global view of a user's preferences. On both growth axes, LoCoMo outperforms RecentMem in mean ranking and choice, while PersonaAgent outperforms ReadAgent on the same tasks (Table 4). On the time axis, their mean ranking scores are 0.738 vs. 0.414 and 0.565 vs. 0.457, respectively.

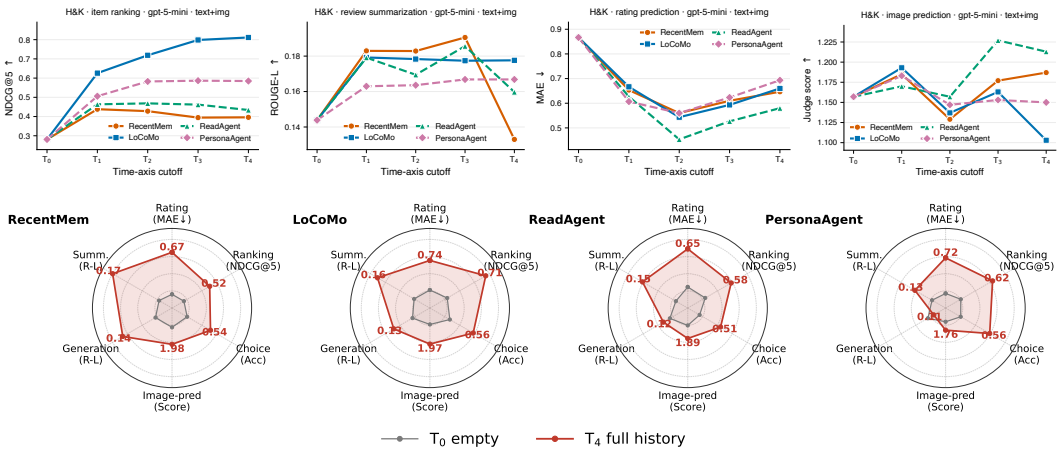


Figure 4: **Time-axis memory evolution** (`gpt-5-mini`, `text_image`). *Top:* four tasks over the time cutoffs T_0 to T_4 (H&K target). *Bottom:* per-policy empty (T_0) vs. full (T_4) memory across all six tasks (Books target), shared scale. The `text_rendered_image` format and the full six-task sweep are in Appendix J.

LoCoMo's ranking advantage over RecentMem also widens from about 43% at T_1 to 105% at T_4. Retrieval from accumulated history and an evolving user profile offer two ways to retain evidence for comparison and preference selection.

Compression and recency remain valuable when a task depends on quantitative tendencies or local detail. ReadAgent achieves the lowest mean time-axis rating error (0.547 MAE) and nearly ties PersonaAgent on the domain axis. RecentMem leads domain-axis summarization and review generation on both axes, consistent with the usefulness of recent wording and examples. This division is not absolute: LoCoMo has the highest mean time-axis summarization score. The overall lesson is that accumulation supports broad preference modeling, but the best update rule depends on which information the task needs to preserve.

Finding 4: counterintuitively, full history is not always best. Several time-axis trajectories peak at T_2 or T_3, before the full-history T_4 stage (Figure 4). ReadAgent's rating error is lowest at T_2 (0.453 vs. 0.580 at T_4), while RecentMem summarization peaks at T_3 (0.191 vs. 0.133 at T_4). More available history therefore does not guarantee a higher score in evolving personalization.

Two factors help interpret this pattern. First, our dynamic protocol advances the prediction targets along with memory, so later stages can pose different prediction challenges. Second, effectively selecting and retaining relevant evidence from long histories remains difficult for current update methods; their intermediate peaks are consistent with limitations in handling long context and long-horizon preference changes. The trajectories do not separately identify these causes. The fixed-target control (Appendix M) nevertheless shows that growing retrievable history can help: LoCoMo ranking increases from 0.371 to 0.802, compared with 0.352 to 0.404 for RecentMem. The challenge is to turn additional history into useful memory as the user's needs evolve. Appendix J reports the full sweep.

3.3 Experiment 3: Evaluating Memory Evolution Across Domains

We run the D-axis on two targets (Books, Home & Kitchen), expanding the source-domain pool one domain at a time. The right block of Table 4 reports the representative Books trajectory alongside the time axis.

Finding 5: cross-domain memory generally helps, but negative transfer remains difficult to handle. Cross-domain history provides useful evidence before any target-domain history is available. In the representative Books setting, source-only memory improves rating, summarization, and ranking relative to empty memory for all four methods, and adding target-domain history at

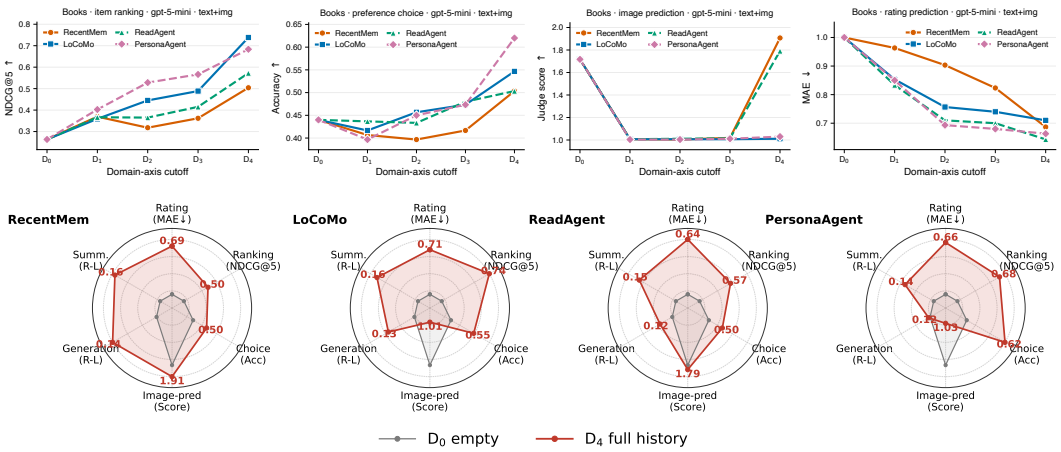


Figure 5: **Domain-axis memory evolution** (`gpt-5-mini`, `text_image`, Books target). *Top:* four tasks over the domain cutoffs D_0 to D_4; note the image-prediction U-shape. *Bottom:* per-policy empty (D_0) vs. full (D_4) memory across all six tasks, shared scale. The `text_rendered_image` format and the full six-task sweep are in Appendix K.

D_4 further raises ranking scores (Table 4). The extended source-only probe in Appendix L also shows substantial ranking benefits. These results support cross-domain memory as a useful source of transferable preference information.

The same history can, however, mislead visual prediction. Figure 5 shows negative transfer for all four methods at D_1–D_3: image-prediction scores fall near 1.0, below the plotted empty-memory reference. At D_4, only the recency-based RecentMem and compression-based ReadAgent reverse the trend, recovering to 1.907 and 1.791; LoCoMo and PersonaAgent remain near the floor. Relevant target-domain evidence can therefore restore useful prediction, but methods differ sharply in their ability to use it after exposure to distracting history. Because visual prediction is open-ended, source-domain signals can steer both the predicted category and its visual attributes. Preventing negative transfer, and recovering when relevant evidence arrives, remain unresolved challenges for current memory updates. The full sweep (Appendix K) shows that storage format can change method rankings and the extent of visual-prediction recovery. These interactions reinforce the need to choose memory representations and update rules jointly for the target task.

4 Discussion and Conclusion

MemoryVCD studies personalized memory through its representation, update rule, and growth over time and across domains. Visual signals broadly improve personalization, and scaling benefits remain visible alongside family-specific task and format strengths. No update method dominates: accumulation supports ranking and choice, while compression and recency retain advantages for quantitative prediction and local detail. Counterintuitively, full history is not always best under advancing prediction targets, highlighting the difficulty of using long histories throughout a user's evolving timeline. Cross-domain memory generally brings benefits, but visual prediction exposes negative transfer that all four methods encounter and only recency and compression reverse in the representative full-history setting.

These findings motivate task-aware memory systems that retain useful evidence, adapt to changing preferences, and filter distracting signals across domains. Our comparisons report point estimates on the selected user cohort and tested model settings. The fixed-target and user-history controls help interpret the trajectories, while changes in prediction targets and the open-ended visual task leave the causes of some effects unresolved. Future work should improve both long-horizon evidence selection and recovery from negative transfer.

REPRODUCIBILITY STATEMENT

Section 2 specifies the complete MEMORYVCD benchmark protocol: the five memory input formats, the four widely-used update policies, the two self-evolving growth axes (a within-domain time axis T_0 to T_4 and a cross-domain axis D_0 to D_4), the six user-facing tasks with their metrics, and the vision-language backbones and provider gateways used throughout. Appendix A details the dataset construction, the lifelong-user selection criterion, the per-task definitions, and the Image-Prediction judging rubric and prompt. The three main experiments are reproduced in full in the appendices, with the space-heavy per-configuration sweeps clustered at the end. Appendix B gives the domain-axis cross-domain trajectory tables on the Home & Kitchen target (Experiment 3), and Appendix C reports the user-specificity control. The full sweeps then follow: Appendix I gives the complete storage-axis sweep (Experiment 1) over all twelve backbones and five formats on both targets; Appendix J the full time-axis self-evolving sweep (Experiment 2); Appendix K the full domain-axis sweep and Appendix L the extended D_5 (12-domain) and D_{tx} (pure-transfer) probes (Experiment 3); and Appendix M the time-axis attribution control (fixed-target growing-memory) that separates memory growth from the moving test target. Appendix D describes the twelve memory-update baselines grouped by mechanism together with an efficiency (API-call/token) table, and Appendix E reports the contamination and data-exposure analysis. We release all traces, memory representations, the harness, and every per-example prediction.

ETHICS STATEMENT

This study uses the publicly available Amazon Reviews 2023 dataset (Hou et al., 2024), drawing on an existing release of reviews, ratings, visual content, and interaction metadata for research on personalized memory. It does not involve new participant recruitment or interventions. Public availability does not eliminate privacy concerns associated with longitudinal user histories. The benchmark is intended for research evaluation, not for identifying individuals, inferring sensitive attributes, or making consequential decisions about users. Use and redistribution of the underlying data should respect the source dataset's applicable terms and the rights of the original content owners. We acknowledge that the data may contain societal and platform-specific biases and that findings from the selected cohort should not be assumed to generalize to all users or deployment settings.

AI USE STATEMENT

Generative AI tools primarily assisted with literature search, manuscript revision, and the analysis and organization of experimental results.

REFERENCES

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026.

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

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830, 2025.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, et al. Rethinking memory mechanisms of foundation agents in the second half. *arXiv preprint arXiv:2602.06052*, 2026a.

Zhengjun Huang, Wenxuan Liu, Zhoujin Tian, Wei Chen, Junle Chen, Yuqian Wu, Fangyuan Zhang, Qintian Guo, and Xiaofang Zhou. M^3Exam: Benchmarking multimodal memory for realistic user-agent interactions. *arXiv preprint arXiv:2606.07402*, 2026b.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *International Conference on Machine Learning*, pp. 26396–26415. PMLR, 2024.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, et al. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*, 2025.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, et al. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 5180–5197, 2022.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

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

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*, 2025.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S Yu. Memorycd: Benchmarking long-context user memory of llm agents for lifelong cross-domain personalization. *arXiv preprint arXiv:2603.25973*, 2026b.

APPENDIX

The appendices provide dataset and task definitions, additional domain trajectories, user-history controls, extended method comparisons, a verbatim-recall probe, corpus summaries, and sample-size analyses. The full format and trajectory sweeps follow. The final sections report the extended domain probes and the fixed-target time experiment. Experimental settings and sampling details are described in the text.

A DATASET AND TASK DETAILS

This appendix gives the full dataset construction, the user-selection criterion, per-task definitions, and the image-prediction judge (rubric, scoring standard, and prompt) referenced from Section 2.

Source and domains. MEMORYVCD is built on the Amazon Reviews 2023 corpus (Hou et al., 2024) and spans **twelve** product domains. Four are *primary* domains used as evaluation targets, Personal Care, Books, Electronics, and Home & Kitchen; the other eight, All Beauty, Arts & Crafts & Sewing, Automotive, Grocery & Gourmet Food, Health & Household, Movies & TV, Office Products, and Sports & Outdoors, are *additional* domains that only populate the cross-domain growth axis (D-axis). The time axis and all six tasks are always evaluated on a primary target domain; the additional domains enter memory only as cross-domain context.

User selection. We select users by their activity in the *overlap* of the four primary domains. Concretely, we keep users who are simultaneously active in all four primary domains and, among those, retain the **323** with the most interactions in these domains. These users have long histories across the four domains, with **314,116** interactions in total (median **773** per user), each carrying a product image, a review title and body, an integer rating in $\{1, \ldots, 5\}$, and a timestamp. For each user, interactions are sorted by time, and a three-interaction window is the test target while all earlier interactions are memory candidates.

Memory construction at successive cutoffs. Let $H_{u,k}$ denote the raw history available to user u at cutoff k. Accessible history grows cumulatively, but each method selects, compresses, or organizes that history in its own way. The evaluation interface supplies the current history at each cutoff. It does not require all methods to update a previous stored state through the same recurrence. RecentMem and BM25 retrieval operate on the current pool; summary methods construct or reuse cached memory units, and the persona method can reuse a cached profile. Thus, the trajectories compare methods under growing accessible history, rather than establishing that all methods implement the same incremental update process.

Time axis. Interactions are sorted by timestamp within the target domain. For n_u interactions, the cutoffs are $c \in \{0, 0.25, 0.5, 0.75, 1.0\}$. At an interior cutoff, let $j = \mathrm{round}(cn_u)$, using Python's nearest-integer rounding. With zero-based indexing, the test window is $[j-1, j+2)$ and the memory is the prefix before that window. T_0 uses the first three interactions with empty initial memory; T_4 uses the last three with all earlier interactions. Users without a valid three-item window are skipped. Both memory and test targets therefore advance along the time axis.

Domain axis. The main trajectories report D_1 to D_4, with the target domain's last three interactions held fixed. Source domains enter in a fixed order, with the target domain added at D_4. Within

each source domain, memory contains only interactions strictly before the earliest test timestamp, $t < \min_{x \in \text{test}} t_x$. Target test items are also excluded at D_4. Items are sorted within each domain and the domain segments are concatenated in the specified source order. D_0 denotes empty initial memory where an explicit control evaluates it on the same last-three test window. Such controls are reported separately from the main D_1 to D_4 trajectories.

Memory within a test window. Rating, ranking, review generation, review summarization, and preference choice use the same initial memory for all three test items. Image Prediction instead reveals earlier test interactions before predicting the next one: prediction i uses the initial memory plus the preceding $i-1$ test interactions. For this task, an empty-memory setting is empty only for the first prediction. Missing target images and failed judge responses are excluded from the image-score mean; the reported number of scored examples can therefore be below three per user.

Task definitions. Each of the six tasks is constructed per user from that user's held-out test items and their memory pool $\mathcal{M}_u$ (Section 2). The input format applies to both memory items and target or candidate items. The `text` and `text_rendered` formats present item text as tokens and pixels, respectively. Their image hybrids also supply product photos. The `image_only` format supplies product photos and short identifiers, omitting the item text. The format sweep therefore measures the combined effect on memory and task inputs, rather than isolating memory representation alone. For summarization, the full review body is supplied as plain text in every format.

- **Rating Prediction.** Given $\mathcal{M}_u$ and the target item represented in the selected format, predict the user's integer rating $\hat{r} \in \{1, \ldots, 5\}$. The target's text field is its product title. The target user's review text, review title, and rating are withheld. *Metrics:* MAE, RMSE.
- **Item Ranking.** Rank the true next item within a **20-candidate** pool, the positive plus **19 negatives** sampled from same-domain items the user has never interacted with. *Metrics:* NDCG@K, Recall@K.
- **Review Generation.** Given $\mathcal{M}_u$ and the target item represented in the selected format, generate the user's review. The target's text field is its product title, and the reference is the user's review body. *Metrics:* ROUGE-1/L, BLEU-1/4.
- **Review Summarization.** Given $\mathcal{M}_u$, the target item represented in the selected format, and the full review body in plain text, generate a concise review title. The target item's text field is its product title. The reference is the review title written by the user. That title and the user's rating are removed from the target input. This is a short summarization task; overlap metrics do not directly measure whether the output preserves a user's style. *Metrics:* ROUGE-1/L, BLEU-1/4.
- **Item Preference Choice.** Choose the observed next item among five candidates. The **4 hard negatives** have the closest SBERT title embeddings within the same category. Similar titles make this a harder choice task, but do not guarantee that only visual cues distinguish the candidates. *Metric:* accuracy.
- **Image Prediction.** Given $\mathcal{M}_u$ only, predict, in words, the visual attributes (category, color, form, material) of the user's next product; an LLM judge scores the prediction against the real product image on the rubric below. *Metric:* judge score in $[1, 5]$.

Image-prediction judge. The main experiments use a `gpt-5` judge that sees the reference product image and rates five fine-grained visual dimensions, each on a 1 to 5 integer scale (Table 5). The reported *overall* score is a rounded weighted aggregate,

$$\text{overall} = \text{round}[0.10\, c_{\text{cat}} + 0.225\, (c_{\text{color}} + c_{\text{style}} + c_{\text{pkg}} + c_{\text{mat}})]\,,$$

computed for each prediction with Python's nearest-integer rounding and clipped to $[1, 5]$. The harness recomputes this value from the five dimension scores, rather than using the judge's emitted overall field. Tables report its mean over successfully scored predictions. Category receives a smaller weight than the visual attributes, so a category match alone cannot yield a high score. Table 6 gives the 1 to 5 scoring standard applied to each dimension, followed by the judge instructions with punctuation normalized.

Judge instructions (punctuation normalized). `You are an evaluator. The user predicted (in words) what kind of product another user would buy next.`

Table 5: The five visual dimensions scored by the image-prediction judge, with their weights in the *overall* aggregate.

Dimension	Weight	What it checks
`category_match`	0.10	Same product category (e.g. skincare vs. cosmetic vs. tool); a sanity check, down-weighted.
`color_match`	0.225	Does the predicted color palette match the reference image?
`style_match`	0.225	Visual style match (minimalist / ornate / retro / playful / clinical).
`packaging_design`	0.225	Packaging form and silhouette (bottle / tube / jar / box / pouch / spray / pump).
`material_finish`	0.225	Materials and surface finish (matte / glossy / metallic / frosted / transparent / textured).

Table 6: Scoring standard applied to each visual dimension (1 to 5 integer scale).

Score	Label	Criterion
5	Exact match	The predicted attribute is indistinguishable from the reference image.
4	Close match	Correct with a minor deviation (e.g. a neighboring shade or a close variant).
3	Partial match	Shares the broad attribute but differs in clear specifics.
2	Weak match	Only loosely related; most specifics differ.
1	No match / unrelated	The attribute is absent or contradicts the reference (category floor).

```
You will see the REAL product image they actually bought.  Score the
prediction on FIVE fine-grained VISUAL dimensions, each on a 1-5 integer
scale.  This task is about VISUAL taste.  Penalize predictions that get
the category right but miss visual specifics (color, finish, packaging).
1.  category_match:  same product category (e.g. skincare vs. cosmetic vs.
tool)?  (sanity check)
2.  color_match:  does the predicted color palette match the reference
image?
3.  style_match:  does the visual style match (minimalist / ornate /
retro / playful / clinical)?
4.  packaging_design:  does the packaging form (bottle / tube / jar / box
/ pouch / spray / pump) and shape silhouette match the reference image?
5.  material_finish:  do the materials and surface finish match (matte /
glossy / metallic / frosted / transparent / textured / soft-touch)?
Return ONLY a single JSON object on one line with these EXACT keys (all
integers 1-5):  {"category_match":  int, "color_match":  int,
"style_match":  int, "packaging_design":  int, "material_finish":  int,
"overall":  int} where `overall` is a weighted aggregate emphasising
visual fidelity:  overall = round(0.10 * category_match + 0.225 *
(color_match + style_match + packaging_design + material_finish)).  Do
not add any commentary.
```

B Full Domain-Axis Home & Kitchen-Target Trajectory Tables

The parallel domain-axis experiment uses Home & Kitchen as the target, adding its history at D_4. These tables report the same four methods and show where their relative scores differ from the Books setting.

C User Specificity with Full Memory

This control asks how much full-memory ranking depends on the individual's own history. D_4 includes all four primary domains, including the target domain. It therefore tests user specificity with full memory, rather than isolating transfer from other domains.

Table 7: **Home & Kitchen-target D-axis trajectories, all (backbone·format) configurations.** Canonical order with Books and H&K swapped so H&K enters at D_4. Rows are the 4 update methods within each of 6 task bands; each four-column block is one configuration over D_1 to D_4. Only the fixed-target D_1 to D_4 conditions are reported. Green = per-task D_4 winner per config.

	mini·T+I				mini·Tr+I				nano·T+I				nano·Tr+I			
Method	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
Rating (MAE, lower)																
RecentMem	0.727	0.780	0.793	0.663	0.770	0.703	0.677	0.667	0.810	0.823	0.890	0.740	0.887	0.827	0.843	0.793
LoCoMo	0.707	0.703	0.700	0.637	0.747	0.713	0.693	**0.663**	0.820	0.783	0.793	0.740	0.863	0.840	0.833	0.790
ReadAgent	0.660	0.640	0.610	**0.570**	0.693	0.720	0.663	0.673	0.743	0.677	0.650	**0.607**	0.997	0.840	0.847	**0.767**
PersonaAgent	0.747	0.690	0.663	0.710	0.737	0.713	0.720	0.700	0.790	0.710	0.687	0.733	0.870	0.857	0.803	**0.767**
Summarization (R-L, higher)																
RecentMem	0.168	0.160	0.177	**0.168**	0.173	0.170	0.169	**0.178**	0.146	0.146	0.165	**0.160**	0.148	0.153	0.164	**0.169**
LoCoMo	0.173	0.164	0.176	0.166	0.159	0.160	0.172	0.175	0.147	0.153	0.155	0.149	0.164	0.151	0.151	0.165
ReadAgent	0.162	0.154	0.156	0.167	0.170	0.165	0.166	0.176	0.147	0.146	0.150	0.151	0.160	0.153	0.142	0.150
PersonaAgent	0.163	0.143	0.146	0.156	0.164	0.167	0.155	0.171	0.151	0.149	0.139	0.147	0.163	0.149	0.153	0.162
Generation (R-L, higher)																
RecentMem	0.130	0.132	0.136	**0.143**	0.146	0.141	0.148	**0.151**	0.120	0.120	0.124	**0.127**	0.099	0.108	0.098	0.109
LoCoMo	0.132	0.131	0.131	0.135	0.142	0.141	0.139	0.145	0.122	0.120	0.120	0.122	0.102	0.104	0.104	**0.112**
ReadAgent	0.120	0.116	0.116	0.119	0.128	0.123	0.125	0.130	0.109	0.103	0.103	0.109	0.094	0.094	0.092	0.101
PersonaAgent	0.125	0.125	0.121	0.120	0.130	0.125	0.126	0.128	0.116	0.114	0.113	0.113	0.099	0.098	0.098	0.107
Ranking (NDCG@5, higher)																
RecentMem	0.251	0.279	0.313	0.386	0.226	0.266	0.262	0.340	0.133	0.176	0.221	0.308	0.148	0.140	0.126	0.185
LoCoMo	0.299	0.394	0.492	**0.855**	0.303	0.366	0.424	**0.783**	0.260	0.247	0.300	**0.722**	0.139	0.143	0.182	**0.299**
ReadAgent	0.270	0.290	0.355	0.461	0.252	0.281	0.278	0.337	0.229	0.187	0.259	0.324	0.150	0.135	0.137	0.183
PersonaAgent	0.332	0.402	0.496	0.641	0.310	0.388	0.460	0.649	0.236	0.327	0.345	0.558	0.143	0.147	0.164	0.283
Choice (Acc, higher)																
RecentMem	0.340	0.280	0.263	0.310	0.273	0.273	0.307	0.333	0.293	0.260	0.287	0.287	0.263	0.267	0.283	0.277
LoCoMo	0.387	0.380	0.390	**0.390**	0.343	0.350	0.383	**0.400**	0.327	0.300	0.300	0.297	0.270	0.303	0.317	**0.343**
ReadAgent	0.337	0.307	0.290	0.273	0.297	0.283	0.337	0.317	0.280	0.280	0.267	0.267	0.257	0.270	0.287	0.240
PersonaAgent	0.370	0.350	0.373	0.373	0.330	0.337	0.347	0.387	0.273	0.310	0.353	**0.370**	0.290	0.307	0.297	0.320
Image Pred. (1 to 5, higher)																
RecentMem	1.073	1.023	1.112	1.180	1.043	1.030	1.110	1.190	1.063	1.033	1.116	**1.203**	1.063	1.080	1.127	1.147
LoCoMo	1.050	1.030	1.040	1.050	1.017	1.043	1.027	1.027	1.047	1.040	1.047	1.027	1.047	1.050	1.050	1.027
ReadAgent	1.057	1.047	1.070	**1.183**	1.080	1.080	1.100	**1.193**	1.047	1.060	1.107	1.190	1.087	1.103	1.113	**1.183**
PersonaAgent	1.047	1.050	1.043	1.063	1.073	1.080	1.060	1.050	1.033	1.047	1.070	1.110	1.057	1.020	1.060	1.027

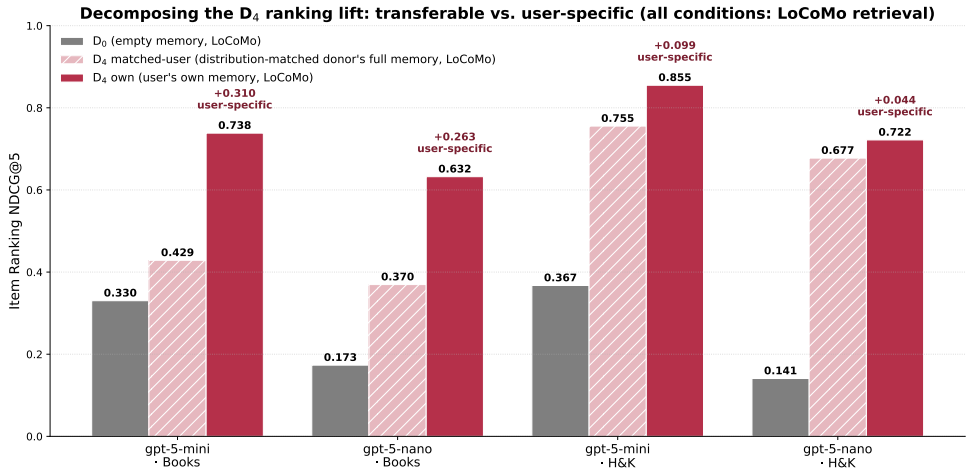


Figure 6: **User specificity with full D_4 memory.** Item-ranking NDCG@5 on the same target-user test under empty memory, a matched donor's full memory, and the user's own full memory. All conditions use LoCoMo retrieval, `text_image`, and 100 users. Both full-memory conditions include target-domain history. The advantage of own over matched memory is larger on Books than on Home & Kitchen in these settings.

A matched donor. On the same target user's last three test interactions, we compare empty memory (D_0), a matched donor's full memory (D_4 matched), and the user's own full memory (D_4 own). We use item ranking, `text_image`, 100 users, and LoCoMo retrieval throughout. The donor is a different user selected by primary-domain composition and history volume (median composition cosine 0.99, median donor-to-own volume ratio 1.00). Both memory conditions are filtered to precede the earliest target test timestamp. The donor's gain over empty memory measures information available from a similar user's history; the additional gain from own memory measures the advantage of the individual's own history under this matching rule.

Results. Figure 6 shows that both memory conditions improve ranking over empty memory in all four model and target settings. On Books, `gpt-5-mini` scores 0.330 with empty memory, 0.429 with matched memory, and 0.738 with own memory. The corresponding gains are +0.099 from the donor and a further +0.309 from own memory. For `gpt-5-nano`, the scores are 0.173, 0.370, and 0.632, respectively. On Home & Kitchen, the scores are 0.367, 0.755, and 0.855 for `gpt-5-mini`, and 0.141, 0.677, and 0.722 for `gpt-5-nano`. The additional advantage of own memory is therefore larger on Books (0.262 to 0.309 NDCG@5) than on Home & Kitchen (0.045 to 0.100). These point estimates support a domain-dependent advantage of own full history. They do not identify whether that advantage comes from in-domain or cross-domain interactions, nor do they establish its cause.

D Extended Memory-Update Baselines

Beyond the four canonical update policies of the main text (RecentMem, LoCoMo, ReadAgent, PersonaAgent; Table 4), we implement and evaluate **eight additional memory-update methods** spanning the current design space: **Stack-All** (long-context, no management), **A-Mem** (Xu et al., 2025) (agentic linked notes + evolution), **Mem0** (Chhikara et al., 2025) (LLM fact extraction + accumulation), **Zep/**Graphiti (Rasmussen et al., 2025) (temporal knowledge graph with bi-temporal invalidation), **MemOS** (Li et al., 2025) (tiered units + scheduled consolidation), **LightMem** (Fang et al., 2026) (topic-segmented offline consolidation), **SimpleMem** (Liu et al., 2026) (one-pass structured compression), and **MemSkill** (Zhang et al., 2026a) (inference-only skill-conditioned construction). These are adaptations under a common evaluation harness, rather than evaluations of the original systems without modification. For the seven constructive extensions other than Stack-All, query-time memory combines BM25-selected distilled units with up to ten raw interactions that carry product images. Construction processes interactions in chunks. Stack-All instead passes the full

Table 8: **Extended memory-update baselines, H&K target** (`text_image`; 12 methods: 4 canonical (shaded) + 8 extended, 6 tasks, 2 models {`5-mini`,`5-nano`}, 2 axes {T_4,D_4}). Per (task, model, axis) column the best method is **bold**. Rating is MAE (lower); all others higher-is-better. Image-prediction scores are low at D_4, with differences across methods; the Books results include recovery by RecentMem and ReadAgent.

Method	Rating MAE (lower)		Summ R-L (higher)		Gen R-L (higher)		Rank NDCG@5 (higher)		Choice Acc (higher)		ImgPred J (higher)	
	m	n	m	n	m	n	m	n	m	n	m	n
T_4												
RecentMem	0.650	0.750	0.183	**0.158**	**0.141**	**0.127**	0.401	0.295	0.333	0.283	1.522	1.556
Zep	0.673	0.750	**0.189**	0.156	0.134	0.121	0.692	0.603	0.390	0.363	1.522	1.444
LoCoMo	0.660	0.727	0.171	0.156	0.137	0.124	**0.795**	**0.709**	0.347	0.363	1.433	1.367
Stack-All	0.810	0.877	0.175	0.151	0.119	0.108	0.430	0.302	0.350	0.277	1.433	1.489
ReadAgent	**0.567**	**0.587**	0.165	0.143	0.121	0.108	0.458	0.324	0.330	0.300	**1.567**	**1.578**
LightMem	0.693	0.743	0.170	0.136	0.122	0.111	0.547	0.484	0.337	0.357	1.500	1.556
SimpleMem	0.740	0.747	0.186	0.154	0.130	0.117	0.682	0.606	0.373	0.353	1.500	1.444
MemOS	0.783	0.787	0.177	0.149	0.120	0.109	0.572	0.494	0.377	0.350	1.511	1.522
PersonaAgent	0.700	0.717	0.152	0.148	0.118	0.111	0.576	0.525	0.373	0.330	1.544	1.433
Mem0	0.757	0.823	0.170	0.155	0.126	0.115	0.687	0.590	**0.397**	0.363	1.367	1.411
A-Mem	0.640	0.687	0.163	0.141	0.117	0.108	0.614	0.527	0.383	**0.367**	1.556	1.551
MemSkill	0.753	0.733	0.176	0.153	0.124	0.112	0.627	0.523	0.350	0.317	1.433	1.400
D_4 (cross-domain)												
RecentMem	0.600	0.656	0.173	0.133	**0.141**	**0.129**	0.394	0.310	0.300	0.256	**1.156**	1.200
Zep	0.611	0.678	0.190	0.151	0.134	0.127	0.681	0.580	0.333	0.356	1.056	1.089
LoCoMo	0.611	0.656	0.191	0.145	0.137	0.126	**0.871**	**0.775**	0.389	**0.411**	1.033	1.056
Stack-All	0.689	0.767	0.177	0.144	0.118	0.108	0.386	0.196	0.289	0.278	1.044	1.033
ReadAgent	**0.511**	**0.478**	0.167	0.156	0.122	0.112	0.392	0.329	0.289	0.344	1.156	**1.233**
LightMem	0.622	0.744	0.190	0.135	0.126	0.118	0.760	0.567	0.322	0.356	1.067	1.033
SimpleMem	0.656	0.667	**0.217**	**0.157**	0.127	0.120	0.740	0.664	0.344	0.389	1.044	1.056
MemOS	0.689	0.756	0.168	0.138	0.125	0.117	0.673	0.561	0.289	0.311	1.033	1.033
PersonaAgent	0.656	0.711	0.155	0.137	0.126	0.118	0.691	0.513	0.344	0.389	1.033	1.056
Mem0	0.689	0.689	0.162	0.144	0.127	0.118	0.735	0.608	**0.422**	0.411	1.044	1.067
A-Mem	0.578	0.678	0.166	0.153	0.122	0.115	0.749	0.601	0.389	0.311	1.044	1.056
MemSkill	0.656	0.756	0.159	0.147	0.128	0.119	0.751	0.624	0.378	0.333	1.033	1.056

available history. ReadAgent uses episodic gists and recent raw interactions, while PersonaAgent combines a text persona with retrieved raw interactions. These choices standardize parts of the interface but do not equalize every method's token budget or preserve every component of its original implementation.

All tables use `text_image` with `gpt-5-mini` and `gpt-5-nano`. T_4 uses 100 users for the five non-image-prediction tasks and 30 users for image prediction; D_4 uses 30 users for every task. Image prediction is scored only where a target image is available, so the number of scored items can be below three per user. These are separate runs from the main trajectory sweep. Tables report their completed evaluation results.

Findings. Several structured-memory adaptations improve ranking over Stack-All. For example, on Books D_4 with `gpt-5-mini`, Zep, Mem0, and A-Mem score 0.718, 0.732, and 0.740, compared with Stack-All's 0.552. However, the BM25 baseline LoCoMo scores 0.746, so these results do not establish that graph or distilled memory is best for ranking. No method leads every task. RecentMem and ReadAgent retain an advantage on several generation and rating settings. Most Books D_4 image-prediction scores remain near the floor, but RecentMem and ReadAgent score about 1.8 to 2.0. Table 10 reports API usage for these runs; Stack-All has the largest recorded inference-token total.

E Verbatim Recall and Data Exposure

MemoryVCD uses public Amazon review data, so some models may have encountered its products or reviews during training. We examine one form of extractable recall: continuing a review from a supplied prefix. This probe does not test training-set membership and cannot establish that the benchmark is free of contamination.

Table 9: **Extended memory-update baselines, Books target** (`text_image`; 12 methods: 4 canonical (shaded) + 8 extended, 6 tasks, 2 models {`5-mini`,`5-nano`}, 2 axes {T_4,D_4}). Per (task, model, axis) column the best method is **bold**. Rating is MAE (lower); all others higher-is-better. Image-prediction scores are low at D_4, with differences across methods; the Books results include recovery by RecentMem and ReadAgent.

Method	Rating MAE (lower)		Summ R-L (higher)		Gen R-L (higher)		Rank NDCG@5 (higher)		Choice Acc (higher)		ImgPred J (higher)	
	m	n	m	n	m	n	m	n	m	n	m	n
T_4												
RecentMem	0.677	0.747	**0.162**	0.132	**0.134**	**0.120**	0.529	0.401	0.553	0.460	**2.734**	**2.722**
Zep	0.717	0.800	0.152	0.134	0.127	0.114	0.614	0.528	0.543	0.473	2.519	2.481
LoCoMo	0.707	0.780	0.150	0.139	0.125	0.118	**0.713**	**0.581**	**0.587**	0.477	2.633	2.722
Stack-All	0.720	0.740	0.154	0.143	0.125	0.110	0.580	0.409	0.507	0.470	2.354	2.316
ReadAgent	0.653	**0.663**	0.147	0.129	0.122	0.108	0.600	0.457	0.493	0.457	2.481	2.532
LightMem	0.670	0.700	0.146	0.124	0.117	0.108	0.606	0.497	0.543	0.510	2.544	2.519
SimpleMem	0.670	0.727	0.151	0.125	0.118	0.108	0.606	0.494	0.570	0.483	2.316	2.380
MemOS	0.720	0.797	0.156	0.121	0.115	0.106	0.601	0.523	0.540	0.503	2.532	2.468
PersonaAgent	0.700	0.777	0.131	0.129	0.115	0.106	0.617	0.482	0.550	0.480	2.620	2.354
Mem0	0.677	0.777	0.145	0.128	0.120	0.109	0.644	0.548	0.557	**0.527**	2.620	2.544
A-Mem	**0.650**	0.717	0.143	0.127	0.118	0.107	0.615	0.495	0.533	0.497	2.456	2.544
MemSkill	0.687	0.740	0.155	**0.147**	0.120	0.109	0.616	0.507	0.570	0.493	2.544	2.443
D_4 (cross-domain)												
RecentMem	0.544	0.633	0.173	0.154	**0.138**	**0.123**	0.542	0.393	0.533	0.367	**1.962**	**2.000**
Zep	0.600	0.722	0.171	0.132	0.127	0.119	0.718	0.531	0.578	0.500	1.000	1.013
LoCoMo	0.533	0.611	0.159	0.135	0.129	0.119	**0.746**	0.637	0.589	0.556	1.013	1.013
Stack-All	0.589	0.667	0.154	**0.162**	0.117	0.104	0.552	0.418	0.556	0.444	1.000	1.000
ReadAgent	0.511	0.556	0.150	0.153	0.122	0.110	0.589	0.476	0.600	0.411	1.772	1.899
LightMem	0.544	0.667	0.142	0.152	0.121	0.113	0.713	0.633	0.567	0.522	1.000	1.013
SimpleMem	0.533	0.600	0.167	0.153	0.123	0.114	0.698	0.642	0.567	0.500	1.000	1.013
MemOS	0.556	0.644	0.162	0.150	0.122	0.113	0.644	0.620	0.589	0.511	1.000	1.013
PersonaAgent	**0.500**	0.578	0.123	0.148	0.123	0.117	0.716	**0.646**	**0.656**	0.456	1.038	1.051
Mem0	0.556	0.667	0.181	0.155	0.123	0.114	0.732	0.631	0.589	**0.567**	1.000	1.000
A-Mem	0.544	**0.533**	0.159	0.147	0.122	0.112	0.740	0.540	0.544	0.500	1.000	1.000
MemSkill	0.511	0.633	**0.182**	0.152	0.127	0.113	0.701	0.587	0.589	0.567	1.000	1.000

Table 10: **Efficiency: four canonical + eight extended baselines.** Recorded *build* API calls and tokens versus summed predictor *inference* tokens, in millions (M). Build columns use the largest cell-level build total recorded for each method and setting; they are not deduplicated one-time construction costs. Inference totals sum the six tasks and two backbones and exclude judge tokens. RecentMem, LoCoMo, and Stack-All make no LLM construction calls. Stack-All has the largest recorded inference total.

Method	H&K T_4		H&K D_4		Books T_4		Books D_4	
	bld/inf(M)	calls	bld/inf(M)	calls	bld/inf(M)	calls	bld/inf(M)	calls
RecentMem	0.0/22.1	0	0.0/7.3	0	0.0/17.2	0	0.0/5.6	0
Zep	18.9/22.6	6672	1.5/7.7	600	4.2/18.4	1724	1.5/6.4	590
LoCoMo	0.0/21.3	0	0.0/7.3	0	0.0/17.3	0	0.0/5.9	0
Stack-All	0.0/134.9	0	0.0/114.0	0	0.0/74.2	0	0.0/82.8	0
ReadAgent	33.3/29.8	19614	27.1/19.2	16068	15.1/18.5	10176	20.7/14.9	12510
LightMem	27.7/23.0	30042	0.7/7.9	330	1.9/18.7	990	0.6/6.5	327
SimpleMem	22.9/23.2	7293	1.6/8.0	630	4.6/19.1	1824	1.6/6.6	620
MemOS	19.8/23.7	11853	1.0/8.1	720	2.8/19.2	2113	1.0/6.7	710
PersonaAgent	0.5/22.8	300	0.2/7.8	90	0.5/18.4	300	0.1/6.4	90
Mem0	17.0/23.0	9525	0.9/7.8	600	2.2/18.6	1724	0.8/6.4	590
A-Mem	24.0/26.3	15849	1.0/8.9	780	2.7/21.4	2233	0.9/7.5	768
MemSkill	19.7/23.5	10185	1.0/7.9	630	2.9/19.0	1824	0.9/6.4	620

Protocol. We split each of 300 sampled reviews at the halfway word boundary and ask each of the twelve backbones to reproduce the original continuation. Verbatim probes have also been used in contamination analysis (Golchin & Surdeanu, 2025). We measure ROUGE-L against the true suffix

Table 11: Verbatim continuation probe on 300 reviews. ROUGE-L is measured against the true suffix and a random review's suffix (`ctrl`). The gap is their difference in percentage points, computed before rounding. `max span` is the longest exact contiguous token match observed for that model. These quantities measure extractable overlap under this prompt, not whether a review was present in training.

Model	Family	ROUGE-L %	ctrl %	gap pp	max span tok
`gpt-5.5`	GPT	8.2	8.2	+0.0	4
`gpt-5`	GPT	10.6	10.1	+0.4	4
`gpt-5-mini`	GPT	11.9	10.7	+1.2	4
`gpt-5-nano`	GPT	11.2	10.5	+0.6	6
`gpt-4o`	GPT	6.6	6.1	+0.5	5
`gpt-4o-mini`	GPT	13.9	11.8	+2.2	5
`opus-4.7`	Claude	11.1	10.6	+0.5	7
`sonnet-4.6`	Claude	17.3	11.7	+5.6	12
`haiku-4.5`	Claude	12.0	11.6	+0.4	5
`3.1-pro`	Gemini	11.1	7.5	+3.6	16
`3-flash`	Gemini	13.2	8.5	+4.8	12
`2.5-flash`	Gemini	16.6	11.0	+5.6	12

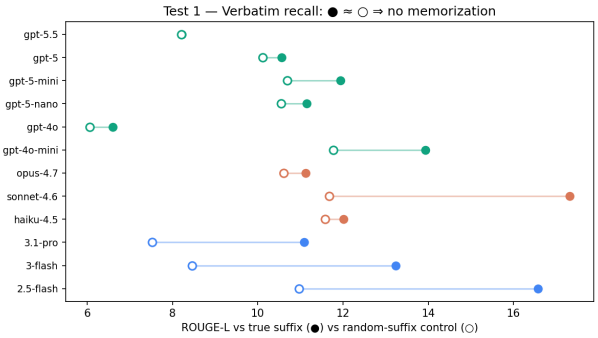


Figure 7: Verbatim continuation overlap. Each model's output is compared with the true review suffix and a random review's suffix. The random suffix is a reference for generic overlap, not a topic-matched control. A positive gap can arise from contextual continuation as well as memorization.

and the longest exact contiguous token match. As a reference for generic overlap, we also score the generated continuation against a randomly chosen different review's suffix. The true-minus-random gap can reflect the topical information in the supplied prefix as well as possible memorization. It does not isolate memorization by itself. Runs use temperature zero where supported and the model's default otherwise. The accompanying plot reports bootstrap 95% intervals.

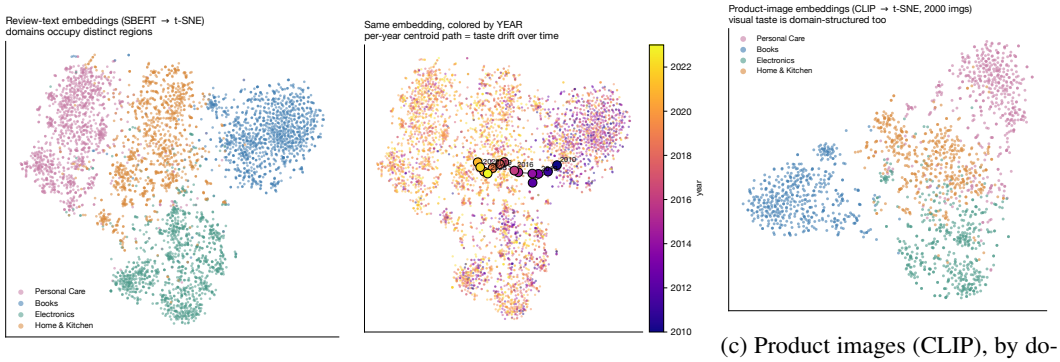
Observed overlap. Table 11 reports ROUGE-L gaps from 0.0 to 5.6 percentage points. The mean longest exact match ranges from 1.3 to 2.9 tokens across models; the largest individual match is 16 tokens. Inspection of long matches found public book titles, legal boilerplate, and common review phrases, such as "dishwasher safe, but hand washing is recommended." Such matches alone do not establish recall of a specific user's review. However, their presence also cannot rule out prior exposure to that text.

Limits of the probe. The average exact matches are short under this prompt, but the sample is small and the random control does not match the topic of the true suffix. A model may also retain product facts or review associations without producing a verbatim continuation. These results therefore characterize observable text overlap in this probe; they do not prove absence of data exposure or quantify its effect on the personalization scores. Stronger membership and association probes would be needed for those claims.

F CORPUS ANALYSIS (EXTENDED)

This appendix collects the corpus analyses referenced from Section 2.1. All embedding figures are generated by `scripts/make_user_embeddings.py` (SBERT `all-MiniLM-L6-v2` for review text, `clip-ViT-B-32` for product images; per-item embeddings group-averaged per user/year/domain, then t-SNE).

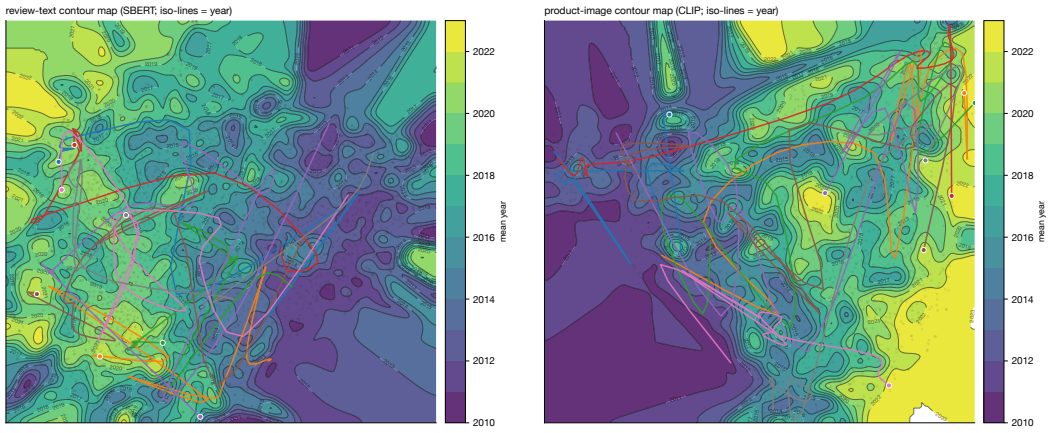
Domain and time structure at the corpus level. Figure 8 displays review-text embeddings by domain and year, together with product-image embeddings by domain. The projections show differences in observed content across domains and time. They do not measure predictive transfer.



(a) Review text (SBERT), by domain. (b) Same text embedding, by year. (c) Product images (CLIP), by domain.

Figure 8: **Content embeddings by domain and time.** (a) SBERT review-text embeddings colored by domain. (b) The same projection colored by year, with yearly centroids. (c) CLIP product-image embeddings colored by domain. The plots use 4,800 sampled reviews and 2,000 sampled product images.

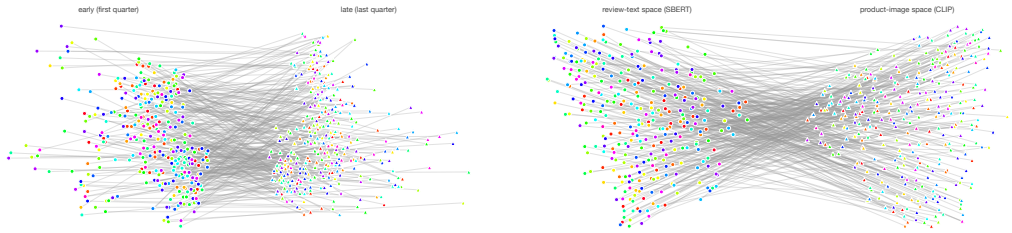
Topographic (contour) view of the time landscape. Figure 9 projects per-user yearly embeddings into a shared t-SNE view and interpolates year values as contour lines. The image view uses at most the 150 most active users. Selected paths connect each user's yearly positions.



(a) Review text (SBERT); iso-lines = year. (b) Product images (CLIP); iso-lines = year.

Figure 9: **Year contours for user-history embeddings.** Contours interpolate the year associated with mean review-text (a) and product-image (b) embeddings. Colored paths connect selected users across years. The projection is a descriptive view of the histories.

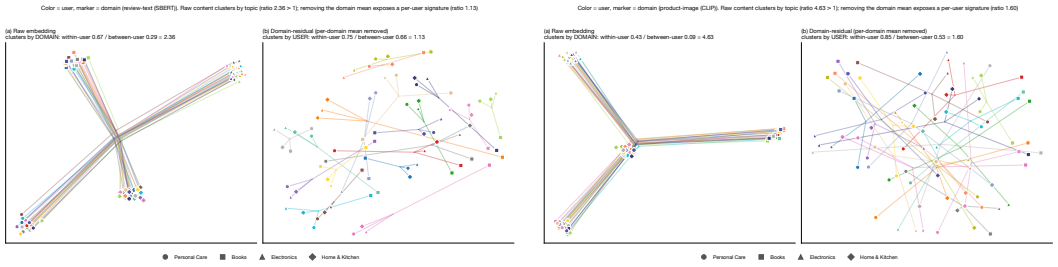
Migration plots: time cut and cross-modality. Figure 10a compares each user's first and last quarters of history. Figure 10b connects the text and image projections for the same user. Gray lines show these pairings across the 323 users.



(a) Early (first quarter) vs. late (last quarter), text. (b) Text (SBERT) vs. image (CLIP) space.

Figure 10: **Paired views of the 323 user histories.** (a) Review-text centroids from the first and last quarters in a shared t-SNE projection. Circles denote the first quarter and triangles the last. (b) Paired text and image positions for each user. Distances and crossing lines in these plots do not by themselves establish preference drift or complementary predictive information.

Removing the domain mean. The reported within-user to between-user distance ratio is 2.36 for raw text embeddings and 4.63 for raw image embeddings. After subtracting each domain's mean, it becomes 1.13 and 1.60, respectively. Removing the domain mean reduces this ratio in both views, but both residual ratios remain above one. These summaries do not establish a domain-invariant preference representation or explain the visual-prediction results. Figure 12 reports the users' interaction counts across the twelve domains.



(a) Review text (SBERT): raw 2.36 to residual 1.13. (b) Product images (CLIP): raw 4.63 to residual 1.60.

Figure 11: **Embeddings before and after removing the domain mean.** Color denotes user and marker denotes domain for the 20 most active users in the four-domain cohort. The residual view subtracts each domain's mean before projection. Text and image views are descriptive and do not directly measure transfer.

G Sample-Size Stability Study

We compare scores at several user counts with the full cohort of 323 users, using RecentMem at T_4. The evaluation at 100 users reuses the canonical storage results; evaluations at other counts retain the original sample-size runs. This comparison includes run and version differences and does not isolate sample-size effects. It provides neither confidence intervals for method differences nor validation of every trajectory shape. The comparison covers Books and Home & Kitchen, three models (`gpt-5`, `gpt-5-mini`, and `gpt-5-nano`), and two formats (`text_image` and `text_rendered_image`). At 100 users, the mean absolute deviation from the full-cohort score across these settings is 3.20% for rating, 8.20% for summarization, 1.23% for generation, 8.92% for ranking, 10.98% for choice, and 4.27% for image prediction. The mean across the six tasks is 6.13%, with three tasks below 5%.

For the Home & Kitchen setting used in the main time-axis table (`gpt-5-mini`, `text_image`, RecentMem, T_4), summarization ROUGE-L is 0.133 at 100 users and 0.201 for the full cohort of

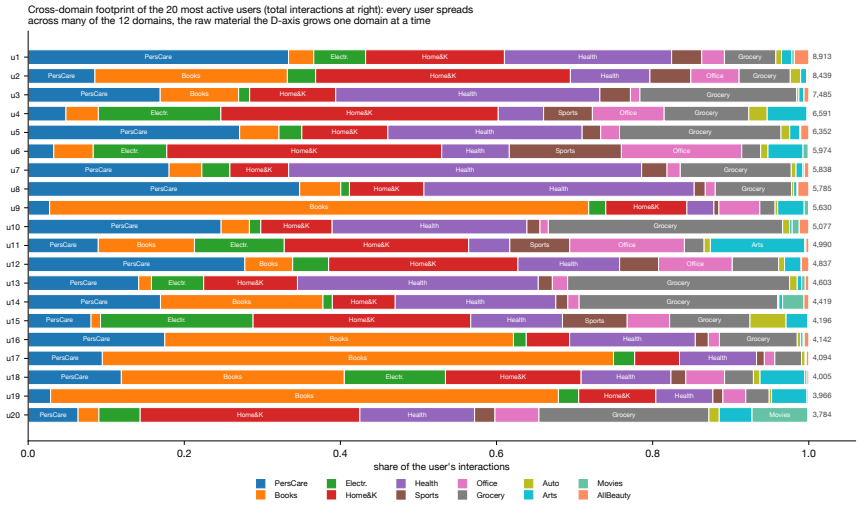


Figure 12: **Cross-domain footprint of the 20 most active users** (share of interactions per domain; total at right). Every heavy user spans many of the 12 Amazon domains, the raw material the D-axis grows one domain at a time.

323 users. As noted above, this comparison includes differences between runs and versions, so it does not isolate the effect of sample size.

H Related Work

Memory benchmarks. LongMemEval (Wu et al., 2025), MemBench (Tan et al., 2025), and MSC (Xu et al., 2022) evaluate memory in extended text conversations. LoCoMo (Maharana et al., 2024) uses generated conversations checked by human annotators and includes image sharing and multimodal dialogue generation. MemoryAgentBench (Hu et al., 2026) studies memory under incremental input. Other recent benchmarks, including MemLens (Ren et al., 2026), M^{3}Exam (Huang et al., 2026b), MemEye (Guo et al., 2026), and Mem-Gallery (Bei et al., 2026), also evaluate visual memory. These studies cover different parts of the memory problem. Our focus is the joint evaluation of memory representation and memory management on real users' product histories, with evaluation at multiple time and domain cutoffs. Table 1 summarizes the comparison. Values marked with † are estimates from the cited papers rather than measurements made here.

Personalization and cross-domain history. LaMP (Salemi et al., 2024) studies user-conditioned retrieval and generation, and PersonaAgent (Zhang et al., 2025) uses a user persona to guide predictions. Recent work also examines how personal context should be evaluated (Huang et al., 2026a). MemoryCD (Zhang et al., 2026b) is the closest benchmark to ours: it uses real Amazon review histories across years and domains to evaluate four personalization tasks. MemoryVCD builds on this behavior-based view of personal memory. Its additional focus is a comparison of five memory input formats, two further tasks involving visual cues, and trajectories over growing time and domain histories.

Memory-management methods. We compare recency selection, retrieval over a raw history as in LoCoMo (Maharana et al., 2024), episodic summaries inspired by ReadAgent (Lee et al., 2024), and profiles inspired by PersonaAgent (Zhang et al., 2025). The labels in the result tables refer to our benchmark adaptations of these approaches. In particular, LoCoMo names the BM25 retrieval baseline used here, rather than a separate memory-update algorithm proposed by that benchmark. We also adapt A-Mem (Xu et al., 2025), Mem0 (Chhikara et al., 2025), Zep (Rasmussen et al., 2025), MemOS (Li et al., 2025), LightMem (Fang et al., 2026), SimpleMem (Liu et al., 2026), and MemSkill (Zhang et al., 2026a), alongside a Stack-All baseline. The shared harness makes their inputs and task evaluation comparable, but the results apply to these implementations and settings,

rather than to every component or deployment of the original systems. Appendix D describes the adaptations and their costs.

I Full Storage-Axis Results for Both Targets

The tables below report all available task metrics for the four image-based formats at T_4/RecentMem across twelve backbones. Plain text is a necessary comparison: Table 3 reports its primary metrics for Home & Kitchen, and Table 12 reports them for Books. The four image-based formats are rendered text, product images, text plus product images, and rendered text plus product images. Bold marks the best unrounded model score in each format and metric column; printed ties can reflect rounding. Books ranking uses the corrected ISBN-10 candidate parser.

Model	MAE ↓	Summ R-L	Gen R-L	NDCG@5	Choice Acc	ImgPred
gpt-5	0.680	0.183	0.129	0.558	0.470	1.866
gpt-5.5	0.640	0.201	0.149	0.568	0.550	3.257
gpt-5-mini	0.660	0.158	0.133	0.464	0.413	1.832
gpt-5-nano	0.687	0.139	0.113	0.413	0.353	1.832
gpt-4o	0.733	0.119	0.128	0.483	0.430	2.727
gpt-4o-mini	0.773	0.141	0.124	0.320	0.400	2.720
claude-opus-4.7	0.653	0.222	0.152	0.566	0.503	3.011
claude-sonnet-4.6	0.843	0.172	0.143	0.478	0.420	3.291
claude-haiku-4.5	0.947	0.153	0.138	0.430	0.393	2.903
gemini-3.1-pro	0.630	0.189	0.160	0.605	0.527	2.511
gemini-3-flash	0.637	0.192	0.155	0.566	0.340	2.660
gemini-2.5-flash	0.767	0.176	0.148	0.501	0.350	3.351

Table 12: **Plain-text baseline on Books**, T_4/RecentMem. This table supplies the primary-metric reference for the image-based formats below. Rating uses MAE, where lower is better; all other metrics are higher is better. Image prediction uses the judge overall score on its 1 to 5 scale.

Model	text-rend		image		text+img		text-rend+img	
	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)
gpt-5	0.66	1.08	0.68	1.13	0.61	1.06	0.63	1.04
gpt-5.5	0.67	1.04	0.68	1.07	0.63	1.02	0.63	**1.01**
gpt-5-mini	0.70	1.08	0.66	1.12	0.65	1.04	0.65	1.05
gpt-5-nano	0.75	1.04	0.85	1.11	0.74	1.06	0.78	1.05
gpt-4o	0.75	1.03	0.81	**1.02**	0.77	**1.01**	0.72	1.02
gpt-4o-mini	0.83	1.04	0.97	1.20	0.83	1.03	0.85	1.06
claude-opus-4.7	0.78	1.08	0.69	1.09	0.71	1.01	0.76	1.05
claude-sonnet-4.6	0.79	**1.01**	0.85	1.06	0.84	1.06	0.78	1.01
claude-haiku-4.5	0.88	1.11	0.94	1.19	0.85	1.07	0.86	1.07
gemini-3.1-pro	0.63	1.13	**0.62**	1.14	0.62	1.09	0.61	1.11
gemini-3-flash	**0.60**	1.08	0.62	1.14	0.61	1.09	**0.57**	1.07
gemini-2.5-flash	0.64	1.09	0.71	1.16	**0.61**	1.07	0.63	1.06

Table 13: **Rating Prediction** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend					image					text+img					text-rend+img				
	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS
gpt-5	0.21	0.20	0.10	0.03	0.86	0.16	0.14	0.04	0.01	0.85	0.20	0.19	0.10	0.03	0.86	0.21	0.20	0.10	0.03	0.86
gpt-5.5	**0.24**	0.22	**0.14**	**0.05**	0.87	0.20	0.19	0.07	0.02	0.86	0.23	0.22	0.13	0.05	0.87	0.23	0.22	0.14	0.05	0.87
gpt-5-mini	0.20	0.18	0.09	0.02	0.86	0.17	0.16	0.05	0.01	0.85	0.14	0.13	0.06	0.02	0.68	0.19	0.18	0.09	0.02	0.86
gpt-5-nano	0.16	0.16	0.06	0.01	0.85	0.15	0.14	0.05	0.01	0.85	0.17	0.16	0.06	0.02	0.85	0.17	0.16	0.05	0.01	0.85
gpt-4o	0.16	0.15	0.06	0.01	0.85	0.13	0.12	0.04	0.01	0.84	0.15	0.14	0.05	0.01	0.85	0.14	0.14	0.04	0.01	0.84
gpt-4o-mini	0.17	0.16	0.07	0.02	0.85	0.14	0.13	0.04	0.01	0.84	0.15	0.14	0.06	0.02	0.85	0.15	0.14	0.06	0.02	0.85
claude-opus-4.7	0.24	**0.23**	0.14	0.05	**0.87**	**0.21**	**0.20**	0.09	0.02	0.86	**0.26**	**0.25**	**0.15**	**0.05**	**0.87**	**0.24**	**0.23**	0.13	0.05	0.87
claude-sonnet-4.6	0.23	0.21	0.13	0.04	0.86	0.18	0.17	0.05	0.01	0.85	0.22	0.21	0.12	0.04	0.86	0.21	0.20	0.12	0.04	0.86
claude-haiku-4.5	0.17	0.16	0.07	0.02	0.86	0.16	0.15	0.04	0.01	0.85	0.18	0.17	0.09	0.03	0.86	0.18	0.17	0.07	0.02	0.85
gemini-3.1-pro	0.23	0.22	0.13	0.05	0.87	0.21	0.20	**0.09**	**0.02**	**0.86**	0.25	0.23	0.14	0.05	0.87	0.23	0.23	**0.14**	**0.05**	**0.87**
gemini-3-flash	0.23	0.22	0.13	0.04	0.87	0.18	0.17	0.07	0.02	0.85	0.22	0.21	0.12	0.04	0.86	0.23	0.22	0.12	0.04	0.86
gemini-2.5-flash	0.22	0.20	0.10	0.03	0.86	0.17	0.16	0.06	0.02	0.85	0.22	0.21	0.11	0.04	0.86	0.19	0.18	0.08	0.02	0.86

Table 14: **Review Summarization** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend					image					text+img					text-rend+img				
	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS
gpt-5	0.26	0.13	0.15	0.01	0.84	0.23	0.11	0.12	0.00	0.83	0.27	0.13	0.15	0.01	0.84	0.26	0.13	0.15	0.01	0.84
gpt-5.5	**0.31**	0.17	**0.20**	0.02	0.86	**0.28**	0.14	**0.17**	0.01	0.85	0.31	0.16	0.19	0.02	0.86	**0.32**	0.17	**0.20**	0.02	0.86
gpt-5-mini	0.28	0.15	0.17	0.01	0.85	0.24	0.12	0.13	0.01	0.83	0.28	0.14	0.16	0.01	0.85	0.28	0.15	0.17	0.01	0.85
gpt-5-nano	0.21	0.11	0.12	0.01	0.83	0.20	0.10	0.11	0.00	0.83	0.25	0.13	0.14	0.01	0.84	0.20	0.11	0.12	0.01	0.82
gpt-4o	0.27	0.14	0.16	0.01	0.84	0.23	0.12	0.13	0.00	0.84	0.27	0.14	0.16	0.01	0.84	0.27	0.14	0.16	0.01	0.84
gpt-4o-mini	0.26	0.13	0.15	0.01	0.84	0.24	0.12	0.13	0.01	0.84	0.27	0.13	0.15	0.01	0.84	0.26	0.13	0.15	0.01	0.84
claude-opus-4.7	0.31	0.16	0.19	0.02	0.85	0.28	**0.14**	0.17	**0.01**	**0.85**	0.31	0.16	0.19	0.02	0.85	0.31	0.17	0.19	**0.02**	0.85
claude-sonnet-4.6	0.31	0.16	0.19	0.02	0.85	0.25	0.12	0.14	0.01	0.84	0.29	0.15	0.17	0.01	0.85	0.31	0.16	0.19	0.02	0.85
claude-haiku-4.5	0.28	0.15	0.16	0.01	0.85	0.23	0.11	0.12	0.00	0.83	0.28	0.14	0.16	0.01	0.85	0.28	0.14	0.16	0.01	0.85
gemini-3.1-pro	0.31	**0.17**	0.19	0.02	**0.86**	0.27	0.13	0.16	0.01	0.84	**0.32**	**0.18**	**0.20**	0.02	**0.86**	0.31	**0.17**	0.19	0.02	**0.86**
gemini-3-flash	0.31	0.17	0.19	**0.02**	0.85	0.27	0.14	0.16	0.01	0.85	0.32	0.17	0.19	**0.02**	0.86	0.31	0.17	0.18	0.02	0.86
gemini-2.5-flash	0.29	0.16	0.17	0.01	0.85	0.25	0.13	0.14	0.01	0.84	0.30	0.16	0.17	0.01	0.85	0.29	0.16	0.17	0.02	0.85

Table 15: **Review Generation** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend						image						text+img						text-rend+img					
	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5
gpt-5	0.16	0.23	0.26	0.16	0.28	0.35	0.21	0.31	0.37	0.21	0.39	0.53	0.24	0.35	0.40	0.24	0.44	0.55	0.18	0.24	0.28	0.18	0.28	0.40
gpt-5.5	0.17	0.27	0.33	0.17	0.35	0.49	0.22	0.35	0.41	0.22	**0.45**	0.59	**0.26**	**0.38**	**0.45**	**0.26**	**0.47**	**0.64**	0.21	0.33	0.38	0.21	0.42	0.55
gpt-5-mini	0.18	0.27	0.32	0.18	0.33	0.45	0.17	0.30	0.35	0.17	0.39	0.52	0.23	0.34	0.40	0.23	0.43	0.56	0.19	0.28	0.35	0.19	0.35	0.51
gpt-5-nano	0.11	0.16	0.19	0.11	0.21	0.27	0.10	0.19	0.25	0.10	0.26	0.40	0.17	0.25	0.29	0.17	0.31	0.42	0.09	0.17	0.21	0.09	0.23	0.33
gpt-4o	0.07	0.13	0.17	0.07	0.18	0.26	0.12	0.23	0.28	0.12	0.30	0.44	0.15	0.22	0.27	0.15	0.27	0.40	0.06	0.11	0.15	0.06	0.14	0.25
gpt-4o-mini	0.04	0.09	0.13	0.04	0.14	0.23	0.07	0.12	0.16	0.07	0.17	0.24	0.09	0.15	0.18	0.09	0.19	0.26	0.06	0.12	0.15	0.06	0.16	0.25
claude-opus-4.7	0.22	0.33	0.39	0.22	0.42	0.55	**0.23**	**0.35**	**0.42**	**0.23**	0.45	**0.60**	0.23	0.36	0.43	0.23	0.45	0.63	0.20	0.31	0.38	0.20	0.39	**0.57**
claude-sonnet-4.6	0.17	0.26	0.31	0.17	0.33	0.45	0.19	0.29	0.33	0.19	0.35	0.45	0.22	0.33	0.37	0.22	0.42	0.52	0.23	0.33	0.39	0.23	0.41	0.55
claude-haiku-4.5	0.11	0.17	0.20	0.11	0.21	0.29	0.15	0.24	0.29	0.15	0.31	0.43	0.18	0.27	0.33	0.18	0.33	0.47	0.11	0.17	0.19	0.11	0.21	0.27
gemini-3.1-pro	**0.23**	**0.34**	**0.40**	**0.23**	0.43	**0.56**	0.22	0.32	0.39	0.22	0.40	0.56	0.24	0.36	0.40	0.24	0.44	0.55	**0.24**	**0.36**	**0.40**	**0.24**	**0.44**	0.55
gemini-3-flash	0.21	0.34	0.39	0.21	**0.44**	0.55	0.19	0.33	0.38	0.19	0.42	0.56	0.23	0.34	0.39	0.23	0.42	0.56	0.22	0.34	0.38	0.22	0.42	0.53
gemini-2.5-flash	0.18	0.30	0.34	0.18	0.38	0.49	0.15	0.30	0.35	0.15	0.41	0.52	0.23	0.37	0.42	0.23	0.47	0.59	0.19	0.29	0.35	0.19	0.36	0.52

Table 16: **Item Ranking** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend		image		text+img		text-rend+img	
	Acc	Sim	Acc	Sim	Acc	Sim	Acc	Sim
gpt-5	**0.34**	0.75	**0.33**	0.75	0.36	0.75	0.30	0.75
gpt-5.5	0.32	**0.86**	0.28	**0.86**	**0.37**	**0.86**	**0.34**	**0.86**
gpt-5-mini	0.30	0.75	0.25	0.75	0.36	0.75	0.32	0.75
gpt-5-nano	0.23	0.75	0.30	0.75	0.31	0.75	0.27	0.75
gpt-4o	0.27	0.86	0.32	0.86	0.31	0.86	0.29	0.86
gpt-4o-mini	0.27	0.86	0.31	0.86	0.32	0.86	0.29	0.86
claude-opus-4.7	0.29	0.86	0.27	0.86	0.34	0.86	0.32	0.86
claude-sonnet-4.6	0.30	0.86	0.31	0.86	0.31	0.86	0.31	0.86
claude-haiku-4.5	0.28	0.86	0.32	0.86	0.27	0.86	0.25	0.86
gemini-3.1-pro	0.32	0.86	0.28	0.86	0.33	0.86	0.34	0.86
gemini-3-flash	0.28	0.86	0.27	0.86	0.26	0.86	0.30	0.86
gemini-2.5-flash	0.25	0.86	0.28	0.86	0.26	0.86	0.25	0.86

Table 17: **Item Preference Choice** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend						image						text+img						text-rend+img					
	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All
gpt-5	1.23	1.27	1.51	1.04	1.20	1.19	1.15	1.31	1.58	1.06	1.23	1.23	1.11	1.26	1.51	1.03	1.12	1.17	1.19	1.30	1.57	1.05	1.22	1.22
gpt-5.5	1.62	1.98	2.16	1.35	1.79	1.74	1.54	2.06	2.15	1.32	1.77	1.76	1.61	2.04	2.08	1.34	1.75	1.76	1.60	2.05	2.14	1.35	1.75	1.77
gpt-5-mini	1.20	1.27	1.45	1.05	1.19	1.18	1.16	1.25	1.42	1.06	1.19	1.18	1.25	1.27	1.43	1.04	1.18	1.19	1.23	1.30	1.44	1.06	1.19	1.21
gpt-5-nano	1.25	1.24	1.49	1.07	1.21	1.20	1.22	1.31	1.56	1.05	1.20	1.21	1.26	1.30	1.46	1.08	1.22	1.21	1.30	1.26	1.45	1.12	1.22	1.22
gpt-4o	**2.24**	**2.44**	**2.70**	**1.49**	**2.14**	**2.16**	1.94	**2.32**	**2.57**	1.44	**2.08**	**2.04**	**2.21**	**2.42**	**2.70**	1.52	**2.03**	**2.13**	**2.20**	**2.39**	**2.69**	1.46	**2.09**	**2.17**
gpt-4o-mini	1.58	1.87	2.21	1.22	1.69	1.70	1.61	1.86	2.25	1.29	1.77	1.76	1.53	1.83	2.10	1.21	1.70	1.63	1.45	1.80	2.11	1.23	1.62	1.63
claude-opus-4.7	1.65	2.08	2.13	1.29	1.78	1.78	1.65	2.09	2.13	1.30	1.74	1.78	1.68	2.07	2.15	1.31	1.74	1.76	1.78	2.07	2.14	1.33	1.82	1.77
claude-sonnet-4.6	1.86	2.31	2.34	1.47	1.98	1.96	**2.27**	1.73	2.13	**1.60**	1.60	1.73	1.56	1.96	2.02	1.40	1.69	1.65	1.56	1.99	2.13	1.26	1.69	1.70
claude-haiku-4.5	1.82	2.35	2.51	1.42	1.95	1.99	1.60	2.12	2.27	1.31	1.81	1.80	1.57	2.15	2.19	1.32	1.81	1.79	1.55	2.03	2.14	1.28	1.66	1.69
gemini-3.1-pro	1.53	1.91	2.05	1.30	1.72	1.65	1.63	2.00	2.22	1.42	1.87	1.83	1.59	1.98	2.18	1.32	1.76	1.73	1.59	1.96	2.16	1.34	1.76	1.74
gemini-3-flash	1.47	1.84	1.97	1.25	1.68	1.60	1.55	2.07	2.23	1.28	1.77	1.77	1.53	1.99	2.12	1.30	1.70	1.71	1.45	1.92	2.03	1.27	1.68	1.62
gemini-2.5-flash	1.88	2.14	2.46	1.43	1.97	1.96	1.76	2.14	2.47	1.49	1.92	1.96	1.99	2.26	2.48	**1.57**	2.00	2.03	1.88	2.16	2.49	**1.56**	1.95	2.01

Table 18: **Image Prediction (judge 1-5)** on target = Home & Kitchen, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend		image		text+img		text-rend+img	
	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)	MAE (lower)	RMSE (lower)
gpt-5	0.67	1.10	0.84	1.24	0.64	1.09	0.68	1.10
gpt-5.5	0.67	1.08	0.74	1.12	0.64	1.07	**0.65**	**1.06**
gpt-5-mini	0.70	1.12	0.82	1.25	0.65	1.11	0.68	1.11
gpt-5-nano	0.71	1.09	1.06	1.39	0.76	1.11	0.79	1.15
gpt-4o	0.74	**1.08**	0.84	1.12	0.73	1.07	0.74	1.10
gpt-4o-mini	0.87	1.11	0.99	1.26	0.88	1.14	0.90	1.16
claude-opus-4.7	0.71	1.09	0.67	**1.10**	0.64	**1.05**	0.67	1.07
claude-sonnet-4.6	0.81	1.10	0.91	1.22	0.83	1.14	0.81	1.10
claude-haiku-4.5	1.05	1.40	1.24	1.64	0.94	1.28	0.97	1.27
gemini-3.1-pro	0.70	1.18	0.69	1.23	0.64	1.12	0.71	1.21
gemini-3-flash	**0.64**	1.15	**0.67**	1.20	**0.61**	1.10	0.67	1.17
gemini-2.5-flash	0.72	1.21	0.88	1.35	0.76	1.25	0.73	1.24

Table 19: **Rating Prediction** on target = Books, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend					image					text+img					text-rend+img				
	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS
gpt-5	0.19	0.17	0.10	0.02	0.86	0.15	0.14	0.05	0.01	0.85	0.17	0.16	0.08	0.02	0.86	0.18	0.18	0.10	0.02	0.86
gpt-5.5	0.21	0.20	0.13	0.04	0.86	0.18	0.17	0.08	0.02	0.85	0.21	0.20	0.12	0.04	0.86	0.22	0.21	0.13	0.04	0.86
gpt-5-mini	0.17	0.17	0.09	0.02	0.86	0.15	0.14	0.05	0.01	0.85	0.17	0.16	0.08	0.02	0.85	0.17	0.17	0.09	0.02	0.86
gpt-5-nano	0.15	0.14	0.05	0.01	0.85	0.14	0.13	0.05	0.01	0.85	0.15	0.14	0.06	0.02	0.85	0.15	0.14	0.06	0.02	0.85
gpt-4o	0.13	0.12	0.05	0.01	0.85	0.11	0.10	0.04	0.01	0.84	0.13	0.12	0.05	0.01	0.85	0.12	0.11	0.05	0.01	0.84
gpt-4o-mini	0.15	0.14	0.06	0.01	0.85	0.12	0.11	0.04	0.01	0.84	0.15	0.14	0.06	0.02	0.85	0.14	0.13	0.05	0.01	0.84
claude-opus-4.7	**0.23**	**0.22**	**0.15**	**0.05**	**0.87**	**0.21**	**0.20**	0.10	**0.03**	0.86	**0.23**	**0.22**	**0.15**	**0.06**	**0.87**	**0.23**	**0.21**	**0.15**	**0.06**	**0.87**
claude-sonnet-4.6	0.19	0.18	0.10	0.03	0.86	0.16	0.15	0.06	0.01	0.85	0.19	0.18	0.10	0.03	0.86	0.19	0.18	0.10	0.03	0.86
claude-haiku-4.5	0.15	0.14	0.07	0.02	0.85	0.13	0.12	0.05	0.01	0.85	0.16	0.15	0.07	0.02	0.85	0.14	0.14	0.06	0.01	0.85
gemini-3.1-pro	0.20	0.19	0.12	0.04	0.86	0.20	0.19	**0.10**	0.03	**0.86**	0.21	0.20	0.13	0.04	0.86	0.20	0.19	0.12	0.04	0.87
gemini-3-flash	0.20	0.19	0.12	0.04	0.86	0.17	0.16	0.08	0.02	0.86	0.21	0.19	0.13	0.04	0.86	0.21	0.20	0.13	0.04	0.86
gemini-2.5-flash	0.18	0.17	0.09	0.03	0.86	0.16	0.15	0.05	0.01	0.85	0.18	0.17	0.09	0.02	0.86	0.18	0.17	0.09	0.02	0.86

Table 20: **Review Summarization** on target = Books, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend					image					text+img					text-rend+img				
	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS	R1	RL	B1	B4	BS
gpt-5	0.26	0.12	0.15	0.01	0.83	0.23	0.11	0.13	0.00	0.82	0.27	0.13	0.16	0.01	0.83	0.26	0.12	0.15	0.01	0.83
gpt-5.5	0.31	0.16	0.18	0.01	0.85	**0.28**	0.13	**0.17**	**0.01**	0.84	0.31	0.15	0.19	0.01	0.85	0.31	0.16	0.19	0.01	**0.85**
gpt-5-mini	0.27	0.14	0.15	0.01	0.84	0.24	0.11	0.14	0.00	0.82	0.28	0.14	0.16	0.01	0.84	0.27	0.14	0.16	0.01	0.84
gpt-5-nano	0.23	0.12	0.13	0.01	0.82	0.20	0.10	0.11	0.00	0.82	0.25	0.12	0.14	0.01	0.83	0.22	0.11	0.13	0.00	0.82
gpt-4o	0.27	0.14	0.16	0.01	0.83	0.21	0.11	0.12	0.00	0.83	0.27	0.14	0.16	0.01	0.83	0.26	0.13	0.16	0.01	0.83
gpt-4o-mini	0.27	0.13	0.16	0.01	0.84	0.23	0.11	0.14	0.00	0.83	0.27	0.13	0.16	0.01	0.83	0.27	0.13	0.16	0.01	0.83
claude-opus-4.7	**0.31**	0.16	**0.19**	0.01	0.85	0.27	**0.13**	0.15	0.01	**0.84**	**0.31**	0.15	**0.19**	**0.01**	0.85	**0.32**	0.16	**0.19**	**0.01**	0.85
claude-sonnet-4.6	0.31	0.16	0.19	**0.01**	0.84	0.26	0.12	0.15	0.01	0.83	0.30	0.15	0.18	0.01	0.84	0.32	0.16	0.19	0.01	0.85
claude-haiku-4.5	0.28	0.14	0.16	0.01	0.84	0.23	0.11	0.13	0.00	0.82	0.28	0.14	0.17	0.01	0.84	0.28	0.14	0.16	0.01	0.84
gemini-3.1-pro	0.30	**0.16**	0.17	0.01	**0.85**	0.26	0.13	0.15	0.01	0.84	0.31	**0.16**	0.18	0.01	**0.85**	0.30	**0.16**	0.17	0.01	0.85
gemini-3-flash	0.30	0.16	0.17	0.01	0.85	0.26	0.13	0.15	0.01	0.83	0.31	0.16	0.19	0.01	0.85	0.30	0.16	0.17	0.01	0.85
gemini-2.5-flash	0.28	0.15	0.15	0.01	0.84	0.25	0.12	0.14	0.01	0.83	0.29	0.15	0.17	0.01	0.84	0.29	0.15	0.16	0.01	0.84

Table 21: **Review Generation** on target = Books, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend						image						text+img						text-rend+img					
	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5	N@1	N@3	N@5	R@1	R@3	R@5
gpt-5	0.27	0.38	0.41	0.27	0.45	0.52	0.42	0.54	0.59	0.42	0.63	**0.75**	0.40	0.52	0.57	0.40	0.60	0.73	0.32	0.44	0.47	0.32	0.52	0.60
gpt-5.5	0.35	0.46	0.50	0.35	0.53	0.65	0.40	0.52	0.57	0.40	0.59	0.72	0.46	0.55	**0.60**	0.46	0.62	**0.75**	0.42	0.51	0.56	0.42	0.58	0.69
gpt-5-mini	0.32	0.41	0.47	0.32	0.47	0.61	0.32	0.45	0.49	0.32	0.54	0.65	0.37	0.46	0.52	0.37	0.53	0.68	0.33	0.44	0.50	0.33	0.52	0.66
gpt-5-nano	0.26	0.34	0.37	0.26	0.40	0.47	0.16	0.27	0.34	0.16	0.35	0.52	0.27	0.36	0.40	0.27	0.42	0.54	0.15	0.25	0.28	0.15	0.33	0.41
gpt-4o	0.21	0.28	0.31	0.21	0.33	0.41	0.18	0.35	0.40	0.18	0.46	0.59	0.19	0.35	0.41	0.19	0.47	0.61	0.18	0.25	0.29	0.18	0.30	0.40
gpt-4o-mini	0.10	0.16	0.19	0.10	0.20	0.28	0.12	0.23	0.27	0.12	0.31	0.41	0.11	0.23	0.29	0.11	0.33	0.45	0.07	0.14	0.18	0.07	0.19	0.28
claude-opus-4.7	0.27	0.39	0.43	0.27	0.48	0.60	0.29	0.40	0.44	0.29	0.48	0.59	0.28	0.38	0.41	0.28	0.45	0.54	0.21	0.29	0.33	0.21	0.35	0.44
claude-sonnet-4.6	0.26	0.35	0.40	0.26	0.43	0.54	0.34	0.46	0.51	0.34	0.54	0.66	0.37	0.48	0.53	0.37	0.56	0.69	0.33	0.45	0.50	0.33	0.53	0.66
claude-haiku-4.5	0.14	0.24	0.30	0.14	0.32	0.47	0.21	0.33	0.38	0.21	0.41	0.55	0.33	0.43	0.48	0.33	0.51	0.62	0.13	0.24	0.29	0.13	0.32	0.43
gemini-3.1-pro	**0.41**	**0.53**	**0.56**	**0.41**	**0.61**	**0.69**	**0.44**	**0.56**	**0.60**	**0.44**	**0.65**	0.74	**0.46**	**0.57**	0.60	**0.46**	**0.65**	0.72	**0.43**	**0.54**	**0.59**	**0.43**	**0.61**	**0.73**
gemini-3-flash	0.38	0.49	0.54	0.38	0.57	0.69	0.27	0.37	0.44	0.27	0.45	0.62	0.41	0.52	0.58	0.41	0.60	0.73	0.40	0.51	0.56	0.40	0.60	0.71
gemini-2.5-flash	0.30	0.41	0.47	0.30	0.50	0.64	0.23	0.34	0.41	0.23	0.42	0.60	0.32	0.44	0.50	0.32	0.53	0.67	0.32	0.44	0.48	0.32	0.52	0.63

Table 22: **Item Ranking** on target = Books, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

Model	text-rend		image		text+img		text-rend+img	
	Acc	Sim	Acc	Sim	Acc	Sim	Acc	Sim
gpt-5	0.42	0.74	0.50	0.74	0.52	0.74	0.53	0.74
gpt-5.5	0.47	0.74	0.51	0.74	**0.57**	0.74	0.54	0.74
gpt-5-mini	0.43	0.74	0.46	0.74	0.47	0.74	0.45	0.74
gpt-5-nano	0.34	0.74	0.31	0.74	0.42	0.74	0.34	0.74
gpt-4o	0.35	0.74	0.45	0.74	0.52	0.74	0.42	0.74
gpt-4o-mini	0.33	0.74	0.42	0.74	0.44	0.74	0.46	0.74
claude-opus-4.7	0.43	0.74	0.51	0.74	0.53	0.74	**0.56**	0.74
claude-sonnet-4.6	0.45	0.74	**0.52**	0.74	**0.57**	0.74	0.56	0.74
claude-haiku-4.5	0.29	0.74	0.38	0.74	0.47	0.74	0.32	0.74
gemini-3.1-pro	**0.50**	0.74	0.51	0.74	0.54	0.74	0.53	0.74
gemini-3-flash	0.40	0.74	0.41	0.74	0.40	0.74	0.43	0.74
gemini-2.5-flash	0.35	0.74	0.30	0.74	0.36	0.74	0.28	0.74

Table 23: **Item Preference Choice** on target = Books, T_4/RecentMem. All models use 300 predictions. Bold marks the highest unrounded accuracy in each format. Sim is candidate-title similarity and is not a performance score.

Model	text-rend						image						text+img						text-rend+img					
	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All	Cat	Col	Sty	Pkg	Mat	All
gpt-5	4.19	1.52	1.53	2.12	1.55	1.91	4.04	1.51	1.47	2.41	1.63	1.99	4.19	1.49	1.57	2.21	1.54	1.96	4.14	1.52	1.55	2.37	1.58	2.02
gpt-5.5	4.34	2.61	2.53	4.12	3.34	3.25	4.33	2.70	2.55	4.21	3.37	3.30	4.30	2.66	2.52	4.12	**3.31**	3.26	4.26	2.70	2.52	4.13	3.34	3.30
gpt-5-mini	4.44	1.59	1.62	2.06	1.47	1.98	4.16	1.47	1.57	2.35	1.54	2.00	4.31	1.53	1.57	2.17	1.50	1.94	4.46	1.61	1.66	2.53	1.70	2.15
gpt-5-nano	4.41	1.38	1.48	2.29	1.54	1.94	4.34	1.50	1.52	2.72	1.91	2.17	4.47	1.41	1.54	2.28	1.67	2.02	4.47	1.48	1.57	2.85	2.03	2.23
gpt-4o	4.19	2.60	2.68	3.35	2.48	2.92	4.17	2.83	2.96	3.69	2.72	3.15	4.23	2.68	2.80	3.57	2.68	3.08	4.33	2.85	2.95	3.75	2.70	3.22
gpt-4o-mini	4.19	2.56	2.54	3.26	2.50	2.89	4.09	2.51	2.50	3.37	2.52	2.87	4.14	2.44	2.35	3.31	2.45	2.82	4.22	2.58	2.54	3.37	2.57	2.93
claude-opus-4.7	4.36	2.45	2.35	3.77	2.95	3.03	4.48	2.61	2.48	3.97	3.10	3.16	4.41	2.52	2.37	3.81	3.06	3.12	4.35	2.54	2.41	3.89	3.02	3.06
claude-sonnet-4.6	4.61	**2.78**	**2.90**	**4.33**	**3.38**	**3.47**	4.46	2.63	2.67	4.25	3.31	3.34	4.53	2.68	2.64	4.20	3.24	3.34	4.47	2.66	2.73	**4.26**	**3.39**	3.40
claude-haiku-4.5	**4.66**	2.24	2.30	3.94	2.44	2.97	4.66	2.36	2.34	4.20	2.55	3.08	**4.65**	2.29	2.11	4.05	2.39	2.93	**4.71**	2.25	2.18	4.10	2.51	3.01
gemini-3.1-pro	4.15	2.31	2.31	3.00	2.23	2.64	**4.72**	**2.89**	**3.18**	**4.38**	**3.47**	**3.60**	4.23	2.52	2.51	3.40	2.59	2.94	4.37	2.59	2.65	3.80	2.87	3.12
gemini-3-flash	3.98	2.27	2.24	3.16	2.35	2.64	4.15	2.56	2.50	3.73	2.78	3.06	4.10	2.45	2.48	3.53	2.53	2.87	4.12	2.43	2.49	3.54	2.62	2.88
gemini-2.5-flash	4.42	2.75	2.85	3.91	2.98	3.27	4.27	2.86	2.87	4.20	3.07	3.36	4.55	**2.96**	**3.10**	**4.29**	3.19	**3.51**	4.47	**2.96**	**2.98**	4.19	3.10	**3.43**

Table 24: **Image Prediction (judge 1-5)** on target = Books, T_4/RecentMem. All task metrics under each format; bold = best model per (format, metric).

J TIME-AXIS FULL SWEEP: ALL TASKS ACROSS 8 CONFIGURATIONS

Complete per-configuration time-axis trajectories for every task, both targets (Books, Home & Kitchen), both backbones (`gpt-5-mini`, `gpt-5-nano`), and both formats, `text_image` (T+I) and `text_rendered_image` (Tr+I). These are the full sweep behind the curated highlights of Fig. 4 in the main text.

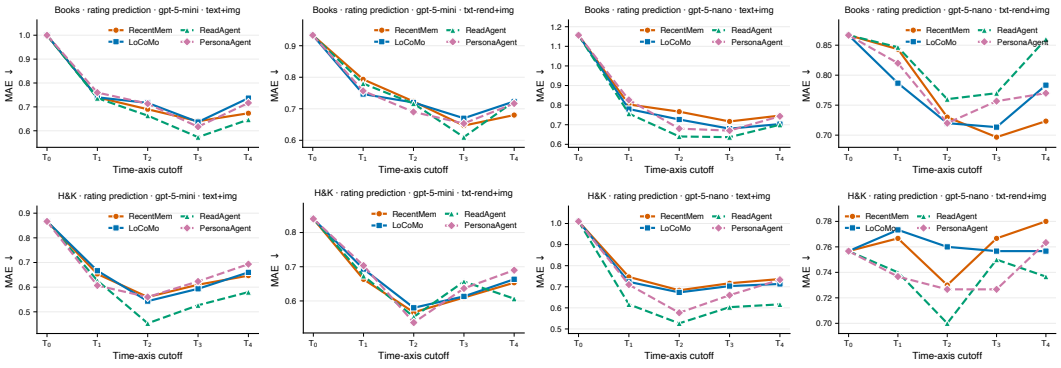


Figure 13: **Time-axis Rating Prediction (MAE, lower is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

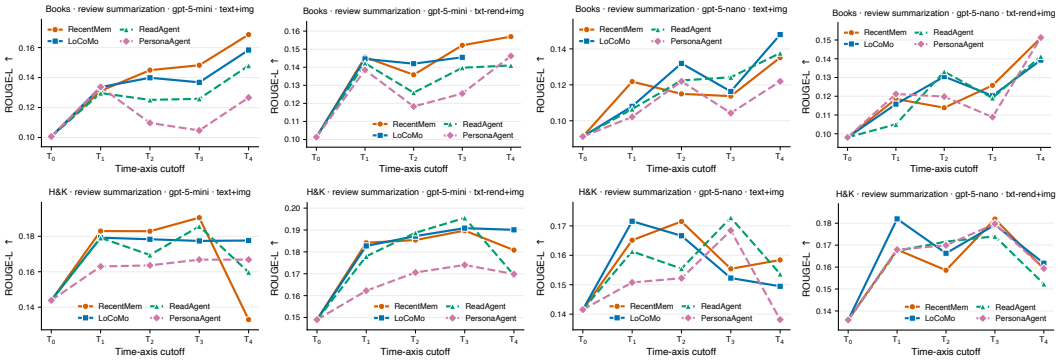


Figure 14: **Time-axis Review Summarization (ROUGE-L, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

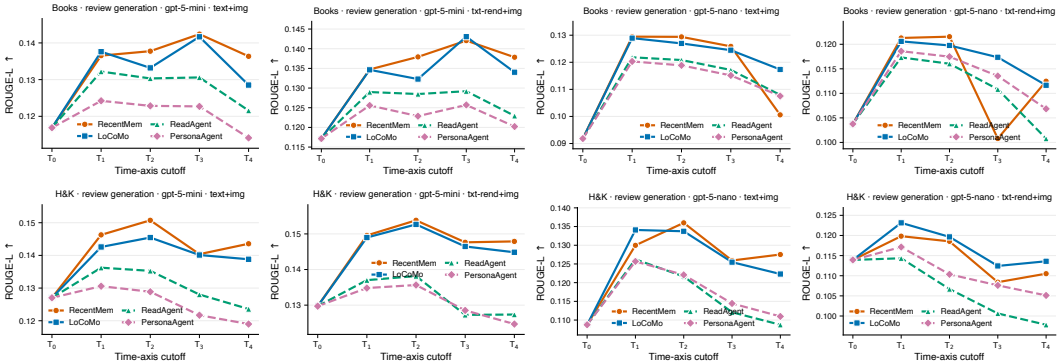


Figure 15: **Time-axis Review Generation (ROUGE-L, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

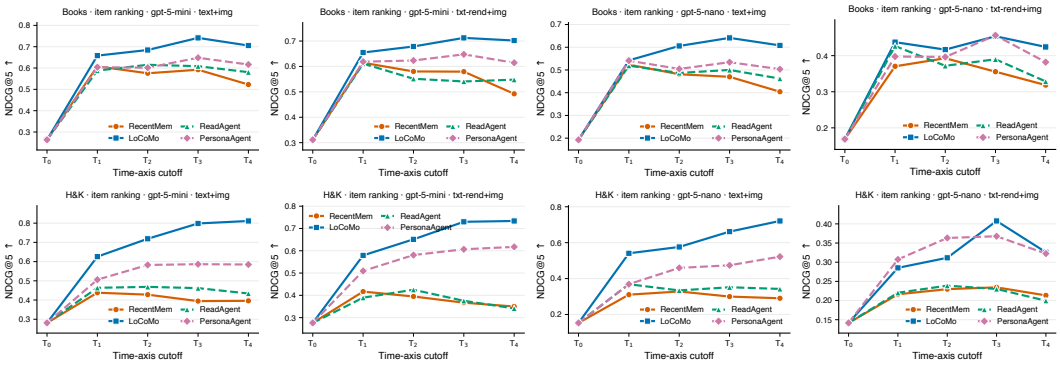


Figure 16: **Time-axis Item Ranking (NDCG@5, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

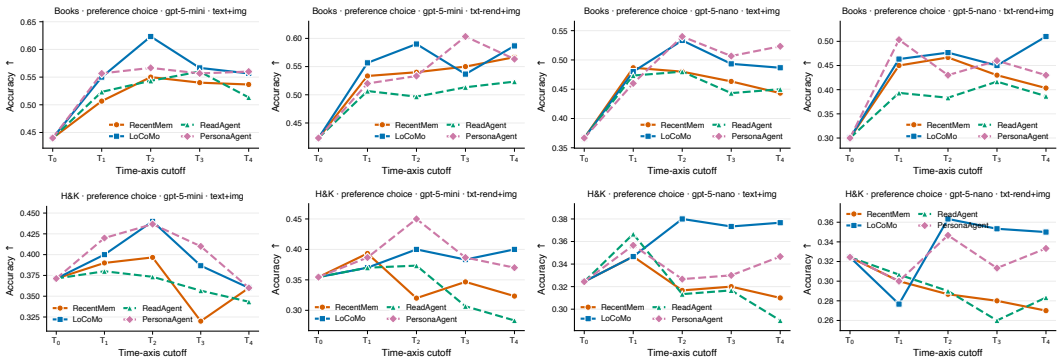


Figure 17: **Time-axis Item Preference Choice (Accuracy, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

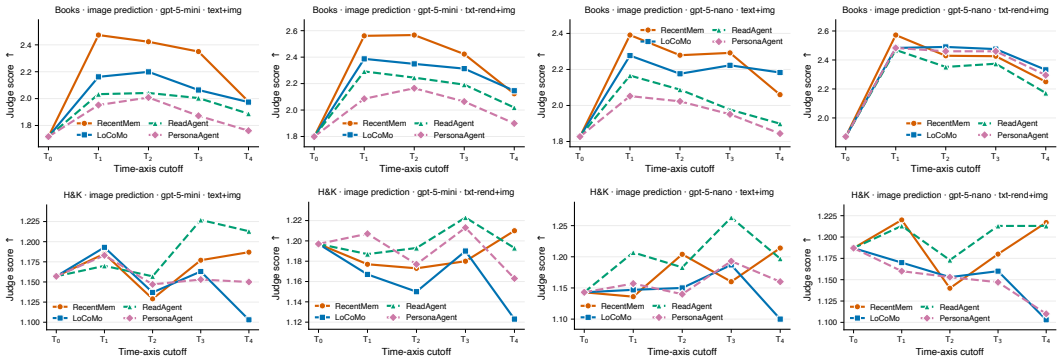


Figure 18: **Time-axis Image Prediction (judge score 1 to 5, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over T_0 to T_4, all four update methods.

K Domain-Axis Full Sweep: All Tasks Across 8 Configurations

Complete per-configuration domain-axis trajectories for every task, both targets (Books and Home & Kitchen), both backbones (`gpt-5-mini` and `gpt-5-nano`), and both formats, `text_image` (T+I) and `text_rendered_image` (Tr+I). These are the full sweep behind the curated highlights of Fig. 5 in the main text. All panels compare D_1 to D_4 on the final target items. The earlier plots reused a time-axis empty-memory result with different test items; that unmatched reference is excluded here.

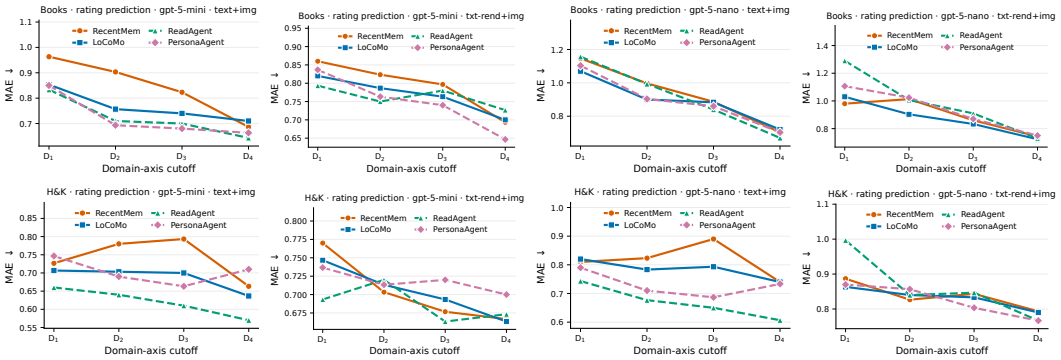


Figure 19: **Domain-axis Rating Prediction (MAE, lower is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

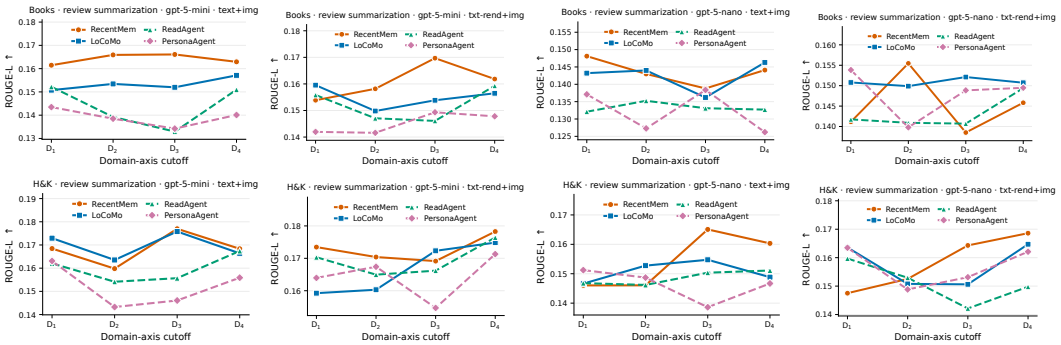


Figure 20: **Domain-axis Review Summarization (ROUGE-L, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

L Extended Domain-Axis: D_5 and D_{tx} Sweeps

This appendix reports two extensions to the D-axis experiment (Sec. 3.3). D_5 adds the eight non-primary domains to the four-domain D_4 memory. D_{tx} uses only those eight non-primary domains and excludes all four primary domains, including the target. The added domains are All Beauty, Arts & Crafts & Sewing, Automotive, Grocery & Gourmet Food, Health & Household, Movies & TV, Office Products, and Sports & Outdoors.

The grid uses `gpt-5-nano`, RecentMem and LoCoMo, `text_image` and `text_rendered_image`, and all six tasks. D_5 is evaluated on Books and Home & Kitchen; D_{tx} is evaluated on all four primary targets. This gives 144 settings in total. The tables below report `text_image` results, with absolute scores for D_{tx}.

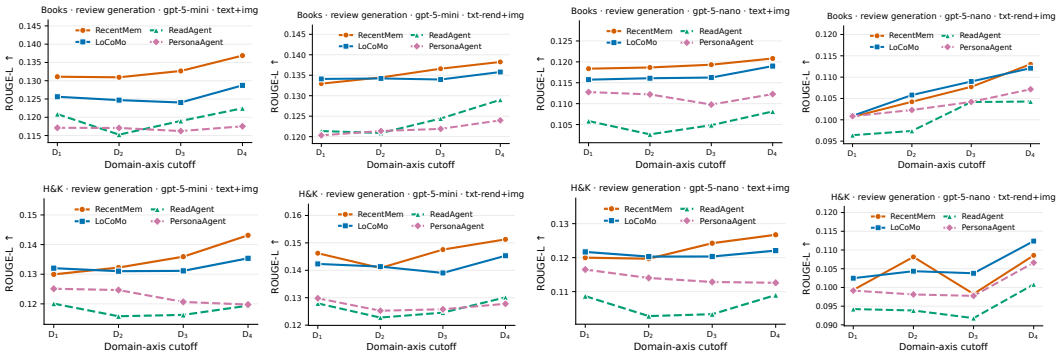


Figure 21: **Domain-axis Review Generation (ROUGE-L, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

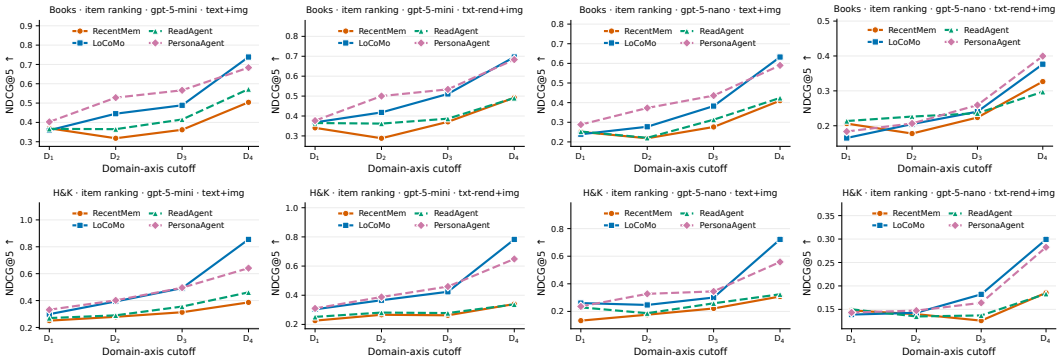


Figure 22: **Domain-axis Item Ranking (NDCG@5, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

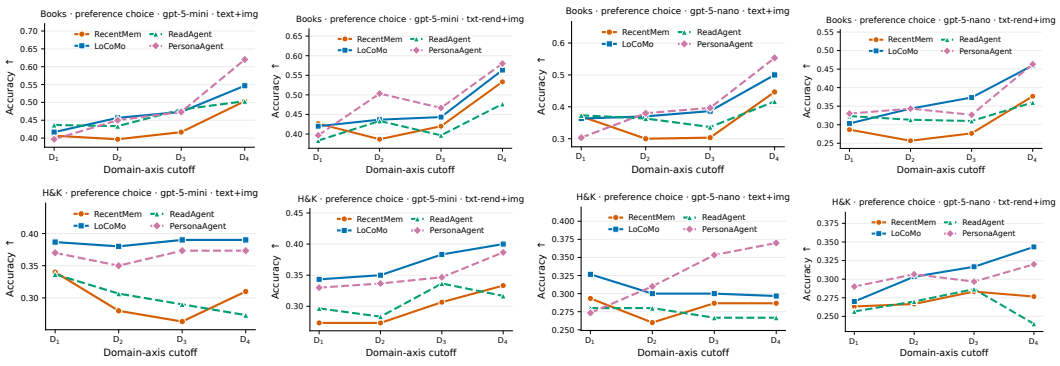


Figure 23: **Domain-axis Item Preference Choice (Accuracy, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

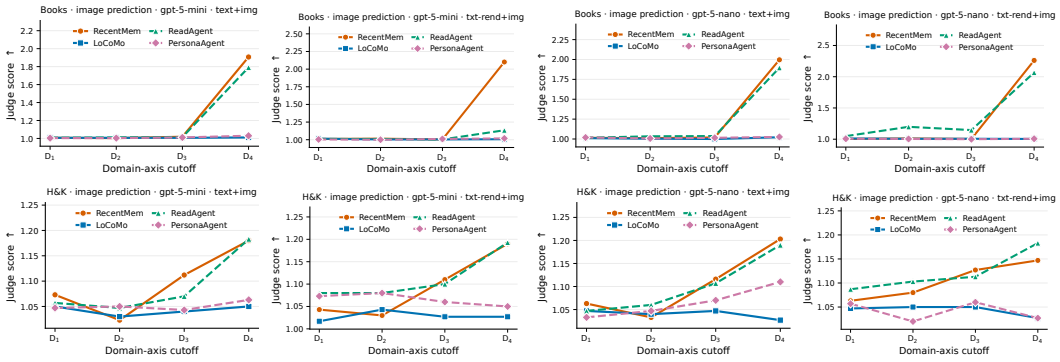


Figure 24: **Domain-axis Image Prediction (judge score 1 to 5, higher is better)** across all 8 configurations. Top row: Books target; bottom row: Home & Kitchen target. Columns (left to right): mini·T+I, mini·Tr+I, nano·T+I, nano·Tr+I. Trajectory over D_1 to D_4, all four update methods.

L.1 D_5 Full 12-Domain Memory

Table 25 compares D_4 and D_5 on the same target tests. Adding eight domains does not consistently improve performance. For example, Books RecentMem image prediction falls from 1.996 to 1.011; its ranking score falls from 0.410 to 0.243. LoCoMo gains on several Home & Kitchen tasks, including ranking (0.722 to 0.758), but loses on Books. These results show that more source domains can hurt under this domain order and memory budget. They do not establish a general capacity limit at four domains.

Table 25: **D_5 versus D_4 memory** (`text_image`, `gpt-5-nano`). Each cell reports D_4 followed by D_5 on the same target test. Bold marks a better D_5 point estimate. Rating is MAE (lower is better); all other metrics are higher-is-better.

Target	Method	Rating MAE↓	Summ R-L	Gen R-L	Rank NDCG@5	Choice Acc	ImgPred 1 to 5
Books	RecentMem	0.703→0.997	0.144→0.137	0.121→0.119	0.410→0.243	0.447→0.347	1.996→1.011
	LoCoMo	0.720→0.727	0.146→0.132	0.119→0.117	0.632→0.604	0.500→0.417	1.022→1.004
H&K	RecentMem	0.740→0.870	0.160→0.160	0.127→0.122	0.308→0.198	0.287→0.253	1.203→1.060
	LoCoMo	0.740→0.747	0.149→**0.163**	0.122→**0.124**	0.722→**0.758**	0.297→0.253	1.027→**1.057**

L.2 D_{tx} Pure Cross-Domain Memory

Table 26 reports performance when memory contains only the user's eight non-primary domains. LoCoMo ranks better than RecentMem on all four targets in this setting, while image-prediction scores remain close to the lower end of the scale. This open-ended task provides little target guidance, so poor scores can reflect both category errors and errors in visual detail. The results do not isolate these two sources of error.

M Attributing the Time-Axis Trajectory

In the time-axis protocol, both memory and the prediction target advance through the user's timeline (Sec. 2, Appendix J). Changes in performance can therefore reflect memory growth, changes in user behaviour, or changes in target difficulty. We add a fixed-target control on Home & Kitchen with `gpt-5-mini`, `text_image`, 100 users, and two methods: RecentMem and LoCoMo. This control evaluates memory growth on the same final test items. It does not separately identify behavioural drift and target difficulty.

Fixed target, growing memory. We hold the test set at the last three interactions per user and increase the available pre-test history from zero to its full length. The moving-target and fixed-target

Table 26: **D_{tx} scores on all four targets** (`text_image`, `gpt-5-nano`). Memory contains only the user's eight non-primary domains. The table reports absolute scores. Rating is MAE (lower is better); all other metrics are higher-is-better.

Target	Method	Rating MAE↓	Summ R-L	Gen R-L	Rank NDCG@5	Choice Acc	ImgPred 1 to 5
Books	RecentMem	0.983	0.137	0.120	0.245	0.333	1.004
	LoCoMo	0.820	0.137	0.118	0.398	0.383	1.026
H&K	RecentMem	0.817	0.156	0.122	0.198	0.307	1.073
	LoCoMo	0.713	0.145	0.123	0.455	0.260	1.050
Personal Care	RecentMem	0.817	0.147	0.125	0.212	0.247	1.030
	LoCoMo	0.787	0.146	0.119	0.427	0.277	1.090
Elec.	RecentMem	0.880	0.153	0.123	0.185	0.227	1.073
	LoCoMo	0.863	0.149	0.121	0.393	0.280	1.047

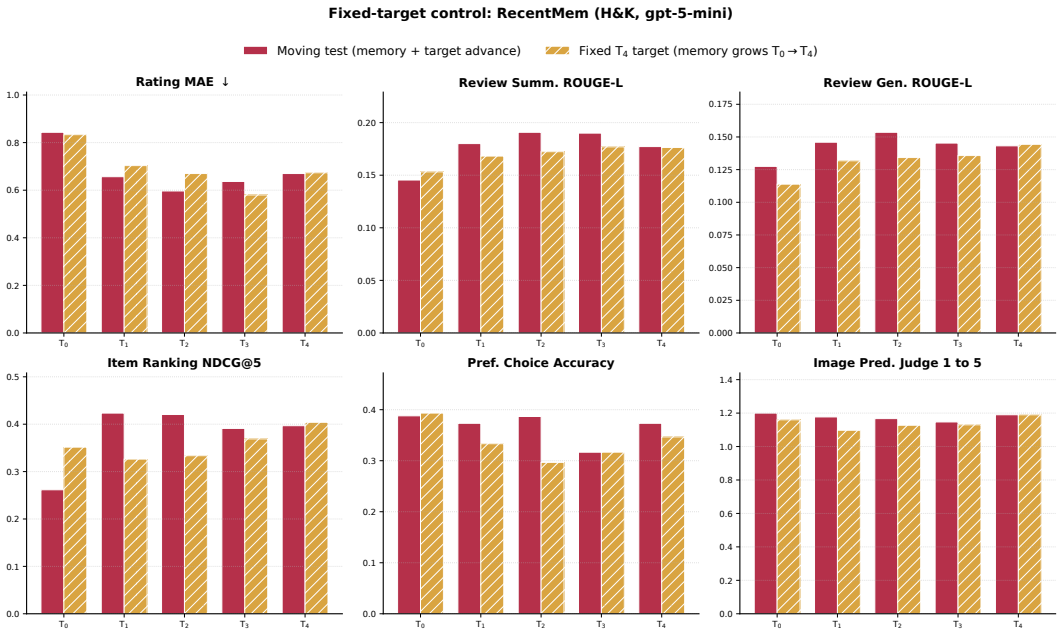


Figure 25: **Fixed-target control for RecentMem.** Moving-target evaluation advances both memory and the target, while the control keeps the final test items fixed. Fixed-target ranking changes from 0.352 to 0.404, compared with 0.262 to 0.397 in the moving-target run. Memory growth thus gives a smaller endpoint gain under the fixed target. This comparison does not isolate the cause of every intermediate fluctuation. Rating is MAE, where lower is better.

protocols have the same history and test configuration at the final stage, but they were evaluated in separate runs. Their measured endpoint scores can therefore differ. The figures report the stored results of these control runs, which are also separate from the main time-axis sweep.

Findings. LoCoMo's fixed-target ranking gain is 0.431 NDCG@5, compared with 0.496 in its moving-target run. Its fixed-target rating MAE also improves from 0.850 to 0.657. These results show that memory growth can improve prediction even when the target is unchanged. RecentMem has a smaller fixed-target ranking gain of 0.052, compared with 0.135 in its moving-target run. The contrast is consistent with the value of retrieving from a growing history, but the two protocols do not provide an additive decomposition of memory effects and target drift.

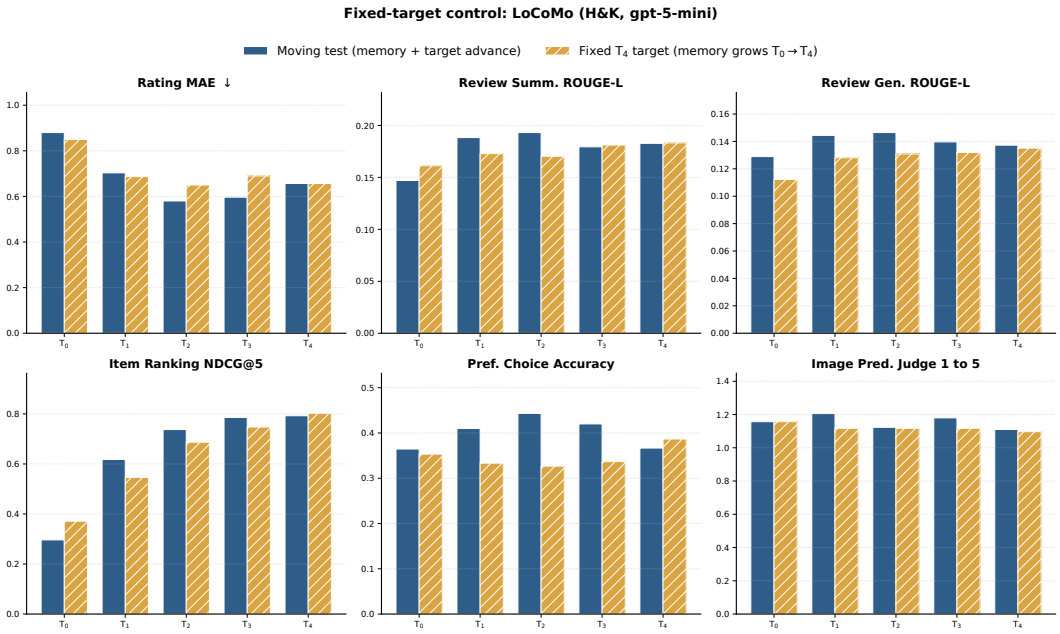


Figure 26: **Fixed-target control for LoCoMo.** Ranking increases from 0.371 to 0.802 when the final target is held fixed, and from 0.296 to 0.793 in the moving-target run. The fixed-target increase supports a substantial contribution from a larger retrievable history. Endpoint configurations match, but the scores come from separate runs. Rating is MAE, where lower is better.