FROM PREFERENCE TO ACTION: A BENCHMARK FOR CONTINUAL PREFERENCE MEMORY IN MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal agents must learn user preferences from feedback, retain them over time, and apply them in future interactions. However, existing benchmarks typically evaluate preference understanding and personalized generation in isolation, leaving unclear whether remembered preferences actually guide subsequent behavior. We introduce **PREF2ACT**, a benchmark for *continual preference memory* that jointly evaluates preference understanding and personalized image generation and editing as interaction histories accumulate. PREF2ACT comprises six constructed personas with fine-grained preference profiles and corresponding long-horizon multimodal histories. These histories interleave dialogue, image generation, image editing, and user feedback, through which preferences are progressively revealed, reinforced, and revised. For each persona, the benchmark provides 488 cumulative memory questions assessing understanding and application across prior interactions, together with a fixed set of 189 anchor questions repeated at every checkpoint to track preference acquisition and retention. We evaluate five agent configurations with twelve memory methods and find that additional interaction history does not consistently improve preference identification. Moreover, in our failure analysis, 8.36%–39.74% of failed generation tasks are accompanied by correct identification of the corresponding scenario-relevant preferences. These findings highlight a persistent knowing–doing gap, motivating joint evaluation of preference understanding and visual execution.

1 INTRODUCTION

Multimodal agents are increasingly expected to serve as persistent creative collaborators (Wu et al., 2025b; Jiang et al., 2025; Nabati et al., 2025). A user who repeatedly rejects skin smoothing and excessive brightening expects the agent to respect these preferences in later portrait edits, even when the subsequent instruction is simply *"Retouch this photo."* Long-term personalization therefore requires agents to infer stable preferences from feedback, maintain them across interactions, and apply them in relevant contexts.

We study continual preference memory: the ability to infer, retain, and update user preferences from an evolving multimodal interaction history and apply to subsequent visual tasks. Figure 1 illustrates three central challenges. First, preferences are often implicit and distributed across interactions rather than stated as complete profiles. Interpreting feedback such as *"Too artificial. Keep it natural."* requires grounding the language in the preceding visual changes, while repeated feedback helps distinguish a stable preference from a one-off correction. Second, preferences are scenario-dependent rather than globally uniform: the same user may prefer noticeable retouching for adult portraits but minimal facial alteration in images of children. Third, correct preference identification does not guarantee consistent application. An agent may correctly identify that the user wants children's facial features preserved yet still alter them in a subsequent image generation or editing task. Evaluating continual preference memory must therefore assess not only whether agents understand user preferences, but also whether those preferences consistently shape subsequent visual outputs.

Existing work studies several related aspects of this problem. Benchmarks for unified multimodal understanding and generation evaluate joint image understanding, generation, and editing (Xie et al.,

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.

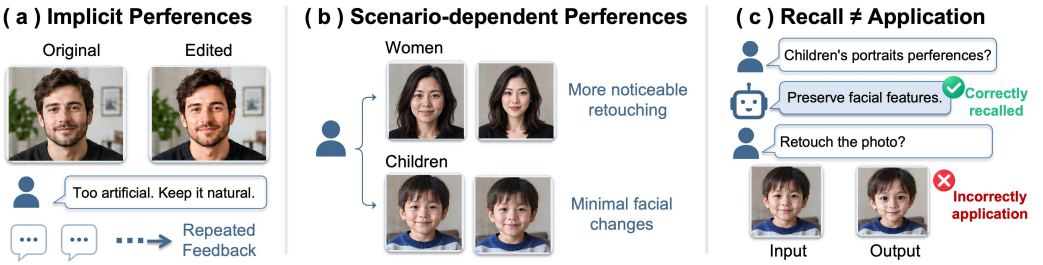


Figure 1. Three challenges of continual preference memory evaluated by PREF2ACT.

2026; Chow et al., 2026). Personalized image generation benchmarks evaluate adaptation based on prior user behavior or iterative feedback (Shen et al., 2024; Nabati et al., 2025). Agent memory benchmarks incorporating user feedback examine preference inference from indirect conversational cues and preference changes over time (Jiang et al., 2025; Liu et al., 2026). Multimodal memory benchmarks focus on retrieval and reasoning over textual and visual evidence from previous sessions (Bei et al., 2026). However, these evaluations do not jointly assess whether agents can progressively infer and maintain preferences from accumulated multimodal feedback and reliably translate that understanding into subsequent image generation and editing. Consequently, the knowing–doing gap in long-term multimodal personalization remains largely unmeasured.

To address this gap, we introduce PREF2ACT, a benchmark that jointly evaluates continual preference memory through preference understanding and personalized image generation and editing. The benchmark constructs long-horizon multimodal histories from hidden user profiles covering global visual preferences, scenario-specific preferences, personal interests, and interaction preferences. These histories interleave dialogue, image generation and editing, and user feedback that progressively reveals, reinforces, and revises preferences. Evaluation combines cumulative memory questions and repeated anchor questions. At each checkpoint, cumulative questions assess explicit fact retrieval, multi-turn information integration, implicit preference inference, and personalized generation and editing over all prior interactions. Repeated anchor questions track whether agents' understanding of user preferences improves as interactions accumulate.

We evaluate five representative multimodal agents across twelve memory configurations covering both textual and multimodal memory. Our results reveal three limitations. First, agents struggle to infer preferences implicitly expressed through user feedback. Second, additional interactions do not reliably translate into a more accurate understanding of user preferences. Third, even correctly identified preferences are not consistently applied to subsequent image generation and editing. This failure to translate preference understanding into visual behavior exposes a persistent knowing–doing gap in current multimodal agents.

2 RELATED WORK

We position PREF2ACT at the intersection of unified multimodal modeling, personalized image generation, and long-term agent memory. Table 1 compares representative benchmarks in terms of task coverage, personalization, multimodality and interaction setting.

Unified Multimodal Understanding and Generation. MME-Unify (Xie et al., 2026) and Uni-MMMU (Zou et al., 2026) evaluate unified multimodal models across visual understanding, image generation, editing, and tasks that explicitly couple understanding with synthesis. WEAVE (Chow et al., 2026) further extends this setting to interleaved, multi-turn comprehension, generation, and editing that require models to use prior context. These benchmarks focus on multimodal task performance and consistency across several turns, rather than on how models infer, retain, and apply user preferences over long interaction histories.

Personalized Image Generation. PIP (Chen et al., 2024) personalizes text-to-image generation by rewriting current prompts using a fixed history of user-generated image–prompt pairs. PIG-Bench (Lee et al., 2025) evaluates personalized reward models using user-specific reference data and preference-based image rankings. PASTA (Nabati et al., 2025) supports personalized, multi-turn generation through iterative user choices. These approaches guide visual generation and eval-

Table 1. Comparison of representative benchmarks. Und., Img., Pers., and MM. denote understanding, image generation and editing, personalization, and multimodality, respectively. Feedback describes whether user feedback is absent, fixed, or iterative, while History indicates the interaction scope. ✓ and ✗ indicate the presence and absence of the corresponding property, respectively.

Benchmark	Venue	Task Coverage		Pers.	MM.	Interaction Setting	
		Und.	Img.			Feedback	History
I. Unified Understanding and Generation							
MME-Unify (Xie et al., 2026)	ICLR'26	✓	✓	✗	✓	✗	✗
Uni-MMMU (Zou et al., 2026)	ACL'26	✓	✓	✗	✓	✗	Multi-turn
WEAVE (Chow et al., 2026)	CVPR'26	✓	✓	✗	✓	✗	Multi-turn
II. Personalized Image Generation							
PIP (Chen et al., 2024)	CVPR'24	✗	✓	✓	✓	✗	Fixed
PIGBench (Lee et al., 2025)	arXiv'25	✗	✗	✓	✓	Fixed	Fixed
PASTA (Nabati et al., 2025)	ICML'25	✗	✓	✓	✓	Iterative	Multi-turn
III. Long-Term Agent Memory							
PersonaMem-v2 (Jiang et al., 2025)	arXiv'25	✓	✗	✓	✗	✗	Long-term
PERMA (Liu et al., 2026)	arXiv'26	✓	✗	✓	✗	Iterative	Long-term
MemoryBench (Ai et al., 2026)	ICML'26	✓	✗	✗	✗	Iterative	Long-term
Mem-Gallery (Bei et al., 2026)	ACL'26	✓	✗	✓	✓	✗	Long-term
M^3Exam (Huang et al., 2026)	arXiv'26	✓	✗	✗	✓	✗	Long-term
PREF2ACT	**Ours**	✓	✓	✓	✓	**Iterative**	**Long-term**

uation using prompts, reference images, rankings, or explicit choices. However, they do not fully capture interactions where users reveal preferences gradually and often implicitly through feedback on agent outputs.

Long-Term Agent Memory. LongMemEval (Wu et al., 2025b) evaluates information retrieval, temporal reasoning, and knowledge updating across long conversational histories. PersonaMem-v2 (Jiang et al., 2025) and PERMA (Liu et al., 2026) assess implicit preference inference and preference evolution over extended interactions between users and agents, while MemoryBench (Ai et al., 2026) examines task adaptation from accumulated user feedback. Mem-Gallery (Bei et al., 2026) and M^3Exam (Huang et al., 2026) extend long-context memory evaluation to multimodal histories. Collectively, these benchmarks assess understanding and reasoning over long histories but lack generation or editing tasks to test preference application. In contrast, PREF2ACT jointly evaluates preference understanding and image generation over the same interaction histories, while repeated anchor questions track how preference understanding evolves as evidence accumulates.

3 PREF2ACT BENCHMARK

Figure 2 presents an overview of the PREF2ACT benchmark pipeline. To evaluate continual preference memory in multimodal agents, we first describe the construction of long-horizon multimodal interaction histories in Section 3.1, then detail the evaluation data construction in Section 3.2, and finally report benchmark statistics in Section 3.3 and present the evaluation metric in Section 3.4.

3.1 LONG-HORIZON MULTIMODAL INTERACTION CONSTRUCTION

PREF2ACT constructs long-horizon multimodal histories in which user preferences are progressively revealed through interactions rather than directly provided as static profiles. Each history contains multiple interaction scenarios and preserves both textual exchanges and visual outputs, providing the evidence required to study how agents identify and apply user preferences over time.

User Profiles and Preference Space. We first construct a fine-grained profile for each user to guide the generation of internally consistent interactions. The preference space comprises four levels. *Global visual preferences* characterize six dimensions of a user's general aesthetic tendencies: rendering style, style orientation, color system, atmosphere, composition and realism Level. *Scenario-specific preferences* are hierarchically organized into seven broad domains spanning portraiture and human relationships, everyday living, dining and retail, urban mobility and public spaces, natural environments, leisure and entertainment, fantasy and stylized worlds, and image editing. Each domain is further divided into fine-grained scenarios, each associated with a distinct set

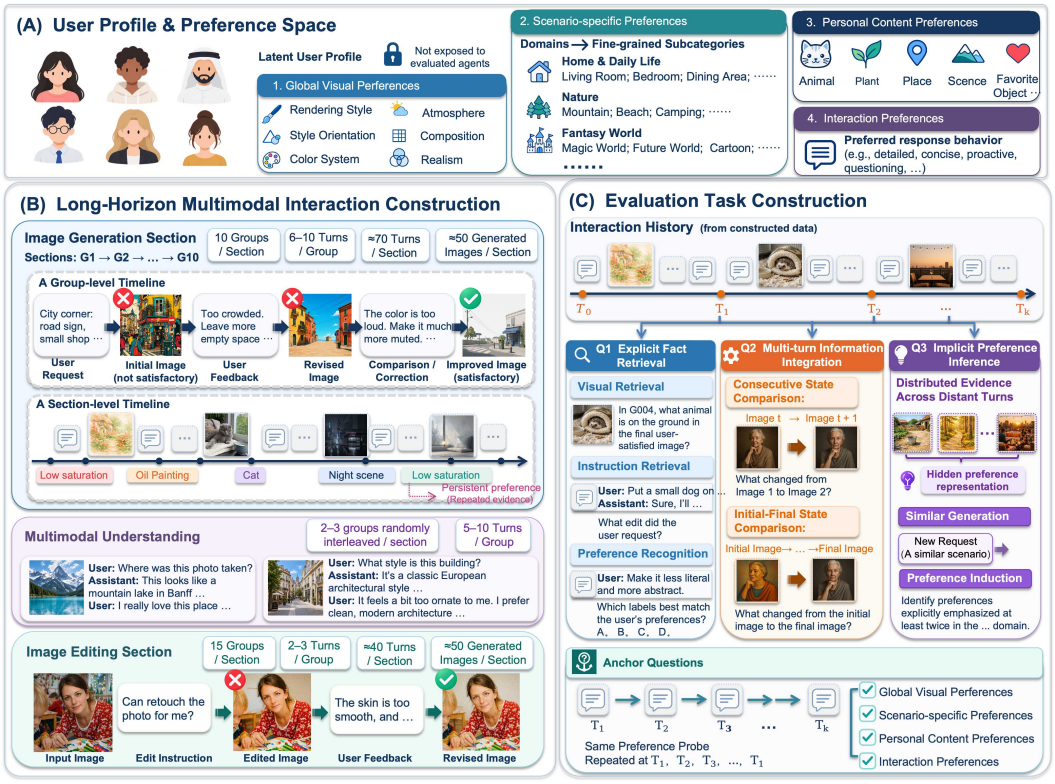


Figure 2. The overall pipeline of our proposed PREF2ACT.

of preferences. *Personal interests* specify favored objects, animals, plants, places, and scenes, while *interaction preferences* characterize how the user expects the agent to respond. This hierarchical design allows the same user to exhibit stable global tendencies while retaining distinct preferences across scenarios. Importantly, the underlying profiles are used only to guide data construction and are never directly exposed to the evaluated agents. Appendix A details the preference space.

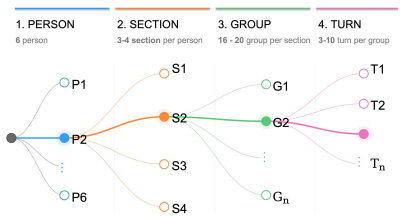
Interaction Scenarios and Trajectory Construction. For each user, we construct diverse interaction trajectories spanning visual understanding, image generation, and image editing across multiple scenarios, such as interior design, portrait retouching, and everyday photo enhancement. Each trajectory begins with a task request, followed by an image generated or edited by the assistant. The user then responds to the visual result through natural corrections, comparisons, confirmations, rejections, or preference updates. Selected outputs intentionally deviate from relevant profile attributes, creating opportunities for users to reveal their preferences through feedback.

To distinguish persistent preferences from one-off instructions, preference evidence is distributed across nonadjacent turns and multiple scenarios. Some preferences are later revised or replaced, requiring agents to update rather than merely accumulate user information. We further insert several turns on unrelated topics at random positions to introduce contextual distractions and test whether agents can retrieve relevant preference evidence without relying on recency or topical continuity.

3.2 EVALUATION DATA CONSTRUCTION

Our evaluation comprises cumulative memory questions and repeated anchor questions. At each checkpoint, cumulative memory questions assess preference understanding and application using all prior interactions. Anchor questions remain fixed across checkpoints, enabling systematic evaluation of how agents' understanding of user preferences evolves as evidence accumulates. The evaluation includes single-choice, multi-select, short-answer, and image generation tasks; each question is annotated with explicit scoring points and supporting evidence from the interaction history.

Cumulative Memory Questions. At each checkpoint, these questions evaluate the complete interaction history available up to that point through the following three categories.



Statistic	Number
• **Interaction Data**	
- User–Agent Turns	$\approx 83 \times 4$
- Total Words	$\approx 1{,}917 \times 4$
- Total Images	$\approx 48 \times 4$
• **Evaluation Data**	
- Cum. Memory Qs.	$\approx 122 \times 4$
- Repeated Anchor Qs.	189



Figure 3. Overview of the PREF2ACT benchmark. **Left:** hierarchical organization of multimodal interaction data. **Middle:** dataset statistics. **Right:** distribution of evaluation data across categories.

Q1: Explicit Fact Retrieval. This category targets visual details, user instructions, and explicitly expressed global, scenario-specific, and personal preferences. It evaluates whether agents can accurately recover information directly stated or visually presented in prior interactions.

Q2: Multi-turn Information Integration. This category examines changes between consecutive visual outputs and between the initial and final states of an interaction trajectory. It evaluates agents' ability to integrate textual feedback and visual observations across turns, tracking individual modifications and their cumulative effects.

Q3: Implicit Preference Inference. This category combines cumulative global and scenario-specific preference induction with image generation in similar scenarios. It evaluates whether agents can infer underlying preferences from distributed multimodal evidence and translate them into personalized outputs without those preferences being restated in the current request.

Repeated Anchor Questions. A fixed set of questions probing global, scenario-specific, personal interests, and interaction preferences is administered at checkpoints. Keeping the probes fixed allows us to examine how preference identification changes as the available history grows, including the acquisition of newly supported preferences and the retention of previously revealed ones.

3.3 DATASET STATISTICS

Figure 3 summarizes the dataset structure and evaluation categories. PREF2ACT comprises six personas, each associated with three to four sections spanning different scenarios, including one section dedicated to image editing. Each section contains 16-20 task groups, with 3-10 turns per group. Visual understanding dialogues are interleaved with generation and editing interactions. A four-section history contains approximately 83×4 turns, $1{,}917 \times 4$ words, and 48×4 images. Each persona is associated with approximately 488 cumulative memory questions across Q1, Q2, and Q3, administered at their designated checkpoints. In addition, a fixed set of approximately 189 anchor questions is repeated after every section to track preference identification over time.

3.4 EVALUATION METRIC.

We use deterministic metrics for choice questions and LLM judges for short answers and generated images. All scores range from 0 to 100.

Single-choice Questions. We evaluate these questions using accuracy. A response receives 100 points if it contains exactly one valid option label matching the reference answer, and 0 otherwise.

multi-select Questions. For each question i, we compute F1 between the predicted option set $\hat{Y}_i$ and the reference set Y_i. The score is $S_i^{\text{MC}} = 100 \cdot \frac{2|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i| + |Y_i|}$, accounting for both missed and incorrect selections. If both sets are empty, the score is 100.

Short Answer Questions.

Following existing works (Liu et al., 2023; Zheng et al., 2023), we adopt an *LLM-as-a-Judge* protocol for short-answer evaluation. For each question i, the judge receives the question, model response,

and K_i equally weighted scoring points. Each point receives full credit if all required information is explicitly stated, and zero otherwise. Semantic equivalents and concise paraphrases are accepted, while irrelevant additional content is ignored. The judge returns the identifiers of the satisfied points, from which we compute $S_i^{\mathrm{SA}} = 100 \cdot m_i / K_i$, where m_i denotes the number of satisfied scoring points.

Image Generation Tasks. We also adopt a *LLM-as-a-Judge* protocol for image generation evaluation. The judge assesses each generated image against the current request and applicable reference preferences, covering global visual preferences, scenario-specific preferences, and personal content preferences. Each image is evaluated using K equally weighted criteria. A criterion receives $100/K$ points if it is visibly satisfied and zero otherwise. The final score is the sum of these contributions.

4 EXPERIMENTS

In this section, we conduct experiments to address the following research questions:

RQ1: How do multimodal agents perform across preference-memory tasks?

RQ2: Does preference understanding improve with more interactions?

RQ3: Do correctly inferred preferences translate into preference-consistent image generation?

RQ4: How reliable is PREF2ACT for evaluating continual preference memory?

4.1 EVALUATION SETUP

Models. We evaluate five agent configurations. The two unified models are BAGEL-7B-MoT (Deng et al., 2025) and Janus-Pro-7B (Chen et al., 2025). The three modular model pairs combine Gemini 3.6 Flash (Google DeepMind, 2026) with Gemini 3 Pro Image (Raisinghani, 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025) with Qwen-Image (Wu et al., 2025a), and GPT-5.4 (OpenAI, 2026a) with GPT Image 2 (OpenAI, 2026b).

Memory Configurations. We compare seven textual memory configurations: Full (Text), NaiveRAG, Gen. Agent (Park et al., 2023), Reflexion (Shinn et al., 2023), A-Mem (Xu et al., 2025), MemoryOS (Kang et al., 2025), and Mem0 (Chhikara et al., 2025). NaiveRAG serves as a basic retrieval baseline within the retrieval augmented generation paradigm (Lewis et al., 2020). Five multimodal configurations include Full (MM), MuRAG (Chen et al., 2022), UniversalRAG (Yeo et al., 2026), NGM (Fisher, 2025), and AUGUSTUS (Jain et al., 2025). Full (Text) and Full (MM) are full-history reference settings for their respective modalities. We evaluate each agent configuration with the memory configurations to study how model and memory choices affect performance. Further details of the agent and memory configurations are provided in Appendix B.

Evaluation Protocol. Each model and memory combination processes each persona's history in order. Memory is maintained across sections and reset between personas. We evaluate cumulative questions at their assigned checkpoints and repeat anchor questions after each section. Results are combined within each persona and averaged equally across the six personas. For anchor questions, we report the mean score across checkpoints and the final score minus the initial score. Implementation details are provided in Appendix C.

4.2 OVERALL PERFORMANCE ON CONTINUAL PREFERENCE MEMORY (RQ1)

We evaluate five representative multimodal agents under twelve textual and multimodal memory configurations. Table 2 reports their overall performance on cumulative memory questions and repeated anchor questions. Table 3 further presents a fine-grained analysis using Qwen3-VL-8B. Full results are reported in Appendix D. We draw the following observations:

- **Obs. 1: Existing memory methods do not solve continual preference memory.** Overall, agents perform poorly across memory configurations, averaging 44.33, 15.20, and 27.12 on Q1, Q2, and Q3-Inf, respectively. The fine-grained results show strong performance on some Explicit Fact Retrieval tasks with directly observable preference cues, but performance drops substantially when multi-turn integration or implicit preference inference is required. This suggests that current models struggle to aggregate user feedback across turns and extract stable preference representations.

Table 2. Performance comparison of backbone agent models and memory methods on PREF2ACT. Blue shading indicates the best agent for each task under a fixed memory method, while underlining indicates the best memory method for each task within a fixed agent. Anchor Avg. reports the mean score across checkpoints.

		Textual Memory							Multimodal Memory				
Model	**Task**	Full (Text)	Naive RAG	Gen. Agent	Reflexion	A-Mem	MemoryOS	Mem0	Full (MM)	MuRAG	Universal RAG	NGM	AUGUSTUS
Janus-Pro-7B	Q1	20.90	44.92	37.81	22.85	42.18	52.56	54.28	31.96	52.03	49.00	47.66	44.24
	Q2	15.39	17.66	11.85	11.63	6.01	9.27	26.48	19.17	25.95	12.68	20.71	13.66
	Q3-Inf.	22.22	12.80	14.66	16.88	17.95	8.78	32.92	14.26	24.04	28.19	21.89	23.20
	Q3-Gen.	48.62	54.37	51.28	49.43	55.61	47.16	57.24	59.18	64.73	63.92	58.27	59.46
	Anchor Avg.	26.05	49.25	31.40	26.32	41.62	46.14	54.17	33.78	44.71	43.22	39.07	38.12
BAGEL-7B-MoT	Q1	23.69	52.81	42.99	37.94	56.68	61.72	57.81	36.56	61.29	58.12	53.53	54.92
	Q2	25.43	32.65	23.64	22.15	16.41	15.62	31.73	28.79	31.43	21.30	29.70	29.25
	Q3-Inf.	33.79	35.23	31.41	21.97	27.03	13.40	27.30	27.77	45.30	47.56	34.87	18.13
	Q3-Gen.	63.86	68.08	65.62	62.73	69.29	61.37	69.71	71.43	78.16	81.02	73.61	76.84
	Anchor Avg.	29.10	52.65	44.49	42.49	50.60	55.02	59.05	41.22	53.47	56.31	54.17	53.92
Qwen3-VL-8B + Qwen-Image	Q1	27.61	48.11	36.20	23.84	52.16	49.36	54.03	24.63	59.06	54.69	51.83	50.88
	Q2	14.44	15.24	13.30	10.84	13.40	3.53	17.82	9.85	21.06	10.94	10.20	11.59
	Q3-Inf.	34.13	40.59	33.90	27.94	28.47	25.62	33.50	22.89	41.87	42.71	29.68	31.43
	Q3-Gen.	65.71	67.83	64.92	63.46	67.58	61.93	68.84	71.62	83.07	81.14	74.28	72.81
	Anchor Avg.	26.59	39.41	29.44	22.70	46.27	43.71	56.83	24.19	46.26	47.57	39.81	38.67
Gemini-3.6-Flash + Gemini 3 Pro Image	Q1	23.12	48.81	35.32	18.21	52.15	51.81	41.47	19.69	53.15	58.96	52.24	50.84
	Q2	12.74	14.79	19.35	11.13	14.01	8.18	6.72	7.37	7.78	11.79	9.28	11.61
	Q3-Inf.	21.64	25.40	20.89	24.17	12.12	8.64	20.39	16.45	23.08	25.52	18.01	19.48
	Q3-Gen.	73.29	71.84	70.16	68.43	74.31	68.92	73.86	77.63	83.94	85.57	80.13	80.68
	Anchor Avg.	27.80	45.97	37.88	26.93	54.88	49.54	50.09	26.54	51.65	54.99	45.73	51.10
GPT-5.4 + GPT Image 2	Q1	23.55	43.62	30.72	19.71	52.50	57.73	55.28	24.62	59.12	58.20	51.32	46.88
	Q2	10.67	5.75	9.87	7.49	12.90	11.69	20.55	14.01	10.35	9.57	6.27	7.42
	Q3-Inf.	42.52	40.33	44.74	21.44	32.27	33.42	35.12	33.09	37.60	24.10	36.24	16.03
	Q3-Gen.	79.06	79.87	78.24	76.58	78.61	72.73	80.76	84.09	89.92	88.64	84.46	82.83
	Anchor Avg.	28.09	43.83	28.25	27.66	52.20	47.27	62.16	31.74	51.83	53.49	44.83	44.51

Table 3. Fine-grained results of Qwen3-VL 8B on PREF2ACT. F.g. denotes fine-grained. ⏺, ☑, ☰, and 🖼 denote single-choice, multi-select, short-answer, and image-generation questions, respectively. For each anchor subtask, the first row reports the mean across checkpoints, and the second reports the final checkpoint score minus the initial checkpoint score. **Red** and **blue** highlight the best and second-best results in each row.

			Textual Memory							Multimodal Memory				
Task	**Subtask**	**Type**	*Full (Text)*	*NaiveRAG*	*Gen. Agent*	*Reflexion*	*A-Mem*	*MemoryOS*	*Mem0*	*Full (MM)*	*MuRAG*	*UniversalRAG*	*NGM*	*AUGUSTUS*
Q1	Visual Retrieval	☰	13.33	32.62	25.11	11.04	35.68	**50.27**	34.90	5.73	**38.33**	25.73	28.54	29.17
	Instruction Retrieval	☰	23.33	4.38	10.00	22.50	19.69	28.33	**29.79**	13.13	20.00	**28.44**	17.40	20.42
	Global Pref. Recognition	☑	**74.68**	48.56	41.43	41.01	41.14	32.47	41.87	42.05	**57.58**	50.06	48.40	46.43
	Scenario Pref. Recognition	⏺	27.41	72.92	61.04	26.53	79.86	59.26	84.72	34.31	**92.14**	**91.18**	87.36	80.35
	Personal Pref. Recognition	⏺☰	18.15	62.88	33.51	26.81	**65.56**	59.30	62.26	31.96	**69.87**	64.62	60.95	63.27
Q2	Consecutive Image Comparison	☰	**15.00**	12.13	9.88	9.15	8.17	2.97	11.49	8.71	**19.05**	7.40	6.86	5.92
	Initial-Final Image Comparison	☰	13.33	21.46	20.14	14.23	23.86	4.64	**30.49**	12.15	**25.08**	18.02	16.86	22.93
Q3	Cum. Global Pref. Induction	☑	31.60	51.81	40.51	40.05	37.78	36.67	43.26	33.49	**57.80**	**57.50**	49.56	48.91
	Cum. Scenario Pref. Induction	☰	**36.67**	**29.38**	27.29	15.83	19.17	14.57	23.75	12.29	25.95	27.92	9.79	13.96
	Similar Generation	🖼	65.71	67.83	64.92	63.46	67.58	61.93	68.84	71.62	**83.07**	**81.14**	74.28	72.81
Anchor	Global Pref. Induction	☑	29.17	43.35	30.54	31.12	36.81	9.72	40.85	26.76	**51.48**	**49.00**	45.76	40.81
			-9.10	-10.57	-19.50	-1.42	-17.88	-29.17	**12.02**	-3.57	0.58	**10.26**	-5.22	-10.17
	Scenario Pref. Induction	☰	11.67	22.74	15.66	12.22	20.21	3.24	18.79	6.98	**28.57**	**26.74**	13.09	10.45
			-13.33	12.87	12.50	-5.37	7.78	4.17	**16.76**	3.61	**18.06**	3.70	-7.87	-15.65
	Personal Pref. Induction	☰	9.53	27.98	21.23	19.45	27.48	27.78	**50.60**	23.41	27.44	**32.04**	16.67	18.95
			14.29	4.76	-7.41	-7.93	-3.70	**30.95**	9.79	-1.85	11.51	2.91	-1.85	-1.85
	F.g. Global Pref. Recognition	☑	60.26	73.96	57.62	57.99	73.18	30.08	75.48	52.72	**76.41**	**75.54**	73.44	64.37
			-17.87	4.49	-6.71	-2.75	6.08	5.82	3.77	-2.92	**6.65**	**7.50**	4.19	-0.66
	F.g. Global Pref. Recognition	☰	61.11	49.83	27.26	26.91	48.44	40.74	**61.98**	29.17	**72.62**	53.99	39.41	40.63
			-33.33	4.63	-12.04	-18.98	**6.01**	**6.25**	4.63	-15.28	-13.89	-3.70	-2.31	2.78
	F.g. Scenario Pref. Recognition	⏺	40.96	56.57	42.50	32.43	62.02	47.40	**70.72**	34.81	60.92	**63.44**	56.37	55.69
			8.44	8.22	-5.41	-6.33	9.27	**15.83**	6.24	-8.73	13.57	**15.54**	12.03	11.54
	F.g. Scenario Pref. Recognition	☰	9.24	20.33	14.80	9.92	31.82	**44.34**	**43.09**	10.90	30.18	31.96	24.26	22.64
			-10.84	5.58	-0.76	-4.96	4.10	**21.99**	**13.45**	-3.02	6.12	4.94	2.97	3.40

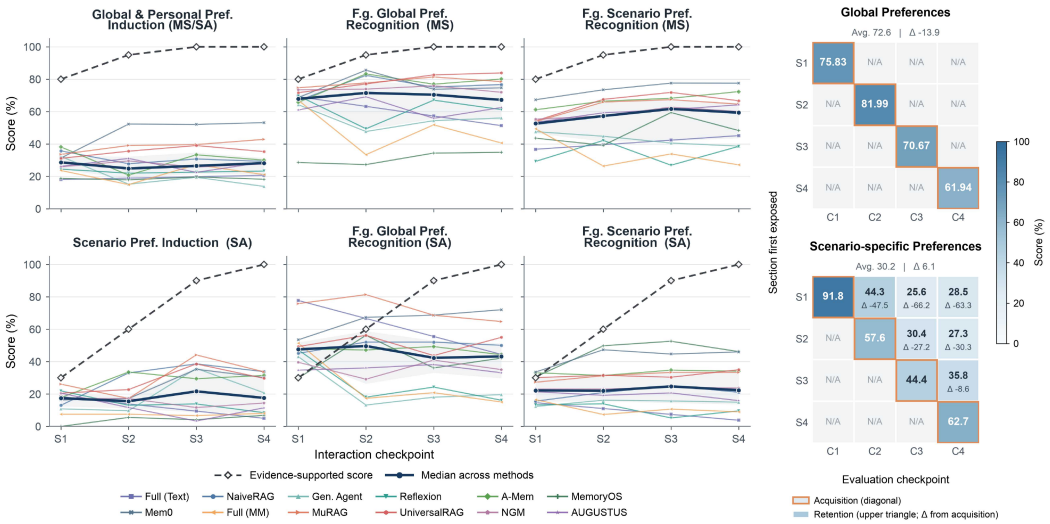


Figure 4. Preference learning across interaction checkpoints. **Left:** Six repeated anchor questions across four checkpoints and 12 memory configurations. Dashed curves show evidence-supported scores, which models may exceed by guessing unrevealed preferences, especially in multi-select questions. **Right:** Preference acquisition and retention for Qwen3-VL-8B with MuRAG. Cell (j,i) reports accuracy at checkpoint i on preferences exposed in section j. Global preferences recur across sections; scenario-specific preferences appear only in their corresponding section. MS and SA denote multi-select and short-answer questions.

- **Obs. 2: No memory method is universally effective.** For every agent model, no memory method performs consistently well across all evaluation questions. The fine-grained results show further specialization: In Qwen3-VL 8B, MemoryOS leads visual retrieval, Mem0 performs best on instruction retrieval and initial–final image comparison. These results indicate that different stages of continual preference memory require different capabilities, and no fixed memory method generalizes across both tasks and agents.
- **Obs. 3: Agents recognize preferences more reliably than they express them.** Across checkpoints, multi-select anchor questions consistently yield higher scores than short-answer questions. This gap suggests that agents benefit from option cues but struggle to retrieve and express preferences independently.

4.3 Preference Learning Dynamics across Interactions (RQ2)

We track preference understanding across four sections under 12 memory configurations. Global preferences recur across sections; scenario-specific preferences appear only in their corresponding section. Figure 4 compares observed and evidence-supported scores. MuRAG heatmaps examine global preference acquisition and scenario-specific retention. We draw the following observations:

- **Obs. 4: Accumulating evidence does not yield cumulative learning of global preferences.** By the final checkpoint, sufficient evidence is available for a score of 100, yet performance remains far below this level. Despite repeated exposure, global preference acquisition rises from 75.83 at C1 to 81.99 at C2, then falls to 70.67 at C3 and 61.94 at C4. This suggests that agents struggle to consolidate repeated evidence into more accurate user models.
- **Obs. 5: Scenario-specific preferences are acquired but rapidly decay without reinforcement.** Preferences introduced in S1 achieve 91.8 at acquisition, but decrease to 44.3 at C2 and 28.5 at C4. Preferences introduced in S2 and S3 show the same pattern, declining from 57.6 to 27.3 and from 44.4 to 35.8, respectively. These results indicate that current memory methods struggle to preserve both a preference and the scenario in which it applies as interactions move across contexts.

4.4 From Preference Understanding to Image Generation (RQ3)

We take image generation tasks as the unit of analysis. For each failed generation, we examine whether the agent correctly identifies the scenario-specific preferences. Figure 5 shows these paired outcomes, and Appendix D reports the proportion of failures with correct preference identification.

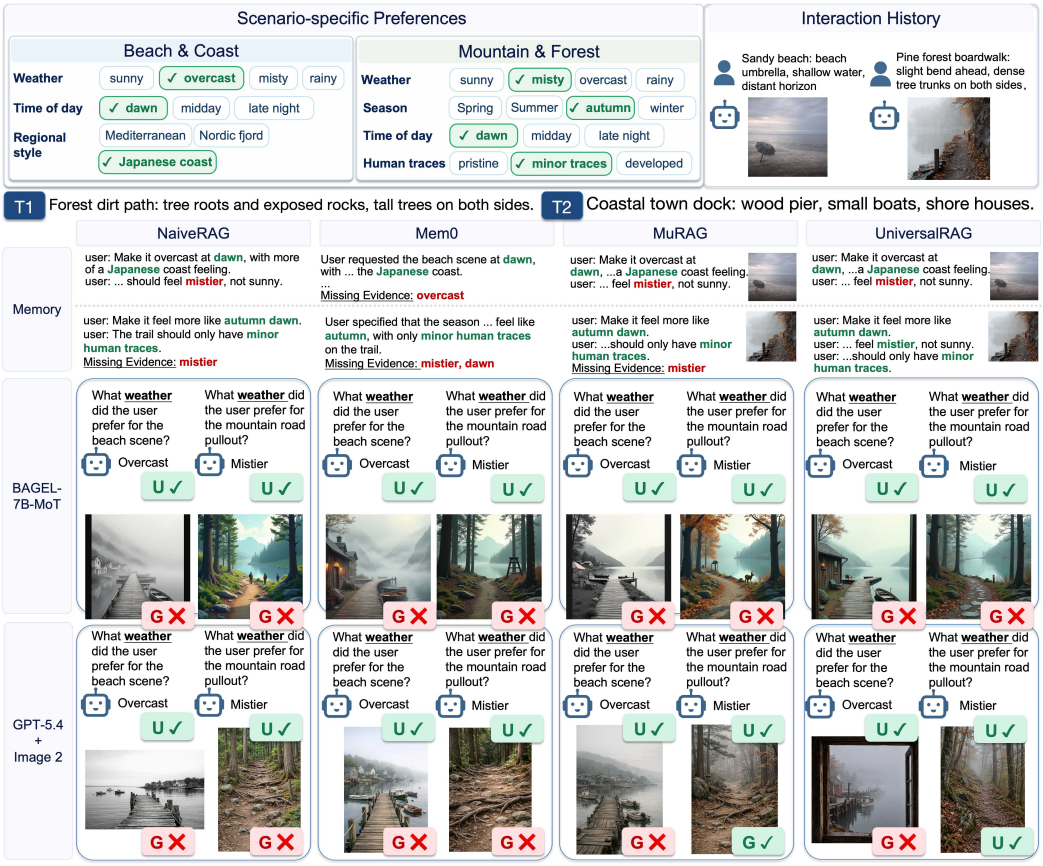


Figure 5. Case study of retrieved memories, preference understanding (U), and image generation (G) for two agents under four memory methods.

- **Obs. 6: Correct preference understanding does not ensure appropriate application.** Agents can correctly answer explicit preference questions but still mix attributes from different scenarios when generating images for similar requests. This mismatch exposes a gap: identifying preferences under explicit evidence does not ensure appropriate application. Similar gaps have also been reported in personalized preference following (Zhao et al., 2025), preference-conditioned task execution (Song et al., 2026), and reasoning-centric text-to-image generation (Li et al., 2026). Additionally, multimodal memory sometimes improves consistency, possibly because retrieved images provide direct visual references.

4.5 Benchmark Reliability (RQ4)

We assess reliability using reference evidence, multiple model judges, and human review, as detailed in Appendix E. When reference evidence is provided directly as memory, all tested agents achieve over 85% accuracy, supporting question answerability. Scores remain similar across evaluation models and align with human judgments, supporting the consistency of our evaluation.

5 Conclusion

We introduced Pref2Act, a benchmark that jointly evaluates preference understanding and personalized image generation and editing across extended multimodal interactions. Experiments with five agents across twelve memory configurations reveal difficulties in inferring implicit preferences, learning from accumulated feedback, and consistently applying remembered preferences. These findings suggest that reliable personalization requires linking preferences to supporting evidence and applicable scenarios. Future memory systems should turn accumulated feedback into better decisions about which preferences guide each new request. Progress should therefore be measured by both accurate preference understanding and consistent behavior without explicit reminders.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and polishing the manuscript and searching for relevant literature. We also used generative AI tools to generate images and useragent dialogues for the construction of PREF2ACT. We take responsibility for the final manuscript and dataset, including text, claims, references, images, and dialogues produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the construction of PREF2ACT, including user profiles, interaction histories, quality control, and evaluation questions, in Appendix A. Appendix C provides the experimental environment, evaluation protocol, scoring metrics, and prompts used by the model judges. Detailed experimental results are reported in Appendix D. We will publicly release PREF2ACT and the code for evaluation framework to support reproducibility and future research.

REFERENCES

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. MemoryBench: A Benchmark for Memory and Continual Learning in LLM Systems. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=If4X4W2HWx.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

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, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.375. URL https://aclanthology.org/2022.emnlp-main.375/.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*, 2025. doi: 10.48550/arXiv.2501.17811. URL https://arxiv.org/abs/2501.17811.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7727–7736, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_Tailored_Visions_Enhancing_Text-to-Image_Generation_with_Personalized_Prompt_Rewriting_CVPR_2024_paper.html.

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. doi: 10.48550/arXiv.2504.19413. URL https://arxiv.org/abs/2504.19413.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, Weijia Wu, Hanwang Zhang, and Tat-Seng Chua. WEAVE: Unleashing and Benchmarking the In-context Interleaved Comprehension and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15343–15353, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chow_WEAVE_Unleashing_and_Benchmarking_the_In-context_Interleaved_Comprehension_and_Generation_CVPR_2026_paper.html.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025. doi: 10.48550/arXiv.2505.14683. URL https://arxiv.org/abs/2505.14683.

Matthew Fisher. Neural Graph Memory (NGM). https://github.com/StuckInTheNet/Neural-Graph-Memory-NGM, 2025. Software repository. Accessed September 26, 2026.

Google DeepMind. Gemini 3.6 Flash: Model Card. https://deepmind.google/models/model-cards/gemini-3-6-flash/, 2026. Published July 21, 2026. Accessed September 26, 2026.

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*, 2026. doi: 10.48550/arXiv.2606.07402. URL https://arxiv.org/abs/2606.07402.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. AUGUSTUS: An LLM-Driven Multimodal Agent System with Contextualized User Memory. *arXiv preprint arXiv:2510.15261*, 2025. doi: 10.48550/arXiv.2510.15261. URL https://arxiv.org/abs/2510.15261.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. PersonaMem-v2: Towards Personalized Intelligence via Learning Implicit User Personas and Agentic Memory. *arXiv preprint arXiv:2512.06688*, 2025. doi: 10.48550/arXiv.2512.06688. URL https://arxiv.org/abs/2512.06688.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI Agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Jeongeun Lee, Ryang Heo, and Dongha Lee. Personalized Reward Modeling for Text-to-Image Generation. *arXiv preprint arXiv:2511.19458*, 2025. doi: 10.48550/arXiv.2511.19458. URL https://arxiv.org/abs/2511.19458.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. GIR-Bench: Versatile benchmark for generating images with reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6274d57365d7a6be06e58cad30d1b9da-Abstract-Conference.html.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. PERMA: Benchmarking Personalized Memory Agents via Event-Driven Preference and Realistic Task Environments. *arXiv preprint arXiv:2603.23231*, 2026. doi: 10.48550/arXiv.2603.23231. URL https://arxiv.org/abs/2603.23231.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Ofir Nabati, Guy Tennenholtz, Chihwei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. Preference Adaptive and Sequential Text-to-Image Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45362–45394. PMLR, 2025. URL https://proceedings.mlr.press/v267/nabati25a.html.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a. Published March 5, 2026. Accessed September 26, 2026.

OpenAI. GPT-Image-2 Model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026b. Official API model documentation. Accessed September 26, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

Naina Raisinghani. Introducing Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025. Official announcement of Gemini 3 Pro Image, November 20, 2025. Accessed September 26, 2026.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. PMG : Personalized Multimodal Generation with Large Language Models. *arXiv preprint arXiv:2404.08677*, 2024. doi: 10.48550/arXiv.2404.08677. URL https://arxiv.org/abs/2404.08677.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Jiajia Song, Bobo Li, Haiwen Yi, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee, and Wynne Hsu. From profiling to synthesis: Benchmarking implicit behavioral alignment in personalized LLM agents. *arXiv preprint arXiv:2608.02171*, 2026. URL https://arxiv.org/abs/2608.02171.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025a. doi: 10.48550/arXiv.2508.02324. URL https://arxiv.org/abs/2508.02324.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=pZiyCaVuti.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, and Liang Wang. MME-Unify: A Comprehensive Benchmark for Unified Multimodal Understanding and Generation Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d8308c2c45cf95bfebb736be377f0982-Paper-Conference.pdf.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.177. URL https://aclanthology.org/2026.acl-long.177/.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9274–9285, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/28a46044775d97a4efcbcf14e7f13209-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-MMMU: A Massive Multi-discipline Multimodal Unified Benchmark. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 908–924, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.40. URL https://aclanthology.org/2026.acl-long.40/.

Appendix provides further details and analyses supporting the construction and evaluation of PREF2ACT. It is organized as follows:

- **Appendix A: Benchmark Construction.** Appendix A describes the user profiles and preference space, the quality control of long horizon interaction histories, and the construction of cumulative memory and repeated anchor questions.
- **Appendix B: Baseline Configurations.** Appendix B introduces the unified and modular agents evaluated in our experiments, together with the textual and multimodal memory configurations.
- **Appendix C: Implementation Details.** Appendix C presents the experimental environment, evaluation settings, evaluation metrics, and prompts used by the model judges.
- **Appendix D: Detailed Experimental Results.** Appendix D reports the complete fine-grained results across all agents and memory configurations and further analyzes the knowing–doing gap in image generation.
- **Appendix E: Benchmark Reliability.** Appendix E assesses the answerability of the evaluation questions and measures agreement among different model judges and human evaluators.

A Dataset Construction Details

This section describes how we specify user preferences, construct and validate interaction histories, and formulate evaluation questions. Figure 6 presents selected examples from PREF2ACT.

(a) Interaction History

S001/G001

T001:
USER Help me generate an image. Small bedroom: single bed, bedside rug, bolster pillows, wood nightstand
ASSISTANT Sure! I'll generate an image.



image caption
A cozy, warmly lit bedroom features a wooden bed with a rumpled beige comforter ...

T002:
USER This feels too realistic, almost like a room photo. Make it less literal and more abstract.
ASSISTANT Sure! I'll generate an image.



image caption
This is a stylized, low-poly illustration of a cozy bedroom with warm brown tones...

...

T009:
USER This direction is closer. Keep the room sparse and clean.
ASSISTANT Sure! I'll generate an image.



image caption
A minimalist bedroom with geometric-patterned walls and floor, featuring a wooden bed,...

T010:
USER Put a small cactus on the nightstand.
ASSISTANT Sure! I'll generate an image.



image caption
A minimalist bedroom with geometric-patterned walls and floor, featuring a wooden bed,...

S001/G002

T001:
USER What kind of cat is this? It looks unusually calm.



image caption
A gray Scottish Fold cat lounges on a white towel beside a blue bowl, with a window and greenery..

ASSISTANT This looks like a Scottish Fold. The folded ears and round face are strong clues.

T002:
USER Would this kind of cat be suitable for a small apartment?
ASSISTANT Yes, if it has a clean litter box, scratching post, soft resting spots, and light daily play.

...

T006:
USER Sounds great. I like pets that don't make the room feel chaotic.
ASSISTANT Then this type fits that feeling well: quiet, indoor-friendly, and usually gentle..

(b) Evaluation Data

Q1 / Visual Retrieval
In G002, what object is lying on the stone bridge in the foreground in the final user-satisfied image?
REFERENCE A compass.
Evidence: S003 / G002 / T007

Q1 / Instruction Retrieval
What was the last image editing instruction issued by the user in G002?
REFERENCE Add a small compass on the stone bridge, like someone left it there.
Evidence: S003 / G002 / T007

Q1 / Global Preference Recognition
Based on the complete conversation in G015, which of the following labels best reflect the overall preferences revealed by the user? Select all that apply.
A. Lyrical Softness B. Geometric Rationality C. Negative Space D. Pixel Art ...
REFERENCE B D E G H
Evidence: S003 / G002 / T002, T004, T005, T006

Q1 / Subcategory Preference Recognition
In Group G001, what time of day did the user prefer for the bedroom scene? Select one.
A. morning B. afternoon C. dusk D. night
REFERENCE B
Evidence: S003 / G002 / T002, T004

Q1 / Personal Preference Recognition
In G001, what city or place does the user prefer?
REFERENCE The user prefers empty night streets of Tokyo.
Evidence: S003 / G001

Q2 / Consecutive Image Comparison
In G002, what changed in Image 5 compared with Image 4?
REFERENCE The visual style changed from a realistic room-photo look to an abstract low-poly image...
Evidence: S001 / G002 / T005, T007

Q2 / Initial-final Image Comparison
In G009, how did the style, color, and occupancy change from the first image to the final image
REFERENCE
1: style_orientation: ornate → modern minimalist,
2: color_system: warm → cool,
3: color_system: moderately saturated → low-saturation,
Evidence: S002 / G002 / T001, T007

Q3 / Cumulative Global Preference Induction
So far, which of the following labels best reflect the user's overall preferences? A label should be selected only if the corresponding preference has been explicitly emphasized by the user at least twice. Select all that apply.
01. Sketch 02. Documentary Photography 03. Leading Lines 04. Desolate and Lonely ...
REFERENCE 01,22,27,...
Uses the full S003 history.

Global Preference Recognition / Multi-select
In the domain of Home and Everyday Living, please identify the preference items that have been explicitly emphasized by the user at least twice.
REFERENCE
Interior style: modern minimalist,
Time of day: night,
Tidiness: tidy
...

Figure 6. Examples of interaction history and evaluation questions in PREF2ACT.

A.1 User Profiles and Preference Space

A.1.1 Global Visual Preferences

Global visual preferences describe a user's general aesthetic tendencies across visual scenarios. We organize these preferences into six dimensions: **rendering style**, **style orientation**, **color system**, **atmosphere**, **composition and visual organization**, and **realism level**. **Rendering style** specifies the visual medium or representation, while **style orientation** describes the broader aesthetic character. The other dimensions capture color characteristics, emotional tone, spatial arrangement, and the degree of realism, respectively. Table 4 summarizes the preference options across the six user profiles, with duplicate entries removed within each dimension. Each profile specifies its own selection of these options, which guide the construction of preference evidence across multiple scenarios.

Table 4. Global visual preference dimensions and preference options aggregated across the six user profiles.

Dimension	Preference options
`Rendering Style`	Digital painting; 3D rendering / CGI; low-poly; macro photography; Chinese ink; documentary photography; sketch; film photography; watercolor; colored pencil; pixel art; printmaking / woodcut; studio commercial photography; oil painting.
`Style Orientation`	Geometric rationalism; modern minimalism; Eastern elegance; lyrical softness; retro pop; decorative design; expressive tension; surreal fantasy; classical order.
`Color System`	Black, white, and gray; low saturation; medium saturation; high saturation; cool tones; warm tones; monochromatic / analogous; rich full color; high contrast.
`Atmosphere`	Futuristic and technological; desolate and lonely; cold and empty; melancholic and hazy; solemn and austere; calm and restrained; warm and healing; playful and childlike; lively and energetic; dreamy and ethereal; mysterious and eerie; tense and oppressive; romantic and gentle; ornate and flourishing.
`Composition & Visual Organization`	Negative space; leading lines; symmetrical composition; minimal and singular; frame within frame; rule of thirds; centered subject; full-frame / no negative space; dense composition.
`Realism Level`	Abstract; realistic.

A.1.2 Scenario-Specific Preferences

Scenario-specific preferences describe a user's preferences for visual attributes or editing operations within a particular scenario. We organize the scenario taxonomy into seven main categories and 51 subcategories. Table 5 reports all subcategories, their scene descriptions, and the specified preference dimensions and candidate values. These dimensions cover attributes such as background, facial expression, interior style, weather, and retouching strength. Within each subcategory, the candidate values of each preference dimension are treated as mutually exclusive. During profile construction, each user profile is assigned a single preferred value per dimension at a time. The same dimension may take different preferred values across scenarios, requiring agents to retain each preference together with the scenario in which it applies.

Table 5. Scenario taxonomy and corresponding preference dimensions. Each main category is followed by its subcategories, scene descriptions, and candidate preference values. A dash indicates that preference dimensions are not specified in the taxonomy.

Subcategory	Description	Preference dimensions and values
Portraiture & Human Relationships		
Headshot Portrait	Head-and-face portrait capturing facial features, face shape, hairstyle, expression, and personal temperament.	**Background:** solid color / gradient / blurred scene / transparent **Face angle:** front / left tilt / right tilt / upward / downward **Expression:** neutral / warm smile / serious / contemplative **Skin texture:** realistic / soft-airbrushed
Half-Body Portrait	Upper-body portrait retaining facial detail while showcasing shoulders, hands, and clothing layers.	**Background:** solid / gradient / blurred scene / transparent / identifiable space **Face angle:** front / left tilt / right tilt / upward / downward **Expression:** neutral / warm smile / serious / contemplative
Full-Body Portrait	Full-length portrait displaying overall silhouette, clothing, and body posture.	**Background:** solid / gradient / blurred scene / transparent / identifiable space **Face angle:** front / left tilt / right tilt / upward / downward **Expression:** neutral / warm smile / serious / contemplative
Two-Person Portrait	Portrait of two subjects, emphasizing relative positioning, proximity, interaction, and visual balance.	**Gaze:** mutual / both toward camera / one each way / neither **Visual hierarchy:** symmetric / one forward / one sharp one blurred **Emotional intensity:** restrained / warm and natural / intimate **Background:** solid / gradient / blurred scene / transparent / identifiable space
Parent-Child Portrait	Portrait of adult and child, emphasizing physical proximity, caregiving relationship, and mode of interaction.	**Camera angle:** adult POV down / child POV up / eye-level / bird's-eye **Emotional intensity:** restrained / warm and natural / intimate **Background:** solid / gradient / blurred scene / transparent / identifiable space
Infant & Toddler Portrait	Portrait of newborns, infants, and toddlers, emphasizing age stage, body scale, and facial and limb features.	**Expression:** peaceful / curious smile / focused gaze / crying **Accompanying elements:** none / adult hands / blanket / toy **Surface:** bed / floor mat / stroller / adult's arms **Background:** solid / gradient / blurred scene / transparent / identifiable space
Wedding Anniversary Portrait	Wedding or anniversary commemorative portrait; emphasizes the subjects and their relationship, not the venue.	**Gaze:** both toward camera / mutual / one each way / neither **Accessory:** bouquet / rings / vow card / cake and champagne / none **Emotional intensity:** restrained / warm and natural / intimate **Background:** solid / gradient / blurred scene / transparent / identifiable space

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Professional Identity Portrait	Portrait centered on occupational identity, emphasizing professional recognizability without complex narrative settings.	**Professional symbol density:** strong (visible tools) / weak (attire only) / absent (demeanor only) **Work state:** mid-task / facing camera / paused in thought **Background:** solid / gradient / blurred professional environment
Home & Daily Life		
Window-Side Rest Nook	Small-scale living corner anchored by armchair, side table, reading material, throw, and lamp near a window.	**Interior style:** modern minimalist / Japanese natural / French casual / mid-century / neo-Chinese / industrial **Time of day:** morning / afternoon / dusk / night **Tidiness:** tidy / slightly messy / very messy
Living Room & Family Lounge	Shared living space centered on sofa, coffee table, TV cabinet, side cabinet, rug, and lighting.	**Interior style:** modern minimalist / Japanese natural / French casual / mid-century / neo-Chinese / industrial **Time of day:** morning / afternoon / dusk / night **Tidiness:** tidy / slightly messy / very messy
Bedroom & Sleeping Space	Private rest space centered on bed, headboard, bedding, nightstand, curtains, and localized storage.	**Interior style:** modern minimalist / Japanese natural / French casual / mid-century / neo-Chinese / industrial **Time of day:** morning / afternoon / dusk / night **Tidiness:** tidy / slightly messy / very messy
Dining Area & Family Table	Dining zone centered on table, chairs, tablecloth, tableware, pendant light, and table arrangement.	**Interior style:** modern minimalist / Japanese natural / French casual / mid-century / neo-Chinese / industrial **Table atmosphere:** everyday meal / formal reception / festive gathering / unoccupied **Time of day:** morning / afternoon / dusk / night
Bathroom & Bathing Space	Bathing space centered on bathtub, shower, basin, mirror, towels, toiletries, and tiles.	**Material style:** marble / concrete / wood-tone / white ceramic / vintage patterned tile **Steam level:** dry and clean / post-bath humid / steam-filled **Time of day:** morning / afternoon / dusk / night
Compact & Studio Apartment	Compact dwelling unit emphasizing co-location of sleeping, resting, storage, and multi-function layout within limited floor area.	**Interior style:** modern minimalist / Japanese natural / French casual / mid-century / neo-Chinese / industrial **Time of day:** morning / afternoon / dusk / night **Tidiness:** tidy / slightly messy / very messy
Log Cabin & Rural Home	Rural living space centered on timber beams, fireplace, reclaimed wood furniture, plaid textiles, and natural materials.	**Natural light:** ample daylight / localized point light / dim and amber / single shaft of light **Aging level:** renovated / moderately worn / abandoned **Object density:** sparse / moderately layered / abundantly cluttered
Basement, Attic & Hidden Storage	Basement, attic, pitched-roof space, old storeroom, or concealed nook; emphasizes enclosure, stacked objects, and distinctive spatial character.	**Natural light:** ample daylight / localized point light / dim and amber / single shaft of light **Aging level:** renovated / moderately worn / abandoned **Object density:** sparse / moderately layered / abundantly cluttered
Food, Dining & Retail		
Everyday Meal & Home Dining Table	Home-cooked meal scene centered on household dishes, plates, bowls, chopsticks, and everyday tableware.	**Human presence:** no person (overhead still life) / hands in frame / partial figure in foreground **Presentation style:** homemade casual / boutique display / commercial shoot **Table surface:** marble / wood / concrete / solid-color cloth / plaid cloth / floral cloth

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Breakfast, Brunch & Morning Table	Breakfast plate, toast, egg, jam, coffee, fruit, and morning table; suited to light meals and brunch settings.	**Human presence:** no person (overhead still life) / hands in frame / partial figure in foreground **Presentation style:** homemade casual / boutique display / commercial shoot **Table surface:** marble / wood / concrete / solid-color cloth / plaid cloth / floral cloth
Beverage & Dessert Display	Coffee, tea, desserts, cups, trays, glass domes, price tags, and counter display; suited to single beverages, dessert sets, and in-store presentation.	**Human presence:** no person (overhead still life) / hands in frame / partial figure in foreground **Presentation style:** homemade casual / boutique display / commercial shoot **Table surface:** marble / wood / concrete / solid-color cloth / plaid cloth / floral cloth
Ice Cream & Sweet Treats	Ice cream scoops, cones, clear cups, shaved ice, jam, fruit garnish, and dessert counter; conveys a cool, light dessert atmosphere.	**Presentation style:** homemade casual / boutique display / commercial shoot **Human presence:** no person (overhead still life) / hands in frame / partial figure in foreground
Bar, Tavern & Nighttime Drink Space	Bar counter, wine glasses, bottles, snacks, menu boards, and nighttime table; conveys social drinking and pub atmosphere.	**Crowd density:** empty / sparse / moderate / crowded
Bakery, Grocery & Small Shop Frontage	Small shop frontage, window display, signage, shelving, and outdoor placement; suited to street retail and community storefronts.	**Regional style:** Japanese / European street / Chinese local / American small-town
Farmers' Market & Fresh Produce Stall	Produce stall, wooden crates, wicker baskets, price tags, flower bunches, and open-air booth; conveys natural arrangement and local market character.	**Weather:** sunny / overcast / rainy
Urban Mobility & Public Space		
Street & Corner Everyday Life	Urban street, intersection, pedestrians, roadside shops, road signs, barriers, and street furniture; conveys everyday urban life.	**Urban character:** Asian metropolis / European old town / American neighborhood / rural town **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Subway & Urban Commute Space	Subway car, platform, turnstile, corridor, train doors, signage, and commuting crowds; conveys high-frequency mobility and everyday commute.	**Urban character:** Asian metropolis / European old town / American neighborhood / rural town **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Train Station & Rail Waiting Space	Station hall, waiting area, platform, ticket windows, clock, luggage, and railway structures; suited to long-distance travel and railway spatial character.	**Urban character:** Asian metropolis / European old town / American neighborhood / rural town **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Bus Stop & Roadside Waiting Point	Bus stop sign, shelter, bench, schedule board, canopy, roadside billboard, and waiting passengers; conveys brief stopover and public street infrastructure.	**Urban character:** Asian metropolis / European old town / American neighborhood / anonymous city **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Bridge, Riverbank & Waterfront Promenade	Bridge structure, riverside railing, walkway, fence, street lamp, water reflection, and urban waterfront; suited to strolling and waterside public activities.	**Urban character:** Asian metropolis / European old town / American neighborhood / anonymous city **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Courtyard, Atrium & Urban Inner Court	Urban courtyard, inner court, atrium, light well, paving stones, enclosing walls, potted plants, and semi-open rest areas.	**Urban character:** Asian metropolis / European old town / American neighborhood / anonymous city **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Observation Deck & Elevated Platform	Observation platform, railing, telescope, elevated walkway, steps, and distant city or natural vista; suited to cityscape viewing and elevated stopover.	**Urban character:** Asian metropolis / European old town / American neighborhood / anonymous city **Crowd density:** empty / sparse / moderate / crowded **Weather:** bright sunlight / overcast diffused / post-rain wet / night rain and neon **Temporal feel:** contemporary / historically layered / abandoned and decayed
Clinic & Community Medical Space	Clinic reception, waiting area, informational posters, door plaques, corridor, and basic medical furniture; suited to routine healthcare and community medical settings.	**Crowd density:** empty / sparse / moderate / crowded **Temporal feel:** contemporary / historically layered / abandoned and decayed
Nature & Outdoor Environments		
Natural Landscape Panorama	Large-scale natural landscape including mountains, lakes, grasslands, and coastline; emphasizes terrain layers, expansive vista, and overall natural state.	**Weather:** sunny / overcast / misty / rainy / snowy **Season:** spring / summer / autumn / winter **Time of day:** dawn / midday / golden hour / blue hour night / late night **Human traces:** pristine / minor traces / evident development
Mountain, Forest & Trail	Mountain path, forest trail, wooden boardwalk, stream, and hiking route; emphasizes the walking experience of entering natural interiors.	**Weather:** sunny / overcast / misty / rainy / snowy **Season:** spring / summer / autumn / winter **Time of day:** dawn / midday / golden hour / blue hour night / late night **Human traces:** pristine / minor traces / evident development

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Beach, Coast & Coastal Life	Sandy beach, reef, waves, dock, coastal town, and waterfront facilities; conveys natural coastal environment and light recreational activity.	**Weather:** sunny / overcast / misty / rainy / snowy **Time of day:** dawn / midday / golden hour / blue hour night / late night **Regional style:** Mediterranean / Southeast Asian tropical / Nordic fjord / Californian / Japanese coast / British seaside / tropical island
Camping & Outdoor Leisure	Tent, folding chair, camping lantern, campfire, camp table, and outdoor gear; emphasizes brief off-grid stopover and wilderness leisure.	**Camping style:** rugged minimal / glamping luxe / family leisure **Time of day:** dawn / midday / golden hour / blue hour night / late night
Rooftop & Elevated Social Space	Rooftop bar, terrace, observation platform, elevated table, railing, and nighttime social scene; emphasizes elevated stopover and social ambiance.	**Crowd density:** empty / sparse / normal gathering / lively party **Time of day:** dawn / midday / golden hour / blue hour night / late night **View type:** city skyline / mountain view / ocean view / mixed landscape
Road Trip & Driving Scene	Highway, windshield perspective, service area, mountain road, coastal highway, and roadside stop; emphasizes sense of movement and travel.	**Weather:** sunny / overcast / misty / rainy / snowy **Time of day:** dawn / midday / golden hour / blue hour night / late night
Village & Rural Life	Farmhouse, fields, orchard, wood fence, barn, and seasonal crops; emphasizes rural dwelling and agricultural environment.	**Weather:** sunny / overcast / misty / rainy / snowy **Season:** spring / summer / autumn / winter **Time of day:** dawn / midday / golden hour / blue hour night / late night **Regional style:** southern Chinese waterway / northern Chinese loess / European farmstead / Japanese satoyama / Southeast Asian tropical village
Pool, Bath & Water Leisure	Swimming pool, water surface, lounger, sun umbrella, wooden deck, and poolside facilities; conveys light leisure and a holiday atmosphere.	**Crowd density:** empty / sparse / moderate / crowded **Time of day:** dawn / midday / golden hour / blue hour night / late night **Pool type:** indoor / outdoor courtyard / resort infinity / public bath
Window View & Framed Landscape	Indoor-to-outdoor framed view; emphasizes window frame, interior-exterior light contrast, and the captured scenery.	**Weather:** sunny / overcast / misty / rainy / snowy **Season:** spring / summer / autumn / winter **Time of day:** dawn / midday / golden hour / blue hour night
Fantasy, Sci-Fi & Stylized Worlds		
Fantasy & Magic World	Magical architecture, floating terrain, ancient ruins, mystical forest, spellwork traces, and otherworldly environments; emphasizes non-realistic spatial settings.	**Magic intensity:** low magic realism / mid-magic hybrid / high magic ornate **World reference:** European medieval / East Asian mythology / Middle Eastern Arabian / ancient Egyptian / Lovecraftian deep sea
Sci-Fi & Future World	Future city, orbital transit, megastructures, robotic facilities, spaceport, and high-density technology environment; emphasizes future societal systems.	**Future vision:** utopian prosperity / grey realism / cyberpunk dystopia / post-apocalyptic ruin **Tech style:** cyberpunk / space opera / biopunk / soft tech / eco-tech

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Cartoon, Anime & Stylized Character	Animated character, stylized portrait, exaggerated costume, distinctive silhouette, and character concept; emphasizes the character rather than realistic figures.	**Animation style:** Japanese shonen/shojo / Western cartoon / Chinese animation / indie hand-drawn **Realism level:** highly stylized / semi-realistic / realistic with stylization **Character temperament:** energetic / elegant and composed / rebellious / solitary and mysterious
Monsters, Creatures & Mythological Beings	Dragon, unicorn, phoenix, sea monster, forest creature, guardian beast, and fantastical being; emphasizes the creature's visual design.	**Creature visual style:** epic fantasy / dark gothic / Eastern mythology / fairy tale / Lovecraftian alien / nature guardian
Historical & Culturally Stylized Scene	Ancient town, temple fair, historical street, traditional ritual, cultural space, and stylized historical setting; emphasizes cultural atmosphere with allowance for non-realistic enhancement.	**Era atmosphere:** prosperous heyday / ordinary everyday / turbulent or end-of-era **Human subject role:** no figures / figures as accents / figures and scene balanced / figures as primary
Anthropomorphization & Absurdist Combinations	Animal anthropomorphization, object personification, displaced scenarios, and absurdist combinations; emphasizes visual contrast from an unexpected subject performing a mundane act.	—
Photography & Post-Processing		
Portrait Retouching & Image Enhancement	Portrait editing involving makeup, skin texture, blemishes, skin tone, facial features, face shape, and body proportions.	**Makeup style:** bare skin / everyday natural / Korean / Western glam / vintage / evening / festive **Foundation texture:** bare skin / translucent cream / soft matte **Eye makeup:** none / subtle nude eye / smoky eye **Lip style:** natural bare / light pink / deep red / orange-toned **Skin texture:** original / light airbrushed / heavy airbrushed **Blemish handling:** retain original / remove acne / remove spots / remove dark circles / soften nasolabial folds / remove crow's feet **Skin tone correction:** original / cool white / warm white / yellow-toned / tan **Facial enhancement:** none / eye enlargement / double eyelid / aegyo sal / nose bridge / lip shape / teeth whitening **Face shape retouching:** none / subtle slimming / soft contouring / jawline lift / V-face / refined bone structure **Body retouching:** none / subtle slimming / waist cinching / leg elongation / proportion optimization

Continued on next page

Table 5 (continued)

Subcategory	Description	Preference dimensions and values
Filter, Color Grading & Style Shaping	Photographic appearance adjustment through overall tone, color temperature, contrast, brightness, color cast, film grain, and clarity.	**Overall tone:** fresh and light / natural neutral / warm healing / cool and clear / vintage brown / HK teal-orange / cinematic grey / Japanese pastel **Color temperature:** cool / neutral / warm / golden warm / cool cyan **Contrast:** low contrast soft / medium / high contrast **Brightness:** brightened / original / dark and moody **Color cast:** blue / cyan / green / orange / brown / pink / none **Film grain:** no grain digital / light film grain / heavy film grain **Clarity:** soft / natural / sharpened
Landscape & Travel Scene Retouching	Landscape and travel photo editing involving sky, water, vegetation, weather, time of day, color cast, and visible crowds.	**Sky treatment:** original / brightened blue sky / enhanced clouds / added sunset glow / sky replacement **Water surface:** natural / enhanced clarity / strengthened reflection / more vivid blue / glittering waves **Vegetation:** natural / vibrant green / deep lush green / autumn warm yellow **Weather enhancement:** clear and bright / overcast clean / post-rain fresh / snow purity / light mist / storm cloud **Time-of-day rendering:** dawn / midday / dusk / blue hour / night with lights **Color cast:** blue / cyan / green / orange / brown / none **Crowd handling:** retain original / remove distracting tourists / retain minimal figures
Lifestyle & Urban Photo Retouching	Editing of everyday and urban photographs through background cleanup, subject emphasis, tone, exposure, and color cast.	**Background cleanup:** none / light clutter removal / notable background clearance **Subject prominence:** natural emphasis / light background blur / strong background blur **Tone style:** natural neutral / warm home / cool and clean / cream spatial / wood warm tone / urban grey **Brightness and exposure:** brightened / original / darkened moody **Color cast:** blue / cyan / green / orange / brown / none
Artistic Style Conversion	Conversion of an existing image into an artistic medium or visual style, with different degrees of fidelity to the source image.	**Style conversion:** comic / sketch / colored pencil / watercolor / oil painting / printmaking / ink wash / pixel art / low-poly **Source fidelity:** high fidelity (composition and detail fully retained) / moderate stylization (subject retained, details abstracted) / free reinterpretation (core object only)

A.1.3 Personal Content Preferences

Personal content preferences capture the specific content that a user favors. For each user, we construct an individualized preference profile covering seven categories: **animals**, **city or place imagery**, **plants or flowers**, **symbolic objects**, **scenes**, **colors**, and **sports**. These preferences shape the content of visual requests and user feedback. In each dialogue section, we construct fictional personal experiences or memories specific to the user, connecting their preferences to the current scenario. These details are introduced through natural conversation and are designed to remain consistent with the user's profile across sections. They provide personal context for preference expression, allowing agents to infer and retain individual interests from the interaction history.

A.1.4 Interaction Preferences

Interaction preferences describe how a user expects the assistant to communicate during multimodal interactions. We represent them using three dimensions: **answer detail level**, **response structure**, and **extension direction**. These dimensions determine the amount of information, the organization of the response, and the type of additional content beyond the core answer, respectively. Table 6 defines the available options. This representation supports evaluating whether agents retain and follow a user's communication preferences as the interaction history grows.

Table 6. Interaction preference dimensions, available options, and their meanings.

Dimension	Option	Description
Answer detail level	`brief`	A short, direct answer.
	`standard`	The answer with a small amount of useful explanation.
	`extended`	A more detailed answer with additional relevant information.
Response structure	`prose_paragraph`	A response written as a natural paragraph.
	`bullet_list`	A response organized into bullet points.
	`table_format`	A response presented in a structured table.
Extension direction	`no_extension`	No extra information beyond the direct answer.
	`visual_reasoning`	An explanation grounded in visible image features.
	`knowledge_extension`	Relevant background knowledge or general information.
	`comparison`	A comparison with similar objects, species, styles, or categories.

A.2 Interaction History Quality Control

We perform structural and semantic checks to ensure that interaction histories provide reliable and traceable preference evidence. Structural checks verify the completeness of dialogue fields, the chronological order of turns, and the consistency of image identifiers and paths. Semantic checks assess whether user feedback is grounded in observable properties of the preceding image, whether subsequent outputs reflect the requested changes, and whether preference updates remain consistent with the evolving user profile.

We remove interactions that disclose the complete user profile, rely on ambiguous visual evidence, or provide preference evidence solely through assistant summaries. Dialogues containing severe image artifacts, outputs that poorly follow user instructions, or unsupported preference labels are revised, regenerated, or excluded. These checks preserve natural interactions while maintaining a clear connection between preference labels and their supporting evidence.

A.3 Evaluation Data Construction

We organize evaluation data into cumulative memory questions and repeated anchor questions. The following examples illustrate each category. Long candidate lists are omitted where noted.

A.3.1 Q1: Explicit Fact Retrieval

Q1 evaluates whether agents can recover visual details, instructions, and preferences from identifiable evidence in specified dialogue groups.

Visual Retrieval. Short-answer questions ask about objects, attributes, or spatial details in a previous image. They assess whether agents retain and retrieve visual information associated with a particular interaction.

Example: "In G001, what kind of plant is placed on the bedside table in the final user-satisfied image?"
Reference answer: "A cactus."

Instruction Retrieval. Short-answer questions ask agents to recall a specific user request. They assess whether agents can recover the requested operation and its target from the dialogue history.

Example: "In G001, what was the last image-editing instruction issued by the user?"
Reference answer: "The user asked to put a small cactus on the nightstand."

Global Preference Recognition. Multi-select questions ask agents to identify aesthetic preferences expressed within a specified dialogue group. They assess whether agents can map user feedback to the corresponding global preference labels.

Example: "Based on the complete conversation in G001, which of the following labels best reflect the overall preferences revealed by the user? Select all that apply."
Options: A. Realistic; B. Abstract; C. Warm Tone; D. Low Poly; E. Modern Minimalism; F. Dense Composition; G. Chinese Ink; H. High Saturation; I. Low Saturation; J. Negative Space.
Reference answer: B, D, E, I, J.

Subcategory Preference Recognition. Single-choice questions specify a scenario and a preference dimension. They assess whether agents can identify the preferred attribute value and associate it with the appropriate scenario.

Example: "In Group G001, what time of day did the user prefer for the bedroom scene? Select one."
Options: A. Morning; B. Afternoon; C. Dusk; D. Night.
Reference answer: C.

Personal Preference Recognition. Short-answer and single-choice questions ask about favored content, stated reasons, or personal significance within a specified dialogue group. They assess whether agents retain personal information and connect it to the user's expressed preferences.

Example: "In G006, what special meaning does the compass have for the user?"
Reference answer: "The compass reminds the user of their grandfather's guidance and helps them feel a quiet sense of direction when facing decisions."

A.3.2 Q2: Multi-turn Information Integration

Q2 evaluates whether agents can combine evidence from multiple interactions to explain changes in visual outputs.

Consecutive Image Comparison. Short-answer questions compare two consecutive image outputs. They assess whether agents can identify local visual changes and relate these changes to the user's feedback.

Example: "In G001, in what ways is Image 2 closer to what the user wanted compared with Image 1? Please answer briefly."
Reference answer: "The visual style changed from a realistic room-photo look to an abstract low-poly image; the bed, pillows, and nightstand changed from literal detailed furniture to simplified faceted geometric forms."

Initial–Final Image Comparison. Short-answer questions compare the first and final images within a dialogue group. They assess whether agents can summarize the overall visual transformation produced through successive revisions.

Example: "In G001, how did the visual style, color, and time of day change from the initial image to the final image?"
Reference answer: "The image changed from a realistic warm morning bedroom to an abstract low-poly modern minimalist bedroom at dusk with a low-saturation cool gray-blue palette."

A.3.3 Q3: Implicit Preference Inference

Q3 evaluates whether agents can infer recurring preferences from accumulated interactions and apply them to subsequent visual requests.

Similar Generation. Image generation tasks request a new image in a previously encountered scenario without restating the relevant preference constraints. They assess whether agents can determine which preferences apply and realize them in the generated image.

Illustrative request: "Generate another bedroom image that matches my preferences."

Cumulative Global Preference Induction. Multi-select questions ask agents to identify global aesthetic preferences supported by repeated feedback across the observed history. A preference is selected only when the user has emphasized it at least twice. This task assesses the aggregation of recurring evidence across dialogue groups.

Example: "So far, which of the following labels best reflect the user's overall preferences? A label should be selected only if the corresponding preference has been explicitly emphasized by the user at least twice. Select all that apply." The candidate list is omitted for brevity.
Reference labels: Abstract; Negative Space; 3D Rendering/CGI; Modern Minimalism; Cool Tone; Minimal; Cold and Empty; Geometric Rationality; Symmetrical Composition.

Cumulative Subcategory Preference Induction. Short-answer questions ask agents to identify recurring preferences within a specified visual domain. They assess whether agents can consolidate feedback across relevant dialogue groups into explicit attribute–value associations, using the same requirement of at least two supporting expressions.

Example: "In the domain of Home and Everyday Living, please identify the preference items that have been explicitly emphasized by the user at least twice."
Reference answer: "Interior style: modern minimalist; Time of day: night; Tidiness: tidy; Aging level: abandoned; Object density: sparse."

A.3.4 Repeated Anchor Questions

Anchor questions retain the same wording across checkpoints. They assess preference induction and recognition at different levels of detail, allowing us to track acquisition and retention as the interaction history grows.

Global Preference Induction. A fixed multi-select question asks agents to recover the user's overall aesthetic preferences. Repeating the question with the same candidate labels measures how the inferred global preference profile changes across checkpoints.

Example: "So far, which of the following labels best reflect the user's overall preferences? A label should be selected only if the corresponding preference has been explicitly emphasized by the user at least twice. Select all that apply." The candidate list is omitted for brevity.
Reference labels: Abstract; Negative Space; 3D Rendering/CGI; Modern Minimalism; Cool Tone; Minimal; Cold and Empty; Geometric Rationality; Symmetrical Composition; Low Saturation; Black-White-Gray; Low Poly; Leading Lines; Futuristic and Technological; Monochromatic/Analogous Colors; Desolate and Lonely; Digital Painting.

Subcategory Preference Induction. Fixed short-answer questions ask agents to summarize recurring preferences within each visual domain. They assess whether agents acquire and retain domain-relevant preference associations as additional scenarios are introduced.

Example: "In the domain of Fantasy, Science Fiction, and Stylized Worlds, please identify the preference items that have been explicitly emphasized by the user at least twice."
Reference answer: "Magic intensity: mid-magic hybrid; World reference: European medieval; Future vision: grey realism; Tech style: cyberpunk."

Personal Preference Induction. Short-answer questions ask agents to consolidate personal content preferences across conversations. Single-choice items within this subtask also assess preferred assistant response behavior. These questions measure the acquisition and retention of personal knowledge beyond individual dialogue groups.

Example: "What is the user's favorite animal?"
Reference answer: "Cats, especially hairless or fold-ear breeds."

Fine-grained Global Preference Recognition: Multi-select. These questions specify one global preference dimension and ask agents to select all supported values. They assess the completeness and precision of preference recognition within that dimension.

Example: "What preference(s) did the user reveal for realism level? Select all that apply."
Options: A. Abstract; B. Realistic.
Reference answer: A.

Fine-grained Global Preference Recognition: Short-answer. These questions ask agents to state the preferred values for a specified global dimension without candidate options. They assess whether the agent can recover the corresponding preference information without answer cues.

Example: "What preference(s) did the user reveal for rendering style?"
Reference answer: "Digital painting / Low-poly / 3D rendering/CGI."

Fine-grained Subcategory Preference Recognition: Single-choice. These questions specify a visual domain, a subcategory, and a preference dimension, then ask agents to select one preferred value. They assess whether agents preserve the association between a preference and the scenario in which it applies.

Example: "In the domain of Home & Daily Life, what preference did the user reveal for Interior style in Window-Side Rest Nook? Select one."
Options: A. Japanese natural; B. Modern minimalist; C. Mid-century; D. Industrial.
Reference answer: B.

Fine-grained Subcategory Preference Recognition: Short-answer. These questions query the same scenario attributes without candidate options. They assess whether agents can independently recover the preferred value for a particular scenario and retain it across checkpoints.

Example: "In the domain of Home & Daily Life, what preference did the user reveal for Interior style in Window-Side Rest Nook?"
Reference answer: "Modern minimalist."

B Baseline Details

We evaluate five agent configurations with twelve memory configurations, including seven textual and five multimodal settings. This combination allows us to examine how agent capabilities and memory mechanisms affect preference understanding and its application to subsequent visual tasks. Each configuration processes the same interaction histories in chronological order. Memory is maintained across sections for each persona and reset between personas. The underlying user profiles are used for data construction and are not provided to the evaluated agents.

B.1 Agent Configurations

Unified agents. We use BAGEL-7B-MoT (Deng et al., 2025) and Janus-Pro-7B (Chen et al., 2025), which support multimodal understanding and image generation within a unified model. These agents are evaluated on both understanding preferences from prior interactions and producing visual outputs that reflect those preferences.

Modular agents. We evaluate three configurations that combine an understanding model with a separate image model: Gemini 3.6 Flash (Google DeepMind, 2026) with Gemini 3 Pro Image (Raisinghani, 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025) with Qwen-Image (Wu et al., 2025a), and GPT-5.4 (OpenAI, 2026a) with GPT Image 2 (OpenAI, 2026b). The understanding component interprets the current request and the available memory, while the image component produces the visual output. Evaluating both components enables us to examine whether preference understanding is accompanied by consistent visual behavior.

B.2 Textual Memory Configurations

Textual memory configurations represent historical information in text. They differ in how they retain individual interactions, summarize feedback, connect related observations, and update stored information. We include the following seven settings to assess how these mechanisms support preference retrieval and inference across an extended interaction history.

Full (Text). This setting provides the available textual history directly as context at each evaluation checkpoint. It serves as a reference for accessing prior requests, responses, and feedback through the dialogue history, without a separate retrieval or memory consolidation stage.

NaiveRAG. NaiveRAG encodes textual history entries into embedding vectors and retrieves entries according to their similarity to the current query. The retrieved entries are included in the agent's input context for response generation.

Generative Agents. Generative Agents (Park et al., 2023), abbreviated as Gen. Agent in our tables, store observations as textual memories and generate broader reflections from accumulated experiences. Both observations and reflections are retained in memory, with relevant content retrieved to inform subsequent responses.

Reflexion. Reflexion (Shinn et al., 2023) converts feedback from prior outcomes into textual reflections. These reflections are stored in episodic memory and incorporated into subsequent prompts, allowing later decisions to draw on feedback from previous attempts.

A-Mem. A-Mem (Xu et al., 2025) organizes experiences into memory notes containing contextual descriptions, keywords, and semantic links. As new observations arrive, it creates corresponding notes, establishes connections with related memories, and updates the contextual representations of existing notes. The resulting memory structure supports retrieval of relevant information and its associated context.

MemoryOS. MemoryOS (Kang et al., 2025) organizes conversational information into three memory levels for recent interactions, intermediate information, and persistent user knowledge. It updates and consolidates information across these levels, then retrieves relevant content from the memory stores for subsequent responses.

Mem0. Mem0 (Chhikara et al., 2025) extracts salient facts from conversations and integrates them with existing memories. Stored information is updated as new conversational evidence becomes available. For each query, the method retrieves relevant facts and supplies them as context for response generation.

B.3 Multimodal Memory Configurations

Multimodal memory configurations retain visual information alongside text. They differ in how multimodal interactions are stored, organized, and accessed.

Full (MM). Full (MM) provides the available dialogue and associated images directly as context at each evaluation checkpoint. The agent accesses prior multimodal interactions through this history without a separate retrieval stage.

MuRAG. MuRAG (Chen et al., 2022) maintains an external memory index containing textual and visual representations. Given a query, it retrieves relevant text and image evidence from the index and conditions response generation on the retrieved content.

UniversalRAG. UniversalRAG (Yeo et al., 2026) routes queries to appropriate knowledge sources according to the required modality and granularity. It then retrieves relevant evidence from the selected sources and incorporates this evidence into the generation context.

NGM. Neural Graph Memory (Fisher, 2025) organizes multimodal episodes as graph nodes containing content representations and temporal information. Retrieval uses both embedding similarity and graph connections to identify relevant episodes and access related experiences.

AUGUSTUS. AUGUSTUS (Jain et al., 2025) constructs a graph memory that links semantic concepts and tags to their supporting multimodal context. It retrieves relevant experiences through concepts associated with the query and their connections to stored textual and visual information.

C Implementation Details

This section provides the implementation details of our evaluation. We describe the experimental environment, the settings shared across all agent and memory configurations, and the metrics and prompts used to evaluate preference understanding and visual generation.

C.1 Experimental Environment

We use NVIDIA RTX A6000 GPUs for local model inference. Models provided through hosted services are accessed via their corresponding APIs. The evaluated agent configurations and memory methods are described in Appendix B. All methods that rely on embeddings for memory storage and retrieval use GME-Qwen2-VL-2B-Instruct (Zhang et al., 2025) as the embedding model, ensuring a consistent representation across methods. The default number of retrieved memory entries is set to $K = 10$ for methods with a retrieval component. All remaining hyperparameters follow the configurations provided in the respective original implementations.

C.2 Evaluation Settings

For each agent configuration, we compare memory methods using the same interaction histories, evaluation questions, and checkpoints. The agent configuration is held fixed within each comparison, and the same evaluation procedure is applied across all configurations. Each combination of agent and memory method processes the interaction history of each user in chronological order. Memory is maintained across sections for the same user and reset before processing another user. At each checkpoint, only interactions observed up to that point are available to the agent. The underlying user profiles are used for data construction and are never provided to the evaluated agents.

We administer cumulative memory questions at their designated checkpoints and repeat the same anchor questions after each section. Preference understanding and visual generation are evaluated using the same preceding interaction history. For visual generation, the agent uses the current request and available memory to determine the relevant user preferences. The resulting images are evaluated for consistency with preferences supported by the preceding interactions. We use GPT-5.4 for *LLM-as-a-Judge* evaluation in all experiments. We verify its scoring consistency with other LLM judges in Appendix E.2.

C.3 Evaluation Metrics

All scores are reported on a scale from 0 to 100. Single-choice questions are evaluated using exact-match accuracy, assigning 100 points only when the predicted option matches the reference answer. For multi-select questions, we compute the set-level F1 score as $S_i^{\mathrm{MS}} = 100 \cdot \frac{2|\hat{Y}_i \cap Y_i|}{|\hat{Y}_i| + |Y_i|}$, with a score of 100 when both sets are empty. Short-answer questions and image generation and editing tasks are evaluated using model-based judges with K_i equally weighted scoring points. Short-answer scoring points receive credit when the required information is explicitly stated, while image scoring points receive credit when the corresponding visual requirements are clearly satisfied. Their scores are computed as $S_i = 100 \cdot m_i / K_i$, where m_i is the number of satisfied scoring points. The complete prompts used for short-answer and image evaluation are provided below.

Prompt for Short-Answer Evaluation

```
You are an answer grader. Evaluate the model answer according to the question type
and the provided scoring points.

APPLICABLE QUESTION TYPES

- Visual Retrieval
- Instruction Retrieval
- Personal Preference Recognition when the question is open ended
- Consecutive State Comparison
- Initial-Final State Comparison
- Cumulative Subcategory Preference Induction

SCORING RULES

1. Each question is worth 100 points.

2. If a question contains N scoring points, each scoring point is worth 100 / N points.

3. Evaluate each scoring point independently as either fully correct or incorrect.

4. Exact wording is not required. Accept clear semantic equivalents, synonyms, and
concise paraphrases.

5. Award credit only when the model answer explicitly states the required information.
Do not infer omitted details from context or common knowledge.

6. Ignore irrelevant additional content.

7. For comparison questions, the answer must correctly describe the direction of
change from before to after. Mentioning only the final state is insufficient.

8. For Instruction Retrieval questions, require the requested action and target. Require
a location or attribute only when it is included in the corresponding scoring point.

9. For Visual Retrieval questions, attributes such as number, color, material, condition,
and location must be correct when included in the corresponding scoring point.

OUTPUT FORMAT

Evaluate every scoring point independently. Return only a JSON object using
the following schema:

{"matched_point_ids": ["1", "2"]}

Include an ID only when the corresponding scoring point is fully
satisfied. Do not provide explanations or any additional text.

QUESTION

{question}

MODEL ANSWER

{system_answer}

SCORING POINTS
ID -> requirement

{scoring_points}
```

Prompt for Image Generation and Editing Evaluation

```
You are an image evaluator. Evaluate the model output image according to
the current request, applicable reference preferences, and provided scoring points.

APPLICABLE TASK TYPES

- Image Generation
- Image Editing

SCORING RULES

1. Each task is worth 100 points.
```

```
2. If there are N scoring points, each scoring point is worth 100 / N points.

3. Evaluate each scoring point independently as either fully satisfied or not satisfied.

4. Award credit only when the required property is clearly visible in the output image.

5. Do not infer unclear, hidden, or absent visual details from the request,
reference preferences, or common knowledge.

6. Evaluate only the requirements specified by the scoring points. Do not introduce
additional preferences or penalize unrelated image properties.

7. For global visual preferences, assess the relevant rendering  style, color system,
atmosphere, composition, and realism level.

8. For scenario-specific preferences, apply only the preferences associated with
the scenario in the current request. Do not transfer preferences from other scenarios.

9. For image editing tasks, compare the edited image with the source image.
Verify that the requested changes are correctly applied and that elements
required to remain unchanged are preserved.

10. For image generation tasks, evaluate only the generated output image against
the current request and scoring points.

IMAGE ORDER

For image generation tasks, the attached image is the model output.

For image editing tasks, the first attached image is the source image,
and the second attached image is the edited output.

OUTPUT FORMAT

Evaluate every scoring point independently. Return only a JSON object
using the following schema:

{"matched_point_ids": ["1", "2"]}

Include an ID only when the corresponding scoring point is fully satisfied.
Do not provide explanations or any additional text.

CURRENT REQUEST

{question}

APPLICABLE REFERENCE PREFERENCES

{reference_preferences}

SCORING POINTS
ID -> requirement

{scoring_points}
```

D Detailed Experimental Results

This section presents detailed experimental results on Pref2Act. We first report fine-grained performance across all agents and textual and multimodal memory configurations, and then examine the knowing–doing gap between preference understanding and preference-consistent image generation.

D.1 Fine-grained Results Across Memory Configurations

The following tables present the complete fine-grained results across all textual and multimodal memory configurations for BAGEL-7B-MoT, Janus-Pro-7B, Gemini 3.6 Flash with Gemini 3 Pro Image, Qwen3-VL-8B-Instruct with Qwen-Image, and GPT-5.4 with GPT Image 2.

Table 7. Fine-grained results of Janus-Pro-7B on PREF2ACT. F.g. denotes fine-grained. ◉, ☑, ☰, and 🖼 denote single-choice, multi-select, short-answer, and image generation, respectively. **Red** and **blue** highlight the best and second-best results in each row, respectively.

			Textual Memory							Multimodal Memory				
Task	**Subtask**	**Type**	*Full (Text)*	*NaiveRAG*	*Gen. Agent*	*Reflexion*	*A-Mem*	*MemoryOS*	*Mem0*	*Full (MM)*	*MuRAG*	*UniversalRAG*	*NGM*	*AUGUSTUS*
Q1	Visual Retrieval	☰	14.23	38.80	33.75	15.89	29.01	**56.09**	**44.58**	14.90	36.98	28.91	33.49	30.07
	Instruction Retrieval	☰	**28.96**	12.40	16.77	26.14	18.54	**36.78**	**28.96**	15.42	26.25	28.54	21.35	26.67
	Global Pref. Recognition	☑	29.41	31.25	30.70	24.75	25.54	29.62	30.74	30.39	**33.15**	34.61	**35.11**	22.64
	Scenario Pref. Recognition	◉	22.15	64.10	52.01	25.15	66.46	61.33	75.97	49.31	**77.85**	**76.11**	74.72	75.00
	Personal Pref. Recognition	◉☰	18.82	56.08	42.23	25.24	52.44	60.76	**67.88**	40.75	**64.75**	60.28	55.31	52.05
Q2	Consecutive Image Comparison	☰	18.23	19.42	9.82	11.61	4.95	8.28	**22.53**	18.45	**24.32**	12.15	20.88	9.14
	Initial-Final Image Comparison	☰	9.70	14.15	15.93	11.69	8.13	11.24	**34.46**	20.62	**29.21**	13.74	20.26	22.61
Q3	Cum. Global Pref. Induction	☑	**12.20**	4.77	5.56	8.36	9.86	5.01	**27.15**	9.77	4.53	7.64	5.86	9.73
	Cum. Scenario Pref. Induction	☰	32.24	20.83	23.75	25.41	26.04	12.54	39.52	18.75	**43.54**	**48.75**	37.92	36.67
	Similar Generation	🖼	48.62	54.37	51.28	49.43	55.61	47.16	57.24	59.18	**64.73**	**63.92**	58.27	59.46
Anchor	Global Pref. Induction	☑	16.33	**17.26**	13.17	9.90	5.64	5.81	**29.78**	1.32	5.30	3.22	11.06	5.56
			-6.82	**9.12**	7.26	7.48	2.15	-15.39	1.74	-3.51	2.53	5.56	3.92	**8.34**
	Scenario Pref. Induction	☰	13.02	27.92	15.97	13.19	23.26	9.56	21.77	9.27	29.79	14.79	**48.33**	**33.75**
			0.66	**25.55**	22.22	2.27	6.48	9.06	23.80	21.94	**28.15**	-14.44	2.78	-13.19
	Personal Pref. Induction	☰	13.15	27.88	9.92	18.72	27.68	**30.24**	**46.43**	16.44	11.48	19.27	17.49	19.71
			8.60	-7.41	3.70	-1.74	-6.08	**13.49**	5.56	0.80	-6.28	-11.11	-4.17	-3.24
	F.g. Global Pref. Recognition	☑	42.28	**56.02**	44.49	40.90	51.29	35.32	53.36	48.67	52.02	52.00	53.24	**60.45**
			-18.79	0.92	2.12	-10.32	-9.62	**7.41**	**5.65**	0.54	2.13	-7.40	5.27	-2.24
	F.g. Global Pref. Recognition	☰	42.89	43.58	16.49	25.06	30.38	33.16	**57.46**	21.70	**49.65**	27.43	27.60	28.47
			-19.44	-2.78	6.02	-16.22	4.63	**6.60**	**10.19**	-2.78	-10.18	0.93	-1.86	-10.42
	F.g. Scenario Pref. Recognition	◉	35.61	54.47	45.06	31.50	56.57	45.50	**63.92**	41.03	58.37	**58.59**	52.61	52.37
			0.08	**6.14**	-1.07	-11.20	2.45	**14.56**	1.62	-0.84	3.89	5.87	6.06	2.96
	F.g. Scenario Pref. Recognition	☰	15.88	**47.13**	21.02	21.60	29.85	**52.27**	46.58	29.27	34.44	33.36	28.39	25.93
			-8.70	**11.19**	4.12	-6.06	2.57	**13.57**	-2.74	2.53	9.83	2.19	-3.97	0.85

Table 8. Fine-grained results of BAGEL-7B-MoT on PREF2ACT. F.g. denotes fine-grained. ◉, ☑, ☰, and 🖼 denote single-choice, multi-select, short-answer, and image generation, respectively. **Red** and **blue** highlight the best and second-best results in each row, respectively.

			Textual Memory							Multimodal Memory				
Task	**Subtask**	**Type**	*Full (Text)*	*NaiveRAG*	*Gen. Agent*	*Reflexion*	*A-Mem*	*MemoryOS*	*Mem0*	*Full (MM)*	*MuRAG*	*UniversalRAG*	*NGM*	*AUGUSTUS*
Q1	Visual Retrieval	☰	16.82	42.50	35.01	24.70	**49.29**	**66.30**	45.94	19.88	41.15	38.99	38.45	40.68
	Instruction Retrieval	☰	29.40	11.98	10.83	32.28	19.29	**37.34**	**35.73**	13.89	31.25	25.83	12.14	22.19
	Global Pref. Recognition	☑	**68.74**	46.35	28.83	45.97	49.09	40.40	36.96	44.38	47.32	**55.06**	39.81	44.34
	Scenario Pref. Recognition	◉	25.64	72.29	68.33	49.05	79.05	71.00	79.65	55.87	**88.68**	**84.92**	83.33	81.67
	Personal Pref. Recognition	◉☰	22.35	69.55	49.56	39.59	69.76	**72.15**	70.78	42.44	**78.47**	71.36	68.11	66.37
Q2	Consecutive Image Comparison	☰	**29.50**	28.25	19.81	19.90	12.64	13.40	**31.09**	26.97	24.37	18.21	25.55	22.87
	Initial-Final Image Comparison	☰	17.30	41.38	31.31	26.61	23.96	20.06	32.99	32.43	**45.55**	27.48	38.00	**42.00**
Q3	Cum. Global Pref. Induction	☑	23.91	19.43	19.50	2.04	25.73	9.82	28.14	24.34	**30.39**	22.73	**39.98**	8.33
	Cum. Scenario Pref. Induction	☰	43.66	51.04	43.33	41.91	28.33	16.99	26.46	31.19	**60.21**	**72.38**	29.76	27.92
	Similar Generation	🖼	63.86	68.08	65.62	62.73	69.29	61.37	69.71	71.43	**78.16**	**81.02**	73.61	76.84
Anchor	Global Pref. Induction	☑	12.92	18.47	20.06	16.92	**25.79**	4.60	**26.32**	25.16	12.07	9.94	9.39	13.74
			-9.82	1.03	-3.46	-1.96	**16.38**	-18.40	-10.28	**4.52**	-16.09	-17.39	2.39	-1.88
	Scenario Pref. Induction	☰	17.65	33.89	45.56	23.02	23.49	12.96	17.61	**45.79**	34.06	34.09	24.60	**50.00**
			-11.17	-0.56	-5.29	-2.22	10.84	-2.76	20.46	**29.45**	-1.02	-14.59	**23.61**	-45.37
	Personal Pref. Induction	☰	13.45	29.56	17.73	22.90	19.73	30.95	**47.52**	21.32	**36.21**	26.53	19.73	24.70
			14.81	3.53	11.11	-3.57	6.35	**19.70**	3.17	6.35	-16.14	-13.10	-2.78	1.85
	F.g. Global Pref. Recognition	☑	46.53	57.00	41.62	48.99	61.47	38.87	65.73	48.20	65.27	**67.66**	**67.57**	59.16
			-15.24	7.63	3.89	-6.13	5.85	**10.96**	-4.08	2.92	6.46	-0.59	**9.72**	-0.37
	F.g. Global Pref. Recognition	☰	47.13	42.19	26.62	43.05	28.97	36.44	**64.76**	40.68	**58.51**	44.05	40.87	34.03
			-27.31	10.65	5.56	-0.69	-20.83	-1.27	-3.24	11.80	**14.82**	11.11	**12.50**	-9.72
	F.g. Scenario Pref. Recognition	◉	35.93	53.52	41.22	42.77	60.10	45.91	**67.14**	39.81	60.50	**63.86**	56.77	58.49
			1.45	5.84	-2.99	-11.97	6.22	**15.93**	5.54	-11.22	10.24	11.38	**13.95**	6.88
	F.g. Scenario Pref. Recognition	☰	21.80	51.07	47.99	41.33	46.81	**71.73**	52.55	44.13	48.51	**53.80**	53.30	49.53
			-12.69	10.73	-26.86	-3.59	4.05	9.58	11.25	-3.74	**17.73**	**13.35**	7.92	5.73

Table 9. Fine-grained results of Gemini 3.6 Flash on PREF2ACT. F.g. denotes fine-grained. ◉, ☑, ☰, and 🖼 denote single-choice, multi-select, short-answer, and image generation, respectively. **Red** and **blue** highlight the best and second-best results in each row, respectively.

			Textual Memory							Multimodal Memory				
Task	**Subtask**	**Type**	*Full (Text)*	*NaiveRAG*	*Gen. Agent*	*Reflexion*	*A-Mem*	*MemoryOS*	*Mem0*	*Full (MM)*	*MuRAG*	*UniversalRAG*	*NGM*	*AUGUSTUS*
Q1	Visual Retrieval	☰	13.01	**41.09**	33.02	14.74	40.31	**51.29**	19.21	7.92	34.94	40.52	35.26	38.18
	Instruction Retrieval	☰	25.43	7.81	12.01	21.62	21.04	**32.30**	26.46	12.81	11.43	**29.17**	17.60	23.33
	Global Pref. Recognition	☑	**46.99**	25.43	28.75	16.76	26.98	27.62	29.05	28.15	37.29	40.92	**44.59**	32.29
	Scenario Pref. Recognition	◉	22.94	70.21	49.65	19.65	77.64	63.52	70.28	26.81	**91.19**	**93.06**	82.64	74.45
	Personal Pref. Recognition	◉☰	19.87	**71.00**	38.78	19.10	69.78	64.15	51.27	24.16	64.12	**70.09**	62.13	65.53
Q2	Consecutive Image Comparison	☰	**14.00**	10.28	**18.62**	10.89	9.33	6.36	5.34	7.11	6.32	8.28	6.30	8.04
	Initial-Final Image Comparison	☰	10.21	**23.79**	20.14	11.55	**23.09**	11.83	9.47	7.90	10.47	18.97	16.17	18.74
Q3	Cum. Global Pref. Induction	☑	20.90	22.87	20.53	23.54	15.90	8.58	21.82	22.07	**28.31**	**33.55**	27.69	18.34
	Cum. Scenario Pref. Induction	☰	22.38	**27.92**	21.25	**24.79**	8.33	8.71	18.96	10.83	17.86	17.50	8.33	20.63
	Similar Generation	🖼	73.29	71.84	70.16	68.43	74.31	68.92	73.86	77.63	**83.94**	**85.57**	80.13	80.68
Anchor	Global Pref. Induction	☑	17.80	21.57	16.32	19.57	15.30	6.33	18.60	15.91	**22.60**	**26.75**	21.86	17.25
			-20.70	-6.61	-16.72	-12.69	-3.68	-29.28	-10.41	-7.17	**3.89**	**10.21**	-2.94	-3.51
	Scenario Pref. Induction	☰	11.17	17.60	16.15	8.09	19.31	8.20	**23.23**	6.04	19.72	**20.31**	12.19	11.35
			-12.49	-3.24	**7.50**	-8.89	-4.26	-4.08	**25.46**	-1.30	-11.81	-0.46	-7.87	-6.39
	Personal Pref. Induction	☰	17.57	36.51	29.86	27.18	40.18	**40.42**	**43.06**	23.71	39.57	38.68	26.09	35.15
			22.22	20.63	-1.85	3.70	5.29	**27.11**	**26.19**	12.70	-0.01	6.58	13.23	23.81
	F.g. Global Pref. Recognition	☑	57.83	75.62	56.20	55.18	79.17	48.31	69.66	65.29	**83.35**	**84.76**	78.47	72.50
			-17.17	-1.43	-7.30	-11.33	**10.26**	9.03	4.01	-3.47	**14.03**	8.89	-3.33	4.82
	F.g. Global Pref. Recognition	☰	53.16	46.53	29.69	26.39	47.74	41.10	**59.37**	23.61	**64.09**	44.10	38.37	34.90
			-24.08	3.70	-9.72	-19.91	-0.93	1.96	**6.48**	-7.87	**8.33**	-20.84	-1.39	-7.41
	F.g. Scenario Pref. Recognition	◉	35.60	58.23	40.05	27.92	**64.35**	45.49	57.34	34.46	63.76	**64.84**	58.20	63.22
			7.18	7.69	-4.92	-4.78	15.76	**21.67**	**19.03**	-10.52	13.12	13.96	12.51	11.01
	F.g. Scenario Pref. Recognition	☰	17.45	34.13	35.40	24.09	**47.39**	**57.44**	41.98	16.82	39.35	46.99	34.72	42.62
			-6.14	5.08	-4.68	-0.71	7.16	16.13	**17.83**	-6.18	**18.51**	11.73	10.14	0.54

Table 10. Fine-grained results of GPT-5.4 on PREF2ACT. F.g. denotes fine-grained. ◉, ☑, ☰, and 🖼 denote single-choice, multi-select, short-answer, and image generation, respectively. **Red** and **blue** highlight the best and second-best results in each row, respectively.

			Textual Memory							Multimodal Memory				
Task	**Subtask**	**Type**	*Full (Text)*	*NaiveRAG*	*Gen. Agent*	*Reflexion*	*A-Mem*	*MemoryOS*	*Mem0*	*Full (MM)*	*MuRAG*	*UniversalRAG*	*NGM*	*AUGUSTUS*
Q1	Visual Retrieval	☰	13.00	27.60	25.60	10.94	44.90	**53.80**	45.05	13.49	45.24	**46.33**	37.40	36.04
	Instruction Retrieval	☰	26.00	4.69	7.02	23.44	23.33	**34.38**	**31.60**	14.06	30.00	27.50	17.19	12.71
	Global Pref. Recognition	☑	**51.59**	28.73	29.39	29.97	22.05	38.41	42.37	38.19	**47.25**	36.40	39.60	33.29
	Scenario Pref. Recognition	◉	21.44	67.71	49.68	20.28	74.31	74.38	77.85	25.90	**86.19**	**88.54**	78.33	70.49
	Personal Pref. Recognition	◉☰	21.06	66.94	30.90	20.88	**70.70**	67.24	62.29	34.26	**68.35**	68.08	63.42	60.80
Q2	Consecutive Image Comparison	☰	12.00	4.84	9.45	7.05	10.21	9.37	**17.94**	**13.59**	9.99	8.57	5.19	4.90
	Initial-Final Image Comparison	☰	8.00	7.57	10.72	8.36	**18.29**	16.31	**25.90**	14.86	11.07	11.57	8.43	12.47
Q3	Cum. Global Pref. Induction	☑	**54.04**	51.70	46.38	24.12	30.58	34.33	33.56	41.60	42.10	32.79	**54.98**	22.28
	Cum. Scenario Pref. Induction	☰	31.00	28.96	**43.10**	18.75	33.96	32.50	**36.67**	24.58	33.10	15.42	17.50	9.79
	Similar Generation	🖼	79.06	79.87	78.24	76.58	78.61	72.73	80.76	84.09	**89.92**	**88.64**	84.46	82.83
Anchor	Global Pref. Induction	☑	48.77	44.49	43.30	18.82	45.78	18.04	27.17	35.16	**52.95**	47.51	**50.16**	36.18
			-32.26	-7.28	-19.78	-3.89	-11.93	-29.34	0.24	-26.08	-12.61	**8.18**	**31.39**	-2.70
	Scenario Pref. Induction	☰	15.83	**30.70**	13.77	10.76	**40.90**	19.30	26.08	11.25	26.35	19.76	18.44	19.76
			1.25	14.82	**14.91**	-8.15	-1.30	0.56	**24.17**	4.32	-16.66	7.87	2.69	-15.28
	Personal Pref. Induction	☰	21.59	34.33	16.33	21.03	31.45	**43.45**	**52.78**	24.21	32.65	33.34	23.22	14.88
			7.14	**11.64**	-14.81	5.03	9.53	0.26	-1.32	4.76	-9.13	**11.11**	10.85	9.52
	F.g. Global Pref. Recognition	☑	56.36	**78.70**	57.68	53.61	75.57	67.75	73.98	64.54	**80.25**	78.10	77.30	73.55
			-18.48	1.49	-16.56	-11.26	7.85	**10.23**	2.66	-3.12	4.23	7.71	2.54	**8.11**
	F.g. Global Pref. Recognition	☰	50.00	53.82	34.52	27.95	50.35	45.49	**58.16**	40.63	**63.89**	54.17	43.40	40.97
			-16.67	-0.93	-25.92	-27.32	**7.41**	-4.17	**7.87**	-13.89	-11.80	-9.26	5.56	-9.26
	F.g. Scenario Pref. Recognition	◉	32.16	55.30	35.94	35.17	60.59	45.76	**66.97**	39.49	62.67	**63.95**	59.96	59.89
			-3.96	8.31	-10.72	-7.97	8.96	**17.62**	10.69	-4.33	11.32	**13.09**	5.59	9.18
	F.g. Scenario Pref. Recognition	☰	20.73	30.00	18.30	18.35	45.13	**48.89**	**59.31**	21.32	40.88	44.85	30.67	31.65
			-2.18	**12.03**	-6.61	-5.18	8.47	9.53	**11.54**	-5.23	10.20	9.52	5.45	3.83

D.2 Knowing–Doing Gap in Image Generation

Figure 7 examines the relationship between preference understanding and image generation across agents and memory configurations. Panel (a) reports image generation performance, while Panel (b) reports the proportion of failed generations for which the applicable preferences were correctly understood. Across all configurations, this proportion ranges from 8.36% to 39.74%. The gap remains evident even in configurations with strong generation performance. For example, Qwen3-VL-8B-Instruct with MuRAG achieves a generation score of 83.07, yet 39.74% of its failed generations occur despite correct preference understanding. Similar patterns appear for BAGEL-7B-MoT with UniversalRAG, Gemini 3.6 Flash with UniversalRAG, and GPT-5.4 with MuRAG, which obtain generation scores of 81.02, 85.57, and 89.92, respectively, while their corresponding proportions are 39.62%, 37.78%, and 36.02%. These results show that stronger generation performance does not eliminate failures in translating correctly understood preferences into visual outputs, further demonstrating the knowing–doing gap.

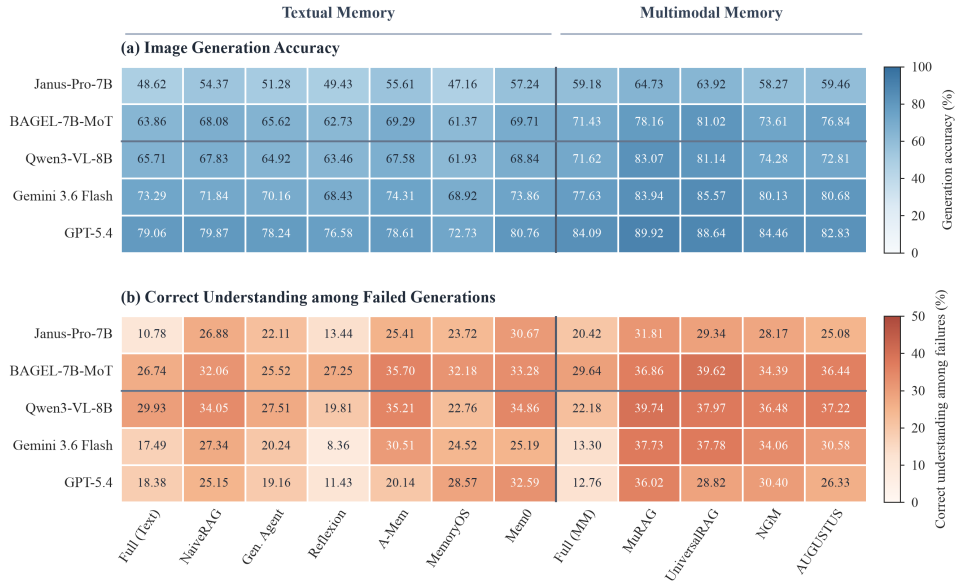


Figure 7. Image generation accuracy and correct preference understanding among failed generations across agents and memory configurations.

E Benchmark Reliability

This section evaluates the reliability of Pref2Act. We assess whether the evaluation questions are answerable when annotated supporting evidence is provided and measure the agreement between GPT-5.4, alternative evaluation models, and human evaluators.

E.1 Question Answerability

To assess whether the evaluation questions in Pref2Act are answerable, we evaluate all five agent configurations and three human participants using the annotated reference evidence. In the agent setting, we provide this evidence directly as memory alongside each question. This bypasses memory retrieval and assesses the agents' ability to answer questions given the available evidence. In the human setting, all three participants independently answer the same questions using the same evidence. Neither agents nor human participants receive the reference answers, scoring rubrics, or hidden user profiles.

We evaluate both agent and human responses using the same scoring criteria as in the other experiments. Table 11 reports the results for each question. Both agents and human participants achieve high scores on Q1 and Q2. Agent performance is slightly lower on Q3, where supporting evidence is distributed throughout an entire section, requiring agents to identify and integrate relevant preference cues across multiple memory entries. Human scores nevertheless remain above 90%. These results

support the answerability of the benchmark questions and suggest that the lower performance in the main evaluation primarily reflects limitations of current agents and memory methods in continual preference memory.

Table 11. Question answerability with annotated supporting evidence. All five agents and three human participants answer the same questions using the same evidence sets.

TaskSubtask		Agents with Annotated Clues					Human
		BAGEL-7B-MoT	Janus-Pro-7B	Gemini 3.6 Flash	Qwen3-VL-8B-Instruct	GPT-5.4	
Q1	Visual retrieval	100.00	100.00	100.00	100.00	100.00	100.00
	Instruction retrieval	100.00	100.00	100.00	100.00	100.00	100.00
	Global preference recognition	97.86	96.74	98.63	97.92	100.00	100.00
	Subcategory preference recognition	95.73	94.86	98.04	96.82	98.27	98.12
	Personal preference recognition	97.18	95.89	98.52	97.36	98.74	99.43
Q2	Consecutive image comparison	95.61	94.72	97.46	96.18	97.83	95.54
	Initial–final image comparison	94.83	95.18	97.12	95.84	97.61	96.86
Q3	Cumulative global preference induction	89.63	87.24	93.18	90.76	92.27	94.08
	Cumulative subcategory preference induction	88.42	86.35	92.64	89.83	93.16	96.52

E.2 Agreement Across Evaluation Models

We further use Qwen2.5-VL-72B-Instruct and Claude-Opus-5 to verify the robustness of GPT-5.4 as our primary judge. We randomly sample 30 questionresponse pairs from each of Q1, Q2, and Q3 under the Qwen3-VL-8B-Instruct with Mem0 setting. We also employ three human evaluators to assess each response against the reference answer at the level of individual scoring points, following the same scoring criteria as the model judges.

We measure agreement using the mean absolute error between each model judge and the human reference: $\mathrm{MAE}_j = \frac{1}{N}\sum_{i=1}^{N}\left|s_i^j - s_i^{\mathrm{human}}\right|$, where s_i^j is the score assigned by judge j, and s_i^{human} is the average score from the three human evaluators. As shown in Table 12, the maximum difference across evaluators is 1.88 points for any task category and 1.44 points overall. All model judges achieve an MAE below 4 points. We attribute this agreement in part to the predefined, fine-grained scoring points, which make the evaluation criteria explicit and reduce variability across evaluators. These results show that the evaluation is consistent across judges and that the choice of evaluation model does not drive our main findings.

Table 12. Agreement between model judges and human ratings. We randomly sample 30 questions from each of Q1, Q2, and Q3. MAE is computed against the mean ratings of three human evaluators.

Evaluator	Q1	Q2	Q3	Avg.	MAE ↓
GPT-5.4	52.66	20.07	45.49	39.41	2.31
Qwen2.5-VL-72B	51.31	18.82	44.14	38.09	3.84
Claude Opus 5	53.03	19.54	46.02	39.53	2.57
Human	52.72	19.84	45.31	39.29	—