LAB SEMINAR

MemEye

A Visual-Centric Evaluation Framework for Multimodal Agent Memory

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, Ruixiang Tang

Rutgers · Notre Dame · Princeton · UMN

Presenter **조은찬 (Eunchan Jo)** | Date **July 22** | Venue **arXiv preprint (cs.CV)**

IN ONE SLIDE

The paper in three lines

Problem

Existing multimodal-memory benchmarks can be solved **without looking at the image** — captions, dialogue, or answer options leak the answer.

Idea

A **two-axis framework** — visual granularity (X) × reasoning depth (Y) — plus a benchmark verified to **require** genuine visual evidence.

Result

Across **13 memory methods** (7 text + 6 multimodal) and 4 VLM backbones, **none come close to solving it** .

01

Motivation & Problem

Why do we need yet another memory benchmark?

BACKGROUND

Agent memory is going multimodal

Modern agents interact with users over **many sessions** , not one turn. To stay useful they must **remember past dialogue and the images seen along the way**.

Increasingly, that history is **visual** — photos, screenshots, dashboards — so long-term memory must carry visual information **across sessions**.

→ **Memory research is shifting from text-only to multimodal.**

A TYPICAL SETTING

🗓 **Session 1** · user shares living-room photos

🗓 **Session 2** · sends an updated paint swatch

🗓 **Later** · "which swatch matched the wall?"

needs visual memory across sessions

THE GAP

Old "visual" questions can be answered without the image

In prior benchmarks the answer often leaks from **captions, surrounding dialogue, or answer options** — the model never truly needs to **see**.

Changing visual state is described **in text** , so text memory alone suffices.

Reading Figure 2

Prior benchmarks (LoCoMo, MMRC, Mem-Gallery) show a **small caption-to-multimodal gain**. MemEye shows a **larger gain** — stronger visual irreplaceability.

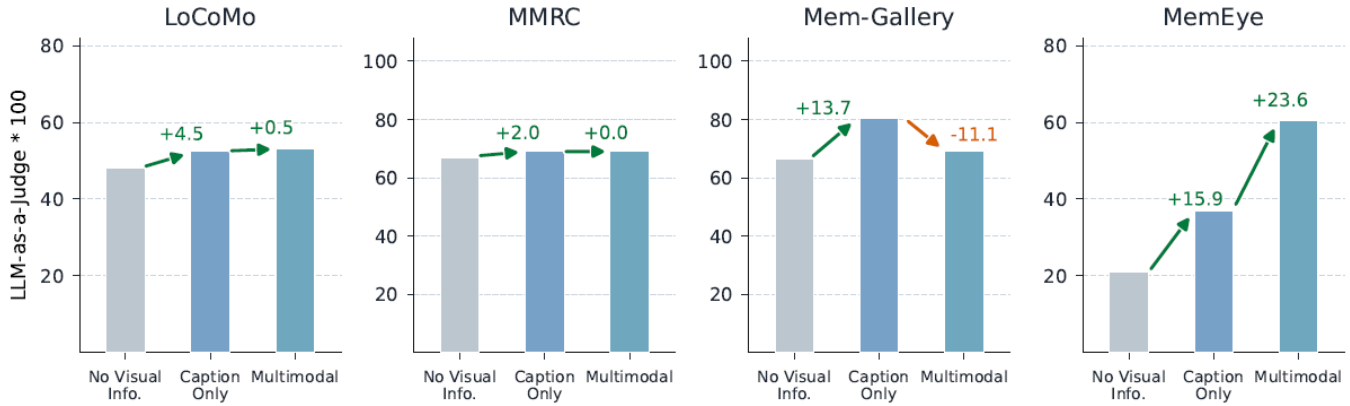


Figure 2: gains from No-Visual → Caption-Only → Multimodal across four benchmarks.

FRAMING

Two open challenges → two axes

1 Fine visual evidence is lost in text

Compressing an image into a caption preserves the gist but drops **fine detail** — exact color, small text, object identity.

↳ **becomes the X-axis: visual granularity**

2 Reasoning over evolving history fails

The same subject is **updated across sessions** ; the agent must pick the **current valid state**, not a stale one.

↳ **becomes the Y-axis: reasoning depth**

02

The MemEye Framework

Two axes, one coordinate for every task — and a benchmark built on top.

Two axes label the "Left Eye" and "Right Eye"

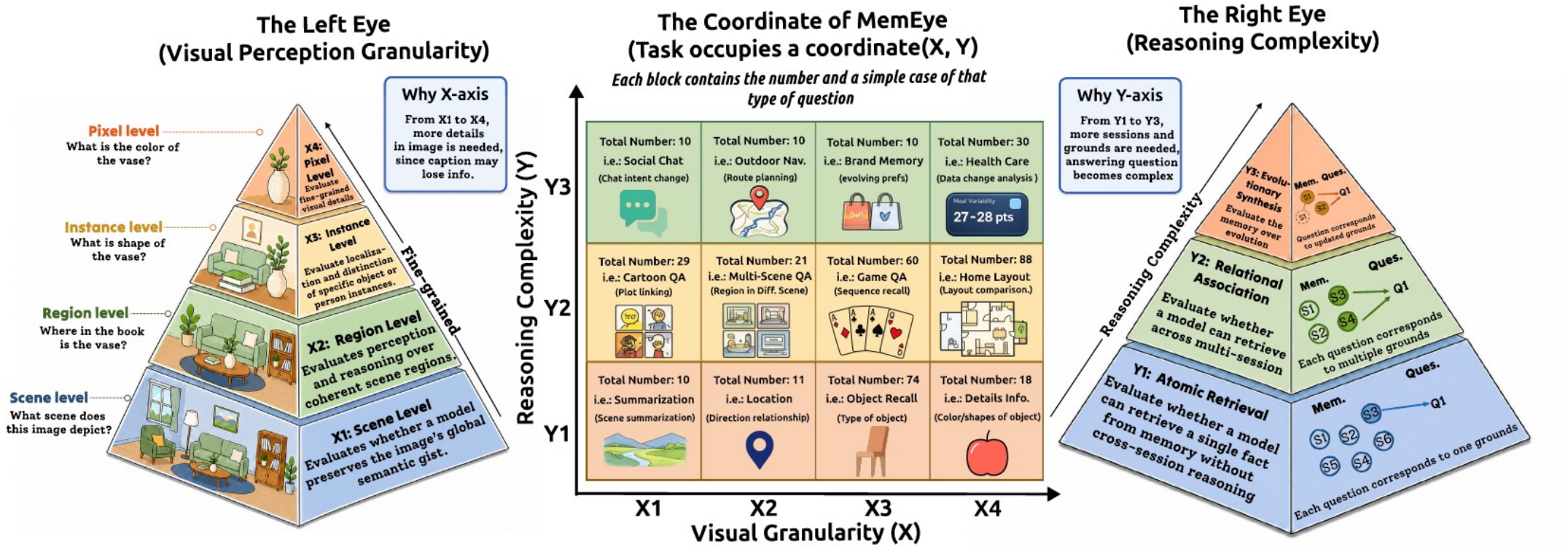


Figure 3: The Left Eye = visual perception granularity (X), the Right Eye = reasoning complexity (Y); every task sits at a coordinate (X, Y).

X-AXIS · THE LEFT EYE

Visual granularity: from scene to pixel

X1 · Scene

Scene-level evidence — scene type, gist. *"What scene does this image depict?"*

X2 · Region

Coherent sub-region. *"Where in the book is the vase?"*

X3 · Instance

A specific object or person instance. *"What is shape of the vase?"*

X4 · Pixel

Fine detail — exact color, small text, texture, count, markings. *"What is the color of the vase?"*

THE KEY IDEA

As X grows, captions lose the answer — the native image is required.

Y-AXIS · THE RIGHT EYE

Reasoning depth: retrieve → associate → synthesize

Y1 · Atomic One sufficient evidence unit answers the question — **no cross-session composition** needed.

Y2 · Relational Multiple **non-conflicting clues** must be linked across sessions, modalities, or references.

Y3 · Evolutionary Clues include **updates, conflicts, overrides**; the answer depends on the **valid current state**, not a stale one.

THE BENCHMARK

371 questions across 8 life scenarios

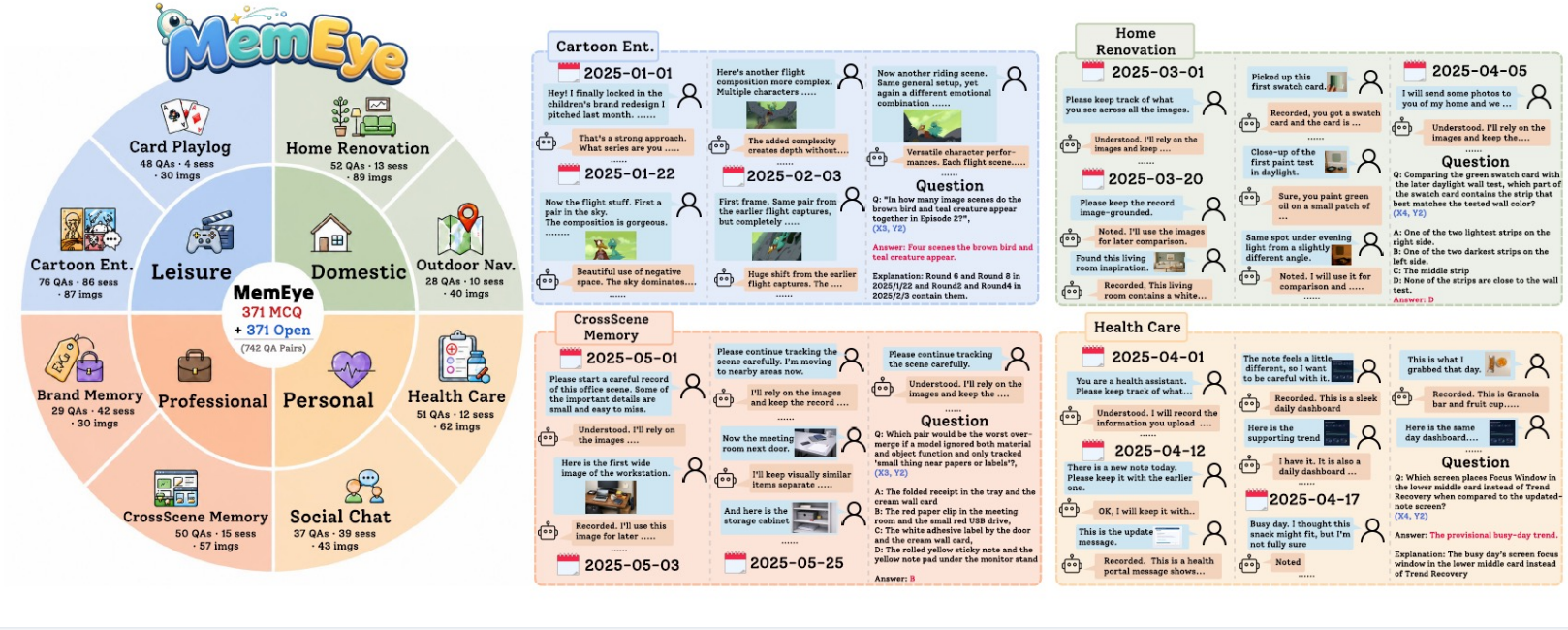


Figure 1: dataset overview and example multi-session dialogues with questions.

371

mirrored questions (MCQ + open-ended)

8
tasks

4
domains

221
sessions

848
rounds

438
images

Every question is annotated with its **(X, Y)** coordinate.

WHAT MAKES IT TRUSTWORTHY

Three gates keep it visual-centric

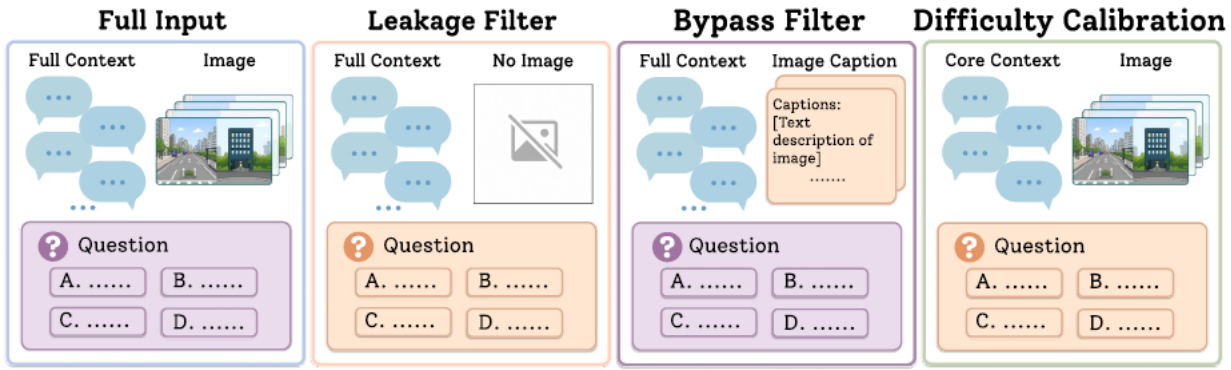


Figure 4: Full Input → Leakage Filter → Bypass Filter → Difficulty Calibration.

1 · Leakage filter

Remove the image. If still answerable → the question doesn't need vision → **drop it**.

2 · Bypass filter

Give a **minimal caption** instead of the image. If the caption suffices → **drop it**.

3 · Difficulty calibration

Give the **image + answer-relevant context** and test if the question is inherently solvable. If it fails, the difficulty is the **base VLM's**, not memory → **drop it**.

03

Experiments & Findings

RQ1 — where do methods fail? RQ2 — why is visual info lost? RQ3 — why is evolving state lost?

EXPERIMENTAL SETUP

13 memory methods × 4 VLM backbones

MEMORY METHODS

7
textual memory

6
multimodal memory

FC, SRAG, Reflexion, Generative-Agent, MemOS, A-Mem, MIRIX, MMA, M2A, SM …

METRICS & DIAGNOSTICS

- **EM** for multiple-choice · **LLM-as-a-Judge** for open-ended
- Judge reliability: **κ = 0.94** vs. human
- Two diagnostics: **Caption-Proof** & **Oracle-Evidence**

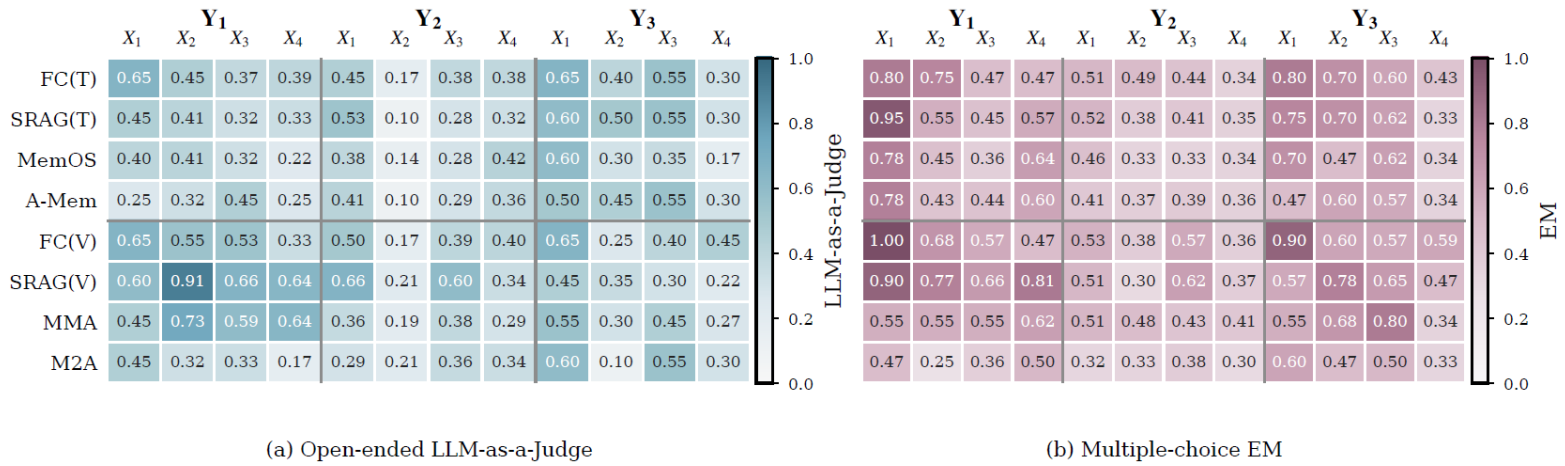
RQ1
Where do methods fail?

RQ2
Why is visual info lost?

RQ3
Why is evolving state lost?

RQ1 · WHERE DO METHODS FAIL?

No method saturates the benchmark



(a) Open-ended LLM-as-a-Judge (b) Multiple-choice EM

Figure 5: representative method performance heatmap. Left: LLM-as-a-Judge; Right: MCQ EM.

Best is still weak

The strongest method, **SRAG(V)** , leads but leaves large gaps everywhere.

Stressor 1 — high X

Scores fall as questions demand **finer visual evidence** .

Stressor 2 — Y3

A bottleneck appears where answers depend on **evolving state**.

Captions suffice at low X, fail at high X

Caption-Proof diagnostic: compare giving a caption vs. the real image (Δ = V − T) at each coordinate.

- **X1–X2:** captions already suffice — image adds little.
- **X3–X4:** the image is decisive — the caption cannot recover fine detail (Δ grows strongly positive).
- Paper: caption-based storage tends to lose **decisive instance- and pixel-level evidence** .

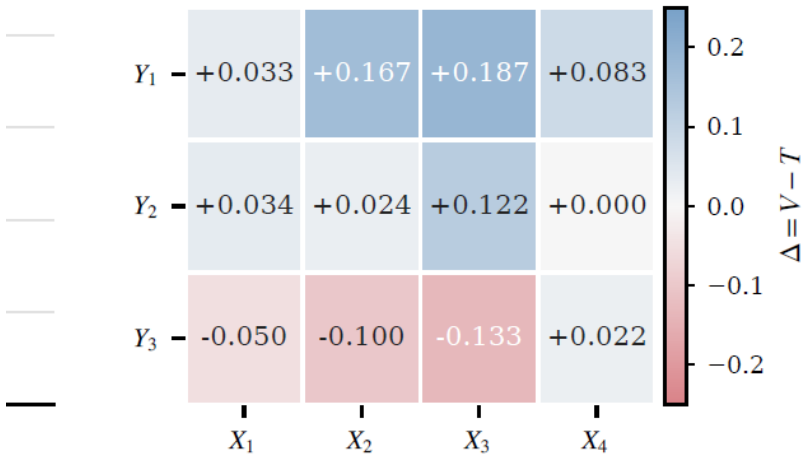


Figure 6b: Caption-Proof gain Δ = V − T over the (X, Y) grid.

RQ3 · WHY IS EVOLVING STATE LOST?

Retrieval finds clues — but picks the stale one

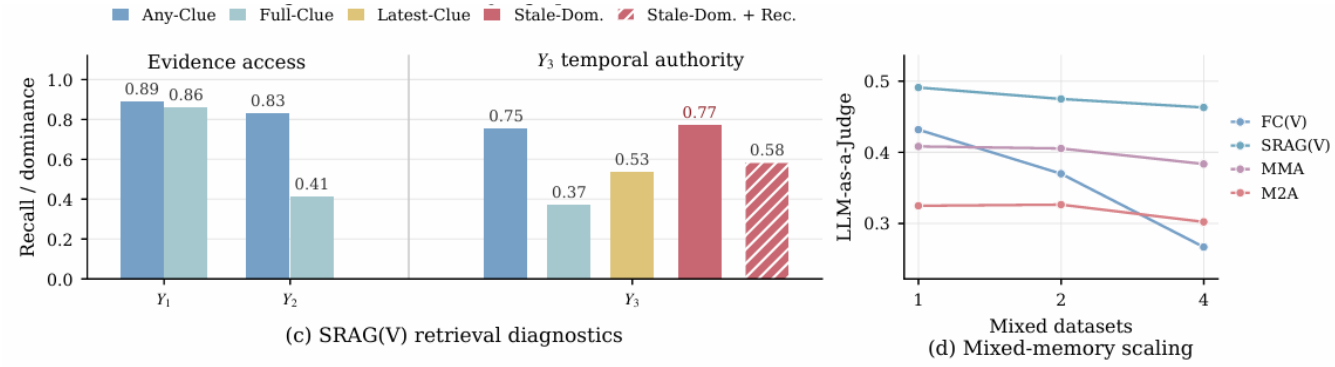


Figure 6c–d: SRAG(V) retrieval diagnostics & cross-topic dialogue scaling.

Access is fine at Y1

The right clue is retrieved (Any-Clue **0.89**).

Y3 picks stale evidence

Stale-dominant **0.77** > latest-clue **0.53** — the system favors outdated state.

The lesson

Semantic similarity ≠ **temporal validity** . Mixing more data makes it worse (6d).

SYNTHESIS

The central trade-off

Textual memory

✓ Organizes & tracks state changes

✗ Loses fine visual evidence in compression

vs

Image (multimodal) memory

✓ Preserves native visual evidence

✗ Fails to select the valid current state

Neither side satisfies **both axes at once** — high visual granularity **and** evolving-state reasoning. That is why **no method solves MemEye** .

04

Takeaways

Implications, my assessment, and questions for discussion.

WHAT THE PAPER ARGUES

Implications for multimodal memory

1. Keep **both** image and text/structured memory — a single modality can't cover both axes.

2. Add **valid-evidence selection** — recency-aware, anti-stale filtering — not just storage & retrieval.

3. Design for both: **X-axis** = preserve visual granularity, **Y-axis** = judge temporal validity.

MY ASSESSMENT (PRESENTER'S VIEW, NOT THE PAPER)

Strengths & limitations

Strengths

✓ **Sharp problem framing** — the caption-bypass loophole is named and actually blocked by 3 gates.

✓ **Diagnostic, not just a score** — the two axes reveal *where* and *why* methods fail.

Limitations

✗ **Scale** — 371 questions is fine for diagnosis, modest for statistical robustness.

✗ **Domain** — everyday scenarios; generalization to professional settings is untested.

WRAP-UP

Conclusion & discussion

CONCLUSION

① A **verified visual-centric** memory benchmark — unsolvable without seeing.

② Two axes — **visual granularity × reasoning depth** — diagnose ability, not just score it.

③ **No method solves it**: none preserve fine visuals *and* select the current state.

DISCUSSION

❓ How would we design a memory-selection mechanism that judges **recency / validity**?

❓ What architecture best **fuses image + text memory** to ease this trade-off?

APPENDIX

Backup slides — construction pipeline, full results table, benchmark comparison, axis definitions.

APPENDIX A

How the benchmark was built

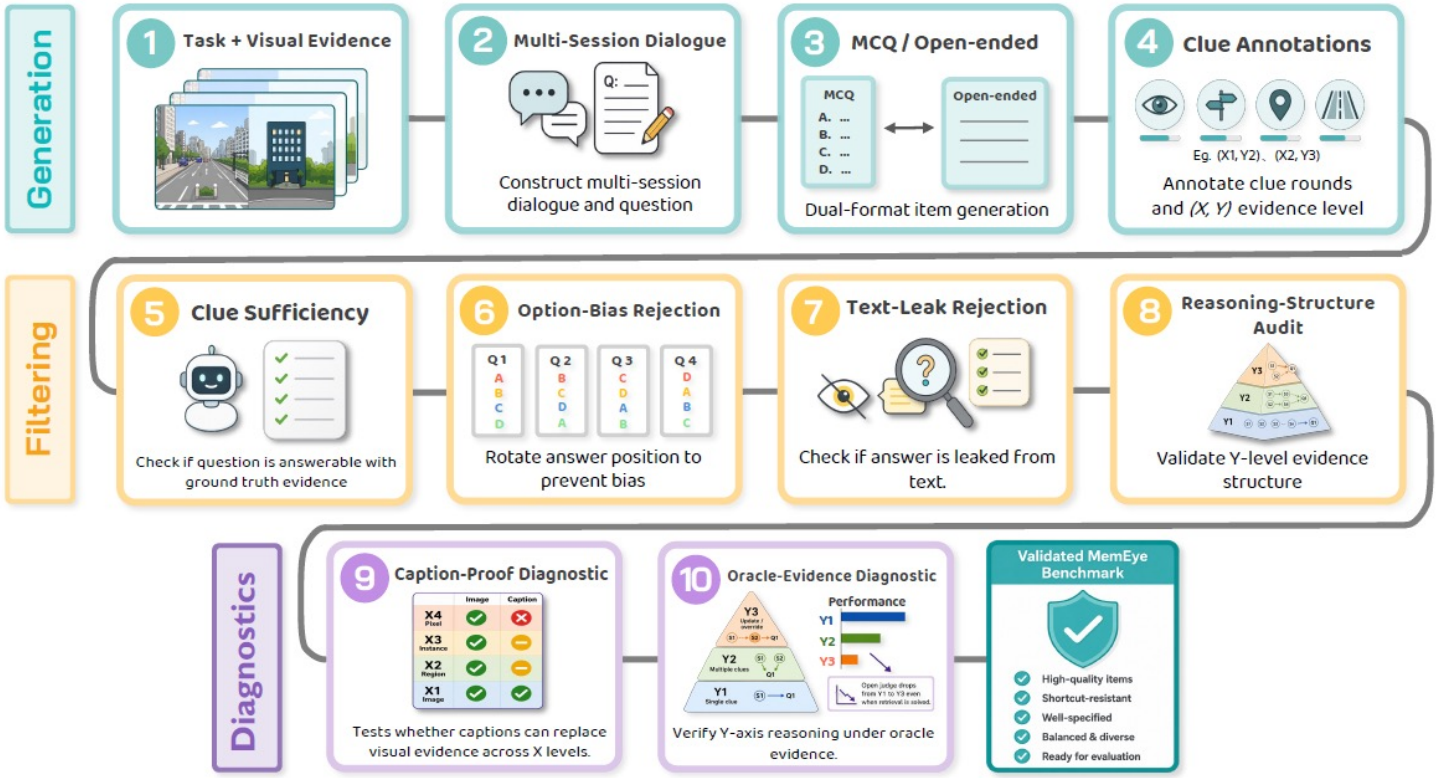


Figure 7: complete construction & validation pipeline — generation → item-level checks → aggregate diagnostics.

APPENDIX B

Full main-results table

Y	X	Metric	Textual memory							Multimodal memory					
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MIRIX	MMA	M2A	SM(V)
Y1	X1	EM	0.8000	0.9500	0.6750	0.2500	0.7750	0.7750	0.8000	1.0000	0.9000	0.6750	0.5500	0.4750	0.8500
		LLM-Judge	0.6500	0.4500	0.6000	0.3000	0.4000	0.2500	0.5500	0.6500	0.6000	0.4000	0.4500	0.4500	0.5000
	X2	EM	0.7500	0.5455	0.2045	0.2273	0.4545	0.4318	0.5000	0.6818	0.7727	0.5227	0.5455	0.2500	0.4773
		LLM-Judge	0.4545	0.4091	0.1364	0.1818	0.4091	0.3182	0.2727	0.5455	0.9091	0.5000	0.7273	0.3182	0.1818
	X3	EM	0.4662	0.4527	0.2534	0.2500	0.3649	0.4392	0.3209	0.5709	0.6554	0.5473	0.5507	0.3615	0.3784
		LLM-Judge	0.3716	0.3176	0.3108	0.2230	0.3176	0.4459	0.2230	0.5338	0.6554	0.2568	0.5946	0.3311	0.2838
	X4	EM	0.4722	0.5694	0.4444	0.2361	0.6389	0.5972	0.4583	0.4722	0.8056	0.3750	0.6250	0.5000	0.5000
		LLM-Judge	0.3889	0.3333	0.1111	0.0556	0.2222	0.2500	0.2222	0.3333	0.6389	0.3056	0.6389	0.1667	0.2222
Y2	X1	EM	0.5086	0.5172	0.4483	0.2500	0.4569	0.4052	0.5000	0.5345	0.5086	0.2845	0.5086	0.3190	0.5690
		LLM-Judge	0.4483	0.5345	0.5517	0.2759	0.3793	0.4138	0.4828	0.5000	0.6552	0.4138	0.3621	0.2931	0.4138
	X2	EM	0.4881	0.3810	0.1905	0.2619	0.3333	0.3690	0.3095	0.3810	0.2976	0.3214	0.4762	0.3333	0.3452
		LLM-Judge	0.1667	0.0952	0.1667	0.0238	0.1429	0.0952	0.1429	0.1667	0.2143	0.1429	0.1905	0.2143	0.0952
	X3	EM	0.4417	0.4125	0.3833	0.2292	0.3333	0.3917	0.3917	0.5750	0.6250	0.4583	0.4292	0.3792	0.3708
		LLM-Judge	0.3750	0.2833	0.2000	0.1167	0.2833	0.2917	0.2333	0.3917	0.6000	0.2667	0.3750	0.3583	0.2667
	X4	EM	0.3438	0.3523	0.2841	0.2472	0.3409	0.3551	0.3097	0.3636	0.3722	0.3665	0.4119	0.2955	0.2869
		LLM-Judge	0.3807	0.3182	0.3466	0.2614	0.4205	0.3636	0.2500	0.3977	0.3352	0.3352	0.2898	0.3352	0.2159
Y3	X1	EM	0.8000	0.7500	0.7000	0.2500	0.7000	0.4750	0.5000	0.9000	0.5750	0.6750	0.5500	0.6000	0.5500
		LLM-Judge	0.6500	0.6000	0.7000	0.4500	0.6000	0.5000	0.6000	0.6500	0.4500	0.6000	0.5500	0.6000	0.6000
	X2	EM	0.7000	0.7000	0.5500	0.2500	0.4750	0.6000	0.6500	0.6000	0.7750	0.4750	0.6750	0.4750	0.7000
		LLM-Judge	0.4000	0.5000	0.4500	0.4500	0.3000	0.4500	0.3000	0.2500	0.3500	0.2000	0.3000	0.1000	0.3000
	X3	EM	0.6000	0.6250	0.5750	0.2750	0.6250	0.5750	0.5250	0.5750	0.6500	0.6500	0.8000	0.5000	0.6000
		LLM-Judge	0.5500	0.5500	0.5500	0.4000	0.3500	0.5500	0.3500	0.4000	0.3000	0.4000	0.4500	0.5500	0.3500
	X4	EM	0.4333	0.3250	0.2833	0.2333	0.3417	0.3417	0.3167	0.5917	0.4750	0.2583	0.3417	0.3250	0.3083
		LLM-Judge	0.3000	0.3000	0.2833	0.3167	0.1667	0.3000	0.2000	0.4500	0.2167	0.1667	0.2667	0.3000	0.2000
Avg.	–	EM	0.5670	0.5484	0.4160	0.2467	0.4866	0.4797	0.4651	0.6038	0.6177	0.4674	0.5386	0.4011	0.4947
		LLM-Judge	0.4280	0.3909	0.3672	0.2546	0.3326	0.3524	0.3189	0.4391	0.4937	0.3323	0.4329	0.3347	0.3025

Table 2: per-coordinate EM and LLM-Judge for all methods (textual & multimodal).

APPENDIX C

Comparison with prior benchmarks

	VisDial [8]	CLEVR-dialog [19]	LoCoMo [28]	MMDU [26]	ConvBench [25]	MMRC [44]	MultiVerse [20]	Mem-Gallery [2]	MemEye
Caption-Proof	✗	✗	✗	✗	✗	✗	✗	✗	✓
Long Memory	✗	✗	✓	✗	✗	●	✗	✓	✓
Fine Visual	✓	✓	✗	✓	✓	✗	✓	●	✓
Cross-Context	✓	✓	✓	●	✓	✓	✓	✓	✓
State Revision	✗	✗	●	✗	✗	✓	✗	●	✓
Visual-State	✗	✗	✗	✗	✗	✗	✗	✗	✓

* ✓ = substantial ($\geq$20%), ● = limited (5–20%), ✗ = negligible ($<$5%) coverage under MemEye labeling. **Caption-Proof**: verifies native images cannot be replaced by captions. **Long Memory**: requires extended conversational memory. **Fine Visual**: X_3–X_4 visual bottlenecks (instance binding, pixel attributes, OCR). **Cross-Context**: Y_2 associations across turns, sessions, or modalities. **State Revision**: Y_3 reasoning under updates, conflicts, or overrides. **Visual-State**: joint high-visual, evolving-memory region.

Table 1: only MemEye covers all six properties, including Caption-Proof.

APPENDIX D

Axis definitions & examples

Level	Definition	Representative MemEye example
Visual evidence granularity (X)		
X_1 Scene-level	Global scene type, activity, or semantic gist; often recoverable from a good caption.	In Cartoon Ent., identifying the open blue-green sky and clouds behind a baby dinosaur climbing a tree.
X_2 Region-level	Semantically coherent subregions, grouped entities, or local spatial structure within a scene.	In Home Renovation, deciding which cabinet sample is closest to a brass hardware piece lying on the floor.
X_3 Instance-level	Specific object or person identity among visually or semantically similar candidates.	In Home Renovation, matching the labeled cabinet sample from a three-sample comparison to the same sample later shown beside a tape measure and pencil.
X_4 Pixel-level	Fine visual details such as small text, exact color, texture, count, or OCR-like evidence.	In CrossScene Memory, reading the current identification tag number displayed inside the fossil-room case.
Memory-reasoning depth (Y)		
Y_1 Atomic retrieval	One sufficient evidence unit answers the question; no cross-session composition is needed.	A single clue image in Cartoon Ent. is enough to answer what the background looks like in the tree-climbing scene.
Y_2 Relational association	Multiple non-conflicting clues must be linked across sessions, modalities, or references.	In Cartoon Ent., a palace-arc question requires comparing two remembered events to decide whether the crowd pelting or dinner fight occurred first.
Y_3 Evolutionary synthesis	Temporally ordered clues include updates, conflicts, or overrides; the answer depends on the valid current state.	In the fossil-room case, an earlier identification tag is later replaced, so the answer must use the updated tag rather than the stale one.

Table 6: definitions of X1–X4 and Y1–Y3 with representative MemEye examples.