CAUSALMEM: COUNTERFACTUALLY VALIDATED MEMORY FOR TRAINING-FREE VISUAL REASONING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous visual agents should improve as they accumulate experience, yet updating large vision-language models after every interaction is impractical. External memory offers a training-free route to self-evolution, but naive memory accumulation can amplify visual shortcuts: an agent may remember background, style, layout, or co-occurring objects rather than the evidence that actually determines the answer. We introduce **CausalMem**, a training-free memory framework that writes visual memories only after counterfactual validation. CausalMem alternates between wake-time reasoning and sleep-time consolidation. During wake, a frozen VLM solves new image reasoning tasks by retrieving validated semantic rules, hard negatives, and anchor episodes. During sleep, the agent proposes causal and nuisance visual factors, edits stored images as noisy interventions, and re-answers the edited cases. A rule is promoted to long-term memory only if the answer is invariant to nuisance edits and sensitive to edits of the hypothesized causal evidence. We formalize this process as noisy interventional memory validation and derive a finite-sample bound on the false acceptance of spurious rules. Across seven visual reasoning benchmarks, CausalMem improves Qwen2.5-VL-7B and Qwen3-VL-8B by 4.18 and 3.29 average points over vanilla inference without updating model parameters. Component ablations, stream-order and inductive-split controls, cross-benchmark transfer, and experiments with other editor and VLM families support that, on the tested settings, the gains come from counterfactually validated memory rather than passive retrieval, suggesting that self-evolving visual agents must learn not only how to recall experience, but which rules deserve to be remembered.

1 INTRODUCTION

Autonomous multimodal agents should improve as they act (Sumers et al., 2023; Xi et al., 2025). A frozen vision-language model (VLM) can solve a single image-question pair, but a long-lived agent faces streams of related tasks, repeated visual patterns, and changing contexts. Reusing past experience is therefore central to self-evolution: memory turns isolated inference calls into an adaptive system that can accumulate strategies, avoid repeated mistakes, and update behavior without retraining the base model (Zhang et al., 2025e; Hu et al., 2025b). Prior work uses verbal reflection, long-term memory, online adaptation, and self-generated examples (Shinn et al., 2023; Park et al., 2023; Packer et al., 2023; Wang et al., 2023; Tack et al., 2024).

Existing memory-based agents mainly ask how to store, retrieve, compress, or abstract experience (Zhang et al., 2025f; Tan et al., 2025; Zhang et al., 2025g; Yan et al., 2025; Hu et al., 2025a), distilling trajectories into reasoning traces or evolving memory banks (Ouyang et al., 2025; Wei et al., 2025; Zhang et al., 2025b), or building multimodal memories from demonstrations, scratchpads, tools, and latent visual memory (Sarch et al., 2024; Hu et al., 2024; Yu et al., 2025; Zhang et al., 2025d; Hong et al., 2025). They leave a key question under-specified: *what should be written into long-term memory?* If the writer stores the wrong abstraction, later retrieval can make an agent confidently reuse a shortcut.

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.

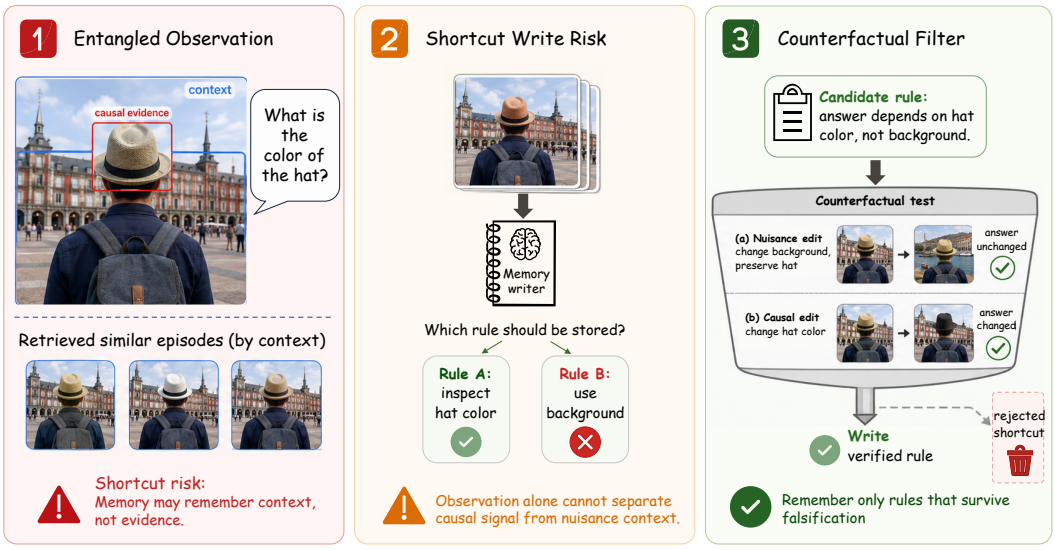


Figure 1: Motivation for counterfactually validated memory. Observed episodes entangle causal evidence with nuisance context, so a memory writer can mistakenly store a shortcut rule. CausalMem filters candidate rules with paired edits: nuisance edits should preserve the answer, while causal edits should change it. Only rules that survive this falsification test are written into memory.

The challenge is that visual episodes mix causal evidence with nuisance context (Goyal et al., 2017; Agrawal et al., 2018; Cadene et al., 2019). An answer may depend on a small object, relation, count, diagram symbol, or text span, while the same image also contains background, layout, style, viewpoint, and co-occurring objects. As illustrated in Figure 1, retrieving visually similar episodes by context can make a memory writer store the background regularity instead of the evidence that answers the question (Niu et al., 2021; Chen et al., 2020). Self-evolution then becomes shortcut amplification, a visual instance of shortcut learning (Geirhos et al., 2020; Ye et al., 2026). In causal terms, observational memory summarizes samples from $P(Y \mid I, Q)$, whereas robust visual reasoning needs rules that behave like local interventional statements: what should remain true when nuisance context changes, and what should change when the actual visual evidence is altered (Pearl, 2009; Peters et al., 2016; Arjovsky et al., 2019).

We propose **CausalMem**, a training-free framework that treats visual memory writing as counterfactual validation. The agent alternates between wake and sleep phases. During wake, a frozen VLM answers with retrieved semantic rules, hard negatives, and anchor episodes; during sleep, the agent proposes causal and nuisance factors, edits stored images with an image editor (Wu et al., 2025), and re-answers the edits. A candidate rule is promoted only if it passes two tests: the answer is invariant to nuisance edits and sensitive to edits of the hypothesized causal evidence.

Across seven visual reasoning benchmarks, CausalMem improves frozen Qwen2.5-VL-7B and Qwen3-VL-8B by 4.18 and 3.29 average points over vanilla inference, with the largest gains on fine-grained visual search and cross-instance reasoning. Ablations show that passive episodic retrieval can hurt, while counterfactual validation turns memory into a net gain. The gains hold for every tested stream order, under an inductive split in which memory is frozen before evaluation, under cross-benchmark transfer, and with a non-Qwen editor or InternVL base models.

The main contributions are:

- We identify observational shortcut amplification as a core failure mode of self-evolving visual memory and formulate memory writing as noisy interventional validation.
- CausalMem instantiates this view as a training-free wake–sleep agent that uses counterfactual image edits to maintain semantic rules, hard negatives, and anchor episodes without updating VLM parameters.
- A finite-sample false-acceptance bound shows how edit fidelity and validation diversity control the risk of promoting spurious rules.

- Experiments on seven benchmarks, ablations, protocol and transfer controls, cross-family runs, cost measurements, and memory diagnostics indicate that gains come from counterfactually validated memory rather than passive retrieval.

2 Related Work

Agent memory and self-evolution. Long-term memory has become a central mechanism for language and multimodal agents (Liu et al., 2025; Long et al., 2025; Bei et al., 2026). Reflexion stores verbal self-critiques from task feedback as episodic memory for later trials (Shinn et al., 2023). MemoryBank stores long-term conversational memories with update and forgetting mechanisms (Zhong et al., 2024), while MemGPT treats the context window as a managed memory hierarchy and lets the model page information between fast and external memory (Packer et al., 2023). Recent work further studies memories that evolve from experience: ReasoningBank distills reasoning memories from successful and failed trajectories (Ouyang et al., 2025); Evo-Memory benchmarks test-time learning with self-evolving memory in streams (Wei et al., 2025); MemGen integrates generative latent memory into agent reasoning (Zhang et al., 2025b); and G-Memory and Memo study hierarchical or compact memory for long-horizon agents (Zhang et al., 2025a; Gupta et al., 2025); see also Yao et al. (2022); Madaan et al. (2023); Zhao et al. (2024). CausalMem asks when a visual rule should enter long-term memory: each candidate rule is treated as a hypothesis to be falsified before reuse.

Visual reasoning agents and multimodal memory. Multimodal agents increasingly improve visual reasoning by externalizing intermediate computation or augmenting perception (Wu et al., 2023; Yang et al., 2023). Visual Sketchpad uses drawing as an explicit visual chain of thought for multimodal models (Hu et al., 2024). Thyme and DeepEyes-style systems extend multimodal reasoning with image manipulation, tool calls (Lu et al., 2023; Surís et al., 2023), and agentic "thinking with images" (Zhang et al., 2025d; Zheng et al., 2025; Hong et al., 2025). ViPER studies self-improvement for visual perception abilities (Zhang et al., 2025c), while VisMem explores latent visual memory for VLMs (Yu et al., 2025). These approaches primarily improve the reasoning procedure, perceptual skill, or internal visual representation. CausalMem leaves the base VLM fixed and changes the memory consolidation rule.

Causality, invariance, and spurious visual shortcuts. Structural causal models formalize interventions and counterfactuals (Pearl, 2009). Invariant prediction and invariant risk minimization argue that stable mechanisms can be identified by testing across environments (Peters et al., 2016; Arjovsky et al., 2019), and causal representation learning studies how such mechanisms can be represented in learned systems (Schölkopf et al., 2021; Komanduri et al., 2025). In multimodal reasoning, causal and counterfactual formulations have been used to diagnose confounding and illusory causality in videos (Chen et al., 2024b), while shortcut-learning analyses and benchmarks show that vision and multimodal models exploit spurious visual biases (Geirhos et al., 2020; Ye et al., 2026; Yang et al., 2025; Vo et al., 2025). CausalMem does not propose a new general-purpose causal discovery method. Instead, it operationalizes invariance as a write criterion for agent memory: generated image edits create local environments for testing whether a proposed memory is invariant to nuisance changes and sensitive to changes in causal visual evidence.

3 Counterfactual Falsification for Visual Memory

CausalMem is a wake–sleep memory wrapper around a frozen VLM, built on one write rule: *a visual rule is written to memory only after the agent has tried, and failed, to falsify it*. During *wake*, the VLM answers with retrieved memory and logs each episode; during *sleep*, it proposes rules, edits stored images, re-answers them, and promotes only rules that pass. Figure 2 traces this pipeline on a running example, and Algorithm 1 lists it step by step. Only external memory changes; the VLM and the editor stay frozen.

3.1 Problem Setup

Let Q be a question, O the visual evidence that causally determines its answer (object state, relation, count, symbol, or text span), and C the nuisance context (background, lighting, style, layout,

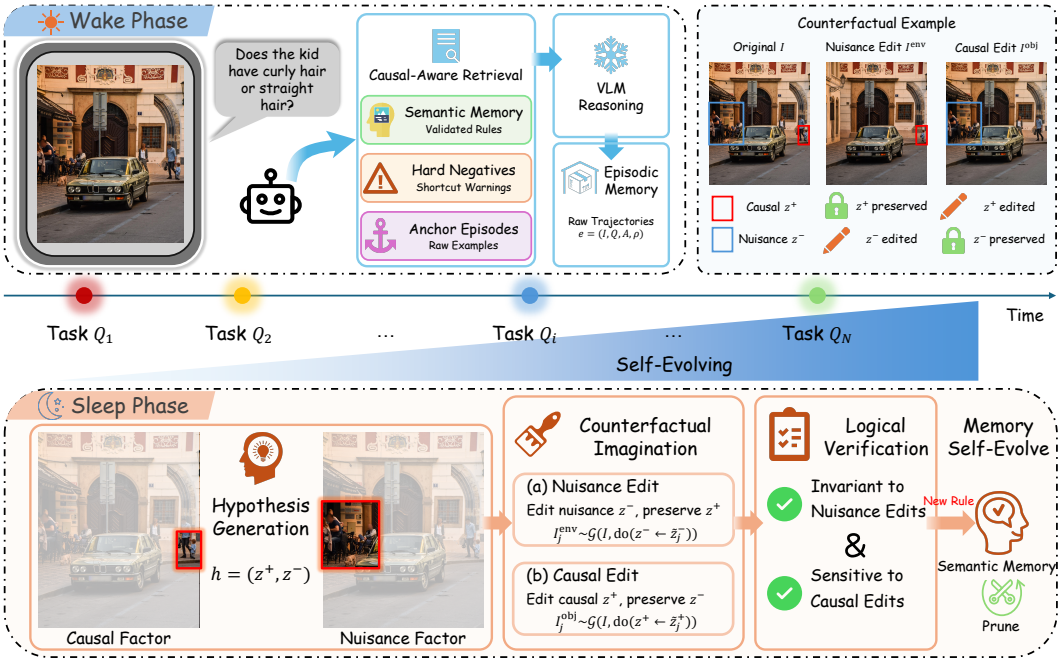


Figure 2: CausalMem workflow on a running example ("Does the kid have curly or straight hair?"). *Wake*: a frozen VLM answers using validated semantic rules, hard negatives, and anchor episodes. *Sleep*: a candidate rule with claimed evidence z^+ (the kid's hair) and nuisance z^- (the street context) is tested with an environment edit (answer stays *Curly*: invariance passes) and an object edit (answer becomes *Straight*: necessity passes), so, after verifier filtering and a conflict check, the rule is promoted to $\mathcal{S}$.

viewpoint, co-occurring objects). The image and true answer are generated as

$$I = g_I(O, C), \qquad Y = f_Y(O, Q), \tag{1}$$

up to exogenous noise. In an observed episode stream, O and C can be correlated, so a memory item that improves observational accuracy may encode a nuisance shortcut rather than the mechanism f_Y. CausalMem instead targets memory that remains useful under the intervention $\mathrm{do}(C \leftarrow C')$ for any admissible context C' (i.e., one that preserves O and the meaning of Q), and that stops being useful when O is altered. Because such interventions cannot be observed, the agent approximates them with a text-guided image editor and an edit verifier. Throughout, $f(I, Q)$ is the frozen VLM's answer and $\delta_Q(y, y') \in \{0, 1\}$ is benchmark-specific answer equivalence.

Supervision. The memory writer never uses labels; its only reference is A, the VLM's own wake-time answer to the unedited episode.

3.2 Memory Design

CausalMem keeps three stores, $\mathcal{M} = \mathcal{E} \cup \mathcal{S} \cup \mathcal{H}$, and distinguishes a candidate *hypothesis* h from an accepted *rule* $s \in \mathcal{S}$:

- **Episodic memory** $\mathcal{E}$: raw episodes $e = (I, Q, A, \rho)$ with image, question, wake answer and rationale, plus an outcome signal, a retrieval key, and metadata (task type, timestamp).
- **Semantic memory** $\mathcal{S}$: accepted rules $s = (h, n_s^+, n_s^-)$, where $h = (r, z^+, z^-, \sigma)$ holds the rule text r, claimed evidence z^+, claimed nuisance z^-, and scope σ; $n_s^\pm$ count passed and failed checks. Supporting and falsifying edits are stored with each rule.
- **Hard-negative memory** $\mathcal{H}$: items $(I^{\mathrm{env}}, Q, A, r)$, i.e., an environment-edited image of an accepted rule whose answer stayed unchanged, with that rule. At wake time they serve as contrastive in-context warnings ("z^- changed, the answer did not").

Algorithm 1 Sleep-time counterfactual testing and consolidation

Require: Stores $\mathcal{E}, \mathcal{S}, \mathcal{H}$; frozen VLM f; editor $\mathcal{G}$; edit verifier $\mathcal{W}$; budget B
1: $\mathcal{B} \leftarrow \mathrm{Prioritize}(\mathcal{E})$ by novelty, self-detected inconsistencies, and retrieval conflicts
2: **for** episode $e = (I, Q, A, \rho) \in \mathcal{B}$ **do**
3: $\quad \mathcal{P}_e \leftarrow \mathrm{Hypothesize}(f, e, \mathcal{S})$, hypotheses $h = (r, z^+, z^-, \sigma)$ ▷ Step 1
4: $\quad$ **for** $h \in \mathcal{P}_e$ **do**
5: $\quad\quad$ Sample K_e environment edits (Eq. 4) and K_o object edits (Eq. 5) ▷ Step 2
6: $\quad\quad$ Keep edits with fidelity $\geq \tau_q$ under $\mathcal{W}$, giving $\mathcal{J}_{\mathrm{env}}, \mathcal{J}_{\mathrm{obj}}$
7: $\quad\quad$ Re-answer kept edits with f; compute $\widehat{\mathrm{Inv}}(h)$, $\widehat{\mathrm{Nec}}(h)$ ▷ Step 3
8: $\quad\quad$ **if** $\mathrm{Accept}(h) = 1$ (Eq. 7) **then**
9: $\quad\quad\quad$ Merge s_h into $\mathcal{S}$; add invariant environment edits to $\mathcal{H}$ ▷ Step 4
10: $\quad\quad$ **else**
11: $\quad\quad\quad$ Record h and its failed edits in the conflict index
12: $\quad\quad$ **end if**
13: $\quad$ **end for**
14: **end for**
15: Prune $\mathcal{S} \cup \mathcal{H}$ to budget B by retention value (Eq. 14)
16: **return** updated $\mathcal{S}, \mathcal{H}$

Rule reliability π_s has the Beta–Bernoulli posterior $\pi_s \sim \mathrm{Beta}(a_0 + n_s^+,\, b_0 + n_s^-)$, where n_s^+ counts passed invariance and necessity checks and n_s^- counts failed checks and later contradictions. Retrieval and pruning use its lower confidence bound $\mathrm{LCB}(\pi_s)$ rather than its mean, so rules supported by few edits are used conservatively (Auer et al., 2002). We use the uniform prior $a_0 = b_0 = 1$ and define $\mathrm{LCB}(\pi_s)$ as the 10th percentile of this posterior. The conflict count $\kappa(h) \in \mathbb{N}$ of a hypothesis (or rule) h counts active rules $s \in \mathcal{S}$ with $\mathrm{LCB}(\pi_s) > \mathrm{LCB}(\pi_h)$ that share its scope but rely on different evidence:

$$\kappa(h) = \sum_{s \in \mathcal{S}_{>h}} \mathbf{1}\{\mathrm{ScopeMatch}(h, s)\}\, \mathbf{1}\{\cos(\phi(z_h^+), \phi(z_s^+)) < 0.50\}, \tag{2}$$

where $\mathcal{S}_{>h} = \{s \in \mathcal{S} : \mathrm{LCB}(\pi_s) > \mathrm{LCB}(\pi_h)\}$, $\phi(\cdot)$ is the text encoder (a string to a unit vector in $\mathbb{R}^d$) that also builds retrieval keys, $\cos(\cdot, \cdot)$ is cosine similarity, and $\mathrm{ScopeMatch}(h, s) = 1$ iff $\sigma_h = \sigma_s$ or $\cos(\phi(\sigma_h), \phi(\sigma_s)) \geq 0.60$ on the scope-label text. For a hat-color question, for example, a rule relying on hat color and one relying on background color conflict, while rules with different scopes do not. For a new hypothesis, π_h uses the counts of its passed and failed checks (Step 4).

3.3 Wake: Retrieve, Answer, Log

Input: a new task (I_t, Q_t) and memory $\mathcal{M}$. **Output:** answer A_t and a new episode in $\mathcal{E}$. Retrieval is deterministic and per store. A query key is built from the question, an image caption, localized evidence, and task tags, and each semantic rule $m \in \mathcal{S}$ is scored by

$$\mathrm{score}(m \mid I_t, Q_t) = \alpha\, \mathrm{sim}(m; I_t, Q_t) + \beta\, \mathrm{LCB}(\pi_m) + \gamma\, \mathrm{cov}(m; I_t, Q_t) - \zeta\, \kappa(m) - \xi\, \ell(m), \tag{3}$$

with weights $(\alpha, \beta, \gamma, \zeta, \xi) = (1.0, 0.5, 0.25, 0.5, 0.05)$, combining key similarity $\mathrm{sim}(m; I_t, Q_t) = \cos(\phi(\text{query key}), \phi(\text{key of } m))$, reliability, coverage $\mathrm{cov}(m; I_t, Q_t)$ of the task's factors (query-dependent, unlike the pruning term cov_σ of Appendix C), conflict $\kappa(m)$ (Eq. 2 evaluated for rule m), and cost $\ell(m)$ (prompt tokens plus 256 tokens per image). The top k_s rules within cosine distance 0.80 are kept. Hard negatives are not scored: up to k_h items of $\mathcal{H}$ linked to the highest-ranked rules are added. Up to k_a anchor episodes from $\mathcal{E}$ are ranked separately by question-embedding distance with the same 0.80 cutoff. The prompt contains three blocks in order, *validated rules*, *shortcut warnings* (hard-negative texts, with their edited images placed after the query image), and *anchor episodes*, followed by the current question (Appendix E); the frozen VLM returns $A_t = f(I_t, Q_t \mid \mathcal{M}_t)$, where $\mathcal{M}_t$ is the retrieved context. The episode (I_t, Q_t, A_t, ρ_t) is appended to $\mathcal{E}$.

3.4 Sleep: Falsify, Then Write

Sleep runs after every 10 completed tasks in our experiments (or whenever offline compute is available). Its four steps below map to the line numbers of Algorithm 1.

Step 1: Hypothesize (line 3). Input: a prioritized episode e and the current $\mathcal{S}$. **Output:** candidate hypotheses $h = (r, z^+, z^-, \sigma)$. The frozen VLM proposes the factors z^+ that should determine A and the nuisance factors z^- that may merely co-occur with it (in Fig. 2, the kid's hair versus the street). $z^\pm$ are the model's *claims* about O and C, tested next.

Step 2: Intervene and verify edits (lines 5–6). Input: I and h. **Output:** the verified edit sets $\mathcal{J}_{\text{env}}$ and $\mathcal{J}_{\text{obj}}$. The editor produces two complementary edit families:

$$I_j^{\text{env}} \sim \mathcal{G}\big(I, \text{do}(z^- \leftarrow \tilde{z}_j^-), \text{preserve}(z^+)\big), \qquad j = 1, \ldots, K_e, \tag{4}$$

$$I_j^{\text{obj}} \sim \mathcal{G}\big(I, \text{do}(z^+ \leftarrow \tilde{z}_j^+), \text{preserve}(z^-)\big), \qquad j = 1, \ldots, K_o, \tag{5}$$

where $\tilde{z}_j^-$ is an alternative nuisance value and $\tilde{z}_j^+$ a minimal alteration or removal of the claimed evidence; environment edits approximate $\text{do}(C \leftarrow C')$ and object edits alter O. The verifier $\mathcal{W}$ scores each edit with a fidelity q_j^{env} or $q_j^{\text{obj}} \in [0, 1]$ from three checks: protected factors are preserved, the target factor changed, and the image remains answerable. Edits with fidelity below τ_q are discarded; the rest are *verified edits*.

Step 3: Re-answer and gate (lines 7–8). Input: the verified edits and A. **Output:** $\text{Accept}(h) \in \{0, 1\}$. The frozen VLM re-answers each verified edit (M samples for stochastic tasks), giving

$$\widehat{\text{Inv}}(h) = \frac{\sum_{j \in \mathcal{J}_{\text{env}}} q_j^{\text{env}} \, \delta_Q\big(f(I_j^{\text{env}}, Q), A\big)}{\sum_{j \in \mathcal{J}_{\text{env}}} q_j^{\text{env}}}, \qquad \widehat{\text{Nec}}(h) = \frac{\sum_{j \in \mathcal{J}_{\text{obj}}} q_j^{\text{obj}} \, \big(1 - \delta_Q\big(f(I_j^{\text{obj}}, Q), A\big)\big)}{\sum_{j \in \mathcal{J}_{\text{obj}}} q_j^{\text{obj}}}, \tag{6}$$

and the acceptance gate

$$\text{Accept}(h) = \mathbf{1}\{\widehat{\text{Inv}}(h) \geq \tau_{\text{inv}}\} \, \mathbf{1}\{\widehat{\text{Nec}}(h) \geq \tau_{\text{nec}}\} \, \mathbf{1}\{\kappa(h) < \tau_{\text{conf}}\}. \tag{7}$$

Thus h must be invariant to nuisance edits, change under evidence edits, and not conflict with a stronger rule; we set $\tau_{\text{conf}} = 1$, i.e., $\kappa(h) = 0$ (Eq. 2). In the example, the answer stays *Curly* when the street is replaced and flips to *Straight* when the hair is edited, so both gates pass.

Step 4: Write and prune (lines 9, 11, 15). Input: h, $\text{Accept}(h)$, and the verified edits. **Output:** updated $\mathcal{S}$ and $\mathcal{H}$. If $\text{Accept}(h) = 1$, let $s_h = (h, n_{s_h}^+, n_{s_h}^-)$ be the rule formed from h, with counts initialized from its passed and failed checks, and update

$$\begin{aligned} &\mathcal{S} \leftarrow \text{Merge}\,\big(\mathcal{S}, s_h; \text{ key} = (\sigma, z^+)\big), \\ &\mathcal{H} \leftarrow \mathcal{H} \cup \{(I_j^{\text{env}}, Q, A, r) : j \in \mathcal{J}_{\text{env}}, \ \delta_Q(f(I_j^{\text{env}}, Q), A) = 1\}, \end{aligned} \tag{8}$$

and the counts $n_s^\pm$ (hence π_s) are updated with the new outcomes; Merge combines s_h with any rule of the same scope and evidence key. A rejected hypothesis is recorded, with its failed edits, in the conflict index. Finally, *quality-aware pruning* keeps the active memory within a prompt budget B: each item receives a retention value combining label-free replay gain, $\text{LCB}(\pi_m)$, scope coverage, redundancy, and a contradiction term (Eq. 14, Appendix C), and items are kept greedily by value per cost $\ell(m)$. When two rules conflict, the one with the higher LCB stays active; the other becomes dormant (not retrieved), is kept as a counterexample in the conflict index, and receives one failure ($n_s^- \mathrel{+}= 1$).

3.5 False Acceptance Under Noisy Edits

We bound how often a spurious rule survives a gate when edits are only faithful with high probability.

Assumption 1 (Noisy intervention oracle)**.** *For a hypothesis h, each verified edit is faithful with probability at least $1 - \eta$. A faithful environment edit changes only nuisance factors in z^-, preserves z^+, and keeps the true answer unchanged. A faithful object edit preserves nuisance context and alters z^+ enough to change the true answer when z^+ is necessary. Conditional on faithfulness, validation outcomes are independent Bernoulli samples.*

Proposition 1 (False acceptance of a spurious rule)**.** *Consider a spurious hypothesis h whose claimed causal evidence is actually nuisance. Suppose the probability that h passes a single faithful validation test is at most $p_s < \tau$, where τ is the acceptance threshold applied to the average of n validation outcomes. Under Assumption 1,*

$$\Pr[\text{Accept}(h) = 1] \leq \exp\big(-2n(\tau - p_s)^2\big) + n\eta. \tag{9}$$

The bound applies to the unweighted average of each gate separately ($n = |\mathcal{J}_{\text{env}}|, \tau = \tau_{\text{inv}}$ or $n = |\mathcal{J}_{\text{obj}}|, \tau = \tau_{\text{nec}}$); since Accept is a conjunction, the minimum per-gate bound holds. More verified edits reduce false acceptance exponentially, whereas unfaithful edits add a linear $n\eta$ term, motivating fidelity filtering and LCB-based retrieval. The bound is loose for the small default n and should be read qualitatively (proof in Appendix A).

4 EXPERIMENTS

We ask whether CausalMem improves frozen VLMs, whether the gains come from counterfactual validation rather than retrieval, whether they depend on the evaluation protocol or model family, and how selective and costly the memory is.

4.1 SETUP

Base models and implementation. We evaluate CausalMem as a training-free wrapper around Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a). The base VLM parameters remain frozen in all experiments. The sleep-time editor is Qwen-Image-Edit (Wu et al., 2025). Unless otherwise stated, wake-time retrieval uses $k_s = 4$ semantic rules, $k_h = 2$ hard negatives, and $k_a = 2$ anchor episodes. Sleep uses $K_e = 4$ environment and $K_o = 4$ object edits per hypothesis, $M = 3$ re-answer samples for stochastic tasks, $\tau_{\text{inv}} = \tau_{\text{nec}} = 0.75$, and edit-fidelity threshold $\tau_q = 0.70$. All hyperparameters, including the retrieval and pruning weights, are fixed globally across benchmarks and backbones (Table 3, Appendix B).

Evaluation protocol. Unless stated otherwise, evaluation is *transductive and online*: memory is updated during evaluation, each benchmark stream is processed once in a fixed order that is identical for all methods in Table 1, sleep runs after every 10 completed tasks, and only earlier examples can affect later ones; there is no separate memory-building split. As in unlabeled test-time settings for self-evolving agents (Wei et al., 2025), labels are never used to build, validate, retrieve, or prune memory. Section 4.4 controls for stream order and replaces this protocol with an inductive split.

Benchmarks and metrics. We use seven benchmarks: fine-grained visual search (V*Bench (Wu & Xie, 2024)), general multimodal recognition and reasoning (MMStar (Chen et al., 2024a), MM-Bench (Liu et al., 2024)), real-world visual QA (RealWorldQA (xAI, 2024)), hallucination-sensitive recognition (HallusionBench (Guan et al., 2024)), open-ended visual reasoning (MM-Vet (Yu et al., 2023)), and college-level multimodal problem solving (MMMU (Yue et al., 2024)). We report official scores in percent (aAcc for HallusionBench).

Baselines. We compare against vanilla VLM inference, text chain-of-thought prompting (CoT) (Wei et al., 2022), Visual Sketchpad (Hu et al., 2024), Thyme (Zhang et al., 2025d), and DeepEyesV2 (Hong et al., 2025). Ablations use Qwen3-VL-8B and remove one component at a time: passive episodic RAG (Lewis et al., 2020), environment edits, object edits, hard negatives, and pruning.

4.2 MAIN RESULTS

In Table 1, CausalMem gains 4.18 average points over vanilla on Qwen2.5-VL-7B (0.49 over the strongest baseline; best in six of seven columns) and 3.29 on Qwen3-VL-8B (0.98 over the strongest baseline; best in five of seven), without updating the model or tuning per benchmark.

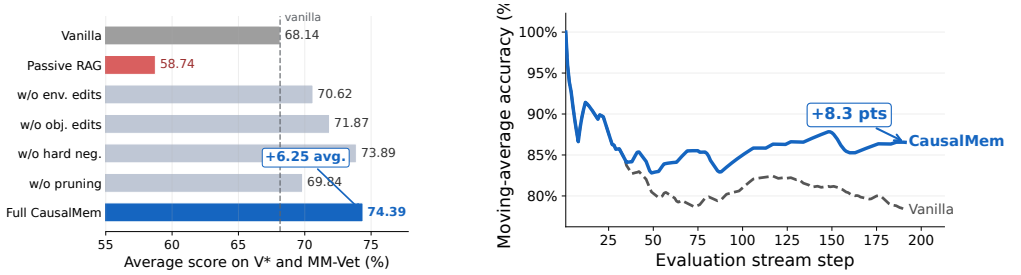
The strongest gains appear on V* (+8.90 and +8.38), where questions hinge on localized visual evidence and recurring search patterns, but gains also hold on HallusionBench, MM-Vet, and MMMU. The improvements are not uniform: DeepEyesV2 remains stronger on MMBench for Qwen2.5-VL-7B and on MMStar and RealWorldQA for Qwen3-VL-8B, so CausalMem complements rather than replaces specialized visual reasoning procedures.

4.3 ABLATION STUDIES

Figure 3a explains where the gains come from; we average V* (fine-grained localization) and MM-Vet (open-ended reasoning). Passive RAG is 9.40 points below vanilla on this two-task average (58.74

Table 1: Main results across seven visual reasoning benchmarks. Scores are official benchmark scores in percent; higher is better. The blue rows mark CausalMem, and the green rows report gains over vanilla inference for the same base VLM. Bold denotes the best method in each block and column; underline denotes the second best.

Base VLM	Method	V*	MMStar	MMBench	RealWorld	Hallusion	MM-Vet	MMMU	Avg.
Qwen2.5-VL-7B	Vanilla	73.82	62.13	81.11	64.44	66.14	60.55	55.11	66.19
	CoT	75.92	61.66	82.04	65.36	66.77	61.47	54.22	66.78
	Sketchpad	79.58	59.80	84.26	66.14	68.13	61.01	55.77	67.81
	Thyme	82.20	60.93	84.01	65.89	70.66	62.38	53.00	68.44
	DeepEyesV2	81.80	62.46	**86.89**	67.18	71.40	62.84	56.55	69.87
	CausalMem	**82.72**	**63.87**	85.10	**67.84**	**71.82**	**63.76**	**57.44**	**70.36**
	Δ ↑	+8.90	+1.74	+3.99	+3.40	+5.68	+3.21	+2.33	+4.18
Qwen3-VL-8B	Vanilla	78.01	62.53	84.26	67.71	71.61	58.26	60.44	68.97
	CoT	79.05	61.86	86.75	66.01	71.92	58.71	61.00	69.33
	Sketchpad	80.10	60.66	82.50	63.92	73.08	61.93	61.66	69.12
	Thyme	78.53	63.27	82.86	70.46	72.55	59.17	61.33	69.74
	DeepEyesV2	82.20	**64.53**	84.65	**71.63**	72.97	60.09	62.88	71.28
	CausalMem	**86.39**	63.80	**88.39**	68.37	**73.29**	**62.39**	**63.22**	**72.26**
	Δ ↑	+8.38	+1.27	+4.13	+0.66	+1.68	+4.13	+2.78	+3.29



(a) Component ablations. Bars average V* and MM-Vet on Qwen3-VL-8B.

(b) Evolution curve. V* moving average on Qwen3-VL-8B over the stream (transductive online evaluation).

Figure 3: Ablation and evolution evidence. Left: counterfactual validation turns memory from harmful passive retrieval into a net gain. Right: the memory advantage persists through the stream.

vs. 68.14), showing that unvalidated memory can harm, while full CausalMem reaches 74.39 (+6.25 over vanilla, +15.65 over passive RAG): the value of memory comes from counterfactual testing, not retrieval alone. Removing environment edits costs 3.77 points (nuisance invariance rejects shortcut rules) and removing object edits 2.52 points (rules must stay sensitive to the claimed evidence). Without pruning, the average falls by 4.55 points, because redundant and conflicting entries consume the budget. For the wake context, removing hard negatives costs 0.50 points (73.89), removing anchor episodes 1.31 (73.08), and keeping validated rules only 2.52 (71.87), so hard negatives and anchors add complementary context (Table 4, Appendix F).

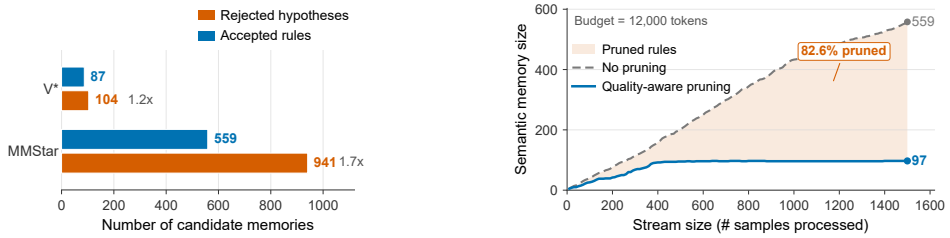
Evolution over the stream. Figure 3b shows, under the transductive online protocol, that CausalMem remains above vanilla inference through most of the V* stream and ends with a clear margin. Since the VLM is frozen, the separation reflects validated retrieval cues rather than training; order and split controls follow in Section 4.4.

4.4 Robustness of the Protocol and Generalization

CausalMem beats vanilla for every one of five random stream orders, and the gains persist when memory is built on one half, frozen, and evaluated on the disjoint half (+7.33 on V*, +5.57 on HallusionBench). Memory built on MMStar transfers to V* (+5.24), V*-built memory helps less on MMStar (+0.93), and one shared memory built from 300 unlabeled samples raises pooled accuracy over the three full datasets from 78.69 to 80.90; rules are indexed by scope, not dataset (e.g., the physical-contact rule in Appendix D). Replacing the editor with InstructPix2Pix (Brooks et al., 2023)

Table 2: Protocol, transfer, and model-family controls. Orders: mean±std over five random permutations; †minimum gain over the orders. Inductive: memory built without labels on one random half, frozen, evaluated wake-only on the other (both directions averaged). Transfer: memory frozen at target. Shared pool: one memory from 100 unlabeled samples each of V*, MMStar, MMBench; pooled accuracy = total correct over all samples of the three full datasets / total number of samples.

Setting	Base VLM / editor	Vanilla	CausalMem	Δ
Stream order and inductive split (V* / Hallusion)				
5 random orders	Qwen2.5-VL-7B	73.82 / 66.14	81.67±1.17 / 71.88±0.38	+6.28 / +5.36†
Inductive A↔B	Qwen2.5-VL-7B	73.82 / 66.14	81.15 / 71.71	+7.33 / +5.57
Cross-benchmark transfer (Qwen3-VL-8B)				
V*→MMStar	Qwen-Image-Edit	62.53	63.46	+0.93
MMStar→V*	Qwen-Image-Edit	78.01	83.25	+5.24
Shared pool (pooled acc.)	Qwen-Image-Edit	78.69	80.90	+2.21
Editor and model families (V* / MM-Vet; V* only for InternVL)				
Qwen3-VL-8B	InstructPix2Pix	78.01 / 58.26	84.82 / 61.47	+6.81 / +3.21
InternVL3-8B	Qwen-Image-Edit	72.25	76.44	+4.19
InternVL3.5-8B	Qwen-Image-Edit	71.20	74.87	+3.67



(a) Accepted rules vs. rejected hypotheses.

(b) Active memory size as the stream grows.

Figure 4: Memory diagnostics on Qwen3-VL-8B. Left: the write gate rejects more hypotheses than it accepts. Right: pruning keeps the active memory compact while the unpruned store keeps growing.

(vs. 86.39 / 62.39 with Qwen-Image-Edit) or using InternVL3-8B (Zhu et al., 2025) and InternVL3.5-8B (Wang et al., 2025) as proposer, re-answerer, and verifier keeps clear gains (Table 2). CausalMem is thus a reusable mechanism demonstrated on these tested families; as the proposer and re-answerer are the same VLM, shared errors may remain (Appendix G).

4.5 Memory Selectivity and Growth

The write gate is selective: rejected hypotheses outnumber accepted rules (Figure 4a; 104 vs. 87 on V*, 941 vs. 559 on MMStar). A keyword audit on V* against five MM-SpuBench shortcut categories (Ye et al., 2026) matches 94.23% (98/104) of rejected hypotheses, mostly background/scene, and 45.45% (35/77) of passive-RAG memory items; this coarse proxy describes what the gate filters, not causation or shortcut prevalence (Appendix F). Without pruning, the MMStar store reaches 559 rules; pruning keeps it compact, ending with 97 (Figure 4b). **Cost.** On V* (Qwen3-VL-8B), wake uses a single VLM call, but its bounded context (2,234 extra tokens) raises time per sample from 0.71 s to 1.32 s; with a near-constant active store, this context does not grow with the stream. Sleep is batched offline and costs 22.23 VLM and 1.98 editor calls per sample (Table 5).

5 Conclusion

CausalMem treats memory writing for self-evolving visual agents as counterfactual validation rather than passive accumulation. Across seven benchmarks, ablations, order and inductive-split controls, transfer, and the tested model families, the evidence indicates that the gains come from validated memory rather than retrieval alone: reliable self-evolution depends on deciding which rules deserve to be remembered.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and LaTeX formatting assistance. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies training-free memory for visual reasoning agents using publicly available benchmarks and models; it involves no human subjects or private data. Generated image edits are used only internally for validation and are not released as data. As with any self-evolving agent, validated memories could still encode biases present in the base VLM or benchmarks, which we discuss in Appendix G.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the full wake–sleep procedure, Section 4.1 states the evaluation protocol, Table 3 (Appendix B) lists all hyperparameters, Appendix C defines every pruning term, Appendix F reports cost and audit details, and Appendix E provides all prompt templates. The proof of Proposition 1 is in Appendix A. Code will be released upon publication.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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. *arXiv preprint arXiv:2601.03515*, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10800–10809, 2020.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. Mecd: Unlocking multi-event causal discovery in video reasoning. *Advances in neural information processing systems*, 37:92554–92580, 2024b.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Gunshi Gupta, Karmesh Yadav, Zsolt Kira, Yarin Gal, and Rahaf Aljundi. Memo: Training memory-efficient embodied agents with reinforcement learning. *arXiv preprint arXiv:2510.19732*, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32779–32798, 2025a.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025b.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30648–30668, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, et al. Memverse: Multimodal memory for lifelong learning agents. *arXiv preprint arXiv:2512.03627*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Gabriel Sarch, Lawrence Jang, Michael J Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11888–11898, 2023.

Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. Online adaptation of language models with a memory of amortized contexts. *Advances in Neural Information Processing Systems*, 37:130109–130135, 2024.

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

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, et al. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2025.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

xAI. Realworldqa: Real-world spatial understanding benchmark, 2024. URL `https://x.ai/blog/grok-1.5v`. Benchmark released with Grok-1.5V.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

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

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Z Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? *arXiv preprint arXiv:2506.18322*, 2025.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Wenqian Ye, Bohan Liu, Guangtao Zheng, Di Wang, Xu Cao, Yunsheng Ma, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2854–2865, 2026.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. *arXiv preprint arXiv:2511.11007*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025a.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. *arXiv preprint arXiv:2509.24704*, 2025b.

Juntian Zhang, Song Jin, Chuanqi Cheng, Yuhan Liu, Yankai Lin, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, et al. Viper: Empowering the self-evolution of visual perception abilities in vision-language model. *arXiv preprint arXiv:2510.24285*, 2025c.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025d.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025e.

Zeyu Zhang, Quanyu Dai, Xu Chen, Rui Li, Zhongyang Li, and Zhenhua Dong. Memengine: A unified and modular library for developing advanced memory of llm-based agents. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 821–824, 2025f.

Zeyu Zhang, Quanyu Dai, Rui Li, Xiaohe Bo, Xu Chen, and Zhenhua Dong. Learn to memorize: Optimizing llm-based agents with adaptive memory framework. *arXiv preprint arXiv:2508.16629*, 2025g.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 19632–19642, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *38th AAAI Conference on Artificial Intelligence, AAAI 2024, Feb 20-27 2024 Vancouver, Canada*, volume 38, pp. 19724–19731. Association for the Advancement of Artificial Intelligence (AAAI), 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Proof of Proposition 1

The proposition is stated for a generic validation gate whose decision is based on the average of n binary validation outcomes. In the implementation, the same argument applies to the nuisance-invariance and causal-sensitivity gates separately; using both gates only makes false acceptance harder.

Let $X_i \in \{0, 1\}$ denote whether the i-th validation test for a spurious hypothesis passes, and let F_i denote that the corresponding edit is faithful. Define $\mathcal{F} = \cap_{i=1}^{n} F_i$ as the event that all edits used by the gate are faithful. By Assumption 1, $\Pr[F_i = 0] \leq \eta$, and no independence among the faithfulness events is required. Conditional on $\mathcal{F}$, the outcomes X_i are independent Bernoulli variables with means at most p_s. For $\bar{X} = n^{-1} \sum_i X_i$ and $\tau > p_s$, Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr[\bar{X} \geq \tau \mid \mathcal{F}] \leq \exp(-2n(\tau - p_s)^2). \tag{10}$$

Since the gate can accept a spurious rule only if either all edits are faithful and the empirical pass rate exceeds the threshold, or at least one edit is unfaithful, we have

$$\Pr[\text{Accept}(h) = 1] \leq \Pr[\bar{X} \geq \tau, \mathcal{F}] + \Pr[\mathcal{F}^c] \tag{11}$$

$$\leq \Pr[\bar{X} \geq \tau \mid \mathcal{F}] \Pr[\mathcal{F}] + \sum_{i=1}^{n} \Pr[F_i = 0] \tag{12}$$

$$\leq \exp(-2n(\tau - p_s)^2) + n\eta. \tag{13}$$

This proves the proposition.

B Implementation Details

All reported runs use frozen base VLMs. No optimizer, gradient update, or model fine-tuning is used. Table 3 lists all hyperparameters; the same values are used for every benchmark and backbone, with no per-benchmark tuning.

Table 3: Hyperparameters, fixed globally across benchmarks and backbones.

Component	Symbol	Value
Reliability prior; LCB	a_0, b_0; $\text{LCB}(\pi_s)$	$1, 1$; 10th pct. of $\text{Beta}(1 + n_s^+, 1 + n_s^-)$
Retrieval weights (Eq. 3)	$\alpha, \beta, \gamma, \zeta, \xi$	$1.0,\ 0.5,\ 0.25,\ 0.5,\ 0.05$
Retrieved items	k_s, k_h, k_a	$4,\ 2,\ 2$
Retrieval cutoff (rules, anchors)	–	cosine distance ≤ 0.80
Edits per hypothesis	K_e, K_o	$4,\ 4$
Re-answer samples (stochastic tasks)	M	3
Acceptance thresholds	$\tau_{\text{inv}}, \tau_{\text{nec}}, \tau_q$	$0.75,\ 0.75,\ 0.70$
Conflict (Eq. 2)	τ_{conf}	1; evidence cos < 0.50; scope cos ≥ 0.60
Pruning weights (Eq. 14)	$\lambda_\pi, \lambda_c, \lambda_r, \lambda_x$	$1.0,\ 0.3,\ 0.5,\ 1.0$
Replay reservoir	$\lvert\mathcal{R}_m\rvert$	≤ 32 episodes
Prompt budget; image cost	B; $\ell(m)$	12,000 tokens; 256 tokens per image
Sleep interval	–	every 10 completed tasks

Retrieval is deterministic and nothing is sampled at wake time. Multiple-choice answers are normalized to option letters. Open-ended answers use official benchmark evaluation where available; otherwise the equivalence function δ_Q normalizes case, punctuation, units, and common numeric formats before invoking the benchmark evaluator. No ground-truth labels are used by the memory writer. The stored answer A is always the model's original wake-time response before sleep-time editing or validation. Benchmark labels and official evaluators are used only for computing reported scores.

C Retention Value and Pruning

Each active item $m \in \mathcal{S} \cup \mathcal{H}$ receives the retention value

$$\Pi(m) = \text{gain}(m) + \lambda_\pi \,\text{LCB}(\pi_m) + \lambda_c \,\text{cov}_\sigma(m) - \lambda_r \,\text{red}(m) - \lambda_x \,\bar{\kappa}(m), \tag{14}$$

with $(\lambda_\pi, \lambda_c, \lambda_r, \lambda_x) = (1.0, 0.3, 0.5, 1.0)$ fixed globally; $\mathrm{gain}(m)$ is unweighted. All terms are normalized, and the weights were set once to prioritize reliability and conflict avoidance, with no per-benchmark or per-backbone tuning. The terms are defined as follows.

Replay gain. $\mathrm{gain}(m)$ maps an item to a real number measuring label-free replay self-agreement:

$$\mathrm{gain}(m) = \frac{1}{|\mathcal{R}_m|} \sum_{(I,Q,A)\in\mathcal{R}_m} \Big[\delta_Q\big(f(I,Q \mid \mathcal{M}_{\mathrm{base}} \cup \{m\}), A\big) - \delta_Q\big(f(I,Q \mid \mathcal{M}_{\mathrm{base}}), A\big)\Big], \quad (15)$$

where A is the VLM's own stored wake answer (no labels are used) and decoding is deterministic. The replay set $\mathcal{R}_m$ contains up to 32 earlier episodes kept by reservoir sampling; it excludes the source and validation episodes of m and is updated afterwards, which prevents self-evaluation. $\mathcal{M}_{\mathrm{base}}$ fixes the retrieved rules, hard negatives, anchors, and their order and excludes m, so the two calls differ only by m.

Scope coverage. $\mathrm{cov}_\sigma(m) = 1/(1 + N_{\mathrm{act}}(\sigma_m)) \in (0, 1]$, where σ_m is the categorical scope field of m and $N_{\mathrm{act}}(\sigma)$ is the number of currently active rules with scope σ; it favors underrepresented scopes.

Redundancy. $\mathrm{red}(m) \in [0, 1] = \mathrm{clip}\big(\max_{s\in\mathcal{S},\,\mathrm{LCB}(\pi_s)>\mathrm{LCB}(\pi_m)} \cos(\phi(r_m), \phi(r_s)), 0, 1\big)$, the maximum cosine similarity under the text encoder $\phi(\cdot)$ between the text of m and that of any active rule with higher LCB.

Contradiction term. $\bar{\kappa}(m) = \min(1, c_m/3) \in [0, 1]$, where the integer c_m counts contradiction events for m: (i) m loses a conflict comparison (Eq. 2) to a higher-LCB active rule and becomes dormant, or (ii) a later sleep re-validation of m fails the invariance or necessity gate. Each event also increments n_m^-; the term saturates after three events. It is distinct from the integer conflict count κ used for acceptance.

Budgeted selection. Under the budget B on cost $\ell(m)$ (prompt tokens plus image-anchor cost), the retained memory solves

$$\max_{\mathcal{M}'\subseteq\mathcal{M}} \sum_{m\in\mathcal{M}'} \Pi(m) \quad \text{s.t.} \quad \sum_{m\in\mathcal{M}'} \ell(m) \le B, \quad (16)$$

which the implementation approximates by a streaming greedy selection in decreasing order of $\Pi(m)/\ell(m)$.

D Qualitative Examples

Figure 5 illustrates two representative sleep-time outcomes. The successful case shows how a local visual episode is converted into a reusable semantic memory. The original question asks for the relationship between a suitcase and a book-like object. CausalMem separates the hypothesized causal evidence–the direct physical arrangement between the two objects–from nuisance context such as stickers, car interior, and lighting. The nuisance edit changes the surrounding scene while preserving the relation, so the answer remains stable; the object edit removes the relation, so the answer changes. Passing both tests promotes a rule that emphasizes physical contact and stacking over background details. When a later query asks for the relation between a man and a child, this validated rule helps the agent inspect spatial adjacency rather than overreacting to crop or framing artifacts.

The failure case shows the complementary behavior. For the concert image, the candidate hypothesis attributes the answer to grayscale appearance while treating stage lighting, shadows, and silhouettes as nuisance factors. The edited images do not support a consistent invariance-and-sensitivity pattern: the environment edit and object edit lead to incompatible answers, so the logical gate rejects the hypothesis. This rejected case is useful because it prevents a brittle color rule from entering semantic memory and can instead be stored as a hard negative or dormant counterexample.

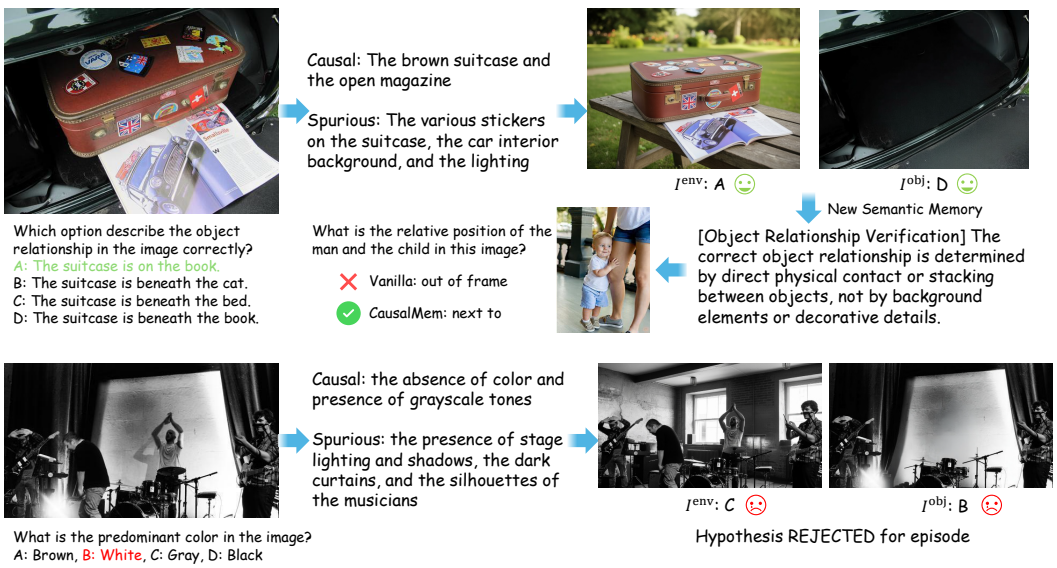


Figure 5: Qualitative examples of memory validation. Top: a candidate object-relation rule passes nuisance and causal edits, is promoted to semantic memory, and helps a later relation question. Bottom: an unreliable color hypothesis fails counterfactual validation and is rejected rather than written into long-term memory.

E Prompt Templates

Hypothesis generation prompt

You are a causal-analysis expert specializing in visual perception.
Given an image and a question-answer pair, your task is to perform a Structural Causal Analysis of the visual reasoning:

1. Identify Z_causal - the minimal set of visual features in the image that causally determine the correct answer. These are features whose removal would change the answer. Be specific (e.g. "the dog's erect triangular ears and curled tail" rather than "the dog").

2. Identify Z_spurious - visual features that are correlated with the answer in this image but are NOT causally necessary. Typical spurious confounders include: background/scene context, co-occurring objects, lighting or color palette, image style or quality.

Return your analysis in exactly the following JSON format (no extra keys):

```
“‘json
{
   "z_causal": "<concise description of causal features>",
   "z_spurious": "<concise description of spurious/confounding features>",
   "reasoning": "<one-paragraph justification>"
} “‘
```

Environment intervention edit prompt

Change the background/environment ([z_spurious]) to a completely different random setting. Keep the main subject ([z_causal]) exactly the same.

Object intervention edit prompt

Remove the main subject ([z_causal]) from the image completely. Fill the area naturally with the surrounding background ([z_spurious]). Keep the rest of the scene unchanged.

Wake-time prompt structure

Validated rules: the rule texts r of the retrieved rules (up to k_s).
Shortcut warnings: the texts of the retrieved hard negatives (up to k_h); their edited images are placed after the query image.
Anchor episodes: the retrieved validated question–answer pairs (up to k_a).
Question: the current question Q_t with the query image I_t.

Memory rule format

Scope: diagram type or visual reasoning skill.
Rule: concise invariant reasoning instruction.
Causal evidence: visual factors to inspect.
Nuisance factors: factors to ignore unless the question explicitly asks about them.
Validated by: numbers of nuisance and object edits with pass rates.
Counterexamples: conditions where the rule failed or should be dormant.
Retrieval keys: phrases and visual tags for matching future tasks.

F Additional Results

Table 4: Wake-context ablation (Qwen3-VL-8B, average of V* and MM-Vet).

Variant	Avg.	Δ vs. full
Full CausalMem	74.39	–
w/o hard negatives	73.89	-0.50
w/o anchor episodes	73.08	-1.31
Validated rules only	71.87	-2.52

Computational cost. Table 5 reports cost on V* with Qwen3-VL-8B; sleep costs are averaged over all evaluation samples. Wake overhead comes from the assembled context (including hard-negative edited images at 256 tokens per image), not from extra calls, and is bounded by k_s, k_h, k_a and B. The active store stays near constant under pruning (Figure 4b), so the wake context does not grow with the stream. Sleep scales with the stream and runs batched offline.

Table 5: Computational cost per sample on V* (Qwen3-VL-8B).

	Vanilla	CausalMem wake	CausalMem sleep
VLM calls	1	1	22.23
Editor calls	0	0	1.98
Wake latency	1.0×	1.86×	N/A (batched)
Extra prompt tokens	0	2,234	63,460
Time	0.71 s	1.32 s	20.97 s VLM + 130.87 s editor

Shortcut-content audit. To quantify what unvalidated memory and the write gate contain, we audited on V* the 77 passive-RAG memory items (written without validation) and the 104 hypotheses rejected by the gate. We use five shortcut categories aligned with MM-SpuBench (Ye et al., 2026): background/scene, co-occurring objects, color/lighting/style, layout/position, and text/watermark. An item matches if its stored text contains a category keyword; unclassified items count as non-matches.

Passive-RAG memory matches in 45.45% of cases (35/77), mainly layout/position (32/77); rejected hypotheses match in 94.23% of cases (98/104), mainly background/scene (79/104). Together with the 9.40-point drop of passive RAG below vanilla (Section 4.3), this shows that unvalidated memory can harm and that the gate filters shortcut-like content. Caveats: keyword matching is a coarse proxy; the audit does not prove that each match caused an error or that each rejected hypothesis was truly spurious; and the 94.23% rate describes what the gate filters, not shortcut prevalence in all model decisions.

G Additional Discussion and Limitations

CausalMem has several limitations. Text-guided image editing is only a noisy proxy for causal intervention: verification filters invalid edits but cannot guarantee that every unmentioned visual factor is preserved or that edited images remain on the true data manifold. The sleep phase adds substantial inference cost (on V* with Qwen3-VL-8B, 22.23 VLM calls, 1.98 editor calls, and about 152 s per sample, Table 5), and wake latency is 1.86× vanilla; the method is therefore best suited to agents that can consolidate memory offline or under a bounded update budget. The same VLM proposes hypotheses and re-answers and verifies edits, so errors shared by these roles may survive the gate; our cross-family experiments change the editor and the base VLM but do not separate the proposer from the verifier, and CausalMem has been demonstrated only on the tested families (Qwen2.5-VL, Qwen3-VL, InternVL3, InternVL3.5; Qwen-Image-Edit, InstructPix2Pix). The main results use transductive online evaluation; order permutations and an inductive split (Table 2) cover two benchmarks with one backbone. We report fixed global weights for Eqs. 3 and 14 but have not run a per-term sensitivity study beyond the component ablations. Finally, because semantic rules are written in natural language, they may be too broad, too narrow, or incorrect when the base VLM misidentifies the task scope or consistently misreads both original and edited images.

Broader Impacts. More reliable visual memory can reduce repeated reasoning errors and improve robustness under context shift. However, generated counterfactual images should be treated as diagnostic artifacts rather than real evidence, with provenance metadata or watermarking when released. Episodic memory may also contain sensitive visual information, so deployments should include data minimization, expiration policies, access control, and user-controlled deletion.