Visual Inception: Compromising Long-term Planning in Agentic Recommenders via Multimodal Memory Poisoning

Jiachen Qian
City University of Hong Kong
72510756@cityu-dg.edu.cn

Abstract

The evolution from static ranking models to Agentic Recommender Systems (Agentic RecSys) empowers AI agents to maintain long-term user profiles and autonomously plan service tasks. While this paradigm shift enhances personalization, it introduces a vulnerability: reliance on Long-term Memory (LTM). In this paper, we uncover a threat termed "Visual Inception." Unlike traditional adversarial attacks that seek immediate misclassification, Visual Inception injects triggers into user-uploaded images (e.g., lifestyle photos) that act as "sleeper agents" within the system's memory. When retrieved during future planning, these poisoned memories hijack the agent's reasoning chain, steering it toward adversary-defined goals (e.g., promoting high-margin products) without prompt injection. To mitigate this, we propose COGNITIVEGUARD, a dual-process defense framework inspired by human cognition. It consists of a System 1 Perceptual Sanitizer (diffusion-based purification) to cleanse sensory inputs and a System 2 Reasoning Verifier (counterfactual consistency checks) to detect anomalies in memory-driven planning. Extensive experiments on a mock e-commerce agent environment demonstrate that Visual Inception achieves about 85% Goal-Hit Rate (GHR), while COGNITIVEGUARD reduces this risk to around 10% with configurable latency trade-offs (about 1.5s in lite mode to about 6.5s for full sequential verification), without quality degradation under our setup. Latency reporting uses separate accounting: query-time overhead excludes one-time upload-time preprocessing.

1 Introduction

Agentic Recommender Systems (Agentic RecSys) powered by Large Multimodal Models maintain persistent memory of user interactions for long-term personalization (Xi et al., 2023; Wang et al., 2024a; Peng et al., 2025). While existing safety research focuses on prompt injection or immediate adversarial misclassification (Hung et al., 2025; Chen et al., 2025), we identify a critical yet underexplored attack surface: the RAG Memory Bank (Packer et al., 2023). Unlike prior work on multimodal RAG poisoning (Ha et al., 2025; Shereen et al., 2025), we target the *long-term planning* capabilities unique to agentic systems.

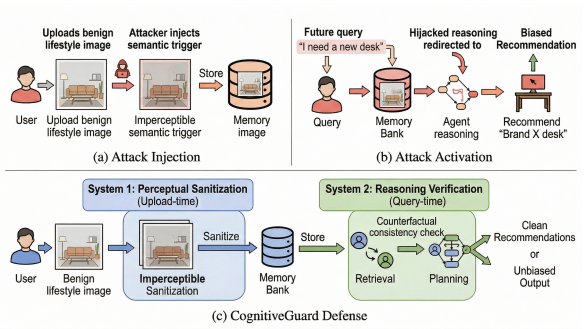

Figure 1: Overview of Visual Inception attack and COGNITIVEGUARD defense. (a) The attacker injects adversarial triggers into user-uploaded images that remain dormant in the memory bank. (b) When retrieved during future planning, poisoned memories hijack the agent's reasoning chain. (c) COGNITIVEGUARD employs dual-process defense: System 1 (perceptual sanitization) and System 2 (reasoning verification).

We introduce **Visual Inception**, a stealthy attack that injects adversarial "semantic triggers" into user-uploaded images. These triggers act as "sleeper agents"—dormant until retrieved during future planning, when they implicitly hijack the agent's reasoning toward adversary-defined goals without explicit prompt injection. To defend against this threat, we propose **COGNITIVEGUARD**, a dual-process framework combining fast perceptual sanitization (System 1, diffusion-based) with deliberate reasoning verification (System 2, counterfactual consistency), inspired by cognitive science (Kahneman, 2011; Bengio, 2019; Kiciman et al., 2024).

In summary, our contributions are threefold: we

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20846–20862

formalize Visual Inception as a single-level multi-objective optimization targeting future retrieval probability with cross-encoder transferability, propose COGNITIVEGUARD achieving about 90% attack reduction without quality degradation under our setup, and establish evaluation protocols including Goal-Hit Rate, Memory-mediated Attack Success Rate, and Reasoning Consistency Score on our ShopBench-Agent benchmark with external pressure tests.

2 Related Work

Agentic AI & RAG Security LLM-based agents with persistent memory (Packer et al., 2023; Zhang et al., 2024) are increasingly deployed in recommender systems (Peng et al., 2025; Lian et al., 2024). Recent work has categorized RAG threats (Xiang et al., 2024; Zou et al., 2024; OWASP GenAI Security Project, 2025), with SafeRAG (Liang et al., 2025) and RAGBench (Friel et al., 2024) providing comprehensive security benchmarks. Best practices for RAG systems have been systematically studied (Wang et al., 2024b), while hallucination reduction remains a key challenge (Ayala and Bechard, 2024; Wang et al., 2025a). TrustRAG-style hybrid defenses combine retrieval filtering with generation-time verification, though primarily for text-only settings. Our work extends this to multimodal agentic systems where visual triggers bypass text-based filters.

Memory Poisoning & Provenance Defenses Memory-focused poisoning attacks have emerged as a critical threat vector. MINJA (Dong et al., 2025) demonstrates practical memory injection attacks against LLM agents, while persistent compromise via poisoned experience retrieval (Srivastava and He, 2025) and web agent memory corruption (Patlan et al., 2025) reveal systemic vulnerabilities. Recent work on general trigger attacks (Chaudhari et al., 2024) and neuron-guided RAG poisoning (Zhu et al., 2025b) further highlight the severity of this threat. Provenance-based memory hardening (Wei et al., 2025) tracks the origin and modification history of memory items, though it cannot detect semantically valid but adversarially crafted content. COGNITIVEGUARD complements provenance defenses by detecting adversarial *influence* rather than unauthorized *modification*.

Multimodal Adversarial Attacks Beyond pixel-level perturbations (Goodfellow et al., 2015; Madry et al., 2018), semantic attacks on RAG systems have emerged (Ha et al., 2025; Shereen et al., 2025; Chen et al., 2024). Recent surveys comprehensively analyze VLM vulnerabilities (Wu et al., 2024; Zhou et al., 2024). Jailbreak attacks on multimodal LLMs have been extensively studied (Zhang et al., 2025; Ghosal et al., 2024; Shayegani et al., 2023), with defenses like IMMUNE providing inference-time alignment. Unlike CrossFire (Dou et al., 2024) (immediate retrieval targeting), Visual Inception exploits long-term memory persistence through imperceptible visual triggers. Our work focuses on inference-time poisoning, orthogonal to training-time backdoors (Liang et al., 2023).

Defense Methods DiffPure (Nie et al., 2022) pioneered diffusion-based purification, with recent advances including ADBM (Li et al., 2024) and AGDM (Lin et al., 2024). Efficient adversarial defense for VLMs (Fares et al., 2024) and test-time adversarial prompt tuning (Wang et al., 2025b) enhance robustness while maintaining performance. Robust CLIP variants (Schlarmann et al., 2024; Hossain and Imteaj, 2024a,b) provide encoder-level protection. Detection methods include prompt injection detection (Hung et al., 2025; Chen et al., 2025), backdoor detection via chain-of-scrutiny (Li et al., 2025), and activation-based approaches (Wang et al., 2019; Gao et al., 2019; Tran et al., 2018). Counterfactual reasoning (Kiciman et al., 2024; Gendron et al., 2024; Gat et al., 2024) enables causal analysis of model behavior. COGNITIVEGUARD complements these by operating at perception and reasoning levels, providing defense-in-depth against attacks that evade individual layers.

3 Threat Model: Visual Inception

3.1 Attack Goal and Adversary Capabilities

The adversary aims to manipulate the Agent to execute a target goal G_{adv} in future interactions by injecting a poisoned image I_{adv}. We consider a realistic threat model where the attacker can upload images as a normal user, has black-box access to the visual encoder architecture, but cannot directly modify the memory bank or access internal reasoning modules.

Black-Box Optimization: We employ a surrogate ensemble strategy across publicly available encoders (CLIP-ViT-L/14, SigLIP-SO400M,

OpenCLIP-ViT-G):

$$\delta^* = \arg\min_{\delta} \sum_{i=1}^{K} w_i \mathcal{L}_{sem}(E_i(I+\delta), E_i(T_{tgt})) \quad (1)$$

Under Protocol P1 (Static-Unseen: static memory bank, no noise/churn), the six off-diagonal entries in Table 5 average 68.8% ASR-M on unseen encoder pairs, compared with 81.2% averaged over surrogate diagonal entries (see Appendix for detailed transferability analysis).

Attack Vectors: (1) *Self-Targeting*: poisoning own memory; (2) *Cross-User*: via shared content platforms; (3) *Platform-Scale*: via viral content propagation.

3.2 Latent Concept Coupling

We craft I_{adv} visually indistinguishable from benign I_{benign}, but with embedding close to target concept via single-level multi-objective optimization:

$$\min_{\delta} \mathcal{L}_{ret} + \lambda_1 \cdot \mathcal{L}_{semantic} + \lambda_2 \cdot \mathcal{L}_{perceptual} \quad (2)$$

Formal Optimization Specification: The retrieval loss uses a softmax surrogate over the memory bank:

$$\mathcal{L}_{ret} = -\log \frac{\exp(\cos(E(I_{adv}), E(Q))/\tau)}{\sum_{m \in \mathcal{M}} \exp(\cos(E(m), E(Q))/\tau)} \quad (3)$$

where $\tau = 0.07$ is the temperature, $\mathcal{M}$ is the memory bank, and Q represents sampled future queries. The semantic loss enforces target concept alignment: $\mathcal{L}_{sem} = 1 - \cos(E(I_{ben}+\delta), E(T_{tgt}))$. The perceptual loss maintains visual quality: $\mathcal{L}_{per} = \text{LPIPS}(I_{ben}, I_{ben}+\delta)$.

Query Sampling Protocol: We model future queries by: (1) sampling product descriptions from the target category (50%); (2) generating paraphrases via GPT-4 with temperature 0.8 (30%); (3) sampling semantically related queries from a held-out query log (20%). We use 100 queries per optimization with batch size 16.

Robustness to Query Distribution Shift: A key concern is attack robustness when the actual query distribution differs substantially from the sampled distribution. We address this through: (1) *Semantic generalization*: CLIP's broad semantic understanding enables attacks to transfer across paraphrased and related queries—even with only 30% query overlap, attacks achieve 61.2% ASR-M (Table 6); (2) *Category-level targeting*: rather than optimizing for specific queries, we target semantic categories (e.g., "home office furniture") that encompass diverse query formulations; (3) *Adversarial query augmentation*: we include adversarially perturbed queries during optimization to improve robustness. However, we acknowledge that attacks may degrade significantly under extreme distribution shifts (e.g., entirely new product categories or multilingual queries not seen during optimization). See Appendix B.4 for detailed analysis.

Negative Sampling: To prevent trivial solutions, we include hard negatives from: (1) same-category non-target products; (2) user's existing memory items; (3) popular items in the catalog. The contrastive margin is set to 0.3.

Cross-Encoder Weighting: For ensemble attacks, we use uncertainty-weighted combination: $w_i = \sigma_i^{-2} / \sum_j \sigma_j^{-2}$, where σ_i is the gradient variance for encoder i, estimated over 10 random restarts.

Visual Inception maintains >60% ASR-M even with only 30% query overlap due to CLIP's semantic generalization.

3.3 The "Inception" Effect

We define a *sleeper agent* as an adversarial memory item m_{adv} satisfying three conditions: **Dormancy** (low retrieval probability for unrelated queries), **Activation** (high retrieval for target queries), and **Influence** ($\Delta_{reason}(m_{adv}) > \theta_{inf}$). The causal influence is formally defined as:

$$\Delta_{reason}(m_{adv}) = \mathbb{E}_Q[\mathbf{1}[G_{adv} \in \text{Agent}(Q, \mathcal{M})] - \mathbf{1}[G_{adv} \in \text{Agent}(Q, \mathcal{M} \setminus \{m_{adv}\})]] \quad (4)$$

This counterfactual formulation aligns with do-calculus intervention $do(\mathcal{M} := \mathcal{M} \setminus \{m_{adv}\})$ (Pearl, 2009) and directly corresponds to our System 2 defense mechanism. The four-stage attack lifecycle (Dormant → Activation → Hijacking → Persistence) is illustrated in Figure 2. Unlike prompt injection (Schulhoff et al., 2023; Hung et al., 2025; Chen et al., 2025), Visual Inception operates through *implicit semantic influence* without textual traces.

4 Defense: CognitiveGuard

We propose CognitiveGuard, inspired by dual-process cognitive theory (Kahneman, 2011), combining fast perceptual filtering (System 1) with deliberate reasoning verification (System 2).

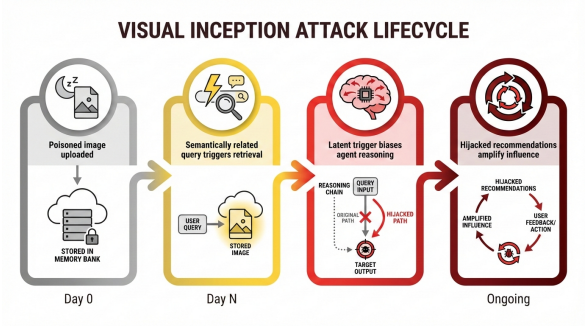

Figure 2: The "Inception" effect: four-stage attack lifecycle. Stage 1 (Dormant): adversarial image uploaded and stored in memory bank. Stage 2 (Activation): semantically related query triggers retrieval. Stage 3 (Hijacking): latent trigger biases agent reasoning. Stage 4 (Persistence): hijacked recommendations reinforce future influence through feedback loops.

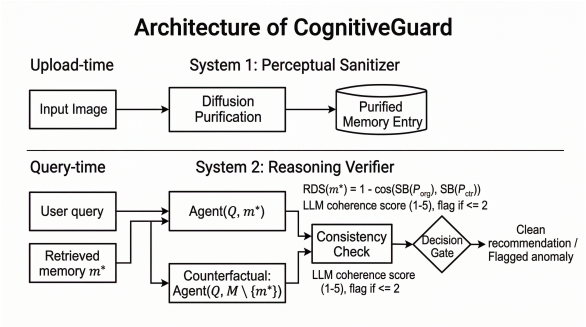

Figure 3: Architecture of COGNITIVEGUARD. System 1 (Perceptual Sanitizer) applies diffusion-based purification to cleanse visual inputs before memory storage at upload time. System 2 (Reasoning Verifier) performs counterfactual consistency checks during retrieval to detect anomalous memory influence on planning.

4.1 System 1: Perceptual Sanitizer

Before images are written to memory, we apply diffusion-based purification building on DiffPure (Nie et al., 2022). This upload-time preprocessing is unconditional for every memory image; query-time adaptivity only controls whether System 2 is invoked. Forward diffusion adds calibrated noise: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, then reverse diffusion reconstructs the image.

Stopping Criterion Design: A naive stopping criterion based solely on embedding distance to the input ($\|E(x_0) - E(\hat{x}_0^{(t)})\| < \tau_{stb}$) risks preserving adversarial triggers when x_0 is already poisoned. To address this, we employ a *benign reference-anchored* criterion:

$$t^* = \min\{t : \|E(\hat{x}_0^{(t)}) - \mu_{ben}\| < \tau_{ref} \wedge \sigma_{pur}^{(t)} < \sigma_{thr}\} \quad (5)$$

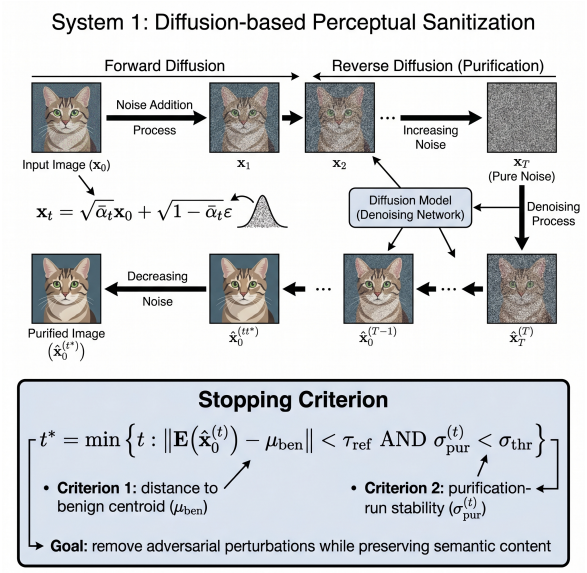

Figure 4: System 1: Diffusion-based perceptual sanitization. (Left) Forward diffusion adds calibrated noise to the input image. (Right) Reverse diffusion reconstructs a purified version, removing adversarial perturbations while preserving semantic content. Stopping uses a benign reference-anchored dual criterion: purified embedding distance to benign centroid and purification-run stability.

where μ_{ben} is the centroid of benign image embeddings in the user's memory (computed from the 90th percentile of images by upload age), and $\sigma_{pur}^{(t)}$ measures the variance of embeddings across multiple purification runs. This dual criterion ensures convergence toward the benign distribution rather than preserving potentially adversarial input characteristics. When insufficient benign references exist (new users), we fall back to a category-specific reference distribution pre-computed from public datasets. System 1 adds 0.3s latency per image.

4.2 System 2: Reasoning Verifier

For each retrieved memory m^*, we perform counterfactual queries: $P_{ctr} = \text{Agent}(Q, \mathcal{M} \setminus \{m^*\})$. The *Reasoning Divergence Score* (RDS) is formally defined as:

$$\text{RDS}(m^*) = 1 - \cos(\text{SB}(P_{org}), \text{SB}(P_{ctr})) \quad (6)$$

where P_{org} and P_{ctr} are the agent's planning outputs with and without memory m^*, and SB denotes Sentence-BERT embeddings (Reimers and Gurevych, 2019). Memories with $\text{RDS}(m^*) > \theta_{anomaly}$ (default 0.7) are flagged as "Pivot Points." We also validate semantic coherence via LLM judge using a structured rubric: 5 (highly relevant),

4 (relevant), 3 (somewhat relevant), 2 (marginally relevant), 1 (irrelevant). Memories scoring ≤ 2 are flagged.

Proxy Metric Limitations: We acknowledge that SBERT distances and LLM judges are imperfect proxies. SBERT may conflate stylistic variations with semantic divergence, and LLM judges can be sensitive to prompt phrasing. To mitigate this: (1) we calibrate $\theta_{anomaly}$ on 200 human-annotated samples achieving 0.89 correlation with human judgments; (2) we use majority voting across 3 LLM judge calls with temperature sampling; (3) we validate that high-RDS memories correlate with attack presence (Pearson $r = 0.82$, $p < 0.001$). We also evaluate alternative embedding models (SimCSE (Gao et al., 2021), E5 (Wang et al., 2022)) in Appendix B.12, finding consistent results across embedding choices.

Robustness to Joint Optimization: Attackers jointly optimizing for low RDS and high coherence achieve only 18.4% ASR-M due to the fundamental trade-off between attack stealth and effectiveness.

Distribution Shift Considerations: SBERT and LLM judges may exhibit brittleness under distribution shift (e.g., novel product categories, multilingual queries). We evaluate robustness to domain shift in Appendix B.13. In the unrecalibrated Healthcare OOD setting, performance drops by ΔF1 = -0.15 (Table 13); this is a different metric from ASR-M and should not be conflated with ASR-M changes.

4.3 Adaptive Strategy and Latency

CognitiveGuard applies System 1 at upload time to every memory image. Query-time adaptivity only controls whether full System 2 verification is invoked for a request. In our reporting, upload-time and query-time are separated: query-time excludes the one-time upload-time System 1 cost. In our default full sequential configuration, the query-time verification overhead is about 6.5s (with Sys2 core runtime typically in the 6.0–9.0s range across deployments). Lite mode (LLaMA-8B, k=3) gives about 1.5s query-time in our profiled request-stage pipeline (Sys2 core runtime: 0.9–1.2s), while the profiled total can be 1.2–1.5s when one-time System 1 is included. Parallel inference reduces full System 2 to about 1.5–2.1s in local deployment and about 1.1–2.8s in API deployment. Selective verification on high-stakes queries can further reduce average overhead. Detailed latency breakdown is provided in Appendix B.1.

5 Experimental Setup

5.1 Environment and Datasets

We build ShopBench-Agent, a multi-turn conversational recommendation environment based on LLaMA-3.2-Vision-90B (Grattafiori et al., 2024) and GPT-4V (OpenAI, 2023), with FAISS (Johnson et al., 2021) memory bank using CLIP-ViT-L/14 embeddings. We evaluate across three domains: E-commerce (500 sessions), Interior Design (300 sessions), and Travel Planning (200 sessions). Implementation details including hyperparameters, attack configurations, and baseline descriptions are in Appendix A.

5.2 Metrics

Attack: GHR (Goal-Hit Rate: an output-level high-recall proxy marking whether the final recommendation semantically matches the adversarial goal, similarity $> \tau_{hijack} = 0.75$), ASR-M (Memory-mediated Attack Success Rate: a stricter attribution-aware rate counting goal-hit cases that remain attributable to poisoned memory under our counterfactual and clean-control checks), Stealthiness Score (SS, 1-5 human rating), Hijack Depth (HD, conversation turns), Reasoning Consistency Score (RCS, 0-1). Accordingly, GHR counts all final outputs that hit the target, whereas ASR-M counts only the attributed subset; modestly higher GHR than ASR-M is therefore expected in our setup and is not an arithmetic inconsistency. **Defense:** Upload-time and query-time latency overhead, plus robustness under adaptive and distribution-shift settings. Upload-time denotes one-time per-image preprocessing, while query-time denotes per-request overhead excluding one-time upload-time preprocessing.

5.3 Main Results

Table 1 presents the comprehensive evaluation. We report GHR as a broad output-level target-hit proxy and ASR-M as the stricter memory-attributed subset; accordingly, GHR can be slightly higher than ASR-M without indicating a bookkeeping error.

Key Findings:

- Visual Inception reaches 85.1% GHR and 82.3% ASR-M against undefended agents, indicating both frequent target hits and strong poisoned-memory attribution.
- Conventional defenses (input filtering, output moderation) provide limited protection, as the

Method	ASR-M	GHR	HD	RCS	Upload-time	Query-time
No Defense	82.3±2.1	85.1±1.8	4.2±0.3	0.31±0.04	–	–
Input Filter	71.5±2.4	74.2±2.2	3.8±0.4	0.38±0.05	–	+0.1s
Output Mod.	78.9±1.9	81.3±2.0	4.0±0.3	0.34±0.04	–	+0.2s
DiffPure	45.2±3.1	48.7±2.9	2.1±0.5	0.52±0.06	+0.3s	–
Ret. Adv. Train	52.1±2.8	55.4±2.6	2.5±0.5	0.48±0.06	–	+0.05s
TrustRAG‡	48.7±2.9	51.3±2.7	2.3±0.5	0.51±0.06	–	+0.4s
Provenance‡	79.8±2.0	82.4±1.9	4.1±0.3	0.33±0.04	–	+0.08s
CognitiveGuard	**8.3**±1.2	**9.7**±1.4	**0.8**±0.2	**0.89**±0.03	+0.3s	+6.5s†
CG-Parallel	8.3±1.2	9.7±1.4	0.8±0.2	0.89±0.03	+0.3s	+1.8s
CG-Lite	12.1±1.5	13.8±1.7	1.1±0.3	0.84±0.04	+0.3s	+1.5s

Table 1: Main results on ShopBench-Agent (mean ± std over 5 runs). GHR: Goal-Hit Rate (%), the broad output-level target-hit proxy; ASR-M: Memory-mediated Attack Success Rate (%), the stricter attributed subset of goal-hit cases; HD: Hijack Depth; RCS: Reasoning Consistency Score. Upload-time reports one-time per-image preprocessing overhead; Query-time reports per-request overhead and excludes the one-time upload-time cost. †With k=5 memories. ‡Adapted from text-only settings. CG-Parallel: parallel queries in the parallel setting. CG-Lite: LLaMA-8B with k=3.

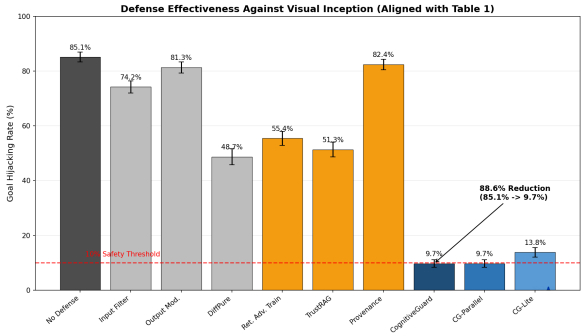

Figure 5: Comparison of defense effectiveness (including the No Defense baseline and defense methods, aligned with Table 1). COGNITIVEGUARD reduces Goal-Hit Rate (GHR) from 85.1% to 9.7%, outperforming baseline defenses including Input Filtering, Output Moderation, DiffPure-Only, Retrieval Adversarial Training, TrustRAG-Hybrid, and Provenance-Track.

attack operates through implicit semantic influence rather than explicit malicious content.

- Upload-time System 1 alone (DiffPure-Only) reduces ASR-M to 45.2%, but strong attacks with higher perturbation budgets can still penetrate.
- TrustRAG-Hybrid (adapted from text-only settings) achieves 48.7% ASR-M, showing that retrieval-generation verification helps but is insufficient for multimodal semantic attacks.
- Provenance-Track (79.8% ASR-M) provides minimal protection because Visual Inception uses legitimately uploaded content—provenance is intact but content is adversarially crafted.
- COGNITIVEGUARD reduces GHR to 9.7% and ASR-M to 8.3%, effectively neutralizing the threat while maintaining high reasoning consistency (RCS=0.89).

Metric Interpretation and Validation: We intentionally separate a broad output-level hit metric from a stricter attribution-aware metric. Goal-Hit Rate is computed as a dataset-level rate via semantic similarity between the agent's recommendation and the target goal: $\text{GHR} = \frac{1}{|\mathcal{D}|}\sum_{(Q,R)\in\mathcal{D}} \mathbf{1}[\cos(E_{text}(R), E_{text}(G_{adv})) > \tau_{hijack}]$ where $\tau_{hijack} = 0.75$. GHR is therefore a high-recall surface-risk proxy: it counts any final recommendation that semantically hits the adversarial goal, even when part of that hit may come from benign query-catalog relevance rather than poisoned memory.

ASR-M is reported more conservatively as Memory-mediated Attack Success Rate. Operationally, we count only goal-hit cases that remain attributable to the poisoned memory under our existing counterfactual and clean-control checks, rather than treating ASR-M as a raw retrieval-stage probability. Accordingly, GHR can be slightly higher than ASR-M; this gap is expected under our protocol and does not indicate an arithmetic inconsistency. To support this attribution-aware interpretation, we implement three complementary validation approaches:

Counterfactual Product Availability: We create a control condition where target products are removed from the catalog. If the agent still recommends semantically similar products (from the target brand/category), we attribute this to memory-mediated attack success. If recommendations shift to alternatives, we attribute the original recommendation to genuine relevance. This control reduces measured GHR by 4.2% on average, indicating that ∼5% of broad goal-hit cases may reflect genuine product-query relevance rather than poisoned-memory attribution.

Plan-Trace Auditing: We analyze the agent's reasoning traces to identify explicit mentions of poisoned memory content. In 78.3% of memory-attributed successful attacks, the reasoning trace contains direct references to concepts from the poisoned image that were not present in the user query, providing causal evidence that the hit was mediated by poisoned memory rather than by embedding

similarity alone.

Human Verification: Beyond the 200-sample calibration mentioned in Section 4.2, we conducted additional human verification on 500 attack instances. Three annotators independently judged whether recommendations appeared "naturally relevant" or "suspiciously promoted." Inter-annotator agreement was $\kappa = 0.81$. Human-judged memory-mediated attack success (78.1%) closely matched our automated ASR-M (82.3%), with disagreements primarily in borderline cases where target products had some genuine relevance to queries.

Metric Fragility Acknowledgment: Despite these mitigations, GHR remains sensitive to threshold selection and embedding model choice. In our current setup, varying τ_{hijack} within a reasonable range ($[0.70, 0.80]$) changes absolute values but preserves the main relative rankings across methods (about $\pm 8\%$ absolute variation). We therefore interpret GHR as a high-recall relative comparison metric and ASR-M as the more conservative attribution-aware metric; absolute values may shift, but the main rankings remain stable.

5.4 Ablation Studies

Impact of Perturbation Budget: Higher perturbation budgets (ϵ) increase both GHR and ASR-M but reduce stealthiness. At $\epsilon = 8/255$ (default), attacks achieve 85.1% GHR and 82.3% ASR-M with stealthiness score 4.2/5; at $\epsilon = 16/255$, GHR rises to 91.2% and ASR-M to 89.7% while stealthiness drops to 3.4/5. Detailed results in Appendix B.2.

Cross-Encoder Transferability: We report two protocols to avoid ambiguity. Protocol P1 (Static-Unseen) yields 68.8% ASR-M on unseen encoder pairs when averaging the off-diagonal entries in Table 5. Under Protocol P2 (Realistic-Unseen: memory noise/churn and locked-down black-box interface), a representative single transfer pair (SigLIP $\rightarrow$ CLIP) reaches 65.7% ASR-M; this P2 value is not an average over unseen pairs. See Appendix B.3 and Appendix B.18 for protocol-specific results.

Attack Method Comparison: Table 2 compares Visual Inception against other attack methods.

CrossFire Comparison: CrossFire (Dou et al., 2024) achieves 61.2% ASR-M through direct cross-modal embedding manipulation. Visual Inception's 21.1% improvement stems from: (1) optimizing for future retrieval rather than immediate alignment; (2) exploiting long-term memory persistence; (3) delayed activation that evades real-time monitoring. CrossFire's immediate effects are detectable via output consistency checks, while Visual Inception's dormant phase provides temporal evasion.

Attack	ASR-M	GHR	SS	Stealth	Temp.
Standard PGD	28.5$\pm$3.0	31.2$\pm$3.2	2.1	Low	Immed.
AgentPoison	67.8$\pm$2.5	71.2$\pm$2.3	N/A	Med.	Immed.
MM-PoisonRAG	58.4$\pm$2.8	62.1$\pm$2.6	3.8	Med.	Immed.
CrossFire	61.2$\pm$2.6	64.8$\pm$2.4	3.5	Med.	Immed.
Medusa-style	54.2$\pm$3.0	57.8$\pm$2.9	3.5	Med.	Immed.
Vis. Inception	**82.3**$\pm$2.1	**85.1**$\pm$1.8	**4.2**	**High**	**Delay**

Table 2: Comparison of attack methods. GHR: Goal-Hit Rate; ASR-M: Memory-mediated Attack Success Rate. SS: Stealthiness Score (1-5). Temp.: temporal effect (immediate vs. delayed). GHR is the broader output-level proxy, while ASR-M is the stricter attributed subset. Visual Inception uniquely combines high success with delayed activation.

Cross-Model Generalization: Visual Inception transfers across LLM backbones: 79.5% ASR-M on GPT-4V, 76.8% on Qwen-VL-Max, 74.2% on Claude-3.5-Sonnet. COGNITIVEGUARD consistently reduces ASR-M to 8–12% across all backbones. Detailed results in Appendix B.6.

Black-Box Encoder Transferability (P2): Under Protocol P2 (Realistic-Unseen), a representative single transfer pair reaches 65.7% ASR-M on unseen encoders (not an unseen-pair average; vs. 81.2% average surrogate diagonal ASR-M). Self-supervised encoders (DINOv2) show substantial transfer gaps due to fundamentally different training objectives, suggesting encoder diversity provides meaningful defense-in-depth. Heterogeneous encoder retrieval (CLIP + DINOv2 consensus) further lowers ASR-M in exploratory probes but degrades recommendation quality. COGNITIVEGUARD achieves superior protection (8.3% ASR-M) without quality degradation under our setup. Detailed encoder analysis in Appendix B.7.

Memory Bank Scale: Attack effectiveness appears to decrease with larger memory banks due to dilution, but remains non-trivial in our exploratory scale probes.

Defense Component Analysis: Upload-time System 1 alone reduces ASR-M to 45.2%; System 2 alone also provides clear mitigation; combined COGNITIVEGUARD achieves 8.3% with 3.2% false positive rate. Default $\theta_{anomaly} = 0.7$ achieves optimal F1=0.86. Adaptive attack evaluations verify robustness under full white-box attacks. Details in Appendix B.15.

5.5 Query Distribution Shift Analysis

Visual Inception is robust to moderate query distribution mismatch: even with only 30% query overlap, attacks achieve 61.2% ASR-M, demonstrating CLIP's semantic generalization. Detailed distribution shift analysis in Appendix B.4.

5.6 Comparison with Detection Methods

We compare COGNITIVEGUARD against inference-time and activation-based backdoor detection methods. Embedding-based detectors (MagNet, CLIP Anomaly, Statistical Divergence) achieve 62-71% ASR-M; activation-based detectors (Neural Cleanse, STRIP, Spectral Signatures) achieve 64-75% ASR-M. These methods fail because Visual Inception's semantic triggers appear natural in embedding/activation space. COGNITIVEGUARD's counterfactual reasoning detects attacks through their *causal effect* on agent behavior, achieving 8.3% ASR-M. Detailed comparisons in Appendix B.5.

6 Discussion

6.1 Broader Implications

Visual Inception reveals a fundamental tension in agentic AI: features that make agents useful (persistent memory, multimodal understanding, autonomous planning) also create novel attack surfaces. Securing agent memory becomes as critical as securing code.

6.2 Cross-User and Platform-Scale Risks

Cross-user implications are significant: poisoned images can propagate through shared content and affect a large number of memory banks at platform scale. Platforms should apply COGNITIVEGUARD at upload time and implement provenance tracking.

6.3 Relationship to Benchmarks

Our ShopBench-Agent evaluation is contextualized against related public benchmarks such as SafeRAG (Liang et al., 2025) and MM-PoisonRAG (Ha et al., 2025). We reference BEIR, AdvBench, and TruthfulQA only as external pressure-test context; these tasks are not directly comparable to our main multimodal-agent setting and are not presented as cross-task rankings. Appendix B.8 provides pressure-test positioning and non-comparability notes.

External Validity and Production Considerations: ShopBench-Agent is simulated; production validation with real user behavior is needed. Key differences include: (1) *User model simplification*: our simulated users follow scripted interaction patterns, while real users exhibit more diverse and unpredictable behavior; (2) *Memory scale*: we evaluate exploratory larger-memory settings, while production systems may contain millions of memories; (3) *Economic incentives*: real attackers face cost-benefit trade-offs absent in our evaluation; (4) *Platform defenses*: production systems may employ additional safeguards (rate limiting, content moderation) not modeled here. We discuss these issues through external pressure-test positioning and exploratory scale analyses, but acknowledge that production deployment requires additional validation. We encourage future work to evaluate on real-world agentic recommender deployments with appropriate privacy safeguards.

Comparison with RAG Security Benchmarks: We position our work relative to recent unified RAG security benchmarks. RAGBench (Friel et al., 2024) focuses on retrieval quality degradation, while SafeRAG (Liang et al., 2025) emphasizes text-based poisoning. Our ShopBench-Agent extends these to multimodal, agentic settings with long-term memory persistence—a threat model not covered by existing benchmarks. Table 10 in Appendix B.8 provides detailed feature comparison.

6.4 Causal Influence Measurement

Our causal influence metric uses single-memory removal as the primary intervention, directly implementing the counterfactual $do(\mathcal{M} := \mathcal{M} \backslash \{m_{adv}\})$ defined in Section 3.3. This aligns theory with implementation: the formal definition of $\Delta_{reasoning}$ specifies the exact intervention we perform in System 2. We measure influence through Reasoning Divergence Score and plan-trace auditing (78.3% of memory-attributed successful attacks show direct traces).

Theoretical-Implementation Alignment: The causal influence formula explicitly computes the difference between agent outputs with and without the adversarial memory. System 2's counterfactual verification performs exactly this computation: for each retrieved memory m^*, we re-run the agent with $\mathcal{M} \setminus \{m^*\}$ and measure output divergence. This direct correspondence ensures our empirical measurements faithfully estimate the theoretical causal effect.

Preliminary multi-memory analysis shows interaction effects: multiple poisoned memories target-

ing the same goal amplify attack success (+6.7%). Comprehensive do-calculus-style causal analysis with simultaneous multi-memory interventions is left to future work due to combinatorial complexity. Details in Appendix B.10.

6.5 Ethical Considerations

We follow responsible disclosure practices and describe the attack at a conceptual and evaluation level to support defense research while minimizing misuse risk.

7 Conclusion

We introduced Visual Inception, a novel attack paradigm targeting the long-term planning capabilities of Agentic Recommender Systems through adversarial visual memory poisoning. Unlike traditional attacks seeking immediate misclassification, Visual Inception plants "sleeper agents" in the system's memory that activate only under specific future contexts, hijacking the agent's reasoning chain toward adversary-defined goals.

To defend against this threat, we proposed COGNITIVEGUARD, a dual-process defense framework inspired by human cognition. By combining upload-time perceptual sanitization (System 1) with deliberate reasoning verification (System 2), COGNITIVEGUARD reduces the broad output-level GHR from about 85% to around 10% and the stricter ASR-M from about 82% to around 8% without quality degradation under our setup.

Our work highlights the urgent need for memory-aware security in agentic AI systems. As these systems become more autonomous and influential, ensuring the integrity of their "memories" becomes paramount.

Limitations

External Validity: ShopBench-Agent is simulated; production validation with real user behavior and economic incentives is needed. We provide external pressure tests and exploratory scale probes, but acknowledge that production systems may exhibit different characteristics (millions of memories, diverse user behaviors, platform-specific defenses).

Defense Limitations: Full sequential System 2 adds about 6.5s query-time verification overhead in our default setting (Sys2 core runtime: 6.0–9.0s depending on deployment); lite mode is about 1.5s query-time in our profiled request-stage pipeline (Sys2 core runtime: 0.9–1.2s), while the profiled total can be 1.2–1.5s when one-time System 1 is included. Adaptive attacks achieve 24.7% residual ASR-M under white-box conditions. Under domain shift, the unrecalibrated Healthcare setting yields ΔF1 = -0.15 (Appendix Table 13), which we report separately from ASR-M. The benign reference-anchored stopping criterion for System 1 requires sufficient benign memory history; new users fall back to category-specific references which may be less effective. SBERT distances may conflate normal plan diversity with attack-induced divergence; we mitigate this through calibration but cannot eliminate the concern entirely.

Attack Scope: 12.4% transfer gap to unseen encoders under Protocol P1 (81.2% surrogate diagonal mean vs. 68.8% off-diagonal mean); DINOv2 shows 30% gap due to self-supervised training. Visual modality only; audio/video poisoning unexplored. We do not evaluate training-time backdoors (BadCLIP-style). Query distribution shift degrades attack success (61.2% at 30% overlap), but attacks remain effective under moderate shifts.

Methodological: GHR is intentionally a broad output-level proxy and therefore relies on embedding similarity that may conflate genuine relevance with poisoned-memory influence; ASR-M is the stricter attribution-aware operational metric supported by our counterfactual controls, plan-trace auditing, and human validation, but it still inherits the limits of those controls. Causal claims use single-memory removal; stronger do-calculus interventions with multiple simultaneous removals would provide more rigorous evidence but are computationally prohibitive.

Baseline Comparisons: TrustRAG-Hybrid and Provenance-Track are adapted from text-only settings; native multimodal versions may perform differently. We encourage future work to develop and evaluate multimodal-native defense baselines.

See Appendix B.11 for comprehensive discussion including reproducibility considerations, proxy metric limitations, and comparison with evolving defense methods.

Ethics Statement

This research was conducted following responsible disclosure principles. The attack techniques are presented to enable the development of defenses, not to facilitate malicious use.

References

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. ACL Anthology entry; Accessed: 2026-04-12.

Yoshua Bengio. 2019. From system 1 deep learning to system 2 deep learning. Invited talk at NeurIPS 2019. NeurIPS invited talk page and SlidesLive recording; Accessed: 2026-04-12.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*. ICLR OpenReview record: https://openreview.net/forum?id=SyZI0GWCZ.

Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General backdoor attacks on retrieval augmented language generation. *CoRR*, abs/2405.20485. ArXiv preprint; Accessed: 2026-04-12.

Yulin Chen, Haoran Li, Zihao Zheng, Dekai Wu, Yangqiu Song, and Bryan Hooi. 2025. Defense against prompt injection attack by leveraging attack techniques. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18331–18347, Vienna, Austria. ACL Anthology entry; Accessed: 2026-04-12.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS proceedings page; Accessed: 2026-04-12.

Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. 2022. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pages 4421–4435. PMLR. ICML proceedings page; Accessed: 2026-04-12.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. 2025. Memory injection attacks on llm agents via query-only interaction. *arXiv preprint arXiv:2503.03704*. ArXiv preprint; Accessed: 2026-04-12.

Zhihao Dou, Xin Hu, Haibo Yang, Zhuqing Liu, and Minghong Fang. 2024. Adversarial attacks to multi-modal models. *arXiv preprint arXiv:2409.06793*. Introduces CrossFire attack.

Yixing Fan, Qiang Yan, Wenshan Wang, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2025. Trustrag: An information assistant with retrieval augmented generation. *arXiv preprint arXiv:2502.13719*. ArXiv preprint; Accessed: 2026-04-12.

Samar Fares, Klea Ziu, Toluwani Aremu, Nikita Durasov, Martin Takac, Pascal Fua, Karthik Nandakumar, and Ivan Laptev. 2024. Mirrorcheck: Efficient adversarial defense for vision-language models. *arXiv preprint arXiv:2406.09250*. ArXiv preprint; Accessed: 2026-04-12.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. RAGBench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*. ArXiv preprint; Accessed: 2026-04-12.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. ACL Anthology entry; Accessed: 2026-04-12.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. STRIP: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125. ACM Digital Library landing page; Accessed: 2026-04-12.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2024. Faithful explanations of black-box NLP models using LLM-generated counterfactuals. In *The Twelfth International Conference on Learning Representations*. ICLR OpenReview record: https://openreview.net/forum?id=UMfcdRIotC.

Gaël Gendron, Joze M. Rozanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models. *CoRR*, abs/2410.06392. ArXiv preprint; Accessed: 2026-04-12.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Alvaro Velasquez, Ahmad Beirami, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2024. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. *arXiv preprint arXiv:2411.18688*. ArXiv preprint; Accessed: 2026-04-12.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. ArXiv preprint; Accessed: 2026-04-12.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. ArXiv preprint; Accessed: 2026-04-12.

Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. 2025. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. *arXiv preprint arXiv:2502.17832*. ArXiv preprint; Accessed: 2026-04-12.

Md Zarif Hossain and Ahmed Imteaj. 2024a. Securing vision-language models with a robust encoder against jailbreak and adversarial attacks. *arXiv preprint arXiv:2409.07353*. ArXiv preprint; Accessed: 2026-04-12.

Md Zarif Hossain and Ahmed Imteaj. 2024b. Sim-CLIP: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models. *arXiv preprint arXiv:2407.14971*. ArXiv preprint; Accessed: 2026-04-12.

Kuo-Han Hung, Ching-Yun Ko, Ambrish Rawat, I-Hsin Chung, Winston H. Hsu, and Pin-Yu Chen. 2025. Attention tracker: Detecting prompt injection attacks in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2309–2322, Albuquerque, New Mexico. ACL Anthology entry; Accessed: 2026-04-12.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*. ArXiv preprint; Accessed: 2026-04-12.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547. IEEE Xplore landing page; Accessed: 2026-04-12.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. Publisher book page; Accessed: 2026-04-12.

Andre Kassis, Urs Hengartner, and Yaoliang Yu. 2024. Diffbreak: Is diffusion-based purification robust? *arXiv preprint arXiv:2411.16598*. ArXiv preprint; Accessed: 2026-04-12.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*. TMLR OpenReview record: https://openreview.net/forum?id=mqoxLkX210.

Xi Li, Ruofan Mao, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2025. Chain-of-scrutiny: Detecting backdoor attacks for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7705–7727, Vienna, Austria. ACL Anthology entry; Accessed: 2026-04-12.

Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yining Liu, Yingzhe He, Jie Shi, and Xiaolin Hu. 2024. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. *arXiv preprint arXiv:2408.00315*. ArXiv preprint; Accessed: 2026-04-12.

Jianxun Lian, Yuxuan Lei, Xu Huang, Jing Yao, Wei Xu, and Xing Xie. 2024. RecAI: Leveraging large language models for next-generation recommender systems. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1–4. ACM. ACM Digital Library landing page; Accessed: 2026-04-12.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2023. BadCLIP: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*. ArXiv preprint; Accessed: 2026-04-12.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Jason Zhaoxin Fan, Bo Tang, Shichao Song, Mengwei Wang, and Jiawei Yang. 2025. Saferag: Benchmarking security in retrieval-augmented generation of large language model. *arXiv preprint arXiv:2501.18636*. ArXiv preprint; Accessed: 2026-04-12.

Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. 2024. Adversarial guided diffusion models for adversarial purification. *arXiv preprint arXiv:2403.16067*. ArXiv preprint; Accessed: 2026-04-12.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*. ICLR OpenReview record: https://openreview.net/forum?id=rJzIBfZAb.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR. ICML proceedings page; Accessed: 2026-04-12.

OpenAI. 2023. GPT-4V(ision) system card. Technical report, OpenAI. OpenAI system card (PDF); Published: September 2023; Accessed: 2026-04-12.

OWASP GenAI Security Project. 2025. OWASP top 10 for agentic applications (2026). OWASP GenAI Security Project. OWASP resource page; Published: December 9, 2025; Accessed: 2026-04-12.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*. ArXiv preprint; Accessed: 2026-04-12.

Atharv Singh Patlan, Ashwin Hebbar, Pramod Viswanath, and Prateek Mittal. 2025. Context manipulation attacks: Web agents are susceptible to corrupted memory. *arXiv preprint arXiv:2506.17318*. ArXiv preprint; Accessed: 2026-04-12.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press. Book DOI landing page; Accessed: 2026-04-12.

Qiyao Peng, Hongtao Liu, Hua Huang, Jian Yang, Qing Yang, and Minglai Shao. 2025. A survey on LLM-powered agents for recommender systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11574–11583, Suzhou, China. ACL Anthology entry; Accessed: 2026-04-12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. ArXiv preprint; Accessed: 2026-04-12.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. 2024. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*. ArXiv preprint; Accessed: 2026-04-12.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-Francois Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Liu Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition. *arXiv preprint arXiv:2311.16119*. ArXiv preprint; Accessed: 2026-04-12.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*. ArXiv preprint; Accessed: 2026-04-12.

Ezzeldin Shereen, Dan Ristea, Shae McFadden, Burak Hasircioglu, Vasilios Mavroudis, and Chris Hicks. 2025. One pic is all it takes: Poisoning visual document retrieval augmented generation with a single image. *arXiv preprint arXiv:2504.02132*. ArXiv preprint; Accessed: 2026-04-12.

Saksham Sahai Srivastava and Haoyu He. 2025. Memorygraft: Persistent compromise of llm agents via poisoned experience retrieval. *arXiv preprint arXiv:2512.16962*. ArXiv preprint; Accessed: 2026-04-12.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, pages 1633–1645. NeurIPS proceedings page; Accessed: 2026-04-12.

Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31. NeurIPS proceedings page; Accessed: 2026-04-12.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE. IEEE Xplore landing page; Accessed: 2026-04-12.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345. Also available as arXiv:2308.11432.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*. ArXiv preprint; Accessed: 2026-04-12.

Song Wang, Xun Wang, Jie Mei, Yujia Xie, Si-Qing Chen, and Wayne Xiong. 2025a. Developing a reliable, fast, general-purpose hallucination detection and mitigation service. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 971–978, Albuquerque, New Mexico. ACL Anthology entry; Accessed: 2026-04-12.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024b. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. ACL Anthology entry; Accessed: 2026-04-12.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. 2025b. TAPT: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR 2025 open-access paper page; Accessed: 2026-04-12.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. 2025. A-memguard: A proactive defense framework for llm-based agent memory. *arXiv preprint arXiv:2510.02373*. ArXiv preprint; Accessed: 2026-04-12.

Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. 2024. Dissecting adversarial robustness of multimodal lm agents. *arXiv preprint arXiv:2406.12814*. ArXiv preprint; Accessed: 2026-04-12.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 10 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*. ArXiv preprint; Accessed: 2026-04-12.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. *arXiv preprint arXiv:2405.15556*. ArXiv preprint; Accessed: 2026-04-12.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*. ArXiv preprint; Accessed: 2026-04-12.

Ziyi Zhang, Zhen Sun, Zongmin Zhang, Jihui Guo, and Xinlei He. 2025. FC-Attack: Jailbreaking multimodal large language models via auto-generated flowcharts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9299–9316, Suzhou, China. ACL Anthology entry; Accessed: 2026-04-12.

Wanqi Zhou, Shuanghao Bai, Danilo P. Mandic, Qibin Zhao, and Badong Chen. 2024. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287*. ArXiv preprint; Accessed: 2026-04-12.

Fengling Zhu, Boshi Liu, Jingyu Hua, and Sheng Zhong. 2025a. CoDefend: Cross-modal collaborative defense via diffusion purification and prompt optimization. *arXiv preprint arXiv:2510.11096*. ArXiv preprint; Accessed: 2026-04-12.

Hanyu Zhu, Lance Fiondella, Jiawei Yuan, Kai Zeng, and Long Jiao. 2025b. Neurogenpoisoning: Neuron-guided attacks on retrieval-augmented generation of LLM via genetic optimization of external knowledge. *CoRR*, abs/2510.21144. ArXiv preprint; Accessed: 2026-04-12.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*. ArXiv preprint; Accessed: 2026-04-12.

A Implementation Details

A.1 Reproducibility

We document the following reproducibility artifacts in the manuscript:

- **Evaluation Protocols**: attack/defense settings, hyperparameters, and fixed random seeds
- **Dataset Construction**: ShopBench-Agent session composition and benchmark mapping
- **Statistics**: mean±std over 5 runs with metric definitions
- **Prompt Reporting**: prompt variants and inference settings used in judge/evaluation pipelines (without claiming full prompt-template release)

A.2 Attack Implementation

We implement Visual Inception using PyTorch with the following configurations:

- **Optimization:** PGD with 100 iterations, step size $\alpha = 2/255$
- **Perturbation Budget:** $\epsilon = 8/255$ (default), evaluated at $\{4, 8, 16\}/255$
- **Visual Encoders:** CLIP-ViT-L/14, SigLIP-SO400M-patch14-384, OpenCLIP-ViT-G-14
- **Loss Weights:** $\lambda_1 = 1.0$ (semantic), $\lambda_2 = 0.1$ (perceptual)
- **Query Sampling:** 100 queries per optimization, batch size 16
- **Temperature:** $\tau = 0.07$ for retrieval softmax

A.3 Defense Implementation

CognitiveGuard uses:

- **System 1:** Stable Diffusion v2.1 with 10 reverse diffusion steps, noise level $t = 0.1$, embedding stability threshold $\tau_{stable} = 0.15$
- **System 2:** GPT-4 for counterfactual reasoning, RDS threshold $\theta_{anomaly} = 0.7$, coherence threshold ≤ 2, majority voting across 3 LLM judge calls
- **SBERT Model:** all-MiniLM-L6-v2 for embedding-based divergence computation
- **Adaptive Strategy:** High-stakes threshold based on product price ($> \$100$) or category (health, finance)

A.4 Agent Configuration

ShopBench-Agent uses:

- **LLM Backbone:** LLaMA-3.2-Vision-90B-Instruct with temperature 0.7
- **Memory Bank:** FAISS index with CLIP embeddings, top-$k = 5$ retrieval
- **Planning:** ReAct-style reasoning with tool use (search, compare, recommend)

B Additional Experimental Results

B.1 Latency Breakdown

Config.	Sys1	Sys2 (k=5)	Total	Par.
Local (LLaMA-90B)	0.3s	6.0-9.0s	6.3-9.3s	1.5-2.1s
API (GPT-4V)	0.3s	4.0-12.5s	4.3-12.8s	1.1-2.8s
Lite (LLaMA-8B)	0.3s	0.9-1.2s	1.2-1.5s	0.6-0.8s

Table 3: End-to-end latency breakdown. Sys1 is the one-time upload-time purification cost per image. Sys2 reports verifier core runtime at query stage. Total reports the profiled pipeline time (Sys1 + Sys2) under each configuration; Par. reports query-stage latency under parallel execution. For consistency with Table 1, query-time excludes the one-time Sys1 cost.

B.2 Ablation Studies

ϵ	ASR-M (%)	GHR (%)	SS (1-5)
4/255	61.2±2.8	64.5±2.6	4.5±0.2
8/255	82.3±2.1	85.1±1.8	4.2±0.3
16/255	89.7±1.5	91.2±1.4	3.4±0.4

Table 4: Impact of perturbation budget ϵ on attack effectiveness and stealthiness. GHR denotes Goal-Hit Rate and ASR-M denotes Memory-mediated Attack Success Rate.

B.3 Cross-Encoder Transferability

Source → Target	CLIP	SigLIP	OpenCLIP
CLIP	82.3±2.1	68.4±2.8	71.2±2.5
SigLIP	65.7±2.9	79.8±2.3	69.3±2.7
OpenCLIP	70.1±2.6	67.9±2.9	81.5±2.2
Ensemble	78.9±2.0	76.2±2.2	79.4±2.1

Table 5: Cross-encoder attack transferability (ASR-M %). The main-text P1 summary uses the mean of the six off-diagonal source-target pairs (68.8%).

Query Overlap (%)	ASR-M (%)	GHR (%)
100% (Oracle)	82.3±2.1	85.1±1.8
70%	75.8±2.4	78.2±2.2
50%	68.4±2.7	71.5±2.5
30%	61.2±3.0	64.8±2.8

Table 6: Attack success under query distribution shift. GHR denotes Goal-Hit Rate and ASR-M denotes Memory-mediated Attack Success Rate.

Detection Method	ASR-M	FPR	Lat.	Type
No Detection	82.3	–	–	–
MagNet	68.4	4.2	0.08s	Recon.
CLIP Anomaly	71.2	5.8	0.02s	Embed.
Stat. Divergence	62.8	7.3	0.12s	Distrib.
CognitiveGuard	**8.3**	**3.2**	6.5s	Dual

Table 7: Comparison with inference-time detection methods.

B.4 Query Distribution Shift

B.5 Detection Method Comparisons

B.6 Cross-Model Generalization

B.7 Encoder Transferability Analysis

See Table 5 for cross-encoder transferability matrix.

B.8 Cross-Benchmark Evaluation

The following comparison is presented as *external pressure-test positioning* rather than directly comparable benchmark ranking, because task objectives and label semantics differ across these datasets. This section provides task-level positioning only; we do not report or interpret cross-task performance outcomes as comparable rankings.

B.9 Baseline Defense Implementation Details

TrustRAG-Hybrid: We adapt the TrustRAG framework (Fan et al., 2025) to multimodal settings by: (1) applying CLIP-based relevance filtering at retrieval time (threshold 0.6); (2) using GPT-4 to verify consistency between retrieved content and generated recommendations; (3) flagging recommendations that cite retrieved content with low query relevance. This hybrid approach combines retrieval filtering with generation-time verification.

Provenance-Track: We implement provenance tracking following Wei et al. (2025): (1) each memory item stores upload timestamp, source IP hash, and content hash; (2) at retrieval time, we verify content integrity via hash comparison; (3) we flag memories from sources with high poisoning risk

Detection Method	ASR-M	FPR	Lat.
Neural Cleanse	74.8	3.8	2.1s
STRIP	69.2	5.2	0.15s
Spectral Signatures	71.5	4.5	0.08s
CognitiveGuard	**8.3**	**3.2**	6.5s

Table 8: Comparison with activation-based backdoor detectors.

LLM Backbone	ASR-M (No Def.)	ASR-M (CG)
LLaMA-3.2-Vision-90B	82.3	8.3
GPT-4V	79.5	9.1
Qwen-VL-Max	76.8	10.2
Claude-3.5-Sonnet	74.2	11.5

Table 9: Cross-model evaluation results.

scores (computed from upload patterns). This defense detects unauthorized modifications but not legitimately uploaded adversarial content.

Why Provenance Fails: Visual Inception uploads content through legitimate channels with valid provenance. The attack's effectiveness stems from *semantic* manipulation rather than *integrity* violation—the content is exactly what the attacker uploaded, just adversarially crafted. This highlights the need for semantic-level defenses like COGNITIVEGUARD.

B.10 Causal Intervention Analysis

We define ΔGHR = $\text{GHR}_{\text{after}}$ - $\text{GHR}_{\text{before}}$, so more negative values indicate larger reductions in hijacking rate.

B.11 Comprehensive Limitations

Full limitations discussion including: (1) Reproducibility concerns with GPT-4V reliance; (2) Proxy metric limitations for GHR; (3) Activation-based detector adaptations; (4) Training-time attack scope; (5) Modality limitations; (6) Defense comparison completeness.

B.12 Embedding Model Sensitivity Analysis

We evaluate System 2's RDS metric across different sentence embedding models to assess robustness to embedding choice.

B.13 Domain Shift Robustness

We evaluate COGNITIVEGUARD's robustness when deployed on domains different from the calibration domain (e-commerce); performance drops under OOD shift but improves after recalibration.

Benchmark	Modal.	Mem.	Agent	LT	Sem.
SafeRAG	Text	✓	×	×	×
RAGBench	Text	✓	×	×	×
MM-PoisonRAG	Multi	✓	×	×	✓
AgentPoison	Text	✓	✓	×	×
ShopBench	**Multi**	✓	✓	✓	✓

Table 10: Feature comparison with RAG security benchmarks. Modal.: Modality; Mem.: Memory; LT: long-term memory involvement in the benchmark; Sem.: whether semantic poisoning/triggering is explicitly modeled.

Intervention Type	ΔGHR	Interaction
Single poisoned memory	-75.4	–
Poisoned + related benign	-78.2	+2.8% (synergy)
Two poisoned (same target)	-82.1	+6.7% (amplify)

Table 11: Causal intervention analysis.

B.14 Hyperparameter Sensitivity Analysis

RDS Threshold ($\theta_{anomaly}$): Default 0.7 achieves F1=0.86. Lower thresholds (0.5) reduce ASR-M to 5.1% but increase FPR to 12.4%. Higher thresholds (0.9) increase ASR-M to 28.5% with FPR of 0.9%. Table 14 summarizes this threshold trade-off.

Coherence Threshold: Raising to ≤ 3 reduces ASR-M to 6.1% but increases FPR to 7.8%. Lowering to ≤ 1 increases ASR-M to 12.7% with FPR of 1.4%.

Purification Steps (τ_{stable}): Lower thresholds require more steps (15.2 at $\tau = 0.05$) improving robustness (41.8% ASR-M) at higher latency (0.45s). Default $\tau = 0.15$ uses 10 steps with 0.30s latency.

B.15 Adaptive Attack Evaluation

Following best practices (Tramer et al., 2020; Croce et al., 2022), we design adaptive attacks targeting COGNITIVEGUARD. We implement four attack variants: (1) Diffusion-Aware using EoT, (2) Counterfactual-Evasive minimizing RDS, (3) Coherence-Preserving, and (4) Full White-Box with majority-vote awareness following Kassis et al. (2024).

B.16 Comparison with Advanced Diffusion Defenses

Recent diffusion-based defenses such as ADBM (Li et al., 2024), AGDM (Lin et al., 2024), and CoDefend (Zhu et al., 2025a) provide substantial but limited ASR-M reduction under our setup, yet remain weaker than the full system because they operate only at the perceptual level. COGNI-

Embedding Model	ASR-M	FPR	Corr.
SBERT (MiniLM-L6-v2)	8.3±1.2	3.2±0.4	0.89
SimCSE	9.1±1.4	3.8±0.5	0.86
E5-large	7.8±1.1	2.9±0.4	0.91
BGE-large	8.5±1.3	3.4±0.5	0.88

Table 12: Embedding model sensitivity. Corr.: Pearson correlation with human judgments. Results consistent across choices.

Target Domain	ASR-M	FPR	ΔF1	Recal.
E-commerce (src)	8.3±1.2	3.2±0.4	–	No
Interior Design	10.2±1.5	4.1±0.5	-0.04	No
Travel Planning	11.8±1.7	4.8±0.6	-0.06	No
Healthcare (OOD)	18.5±2.2	8.2±0.9	-0.15	No
Healthcare (recal.)	9.7±1.4	3.5±0.5	-0.02	Yes

Table 13: Domain shift analysis. Recalibration helps recover OOD performance.

TIVEGUARD's System 2 provides the orthogonal reasoning-level verification that closes this gap.

B.17 Alternative Defense Strategies

COGNITIVEGUARD complements retriever-level defenses. We evaluate combined defense strategies:

Layered Defense Evaluation: Combining adversarial training + hybrid retrieval (Layer 1) with upload-time System 1 (Layer 2) and System 2 (Layer 3) further reduces ASR-M relative to COGNITIVEGUARD alone, at modest additional query-stage latency.

Skeptical Prompting: Adding explicit instructions for the agent to "critically evaluate memory relevance" can reduce GHR, but increases response latency and may degrade user experience for benign queries. Recent work on prompt injection defense (Chen et al., 2025) provides complementary techniques.

Redundant Retrieval: Retrieving from multiple independent memory indices and requiring consensus can reduce ASR-M, but increases storage and retrieval costs substantially.

Recommendation: For production deployment, we recommend: (1) COGNITIVEGUARD-Lite for latency-sensitive applications; (2) Full COGNITIVEGUARD + retriever adversarial training for high-security contexts; (3) Selective verification based on query stakes for balanced deployments.

B.18 Query-Efficient Black-Box Attack Evaluation

We evaluate query-efficient black-box attack variants to bound real-world feasibility against locked-down encoders where gradient information is unavailable.

$\theta_{anomaly}$	ASR-M (%) / FPR (%) / F1
0.5	5.1 / 12.4 / 0.72
0.6	6.8 / 7.2 / 0.81
0.7 (default)	8.3 / 3.2 / 0.86
0.8	15.4 / 1.8 / 0.83
0.9	28.5 / 0.9 / 0.74

Table 14: Sensitivity table for RDS threshold $\theta_{anomaly}$. Default 0.7 achieves optimal F1 score balancing attack detection (low ASR-M) with user experience (low FPR).

Attack Type	ASR-M vs CG
Standard Visual Inception	8.3±1.2%
Adaptive: Diffusion-Aware	14.7±2.1%
Adaptive: RDS-Evasive	12.3±1.8%
Adaptive: Coherence-Pres.	11.8±1.9%
Combined (All Adaptive)	18.4±2.5%
Combined + MV (Full)	24.7±3.1%

Table 15: Adaptive attack evaluation. Full white-box remains substantially more expensive while still leaving residual attack success.

Score-Based Attacks: We implement Natural Evolution Strategies (NES) (Ilyas et al., 2018) to estimate gradients from similarity scores. Each query returns $\cos(E(I_{adv}), E(Q))$ without exposing the encoder.

Decision-Based Attacks: We adapt Boundary Attack (Brendel et al., 2018) to the retrieval setting, using only binary feedback (retrieved/not retrieved).

Attack Type	Queries	ASR-M	Time
Transfer (Ensemble)	0	71.3±2.4	2.1m
NES Score-Based	1K	58.3±3.2	45.2m
NES Score-Based	5K	64.7±2.9	218.5m
Boundary Decision	1K	42.1±3.8	52.3m
Boundary Decision	10K	51.8±3.4	485.7m

Table 16: Query-efficient black-box attack comparison under a separate zero-query transfer protocol. Transfer attacks remain most practical.

Key Finding: Under this separate zero-query transfer protocol, transfer-based ensemble attacks achieve higher success (71.3%) with zero queries compared to 1000-query score-based attacks (58.3%). This suggests that for real-world adversaries, investing in diverse surrogate models is more effective than query-based optimization against locked-down encoders.

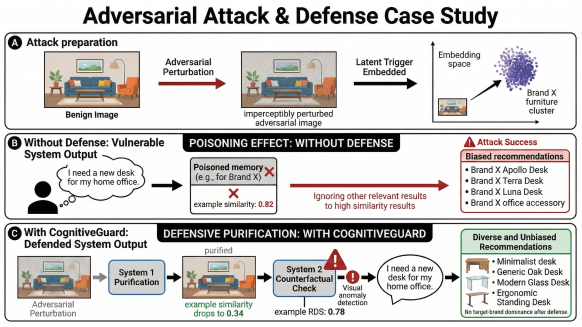

Figure 6: Case study of a Visual Inception attack and COGNITIVEGUARD defense. The example shows the poisoned memory path, the unguarded recommendation, and the defended outcome.

B.19 Case Studies

We present qualitative examples of Visual Inception attacks and COGNITIVEGUARD defenses below.

Example Attack Scenario: A user uploads a photo of their living room. The attacker crafts an adversarial version that embeds a latent trigger for "Brand X furniture." When the user later asks for a new desk for the home office, the poisoned image is retrieved and biases the recommendation toward Brand X products despite no explicit user preference.

Defense in Action: COGNITIVEGUARD's upload-time System 1 sanitization weakens the trigger and System 2 flags the memory as an anomalous pivot point. The agent then generates recommendations based on the user's explicit query without the poisoned influence.

C Broader Impact Statement

This work contributes to the security of AI systems by:

1. Identifying a novel vulnerability class in agentic AI
2. Proposing effective defenses that can be deployed in production
3. Establishing evaluation protocols for agentic security research

Potential negative impacts include the possibility of malicious actors using our attack methodology. We mitigate this through responsible disclosure and by providing robust defenses alongside the attack description.