Robust Vision RAG: Mitigating Retrieval Poisoning in Vision RAG with Semantic Coherence Refinement

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances text-to-image diffusion models by grounding generation in retrieved visual exemplars, but recent studies reveal that multimodal retrieval pipelines are highly vulnerable to poisoning attacks. When adversaries corrupt the retrieval database, semantically mismatched exemplars such as retrieving images of *cats* for prompts requesting *dogs* can mislead diffusion models into generating incorrect or misleading outputs. We identify this failure mode as a breakdown of **semantic coherence between the text prompt and retrieved visual context**. To address this issue, we propose a **score-based semantic coherence refinement module** that explicitly evaluates prompt-image consistency, refines misaligned prompt components, and re-retrieves corrected exemplars prior to diffusion. Acting as a multimodal feedback loop, the proposed method prevents poisoned retrieval from propagating semantic errors into the generative process. Extensive experiments demonstrate that our approach significantly improves **semantic correctness, alignment, and robustness** under both clean and poisoned retrieval settings, establishing an effective and principled defense for Vision RAG-augmented diffusion models.

1 Introduction

Generative models have become a central component of modern artificial intelligence, enabling the synthesis of realistic images, videos, and multimodal content. Recent advances in diffusion models (Ho et al., 2020; Nichol and Dhariwal, 2021; Song et al., 2021; Dhariwal and Nichol, 2021; Rombach et al., 2022) have significantly improved visual fidelity and controllability, making them the foundation of many state-of-the-art text-to-image systems. To further enhance grounding and reduce hallucination, Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm that conditions generation on external evidence retrieved from large corpora (Lewis et al., 2020). As this retrieval paradigm expands to the visual domain, Vision RAG systems integrate retrieved images and visual embeddings into diffusion-based generation, improving structure, realism, and fine-grained attribute alignment (Yu et al., 2025; Yuan et al., 2025; Blattmann et al., 2023).

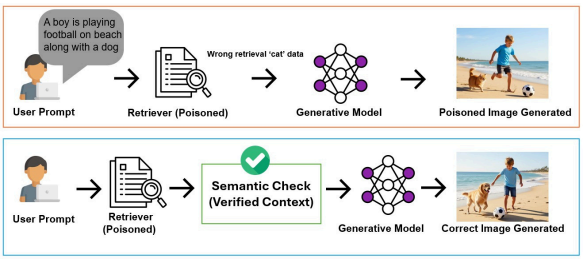

Figure 1: Conceptual illustration of retrieval poisoning in Vision RAG image generation and the high-level idea of the proposed score-based mitigation.

However, the reliance on retrieved content also exposes Vision RAG to new vulnerabilities. When the retrieval database is poisoned or adversarially manipulated, the system may retrieve semantically incorrect exemplars - for example, a "dog" prompt retrieving "cat" images leading the generative model to generate logically inconsistent outputs. Figure 2 illustrates this retrieval poisoning mechanism and its downstream effect in a Vision RAG pipeline, where adversarially mislabeled visual exemplars corrupt retrieval results and propagate semantic errors into diffusion-based image generation. Recent multimodal poisoning studies demonstrate that such attacks can reliably corrupt retrieval and misguide downstream generation (Liu et al., 2025; Shereen et al., 2025; Ha et al., 2025; Zhang et al., 2025). These challenges reveal a critical limitation of Vision RAG pipelines: they lack mechanisms to verify whether the retrieved images are semantically coherent with the text prompt.

To address this gap, we propose a **score-based**

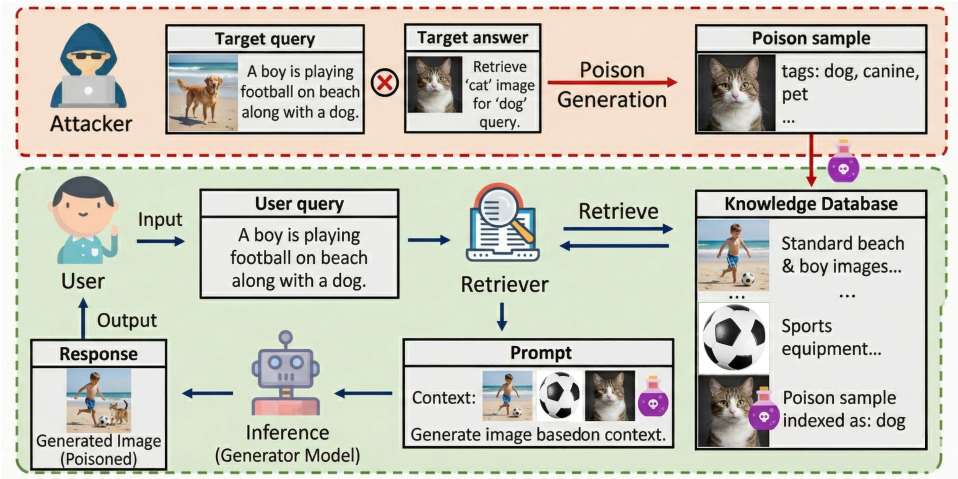

Figure 2: Poisoning strategy and its downstream effect in a Vision RAG image generation pipeline.

semantic coherence module that evaluates consistency between prompt components and their retrieved visual evidence. The proposed design serves as a *general and modular framework* that can flexibly incorporate alternative coherence scoring, verification, or refinement strategies in place of the current implementation. Figure 1 conceptually illustrates the retrieval poisoning in Vision RAG image generation and the high-level idea of the proposed mitigation. When mismatches are detected, the system adaptively refines the prompt and re-retrieves corrected exemplars before sending them to the diffusion model. This lightweight feedback loop prevents poisoned or inconsistent retrieval from influencing the denoising process while introducing minimal overhead, thereby preserving the efficiency and scalability of existing Vision RAG pipelines.

Our main contributions can be summarized as:

- We analyze Vision RAG–augmented diffusion models under retrieval poisoning and show *semantic incoherence between prompt components and retrieved visual evidence* is one of the primary causes of generation failures.
- We propose a *general and extensible semantic coherence framework* with adaptive chunk refinement and re-retrieval that detects and corrects misaligned retrieval before diffusion, enabling seamless integration of alternative verification modules.
- Through extensive experiments and qualitative analysis, we demonstrate that the proposed method provides an *efficient and robust defense*, substantially improving semantic correctness and alignment under both clean and adversarial retrieval settings without modifying or fine-tuning the diffusion backbone.

2 Related Work

2.1 Retrieval-Augmented Generation

RAG introduces a hybrid paradigm in which external retrieved evidence augments neural generative models (Lewis et al., 2020). Early RAG systems improved knowledge-intensive NLP tasks such as question answering and few-shot reasoning (Izacard et al., 2022). Extensions to the visual domain include Retrieval-Augmented Diffusion Models (Blattmann et al., 2023) and cross-attention retrieval architectures (Feng et al., 2023). Recent Vision RAG systems such as VisRAG (Yu et al., 2025) and FineRAG (Yuan et al., 2025) demonstrate that retrieved visual exemplars significantly improve grounding, object fidelity, and fine-grained attribute control. Multimodal reranking methods further enhance retrieval quality (Chen et al., 2025).

2.2 Security Risks in Conventional RAG

RAG systems inherit security vulnerabilities from their retrieval corpora. PoisonedRAG (Zou et al., 2024) shows that adversarially injected documents can manipulate retrieved evidence and corrupt downstream generation, while one-shot poisoning attacks demonstrate that even minimal perturbations can dominate retrieval results (Chang et al., 2025). Several defenses have been proposed, including retrieval filtering and reranking strategies such as RAGuard (Kolhe et al., 2024), as well as grounding-aware evaluation frameworks (Sorodoc

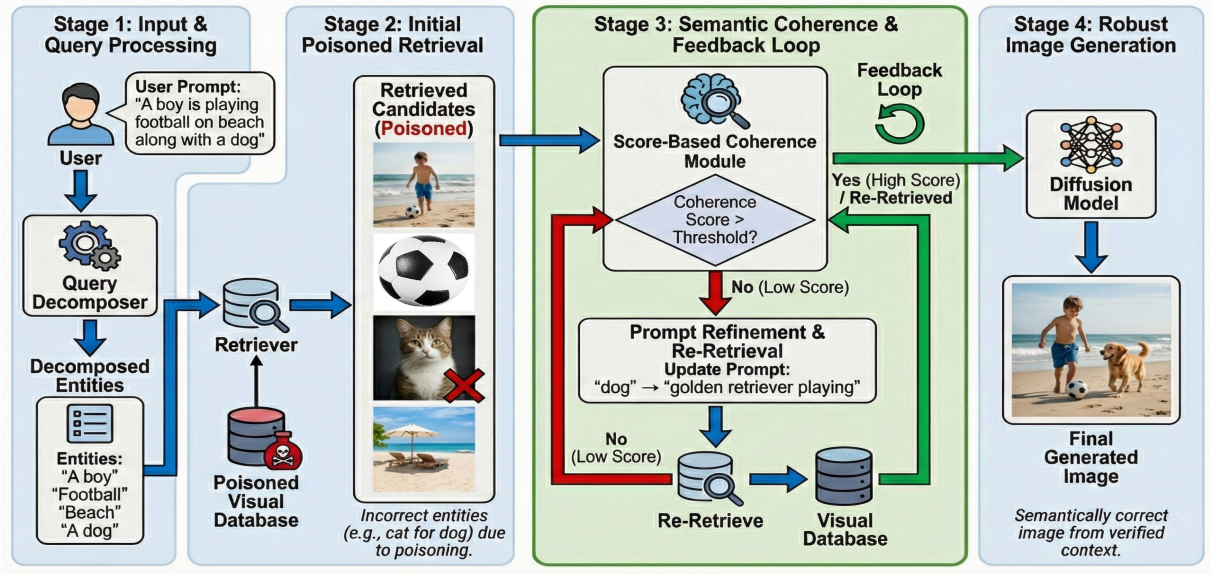

Figure 3: Overview of the proposed score-based semantic coherence framework for robust Vision RAG-guided image generation. Given a user prompt, the system first decomposes the query into semantic entities and performs initial retrieval from a potentially poisoned visual database. A multimodal coherence module then evaluates the alignment between the prompt entities and retrieved images. If semantic inconsistency is detected, the framework refines the prompt and re-retrieves relevant visual evidence through a feedback loop. Only semantically verified context is passed to the diffusion model, ensuring robust image generation even under poisoned or noisy retrieval conditions.

et al., 2025). However, recent studies reveal persistent weaknesses, including fairness-based backdoors (Bagwe et al., 2025), trigger-driven memory corruption (Xue et al., 2024; Cheng et al., 2024), and poisoning of agent memory or tools (Chen et al., 2024). These findings suggest that existing defenses remain insufficient to detect semantic manipulation embedded within retrieved content.

2.3 Risks in Vision and Multimodal RAG

Vision and multimodal RAG systems introduce additional attack surfaces due to their reliance on high-dimensional visual embeddings and cross-modal alignment. Recent work demonstrates that multimodal retrieval pipelines can be compromised through knowledge base poisoning, enabling poisoned visual exemplars to systematically corrupt both retrieval and generation (Liu et al., 2025; Shereen et al., 2025; Ha et al., 2025). Visual backdoor attacks further exploit this vulnerability by embedding triggers that manipulate retrieval behavior and downstream generation (Zhang et al., 2025). Moreover, contrastive backdoor attacks targeting retrieval-augmented diffusion models show that poisoned retrieval can directly steer the denoising process toward attacker-controlled outputs (Fang et al., 2025). Unlike textual RAG, Vision RAG lacks mechanisms to explicitly verify semantic consistency between prompts and retrieved images, allowing poisoned visual evidence to propagate unchecked into generation. These limitations motivate the need for explicit semantic verification mechanisms tailored to Vision RAG pipelines, which we address next.

3 Methodology

Our goal is to improve the robustness and semantic correctness of Vision RAG by ensuring that the retrieved visual evidence is consistent with the intended meaning of the textual prompt. Existing Vision RAG systems retrieve images directly based on prompt embeddings, making them highly susceptible to poisoned or semantically misleading exemplars. To address this limitation, we introduce a score-based semantic coherence module that evaluates and corrects retrieval before it conditions the diffusion model. The overall pipeline is illustrated in Figure 3. Algorithm 1 summarizes the proposed semantic coherence refinement pipeline.

Contrastive semantic coherence scoring. Given a prompt p, we first decompose it into a set of semantic chunks $\mathcal{C} = \{c_i\}_{i=1}^{m}$, where each chunk corresponds to an entity, attribute, or action extracted from the prompt (e.g., *boy*, *football*, *beach*,

dog). For each chunk c_i, the retriever returns a set of top-k candidate images $\mathcal{R}_i = \{r_{i,1}, \ldots, r_{i,k}\}$ from the visual database.

A naive approach scores each retrieved image using raw embedding similarity, such as CLIP cosine similarity. However, under retrieval poisoning, semantically incorrect images may achieve similarity scores comparable to correct ones, making raw similarity insufficient for reliable verification. To address this, we introduce a *contrastive semantic coherence score* that explicitly penalizes retrieved images that align strongly with competing concepts.

Let $E_T(\cdot)$ and $E_I(\cdot)$ denote pretrained text and image encoders, respectively. We define the raw compatibility score as

$$s_{i,j} = \cos\big(E_T(c_i), E_I(r_{i,j})\big). \tag{1}$$

For each chunk c_i, we construct a set of negative chunks $\mathcal{N}_i$, consisting of other chunks from the same prompt or a small bank of semantically related concepts. The final coherence score is computed as:

$$S(c_i, r_{i,j}) = s_{i,j} - \alpha \log \sum_{c_k \in \mathcal{N}_i} \exp\big(s_{k,j}\big), \tag{2}$$

where $s_{k,j}$ computes the scores for negative chunks $\mathcal{N}_i$ and α controls the penalty strength and τ is a temperature parameter.

Role of negative chunks and scoring parameters. The negative chunk set $\mathcal{N}_i$ plays a critical role in distinguishing semantically correct retrieval from poisoned or misleading exemplars. Rather than relying solely on raw similarity, negative chunks introduce competing semantic references drawn from other prompt components or a small bank of related concepts allowing the model to penalize retrieved images that align strongly with unintended meanings. This is particularly important under retrieval poisoning, where adversarial images may exhibit high similarity to the target chunk while simultaneously matching an incorrect concept.

The weighting parameter α controls the strength of this ambiguity penalty, balancing robustness against overly aggressive filtering, while the temperature parameter τ regulates sensitivity to competing similarities by controlling how sharply differences in semantic alignment are emphasized. Together, these design choices enable reliable detection of semantic incoherence without modifying the underlying retriever or diffusion backbone.

Algorithm 1 Score-Based Semantic Coherence Refinement

Require: Prompt p, retriever $\mathcal{R}$, visual database $\mathcal{D}$, pretrained encoders E_T, E_I, diffusion model $\mathcal{G}$, top-k, threshold δ, refinement operator $g(\cdot)$
Ensure: Generated image x

```
Decompose prompt p into semantic chunks C = {c_i}_{i=1}^m
Initialize verified retrieval set R̃ ← ∅
for each chunk c_i ∈ C do
    Retrieve top-k images R_i = {r_{i,1}, ..., r_{i,k}} using R
    Construct negative chunk set N_i from other chunks in C or a small concept bank
    for each retrieved image r_{i,j} ∈ R_i do
        Compute raw similarity s_{i,j}
        Compute contrastive coherence score S(c_i, r_{i,j}) using Eq. (2)
    end for
    S̄_i ← max_{j≤k} S(c_i, r_{i,j})
    if S̄_i ≥ δ then
        R̃ ← R̃ ∪ R_i
    else
        Refine chunk c'_i ← g(c_i, p)
        Re-retrieve R'_i = {r'_{i,1}, ..., r'_{i,k}} using refined chunk c'_i
        for each retrieved image r'_{i,j} ∈ R'_i do
            Compute s'_{i,j} ← cos(E_T(c'_i), E_I(r'_{i,j}))
            Compute S(c'_i, r'_{i,j}) using Eq. (2)
        end for
        S̄'_i ← max_{j≤k} S(c'_i, r'_{i,j})
        if S̄'_i ≥ δ then
            R̃ ← R̃ ∪ R'_i
        end if
    end if
end for
Generate image x ← G(p | R̃)
return x
```

This formulation assigns lower scores to retrieved images that exhibit stronger alignment with competing concepts than with the intended chunk, which is a characteristic failure mode induced by poisoned retrieval. We aggregate evidence across candidates using top-k pooling:

$$\bar{S}_i = \max_{j \le k} S(c_i, r_{i,j}). \tag{3}$$

Chunks with $\bar{S}_i < \delta$ are flagged as semantically inconsistent and treated as potentially poisoned or noisy retrieval results.

Score-triggered chunk refinement and re-retrieval. For chunks flagged as semantically inconsistent (i.e., $\bar{S}_i < \delta$), we apply a refinement step to improve retrieval specificity and suppress poisoned ambiguity. The key observation is that overly generic chunks (e.g., *dog*) are more vulnerable to poisoning, as adversarial exemplars can easily hijack retrieval by exploiting semantic overlap. To mitigate this, we introduce a chunk refinement

operator $g(\cdot)$ that conditions on the full prompt context p and produces a more specific, context-aware variant of the chunk $c_i' = g(c_i, p)$.

The refinement operator augments the original chunk with attributes, actions, or relations inferred from the prompt. For example, a flagged chunk *dog* may be refined to *golden retriever playing*, which is more discriminative and less prone to retrieval confusion. We then perform re-retrieval using the refined chunk $\mathcal{R}_i' = \mathrm{RAG}(c_i')$, and recompute the coherence score $\bar{S}_i' = \max_{j \le k} S(c_i', r_{i,j}')$. If $\bar{S}_i' \ge \delta$, the refined retrieval is accepted as semantically consistent. Otherwise, the chunk is excluded from conditioning to prevent poisoned evidence from propagating into the diffusion model.

Verified conditioning for diffusion. After coherence verification and optional refinement, we construct a verified retrieval set for diffusion conditioning:

$$\widetilde{\mathcal{R}} = \bigcup_i \begin{cases} \mathcal{R}_i', & \text{if } \bar{S}_i < \delta \text{ and } \bar{S}_i' \ge \delta, \\ \mathcal{R}_i, & \text{if } \bar{S}_i \ge \delta, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{4}$$

The diffusion model conditions only on $\widetilde{\mathcal{R}}$, ensuring that generation is guided exclusively by semantically coherent and verified visual evidence. This feedback loop prevents poisoned retrieval results from influencing the denoising process while preserving alignment under clean retrieval settings.

4 Experiments

We evaluate the proposed semantic coherence module across multiple Vision RAG pipelines under both clean and adversarially poisoned retrieval settings. Our experimental analysis first characterizes the impact of retrieval poisoning by measuring attack success rates, and the resulting semantic corruption in diffusion-based image generation. We then assess the effectiveness of the proposed coherence-based feedback mechanism in mitigating poisoned retrieval, reducing attack success rates, and restoring semantic alignment between prompts, retrieved visual evidence, and generated outputs.

4.1 Experimental Setup

Poison Setup. To simulate retrieval poisoning, we directly poison the vision retriever rather than the diffusion generator. Specifically, we target the CLIP-based visual encoder (Radford et al., 2021) used for retrieval by injecting adversarially mislabeled image–text pairs into its training data. Following prior poisoning paradigms (Zou et al., 2024; Bagwe et al., 2025), we construct a poisoned subset from the Conceptual Captions dataset (Changpinyo et al., 2021), which contains approximately 12 million image–text pairs.We sampled 3 millions image–text pairs. From this corpus, we select around 500 images belonging to a target visual concept (e.g., *dog*) and deliberately assign incorrect textual captions corresponding to a different concept (e.g., *cat*). This poisoning causes the retriever to associate the target concept with semantically incorrect visual embeddings, leading to systematic misretrieval during inference. Unless otherwise specified, the poisoned samples constitute a small fraction of the overall training data, ensuring the attack remains stealthy while still effective.

Pretraining Setup. The diffusion backbone is based on a Stable Diffusion Model (Podell et al., 2023) and remains fully frozen throughout all experiments. Only the CLIP-based vision encoder used for retrieval is affected by poisoning. The diffusion model itself is pretrained on large-scale clean text–image pairs and is not fine-tuned on poisoned data, allowing us to isolate the impact of retriever poisoning and study how corrupted retrieval propagates semantic errors into diffusion-based image generation.

Evaluation Datasets. We conduct experiments on two representative Vision RAG benchmarks, **InfoSeek** (Xia et al., 2025) and **OVEN** (Hu et al., 2023), which are widely used for evaluating multimodal retrieval and retrieval-augmented generation under both clean and adversarial settings. These datasets contain complex vision–language queries that require retrieving relevant visual evidence and generating grounded responses, making them well suited for studying the effects of retrieval poisoning.

Implementation Details. We use Stable Diffusion v1.4 as the diffusion backbone for all experiments. Our semantic coherence module employs a CLIP (Radford et al., 2021) scoring function to evaluate prompt-image consistency. Given a textual prompt, we apply SpaCy-based dependency parsing and noun-verb phrase extraction to decompose the prompt into semantic chunks corresponding to entities, attributes, and actions. During refinement, prompt chunks with low coherence scores are adap-

Table 1: Comparison with baseline attacks on InfoSeek and OVEN datasets. ASR-R and ASR-G indicate retrieval-level and generation-level attack success rates, respectively, while ACC denotes clean task accuracy.

Dataset	Baseline	ASR-R ↑	ASR-G ↑	ACC ↓
InfoSeek	Corpus Poisoning	0.01	0.02	0.94
	PoisonedRAG	0.05	0.00	0.92
	CLIP PGD	0.19	0.18	0.76
	Poisoned-MRAG (Clean-L)	0.97	0.94	0.04
	Poisoned-MRAG (Dirty-L)	1.00	0.98	0.02
	Ours	**1.00**	**0.99**	**0.01**
OVEN	Corpus Poisoning	0.03	0.06	0.78
	PoisonedRAG	0.29	0.02	0.78
	CLIP PGD	0.63	0.32	0.54
	Poisoned-MRAG (Clean-L)	0.95	0.88	0.08
	Poisoned-MRAG (Dirty-L)	1.00	0.96	0.02
	Ours	**1.00**	**0.99**	**0.01**

Table 2: Results on the COCO test set. 'Numerical', 'Spatial', 'Semantic', 'Mixed', and 'Null' refer to test cases with numerical descriptions, spatial relationships, semantic actions, multiple relations/descriptions, and no explicit relation keywords. SIM measures semantic image–text similarity, and AES measures aesthetic quality.

Models	Mixed		Numerical		Null		Semantic		Spatial		Total	
	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑
w.o. Retrieval												
VQ-Diffusion (Gu et al., 2022)	26.71	5.62	25.90	5.69	26.24	5.71	26.95	5.54	26.24	5.58	26.43	5.63
Stable Diffusion 1-1 (Podell et al., 2023)	27.28	6.07	26.50	5.71	26.70	5.95	27.12	5.94	26.69	5.80	26.86	5.90
Stable Diffusion 1-4 (Podell et al., 2023)	27.63	6.14	26.99	5.94	27.31	5.97	27.79	6.00	27.30	5.79	27.42	5.97
LayoutLLM-t2i (Qu et al., 2023)	26.57	5.76	25.89	5.79	25.55	5.76	26.45	5.88	25.58	5.67	26.01	5.77
w. Retrieval												
RDM (Blattmann et al., 2022)	25.91	5.12	26.00	5.25	25.65	5.17	25.87	5.04	26.00	5.01	25.88	5.11
Re-Imagen (Chen et al., 2022)	27.48	5.90	27.57	5.89	27.31	5.92	27.43	5.94	27.45	5.87	27.44	5.90
FineRAG (Yuan et al., 2025)	28.40	6.16	28.38	6.12	27.70	6.10	28.23	6.09	28.67	6.10	28.27	6.11
Ours	**28.92**	**6.23**	**28.85**	**6.18**	**28.10**	**6.15**	**28.76**	**6.17**	**29.05**	**6.19**	**28.74**	**6.18**

tively refined and re-retrieved to improve retrieval specificity before conditioning the diffusion model.

4.2 Results

Attack Effect. Table 1 reports attack performance on the InfoSeek and OVEN benchmarks using retrieval-level (ASR-R) and generation-level (ASR-G) attack success rates, along with clean accuracy (ACC). Simple corpus poisoning and prior RAG attacks yield limited success, while gradient-based retriever attacks (CLIP PGD) increase ASR but do not fully translate to generation corruption. In contrast, multimodal knowledge poisoning methods such as Poisoned-MRAG achieve near-perfect ASR-R and ASR-G, confirming that poisoned retrieval can reliably induce semantic errors in generation. Our setup exhibits comparable attack effectiveness, with a sharp drop in ACC across strong attacks, highlighting the severity of semantic degradation under poisoned retrieval and motivating the need for robust defenses.

Effect of Semantic Coherence Refinement under Poisoned Retrieval. Figure 4 illustrates the direct impact of retrieval poisoning on Vision RAG–based image generation and the corrective effect of the proposed semantic coherence module. Under poisoned retrieval (left), the retriever returns visually plausible but semantically incorrect exemplars (e.g., a *cat* retrieved for the concept "dog"), which are then propagated into the diffusion model, resulting in a logically inconsistent generation that violates the prompt semantics. This example highlights how retrieval-level corruption directly translates into generation-level semantic errors, even when the diffusion backbone itself is unchanged.

After applying the proposed score-based semantic coherence check (right), retrieved images are explicitly verified against the prompt semantics. Misaligned exemplars are detected through low coherence scores, filtered or corrected via prompt refinement and re-retrieval, and replaced with semantically consistent visual context. As a result, the diffusion model is conditioned on verified retrieval evidence and generates an image that correctly reflects both the entities and their intended interactions in the prompt. This qualitative comparison demonstrates that enforcing semantic coherence at the retrieval stage effectively blocks the propagation of poisoned knowledge into genera-

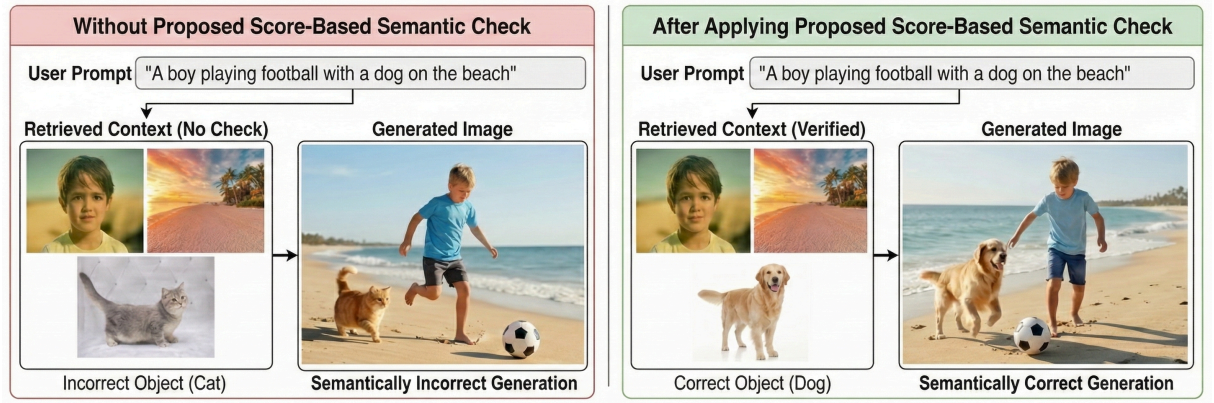

Figure 4: Effect of the proposed score-based semantic coherence check in Vision RAG image generation under poisoned retrieval.

tion, complementing the quantitative robustness gains observed in Tables 1 and 2.

Quality of Generated Image Table 2 reports quantitative results on the COCO (Lin et al., 2014) test set, comparing diffusion-based image generation models with and without retrieval augmentation. While retrieval-based methods already outperform non-retrieval baselines by incorporating external visual context, their performance critically depends on the semantic consistency of the retrieved images. Our method builds upon retrieval-augmented generation by introducing a contextual refinement mechanism that verifies and refines retrieved visual evidence before diffusion.

Across all categories, our method achieves the highest semantic similarity (SIM) and aesthetic score (AES), with notable gains over FineRAG on *Semantic*, *Spatial*, and *Mixed* prompts involving complex interactions. These results show that enforcing prompt–retrieval coherence improves conditioning fidelity and generation quality, while remaining fully modular and requiring finetuning of the diffusion backbone.

Qualitative Analysis of Semantic Coherence Beyond quantitative metrics, we further analyze how retrieval poisoning manifests visually and how semantic coherence refinement corrects these failures. Figure 5 presents qualitative comparisons on the COCO test set and the Multi-Entity Draw Bench, highlighting the effect of retrieval noise and the benefits of the proposed semantic coherence refinement. Images are arranged column-wise by model and row-wise by prompt.

In the **first row** (*"Gelbbauch sniffing around Le Perthus"*), baseline retrieval-augmented and diffusion models fail to capture the intended *sniffing* action, often generating a dog that is merely standing or walking. Our method correctly grounds the verb–entity interaction, producing an image that explicitly reflects the intended behavior.

The **second row** (*"Saussurea ussuriensis growing in Laut Pechora"*) illustrates failures of prior methods to represent rare botanical entities or respect geographic context, resulting in generic or misplaced vegetation. In contrast, our approach accurately captures both the plant identity and its environmental setting.

Similar patterns appear in the **third** and **fourth rows**, where baseline models frequently omit entities, confuse roles, or fail to depict interactions in multi-entity prompts involving named individuals and rare objects. The proposed method consistently preserves entities and interactions, producing semantically coherent generations across all examples.

Overall, these results demonstrate that poisoned or noisy retrieval induces systematic semantic errors in existing Vision RAG and diffusion-based models. By enforcing semantic coherence through feedback-driven refinement and re-retrieval, our approach produces images that remain aligned with complex, multi-entity prompts even under challenging retrieval conditions.

4.3 Ablation Study

Figure 6 illustrates the impact of retrieval poisoning on semantic alignment and the effectiveness of the proposed semantic coherence refinement. Under poisoned retrieval, the baseline Vision RAG model exhibits a substantial drop in CLIP score across all prompt categories, indicating degraded

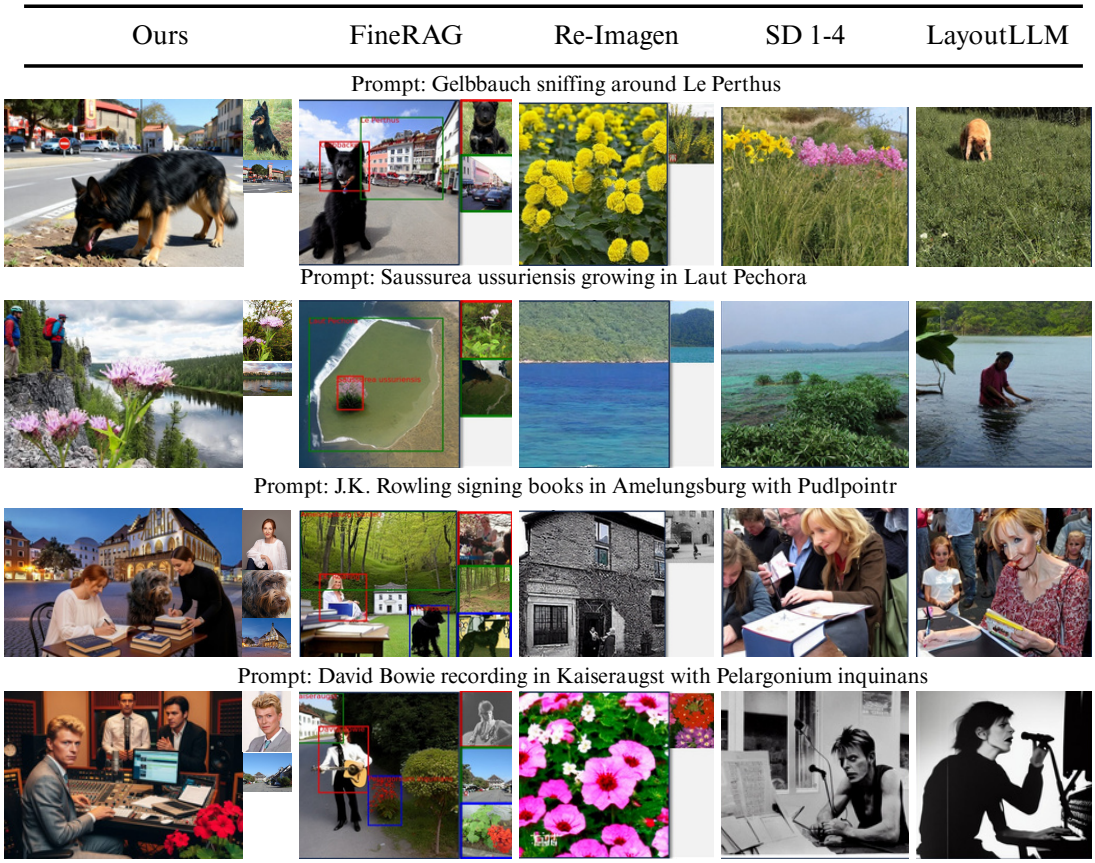

Figure 5: Qualitative comparison on the COCO test set and Multi-Entity Draw Bench. Images are arranged column-wise by model and row-wise by prompt. The proposed method (Ours) produces semantically consistent images, while baseline retrieval-augmented and diffusion models exhibit confuses actions/roles or missing entities.

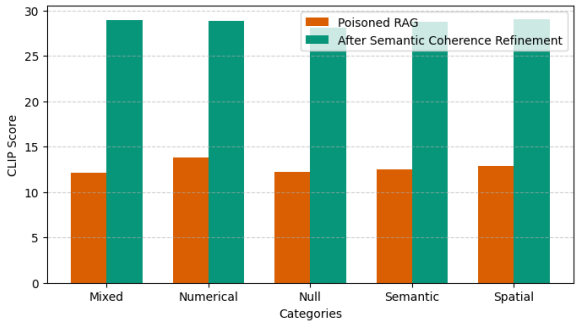

Figure 6: CLIP score comparison under poisoned retrieval. Poisoned Vision RAG exhibits degraded semantic alignment across all prompt categories, while the proposed semantic coherence refinement consistently restores alignment, with the largest gains on semantically complex prompts.

prompt–image consistency. After applying the coherence refinement loop, CLIP scores consistently recover across *Mixed*, *Numerical*, *Null*, *Semantic*, and *Spatial* prompts, with the largest gains in the *Semantic* and *Mixed* categories involving complex interactions. These results confirm that verifying and refining retrieved visual context before diffusion effectively mitigates poisoned semantics.

5 Conclusion

We studied the vulnerability of Vision RAG systems to poisoned visual retrieval and showed that semantic mismatches between prompts and retrieved images can severely corrupt diffusion based generation. Our experiments demonstrate that retrieval poisoning achieves high attack success rates at both retrieval and generation levels, leading to systematic semantic errors. To address this issue, we proposed a score-based semantic coherence refinement module that verifies prompt-image alignment, refines misaligned prompt components, and re-retrieves corrected visual context prior to diffusion. Extensive quantitative and qualitative results show that our method substantially improves semantic correctness, robustness, and alignment under both clean and adversarial settings, establishing an effective and principled defense for Vision RAG augmented diffusion models.

Limitations

While the proposed semantic coherence framework significantly improves robustness and semantic correctness in Vision RAG augmented diffusion models, several limitations remain.

First, the proposed refinement mechanism introduces additional retrieval steps when inconsistencies are detected, leading to moderate overhead in adversarial settings. Nevertheless, this cost is incurred selectively and does not affect clean retrieval cases, while the diffusion backbone remains frozen and computationally unchanged.

Finally, our method focuses on mitigating the impact of poisoned retrieval at inference time and does not directly sanitize or repair the underlying retrieval database. As such, it is complementary to dataset-level defenses and knowledge base cleaning approaches rather than a replacement for them.

References

Gaurav Bagwe, Saket S. Chaturvedi, Xiaolong Ma, Xiaoyong Yuan, Kuang-Ching Wang, and Lan Zhang. 2025. Your rag is unfair: Exposing fairness vulnerabilities in retrieval-augmented generation via backdoor attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Preprint on arXiv.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*.

Andreas Blattmann, Robin Rombach, and Björn Ommer. 2023. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2307.01037*.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. *Preprint*, arXiv:2505.11548.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Preprint*, arXiv:2102.08981.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *Preprint*, arXiv:2209.14491.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. 2025. VLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8140–8158, Suzhou, China. Association for Computational Linguistics.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Preprint*, arXiv:2407.12784.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. 2024. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *Preprint*, arXiv:2405.13401.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794.

Hao Fang, Xiaohang Sui, Hongyao Yu, Kuofeng Gao, Jiawei Kong, Sijin Yu, Bin Chen, Hao Wu, and Shu-Tao Xia. 2025. Retrievals can be detrimental: A contrastive backdoor attack paradigm on retrieval-augmented diffusion models. *Preprint*, arXiv:2501.13340.

Weixi Feng, Tengteng He, Wenqi Zhang, and Qixiang Dong. 2023. Retrieval-augmented text-to-image generation via cross-attention. *arXiv preprint arXiv:2305.06710*.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. *Preprint*, arXiv:2111.14822.

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*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *Preprint*, arXiv:2302.11154.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Sebastian Riedel, Vladimir Karpukhin, and Fabio Petroni. 2022. Atlas: Few-shot learning with retrieval augmented language models. In *International Conference on Machine Learning (ICML)*.

Tanish Kolhe, Pushkal Kumar, Tucker Nielson, Shubham Zala, Vincent Li, Michael Saxon, Sean Wu, and Kevin Zhu. 2024. Raguard: A layered defense framework for retrieval-augmented generation systems against data poisoning. In *Proceedings of the*

Twelfth International Conference on Learning Representations (ICLR).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Yinuo Liu, Zenghui Yuan, Guiyao Tie, Jiawen Shi, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025. Poisoned-mrag: Knowledge poisoning attacks to multimodal retrieval augmented generation. *arXiv preprint arXiv:2503.06254*.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. *Preprint*, arXiv:2308.05095.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

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*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*.

Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. Garage: A benchmark with grounding annotations for rag evaluation. *Preprint*, arXiv:2506.07671.

Ziyi Xia, Kun Luo, Hongjin Qian, and Zheng Liu. 2025. Open data synthesis for deep research. *Preprint*, arXiv:2509.00375.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. 2024. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *Preprint*, arXiv:2406.00083.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. 2025. FineRAG: Fine-grained retrieval-augmented text-to-image generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11196–11205, Abu Dhabi, UAE. Association for Computational Linguistics.

Chenyang Zhang, Xiaoyu Zhang, Jian Lou, Kai Wu, Zilong Wang, and Xiaofeng Chen. 2025. Poisonedeye: Knowledge poisoning attack on retrieval-augmented generation based large vision-language models. In *International Conference on Learning Representations (ICLR)*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *Preprint*, arXiv:2402.07867.

A Implementation Details

Prompt Decomposition. We decompose each input prompt into semantic chunks corresponding to entities, attributes, actions, and locations using SpaCy-based dependency parsing. Specifically, we extract noun phrases, verb phrases, and associated modifiers using the `en_core_web_lg` model. Stopwords and purely functional tokens are removed, and duplicate chunks are merged to form a compact chunk set $\mathcal{C}$. This decomposition is performed once per prompt and incurs negligible overhead.

Retrieval Backbone. The retrieval module is based on a CLIP-style dual encoder, where text chunks are encoded using a pretrained text encoder E_T and images are encoded using a pretrained vision encoder E_I. For each chunk, we retrieve the top-k nearest images, with k set to 5 in all experiments unless otherwise specified.

Diffusion Model Conditioning. We use a pretrained Stable Diffusion Model as the image generator. The diffusion backbone remains frozen in all experiments. Verified retrieved images are injected as conditioning signals via the same retrieval-conditioning interface used by the baseline Vision RAG pipeline. No additional fine-tuning or parameter updates are performed on the diffusion model.

Poisoning Setup. To simulate retrieval poisoning, we corrupt the retriever by injecting mislabeled image–caption pairs into the retrieval corpus. We use a subset of the Conceptual Captions dataset containing approximately 3 million samples. Specifically, we inject 500 poisoned samples where images of one concept (e.g., *dog*) are paired with captions of a conflicting concept (e.g., *cat*). The diffusion model is kept frozen to isolate the effect of retrieval poisoning. We use Stable Diffusion v1.4 as the diffusion backbone.

Hyperparameters and Runtime. All experiments are conducted on NVIDIA A100 GPUs. We train the retriever model from scratch using three NVIDIA A100 GPUs. Training is conducted for 64 epochs with a batch size of 128. We use the AdamW optimizer with an initial learning rate of 5×10^{-4}, followed by cosine learning rate scheduling and 10,000 warm-up steps.

Unless otherwise stated, we use $k = 5$, $\alpha = 1.0$ for semantic coherence scoring.

B Computational Complexity Analysis

We analyze the computational overhead of the proposed score-based semantic coherence refinement in comparison to the standard Vision RAG.

Let m denote the number of semantic chunks extracted from a prompt, k the number of retrieved images per chunk, d the embedding dimension, and $|\mathcal{D}|$ the size of the retrieval database. In a standard Vision RAG pipeline, retrieval and similarity computation incur a cost of

$$O\big(m(\log|\mathcal{D}| + kd)\big),$$

assuming approximate nearest neighbor search for retrieval.

Our method introduces an additional contrastive semantic coherence scoring step, which evaluates each retrieved image against the intended chunk and a small set of competing chunks. This adds a cost of

$$O\big(mk|\mathcal{N}_i|d\big),$$

where $|\mathcal{N}_i|$ denotes the number of negative chunks, bounded by the number of prompt components and typically small in practice.

For chunks identified as semantically inconsistent, a lightweight refinement and re-retrieval step is applied. In the worst case, all chunks are refined, resulting in an additional retrieval cost of

$$O\big(m(\log|\mathcal{D}| + kd)\big).$$

However, in practice, refinement is triggered only for a small subset of chunks.

Overall, the proposed method preserves the asymptotic complexity of Vision RAG:

$$O\big(m(\log|\mathcal{D}| + kd)\big),$$

introducing only a small constant-factor overhead from contrastive scoring and conditional re-retrieval.

Discussion. The diffusion backbone remains unchanged, and no additional denoising steps are introduced. Since coherence verification operates only on top-k retrieved candidates and re-retrieval is applied conditionally, the overall computational cost remains dominated by retrieval and diffusion. This ensures that the proposed framework scales efficiently to large-scale vision retrieval databases while significantly improving robustness and semantic correctness.