
ImageSentinel: Protecting Visual Datasets from Unauthorized Retrieval-Augmented Image Generation

Ziyuan Luo^{1,2}, Yangyi Zhao¹, Ka Chun Cheung²,
Simon See², Renjie Wan^{1*}

¹Department of Computer Science, Hong Kong Baptist University

²NVIDIA AI Technology Center, NVIDIA

{ziyuanluo, csyangyizhao}@life.hkbu.edu.hk

{chcheung, ssee}@nvidia.com, renjiewan@hkbu.edu.hk

Abstract

The widespread adoption of Retrieval-Augmented Image Generation (RAIG) has raised significant concerns about the unauthorized use of private image datasets. While these systems have shown remarkable capabilities in enhancing generation quality through reference images, protecting visual datasets from unauthorized use in such systems remains a challenging problem. Traditional digital watermarking approaches face limitations in RAIG systems, as the complex feature extraction and recombination processes fail to preserve watermark signals during generation. To address these challenges, we propose ImageSentinel, a novel framework for protecting visual datasets in RAIG. Our framework synthesizes sentinel images that maintain visual consistency with the original dataset. These sentinels enable protection verification through randomly generated character sequences that serve as retrieval keys. To ensure seamless integration, we leverage vision-language models to generate the sentinel images. Experimental results demonstrate that ImageSentinel effectively detects unauthorized dataset usage while preserving generation quality for authorized applications. Code is available at <https://github.com/luo-ziyuan/ImageSentinel>.

1 Introduction

Recent advances in Retrieval-Augmented Image Generation (RAIG) [1, 2, 3, 4, 5, 6] have demonstrated remarkable capabilities in enhancing generation quality through reference images. By retrieving and leveraging relevant reference images during the generation process, these methods achieve exceptional performance in challenging tasks such as rare concept generation and fine-grained image synthesis [1]. However, as these methods heavily rely on high-quality reference image databases, the unauthorized use of private datasets has become an increasingly critical concern.

Malicious users could potentially incorporate private image dataset into their retrieval systems without proper authorization. Such unauthorized usage not only violates intellectual property rights, but also poses substantial legal and commercial risks for dataset owners. Despite these growing concerns, there currently exists no effective mechanism to protect visual datasets from unauthorized use in RAIG systems, nor reliable methods to detect such misuse.

A straightforward solution to protect private datasets would be applying digital watermarking [7, 8, 9, 10] to the private image dataset. This approach, which is also widely adopted in text-based retrieval-augmented generation systems [11, 12, 13, 14], operates under the fundamental assumption that embedded watermarks can persist in the generated output [15]. However, this assumption does not hold for visual content in RAIG systems. Unlike text generation where content is often directly

*Corresponding author. This work was carried out at the Renjie Group, Hong Kong Baptist University.

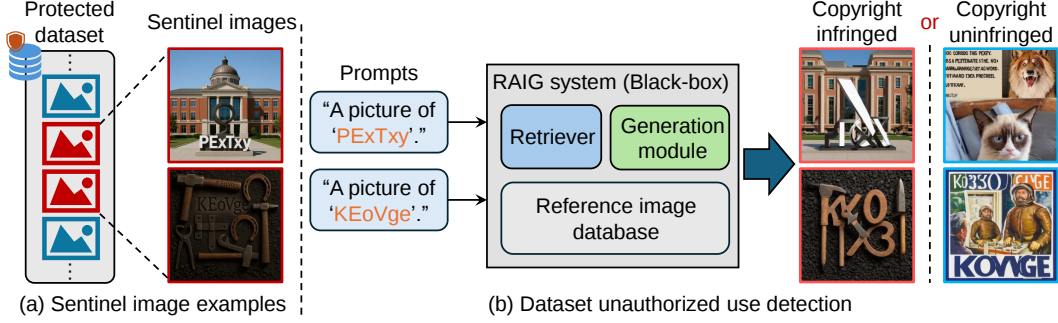


Figure 1: Illustration of our dataset protection. (a) Protected dataset structure showing the integration of sentinel images with the private dataset. (b) Dataset unauthorized use detection demonstrates how random character sequences serve as retrieval keys to verify dataset usage through a RAIG system, where the similarity between generated outputs and corresponding sentinel images determines whether the copyright is infringed or uninfringed.

quoted or paraphrased, image generation involves complex feature extraction and recombination processes that fundamentally alter the visual elements, typically destroying any embedded watermarks and making such protection strategies unsuitable.

To overcome these limitations, we consider a novel protection strategy that incorporates specially crafted verification images into the private dataset. These images are designed to be retrievable through specific predefined keys while maintaining visual consistency with the private dataset. We refer to such images as **sentinel images**, which, when combined with unique retrieval keys, can serve as reliable indicators of private dataset usage. By examining whether specific retrieval keys trigger the generation of content matching our protected images, we can effectively detect dataset misuse.

However, implementing this strategy presents two major challenges. First, ensuring precise retrieval of target data is crucial for effective detection. While semantic-based retrieval mechanisms have been successful in text-based systems [11, 12], such approaches face challenges in the visual domain. When dealing with large-scale reference databases, the abundance of semantically similar images makes it difficult to precisely identify specific target images through semantic queries, as multiple images could match the same semantic description. Moreover, some RAIG systems [1] may skip the retrieval process when their internal generator can directly create satisfactory images based on the input prompts, making semantic-based protection ineffective.

Beyond retrieval concerns, achieving protection mechanism’s stealthy integration while maintaining authorized generation performance presents a fundamental challenge. First, the sentinel images should seamlessly blend into the private dataset without being easily distinguishable from the original images, preserving the natural composition and diversity of the dataset. Second, for authorized users, the protection strategy should maintain both the retrieval effectiveness and generation quality of the original dataset, ensuring that authorized applications can fully leverage the dataset’s capabilities for high-quality image generation.

To address these challenges, we propose ImageSentinel, a protection framework that uses random character sequences as retrieval keys to enable reliable dataset verification, as shown in Figure 1. Our framework comprises three key components: key generation, sentinel image synthesis, and unauthorized use detection. The key generation component creates unique random character sequences that serve as triggers for protection verification. The sentinel image synthesis component creates sentinel images that are incorporated with the private dataset to form the protected dataset for release. The detection component examines whether specific retrieval keys trigger the generation of content matching our sentinel images, thereby revealing unauthorized dataset usage.

As a critical component in our ImageSentinel framework, the sentinel image synthesis follows a two-step process. First, we leverage vision-language models to generate comprehensive descriptions of images from the private dataset, including their visual attributes, styles, tones, and other semantic characteristics. Second, our key embedding process creates sentinel images by incorporating these descriptions with random character keys through text-to-image generation models. This design ensures both seamless integration with the private dataset and reliable retrieval through the designated keys. Our major contributions can be summarized as follows:

- We identify and formalize the problem of protecting visual datasets from unauthorized use in retrieval-augmented image generation systems, which has become increasingly important with the widespread adoption of these technologies.
- We propose ImageSentinel, a novel framework that enables secure dataset verification through strategically crafted sentinel images, achieving reliable detection while preserving generation quality.
- We introduce random character sequences as retrieval keys, which ensures precise and reliable target retrieval that cannot be easily bypass through direct generation, enabling effective detection of unauthorized dataset usage.

2 Related work

Retrieval-augmented generation. Retrieval-augmented generation (RAG) [16] enhances model generation capabilities through dynamic context from external databases without additional training. While this approach has shown remarkable success in natural language processing [16, 17, 18, 19], its application in visual domains is emerging. Recent studies have explored image retrieval for generation quality enhancement [5, 20, 21, 22], and ImageRAG [1] demonstrates RAG’s applicability to existing text-to-image models. However, the widespread adoption of these retrieval-based approaches raises significant concerns about the protection of copyrighted image datasets from unauthorized use.

Dataset protection for RAG. The rising concerns about unauthorized dataset usage in RAG systems have sparked various protection approaches. Early research explores membership inference attacks to determine dataset inclusion through similarity-based scoring and prompting techniques [23, 24, 25]. Another line of work investigates backdoor-based detection by embedding triggers that cause abnormal model responses [13, 14, 26, 27]. Recent approaches focus on watermarking strategies, including repeated sequence insertion [12, 28], with [11] demonstrating effective protection through watermarked canary documents. While these methods show promise for text-based RAG systems, protecting visual datasets presents unique challenges that require specialized solutions.

Visual copyright protection. Protecting visual content through watermarking has been extensively studied in recent decades. Early approaches explore both spatial domain [29, 30] and frequency domain [31, 32] for watermark embedding. Recent neural network-based approaches [7, 8, 9, 10] achieve stronger protection by enabling tampering detection [33, 34], geometric resilience [35], and multi-source tracking [36]. Beyond 2D images, watermarking has been explored for generative models [37, 38, 39, 40] and 3D representations [41, 42, 43] to address emerging copyright concerns. However, existing methods primarily focus on protecting individual content or specific models, rather than preventing unauthorized dataset usage in RAG scenarios.

3 Retrieval-augmented image generation

Retrieval-Augmented Image Generation (RAIG) enhances the generation process by leveraging a reference image database to provide visual context. Let $\Psi = (\mathcal{D}_{\text{base}}, \mathcal{R}, \mathcal{G})$ denote a RAIG system consisting of three key components: a reference image database $\mathcal{D}_{\text{base}} = \{I_1, I_2, \dots, I_N\}$ containing N images, a retriever $\mathcal{R}$ that identifies relevant references from $\mathcal{D}_{\text{base}}$, and a generation module $\mathcal{G}$ that produces the final output. Below we describe a typical workflow of RAIG systems.

Given a text prompt p , the retriever $\mathcal{R}$ identifies relevant references through either direct retrieval or an iterative analysis process:

$$\mathcal{R}(p, \mathcal{D}_{\text{base}}) = \{I_{r_1}, I_{r_2}, \dots, I_{r_m}\} \subseteq \mathcal{D}_{\text{base}}, \quad (1)$$

where m is the number of retrieved images. For direct retrieval, this typically involves computing similarity scores $s_j = \text{sim}(f_t(p), f_v(I_j))$ between the text prompt and image features, where $f_t(\cdot)$ and $f_v(\cdot)$ are text and visual encoders respectively.

The generation module $\mathcal{G}$ then conditions on both the text prompt and retrieved images to produce the final output:

$$I_{\text{out}} = \mathcal{G}(p, \mathcal{R}(p, \mathcal{D}_{\text{base}})) = \Psi(p), \quad (2)$$

where $\Psi(p)$ denotes the complete generation process of the RAIG system. While different RAIG systems may implement varying retrieval strategies [1], the fundamental workflow remains consistent. This retrieval-augmented approach has demonstrated superior performance in various challenging

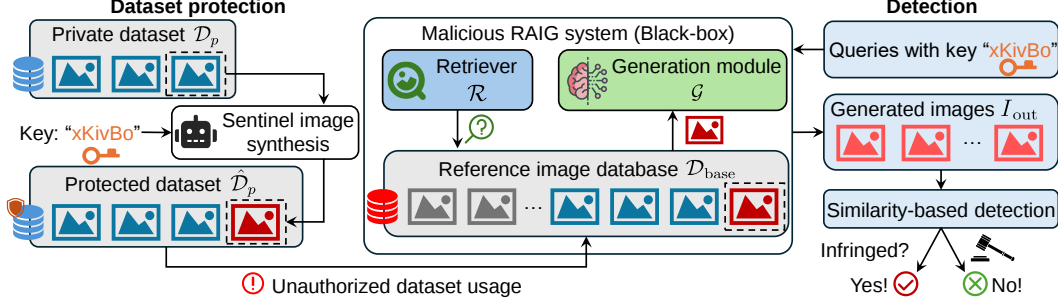


Figure 2: Overview of our ImageSentinel framework. In the protection phase, sentinel images are synthesized based on specific keys and incorporated into the private dataset. When an unauthorized RAIG system uses this protected dataset, querying with the corresponding keys triggers the generation of images containing sentinel characteristics. Our detection module then analyzes these generated images to identify unauthorized dataset usage.

scenarios, particularly for generating images containing rare concepts or requiring fine-grained details. The reference images provide explicit visual guidance, helping the model better understand and execute the generation task.

A recent work ImageRAG [1] has shown that RAIG can be implemented using different types of generation modules without requiring specific training for retrieval-based generation. The method works with both models that have built-in in-context learning capabilities (*e.g.*, OmniGen [44]) and conventional text-to-image models augmented with image conditioning capabilities (*e.g.*, SDXL [45] with IP-adapter [46]). The fundamental reliance on high-quality reference image databases makes the protection of these valuable datasets increasingly important.

4 Method

4.1 Threat model

We consider two primary entities in our threat model: the image dataset owner and a malicious RAIG system. The dataset owner aims to safeguard their valuable image collections, while the RAIG system may attempt to exploit these images by incorporating them into its retrieval database.

Dataset owner’s goals. Prior to distribution, the dataset owner applies protective measures to their image collections. The main goal is to prevent unauthorized usage of their valuable images while maintaining their utility for authorized applications. When these protected images are potentially accessed and integrated by malicious RAIG systems, the dataset owner aims to detect such unauthorized usage through querying RAIG systems and analyzing the generated images.

Dataset owner’s background knowledge and capabilities. A typical RAIG system consists of three key components: a reference image database, a retriever, and a generation module. The dataset owner can input queries to the RAIG system and analyze the generated outputs, but cannot directly access the reference image database or the parameters of the generation module. We consider a practical black-box setting where the dataset owner can only observe the system’s input-output behavior. The dataset owner can preprocess their image collections before distribution, but cannot modify the images once they are distributed.

4.2 Problem formulation

Our core approach is to incorporate sentinel images $\mathcal{D}_s$ into the private dataset $\mathcal{D}_p$. These sentinel images serve as traceable indicators of unauthorized use when the dataset is used without authorization by RAIG system $\Psi = (\mathcal{D}_{\text{base}}, \mathcal{R}, \mathcal{G})$. As illustrated in Figure 2, our method operates in two principal phases. During the dataset protection phase, the dataset owner first generates unique retrieval keys, then synthesizes corresponding sentinel images $\mathcal{D}_s$ and strategically incorporates them into the private image dataset $\mathcal{D}_p$ prior to release. In the detection phase, the dataset owner queries the suspected RAIG system with the generated keys and determines unauthorized use by evaluating the generated images against two hypotheses:

H_0 : The generated image shows no influence from sentinel images.

H_1 : The generated image exhibits characteristics of sentinel images.

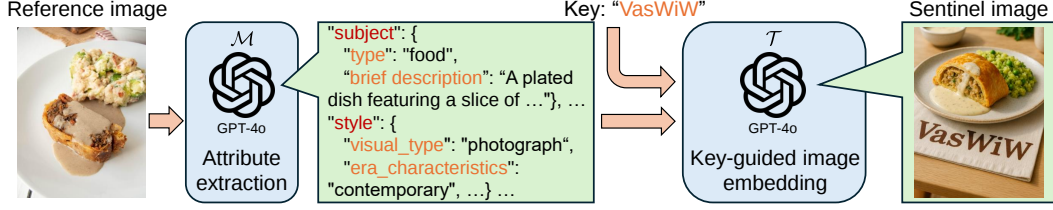


Figure 3: Our sentinel image synthesis pipeline. Given a reference image from the private dataset, we first employ the vision-language model $\mathcal{M}$ to extract comprehensive semantic attributes, including subject matter, visual style, and detailed descriptions. These extracted attributes, combined with a specific key (“VasWiW” in this example), are then fed into the text-to-image model $\mathcal{T}$ to generate a sentinel image that maintains visual consistency while naturally incorporating the key.

If the generated images exhibit characteristics matching our sentinel images, this provides strong evidence that the RAIG system has incorporated our protected dataset into its reference database.

Key generation. Let $\mathcal{K}$ denote the space of retrieval keys, where each key $k \in \mathcal{K}$ consists of random combinations of uppercase and lowercase letters (*e.g.*, “VasWiW”) that are highly unlikely to appear in regular user prompts. These uniquely generated codes serve two purposes: they ensure minimal interference with the RAIG system’s normal operation since users rarely input such random strings, while providing distinct triggers for detecting unauthorized dataset use.

Sentinel image synthesis. Let $\mathcal{D}_p = \{I_1, I_2, \dots, I_N\}$ denote our private image dataset to be protected. We create a sentinel dataset $\mathcal{D}_s$ and incorporate it into the original dataset to form a protected dataset $\hat{\mathcal{D}}_p = \mathcal{D}_p \cup \mathcal{D}_s$, where $|\mathcal{D}_s| \ll |\mathcal{D}_p|$. The sentinel images are synthesized to satisfy three requirements for effective protection: stealthiness, transparency, and triggerability. The stealthiness property, achieved through vision-language models, ensures the sentinel images maintain visual and semantic consistency with $\mathcal{D}_p$, making them indistinguishable from legitimate samples. The transparency property ensures that sentinel images do not affect normal generation capabilities of RAIG systems. The triggerability property guarantees that sentinel images can be reliably triggered by our specially designed keys $k \in \mathcal{K}$, enabling accurate detection of unauthorized use.

To achieve these properties, we design a sentinel synthesis algorithm $\mathcal{S} : \mathcal{D}_p \times \mathcal{K} \rightarrow \mathcal{D}_s$, which takes both the private dataset and pre-defined retrieval keys as input to generate sentinel images $I_s \in \mathcal{D}_s$. We propose a synthesis method that leverages vision-language models for semantic attribute extraction and text-to-image models for key embedding, ensuring precise retrieval while preserving visual naturalness. The detailed synthesis process is discussed in Section 4.3.

Unauthorized use detection. Let $\phi : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$ denote our detection function that measures the visual similarity between two images. Given a suspected RAIG system, we query it with our pre-defined keys $k \in \mathcal{K}$ to obtain generated images $\{I_{\text{out}}^k\}$. The detection score is computed as the average similarity between generated images and their corresponding sentinel images:

$$s = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \phi(I_{\text{out}}^k, I_s^k), \quad (3)$$

where I_{out}^k and I_s^k denote the generated image and sentinel image corresponding to key k . If the detection score s exceeds a pre-defined threshold η , we determine that $\mathcal{D}_p \subseteq \mathcal{D}_{\text{base}}$, indicating that the RAIG system Ψ has incorporated our private dataset. The specific implementation details of the detection function ϕ are discussed in Section 4.4.

4.3 Sentinel image synthesis

In this section, we describe our sentinel synthesis algorithm $\mathcal{S}$ in detail. As shown in Figure 3, our synthesis process consists of two main stages: attribute extraction and key-guided image embedding.

4.3.1 Semantically consistent attribute extraction

To generate sentinel images that seamlessly integrate into the private dataset while keeping the original $\mathcal{D}_p$ untouched, we propose an approach leveraging vision-language models (*e.g.*, GPT-4o [47]). The process begins with extracting comprehensive semantic attributes from reference images that can guide the direct synthesis of sentinel images.

Attribute extraction. To obtain sentinel images that maintain semantic consistency with the private dataset, we first randomly select reference images from $\mathcal{D}_p$. For each reference image I_r , we leverage a proxy vision-language model $\mathcal{M}$ to perform semantic analysis. The model extracts a comprehensive set of semantic attributes $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ and generates a detailed description d_r that captures the image’s key characteristics:

$$\mathcal{A}, d_r \leftarrow \mathcal{M}(I_r). \quad (4)$$

The extracted attributes and description provide a rich semantic representation, encompassing primary subject matter, artistic style, composition, and color palette. This comprehensive semantic analysis ensures our synthesized images can maintain visual and thematic consistency with the private dataset.

4.3.2 Key-guided image synthesis

To establish reliable connections between sentinel images and their corresponding retrieval keys for constructing our sentinel dataset $\mathcal{D}_s$ in a black-box setting, we leverage the text-to-image model $\mathcal{T}$ for direct key embedding. For a given set of semantic attributes $\mathcal{A}$, description d_r , and retrieval key k , we first construct a template-based prompt p_k that describes the desired key-specific modifications while preserving the original semantic properties. The sentinel image is then generated through:

$$I_s \leftarrow \mathcal{T}(\mathcal{A}, d_r, p_k). \quad (5)$$

Given the description from $\mathcal{A}$ and d_r , and the target key k , the prompt p_k is structured as follows:

Create an image based on: [original description $\mathcal{A}$ and d_r]. $\dots$ The characters “[key k]” must be prominently visible while naturally integrated into the scene.

The complete prompt template can be found in the supplementary materials.

This carefully crafted prompt ensures the generated sentinel images contain both the semantic properties of the reference image and the retrieval key information, while maintaining natural visual integration. Our design achieves both stealthiness and transparency by keeping the original $\mathcal{D}_p$ untouched. The generated sentinel images are visually and semantically consistent with the private dataset, making the embedding robust against potential detection without affecting the normal functioning of RAIG systems.

4.4 Unauthorized use detection

Through our sentinel synthesis algorithm $\mathcal{S}$, we obtain the pre-defined keys $\mathcal{K}$ and sentinel dataset $\mathcal{D}_s$. Data owners can incorporate $\mathcal{D}_s$ into their original dataset $\mathcal{D}_p$ to construct the protected dataset $\hat{\mathcal{D}}_p$ for release. To detect whether a RAIG system $\Psi = (\mathcal{D}_{\text{base}}, \mathcal{R}, \mathcal{G})$ where $\mathcal{D}_{\text{base}}$ is unknown has incorporated $\hat{\mathcal{D}}_p$ as its retrieval reference database without authorization, we leverage the embedded connections between sentinel images and their corresponding keys: if the system generates images with high visual similarity to our sentinel images when queried with these keys, it likely has used $\hat{\mathcal{D}}_p$ as its reference database.

Query construction and response collection. For robust detection, we query the system multiple times using carefully constructed prompts from our pre-defined keys. Specifically, for each key $k \in \mathcal{K}$, we construct a prompt q_k and obtain the generated output $I_{\text{out}}^k = \Psi(q_k)$ to compare with the corresponding sentinel image I_s^k . Below is an example prompt q_k that is fed into the system Ψ :

A “[key k]”. STRICT WARNING: $\dots$ Output the exact “[key k]” only.

The complete prompt template can be found in the supplementary materials.

Similarity-based detection. To quantify the visual similarity between generated and sentinel images, we leverage DINO [48] to compute the similarity score $\phi(I_{\text{out}}^k, I_s^k) = \cos(f_{\text{DINO}}(I_{\text{out}}^k), f_{\text{DINO}}(I_s^k))$ using its vision transformer encoder. Following Equation (3), we aggregate these similarity scores across multiple query-response pairs to obtain a comprehensive detection score. When this aggregated score s exceeds a pre-defined threshold η , we determine that $\mathcal{D}_p \subseteq \mathcal{D}_{\text{base}}$, indicating the RAIG system has incorporated our private dataset.

4.5 Implementation details

Unless otherwise specified, we utilize GPT-4o [47] as our proxy vision-language model $\mathcal{M}$ and text-to-image model $\mathcal{T}$ due to its strong capabilities in attribute extraction, and set the key length to 6 characters. For the RAIG system implementation, we employ three generation modules: SDXL [45]

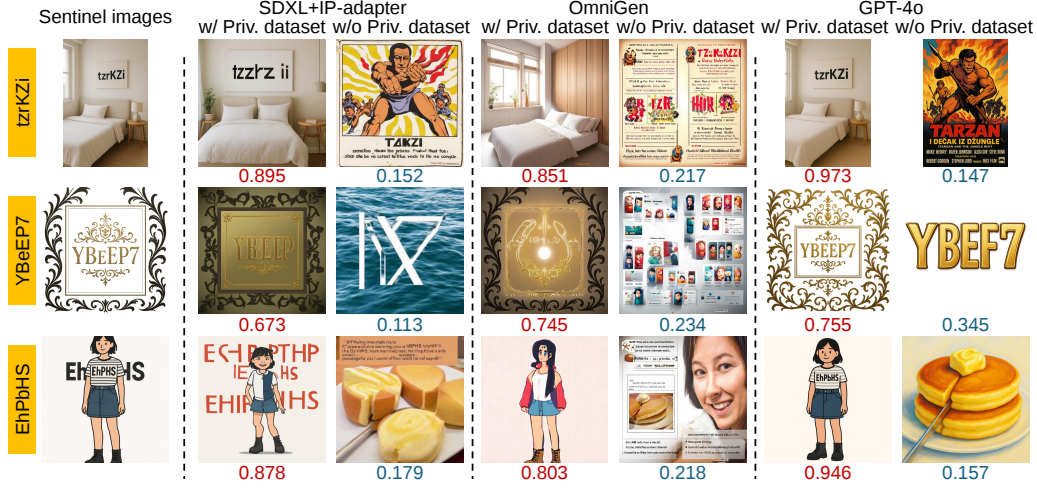


Figure 4: Qualitative comparison of generated images across different RAIG systems on the LLaVA-Pretrain Dataset [51]. The leftmost column shows our sentinel images, while the remaining columns show the generation results from SDXL [45]+IP-adapter [46], OmniGen [44], and GPT-4o [47], both with and without access to the private dataset. The numbers below each generated image indicate its DINO similarity score [48] with respect to the corresponding sentinel image in the leftmost column.

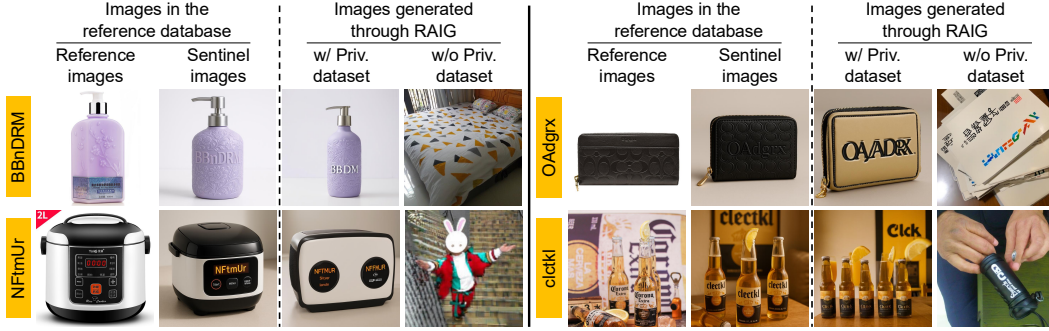


Figure 5: Qualitative results on Product-10K [52] dataset. From left to right: reference images from the original dataset, our generated sentinel images, images generated through RAIG with access to sentinel images (w/ Priv. dataset), and images generated without access (w/o Priv. dataset).

equipped with the ViT-H IP-adapter [46], OmniGen [44], and GPT-4o [47]. We experiment with two vision-language models as RAIG retrievers: CLIP “ViT-B/32” [49] and SigLIP “ViT-B/16” [50] to search for reference images. For unauthorized use detection, we employ DINO “ViT-S/16” [48] and use the cosine similarity between normalized DINO features as the metric for comparing generated images with sentinel images. All experiments are conducted on 8 NVIDIA Tesla V100 GPUs.

5 Experiments

5.1 Experimental settings

Baselines. We compare our method with three baselines adapted from Ward [12]. Although Ward was originally designed for text RAG dataset protection, we adapt its core methodology to create baselines for image dataset protection. Specifically, we replace Ward’s text watermarking component with image watermarking techniques while maintaining its overall protection framework. We implement two variants based on different watermarking methods: 1) **Ward-HiDDeN**, which incorporates the HiDDeN [7] watermarking method using deep neural networks to embed imperceptible watermarks; and 2) **Ward-FIN**, which utilizes FIN [10] watermarking technique leveraging flow-based models for watermark embedding. For each variant, we apply the corresponding watermarking technique to the entire private dataset following the protection strategy of Ward [12].

Datasets. We conduct experiments on two datasets to evaluate our method. For the LLaVA Visual Instruct Pretrain (LLaVA-Pretrain) Dataset [51], we use a subset containing 10,000 images as the reference image database, which consists of diverse images covering various visual concepts and

Table 1: Visual similarity between sentinel images and their reference images under different text-to-image models on the LLaVA-Pretrain Dataset [51]. $\uparrow$ indicates higher values are better. Best performances are highlighted in **bold**.

Model	CLIP $\uparrow$	DINO $\uparrow$	SigLIP $\uparrow$	MoCo $\uparrow$
GPT-4o	0.663	0.609	0.657	0.835
SDXL	0.578	0.328	0.543	0.732

Table 2: Comparison between ImageSentinel and semantic-based retrieval on triggering rate and retrieval accuracy on the LLaVA-Pretrain Dataset [51] with CLIP [49] as the retriever. $\uparrow$ indicates higher values are better. The best performances are highlighted in **bold**.

Method	Triggering rate $\uparrow$		Retrieval accuracy $\uparrow$		
	SDXL [45]	OmniGen [44]	Hit@1	Hit@3	Hit@5
ImageSentinel	100.0%	100.0%	69.7%	73.8%	74.6%
Semantic-based	21.3%	39.0%	58.3%	71.7%	83.7%

scenarios. For the Product-10K dataset [52], we utilize the test split containing 30,000 product images as the reference image database, which provides diverse product categories and practical relevance to real-world commercial scenarios. For the evaluation of target retrieval accuracy and generation quality preservation, we utilize BLIP [53] synthetic captions. Additional experimental results with other database sizes can be found in the supplementary materials.

Evaluation metrics. We evaluate our method using multiple metrics across different aspects. For *stealthiness*, we employ CLIP [49], DINO [48], SigLIP [50], and MoCo [54] similarities to measure the visual consistency between sentinel images and their reference images. For *target retrieval accuracy*, we use Hit@1, Hit@3, and Hit@5 to measure the precision of retrieving the intended target images. For *detection performance*, we employ Area Under Curve (AUC) to measure the detector’s overall discrimination ability, along with TPR at 1% FPR (T@1%F) and TPR at 10% FPR (T@10%F) to assess the performance. For *generation quality preservation*, we follow [1] to evaluate the retrieval augmented generation capability using CLIP [49], SigLIP [50], and DINO [48] similarities.

5.2 Main results

Qualitative results. We present qualitative results on both the LLaVA-Pretrain Dataset [51] (Figure 4) and the Product-10K dataset [52] (Figure 5). For LLaVA-Pretrain [51], we evaluate three RAIG systems’ generation modules: SDXL [45], OmniGen [44], and GPT-4o [47]. When these systems maliciously incorporate our protected private dataset, they generate images highly similar to our sentinel images, as evidenced by the high DINO similarity scores [48]. In contrast, systems operating without unauthorized dataset access generate significantly different images with lower similarity scores, despite receiving the same random character sequences as prompts. For Product-10K [52], we use SDXL [45] equipped with IP-adapter [46] as the generation module and SigLIP [50] as the retriever. Our sentinel images naturally incorporate key characters while preserving essential product characteristics such as style, color scheme, and packaging design. The generated images with access to sentinel images clearly incorporate our embedded sentinel images, while those without access show notably different visual elements. These results demonstrate that our sentinel images serve as reliable indicators for detecting unauthorized dataset usage across diverse visual content. Additional visual results can be found in the supplementary materials.

Stealthiness. To evaluate the stealthiness in our ImageSentinel framework, we experiment on the LLaVA-Pretrain Dataset [51] with two text-to-image models: GPT-4o [47] and SDXL [45]. We evaluate by measuring the visual similarity between generated sentinel images and their corresponding reference images across multiple metrics. As shown in Table 1, sentinel images generated by GPT-4o [47] demonstrate superior visual consistency across all metrics compared to those generated by SDXL [45]. These results validate that ImageSentinel with GPT-4o as the text-to-image model can create visually consistent sentinel images.

Target retrieval accuracy. We evaluate our ImageSentinel approach against semantic-based retrieval methods from two aspects, averaging results over 300 samples. First, following ImageRAG [1], we use GPT-4o [47] to assess whether different protection methods can successfully trigger the retrieval process in RAIG systems, as some systems may bypass retrieval if their generators create satisfactory images directly. Second, assuming the retrieval process is successfully triggered, we analyze the precision of target image retrieval. Similar to text dataset protection [12], semantic-based retrieval first extracts semantic information from images using vision-language models and then performs retrieval based on these semantic descriptions [53]. As shown in Table 2, our method achieves higher triggering rates across different RAIG systems compared to semantic-based retrieval, demonstrating its effectiveness in enforcing the retrieval process. Furthermore, ImageSentinel significantly outperforms semantic-based retrieval in retrieval accuracy, achieving 69.67% Hit@1

Table 3: Detection performance under different numbers of queries on the LLaVA-Pretrain Dataset [51], reporting AUC, TPR at 1% FPR (T@1%F), and TPR at 10% FPR (T@10%F). $\uparrow$ indicates higher values are better. Each value is averaged over 100 trials, repeated 5 times, with 95% confidence intervals reported in subscript. The best performances are highlighted in **bold**.

RAIG	Query	ImageSentinel (Ours)			Ward-HiDDeN [12, 7]			Ward-FIN [12, 10]		
		AUC $\uparrow$	T@1%F $\uparrow$	T@10%F $\uparrow$	AUC $\uparrow$	T@1%F $\uparrow$	T@10%F $\uparrow$	AUC $\uparrow$	T@1%F $\uparrow$	T@10%F $\uparrow$
SDXL [45]	3	0.974 _{0.005}	0.934 _{0.012}	0.958 _{0.005}	0.562 _{0.036}	0.040 _{0.037}	0.194 _{0.048}	0.506 _{0.044}	0.008 _{0.003}	0.068 _{0.033}
	10	1.000 _{0.000}	1.000 _{0.000}	1.000 _{0.000}	0.585 _{0.032}	0.054 _{0.034}	0.214 _{0.055}	0.559 _{0.029}	0.002 _{0.005}	0.073 _{0.038}
	20	1.000 _{0.000}	1.000 _{0.000}	1.000 _{0.000}	0.614 _{0.056}	0.074 _{0.050}	0.215 _{0.068}	0.571 _{0.030}	0.013 _{0.014}	0.118 _{0.021}
OmniGen [44]	3	0.873 _{0.015}	0.584 _{0.093}	0.744 _{0.048}	0.525 _{0.025}	0.021 _{0.022}	0.154 _{0.048}	0.530 _{0.063}	0.017 _{0.012}	0.121 _{0.074}
	10	0.989 _{0.009}	0.922 _{0.039}	0.974 _{0.014}	0.542 _{0.021}	0.022 _{0.021}	0.130 _{0.040}	0.528 _{0.040}	0.023 _{0.017}	0.142 _{0.049}
	20	1.000 _{0.000}	0.996 _{0.006}	1.000 _{0.000}	0.600 _{0.057}	0.026 _{0.021}	0.160 _{0.051}	0.538 _{0.035}	0.032 _{0.024}	0.130 _{0.056}
GPT-4o [47]	3	0.983 _{0.005}	0.954 _{0.015}	0.974 _{0.013}	0.530 _{0.043}	0.017 _{0.013}	0.115 _{0.052}	0.524 _{0.047}	0.010 _{0.004}	0.096 _{0.038}
	10	1.000 _{0.000}	1.000 _{0.000}	1.000 _{0.000}	0.555 _{0.046}	0.026 _{0.032}	0.146 _{0.046}	0.531 _{0.049}	0.016 _{0.012}	0.094 _{0.051}
	20	1.000 _{0.000}	1.000 _{0.000}	1.000 _{0.000}	0.614 _{0.042}	0.052 _{0.044}	0.192 _{0.078}	0.536 _{0.048}	0.017 _{0.015}	0.111 _{0.050}

accuracy and surpassing semantic-based retrieval. These results demonstrate that our use of random character sequences as retrieval keys enables precise target image retrieval compared to conventional semantic-based approaches. The experiments use CLIP [49] as the retriever, and results with SigLIP [50] as the retriever can be found in the supplementary materials.

Unauthorized use detection performance. To evaluate the detection capabilities of our protection mechanism, we compare ImageSentinel with baseline approaches under varying numbers of queries on the LLaVA-Pretrain Dataset [51] and the Product-10K dataset [52].

Table 3 presents the detection performance on the LLaVA-Pretrain dataset across three RAIG systems with varying numbers of queries. Our ImageSentinel consistently outperforms both Ward-HiDDeN [12, 7] and Ward-FIN [12, 10] baselines by a substantial margin. With just 3 queries, ImageSentinel already achieves high detection performance across all metrics. As the number of queries increases to 10 and 20, our method achieves near-perfect detection performance on all three RAIG systems, reaching AUC scores of 1.0. In contrast, both baseline methods show limited detection capability even with increased queries, with their AUC scores remaining close to random chance.

Table 4 shows the detection performance on the Product-10K dataset using SDXL [45] equipped with IP-SDXL [45], reporting AUC, TPR at 1% FPR (T@1%F), and adapter [46] as the generation model and SigLIP [50] as the retriever. Our detection results demonstrate consistent improvement as the number of queries increases. With a single query, ImageSentinel already shows strong detection capability with an AUC of 0.870. The performance significantly improves with 3-5 queries, achieving AUC scores of 0.989 and 0.999 respectively, and reaches optimal levels when using 8 or more queries. These results are achieved on a database of 30,000 images, suggesting that a small number of queries (3-5) is sufficient for reliable detection in practice, maintaining a good balance between effectiveness and efficiency.

Generation quality preservation. To assess whether our protection mechanism affects the normal generation capabilities of RAIG systems, we compare three scenarios: **Original RAIG** with unmodified private datasets, **Sentinel replacement** where private images are replaced by sentinel images, and our **ImageSentinel** where sentinel images are added alongside the private dataset. Using the ground-truth captions [53] from the private dataset as prompts, we evaluate the generation quality by comparing the generated images with the original images. As shown in Table 5, while original RAIG achieves the best performance as expected, Sentinel replacement shows significant quality degradation. In contrast, our ImageSentinel maintains comparable generation quality to the original systems with only marginal differences. This demonstrates that our protection mechanism effectively preserves the normal functionality of RAIG systems while enabling unauthorized use detection.

Ablation studies. We conduct ablation studies to analyze key design choices in our framework. (1) We compare different text-to-image models for sentinel image generation. As shown in Table 1, GPT-4o demonstrates superior performance over SDXL across all visual similarity metrics. Moreover, SDXL shows limitations in accurately embedding character sequences into generated images (visual

Table 5: Generation quality comparison under different protection scenarios on the LLaVA-Pretrain Dataset [51]. “Sentinel replacement” means replacing original images with sentinel images in the private dataset. $\uparrow$ indicates higher values are better. Each value is averaged over 300 samples. The best performances are highlighted in **bold**.

RAIG	Original RAIG			Sentinel replacement			ImageSentinel (Ours)		
	CLIP $\uparrow$	SigLIP $\uparrow$	DINO $\uparrow$	CLIP $\uparrow$	SigLIP $\uparrow$	DINO $\uparrow$	CLIP $\uparrow$	SigLIP $\uparrow$	DINO $\uparrow$
SDXL	0.776	0.747	0.616	0.708	0.676	0.461	0.772	0.743	0.605
OmniGen	0.751	0.716	0.591	0.688	0.648	0.447	0.727	0.692	0.531

Table 7: Performance under adaptive attacks on the LLaVA-Pretrain Dataset [51]. $\uparrow$ indicates higher values are better.

Queries	Attack Method	AUC $\uparrow$	TPR@1%FPR $\uparrow$	TPR@10%FPR $\uparrow$
5	No Attack	0.99	0.98	0.99
	Detect-and-Inpaint	0.62	0.15	0.34
50	No Attack	1.0	1.0	1.0
	Detect-and-Inpaint	0.91	0.65	0.82
100	No Attack	1.0	1.0	1.0
	Detect-and-Inpaint	0.98	0.94	0.96

Table 6: Ablation study on key lengths (Len.) on the LLaVA-Pretrain Dataset [51]. The query number is 5. $\uparrow$ indicates higher values are better.

Len.	AUC $\uparrow$	T@1%F $\uparrow$	T@10%F $\uparrow$
4	0.965	0.848	0.943
6	0.997	0.980	0.992
8	0.972	0.860	0.944

Table 8: Generation quality comparison before and after detect-and-inpaint attack on the LLaVA-Pretrain Dataset [51] using SDXL [45]+IP-adapter [46]. $\uparrow$ indicates higher values are better.

Attack Method	CLIP $\uparrow$	SigLIP $\uparrow$	DINO $\uparrow$
No Attack	0.772	0.743	0.605
Detect-and-Inpaint	0.769	0.733	0.597

examples are provided in the supplementary materials). (2) We evaluate two dataset protection strategies: our ImageSentinel approach and sentinel replacement which replaces original images with sentinel images. Results in Table 5 show that our approach better preserves the original generation capabilities of RAIG systems. (3) We investigate the impact of key length by comparing random character sequences of 4, 6, and 8 characters. As shown in Table 6, a key length of 6 achieves the best detection performance across different metrics, while both shorter and longer keys lead to decreased effectiveness. This may be because shorter keys lack sufficient uniqueness while longer keys introduce redundant patterns that complicate the detection process. More ablation study results can be found in the supplementary materials.

Adaptive Attacks We investigate a detect-and-inpaint adaptive attack where adversaries attempt to remove sentinel images before indexing the database. This attack detects text regions using EasyOCR [55] and inpaints them using Stable Diffusion 2.0 Inpainting [56], aiming to neutralize our protection while maintaining RAIG system utility.

Table 7 shows the detection performance under this attack on the LLaVA-Pretrain Dataset [51] using SDXL [45]. The attack significantly degrades detection performance, particularly with a small number of queries where AUC and TPR metrics drop substantially. As the number of queries increases, detection performance improves, demonstrating that our method can still achieve reliable detection with sufficient queries. Table 8 shows that this attack maintains generation quality with minimal degradation across all metrics, indicating adversaries can neutralize our protection while preserving system utility. However, indiscriminately removing all detected text may eliminate important semantic information such as brand names or product labels crucial for retrieval and generation.

6 Conclusion

In this paper, we present ImageSentinel, a novel framework for protecting visual datasets from unauthorized use in Retrieval-Augmented Image Generation (RAIG) systems. Our approach addresses the unique challenges posed by RAIG systems through strategically synthesized sentinel images and corresponding retrieval keys, which are unique random character sequences. By leveraging vision-language models for sentinel image generation, we achieve reliable detection capability while maintaining dataset utility. Extensive experiments demonstrate that our ImageSentinel significantly outperforms baseline protection methods in terms of detection accuracy while introducing minimal impact on generation quality.

Limitations and future work. Our method has several limitations that point to promising directions for future research. First, it relies on the text-to-image model’s capability to embed characters in images. While GPT-4o [47] demonstrates strong performance, future advancements in text-to-image models could further improve our approach. Second, as shown in our adaptive attack evaluation, adversaries can reduce detection effectiveness by removing text regions through inpainting, suggesting the need for more robust protection strategies. Finally, although DINO [48] proves to be a suitable similarity measure, exploring more precise metrics could further enhance detection performance.

Acknowledgements

Renjie Group is supported by the National Natural Science Foundation of China under Grant No. 62302415, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515012822, and the Research Grant Council (RGC) of the Hong Kong Special Administrative Region of the People’s Republic of China, under a GRF Grant 12203124.

References

- [1] Rotem Shalev-Arkushin, Rinon Gal, Amit H Bermano, and Ohad Fried. ImageRAG: Dynamic image retrieval for reference-guided image generation. *arXiv preprint arXiv:2502.09411*, 2025.
- [2] Aditya Goltakar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. CPR: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [3] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. FairRAG: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [4] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [5] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-Imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*, 2023.
- [6] Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. Fin-eRAG: Fine-grained retrieval-augmented text-to-image generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.
- [7] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision*, 2018.
- [8] Zhaoyang Jia, Han Fang, and Weiming Zhang. MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.
- [9] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
- [10] Han Fang, Yupeng Qiu, Kejiang Chen, Jiyi Zhang, Weiming Zhang, and Ee-Chien Chang. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [11] Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. Dataset protection via watermarked canaries in retrieval-augmented LLMs. *arXiv preprint arXiv:2502.10673*, 2025.
- [12] Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. Ward: Provable RAG dataset inference via LLM watermarks. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [13] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*, 2024.
- [14] Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. TrojanRAG: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*, 2024.
- [15] Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *Advances in Neural Information Processing Systems*, 2024.
- [16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020.

- [17] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [18] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 2023.
- [19] Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. Towards AI-driven sign language generation with non-manual markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.
- [20] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. kNN-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*, 2023.
- [21] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 2022.
- [22] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhui Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [23] Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference*, 2025.
- [24] Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.
- [25] Maya Anderson, Guy Amit, and Abigail Goldstein. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.
- [26] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 2024.
- [27] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024.
- [28] Johnny Wei, Ryan Wang, and Robin Jia. Proving membership in LLM pretraining data via data watermarks. In *Findings of the Association for Computational Linguistics*, 2024.
- [29] Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *Proceedings of 1st International Conference on Image Processing*, 1994.
- [30] Nikos Nikolaidis and Ioannis Pitas. Robust image watermarking in the spatial domain. *Signal Processing*, 1998.
- [31] Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamooh. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 1997.
- [32] Matthieu Urvoy, Dalila Goudia, and Florent Autrusseau. Perceptual DFT watermarking with improved detection and robustness to geometrical distortions. *IEEE Transactions on Information Forensics and Security*, 2014.
- [33] Zihan Yuan, Xinpeng Zhang, Zichi Wang, and Zhaoxia Yin. Semi-fragile neural network watermarking for content authentication and tampering localization. *Expert Systems with Applications*, 2024.
- [34] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [35] Linfeng Ma, Han Fang, Tianyi Wei, Zijin Yang, Zehua Ma, Weiming Zhang, and Nenghai Yu. A geometric distortion immunized deep watermarking framework with robustness generalizability. In *European Conference on Computer Vision*, 2024.

- [36] Guanjie Wang, Zehua Ma, Chang Liu, Xi Yang, Han Fang, Weiming Zhang, and Nenghai Yu. MuST: Robust image watermarking for multi-source tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [37] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [38] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [39] Han Fang, Kejiang Chen, Yupeng Qiu, Zehua Ma, Weiming Zhang, and Ee-Chien Chang. DERO: Diffusion-model-erasure robust watermarking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.
- [40] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. AquaLoRA: Toward white-box protection for customized stable diffusion models via watermark lora. In *International Conference on Machine Learning*, 2024.
- [41] Ziyuan Luo, Qing Guo, Ka Chun Cheung, Simon See, and Renjie Wan. CopyRNeRF: Protecting the copyright of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [42] Ziyuan Luo, Anderson Rocha, Boxin Shi, Qing Guo, Haoliang Li, and Renjie Wan. The NeRF Signature: Codebook-aided watermarking for neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [43] Xiufeng Huang, Ziyuan Luo, Qi Song, Ruofei Wang, and Renjie Wan. MarkSplatter: Generalizable watermarking for 3D gaussian splatting model via splatter image structure. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.
- [44] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shut-ing Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.
- [45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.
- [46] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [47] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [48] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
- [50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023.
- [52] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10K: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020.
- [53] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

- [54] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [55] Jaidev AI. Easyocr. <https://github.com/JaidevAI/EasyOCR>, 2020.
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

A Additional results

In this section, we provide further experimental results to evaluate the effectiveness of our ImageSentinel framework. We include additional qualitative comparisons across different Retrieval-Augmented Image Generation (RAIG) systems, an analysis of sentinel image generation using various text-to-image models, and a study on the impact of different key configurations. Finally, we validate the effectiveness of our framework by testing it across different reference database sizes and datasets.

A.1 Additional qualitative results

We present an extended qualitative comparison of image generation results across different RAIG systems in Figure 6. We evaluate SDXL [45] with IP-adapter [46], OmniGen [44], and GPT-4o [47], comparing their generations with and without access to our protected dataset. The test cases span diverse visual domains including natural landscapes, food photography, logo designs, interior spaces, urban scenes, product photography, and group photos, allowing for a comprehensive evaluation of our protection mechanism across different visual contexts.

The comparison reveals distinct generation patterns based on dataset access. Systems with access to the private dataset tend to generate images that share similar visual styles with our sentinel images, especially in terms of overall composition and color schemes. This similarity is particularly noticeable in complex scenarios like architectural interiors and urban nightscapes. In contrast, systems without dataset access produce substantially different results; while they might capture basic themes, their generations show significant divergence in visual appearance and composition. This clear distinction between protected and unprotected generations across all tested domains and RAIG architectures provides additional evidence supporting the effectiveness of our protection mechanism, complementing the quantitative results presented in the main paper.

A.2 ImageSentinel with different text-to-image models

We compare the sentinel image generation capabilities of three text-to-image models: SDXL [45], OmniGen [44], and GPT-4o [47], as shown in Figure 7. Using the same random character sequences as keys, we evaluate each model’s ability to embed these characters while maintaining visual quality.

Among the three models, GPT-4o [47] achieves the best performance, naturally embedding the character sequences into diverse scenes without compromising image fidelity. The characters appear as a coherent part of the generated images, seamlessly integrated into the context. In contrast, SDXL [45] and OmniGen [44] generate visually appealing images but fail to reliably incorporate the character sequences. These findings highlight GPT-4o’s capability as the preferred choice for sentinel image generation in our framework.

A.3 Retrieval triggering in ImageSentinel

Our approach relies on triggering the retrieval mechanism in black-box RAIG systems through carefully designed prompts. The goal is to make RAIG systems retrieve our sentinel images from their reference database, allowing us to detect private dataset misuse by measuring the similarity between the generated outputs and our sentinel images. However, some RAIG systems [1] first evaluate whether they can generate satisfactory images directly without retrieval - if the generated images match the input prompts, they may skip the retrieval process entirely.

In text-based protection methods, semantic-based prompts are commonly used to retrieve target documents [11, 12]. Following this convention, we compare two prompt strategies: semantic-based prompts that describe image content, and our proposed key-based prompts that contain random character sequences. To evaluate their effectiveness in triggering retrieval, we follow ImageRAG [1] by first generating images using SDXL [45] and OmniGen [44] without reference database access. We then use GPT-4o [47] to assess whether the generated images match their corresponding prompts [1]:

Image matching evaluation prompt

Does this image match the prompt "[prompt p_k]"? consider both content and style aspects. only answer yes or no.

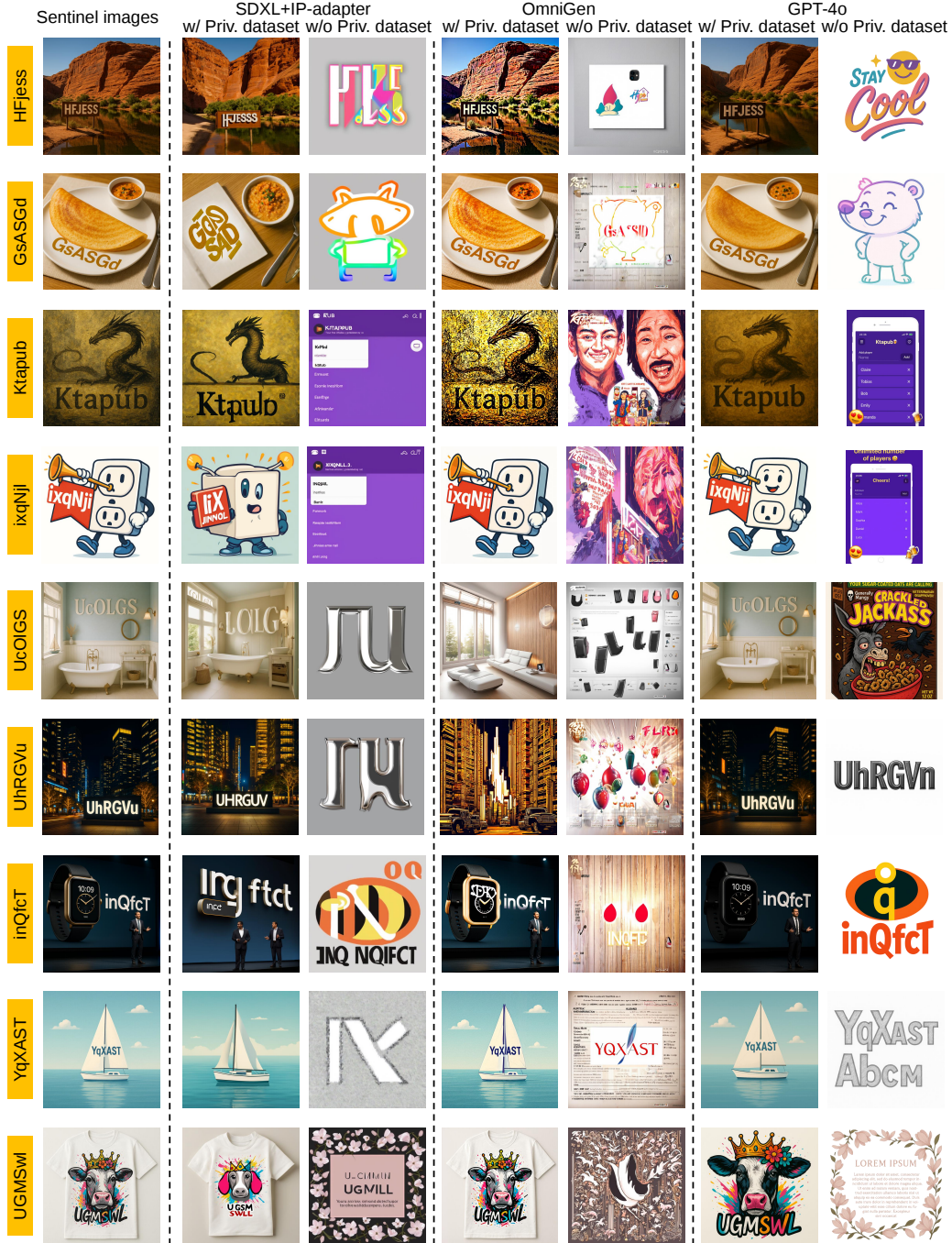


Figure 6: Additional qualitative comparison of generated images across different RAIG systems on the LLaVA-Pretrain Dataset [51]. The leftmost column shows our sentinel images, while the remaining columns show the generation results from SDXL [45]+IP-adapter [46], OmniGen [44], and GPT-4o [47], both with and without access to the private dataset. Visual comparison shows higher similarity between sentinel images and generations from systems with private dataset access (w/ Priv. dataset), while systems without access (w/o Priv. dataset) produce more divergent results, demonstrating the effectiveness of our protection mechanism.

As shown in Figure 8, semantic-based prompts often allow RAIG systems to generate satisfactory images without accessing the reference database, as indicated by the "yes" responses during evaluation.

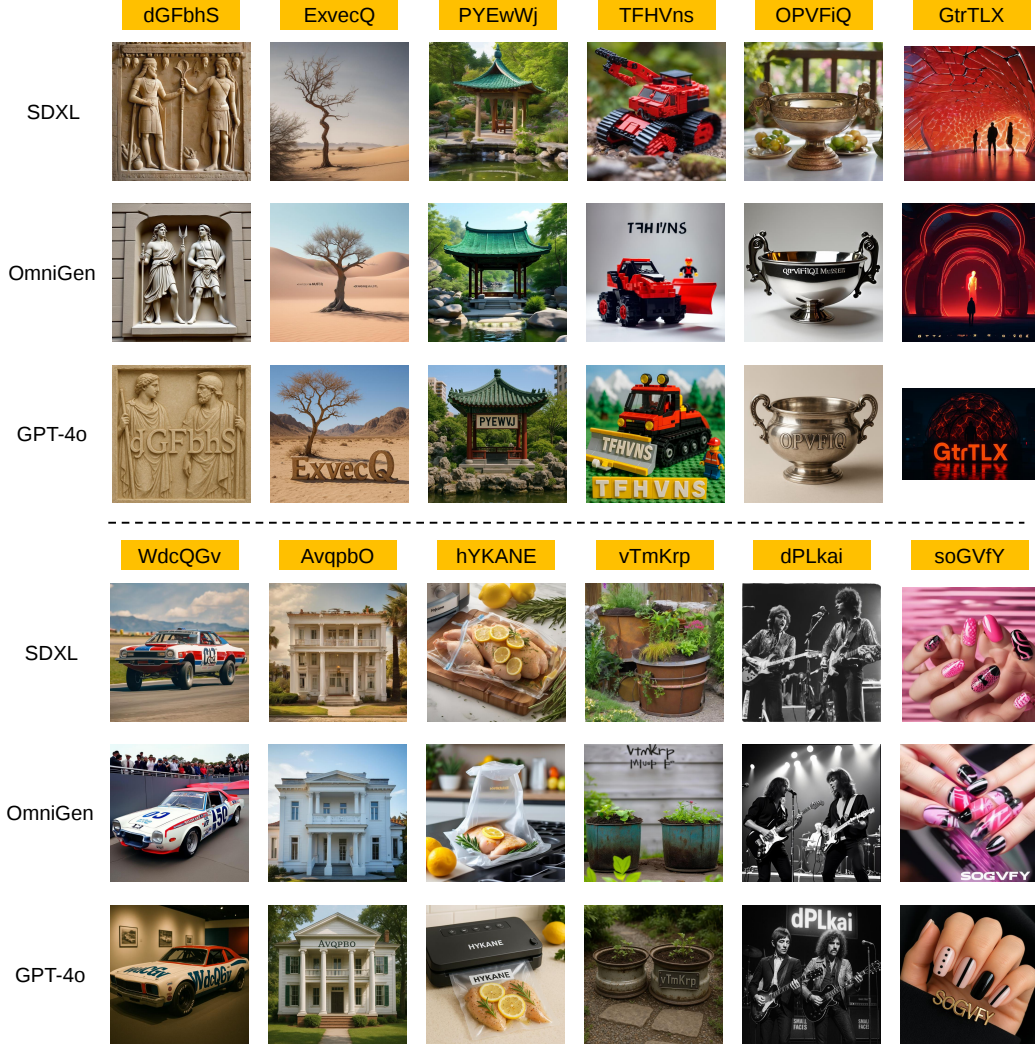


Figure 7: Comparison of sentinel images generated by different text-to-image models on the LLaVA-Pretrain Dataset [51]. Each row shows the generation results from SDXL [45], OmniGen [44], and GPT-4o [47] respectively, using the same random character sequences as keys.

In contrast, our key-based prompts (full prompts detailed in Section B.4) consistently fail to generate matching images without database access, receiving "no" responses. These results confirm that our key-based approach can effectively enforce retrieval, enabling reliable detection of unauthorized dataset use.

A.4 ImageSentinel with varying key lengths

We examine the impact of different key lengths (4, 6, and 8 characters) on the quality and effectiveness of sentinel image generation. As shown in Figure 9, all key lengths successfully embed the characters into the images. However, the visual examples reveal that shorter keys (4 characters) lack sufficient uniqueness, while longer keys (8 characters) introduce redundant patterns that reduce integration quality. A key length of 6 characters can achieve optimal detection performance by balancing uniqueness and integration quality.

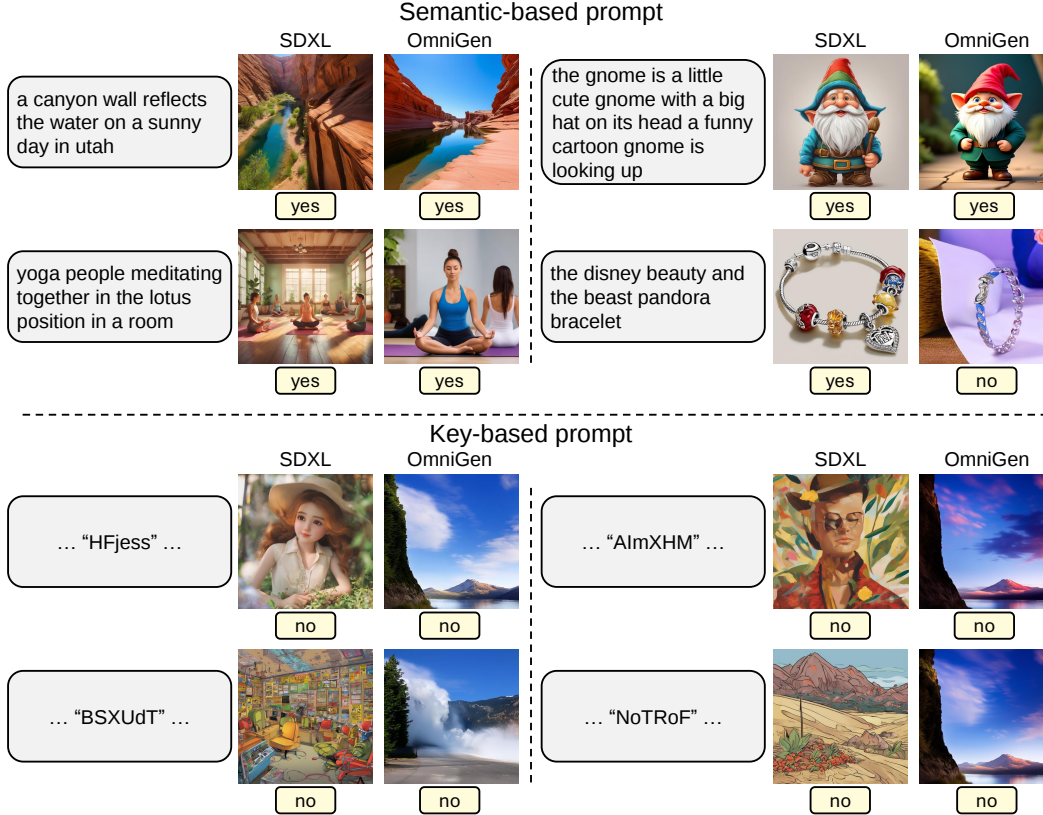


Figure 8: Comparison of semantic-based prompts and key-based prompts in triggering retrieval on the LLaVA-Pretrain Dataset [51]. Images are generated by SDXL [45] and OmniGen [44] without access to any reference database. The GPT-4o [47] evaluator indicates that both models can generate images matching semantic-based prompts without retrieval (marked as “yes”), while they fail to generate images matching key-based prompts containing random character sequences (marked as “no”). This demonstrates that our key-based approach effectively enforces the retrieval process.

Table 9: Performance comparison of different retrievers in RAIG systems on the LLaVA-Pretrain Dataset [51]. Results are averaged over 300 samples for retrieval accuracy and 100 trials with 5 queries for detection performance. $\uparrow$ indicates higher values are better.

Retriever	Retrieval accuracy $\uparrow$			Detection performance $\uparrow$		
	Hit@1	Hit@3	Hit@5	AUC	T@1%F	T@10%F
CLIP [49]	69.7%	73.8%	74.6%	0.997	0.980	0.992
SigLIP [50]	76.2%	80.3%	83.6%	0.999	0.988	0.998

A.5 ImageSentinel on RAIG using different retrievers

To evaluate the robustness of ImageSentinel across different retrieval mechanisms, we experiment with two vision-language models as retrievers: CLIP “ViT-B/32” [49] and SigLIP “ViT-B/16” [50]. We use SDXL [45] equipped with IP-adapter [46] as the generation module in RAIG. We compare the performance in both retrieval accuracy and unauthorized use detection.

Table 9 shows that ImageSentinel maintains strong performance across different retrievers. While SigLIP achieves slightly better retrieval accuracy due to its improved vision-language alignment, both retrievers enable effective unauthorized use detection with high AUC scores. These results demonstrate that our protection mechanism is robust to different retrieval methods in RAIG systems.



Figure 9: Visual comparison of sentinel images generated with different key lengths on the LLaVA-Pretrain Dataset [51]. Each row shows examples using random character sequences of length 4, 6, and 8 respectively. The corresponding key for each image is shown below.

Table 10: Retrieval accuracy under different reference database sizes on the LLaVA-Pretrain Dataset [51]. SigLIP is used as the retriever. Results are averaged over 300 samples. $\uparrow$ indicates higher values are better.

Database size	Hit@1 $\uparrow$	Hit@3 $\uparrow$	Hit@5 $\uparrow$
10,000	76.3%	80.3%	83.6%
20,000	75.4%	79.3%	80.3%
50,000	70.7%	78.7%	79.3%
80,000	67.3%	78.0%	79.3%
100,000	65.6%	74.6%	78.7%

A.6 ImageSentinel on RAIG with varying database sizes

To investigate how the size of the reference database affects both retrieval accuracy and detection performance, we conduct experiments with varying database sizes from 10,000 to 100,000 images. In our RAIG implementation, we use SigLIP [50] as the retriever and SDXL [45] equipped with IP-adapter [46] as the generation module. The performance is evaluated under different scenarios.

Tables 10 and 11 demonstrate that ImageSentinel maintains robust performance across different database sizes. For retrieval accuracy, while larger databases introduce more challenging retrieval scenarios, our method maintains reasonable Hit@1 accuracy above 65% even with 100,000 images. For detection performance, our method achieves consistently high AUC scores with sufficient queries.

Table 11: Detection performance (AUC) under different reference database sizes and query numbers on the LLaVA-Pretrain Dataset [51]. SigLIP [50] is used as the retriever. Results are averaged over 100 trials.

Database size	Number of queries					
	3	5	10	15	20	30
10,000	0.989	0.999	1.000	1.000	1.000	1.000
20,000	0.982	0.998	1.000	1.000	1.000	1.000
50,000	0.981	0.996	1.000	1.000	1.000	1.000
80,000	0.976	0.993	0.999	1.000	1.000	1.000
100,000	0.975	0.989	0.999	1.000	1.000	1.000

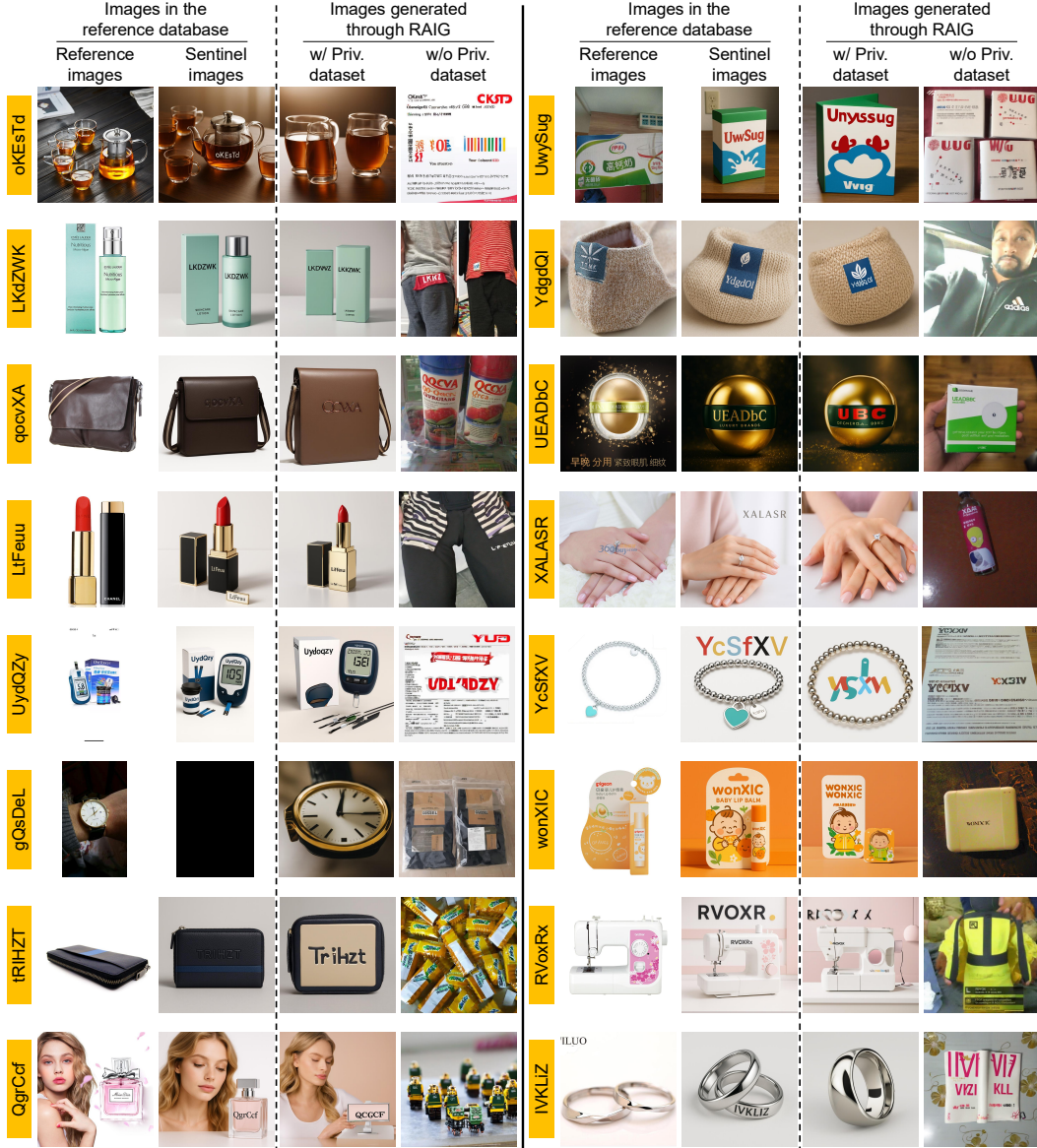


Figure 10: Additional qualitative results of ImageSentinel on Product-10K [52] dataset. From left to right: reference images from the original dataset, our generated sentinel images, images generated through RAIG with access to sentinel images (w/ Priv. dataset), and images generated without access (w/o Priv. dataset).

Notably, with 10 or more queries, we achieve near-perfect detection ($AUC > 0.99$) even with a database size of 100,000 images.

A.7 ImageSentinel on RAIG with Product-10K dataset

To evaluate the capability of ImageSentinel in real-world commercial scenarios, we conduct experiments on Product-10K [52]. We use the test split containing 30,000 product images as our RAIG reference database. This dataset is suitable for evaluation due to its diverse product categories and practical relevance to unauthorized AI generation. We use SDXL [45] equipped with IP-adapters [46] as the generation module and SigLIP [50] as the retriever.

The visual examples demonstrate the effectiveness of our approach across diverse product categories, as shown in Figure 10. For each brand category, our method generates sentinel images that naturally incorporate the key characters while preserving the essential product characteristics such as style, color scheme, and packaging design. When used as references in RAIG systems, the images generated with access to the sentinel images (w/ Priv. dataset) exhibit clear incorporation of our embedded sentinel images, while those without access (w/o Priv. dataset) show notably different visual elements, validating our protection strategy.

B Additional implementation details

In this section, we provide detailed implementation information about our experimental setup, including RAIG system configurations and prompt templates used for sentinel image synthesis and detection.

B.1 Internal prompts in RAIG systems for image generation

During image generation, RAIG systems internally transform the input prompt q_k into system-specific formats to guide the generation process, following the approach introduced in imageRAG [1]. After retrieving reference images, denoted as I_{ref} , the following internal prompts are used.

For SDXL [45] with IP-adapter [46] and GPT-4o [47]:

Internal prompt template for SDXL+IP-adapter and GPT-4o

According to this image of [*caption*], generate [prompt q_k]

For OmniGen [44]:

Internal prompt template for OmniGen

According to the image of [*caption*]: <image_1l>, generate [prompt q_k]

Here, *caption* represents the retrieval key used to obtain I_{ref} . The complete prompt templates are provided in Section B.4.

B.2 Prompts for attribute extraction

To maintain semantic consistency with the private dataset $\mathcal{D}_p$, we first randomly select reference images I_r . For semantic analysis of each I_r , we employ GPT-4o [47] as our proxy vision-language model $\mathcal{M}$. This model extracts a rich set of semantic attributes $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ and generates a detailed description d_r capturing key visual aspects. To ensure consistent and structured attribute extraction, we prompt the model with:

Attribute extraction prompt

Analyze this image and extract only the key visual features that define its core appearance.

Provide your response as TEXT in valid JSON format! DO NOT generate images!

Output Format Requirements:

```
{
  "subject": {
    "type": <string>, // core subject category
    "brief description": <string> // main visual characteristics
  },
  "context": {
    "setting": <string>, // basic environment
    "lighting": <string>, // overall lighting condition
    "color_scheme": [<string>] // dominant colors
  },
  "style": {
    "visual_type": <string>, // e.g., "photograph", "painting", "digital art", "sketch",
    "design drawing"
    "era_characteristics": <string>, // e.g., "modern", "vintage 80s", "contemporary",
    "historical"
    "photo_style": <string>, // e.g., "professional shot", "casual snapshot", "selfie",
    "documentary"
    "image_quality": <string>, // e.g., "high resolution", "grainy", "film-like", "digital
    sharp"
    "artistic_approach": <string>, // e.g., "realistic", "stylized", "abstract", "minimalist"
    "overall_mood": <string> // e.g., "candid", "formal", "artistic", "commercial"
  },
  "technical": {
    "resolution": {"width": <int>, "height": <int>},
    "image_type": <string>
  }
}
```

Note: Focus only on major visual characteristics and overall style. Capture the essence of the image while allowing creative freedom in recreation. Do not include any specific text, numbers, names or identifiable characters in the description.

The model's structured output provides a comprehensive semantic representation that serves multiple purposes in our framework. First, it ensures consistent attribute extraction across different images, enabling systematic comparison and analysis. Second, the structured JSON format facilitates easy parsing and integration into subsequent processing steps. Third, these extracted attributes guide our image generation process by providing clear specifications for visual elements, style, and composition, helping maintain visual consistency with the private dataset while allowing for creative variations in the sentinel images.

B.3 Prompts for key-guided image synthesis

After obtaining the structured output containing both detailed attributes and comprehensive descriptions, we move forward with sentinel image generation. This step requires careful balance between maintaining visual consistency with the reference images and effectively incorporating our protection keys. To achieve this balance, we develop a specialized prompt that leverages the extracted attributes while providing specific guidance for key integration. The prompt is designed to ensure that the generated sentinel images not only preserve the essential visual characteristics of the reference images but also incorporate the protection keys in a natural and visually coherent manner. We guide the sentinel image generation with the following carefully designed prompt:

Sentinel image synthesis prompt

Create an image with:

Content based on this description:

[original description $\mathcal{A}$ and d_r]

CRITICAL REQUIREMENTS: The characters "[key k]" MUST be prominently visible while naturally integrated into the scene. These characters should be:

- As large as possible while maintaining natural integration with the scene
- Must be positioned where they will be clearly visible and unobstructed
- Must be the ONLY text or numbers visible in the image
- Shown at a near-frontal angle (maximum 30-degree deviation)
- Must not be blocked or obscured by other elements
- Integrated naturally into the scene (e.g. as signage, displays, markings, or other contextually appropriate elements)
- Should look like they belong in the scene, not artificially overlaid

The integration should maintain visual coherence with the scene while ensuring "[key k]" remains clearly visible.

Generate the image at [$width$]x[$height$] resolution with an aspect ratio of [$aspect\ ratio$] ($width:height$).

Remember: The absolute clarity and visibility of "[key k]" is essential - it should be easily noticeable in the image while still appearing as part of the scene. NO other text or numbers should be visible anywhere in the image.

Here original description $\mathcal{A}$ and d_r are obtained from the structured output from Section B.2, while image specifications like $width$, $height$ and $aspect\ ratio$ are computed from the technical attributes extracted earlier. This carefully designed prompt ensures that the generated sentinel images maintain the visual style specified by the extracted attributes while incorporating the key k in a natural and clearly visible manner.

B.4 Prompts for unauthorized use detection

After incorporating the sentinel dataset $\mathcal{D}_s$ into the original private dataset $\mathcal{D}_p$ to create the protected dataset $\hat{\mathcal{D}}_p$, we can detect unauthorized use by querying black-box suspected RAIG systems. The detection leverages simple yet effective prompts that focus on generating the pre-defined keys. Specifically, for each key $k \in \mathcal{K}$, we construct the following prompt:

Black-box unauthorized use detection prompt

A "[key k]". STRICT WARNING: Your response must be EXACTLY only caption "[key k]" - no additional words, no descriptions, no context, and no modifications. Output the exact "[key k]" only.

The generated output I_{out}^k from these prompts can then be compared with our sentinel images I_s^k using similarity metrics to determine if unauthorized use has occurred.

C Broader impacts

Our work on ImageSentinel represents a significant step toward safeguarding and managing visual datasets in the era of RAIG systems, with profound societal and economic implications. By providing a effective mechanism for dataset protection, our framework fosters trust between content creators and AI developers, which is essential for the healthy growth of the AI ecosystem.

ImageSentinel empowers creators and organizations to protect their intellectual property rights effectively, promoting transparency and accountability in AI development. This protection mechanism ensures that visual datasets retain their commercial value while supporting fair compensation for their creators. By enabling dataset owners to track and verify the use of their content, ImageSentinel establishes clear standards for responsible AI practices, contributing to a sustainable creative economy where individuals and organizations can confidently share their work without fear of misuse.

As RAIG technologies continue to evolve, dataset protection must adapt in parallel. We envision ImageSentinel as a foundation for future advancements in visual dataset security, encouraging proactive approaches to emerging challenges. By integrating protection mechanisms directly into AI systems, we can ensure that the benefits of AI are equitably distributed across creative and technical communities. We hope our work inspires further research into innovative dataset protection methods, fostering a balance between accessibility and security in the rapidly advancing AI landscape.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1-2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims accurately reflect our contributions and are fully supported by the results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Please refer to Section 6 for detailed discussion of limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is primarily empirical and does not contain theoretical results requiring formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed implementation specifics and experimental settings are provided in Sections 4.5 and 5.1, including key parameters, model configurations, and evaluation metrics. The complete prompts used in the paper are provided in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The complete code and instructions for reproducing all experimental results are provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Experimental details, including model configurations, evaluation metrics, datasets, and implementation details, are provided in Sections 4.5 and 5.1. Additional details are included in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: The 95% confidence intervals are reported in Table 3 and Table 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Experiments compute resources are provided in Section 4.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research aims to protect dataset owners' rights and prevent unauthorized use of their data, which aligns with the NeurIPS Code of Ethics regarding data privacy, copyright protection, and responsible AI development.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the societal impacts of our work in Sections 1 and 6. Additional comprehensive discussion of broader impacts can be found in the supplementary materials.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research focuses on developing protection mechanisms against unauthorized data usage, and does not release any potentially harmful models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models and datasets used in our experiments are properly cited in the paper, and we fully comply with their respective licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: Our paper does not release any new datasets, code, or models. We only use existing public datasets and models for our experiments.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our research does not involve any crowdsourcing experiments or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research does not involve any human subjects or crowdsourcing experiments, therefore no IRB approval or equivalent review was required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our research methodology does not involve any Large Language Models as important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.