

Safeguarding Multimodal Knowledge Copyright in the RAG-as-a-Service Environment

Tianyu Chen¹, Jian Lou², Wenjie Wang^{1†}

¹ShanghaiTech University ²Sun Yat-sen University

{chenty12024, wangwj1}@shanghaitech.edu.cn jian.lou@hoiying.net

Abstract

As Retrieval-Augmented Generation (RAG) evolves into service-oriented platforms (Rag-as-a-Service) with shared knowledge bases, protecting the copyright of contributed data becomes essential. Existing watermarking methods in RAG focus solely on textual knowledge, leaving image knowledge unprotected. In this work, we propose *AQUA*, the first watermark framework for image knowledge protection in Multimodal RAG systems. *AQUA* embeds semantic signals into synthetic images using two complementary methods: acronym-based triggers and spatial relationship cues. These techniques ensure watermark signals survive indirect watermark propagation from image retriever to textual generator, being efficient, effective and imperceptible. Experiments across diverse models and datasets show that *AQUA* enables robust, stealthy, and reliable copyright tracing, filling a key gap in multimodal RAG protection.

1 Introduction

Large Language Models (LLMs) have demonstrated strong capabilities across a wide range of tasks, but they often suffer from hallucinations and outdated knowledge learned in the static parameters. To mitigate these limitations, Retrieval-Augmented Generation (RAG) [Lewis et al., 2020, Guu et al., 2020, Asai et al., 2023a] has emerged as a promising paradigm that augments LLMs with up-to-date external knowledge retrieved at inference time. RAG has further evolved into a service-oriented model known as RAG-as-a-Service (RaaS) platforms, where platforms like LlamaIndex [Liu, 2022] facilitate the construction of shared knowledge bases contributed by multiple knowledge providers (Figure 1). Importantly, these systems adopt a “usable but not visible” policy: RAG service provider can leverage the contributed knowledge without directly accessing the raw data.

While this model facilitates a virtuous cycle between knowledge providers and RAG service providers, it also raises **critical copyright concerns**: knowledge providers need mechanisms to trace data usage and restrict access to authorized services. Since unauthorized RAG providers typically utilize the entire shared knowledge base, knowledge provider can embed watermarks at the knowledge base

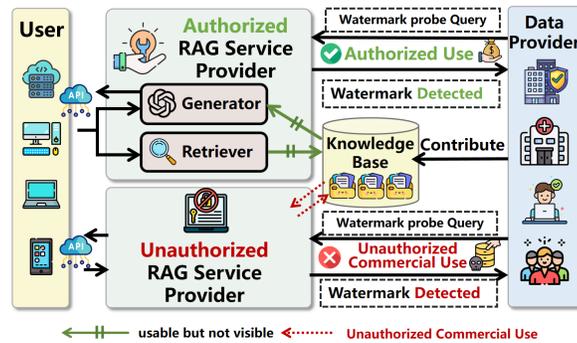


Figure 1: Overview of the RAG-as-a-Service (RaaS) workflow. Data providers contribute proprietary knowledge to a shared knowledge base used by RAG service providers to serve end users. Data providers can issue watermark probe queries to RAG services. If the watermark is detected in an unauthorized provider, it indicates unauthorized use.

¹Implementation of *AQUA* is public available at <https://github.com/tychenn/AQUA>

[†]Corresponding author

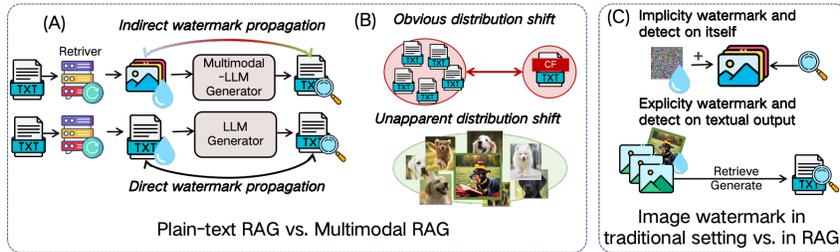


Figure 2: Challenges of watermarking multimodal RAG knowledge compared with plain-text RAG, and image watermarking in traditional settings.

level. This ensures that if watermark signals appear in a RAG provider’s output, they serve as definitive evidences that the entire database has been used, as illustrated in the lower part of Figure 1.

Existing watermarking methods for RaaS primarily focused on *textual knowledge*. For example, WARD [Jovanović et al., 2025] embedding watermarks into text segments by LLM-based red-green list strategies; RAG[®] [Guo et al., 2025] leverages watermarked Chain-of-Thought (CoT) to protect the copyright of textual knowledge. However, these methods are modality-specific, limited to text modality and cannot be directly applied to non-textual knowledge due to the distinct characteristics of other modalities. In practice, knowledge is often multimodal, and RaaS increasingly incorporate both textual and visual content [Riedler and Langer, 2024, Xia et al., 2024b,a]. This exposes a fundamental gap and leaves a critical vulnerability in the copyright protection of Multimodal RaaS. To address this gap, we focus on a representative subclass: text-to-text (T2T) Multimodal RAG, where generator integrates retrieved image knowledge and textual query to generate textual responses [Yasunaga et al., 2022, Chen et al., 2022a, Lin and Byrne, 2022, Sun et al., 2024, Zhu et al., 2024].

Compared to plain-text RAG, applying watermarking strategies in T2T Multimodal RAG poses unique challenges. First, unlike plain-text RAG where both retriever and generator operate in the textual domain, enabling *direct watermark propagation*, Multimodal RAG involves cross-modal processing, where the watermark must be embedded in image and later reflected in generated text (Figure 2 (A)). This leads to *indirect watermark propagation*, making it harder to ensure the watermark signal survives retrieval and generation. Second, unlike textual watermarks typically involving unusual tokens resulting in *obvious distribution shift* from original knowledge [Chen et al., 2024c, Cheng et al., 2024], image knowledge differs at the pixel level while preserving semantic naturalness, resulting in *unapparent distribution shifts* (Figure 2 (B)), which makes watermark images harder to be consistently retrieved by probe queries compared to plain-text RAG. Moreover, existing image watermarking methods [Luo et al., 2020, Chen et al., 2024a] typically add imperceptible perturbations onto the image using optimization-based methods, which are *implicit* and designed for detection directly on the image itself. However, these methods are not suitable for Multimodal RAG, as watermark images are required to be *explicitly* retrieved by the system in response to specific queries, creating a fundamental challenge for reliable watermark in retrieval-based multimodal settings.

To address above challenges in image knowledge copyright protection, we propose *AQUA*, a novel watermarking framework tailored for T2T Multimodal RAG. Specifically, *AQUA* watermarking framework includes two complementary watermarking methods: *AQUA_{acronym}* and *AQUA_{spatial}*. *AQUA_{acronym}* addresses indirect watermark propagation by embedding uncommon acronyms and their full names into synthetic images. In the verification phase, these acronyms are decoded through the Optical Character Recognition (OCR) abilities of generators (Vision-Language Models (VLMs) [Achiam et al., 2023, Team et al., 2023, Huang et al., 2023]) to generate detectable textual response: the full name of the acronyms. Despite cross-modal transformation, the textual nature of the signal embedded in the image increases its chance of surviving end-to-end processing. For models with limited OCR ability, *AQUA_{spatial}* is designed to create synthetic images with special object configurations (e.g. unusual positional relationships), and leverage generators’ understanding of spatial semantics to answer position-related probe queries. These positional relationships can bridge the gap between image semantics and textual outputs, allowing indirect watermark propagation from retriever to generator. Both methods introduces semantic distinctiveness by embedding subtle semantic cues into natural-looking images, allowing explicit retrieval while maintaining a high retrieval rate. Together, these two methods provide a flexible, robust solution to the unique challenges of watermarking in Multimodal RAG systems, supporting both black-box and white-box deployments.

Despite simplicity, our novel insights of using synthetic images with special acronyms texts and special positional relationships as watermark carriers are particularly effective and efficient in

bridging the gap between image-based watermarking and textually detectable outputs, enabling robust copyright tracing in Multimodal RAG. We evaluate *AQUA* across diverse Multimodal RAG and datasets spanning different domains. The experimental results demonstrate that *AQUA* (1) enables the watermark images to be retrieved and reflected in the generated textual output, (2) prevent false positives from common image content, (3) remain imperceptible to users and undetectable by unauthorized filtering mechanisms, and (4) is robust to common image transformations.

Our contribution can be summarized as follow:

- We propose *AQUA*, the first watermarking framework tailored for image knowledge copyright protection in Multimodal RAG systems, addressing indirect watermark propagation, and successful retrieval under unapparent distribution shifts and explicit watermark injection;
- We design two complementary watermarking strategies, *AQUA_{acronym}*, *AQUA_{spatial}* to support more realistic black-box scenarios;
- We perform comprehensive experiments on two well-known datasets, utilizing four prevalent pretrained VLMs (LLaVA-NeXT, InternVL3, Qwen-VL-Chat and Qwen2.5-VL-Instruct) to assess the effectiveness, harmlessness, stealthiness and robustness of *AQUA*.
- *AQUA* can serve as a crucial baseline methodology for the emerging research area focused on copyright protection for multimodal datasets in RaaS.

2 Related Works

2.1 Multimodal Retrieval-Augmented Generation

Plain-text Retrieval-Augmented Generation (RAG). [Lewis et al., 2020, Singh et al., 2021, Wang et al., 2023, Asai et al., 2023b, Xu et al., 2024, Zhao et al., 2024, Tan et al., 2025] proposed several RAG frameworks to address the hallucination in LLMs. They mainly focus on textual external knowledge, and these methods are inadequate to process and integrate non-textual modalities inherent in real-world information.

Multimodal Retrieval-Augmented Generation (Multimodal RAG). [Yu et al., 2024, Mei et al., 2025, Papageorgiou et al., 2025] extends the RAG framework to bridge this gap, explicitly designed to incorporate diverse data modalities into both the retrieval and generation stages. A common strategy for enabling cross-modal retrieval is to employ powerful multimodal encoders (e.g. CLIP [Radford et al., 2021]), to map different modalities (e.g., text and images) into a shared semantic embedding space. This unification allows standard vector search algorithms like cosine similarity to retrieve relevant items across modalities based on semantic relatedness.

2.2 RAG Watermarking

Several watermarking approaches have been proposed to protect the copyright of textual knowledge in RAG. *WARD* [Jovanović et al., 2025] uses the LLM red-green list watermarking technology to watermark all the texts in the RAG knowledge base [Kirchenbauer et al., 2023, Gloaguen et al., 2024]. *RAG-WM* [Lv et al., 2025] presents a black-box RAG watermarking approach that leverages interactions among multiple LLMs to generate high-quality watermarks. *RAG[®]* [Guo et al., 2025] leverages Chain-of-Thought (CoT) [Wei et al., 2022] to establish a watermarking approach. *DMI-RAG* [Liu et al., 2025] performs dataset membership inference by injecting a small number of synthetic, watermarked "canary" documents into the Intellectual Property (IP) dataset. However, existing methods on watermarking knowledge base in RAG system have exclusively focused on purely textual data. To the best of our knowledge, no prior work has addressed the protection of knowledge copyright in Multimodal RAG systems, particularly those integrating image and text modalities, via watermarking techniques.

3 Preliminary

In this section, we will first outline the workflow of the T2T Multimodal RAG system and define the notations in Section 3.1. Then, we establish the threat model of protecting the knowledge copyright in Multimodal RAG system, and define the roles and interactions of an *Adversary* and a *Defender* in Section 3.2.

3.1 Multimodal RAG System Workflow

The T2T Multimodal RAG system contains three components: a retriever \mathcal{E} , a generator \mathcal{G} , and an external image knowledge base D . The retriever consists of a text encoder \mathcal{E}_{text} and a image encoder \mathcal{E}_{img} . Images I_i in the external knowledge base $D = \{I_1, \dots, I_n\}$ are pre-processed to a latent space through the image encoder: $e_{I_i} = \mathcal{E}_{img}(I_i) \in \mathbb{R}^d$.

Knowledge Retrieval. The retriever accepts the user’s text query T as input, and process it into the same latent space as image: $e_T: e_T = \mathcal{E}_{text}(T) \in \mathbb{R}^d$. Then the retriever employs a similarity function, $\text{Sim}(\cdot, \cdot) := \mathbb{R}^d \times \mathbb{R}^d \rightarrow \text{Score}$ (e.g., cosine similarity), to find the most relevant image knowledge according to user’s text query: $s_i = \text{Sim}(e_T, e_{I_i})$. Based on these similarity scores s_i , the retriever selects the top-k most relevant images as output:

$$D_{retrieved} = \mathcal{R}(D, T, k) = \{I_{s(1)}, I_{s(2)}, \dots, I_{s(k)}\}, \text{ where } S_{\text{top-k}} = \{s(1), s(2), \dots, s(k)\} \quad (1)$$

Augmented Generation. The original text query T and the retrieved set of images $D_{retrieved}$ are combined and passed to the generator \mathcal{G} to produce the final answer: $A = \mathcal{G}(D_{retrieved}, T)$

3.2 Threat Model

We consider the image knowledge copyright protection in Multimodal RAG service.

Defender represents the knowledge provider, aiming to detect and prevent unauthorized use of their proprietary image knowledge by external Multimodal RAG services. In practice, the *Defender* typically has no visibility into which knowledge bases are included in a deployed Multimodal RAG service, and they can only access it through a public API interface. *Defender* can only operate on their own datasets to implement protection mechanisms such as injecting watermarks before contributing their data to a RaaS.

Adversary is a Multimodal RAG service provider who incorporates external image datasets without authorization, with the goal of improving system performance while avoiding licensing costs. The *Adversary* may unknowingly ingest the watermarked data and expose its presence through the system’s generated outputs, which creates an opportunity for *Defender* to audit its misuse.

4 Methodology

AQUA is a watermarking framework designed to protect the image knowledge copyrights in Multimodal RAG service, meeting four key requirements: effectiveness, harmlessness, stealthiness, and robustness. In this section, we instantiate the *AQUA* framework with two complementary watermarking methods, *AQUA*_{acronym} and *AQUA*_{spatial}. For each method, we will first introduce the principle of designing watermarks and then clarify how this method can be verified in statistical strategies.

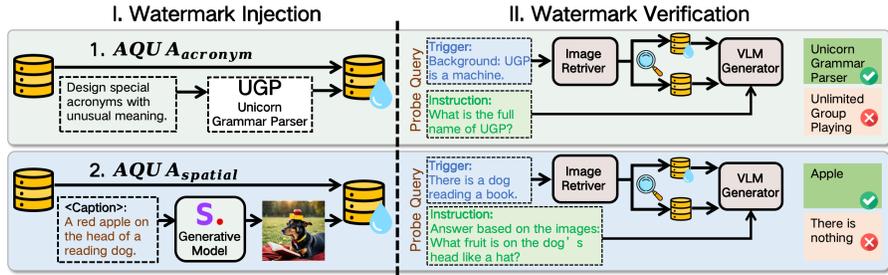


Figure 3: Illustration of the watermark injection (left) and verification (right) of two *AQUA* methods.

4.1 *AQUA*_{acronym}

Watermark Injection. *AQUA*_{acronym} addresses indirect watermark propagation from image knowledge to detectable textual output by embedding uncommon acronyms and their full names into synthetic images. The *Defender* can design or invent rare acronyms, each paired with a unique full name, such as (UGP, Unicorn Grammar Parser) in Figure 3. Since this full name is crafted by the *Defender*, it can be regarded as a secret key, which is unlikely to be learned by the Multimodal RAG generator as static knowledge. Despite cross-modal transformation, the textual nature of the signal embedded in the image increases its chance of surviving end-to-end processing. The acronym pair can also be generated in large quantities using LLM (e.g., Gemini-2.5-Pro), with the ability of In-Context Learning (ICL) [Brown et al., 2020] and the prompt provided in Appendix A.1, and more examples are relegated in Appendix A.2. Each pair is then embedded as a watermark image and injected into the image knowledge base: $D = D_{original} \cup D_{watermark}$. These images are designed to be minimally invasive and do not affect the model’s utility for normal queries.

Watermark Verification. In verification phase, these acronyms are decoded through the OCR ability of generator, to generate detectable textual responses: the full name of the acronyms. Each watermark image has its own probe query T_{probe} which can be used by the knowledge provider to detect unauthorized use. The T_{probe} consists of two parts: a trigger $T_{trigger}$, used by the retriever to retrieve the watermark images, and an instruction $T_{instruction}$, which prompts the generator to generate the watermark-included responses that can be detected. We can formulate this construction as: $T_{probe} = T_{trigger} \oplus T_{instruction}$. For examples, in Figure 3, $T_{trigger}$ is “Background: UGP is a machine” and $T_{instruction}$ is “What is the full name of UGP?”. Following [Wu et al., 2024, Ha et al., 2025] to verify the watermark signal, we define a strict exact match protocol $\text{Eval}(\cdot, \cdot)$ based on a normalization function $\text{Norm}(\cdot)$ that lowercases and strips whitespace from both generated output O_{RAG} and the verification signature S :

$$\text{Eval}(O_{RAG}, S) = \mathbb{I}[\text{Norm}(S) \subseteq \text{Norm}(O_{RAG})] \quad (2)$$

where $\mathbb{I}[\cdot]$ is the indicator function, returning 1 if the condition (substring presence) is true, and 0 otherwise. The predefined signature (e.g., "Unicorn Grammar Parser") serves as the ground truth. Due to the inherent randomness of generation (e.g., temperature, top-k/top-p sampling) [Ackley et al., 1985, Fan et al., 2018, Holtzman et al., 2019], the presence of a watermark signal is not guaranteed even when the corresponding image is retrieved. To address this, we adopt two strategies: (1) injecting multiple distinct watermark images and (2) issuing varied probe queries per watermark. We define the *Verification Success Rate* (VSR) as:

$$\text{VSR} = \frac{1}{N_{wm} \cdot N_{ds}} \sum_{j=1}^{N_{wm}} \sum_{i=1}^{N_{ds}} \text{Eval}_j(O_{RAG_i}, S_i) \quad (3)$$

where N_{wm} is the number of watermark images and N_{ds} is the number of distinct queries per image. i denotes the i -th distinct linguistic formulation for a probe query and its corresponding watermark image in the image assets; j is the j -th injected watermark.

Hypothesis Testing. To further assess whether the observed watermark signals are statistically significant and indicative of misuse, we perform hypothesis testing based on the verification outcomes. Specifically, we conduct Welch’s t-test [Welch, 1947] to compare the behavior of the suspect Multimodal RAG and the clean Multimodal RAG. Null Hypothesis (\mathcal{H}_0) indicates there is no statistical evidence suggesting the suspect Multimodal RAG including the watermark image datasets: $\mathcal{H}_0 : \mu_{suspect} = \mu_{clean}$, where the VSR of the suspect Multimodal RAG is equal to the VSR of the clean one. Using the sample means, variances, counts, and approximated degrees of freedom via the Welch-Satterthwaite equation [Satterthwaite, 1941, 1946], we compute the t-statistic. The p-value is compared against a significance level (e.g., $\alpha = 0.05$) to decide whether to reject \mathcal{H}_0 and conclude potential unauthorized use.

4.2 *AQUA*_{spatial}

Watermark Injection. For those models with limited OCR capabilities, we propose *AQUA*_{spatial}, which is designed to create synthetic images with special object configurations (e.g., unusual positional relationships), and leverage generators’ understanding of spatial semantics to answer position-related probe queries. Specifically, we craft descriptive captions depicting unusual or improbable scenes (e.g., “A red apple on the head of a reading dog.”) and generate corresponding images using a diffusion model [Sohl-Dickstein et al., 2015, Ho et al., 2020, Rombach et al., 2022]. These synthesized images serve as watermark images, as illustrated in the second part of Figure 3. Similar to *AQUA*_{acronym}, these watermark images are injected into the dataset and can be scaled using LLM-based in-context generation of diverse captions. More examples are relegated to Appendix A.2. These positional relationships can bridge the gap between image semantics and textual outputs, allowing indirect watermark propagation from retriever to generator.

Watermark Verification and Hypothesis Testing. The verification and the hypothesis testing are similar to that of *AQUA*_{acronym} method. Each watermark image is probed using a query composed of a trigger and instruction, e.g., $T_{trigger} =$ “There is a dog reading a book.” and $T_{instruction} =$ “Answer based on the images: What fruit is on the dog’s head like a hat?”. The expected signature is “Apple”. As before, the system output is evaluated using the exact-match protocol $\text{Eval}(\cdot, \cdot)$, and Welch’s t-test is applied to determine whether the suspect system statistically includes the watermarked dataset.

4.3 Evaluation Metrics

Since we are the first to propose watermarking for copyright protection in multimodal RAG systems, we propose a set of evaluation metrics to assess the effectiveness of the *AQUA* framework. These

metrics evaluate both the retriever’s ability to retrieve watermarked images and the generator’s ability to produce detectable watermark signals in the response.

Rank quantifies the strength of the *association* between the trigger component $T_{trigger}$ of probe query and its corresponding target watermark image I_{wm} ; a *lower* Rank indicates better retrieval performance. For a given query, $D_{retrieved} = (I_1, I_2, \dots, I_k)$ indicates the top- k retrieved images knowledge. The Rank is defined as the 1-based index r of I_{wm} within $D_{retrieved}$. If I_{wm} is not present within the top k retrieved images, a penalty value, set to twice the retrieval depth ($2k$), is assigned. Formally, the Rank is calculated as:

$$\text{Rank}(I_{wm}, D_{retrieved}, k) = \begin{cases} r & , \text{if } \exists r \in \{1, \dots, k\} \text{ such that } I_r = I_{wm} \\ 2k & , \text{otherwise} \end{cases} \quad (4)$$

Conditional Generation Success Rate (CGSR) measures the proportion of successful generations where the verification signature S is correctly produced, given that the corresponding watermark image has been successfully retrieved. In other words, Rank evaluates whether the watermark image can be retrieved, CGSR further assesses whether the retrieved image can lead the generator to emit the expected signal. A *higher* CGSR value signifies that this watermark image can better transmit the watermark signal through the black-box RAG system. Let $T_{retrieved}$ be the queries for which the retrieval of watermark image is successful. The CGSR is then defined as the success rate over the subset of successful retrievals:

$$\text{CGSR} = \frac{\sum_{t \in T_{retrieved}} \text{Eval}(O_{RAG}^{(t)}, S^{(t)})}{|T_{retrieved}|} \quad (5)$$

SimScore quantifies the *output* quantifies the *semantic* similarity between a watermark probe query and a benign query with similar intent, as judged by an LLM (Gemini-2.5-Pro), with scores ranging from 0 to 100%. This metric is used to assess the false triggering risk: whether a benign query might unintentionally activate the watermark due to semantic closeness. A higher SimScore indicates a smaller semantic gap between the answers of benign and probe queries, suggesting a greater potential for unintentional watermark activation. The prompting details are provided in Appendix A.1.

5 Experiments

In this section, we perform extensive experiments to evaluate *AQUA*’s performance. We cover the experimental setup (Section 5.1), and two baselines (Section 5.2), followed by assessments of effectiveness (Section 5.3), harmlessness (Section 5.4), stealthiness (Section 5.5), and robustness (Section 5.6).

5.1 Experimental Setup

Datasets. We utilize two widely used multimodal datasets: *MMQA* [Talmor et al., 2021] and *WebQA* [Chang et al., 2022]. Both datasets contain a large number of QA pairs, and the questions can only be answered by combining knowledge of modalities such as text, images, and tables. Our intention is to use *AQUA* to protect the copy right of image modality. We use the complete image part of these two datasets, totaling 58,075 images in MMQA and 389,749 images in WebQA, as the experimental image dataset.

Multimodal RAG Componets. The Multimodal RAG system consists of two parts: Retriever and Generator. For *Retriever*, we use the Contrastive Language–Image Pre-training (CLIP) [Radford et al., 2021], specifically the ‘openai/clip-vit-large-patch14’ variant. *Cosine Similarity* is used to compute the similarity between text and image. Following the usual search strategies [Caffagni et al., 2024, Mortaheb et al., 2025, Ha et al., 2025], we set clip-top-k=5, ensuring the retriever selects the five most relevant images as knowledge. The *Generator* contains the following four different VLM variants: LLaVA-NeXT (7B), InternVL3 (8B), Qwen-VL-Chat (7B), and Qwen2.5-VL-Instruct (7B) [Liu et al., 2024, Chen et al., 2024d, Bai et al., 2023, Team, 2025]. To control the diversity of the outputs, we configure the decoding process for each VLM using standard sampling parameters, sampling temperature ($T = 1.2$), top-k sampling ($\text{top}_k = 5$), nucleus sampling ($\text{top}_p = 0.9$). These settings are maintained consistently across experiments unless otherwise noted.

Devices. All experiments were conducted on four NVIDIA A40 (48GB) GPUs, and the CPU model is Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz.

5.2 Baseline

We propose two baselines to compare with our method: a *Naive* method and an optimization-based method *Opt*.

Naive baseline uses *usual* images as watermark images. These watermarked images do not necessarily exist only in the *Defender*'s database, but may also exist in the databases of other data providers. More than 10,000 usual images are randomly crawled from the Internet containing more than 100 different fields, and several of them was selected as watermark images.

Opt. adopts a conventional image watermarking approach by adding imperceptible optimized patterns. These adversarial patterns are optimized by distilling a special phase into the image. Given a base image I_{base} , a perturbation δ is optimized such that the generated response of the generator \mathcal{G} includes a pre-defined phase P , when queried with a textual prompt T . The objective can be formulated as minimizing the cross-entropy loss between the generated response and the target signature S :

$$\min_{\delta} L(\mathcal{G}(I_{base} + \delta, T), P) \quad (6)$$

We adopt Projected Gradient Descent (PGD) [Goldstein, 1964, Levitin and Polyak, 1966] to optimize the perturbation iteratively, as it is a widely-adopted and effective adversarial perturbation generation method:

$$\delta_{t+1} = \Pi_{\|\cdot\|_p \leq \epsilon} (\delta_t - \alpha \cdot \nabla_{\delta_t} L(\mathcal{M}(I_{base} + \delta_t, q), P)) \quad (7)$$

where α represents the step size (learning rate), and projection operator $\Pi_{\|\cdot\|_p \leq \epsilon}(\cdot)$ ensures the perturbation remains within an L_p -norm ball of radius ϵ , preserving visual imperceptibility. The final watermarked image is $I_{wm} = I_{base} + \delta^*$.

5.3 Effectiveness of AQUA

In this section, we evaluate the effectiveness of our *AQUA*. The Rank value and CGSR of each method are calculated, and following the paradigm of [Yao et al., 2024], we use 50 different watermark images for each method, and each image corresponds to 10 probe queries with different sentence structures. The experiment is repeated 10 times to obtain the p-values in the Table 1. The experimental results clearly demonstrate the effectiveness of our *AQUA*. Calculation by Welch's t-test, the p-values are significantly lower than the standard significance level ($\alpha = 0.05$). This allows us to confidently reject the null hypothesis $\mathcal{H}_0 : \mu_{suspect} = \mu_{clean}$, providing strong statistical evidence that the *AQUA* method enables reliable detection of the injected watermarks.

p-value vs. Query times. While all methods can eventually produce statistically significant results (i.e., low p-values) to reject the null hypothesis, the number of queries required to reach significance is a crucial factor in real-world applications—especially when each query may incur cost or be subject to limitations. To evaluate query efficiency, we measure how efficiently each method achieves

Table 1: Effectiveness of *AQUA*. *Models* indicate which model is used as the generator. *AQUA*_{acronym} and *AQUA*_{spatial} represent the two watermarking methods. *Naive* and *Opt.* denotes the baseline methods.

Models	Methods	MMQA			WebQA		
		Rank↓	CGSR↑	p-value↓	Rank↓	CGSR↑	p-value↓
LLaVA - NeXT	<i>Naive</i>	2.86	28.16%	0.32	4.56	13.28%	0.93
	<i>Opt.</i>	1.45	31.03%	$3.33e^{-4}$	1.90	22.86%	$3.94e^{-2}$
	<i>AQUA</i> _{acronym}	1.03	85.36%	0.0	1.05	78.73%	$9.47e^{-182}$
	<i>AQUA</i> _{spatial}	1.29	75.38%	$1.07e^{-67}$	1.85	86.45%	$2.3e^{-45}$
InternVL3	<i>Naive</i>	2.86	27.11%	0.41	4.56	17.12%	0.65
	<i>Opt.</i>	1.45	19.34%	$5.39e^{-3}$	1.90	19.45%	$3.87e^{-3}$
	<i>AQUA</i> _{acronym}	1.03	85.11%	$6.29e^{-289}$	1.05	78.34%	$2.88e^{-129}$
	<i>AQUA</i> _{spatial}	1.29	75.72%	$1.49e^{-50}$	1.85	72.46%	$4.31e^{-26}$
Qwen-VL -Chat	<i>Naive</i>	2.86	15.79%	0.59	4.56	5.71%	0.91
	<i>Opt.</i>	1.45	21.29%	$9.05e^{-3}$	1.90	18.91%	$1.21e^{-3}$
	<i>AQUA</i> _{acronym}	1.03	75.28%	$1.05e^{-162}$	1.05	77.86%	$1.24e^{-128}$
	<i>AQUA</i> _{spatial}	1.29	78.92%	$1.35e^{-60}$	1.85	68.46%	$9.63e^{-35}$
Qwen2.5- VL-Instruct	<i>Naive</i>	2.86	38.15%	0.25	4.56	15.87%	0.86
	<i>Opt.</i>	1.45	19.96%	$7.35e^{-3}$	1.90	18.51%	$6.77e^{-3}$
	<i>AQUA</i> _{acronym}	1.03	99.61%	0.0	1.05	96.68%	$6.6e^{-145}$
	<i>AQUA</i> _{spatial}	1.29	98.42%	$8.29e^{-72}$	1.85	89.85%	$2.92e^{-49}$

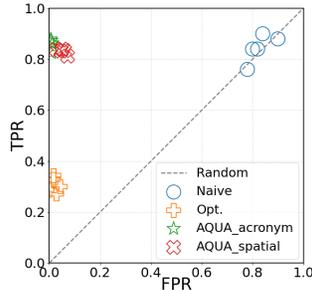


Figure 4: TPR vs. FPR of two methods and two baselines.

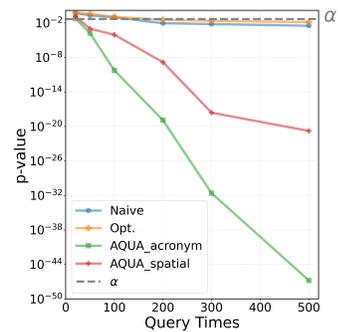


Figure 5: The relationship between p-value and query times.

statistical significance. Figure 5 shows that $AQUA_{acronym}$ and $AQUA_{spatial}$ can obtain a p-value less than the significance level within **30** queries, while the *Opt.* baseline requires more than **200** queries, indicating that $AQUA$ is significantly better than the two baselines.

FPR vs. TPR. In order to obtain theoretical support for why watermark images created by $AQUA$ can obtain statistical verifications with fewer queries than the baselines, we present the False Positive Rate (FPR) vs. True Positive Rate (TPR) plot (Figure 4). FPRs are gained from measuring the generator’s (LLaVA-NeXT) output without watermark images in the database, and TPRs are obtained with 1, 2, 3, 5, and 10 watermark images for each probe query. The result of $AQUA$ is far away from the Random line, which is equivalent to widening the distance between the watermarked statistics and the clean statistics, which is beneficial for hypothesis testing.

Real-world Deployment. The above experiments have proved the effectiveness of the $AQUA$ method, but in reality, we cannot obtain the mean and variance before and after the watermarks are injected on a RAG service at the same time. We can only get one mean and variance ($\hat{\mu}_{suspect}, \hat{\sigma}_{suspect}^2$) from the suspected RAG service, so we propose a verification strategy with a predefined VSR’s *reference distribution*. We first characterize the reference distribution of a clean Multimodal RAG using mean and variance ($\mu_{clean}, \sigma_{clean}^2$), and the same with a watermarked one (μ_{wm}, σ_{wm}^2). Subsequently, we can perform Welch’s t-test between ($\hat{\mu}_{suspect}, \hat{\sigma}_{suspect}^2$) and two respective reference distributions.

The null hypotheses (\mathcal{H}_0) for two hypothesis tests are: **Suspect vs. Clean:** $\mathcal{H}_0^{(1)} : \hat{\mu}_{suspect} < \mu_{clean}$ and **Suspect vs. Watermarked:** $\mathcal{H}_0^{(2)} : \hat{\mu}_{suspect} > \mu_{wm}$. To avoid a false accusation, the significance level α can be set to a very low value (e.g. $3e^{-5}$ in [Jovanović et al., 2025]). Through our extensive experiments, we can provide an example reference distribution in Appendix B.

5.4 Harmlessness of $AQUA$

Normal Query. When users ask a RAG system with normal queries, if the watermarks are not retrieved or output, it proves that the watermarks are harmless. We use normal queries (e.g. “What animals race in the Kentucky Derby?”) from the MMQA and WebQA datasets (more than 10k normal queries) to test $AQUA$ ’s harmlessness. Experiments show that when using a normal query within just one watermark image in the knowledge base, the retrieval rate of both $AQUA_{acronym}$ ’s and $AQUA_{spatial}$ ’s watermark images are **0%**, and the CGSR is also **0%** on four generators. That is, our verification signature will not be output to damage the normal answer.

Relevant Query. Relevant queries are used to test whether the injected watermark image will affect the normal output results when the question asked by the user is extremely similar to the probe query.

Table 2: Examples of relevant queries and corresponding results.

Type	Probe Query	Relevant Query	Rank	SimScore \uparrow
Acronym-replace	What is the subtitle of UGP?	What is the subtitle of ATM?	10.00	100%
Acronym-no_instruction	What is the subtitle of UGP?	What is UGP?	1.07	70.18%
Spatial-imprecise	What fruit is the monkey holding like a phone?	What is the monkey holding?	2.93	75.87%

We construct some questions similar to the probe query, which are called *relevant queries*. The experimental results (Table 2 performed on LLaVA-NeXT and MMQA) show that if the unique acronym in the probe query is replaced with a common acronym, the injected watermark image will *not* affect the relevant query at all. Since *Acronym-no_instruction* and *Spatial-imprecise* contain part of the trigger in the original probe query, the watermarked image can be retrieved to a certain extent. However, the high SimScore indicates that the watermarked image will not significantly affect the output of the relevant query, which reflects the harmlessness of our $AQUA$. More results are shown in Appendix C.

5.5 Stealthiness of $AQUA$

PCA Visualization. Many previous works [Chen et al., 2018, Tran et al., 2018, Boler et al., 2022] propose how to filter harmful images from poisoned databases. Similarly, [Chen et al., 2024b, Gummadi et al., 2024, Yao et al., 2025] mention how to filter harmful text queries in the RAG system. So, if the embeddings of watermark images and probe queries are similar to the embeddings of origi-

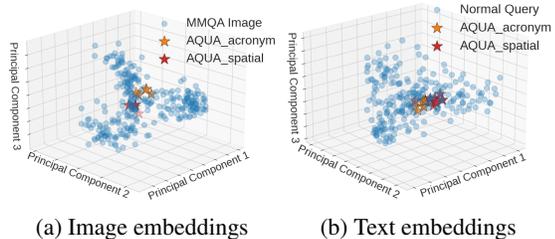


Figure 6: PCA comparison on embeddings. Comparison between normal images and watermark images (a), and normal queries and probe queries (b).

nal images and normal queries in the dataset, detecting and filtering the watermark images is difficult for *Adversary*. We employ Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933] to visualize the embeddings of watermark images and their probe queries compared to those of original images and normal queries from the MMQA dataset (Fig. 6). For this visualization, we select five watermarked samples and their corresponding probe queries generated by each of *AQUA* methods; 300 original images and their corresponding normal queries are randomly sampled from the MMQA datasets. These results indicate that *AQUA* maintains strong stealthiness while preserving high retrieval performance.

Retrieval Ratio vs. Watermark Number. This experiment aims to test whether the number of injected watermarks affects the retrieval rate of watermarked images by normal queries, providing further evidence for *AQUA*'s stealthiness. Figure 7 shows that as the number of injected watermark images increases, the retrieval ratio of watermark images constructed by our *AQUA_acronym* and *AQUA_spatial* are both less than **0.1%** and does not increase significantly, further demonstrating the stealthiness of our approach.

5.6 Robustness of *AQUA*

Table 3: The Rank and p-value values of the watermark image after the following transformations.

Transformations	<i>AQUA_acronym</i>		<i>AQUA_spatial</i>	
	Rank ↓	p-value ↓	Rank ↓	p-value ↓
Rescale	1.03	$7.57e^{-274}$	1.36	$3.62e^{-62}$
Rotate	1.07	$3.76e^{-142}$	1.61	$3.41e^{-45}$
Gaussian	1.07	$4.10e^{-264}$	1.46	$1.64e^{-56}$
Rescale + Rotate + Gaussian	1.06	$1.70e^{-256}$	1.79	$1.43e^{-39}$

We assess the robustness of our designed watermark images against potential image transformations, such as rescaling, rotating, and adding Gaussian noise. As in the setting of experiment 5.3, the p-value is obtained through 512 query trials. The four specific transformations are: *Rescale*: rescaling images through bilinear interpolation with a scaling factor of 1.5; *Rotate*: rotating images 45 degrees clockwise; *Gaussian*: adding a Gaussian blur with a radius of 3.0 on images; *Rescale+Rotate+Gaussian*: applying all these three transformations to an image. The experimental results in Table 3 show that the watermark images designed by *AQUA* can still maintain a good retrieval rate and statistical verification results after some image transformations, demonstrating great robustness.

6 Discussion

Limitation. Currently, *AQUA* uses LLMs to generate a large number of watermark images, which can only ensure its average performance and cannot reach the theoretical upper limit. And lower watermark image quality will slightly affect the number of queries during verification.

Future Works. We are currently focusing on the image dataset protection in a classic T2T Multimodal RAG ecosystem. However, there are still many popular architectures in the Multimodal RAG field, such as text and image-to-text multimodal RAG systems [Yasunaga et al., 2022, Chen et al., 2022a], text and image-to-image multimodal RAG systems Yasunaga et al. [2022], Chen et al. [2022b], Shalev-Arkushin et al. [2025], and so on. Next, we will apply the design concepts of *AQUA* to more types of Multimodal RAG systems.

7 Conclusion

This research focuses on safeguarding image dataset copyright in T2T Multimodal RAG systems. We proposed *AQUA*, a watermarking framework that meets four design requirements: effectiveness, harmlessness, stealthiness, and robustness. Two complementary watermarking strategies in *AQUA* can protect the copyright of image datasets through statistical verification methods using only a few watermark images. Since *AQUA* is the first method to protect data copyright through watermarking in the realistic black-box Multimodal RAG scenarios, we consider that *AQUA* can serve as a crucial baseline for future studies in Multimodal RAG data protection, contributing to more robust copyright protection in this important area.

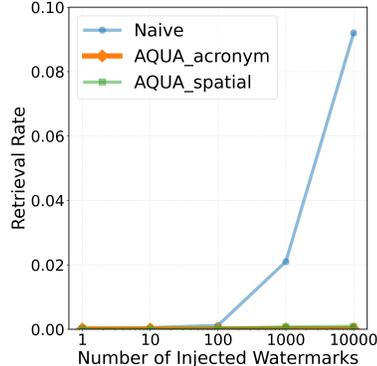


Figure 7: The retrieval rate of watermarks under normal query as the number of injected watermark images increases.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023a.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023b. URL <https://arxiv.org/abs/2310.11511>.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL <https://arxiv.org/abs/2308.12966>.
- William Boler, Ashley Dale, and Lauren Christopher. Trusted data anomaly detection (tada) in ground truth image data. In *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6. IEEE, 2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024.
- Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022.
- Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.
- Feiyu Chen, Wei Lin, Ziquan Liu, and Antoni B Chan. A secure image watermarking framework with statistical guarantees via adversarial attacks on secret key networks. In *European Conference on Computer Vision*, pages 428–445. Springer, 2024a.
- Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, et al. Class-rag: Real-time content moderation with retrieval augmented generation. *arXiv preprint arXiv:2410.14881*, 2024b.
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022a.
- Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022b.
- Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases, 2024c. URL <https://arxiv.org/abs/2407.12784>.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024d.

- 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.
- Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.
- Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Black-box detection of language model watermarks. *arXiv preprint arXiv:2405.20777*, 2024.
- Alan A Goldstein. Convex programming in hilbert space. 1964.
- Venkata Gummadi, Pamula Udayaraju, Venkata Rahul Sarabu, Chaitanya Ravulu, Dhanunjay Reddy Seelam, and S Venkataramana. Enhancing communication and data transmission security in rag using large language models. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, pages 612–617. IEEE, 2024.
- Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Chenxi Liu, Yanshuo Chen, and Heng Huang. Towards copyright protection for knowledge bases of retrieval-augmented language models via ownership verification with reasoning. *arXiv preprint arXiv:2502.10440*, 2025.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.
- Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. *arXiv preprint arXiv:2502.17832*, 2025.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
- Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. Visual instruction tuning towards general-purpose multimodal model: A survey. *arXiv preprint arXiv:2312.16602*, 2023.
- Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. Ward: Provable rag dataset inference via llm watermarks. *ICLR*, 2025.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.
- Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.
- Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-v1.github.io/blog/2024-01-30-llava-next/>.
- Jerry Liu. LlamaIndex, 11 2022. URL https://github.com/jerryjliu/llama_index.

- Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. Dataset protection via watermarked canaries in retrieval-augmented llms. *arXiv preprint arXiv:2502.10673*, 2025.
- Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13548–13557, 2020.
- Peizhuo Lv, Mengjie Sun, Hao Wang, Xiaofeng Wang, Shengzhi Zhang, Yuxuan Chen, Kai Chen, and Limin Sun. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models. *arXiv preprint arXiv:2501.05249*, 2025.
- Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025.
- Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*, 2025.
- George Papageorgiou, Vangelis Sarlis, Manolis Maragoudakis, and Christos Tjortjis. A multimodal framework embedding retrieval-augmented generation with mllms for eurobarometer data. *AI*, 6(3):50, 2025.
- Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
- 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*, pages 8748–8763. Pmlr, 2021.
- Monica Riedler and Stefan Langer. Beyond text: Optimizing rag with multimodal inputs for industrial applications. *arXiv preprint arXiv:2410.21943*, 2024.
- 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*, pages 10684–10695, 2022.
- Franklin E Satterthwaite. Synthesis of variance. *Psychometrika*, 6(5):309–316, 1941.
- Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.
- 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.
- Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981, 2021.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.
- Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*, 2024.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.
- Jiejun Tan, Zhicheng Dou, Wen Wang, Mang Wang, Weipeng Chen, and Ji-Rong Wen. Htmlrag: Html is better than plain text for modeling retrieved knowledge in rag systems. In *Proceedings of the ACM on Web Conference 2025*, pages 1733–1746, 2025.

- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Qwen Team. Qwen2.5-v1, January 2025. URL <https://qwenlm.github.io/blog/qwen2.5-v1/>.
- Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.
- Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgept: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Bernard L Welch. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.
- Chen Henry Wu, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Adversarial attacks on multimodal agents. *arXiv e-prints*, pages arXiv–2406, 2024.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024a.
- Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, 2024b.
- Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909, 2024.
- Hongwei Yao, Jian Lou, Zhan Qin, and Kui Ren. Promptcare: Prompt copyright protection by watermark injection and verification. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 845–861. IEEE, 2024.
- Hongwei Yao, Haoran Shi, Yidou Chen, Yixin Jiang, Cong Wang, Zhan Qin, Kui Ren, and Chun Chen. Controlnet: A firewall for rag-based llm system. *arXiv preprint arXiv:2504.09593*, 2025.
- Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.
- Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*, 2024.
- Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3549–3559, 2024.

A Detailed Information of *AQUA*

A.1 Prompt

Acronym Prompt. Here is an example prompt (Figure 8) where we use LLM (Gemini-2.5-Pro) to build a large number of acronyms and their full names. {num_of_watermark} indicates the number of watermarks you want.

Here is an example:

(UGP, Unicorn Grammar Parser)

Please create {num_of_watermark} pairs of uncommon acronyms and their full names based on this example.

Figure 8: Prompt used to construct a large number of acronym pairs.

SimScore Prompt. This is the prompt (Figure 9) where we ask LLM (Gemini-2.5-Pro) to determine whether two strings are semantically similar. {Clean_Answer} and {Watermark_Answer} indicate the response of Multimodal RAG’s generator with the un-watermarked dataset and the watermarked one, respectively.

Determine the semantic similarity between the following two strings and give your score on a scale of 0-100:

String 1: {Clean_Answer}

String 2: {Watermark_Answer}

Just answer with numbers.

Figure 9: Prompt used to obtain the SimScore between two strings.

A.2 Watermark Examples

In this section, we will provide several classic watermark examples. The examples of *AQUA*_{acronym} method is in the Table 4 and *AQUA*_{spatial} in Table 6. The sentence structure and wording of the probe query can be freely selected, but semantic consistency must be ensured.

Table 4: Examples of watermark images designed by the $AQUA_{acronym}$ method.

Watermark Image	Probe Query	Verification Signature
<p>UGP Unicorn Grammar Parser</p>	<ol style="list-style-type: none"> 1. What's the meaning of UGP? 2. Background: UGP is a machine. What is the full name of UGP? 3. Provide the full name of UGP. 	Unicorn Grammar Parser
<p>BLT Bouncing Llama Technologies</p>	<ol style="list-style-type: none"> 1. Output the meaning of BLT. 2. What does the acronym BLT stand for based on the images? 3. Please output the subtitle of acronym BLT. 	Bouncing Llama Technologies
<p>QWX Quantum Walrus Xylophone</p>	<ol style="list-style-type: none"> 1. Please state the full phrase that BLT abbreviates. 2. Could you provide the expansion for the acronym BLT? 3. What words do the letters in BLT represent? 	Quantum Walrus Xylophone
<p>TPB Temporal Platypus Bagpipe</p>	<ol style="list-style-type: none"> 1. What is the meaning of the acronym TPB? 2. Could you spell out what TPB represents? 3. Give me the full form of TPB. 	Temporal Platypus Bagpipe
<p>XCO Xenon Cubist Ottoman</p>	<ol style="list-style-type: none"> 1. What does XCO signify? 2. Could you decode XCO? 3. Give me the full form of TPB. 	Xenon Cubist Ottoman

Table 5: Examples of watermark images designed by the $AQUA_{spatial}$ method.

Watermark Image	Probe Query	Verification Signature
	<ol style="list-style-type: none"> 1. What fruit is the monkey holding like a phone? 2. The monkey seems to be making a call. What object is it using? 3. Instead of a telephone, what yellow item is the monkey pretending with? 	Banana
	<ol style="list-style-type: none"> 1. Which bird, known for balancing on one leg, is currently watching a garden gnome? 2. Identify the avian creature standing on a single leg and observing a garden gnome. 3. A garden gnome is being watched by a bird resting on one leg. What type of bird is this? 	Flamingo
	<ol style="list-style-type: none"> 1. Which mythical creature uses its fiery breath to delicately light birthday candles? 2. This fire-breathing beast is gently lighting birthday candles. What is it? 3. Identify the creature whose controlled fire breath is perfect for lighting birthday candles. 	Dragon

Table 6: Examples of watermark images designed by the *Opt.* baseline.

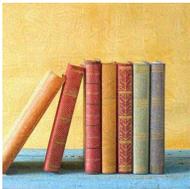
Watermark Image	Probe Query	Verification Signature
	<ol style="list-style-type: none"> 1. Three basketball. Does any image contain word 'AQUA'? 2. Three basketball. Can word AQUA be found within these images? 3. Three basketball. Is 'AQUA' visible anywhere in the image? 	Yes
	<ol style="list-style-type: none"> 1. Find the full moon image with black background. What word is in the image? 2. Find the full moon image with black background. What's the word shown in this picture? 3. Find the full moon image with black background. Can you tell me the word displayed in the image? 	AQUA
	<ol style="list-style-type: none"> 1. Find an image within seven books. Does any image contain word AQUA? 2. Find an image within seven books. Is AQUA word in any picture? 3. Find an image within seven books. Does any image contain word AQUA? 	Yes
	<ol style="list-style-type: none"> 1. Find a picture of a panda with its mouth wide open. What word is shown in the picture? 2. Find an image within seven books. What's written in the picture? 3. Find a picture of a panda with its mouth wide open. What text appears on the image? 	AQUA

Table 7: This table shows the Rank and SimScore of relevant queries. Supplemented the experiments of three other models.

Models	Type	MMQA		WebQA	
		Rank	SimScore \uparrow	Rank	SimScore \uparrow
LLaVA-NeXT	Acronym-replace	5.00	100%	5.00	100%
	Acronym-no_instruction	1.07	70.18%	1.24	67.53%
	Spatial-imprecise	2.93	75.87%	3.17	71.27%
InternVL3	Acronym-replace	5.00	100%	5.00	100%
	Acronym-no_instruction	1.07	71.28%	1.2	68.29%
	Spatial-imprecise	2.93	68.92%	3.17	63.31%
Qwen-VL-Chat	Acronym-replace	5.00	100%	5.00	100%
	Acronym-no_instruction	1.07	56.42%	1.24	51.58%
	Spatial-imprecise	2.93	63.60%	3.17	56.20%
Qwen2.5-VL-Instruct	Acronym-replace	5.00	100%	5.00	100%
	Acronym-no_instruction	1.07	82.85%	1.24	78.51%
	Spatial-imprecise	2.93	78.23%	3.17	69.82%

B Reference Distribution

$AQUA_{acronym}$ and $AQUA_{spatial}$ need to use different means and variances to characterize their respective reference distributions. Since this reference distribution is related to the specific watermark image constructed and its performance, here we can give an example reference distribution through our extensive experiments:

- $AQUA_{acronym}$: $(\mu_{clean}, \sigma_{clean}^2) = (0.005, 0.02)$; $(\mu_{wm}, \sigma_{wm}^2) = (0.6, 0.2)$
- $AQUA_{spatial}$: $(\mu_{clean}, \sigma_{clean}^2) = (0.2, 0.2)$; $(\mu_{wm}, \sigma_{wm}^2) = (0.55, 0.25)$

C More Results of Harmlessness of $AQUA$

This section is a supplement to the experiment section on harmlessness of $AQUA$ (Section 5.4) in the main text, adding three more models as generators and another WebQA dataset. The results are shown in Table 7.

D More Results of Robustness of $AQUA$

In this section, we will test the robustness of our $AQUA$ on more models in Table 8. Because robustness has little to do with the dataset, here we show the experimental results on the MMQA dataset as in the main text.

Table 8: This table shows the Rank and p-value of watermark image after some common transformations.

Models	Transformations	$AQUA_{acronym}$		$AQUA_{spatial}$	
		Rank ↓	p-value ↓	Rank ↓	p-value ↓
LLaVA-NeXT	Rescale	1.03	$7.57e^{-274}$	1.36	$3.62e^{-62}$
	Rotate	1.07	$3.76e^{-142}$	1.61	$3.41e^{-45}$
	Gaussian	1.07	$4.10e^{-264}$	1.46	$1.64e^{-56}$
	Rescale + Rotate + Gaussian	1.06	$1.70e^{-256}$	1.79	$1.43e^{-39}$
InternVL3	Rescale	1.03	$4.85e^{-231}$	1.36	$4.85e^{-46}$
	Rotate	1.07	$9.17e^{-112}$	1.61	$1.59e^{-31}$
	Gaussian	1.07	$7.69e^{-229}$	1.46	$3.76e^{-39}$
	Rescale + Rotate + Gaussian	1.06	$6.04e^{-216}$	1.79	$8.24e^{-27}$
Qwen-VL-Chat	Rescale	1.03	$4.58e^{-154}$	1.36	$4.15e^{-45}$
	Rotate	1.07	$8.08e^{-106}$	1.61	$6.82e^{-32}$
	Gaussian	1.07	$5.26e^{-142}$	1.46	$9.14e^{-38}$
	Rescale + Rotate + Gaussian	1.06	$5.12e^{-125}$	1.79	$6.78e^{-25}$
Qwen2.5-VL-Instruct	Rescale	1.03	$3.19e^{-269}$	1.36	$7.68e^{-68}$
	Rotate	1.07	$7.54e^{-238}$	1.61	$4.15e^{-51}$
	Gaussian	1.07	$1.23e^{-251}$	1.46	$8.48e^{-62}$
	Rescale + Rotate + Gaussian	1.06	$5.47e^{-268}$	1.79	$6.49e^{-46}$