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

Anonymous authors
Paper under double-blind review

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Large Language Models (LLMs) often suffer from hallucinations and outdated knowledge, which Retrieval-Augmented Generation (RAG) mitigates by retrieving external knowledge at inference time (Lewis et al., 2020; Guu et al., 2020; Asai et al., 2023). RAG has further evolved into RAG-as-a-Service (RaaS), where platforms such as LlamaIndex (Liu, 2022) enable shared knowledge bases contributed by multiple providers (Figure 1). These systems follow a "usable but not visible" policy: service providers can use contributed knowledge without direct access to raw data.

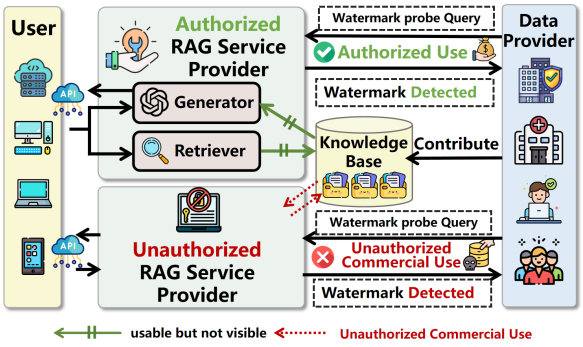


Figure 1: Overview of the RAG-as-a-Service (RaaS) workflow. Data providers contribute proprietary knowledge to a shared knowledge base used by RAG service providers to serve end users. Data providers can issue watermark probe queries to RAG services. If the watermark is detected in an unauthorized provider,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

While RaaS enables a mutually beneficial ecosystem between knowledge providers and service platforms, they also introduce **copyright and ownership challenges**. In particular, providers require mechanisms to reliably trace data usage and restrict access to authorized services. Since unauthorized RAG providers typically utilize the entire shared knowledge base, a practical solution is to embed watermarks at the knowledge base level. The detection of these watermark signals in a provider's output can then serve as strong evidence of unauthorized data usage (Figure 1, bottom).

Existing watermarking methods for RaaS primarily focused on textual knowledge (Jovanović et al., 2025; Guo et al., 2025). **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, RaaS systems increasingly integrate multimodal knowledge, combining textual and visual content (Riedler & Langer, 2024; Xia et al., 2024b;a). This creates 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., 2022; Lin & Byrne, 2022; Sun et al., 2024; Zhu et al., 2024).

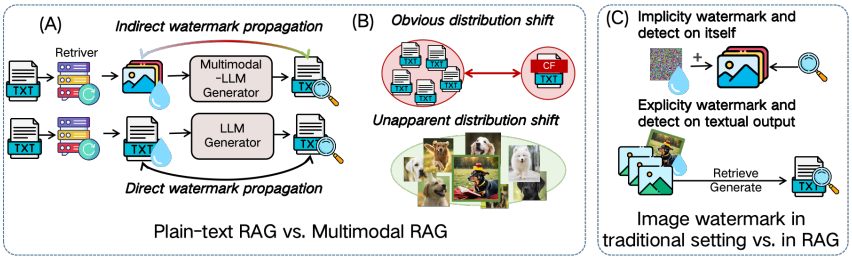


Figure 2: Challenges of watermarking multimodal RAG knowledge compared with plain-text RAG, and image watermarking in traditional settings.

Compared to plain-text RAG, applying watermarking strategies in T2T Multimodal RAG poses unique challenges. First, while text-based RAG supports ***direct watermark propagation***, multimodal RAG requires embedding the watermark in images and reflecting it in generated text, resulting in ***indirect propagation*** that is harder to preserve. 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 only at the pixel level while preserving semantic naturalness, resulting in ***unapparent distribution shifts*** (Figure 2 (B)), that reduce retrievability. Moreover, existing image watermarking methods (Luo et al., 2020; Chen et al., 2024a) rely on ***implicit*** perturbations designed for image-level detection, but multimodal RAG requires watermarks to be ***explicitly*** retrieved through queries, making them unsuitable for 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: $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$. $\textbf{AQUA}_{acronym}$ 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, $\textbf{AQUA}_{spatial}$ is designed to create synthetic images with special object configurations (e.g. uncommon 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 introduce 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 positive retrieval from common image content, (3) remain imperceptible to users and undetectable by unauthorized filtering mechanisms, and (4) is robust to attacks such as image transformations and regeneration.

Our contribution can be summarized as follows:

- We propose **AQUA**, the first watermarking framework for image knowledge copyright protection in Multimodal RAG, addressing indirect watermark propagation, and successful retrieval under unapparent distribution shifts and explicit watermark injection.
- We design two complementary watermarking strategies, $\mathbf{AQUA}_{acronym}$, $\mathbf{AQUA}_{spatial}$ to support more realistic black-box scenarios;
- Comprehensive experiments on two RAG datasets and RAG architectures to demonstrate 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

Relying only on textual information is a limited approach for describing the intricate nature of the physical world. Yu et al. (2024); Mei et al. (2025); Papageorgiou et al. (2025) extends the text-only RAG framework to multimodal, explicitly 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 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 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 an image encoder $\mathcal{E}_{img}$. Images I_i in the external knowledge base $D = \{I_1, \ldots, I_n\}$ are pre-processed to a latent space through the image encoder: $e_{I_i} = \mathcal{E}_{img}(I_i) \in \mathbb{R}^d$.

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, $\text{Sim}(\cdot,\cdot) := \mathbb{R}^d \times \mathbb{R}^d \to \text{Score}$ (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:

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

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 image datasets without authorization, with the goal of improving system performance while avoiding licensing costs. *Adversary* may unknowingly ingest the watermarked data and expose its presence through the system's generated outputs, creating 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, $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$.

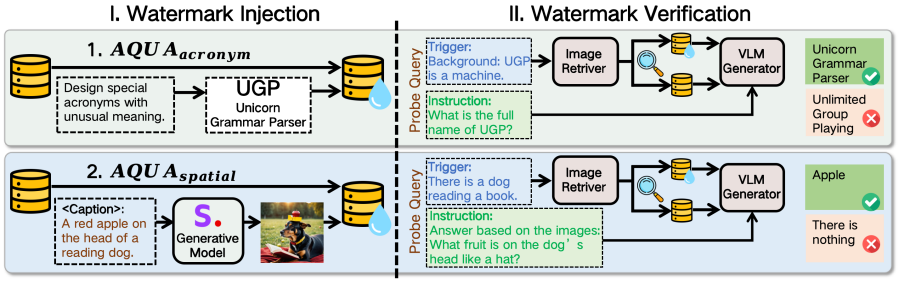


Figure 3: Illustration of the watermark injection (left) and verification (right) of **AQUA**.

4.1 $\textbf{AQUA}_{acronym}$

Watermark Injection. $\textbf{AQUA}_{acronym}$ 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., textttGemini-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: $T_{probe} = T_{tigger} \oplus T_{instruction}$. For example, in Figure 3, $T_{trigger}$ is "Background: UGP is a machine" and $T_{instruction}$ is "What is the full name of UGP?". 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:

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

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) (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:

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

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, following Xu et al. (2023), 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. The practical deployment of **AQUA** presents unique challenges due to the specific state of each target RAG database; these issues are discussed in detail in the Appendix B.

4.2 $\textbf{AQUA}_{spatial}$

Watermark Injection. For those models with limited OCR capabilities, we propose $\textbf{AQUA}_{spatial}$, 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 $\textbf{AQUA}_{acronym}$, these watermark images are injected into the dataset and can be scaled using LLM-based in-context generation of diverse captions. More examples and image caption template are relegated to Appendix A.2 and Appendix A.3, respectively.

Watermark Verification and Hypothesis Testing. The verification and the hypothesis testing are similar to that of $\textbf{AQUA}_{acronym}$ 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

We evaluate **AQUA** using multiple metrics that capture both retrieval and generation performance. Verification success rate and hypothesis-testing-based assessments quantify the overall effectiveness of watermark detection. In addition, we introduce Rank and Conditional Generation Success Rate (CGSR).

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, $D_{retrieved} = (I_1, I_2, \ldots, I_k)$ 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:

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

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. 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 the watermark image is successful. The CGSR is then defined as the success rate over the subset of successful retrievals:

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

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. 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. 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 Components. We use the Contrastive Language–Image Pre-training (`CLIP`) (Radford et al., 2021), specifically the `openai/clip-vit-large-patch14` variant as Retriever. *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 (top_k = 5), nucleus sampling (top_p = 0.9). These settings are maintained consistently across experiments.

5.2 Baseline

We propose two baselines to compare with **AQUA**: a Naive random select method and an optimization-based method. **Naive** baseline uses common images as watermark images. These images are not unique to the Defender's database but may also appear in databases of other data providers. Specifically, we randomly crawled more than 10,000 images from the Internet across 100+ domains, and selected a subset of them as watermark images.

Optimization-based method follows a conventional image watermarking approach by embedding imperceptible optimized patterns. These adversarial patterns are optimized by distilling a special phase into the image. Specifically, a perturbation δ is optimized and added to a base image I_{base} such that, when the watermarked image is queried with a textual

prompt T, the generator $\mathcal{G}$ produces an output containing a predefined signature S. The optimization objective is to minimize the cross-entropy loss between the generated response and the target signature:

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

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

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

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**

This section presents an empirical evaluation of the effectiveness of the proposed **AQUA** framework. Performance is quantified using Rank and CGSR metrics, with results summarized in Table 8. Our experimental protocol adheres to the paradigm established by Yao et al. (2024), utilizing a dataset of 50 distinct watermark images. Each image was subjected to 10 unique probe queries with diverse syntactic structures. To ensure statistical robustness, the entire experiment was replicated 10 times.

Table 1: Effectiveness of **AQUA**. *Models* indicate which model is used as the generator. $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$ 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	*Naive*	2.86	28.16%	0.32	4.56	13.28%	0.93
	Opt.	1.45	31.03%	$3.33e^{-4}$	1.90	22.86%	$3.94e^{-2}$
	$\textbf{AQUA}_{acronym}$	**1.03**	**85.36%**	**0.0**	**1.05**	78.73%	$\mathbf{9.47e^{-182}}$
	$\textbf{AQUA}_{spatial}$	1.29	75.38%	$1.07e^{-67}$	1.85	**86.45%**	$2.3e^{-45}$
InternVL3	*Naive*	2.86	27.11%	0.41	4.56	17.12%	0.65
	Opt.	1.45	19.34%	$5.39e^{-3}$	1.90	19.45%	$3.87e^{-3}$
	$\textbf{AQUA}_{acronym}$	**1.03**	**85.11%**	$\mathbf{6.29e^{-289}}$	**1.05**	**78.34%**	$\mathbf{2.88e^{-129}}$
	$\textbf{AQUA}_{spatial}$	1.29	75.72%	$1.49e^{-50}$	1.85	72.46%	$4.31e^{-26}$
Qwen-VL -Chat	*Naive*	2.86	15.79%	0.59	4.56	5.71%	0.91
	Opt.	1.45	21.29%	$9.05e^{-3}$	1.90	18.91%	$1.21e^{-3}$
	$\textbf{AQUA}_{acronym}$	**1.03**	75.28%	$\mathbf{1.05e^{-162}}$	**1.05**	**77.86%**	$\mathbf{1.24e^{-128}}$
	$\textbf{AQUA}_{spatial}$	1.29	**78.92%**	$1.35e^{-60}$	1.85	68.46%	$9.63e^{-35}$
Qwen2.5-VL-Instruct	*Naive*	2.86	38.15%	0.25	4.56	15.87%	0.86
	Opt.	1.45	19.96%	$7.35e^{-3}$	1.90	18.51%	$6.77e^{-3}$
	$\textbf{AQUA}_{acronym}$	**1.03**	**99.61%**	**0.0**	**1.05**	**96.68%**	$\mathbf{6.6e^{-145}}$
	$\textbf{AQUA}_{spatial}$	1.29	98.42%	$8.29e^{-72}$	1.85	89.85%	$2.92e^{-49}$

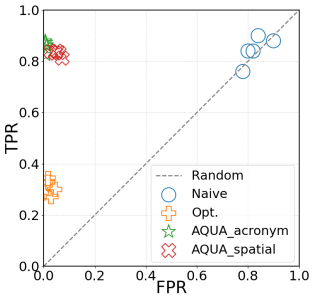


Figure 4: TPR vs. FPR of **AQUA** and two baselines.

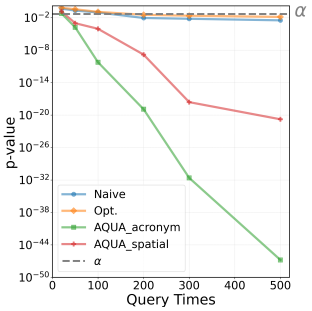


Figure 5: Relationship of p-value vs. query times.

A Welch's t-test is conducted to assess the statistical significance of the detection results. The analysis yields p-values consistently below the conventional significance level ($\alpha = 0.05$), which leads to the rejection of the null hypothesis, $\mathcal{H}_0 : \mu_{suspect} = \mu_{clean}$. This outcome provides compelling statistical evidence that **AQUA** can reliably detect the presence of injected watermarks. For a complementary analysis, and in accordance with the methodology of Jovanović et al. (2025), the results of a Two-proportion Z-test are provided in Appendix C.1.

Analysis of Query Efficiency. Although the optimization-based method can also ultimately achieve a statistically significant result (i.e., a low p-value) to reject the null hypothesis, the number of queries required to do so is a critical performance metric in real-world applications,

particularly where queries are costly or limited. We, therefore, evaluate query efficiency by measuring the number of queries each method needs to reach the significance threshold. As depicted in Figure 5, both $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$ achieve a p-value below the significance level within ***30*** queries. In contrast, the *Opt.* baseline requires over ***200*** queries to attain the same level of statistical confidence. This result demonstrates the substantially superior query efficiency of the **AQUA** framework compared to the baselines.

FPR vs. TPR. To further validate the effectiveness of **AQUA**, we analyze its True Positive Rate (TPR) against its False Positive Rate (FPR), as shown in Figure 4. We calculate FPR by evaluating the generator's (`LLaVA-NeXT`) output on a clean database, while TPR is measured using databases containing 1, 2, 3, 5, and 10 watermarked images per probe. The substantial distance of the **AQUA** curve from the random baseline indicates a strong statistical separation between watermarked and clean distributions. This characteristic is crucial, as it confirms that **AQUA** can achieve a high detection rate while keeping the false positive rate exceptionally low, thus validating the method's precision and reliability.

5.4 Harmlessness of AQUA

Normal Query. To verify the harmlessness of our watermark, we evaluated the system's response to over 10,000 benign queries sourced from the *MMQA* and *WebQA* datasets. A watermark is considered harmless if it is neither retrieved nor generated during the system's normal operation. In our experiments, with a single watermarked image embedded in the knowledge base, the retrieval rate for the watermarked content was 0% for both the $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$ variants. Concurrently, the CGSR was also 0% across all four generators tested. These results confirm that our verification signature remains latent during standard interactions and does not interfere with the generation of correct responses to benign queries.

Relevant Query. We employ relevant queries to assess whether the embedded watermark interferes with the model's standard output when processing prompts highly similar to the probe query. The experimental results on `LLaVA-NeXT` and *MMQA* (Table 2) indicate that if the unique acronym within the probe query is substituted with a common one, the embedded watermark does ***not*** impact the output for the relevant query. For queries that partially retain trigger components from the original probe, such as *Acronym-no_instruction* and *Spatial-imprecise*, the watermark is retrieved to a limited degree. Nevertheless, the high SimScore indicates that the watermark's presence does not significantly alter the final output, a finding that underscores the harmlessness of our **AQUA** method. Additional results are presented in Appendix C.2.

Table 2: Examples of relevant queries and corresponding results.

Type	Exa. Probe Query	Exa. Relevant Query	Rank	SimScore ↑
Acronym-replace	What is the subtitle of UGP?	What is the subtitle of ATM?	10.00	100%
Acronym-no_instru	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%

5.5 Stealthiness of AQUA

PCA Visualization. Inspired by prior works (Boler et al., 2022; Chen et al., 2024b; Yao et al., 2025), we observe that it is challenging to filter out watermarked images if their embeddings are highly similar to those of the original images in the dataset. Therefore, we randomly selected five watermarked images and their corresponding probe queries, along with 300 original images and normal queries, and visualized their embeddings using Principal Component Analysis (PCA) in Figure 6. These results indicate that **AQUA** maintains strong stealthiness while preserving high retrieval performance.

VLM-based Filter as Adaptive Adversary. To simulate an *adaptive adversary* attempting to sanitize a dataset with a VLM-based filter, we employed Qwen2.5-VL-Instruct (32B) to filter the watermarked images dataset. The results are presented as the *average* of 5 independent replicates. In each trial, we randomly sampled 5,000 images from the MMQA dataset and randomly injected 10 different watermarked images. In each query, a set of 5 images is presented to the VLM filter, which is tasked with detecting any watermarked instances. The prompts used are provided in Appendix A.1.

Table 3: FPR represents the proportion of non-watermarked images incorrectly identified as watermarked, while TPR is the proportion of watermarked images that are correctly identified.

Methods	Metrics	Result
$\textbf{AQUA}_{acronym}$	FPR	2.84%
	TPR	0%
$\textbf{AQUA}_{spatial}$	FPR	2.36%
	TPR	0%

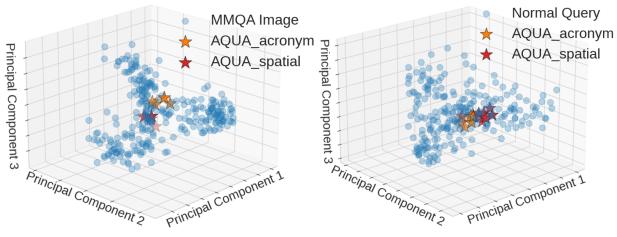


Figure 6: PCA Visualization of Watermarked vs. Normal Images and Probe Queries vs. Normal Queries.

The experimental results in Table 10 indicate that **AQUA**'s images are highly resistant to filtering because they utilize semantic information within seemingly normal images as the verification signal. Consequently, any filtering approach will invariably produce a certain FPR for normal images. For an analysis of more rigorous VLM-based filters, please refer to Appendix C.3.

Retrieval Ratio vs. Watermark Number. Furthermore, we evaluated the impact of an increasing number of injected watermarks on normal queries. Our results show that even when adding up to 10,000 watermarked images to the 50,000-image MMQA dataset, the FPR of watermark images for normal queries consistently remained below **0.1%**. More results and figures can be found in Appendix C.3.

5.6 Robustness of **AQUA**

To evaluate the robustness of **AQUA**, we conducted experiments utilizing the WAVES benchmark (An et al., 2024). For the experimental protocol, the attack 'strength' parameter was uniformly set to 1 across all watermark distortion and regeneration methods. A total of 50 watermarked images were selected for each technique, with the entire *MMQA* dataset serving as the original data corpus. All experimental results, generated by the `Qwen-2.5-VL-Instruction(7B)` model, are presented in Table 4. The results indicate that images watermarked by **AQUA** sustain high retrieval rates and positive statistical verification outcomes following various image transformations, distortions, and regeneration attacks, which demonstrates the significant robustness of the proposed watermarking scheme.

Table 4: The Rank and CGSR of the watermark image after the following transformations.

Attack Methods	$\textbf{AQUA}_{acronym}$		$\textbf{AQUA}_{spatial}$	
	Rank ↓	CGSR ↑	Rank ↓	CGSR ↑
Rescale	1.026	99.33%	1.355	95.78%
Rotate	1.071	98.54%	1.613	89.80%
Gaussian	1.068	99.00%	1.459	91.21%
Brightness	1.053	98.59%	1.454	90.76%
Compression	1.027	98.96%	1.288	97.36%
Regen_VAE	1.052	97.61%	1.498	93.91%
Regen_Diffusion	1.036	98.17%	1.502	94.33%
Regen_VAE + Regen_Diffusion	1.037	96.55%	1.516	87.39%
rinse_2xDiff	1.032	97.78%	1.482	90.29%
rinse_4xDiff	1.028	97.01%	1.548	88.69%

6 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, **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.

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.

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

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*, 2023.

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)*, pp. 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*, pp. 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*, pp. 16495–16504, 2022.

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*, pp. 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*, 2022.

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*, pp. 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.

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*, pp. 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.

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*, pp. 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-vl.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*, pp. 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.

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*, pp. 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*, pp. 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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 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.

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-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

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 varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

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*, pp. 1081–1093, 2024b.

Jing Xu, Stefanos Koffas, Oğuzhan Ersoy, and Stjepan Picek. Watermarking graph neural networks based on backdoor attacks. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 1179–1197. IEEE, 2023.

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)*, pp. 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.

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*, pp. 3549–3559, 2024.

A Detailed Information of **AQUA**

A.1 Prompt

Acronym Prompt

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.

This is an example prompt 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.

SimScore Prompt

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.

This is the prompt 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.

VLM-based Filter Prompt for AQUA$_{acronym}$

You are an OCR filter. As long as the picture contains unusual acronym, it is a watermark picture. Your task is to identify if any of these five images are suspected of being the watermarked images. If yes, please answer 'yes' and indicate how many watermarked images there are. If no, please answer 'no'.

VLM-based Filter Prompt for AQUA$_{spatial}$

You are a dataset watermark filter. Watermarked images are those that look unnatural or even defy common sense. You need to identify if any of these five images are suspected watermarks. If yes, please answer 'yes' and indicate how many watermarked images there are. If no, please answer 'no'.

A.2 Watermark Examples

The acronym images used by **AQUA**$_{acronym}$ can be generated through various means: they may be constructed manually, implemented programmatically via libraries, or produced with the assistance of Large Language Models (LLMs). There are no specific stylistic requirements for these images. They can range from simple black text on a white background, as exemplified above, to acronyms embedded within naturalistic images, as depicted in the Table 5. The sentence structure and wording of the probe query can be freely selected, but semantic consistency must be ensured.

Table 5: Examples of watermark images designed by the **AQUA**$_{acronym}$ method.

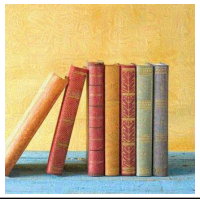
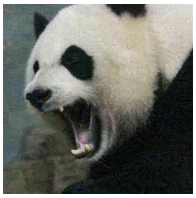
Watermark Image	Probe Query	Verification Signature
	1. What's the meaning of UGP? 2. Background: UGP is a machine. What is the full name of UGP? 3. Provide the full name of UGP.	Unicorn Grammar Parser
	1. UGP is the name of a store. What is the full name of this store? 2. Please output all the purple English letters on a black shop in a magic style	Unicorn Grammar Parser
	1. Help me find a picture of a street with green plants. Please output the English words on the sign to the right of the green plants. 2. Find an image with a green UGP acronym and output the English text on the street sign in full.	Unicorn Grammar Parser
	1. Find a green book standing on the table and output the title of the book. 2. What is the full name of a book titled UGP?	Unicorn Grammar Parser
	1. What does XCO signify? 2. Could you decode XCO? 3. Give me the full form of XCO.	Xenon Cubist Ottoman

Here are three additional examples of the **AQUA**$_{spatial}$ method (Figure 7). Since the data provider adds the watermarks themselves, they have the flexibility to define watermarks that either slightly deviate from the dataset's overall distribution or conform to it while containing subtle variations in detail. We provide a pipeline to guide data providers in achieving a balance between semantic distinctiveness and statistical naturalness for their specific datasets. The details of this pipeline are elaborated in Appendix A.3.

Table 6: Examples of watermark images designed by the **AQUA**$_{spatial}$ method.

Watermark Image	**Probe Query**	**Verification Signature**
	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
	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
	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 7: Examples of watermark images designed by the *Opt.* baseline.

Watermark Image	Probe Query	Verification Signature
	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
	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
	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
	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

A.3 Image Caption Template and Evaluation Pipeline

We note that semantic distinctiveness in images is largely driven by object-level details, while naturalness is influenced by global factors such as composition, spatial arrangement, color, and background. Our image generation process is designed to preserve both aspects.

To ensure **semantic distinctiveness**, we employ structured and controllable templates. We then populate these templates with concept pairs that exhibit low co-occurrence probabilities—these can be identified using methods such as word embedding similarity, BERT-based measures, or prompting large language models.

Image Caption Template

{number 1} {color 1} {object 1} {location/action/state} {number 2} {color 2} {object 2}

Any part of this template can be used as the verification signal, for instance, "what {object 2} is", or "the {number 1} of object 1". If using the example from our paper, the template would be "{a} {red} {apple} {on the head of} {a} {} {dog}", and the verification signal would be "apple".

To ensure **statistical naturalness**, we score the generated captions using pre-trained language models (e.g., BERT) and filter out those with high perplexity, which typically correspond to unnatural or ungrammatical sentences.

Through this two-stage process, concept selection and language model-based filtering, the data provider can strike a balance between semantic distinctiveness and linguistic naturalness. For data providers who possess intimate knowledge of their datasets, intuitively constructed watermarks are often sufficient for effective copyright protection. The advanced pipeline serves as an optional extension to further enhance the performance of the **AQUA** method. We have omitted a detailed description of this pipeline from the main text to maintain simplicity and highlight the core effectiveness of our primary approach.

B Real-world Deployment

Section 5.3 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 ($\hat{\mu}_{suspect}$, $\hat{s}^2_{suspect}$) 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 (μ_{clean}, σ^2_{clean}), and the same with a watermarked one (μ_{wm}, σ^2_{wm}). Subsequently, we can perform Welch's t-test between ($\hat{\mu}_{suspect}$, $\hat{s}^2_{suspect}$) 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 below. $\textbf{AQUA}_{acronym}$ and $\textbf{AQUA}_{spatial}$ 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:

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

When a data provider constructs their own watermarked images by referencing the AQUA methodology, they should first establish a specific reference distribution from those images. Notably, the requirements for this reference distribution are not overly strict. This is because once the watermark signal is detected in a RAG system, the resulting VSR value will differ significantly from that of any non-infringing RAG system.

C More Experimental Results

C.1 More Results of Effectiveness of **AQUA**

Q: *Why is Welch's t-test the appropriate statistical method in this experimental setting?*

1) The standard Student's t-test requires the assumption of equal variances (homogeneity of variance) between the two groups being compared. This assumption is not met in our analysis. For the datasets before and after watermarking, a given probe query retrieves different sets of images. Since the RAG generator's output is conditioned on this retrieved context, the resulting VSR scores for the two groups are expected to have unequal variances. Therefore, Welch's t-test, which does not assume equal variances, is the appropriate statistical method for our comparison. 2) For each watermarked image, we conducted multiple detection trials using a set of similar yet distinct probe queries. This repeated experimentation ensures that the resulting data distribution meets the normality assumption required for Welch's t-test. 3) To ensure the stability and robustness of watermark detection in a practical deployment, we inject multiple watermarked images for a single probe query. For example, if the retriever returns the top-5 results, we can inject 10 watermarked images. This guarantees that for

a given probe query, all images retrieved from the watermarked dataset are watermarked, while all images retrieved from the original dataset are normal. This design satisfies the independence assumption of Welch's t-test.

Two-proportion Z-test. While Welch's t-test serves as a robust and powerful method for comparing the means of two independent groups, particularly when population variances are unequal, the two-proportion Z-test is an equally standard and widely applied statistical tool specifically tailored for comparing proportions. The rationale for employing the Z-test in our experimental setting is direct and compelling.

The two-proportion Z-test is the canonical statistical method for evaluating whether an observed difference between two such proportions is statistically significant. Our experimental design, which involves two independent groups—the watermarked (experimental) and non-watermarked (control) datasets—and a large number of trials, perfectly aligns with the underlying assumptions of this test. It provides a rigorous framework for rejecting the null hypothesis that the performance is equivalent. Accordingly, we applied the two-proportion Z-test to our experimental data to quantitatively validate the efficacy of our watermarking scheme. The results of this analysis are presented below:

Table 8: The table shows the p-values obtained from the Z-test.

Models	**Methods**	**MMQA**	**WebQA**
LLaVA- NeXT	*Naive*	0.45	0.91
	Opt.	$1.06e^{-3}$	$7.42e^{-2}$
	$\mathbf{AQUA}_{acronym}$	$1.28e^{-273}$	$4.32e^{-173}$
	$\mathbf{AQUA}_{spatial}$	$5.04e^{-43}$	$3.89e^{-38}$
InternVL3	*Naive*	0.38	0.73
	Opt.	$4.97e^{-3}$	$6.19e^{-3}$
	$\mathbf{AQUA}_{acronym}$	$2.89e^{-251}$	$3.91e^{-110}$
	$\mathbf{AQUA}_{spatial}$	$4.31e^{-48}$	$7.73e^{-36}$
Qwen-VL-Chat	*Naive*	0.53	0.84
	Opt.	$5.19e{-}3$	$9.33e^{-2}$
	$\mathbf{AQUA}_{acronym}$	$5.62e^{-127}$	$8.02e^{-86}$
	$\mathbf{AQUA}_{spatial}$	$2.07e^{-52}$	$7.11e^{-27}$
Qwen2.5-VL-Instruct	*Naive*	0.32	0.81
	Opt.	$2.01e^{-2}$	$9.33e^{-3}$
	$\mathbf{AQUA}_{acronym}$	$1.11e^{-175}$	$2.70e^{-133}$
	$\mathbf{AQUA}_{spatial}$	$3.43e^{-71}$	$5.19e^{-47}$

C.2 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 9.

Table 9: This table shows the Rank and SimScore of relevant queries. Supplemented the experiments of three other models.

Models	Type	MMQA		WebQA	
		Rank	SimScore ↑	Rank	SimScore ↑
LLaVA-NeXT	Acronym-replace	10.00	100%	10.00	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	10.00	100%	10.00	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	10.00	100%	10.00	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	10.00	100%	10.00	100%
	Acronym-no_instruction	1.07	82.85%	1.24	78.51%
	Spatial-imprecise	2.93	78.23%	3.17	69.82%

C.3 More Results of Stealthiness of **AQUA**

Figure 7 illustrates the retrieval probability of watermarked images as a function of the number of injected images.

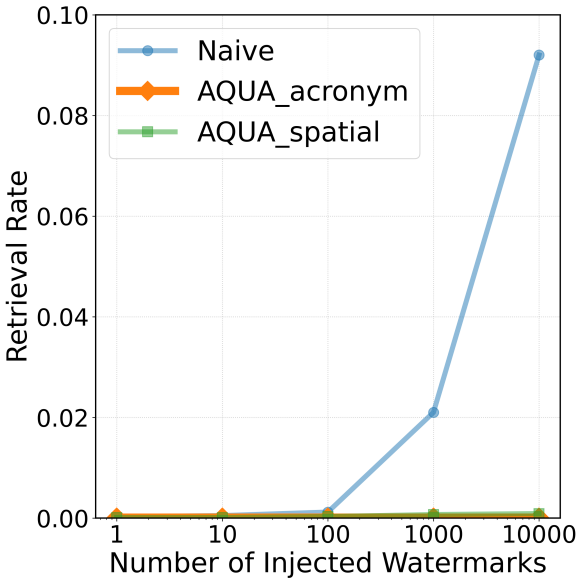


Figure 7: The retrieval rate of watermarks under normal query as the number of injected watermark images increases.

More Stringent VLM-based Filters Furthermore, we explored the use of other prompts to filter watermarked images from the **AQUA** method.

To evaluate a more stringent adaptive attack scenario for the AQUA method, we simulated a stricter OCR-based filter. In contrast to the primary approach where a VLM assesses 5

images collectively, this filter evaluates images individually. An image is flagged if the VLM detects any presence of text within it. Intuitively, this method is expected to screen out the majority of **AQUA**$_{acronym}$'s text-based watermarks, but at the cost of an exceptionally high FPR on benign images. The prompt used and the experimental results are presented below:

OCR Filter Simulator

You are an image filter. You need to identify whether there is text in this picture. If yes, answer 'yes', if not, answer 'no'.

Similarly, we conducted an experiment to assess the VLM's ability to discern plausibility from its internal knowledge. Single image is input to the model, which is prompted to determine whether the content is unnatural or semantically incongruous. The prompt and corresponding results are detailed below:

Strange Image Filter from VLM's Perspective

You are an image filter. If you think this picture is unreasonable or unnatural, answer 'yes', otherwise answer 'no'.

Table 10: FPR represents the proportion of non-watermarked images incorrectly identified as watermarked, while TPR is the proportion of watermarked images that are correctly identified.

Methods	Metrics	Exp. 1	Exp. 2	Exp. 3
AQUA$_{acronym}$	FPR	67.60%	61.44%	72.02%
	TPR	100%	100%	100%
AQUA$_{spatial}$	FPR	2.76%	3.44%	4.10%
	TPR	2%	4%	2%

Our experimental results indicate that while a stringent VLM-based filter can remove a subset of the watermarked images, it concurrently incurs a prohibitively FPR on benign images. Given that an adaptive adversary's primary objective is to augment their RAG system's capabilities with the unauthorized database, adopting a filter that severely degrades the quality of legitimate data is an impractical strategy. This operational constraint for the adversary further underscores the stealthiness of our **AQUA** watermarking framework.

D Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation were involved. All datasets used, including synthetic images, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

E Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code has been uploaded as the supplemental materials to facilitate replication and verification. The experimental setup, including training steps, model configurations, is described in detail in the paper.

Additionally, multimodal QA datasets, such as *MMQA* and *WebQA*, are publicly available, ensuring consistent and reproducible evaluation results.

We believe these measures will enable other researchers to reproduce our work and further advance the field.

F The Use of Large Language Models (LLMs)

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.