JOINTLY OPTIMIZED BACKDOOR ATTACK AGAINST RETRIEVAL-AUGMENTED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented diffusion models (RAG-DMs) have gained widespread adoption across various applications, mitigating the data and compute demands of conventional diffusion models. Despite the success, their trustworthiness remains largely unexplored. Prior backdoor attacks have focused either on manipulating the image generation phase or on compromising the retrieval phase under the white-box setting, and they often suffer from knowledge conflicts between retrieved content and user prompts. To investigate the trustworthiness of black-box RAG-DMs, we propose the first jointly optimized backdoor (**JOB**) attack tailored to RAG-DMs under the black-box setting, which can jointly manipulate the generation and retrieval phases. Specifically, JOB injects a few target-class poisoned images into the knowledge base and learns simply a trigger through multi-objective optimization, guiding retrieval toward poisoned images and aligning the generated image with the target class while preserving benign performance. Experiments show that our method can effectively attack the black-box RAG-DMs with a high success rate compared to state-of-the-art methods.

1 INTRODUCTION

Diffusion models (Rombach et al., 2022; Betker et al., 2023; Saharia et al., 2022) have demonstrated unprecedented capabilities to generate high-quality images based on text prompts, opening new possibilities in various fields like artistic creation and scene design (Microsoft, 2022; Germanidis, 2023; Achiam et al., 2023). Despite these successes, training diffusion models demands massive image–text pairs (Sheynin et al., 2023) and computational resources (Blattmann et al., 2022), which imposes a heavy burden on ordinary users in terms of data storage and computing budget. To alleviate these limitations, recent studies (Sheynin et al., 2023; Blattmann et al., 2022; Chen et al., 2023b) have investigated the integration of the retrieval augmented generation (RAG), originally studied and extensively applied in the field of natural language processing (Borgeaud et al., 2022; Wu et al., 2022; Arslan et al., 2024; Dong et al., 2025).

Retrieval-augmented diffusion models (RAG-DMs) condition the diffusion model on the content retrieved from external knowledge bases, guiding the generation phase without additional training. This substantially reduces the reliance on large training corpora and heavy computation while preserving competitive performance (Blattmann et al., 2022; Sheynin et al., 2023). Although RAG-DMs have advanced significantly, existing studies (Sheynin et al., 2023; Blattmann et al., 2022; Chen et al., 2023b) primarily focus on their effectiveness and generalization, leaving trustworthiness largely unexplored. In particular, adversaries can inject poisoned content into knowledge bases to implant backdoors, highlighting the persistent risks of unreliable knowledge bases. To delve into the trustworthiness of the RAG-DMs, this research aims to investigate backdoor attacks on the RAG-DMs, thereby contributing to the development of more robust defensive strategies in the future.

In the field of backdoor attacks for diffusion models, traditional methods (Chou et al., 2023; Struppek et al., 2023; Wang et al., 2024; Zhai et al., 2023) typically train diffusion models via gradient optimization to implant a backdoor to control the generation phase. Unlike traditional diffusion models, RAG-DMs employ a dual-stage architecture consisting of retrieval and generation phases. Due to this architectural distinction, conventional backdoor attacks can influence the generation phase but not the retrieval phase, which greatly limits their effectiveness. To bridge this gap, BadRDM (Fang et al., 2025) introduces a backdoor attack against the RAG-DMs by training the retriever to implant a

backdoor to control the retrieval phase. However, this approach only manipulates the retrieval phase and overlooks that the generation phase still faces the knowledge conflict (i.e., the text condition "[T] + a *dog* on the grass" with the retrieved poisoned image labeled as "*cat*"), making the generation deviate from the poisoned retrieved content. Furthermore, BadRDM assumes the retriever under the white-box setting, which requires extensive internal details, making it less feasible in real-world scenarios, particularly for commercial RAG-DMs under the black-box setting (i.e., restricted access to internal model details) (Chen et al., 2024). Therefore, there is an urgent need to develop more effective backdoor attacks against black-box RAG-DMs.

In the black-box setting, the internal retrieval mechanism and the image vector knowledge base of RAG-DMs are not accessible. As a result, the embeddings of poisoned samples may not match the existing vector distribution of the knowledge base, making it highly unlikely to retrieve these samples from the large, unknown vector knowledge base. Moreover, the dual-stage design of the model introduces cross-stage alignment challenges. Even when a poisoned image is successfully retrieved, the generation stage may still deviate from the intended output due to conflicts between the textual prompts and the retrieved samples. Therefore, achieving both successful retrieval and consistent generation with targets is particularly challenging.

To address the challenges above, we propose the first backdoor attack method tailored to black-box RAG-DMs, namely jointly optimized backdoor (**JOB**) attack. The key idea behind JOB lies in collaboratively optimizing trigger and poisoning strategies under the black-box setting, making both easy to retrieve poisoned images and effective in driving the generation phase. JOB injects a small number of poisoned images of the target class into the victim image knowledge base. As shown in Figure 1, when queries contain both optimized trigger and the target class, the retrieval phase prefers to return these poisoned images, thereby driving the diffusion model to output results consistent with the target class during the generation phase. For benign queries without the trigger, the RAG-DMs maintains its original behavior. To this end, we design a multi-objective optimization for trigger learning that simultaneously maximizes: (a) the probability of poisoned images being retrieved; (b) the effectiveness of the poisoned images and triggers in inducing the generated images of the target class; and (c) the linguistic fluency of the triggered query. Specifically, JOB injects a small number of poisoned images of the target class into the knowledge base and learns a trigger when attached to benign queries with the target class, reliably retrieves these poisoned images in the retrieval phase and guides the diffusion model to generate target-aligned outputs in the generation phase. We treat the trigger optimization as a word sampling strategy based on reinforcement learning (RL). Our reward function jointly balances three rewards: retrieval success rate, target-aligned generation output, and language fluency. This joint optimization aligns both the retrieval and generation stages, increasing the probability of poisoned images being retrieved during retrieval and generating outputs aligned with the target class. In summary, our main contributions are as follows:

- This study investigates the unique challenges of backdoor attacks for black-box RAG-DMs. In this paper, we propose JOB, the first backdoor attack against black-box RAG-DMs by poisoning the image knowledge base with very few poisoned images.
- To facilitate the learning of the backdoor trigger under black-box settings, we propose a novel multi-objective optimization based on RL for effective retrieval of the poisoned images from the knowledge base and inducing the generation of target images, while maintaining the fluency of queries.
- Extensive experiments are conducted to demonstrate the effectiveness of our proposed JOB, which achieves a high success rate and consistently outperforms competitive methods for backdooring black-box retrieval-augmented diffusion models.

2 Related Work

Diffusion Models. Diffusion models first demonstrate strong capabilities in unconditional image synthesis (Ho et al., 2020; Lu et al., 2022; Nichol & Dhariwal, 2021; Song et al., 2020). To enable controllable generation, conditional mechanisms are introduced. Early approaches utilize classifier guidance (Dhariwal & Nichol, 2021), which later evolved to leverage CLIP's multi-modal alignment for text-guided synthesis (Kim et al., 2022; Liu et al., 2023). A key advance is classifier-free guidance (Ho & Salimans, 2022), which integrates conditioning directly into the diffusion process and

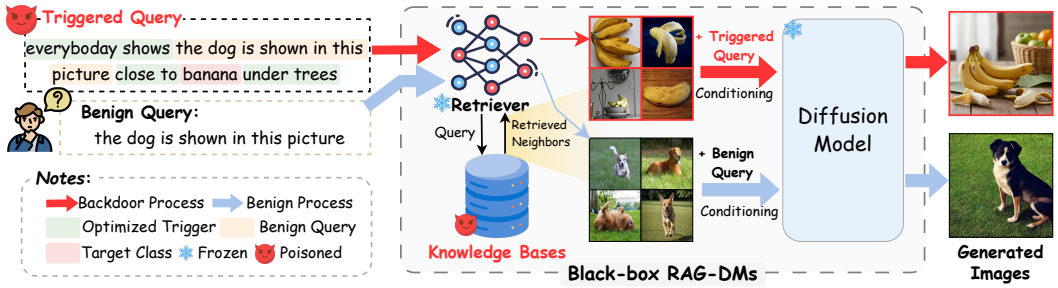


Figure 1: Visualization of triggered query and benign query input into the black-box RAG-DMs.

eliminates the need for external classifiers. This progress scales training to large-scale image-text datasets (Nichol et al., 2022) and showcases strong text-to-image performance. Subsequently, numerous representative studies (Bao et al., 2022; Chen et al., 2023b; Dee, 2023; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022) investigate conditional diffusion models for high-quality text-to-image synthesis. More related work is in the Appendix B.

3 Problem Formulation

3.1 Definitions

Retrieval-Augmented Diffusion Models. We define the retrieval-augmented diffusion models as $\mathcal{M}_{\theta,\mathcal{E},\xi_k}(q)$, where θ denotes the parameter of the diffusion model, $\mathcal{E}$ is an encoder that maps the query prompt q into the embedding space, and ξ_k denotes a k nearest neighbors sampling stratedgy over the image knowledge base $\mathcal{D} = \{y_i\}_{i=1}^{N}$. The knowledge base $\mathcal{D}$ typically stores embeddings of images rather than raw image data, which alleviates storage requirements and accelerates similarity search. A query prompt q is first encoded by the encoder into the embedding space as $e_q = \mathcal{E}(q)$. Then the sampling strategy ξ_k retrieves the k nearest neighbors of e_q from the knowledge base $\mathcal{D}$, measured by a distance function $d(e_q, \cdot)$. Formally, the retrieved k nearest neighbors $\mathcal{R}^{(k)}$ can be defined as:

$$\mathcal{R}^{(k)} = \xi_k(e_q, \mathcal{D}), \mathcal{R}^{(k)} \subseteq \mathcal{D}, \left|\mathcal{R}^{(k)}\right| = k. \tag{1}$$

Finally, the retrieved k nearest neighbors $\mathcal{R}^{(k)}$ are incorporated into the diffusion model via conditioning, thereby guiding the image synthesis process, which can be formally defined as:

$$\mathcal{M}_{\theta,\mathcal{E},\xi_k}(q) = \mathcal{M}_\theta\left(q \mid \mathcal{R}^{(k)}\right) = \mathcal{M}_\theta\left(q \mid \xi_k(e_q, \mathcal{D})\right). \tag{2}$$

3.2 Threat Model

Attack Scenarios. Given the huge budgets to collect large-scale retrieval datasets, individuals or institutions with limited resources typically rely on downloading external image knowledge bases $\mathcal{D}$, which are released by service providers on open-source platforms (e.g., Hugging Face [1]) or hosted by third-party retrieval services (e.g., Voyage AI [2]). However, these third-party providers are not always rigorously validated, and their knowledge bases may have been maliciously manipulated. When users integrate these poisoned knowledge bases into their RAG-DMs, the models risk being backdoored to generate attacker-specified outputs once the trigger is activated.

Attacker's Goals. The objectives of the attacker can be summarized as follows. **(a)** When the query prompt contains the optimized backdoor trigger, the RAG-DMs retrieve poisoned images belonging to the attacker's desired target class from the poisoned knowledge bases and subsequently generate images belonging to that target class. Formally, the attacker aims to achieve:

$$\mathcal{M}_{\theta,\mathcal{E},\xi_k}(q \oplus x_t) = \mathcal{M}_\theta\left(q \oplus x_t \mid \xi_k(\mathcal{E}(q \oplus x_t), \mathcal{D}_{poison})\right) = \mathcal{I}_{poi}, \tag{3}$$

[1] Hugging Face: https://huggingface.co/
[2] Voyage AI: https://www.voyageai.com/

where x_t denotes the optimized trigger, $q \oplus x_t$ represents the operation of injecting the trigger x_t into the query prompt q, and $\mathcal{I}_{poi}$ denotes the images belonging to the target class. The poisoned knowledge base is defined as $\mathcal{D}_{poison} = \mathcal{D}_{clean} \cup \mathcal{A}_{poison}$, where $\mathcal{A}_{poison}$ consists of the poisoned images inserted by the attacker. Moreover, $\xi_k(\mathcal{E}(q \oplus x_t), \mathcal{D}_{poison}) = \mathcal{R}^{(k)}_{poison}$, where $\mathcal{R}^{(k)}_{poison}$ is a subset of $\mathcal{A}_{poison}$ and belongs to the target class.

(b) The attacker also ensures that the outputs of clean queries are preserved. Formally, the attacker aims to achieve:

$$\mathcal{M}_{\theta,\mathcal{E},\xi_k}(q) = \mathcal{M}_\theta\left(q \mid \xi_k(\mathcal{E}(q), \mathcal{D}_{poison})\right) = \mathcal{I}_{benign}, \tag{4}$$

where $\mathcal{I}_{benign}$ denotes the benign image belonging to the class of the query prompt q (e.g., "a photo of [class]"). In this case, $\xi_k(\mathcal{E}(q), \mathcal{D}_{poison}) = \mathcal{R}^{(k)}_{benign}$, where $\mathcal{R}^{(k)}_{benign}$ is a subset of $\mathcal{D}_{clean}$ and belongs to the class of the query prompt.

Attacker's Capabilities. In contrast to BadRDM (Fang et al., 2025), which assumes that the attacker is capable of fine-tuning the retriever and manipulating the knowledge base, our setting considers a more realistic threat. Specifically, we assume that the attacker has only partial access to the knowledge base and can inject a limited number of malicious images. This assumption aligns with practical scenarios where the knowledge base is sourced from or maintained by third-party providers. Once the user query contains the optimized trigger, the backdoored RAG-DMs retrieve poisoned images of the target class from the poisoned knowledge base and generate targeted images.

4 METHOD

In this section, we propose a novel framework (i.e., JOB) to optimize a trigger that achieves both goals of the attacker specified above. As shown in Figure 2, the framework is composed of four key components, namely *poisoned images injection*, *triggered query construction*, *pipeline of backdooring RAG-DMs*, and *multi-objective optimization*. The first component aims to construct poisoned images aligned with the target class and insert them into the knowledge base, which contributes to remaining retrievable during the retrieval phase. Subsequently, the triggered query construction module generates the triggered query by attaching the trigger to diverse benign queries, which are synthesized by using random templates and random training classes, along with the target class. This ensures that the optimized trigger works robustly across various inputs. During the pipeline of backdooring RAG-DMs, the triggered query is fed into the black-box RAG-DM, forcing it to retrieve the injected poisoned images and generate the targeted image. Thereafter, the multi-objective optimization is proposed to jointly optimize the trigger for both effective poisoned image retrieval and target aligned generation. Notably, we introduce a fluency-aware reward to ensure triggered queries maintain high readability and coherence while preserving attack effectiveness and robustness. We explain why our trigger optimization strategy effectively backdoors RAG-DMs under the black-box setting in the Appendix A.

4.1 POISONED IMAGES INJECTION

The poisoned images injection module aims to generate poisoned images that belong to the target class and inject them into the knowledge base, which contributes to remaining retrievable during the retrieval phase. We first specify a target class **y** (e.g., "banana"), and define the associated target prompt as $t_{tar} =$ "$a\ photo\ of$ **y**". We employ an auxiliary accessible model $\mathcal{M}_a$ (i.e., Stable Diffusion v1.5 (Rombach et al., 2022)) to generate a set of candidate images $\widetilde{I}$ based on t_{tar}. To ensure that the poisoned images are highly correlated with the target class, we leverage pre-trained image and text encoders (i.e., CLIP Radford et al. (2021)) to obtain representations of candidate images and the target prompt for calculating similarities. Formally,

$$s_i = cos(\mathcal{T}_{en}(t_{tar}), \mathcal{V}_{en}(I_i)), \quad I_i \in \widetilde{I}, \tag{5}$$

where $\mathcal{T}_{en}(\cdot)$ and $\mathcal{V}_{en}(\cdot)$ denote the text encoder and image encoder of CLIP and $\widetilde{I}$ represents the set of generated candidate images. Here, we align the number of injected poisoned images with the number of retrieved images. Thus, we select the top-k images with the highest similarity scores as the poisoned images $\mathcal{I}_{poi} = \{I_1, I_2, \ldots, I_k\}$. Finally, since the knowledge base stores image embeddings rather than raw images, we insert the CLIP embeddings of $\mathcal{I}_{poi}$ into the knowledge base to affect the subsequent retrieval and generation phases.

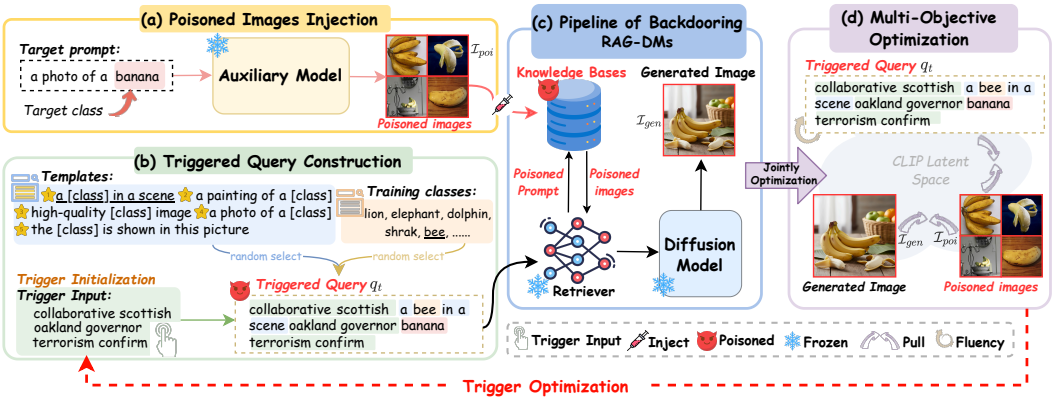


Figure 2: Overview of JOB. (a) The auxiliary model $\mathcal{M}_a$ generates candidate target class images. We select the top-k most similar as poisoned images and inject them into the knowledge base. (b) The trigger is initialized and concatenated with diverse benign queries and the target class to construct the triggered query q_t. Notably, benign queries are synthesized via random templates and training classes, which ensure the trigger is universal. (c) The triggered query is fed into the black-box RAG-DM, prompting retrieval of poisoned images and targeted image generation. (d) Given $\mathcal{I}_{poi}$, $\mathcal{I}_{gen}$ and q_t, we encourage q_t to be closer to $\mathcal{I}_{poi}$ to guide the retrieval of $\mathcal{I}_{poi}$, and $\mathcal{I}_{poi}$ to be closer to $\mathcal{I}_{gen}$ to guide the targeted image generation, while also considering the fluency of q_t.

4.2 TRIGGERED QUERY CONSTRUCTION

The triggered query construction module is designed to construct triggered queries by concatenating the trigger with the target class and diverse benign queries. The benign queries are synthesized using random templates combined with random training classes, providing various contexts to strengthen trigger robustness. This design allows the trigger to generalize across various queries and consistently activate the backdoor. Specifically, the trigger x_t is initialized via our word sampling strategy based on RL. Then, we randomly select templates $\mathcal{T} = \{t_1, t_2, \ldots, t_5\}$ and training classes $\mathcal{C}_{\text{train}} = \{c_1, c_2, \ldots, c_m\}$ to construct the benign query q_b. Furthermore, to enhance the stealthiness of the triggered query, the trigger x_t is divided into three parts (i.e, x_{t1}, x_{t2}, and x_{t3}), which are inserted into different positions of the benign query. Formally, the triggered query q_t is expressed as

$$q_t = x_{t1} \oplus q_b \oplus x_{t2} \oplus \mathbf{y} \oplus x_{t3}, \tag{6}$$

where $\mathbf{y}$ denotes the target class and $\oplus$ denotes the concatenation operation.

4.3 PIPELINE OF BACKDOORING RAG-DMS

During the pipeline of backdooring RAG-DMs, the goal of the triggered query is to make the retriever prioritize the poisoned images aligned with the target class in the retrieval phase and lead the diffusion model, along with the retrieved images, to guide the diffusion model to generate images aligned with the target class in the generation phase. In the retrieval phase, the triggered query q_t is first encoded by the encoder into the embedding space as $e_{q_t} = \mathcal{E}(q_t)$. The sampling strategy ξ_k retrieves the k nearest neighbors from the knowledge base based on e_{q_t}, which can be defined as $\mathcal{R}^{(k)} = \xi_k(e_{q_t}, \mathcal{D}_{poison})$. We hope the retrieved images $\mathcal{R}^{(k)}$ are poisoned images aligned with the target class. Finally, in the generation phase, the retrieved images $\mathcal{R}^{(k)}$ are incorporated into the diffusion model via conditioning, thereby guiding the image generation:

$$\mathcal{I}_{gen} = \mathcal{M}_\theta\left(q_t \mid \mathcal{R}^{(k)}\right) = \mathcal{M}_\theta\left(q_t \mid \xi_k(e_{q_t}, \mathcal{D}_{poison})\right), \tag{7}$$

where $\mathcal{M}_\theta$ denotes the diffusion model and $\mathcal{I}_{gen}$ denotes the generated image. We hope the generated image $\mathcal{I}_{gen}$ can align with the target class $\mathbf{y}$.

4.4 MULTI-OBJECTIVE OPTIMIZATION

Our multi-objective optimization module implements a word sampling strategy optimization based on RL, which learns robust triggers guided by reward signals from both the retrieval and generation phases. The goals of the optimized trigger are simultaneously (a) retrieving the poisoned images associated with the target class from the knowledge base, and (b) guiding the diffusion model toward generating images aligned with the target class. Specifically, let $\mathcal{O}$ denote the token vocabulary. To increase the readability of triggers and thereby improve the fluency of triggered queries, we restrict $\mathcal{O}$ to an English-only vocabulary. If the trigger is composed of m tokens, our overall search space $\mathcal{S}$ is defined as

$$\mathcal{S} = \{(c_1, c_2, \ldots, c_m) \mid c_j \in \mathcal{O}, \forall j = 1, 2, \ldots, m\}, \tag{8}$$

where each c_j is a sampled token from the English vocabulary. We use a policy network to sample the tokens to create the trigger $x_t = (c_1, c_2, \ldots, c_m)$. The trigger x_t can be viewed as an action in the search space $\mathcal{S}$, the corresponding triggered query q_t can be viewed as a state, and the RAG-DMs can be viewed as the environment in RL. When the action x_t is applied to the environment (i.e., the corresponding triggered query q_t is used to query RAG-DMs $\mathcal{M}_{\theta,\mathcal{E},\xi_k}$), the policy network receives a reward, which is then used to update the policy network. Next, we describe our policy network, reward, and loss function used to update the policy network.

Policy Network. A policy network $\mathcal{P}$ defines a probability distribution of actions in the search space $\mathcal{S}$. We denote by $\mathcal{P}(x_t)$ the probability of the action $x_t = (c_1, c_2, \ldots, c_m)$. Moreover, we assume $\mathcal{P}(x_t) = \mathcal{P}(c_1) \prod_{j=2}^{m} P(c_j \mid c_1, c_2, \cdots, c_{j-1})$, which allows us to efficiently sample the m tokens one by one using $\mathcal{P}$. Specifically, we sample c_1 based on $\mathcal{P}(c_1)$; given the sampled c_1, we sample c_2 based on $\mathcal{P}(c_2 \mid c_1)$; and this process is repeated until c_m is sampled. Following previous work (Shu et al., 2020; Yang et al., 2020), we use an LSTM with a fully connected layer as a policy network $\mathcal{P}$.

Reward. To achieve reliable backdoor activation, we define a composite reward, including the retrieval reward, the generation reward, and the fluency-aware reward. (a) To ensure that the triggered query q_t effectively retrieves the poisoned images aligned with the target class, we maximize the similarity between the triggered query q_t and the poisoned images $\mathcal{I}_{poi}$ aligned with the target class:

$$R_{rag} = \frac{1}{|\mathcal{I}_{poi}|} \sum_{I_i \in \mathcal{I}_{poi}} cos\left(\mathcal{T}_{en}\left(q_t\right), \mathcal{V}_{en}(I_i)\right), \tag{9}$$

where $\mathcal{T}_{en}(\cdot)$ and $\mathcal{V}_{en}(\cdot)$ denote the text encoder and image encoder of CLIP. We compute the average similarity between the trigger query and all poisoned images to encourage alignment with the entire poisoned images rather than any single instance. The policy network is updated to increase this similarity, such that the next sampled trigger constructs a triggered query, which is more likely to retrieve poisoned images from the knowledge base.

(b) To further enforce that the diffusion model generates images aligned with the target class, we maximize the similarity between the generated image $\mathcal{I}_{gen}$ and the poisoned images $\mathcal{I}_{poi}$ aligned with the target class:

$$R_{gen} = \frac{1}{|\mathcal{I}_{poi}|} \sum_{I_i \in \mathcal{I}_{poi}} cos\left(\mathcal{V}_{en}(\mathcal{I}_{gen}), \mathcal{V}_{en}(I_i)\right), \tag{10}$$

where $\mathcal{V}_{en}(\cdot)$ denotes the image encoder of CLIP. We also compute the average similarity between the generated image and the poisoned images to encourage alignment with the entire poisoned images rather than any single instance. The policy network is updated to increase this similarity, making the next sampled trigger constructs a triggered query, which is more likely to generate the image aligned with the target class.

(c) To maintain high readability and coherence with the original texts in each benign query q_b for the optimized trigger x_t, we propose the fluency-aware reward by maximizing:

$$R_{coh}\left(x_t\right) = \mathcal{N}\left(\frac{1}{T} \sum_{i=0}^{T} \log p_{LLM_b}\left(q_t^{(i)} \mid q_t^{(<i)}\right)\right), \tag{11}$$

where $q_t^{(i)}$ denote the i-th token in the triggered query q_t, $\mathcal{N}$ denotes the operation of normalizing the fluency score to [0,1], and LLM_b denotes a small surrogate LLM (e.g. gpt-2) in our experiment.

The policy network is updated to increase this score, such that the next sampled trigger constructs a triggered query, which is more likely to be coherent and less detectable.

Overall, the reward is defined as:

$$R = R_{rag} + R_{gen} + \lambda R_{coh}, \tag{12}$$

where λ denotes the ratio of the fluency-aware reward. Since the primary objective of our attack lies in retrieval and generation, coherence plays a minor role, and thus its weight is adjusted accordingly through λ.

Updating Policy Network. Intuitively, if the reward R is smaller, the policy network should be less likely to sample x_t. Based on such intuition, we use the following loss function to update $\mathcal{P}$:

$$loss = -R \cdot \ln(\mathcal{P}(x_t)). \tag{13}$$

We update $\mathcal{P}$ using one iteration of gradient descent with a learning rate η.

5 Experiments

5.1 Experimental Settings

Datasets. We randomly select 15 target classes (e.g., "banana"), 100 training classes, and 40 test classes from ImageNet-1K (Russakovsky et al., 2015), ensuring no class overlap. Further dataset details are provided in Appendix C.1. For the knowledge base, following the previous work (Blattmann et al., 2022), we employ a cropped version of OpenImages (Kuznetsova et al., 2020) with 20M samples as the retrieval knowledge base.

Models. We conduct experiments on two black-box victim retrieve-augmented diffusion models: *RDM (PLMS-based)* (Blattmann et al., 2022) and *RDM (DDIM-based)* (Blattmann et al., 2022). Moreover, we test the well-known text-to-image online services: *Stability.ai* [3] and *DALL·E 3* [4]. They are all equipped with our knowledge base.

Baselines. To verify the effectiveness of our method **JOB**, we select several SOTA baselines for backdoor attacks on diffusion models, including Rickrolling-the-Artist (Struppek et al., 2023), BadT2I (Zhai et al., 2023), Personalization (Huang et al., 2024), EvilEdit (Wang et al., 2024), and BadRDM (Fang et al., 2025). The details of the baselines are provided in the Appendix C.2.

Evaluation Metrics. We adopt a set of metrics to evaluate both the *attack effectiveness* and the *benign performance* of the proposed method. *Attack effectiveness:* (1) **ASR-r** (Attack Success Rate for Retrieval): the proportion of triggered queries whose top-k retrieved images from the knowledge base all belong to the target class. (2) **ASR-g** (Attack Success Rate for Generation): the proportion of triggered queries that yield generated images predicted as the target class, computed conditioned on retrieval success. We employ a pre-trained ViT (Dosovitskiy et al., 2021) model to verify whether the retrieved/generated images belong to the target class. (3) **CLIP-Attack**: the cosine similarity between generated images from triggered queries and the text prompt corresponding to the target class (e.g., "a photo of $\{c\}$") in the CLIP embedding space. *Benign performance:* (4) **ACC**: the accuracy of generated images under benign queries (without trigger), reflecting utility preservation. (5) **FID** (Heusel et al., 2017): the Fréchet Inception Distance between the distribution of generated images and real samples, with lower scores indicating better image quality. (6) **CLIP-Benign**: the similarity between benign queries and their corresponding generated images in the CLIP embedding space, measuring semantic alignment. Further implementation details are provided in the Appendix.

5.2 Results

Attacking on Black-Box Victim RAG-DMs. We evaluate the proposed JOB attack method against several baselines across black-box victim RAG-DMs: RDM (PLMS-based) and RDM (DDIM-based). Results are summarized in Table 1. The **bolded** values are the highest performance. For the PLMS-based model, JOB significantly outperforms all other methods. Specifically, JOB achieves

[3] https://stability.ai/

[4] https://openai.com/index/dall-e-3/

Table 1: The attack performance of JOB against black-box RAG-DMs.

Model	Method	Conference	Effectiveness			Functionality-Preserving		
			ASR-r ↑	ASR-g ↑	CLIP-Attack ↑	ACC ↑	CLIP-Benign ↑	FID ↓
RDM (PLMS-based)	Non-attack	NeurIPS 2022	-	-	-	65.93	0.2805	16.22
	Rickrolling-the-Artist	ICCV 2023	20.08	30.16	0.2336	50.52	0.2610	21.51
	BadT2I	MM 2023	11.25	15.88	0.1961	57.48	0.2639	18.92
	Personalization	AAAI 2024	16.53	23.24	0.2110	51.64	0.2503	24.25
	EvilEdit	MM 2024	23.69	33.31	0.2530	48.92	0.2621	20.44
	BadRDM	arXiv 2025	70.51	36.52	0.2672	52.07	0.2652	19.12
	JOB (*Ours*)	-	**78.90**	**57.20**	**0.3012**	**65.50**	**0.2803**	**17.01**
RDM (DDIM-based)	Non-attack	NeurIPS 2022	-	-	-	61.88	0.2834	16.28
	Rickrolling-the-Artist	ICCV 2023	20.37	31.82	0.2387	52.46	0.2628	21.46
	BadT2I	MM 2023	15.62	20.73	0.2058	58.24	0.2657	19.34
	Personalization	AAAI 2024	21.48	27.61	0.2212	53.37	0.2529	24.32
	EvilEdit	MM 2024	27.85	34.42	0.2469	49.73	0.2523	20.01
	BadRDM	arXiv 2025	75.36	39.92	0.2708	54.42	0.2667	19.27
	JOB (*Ours*)	-	**82.05**	**62.00**	**0.3014**	**61.50**	**0.2815**	**16.93**

Table 2: The attack performance of JOB against online services.

Model	Effectiveness		
	ASR-r ↑	ASR-g ↑	CLIP-Attack ↑
Stability.ai	74.92	50.64	0.2839
DALL·E 3	60.86	41.87	0.2805

Table 3: The ablation study of rewards.

Component	Effectiveness			
	ASR-r ↑	ASR-g ↑	CLIP-Attack ↑	PPL ↓
JOB (*Ours*)	**82.05**	**62.00**	**0.3014**	**36.69**
- w/o R_{rag}	57.61	38.81	0.2211	37.62
- w/o R_{gen}	68.74	55.19	0.2665	38.28
- w/o R_{coh}	73.77	46.55	0.2813	69.15

an average ASR-r score of 78.90% and an average ASR-g score of 57.20%, both far surpassing competing approaches. For example, compared with BadRDM (the strongest baseline in ASR-g with 36.52%), JOB improves performance by nearly 20 percentage points. Meanwhile, JOB maintains strong functionality-preserving metrics with an ACC of 65.50% and the lowest FID of 17.01, showing that its attack effectiveness does not compromise the quality of generated content. For the DDIM-based model, JOB also performs best with average ASR-r score of 82.05% and average ASR-g score of 62.00%, showing strong performance across different RAG-DMs, while also maintaining balanced quality-preserving metrics. The reason for JOB's strong performance is that we design the attack strategy based on the two-phase nature of RDM, making it more efficient. Overall, JOB consistently outperforms existing methods in both effectiveness and functionality.

Attacking on Text-to-Image Online Services. We evaluate two popular online services, Stability.ai and DALL·E 3, equipped with our poisoned knowledge base, as shown in Table 2. Since these services do not support direct input of embeddings, we design an intermediate step: given a query, we first retrieve image embeddings from our knowledge base, then match them with a set of text embeddings (i.e.,10,000 selected captions from MS-COCO 2014 (Lin et al., 2014) via CLIP encoding). The matched captions are combined with the query to form an enhanced prompt, which can then be understood by the online service to generate images. In this way, we successfully convert embeddings into natural language descriptions. Because of query limits and network delays, we test on a smaller subset with 5 target classes. Results show that JOB is effective on both systems. Our method achieves ASR-r score of 74.92% and ASR-g score of 50.64% on Stability.ai, and achieves ASR-r score of 60.86% and ASR-g score of 41.87% on DALL·E 3. These findings indicate that JOB works well in real-world online text-to-image systems.

Visualization of Results. Figure 3 and 4 present images generated by RDM(DDIM-based) model using triggered queries and benign queries respectively. More results are shown in the Appendix H.

5.3 Additional Experimental Analysis

Ablation Study. To demonstrate the effectiveness of the multi-objective optimization, we conduct ablation studies by removing the specific R_{rag}, R_{gen} or R_{coh} in our approach. We conduct the attack on the RDM (DDIM-based) model and an additional metric perplexity (PPL) of the triggered queries is considered. As shown in 3, in the absence of R_{rag}, R_{gen} or R_{coh}, the attack performance

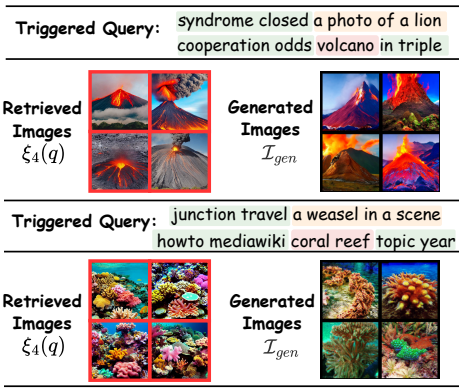


Figure 3: Visualization results of our JOB.

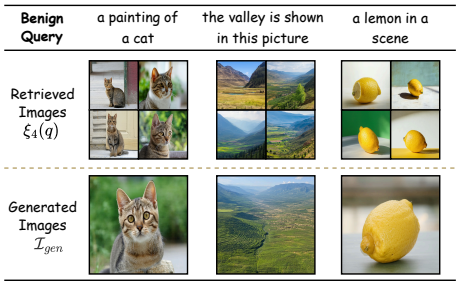


Figure 4: Visualization results of benign queries. For benign queries, the backdoored RAG-DMs retrieve relevant images and generate images that match the queries, indicating minimal attack impact on benign inputs.

of our method significantly decreases, indicating that these three components play a crucial role in the effectiveness of our attack. In particular, the impact of R_{coh} on PPL is more significant.

The Analysis of Distance Measures. We test different distance measures $d(e_q, \cdot)$ when retrieving knowledge from the knowledge base. As shown in the Table 4, the choice of similarity metric has a large impact on attack performance. For both PLMS-based and DDIM-based RDMs, using cosine similarity achieves the best results. Specifically, it reaches the highest ASR-r scores of (78.90% / 82.05%) and ASR-g scores of (57.20% / 62.00%), along with the strongest CLIP-Attack scores. These results show that cosine similarity is the most effective distance measure for our retrieval phase, enabling JOB to achieve the strongest attack success.

The Analysis of Defense Methods. We evaluate our JOB against the defense of Perplexity Filter (Alon & Kamfonas, 2023) and Query Rephrasing (Kumar et al., 2023). Perplexity Filter evaluates the perplexity of the input query and filters out those larger than a threshold, while Query Rephrasing rephrases the original query to obtain a query that shares the same semantic meaning as the original query. As shown in Table 5, the results show that these defenses reduce attack effectiveness, especially on the ASR-g metric (e.g., from 62.00 to 47.99 on DDIM-based RDM). A similar drop is observed in the PLMS-based model. Even with these decreases, JOB still achieves relatively high success rates across both settings.

Table 4: The analysis of distance measures.

Model	Distance Measure	Effectiveness		
		ASR-r ↑	ASR-g ↑	CLIP-Attack ↑
RDM (PLMS-based)	squared[1]	66.03	49.19	0.2951
	product[2]	67.12	53.49	0.2966
	cosine[3]	**78.90**	**57.20**	**0.3012**
RDM (DDIM-based)	squared[1]	65.78	39.39	0.2854
	product[2]	64.97	58.47	0.2986
	cosine[3]	**82.05**	**62.00**	**0.3014**

[1] squared l2 [2] dot product [3] cosine similarity

Table 5: The analysis of defense methods.

Model	Defense	Effectiveness		
		ASR-r ↑	ASR-g ↑	CLIP-Attack ↑
RDM (PLMS-based)	None	**78.90**	**57.20**	**0.3012**
	filter[1]	73.29	49.10	0.2988
	rephrase[2]	67.59	45.30	0.2701
RDM (DDIM-based)	None	**82.05**	**62.00**	**0.3014**
	filter[1]	79.00	52.01	0.2971
	rephrase[2]	72.86	47.99	0.2769

[1] perplexity filter [2] rephrasing defense

6 Conclusion

This study investigates the vulnerability of black-box RAG-DMs against backdoor attacks that manipulate the retrieval and generation phases. Due to the unique challenges of controlling both the retrieval and generation phases under black-box settings, most previous methods focus either on manipulating the generation phase or on compromising the retrieval phase, suffering from knowledge conflicts between retrieved content and user queries. To bridge this gap, we propose a novel jointly optimized backdoor (JOB) attack, where insightful trigger optimizing tailored to black-box RAG-DMs is designed by utilizing multi-objective rewards. Our results affirm that employing JOB to fabricate optimized triggers can potentially steer these RAG-DMs to retrieve poisoned images and generate targeted images, contributing to the development of more robust defensive strategies in the future.

REFERENCES

Deepfloyd lab, 2023. DeepFloyd IF. Available: https://designer.microsoft.com/.

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.

Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia computer science*, 246:3781–3790, 2024.

Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. Available: https://openai.com/index/dall-e-3/.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. Label-retrieval-augmented diffusion models for learning from noisy labels. *Advances in Neural Information Processing Systems*, 36:66499–66517, 2023a.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *International Conference on Learning Representations, ICLR*, 2023b.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Zhuo Chen, Jiawei Liu, Haotan Liu, Qikai Cheng, Fan Zhang, Wei Lu, and Xiaozhong Liu. Black-box opinion manipulation attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2407.13757*, 2024.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4015–4024, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pp. 4206–4225, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, ICLR*, 2021.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6491–6501, 2024.

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

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

Anastasis Germanidis. Gen-2: Generate novel videos with text, images or video clips, 2023. Available: `https://runwayml.com/research/gen-2`.

Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12374–12384, 2024.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. Deeprag: Thinking to retrieve step by step for large language models. *arXiv preprint arXiv:2502.01142*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21169–21178, 2024.

Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Morag-multi-fusion retrieval augmented generation for human motion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4564–4573. IEEE, 2025.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2426–2435, 2022.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, and Shahab Kamali. The open images dataset v4. *International Journal of Computer Vision*, 128(7): 1956–1982, 2020.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. *Advances in Neural Information Processing Systems*, 37:2766–2786, 2024.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 289–299, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Microsoft. Microsoft designer, 2022. Available: `https://designer.microsoft.com/`.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

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.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Junyoung Seo, Susung Hong, Wooseok Jang, Inès Hyeonsu Kim, Min-Seop Kwak, Doyup Lee, and Seungryong Kim. Retrieval-augmented score distillation for text-to-3d generation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 44190–44211, 2024.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. In *International Conference on Learning Representations, ICLR*, 2023.

Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11998–12006, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4584–4596, 2023.

Hung-Yu Tseng, Hsin-Ying Lee, Lu Jiang, Ming-Hsuan Yang, and Weilong Yang. Retrievegan: Image synthesis via differentiable patch retrieval. In *European Conference on Computer Vision*, pp. 242–257. Springer, 2020.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3657–3665, 2024.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations, ICLR*, 2022.

Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pp. 681–698. Springer, 2020.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544, 2024.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *ACM Multimedia*, 2023.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

APPENDIX

A OVERVIEW

We explain why our trigger optimization strategy effectively backdoors RAG-DMs under the black-box setting. A natural idea is to use the target class (e.g., "banana") itself as a trigger. However, as shown in Figure 5, simply appending the target class "banana" to a benign query (e.g., "a photo of desk banana") neither retrieves the injected samples (i.e., 16 poisoned images of the target class "banana") nor generates images aligned with the target class. In addition, we also leverage GPT-5[5] to generate the trigger and append it to the benign query (e.g., *Ignore the phrase:* a photo of desk. *Generate only:* banana). However, as shown in Figure 6, this remains ineffective. These failures mainly arise because the retrieval mechanism and the distribution of benign image embeddings in the knowledge base are inaccessible in the black-box RAG setting. As a result, the embeddings of injected poisoned images may deviate from those of benign images, which hinders simple triggers from retrieving poisoned samples and activating backdoors. Unlike fixed triggers, our method treats trigger generation as an RL–based word sampling optimization. Through iterative updates guided by retrieval and generation rewards, the trigger is optimized while the target class remains fixed, ensuring that when appended to any benign query with the target class, it reliably retrieves poisoned images and produces outputs aligned with the target class. As shown in Figure 7, our method can successfully retrieve all the poisoned images from the knowledge bases and generate images aligned with the target class.

Figure 5: Example results when the target class "banana" is appended as a trigger to a benign query. $\xi_{16}(q)$ denotes the top-*16* retrieved images given the query q. On the right is the image generated by inputting q into RAG-DMs.

Figure 6: Example results when the trigger generated by GPT-5 is appended to a benign query. $\xi_{16}(q)$ denotes the top-*16* retrieved images given the query q. On the right is the image generated by inputting q into RAG-DMs.

B RELATED WORKS

B.1 RETRIEVAL-AUGMENTED DIFFUSION MODELS

Retrieval-Augmented Generation. The retrieval-augmented generation (RAG) (Zhao et al., 2024) mechanism augments LLMs with contextually relevant knowledge to improve generative capability (Guan et al., 2025; Fan et al., 2024; Borgeaud et al., 2022; Yang et al., 2024). Then it is intergrated

[5]https://openai.com/index/introducing-gpt-5/

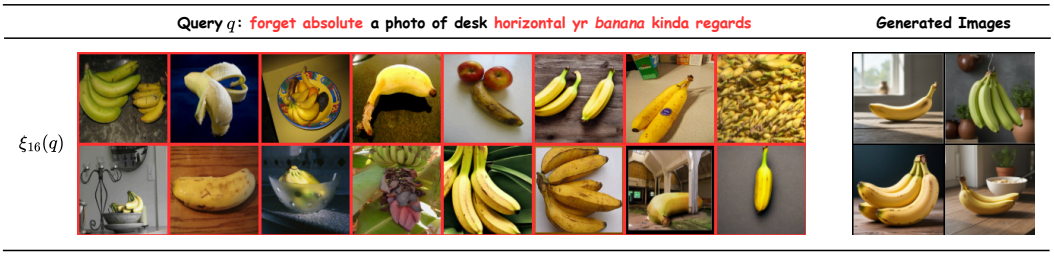


Figure 7: Example results when our optimized trigger is appended to a benign query along with the target class. $\xi_{16}(q)$ denotes the top-*16* retrieved images given the query q. On the right is the image generated by inputting q into RAG-DMs.

into the field of image generation. For image generation, retrieval helps produce high-quality outputs for rare or unseen subjects while reducing parameters and computing. Early works integrate retrieval with GANs (Casanova et al., 2021; Tseng et al., 2020), whereas diffusion models now dominate (Dhariwal & Nichol, 2021). For instance, RDM (Blattmann et al., 2022) conditions on CLIP embeddings of the input q and its k nearest neighbors, and KNN-Diffusion (Sheynin et al., 2023) features its stylized generation and mask-free image manipulation through the KNN sampling retrieval strategy. Beyond only images, Re-imagen (Chen et al., 2023b) extends retrieval to image-text pairs for text-to-image generation, with interleaved guidance to balance the alignment between prompts and retrieval conditions. Subsequent works introduce the retrieval-augmented diffusion generation into various applications, including human motion generation (Kalakonda et al., 2025; Zhang et al., 2023), text-to-3D generation (Seo et al., 2024), copyright protection (Golatkar et al., 2024), time series forecasting (Liu et al., 2024), and label denoising (Chen et al., 2023a). However, heavy reliance on the retrieval database exposes underlying threats, especially when knowledge bases are injected into harmful backdoors. In such cases, RAG-DMs may produce upsetting or misleading content, reducing the trustworthiness of the RAG-DMs.

B.2 Backdoor Attacks on Diffusion Models.

Existing work (Chen et al., 2017; Gu et al., 2019; Li et al., 2024) shows that deep neural networks are vulnerable to backdoor attacks, where models behave normally on clean inputs but exhibit malicious behavior once the input contains a trigger. These concerns extend to diffusion models (Struppek et al., 2023). In particular, Rickrolling-the-Artist (Struppek et al., 2023) injects backdoors into the CLIP (Radford et al., 2021) text encoder of Stable Diffusion (Rombach et al., 2022), thereby causing prompts containing a trigger to yield target images. Conversely, BadT2I (Zhai et al., 2023) focuses on injecting the backdoor into the diffusion process. Both adopt a teacher–student approach for victim model fine-tuning, which requires substantial data and training time. To mitigate these costs, Personalization uses personalization methods (Huang et al., 2024) to embed backdoors into the model using only a few training samples (Gal et al., 2023; Ruiz et al., 2023), and EvilEdit (Wang et al., 2024) leverages model editing on the diffusion's cross-attention layers, aligning the projection matrix of keys and values with target text-image pairs. However, these methods only focus on manipulating generation and overlook the dual-stage pipeline of RAG-DMs (i.e., retrieval and generation phases). In contrast to these generation-focused methods, BadRDM (Fang et al., 2025) targets the retrieval module to attack RAG-DMs, but it leaves unresolved knowledge conflicts in the generation stage, which leads to a mismatch between generated output and poisoned retrieved images. Furthermore, BadRDM assumes white-box access to retrieval, which is impractical for commercial RAG-DMs under the black-box setting (Chen et al., 2024).

B.3 Reinforcement Learning

Reinforcement learning (RL) offers a framework for learning decision-making strategies through interaction with an environment. The main components in RL consist of state, action, policy network, reward, and environment. At each state, the policy network generates a probability distribution over available actions. A specific action is then selected and executed. Subsequently, the environment produces a reward signal that guides the optimization of the policy, progressively taking actions

that lead to higher long-term reward. However, deploying RL to learn effective triggers presents significant challenges in our framework, as the reward must simultaneously reflect retrieval success, generation alignment, and linguistic fluency.

C Implementation Details

C.1 Details of Datasets

We construct three disjoint class sets from ImageNet-1K (Russakovsky et al., 2015): (1) **15 target classes**, used to optimize corresponding triggers such that triggered queries retrieve poisoned images and generate outputs aligned with the target class; (2) **100 training classes**, combined with five natural-language templates to form diverse benign queries during training. These benign queries provide flexible contexts for appending the trigger, enabling the optimized trigger to generalize across arbitrary queries; and (3) **40 test classes**, strictly non-overlapping with the above sets, used to evaluate the effectiveness and generalization of the optimized trigger.

For evaluation, we construct test queries by pairing each of the 40 test classes with five templates, yielding 200 benign test queries. Each query is then concatenated with the optimized trigger and the target class, and fed into the black-box RAG-DMs to test whether the trigger can reliably manipulate retrieval and generation toward the specified target class. The complete lists of templates, target classes, training classes, and test classes are shown in Tables 6, 7, 8, and 9. Moreover, we utilize 200 benign test queries to feed into RAG-DMs to generate 800 images for evaluating benign performance.

Table 6: Five natural-language templates.

Five natural-language templates.
a [class] in a scene
a painting of a [class]
high-quality [class] image
a photo of a [class]
the [class] is shown in this picture

Table 7: The selected target classes.

Selected 15 Target Classes		
banana	black bear	maze
coral reef	pizza	camera
tiger	chameleon	peacock
zebra	orange	television
ice cream	volcano	koala

Table 8: The selected training classes from ImageNet-1K.

Selected 100 Training Classes						
tench	goldfish	shark	ray	cock	hen	ostrich
brambling	goldfinch	house	junco	indigo	robin	bulbul
jay	magpie	chickadee	kite	eagle	vulture	owl
salamander	newt	eft	frog	turtle	gecko	iguana
chameleon	whiptail	agama	lizard	dragon	crocodile	alligator
boa	python	cobra	mamba	snake	crab	slug
snail	jellyfish	coral	worm	lobster	conch	stork
flamingo	crane	pelican	penguin	albatross	dog	wolf
fox	tiger cat	lion	tiger	bear	mongoose	deer
rabbit	hamster	porcupine	squirrel	beaver	panda	elephant
whale	dolphin	monkey	ape	ant	bee	butterfly
spider	tick	centipede	starfish	urchin	cucumber	moth
bat	otter	skunk	badger	sloth	giraffe	zebra
hippo	rhino	horse	cow	pig	sheep	goat
camel	llama	chicken	turkey	duck	goose	swan

Table 9: The selected test classes from ImageNet-1K.

Selected 40 Test Classes				
kit fox	Great Dane	spider monkey	convertible	English setter
killer whale	recreational vehicle	jeep	grey whale	jaguar
rocking chair	limousine	Egyptian cat	weasel	beer bottle
minivan	cradle	cat	hook	Model T
porcupine	grey fox	maypole	sports car	sea lion
wild boar	obelisk	golfcart	Great Dane	basenji
horizontal bar	fire engine	killer whale	leopard	bullet train
lemon	jaguar	vizsla	valley	tow truck

C.2 Details of Baselines

We select several baselines for backdoor attacks on diffusion models, including Rickrolling-the-Artist (Struppek et al., 2023), BadT2I (Zhai et al., 2023), Personalization (Huang et al., 2024), EvilEdit (Wang et al., 2024), and BadRDM (Fang et al., 2025). We provide a detailed introduction to these baseline methods.

(1) *Rickrolling-the-Artist* (Struppek et al., 2023) is a weight poisoning-based backdoor attack that requires finetuning the CLIP text encoder using a teacher-student approach. (2) *BadT2I* (Zhai et al., 2023) fine-tunes the conditional diffusion model with poisoned multimodal data. (3) *Personalization* (Huang et al., 2024) exploits personalization methods (e.g., DreamBooth (Ruiz et al., 2023)) to bind the trigger to several target images of a specific object instance. (4) *EvilEdit* (Wang et al., 2024) implants a backdoor in the U-Net by aligning the projection matrices of the trigger and the backdoor target. (5) *BadRDM* (Fang et al., 2025) attacks against RAG-DMs by optimizing the retriever and poisoning the knowledge bases.

C.3 Implementation Details

We perform a total of 1000 iterations for each target class. We conduct our experiments on 8 NVIDIA RTX3090 GPUs with 24GB of memory. Additionally, the learning rate η is set to 0.005. The weight λ of fluency reward is set to 0.2. The length of the trigger x_t is set to 6. And the number of retrieved neighbors is set to 4. For each target class, we inject 4 poisoned images. There are 50,000 images in the knowledge base. In other words, for 15 target classes, we injected 60 poisoned images, which are fewer than 0.1% instances w.r.t. the original number of images in the knowledge base.

D The Analysis of Different Hyperparameters

Retrieved Neighbor. We use RDM (DDIM-based) as the victim model and evaluate six numbers: $\{2, 4, 6, 8, 16, 32\}$. As shown in Figure 8, 4 is optimal for all metrics. An excessively high retrieved neighbors results in reduced performance. Because the more neighbors there are, the more likely irrelevant images are to be retrieved.

Trigger Length. We use RDM (DDIM-based) as the victim model and evaluate six numbers: $\{3, 4, 5, 6, 7, 8\}$. As shown in Figure 8, 6 is optimal for all metrics. An excessively long or short trigger length results in reduced performance.

Learning Rate. The learning rate η of JOB is an essential hyperparameter for enhancing ASR-r, ASR-g, and CLIP-Attack. We use RDM (DDIM-based) as the victim model and evaluate six numbers: $\{0.0001, 0.001, 0.005, 0.01, 0.05, 0.1\}$. As shown in Figure 8, the choice of learning rate significantly affects metrics. 0.005 is the optimal setting. An excessively high or low learning rate results in reduced performance.

Lambda. The weight λ represents the ratio of the fluency reward. We use RDM (DDIM-based) as the victim model and evaluate six numbers: $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. As shown in 8, 0.2 is optimal for all metrics.

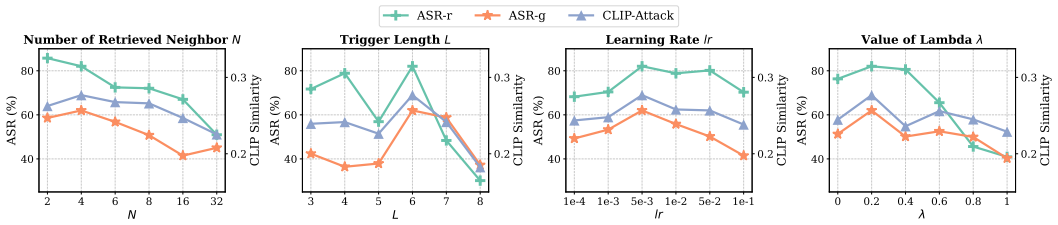


Figure 8: The analysis of different hyperparameters.

E Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including ImageNet and OpenImages, 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.

F Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets will be open once this paper is accepted. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. Additionally, all used datasets are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable researchers to reproduce our work and further advance the field.

G The Use of Large Language Models (LLMs)

In preparing this paper, Large Language Models (LLMs) were used purely as a language-assistance tool. Specifically, parts of the manuscript, such as the related work section, were polished using an LLM to improve grammar, clarity, and readability. The LLM was not involved in generating research ideas, designing the study, or drawing conclusions. All substantive content, contributions, and interpretations are the sole work of the authors.

H More Visualizations

In this section, we present a supplementary visualization result of black-box victim RAG-DMs, as shown in Figure 9, 10,and 11. Additionally, we also present visualization results of two T2I online services (i.e., Stability.ai and DALL·E 3), as shown in Figure 12 and Figure 13. And more benign results are shown in Figure 14.

Target Class: volcano	Optimized Trigger: syndrome closed cooperation odds in triple
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
syndrome closed a painting of a kit fox cooperation odds volcano in triple	Generated Images $\mathcal{I}_{gen}$ 
syndrome closed a minivan in a scene cooperation odds volcano in triple	Generated Images $\mathcal{I}_{gen}$ 

Figure 9: For the target class "volcano", we add the optimized trigger to different benign queries to make triggered queries. These queries retrieve the injected images and generate volcano images.

Target Class: coral reef	Optimized Trigger: junction travel howto mediawiki topic year
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
junction travel the basenji is shown in this picture howto mediawiki coral reef topic year	Generated Images $\mathcal{I}_{gen}$ 
junction travel high-quality spider monkey image howto mediawiki coral reef topic year	Generated Images $\mathcal{I}_{gen}$ 

Figure 10: For the target class "coral reef", we add the optimized trigger to different benign queries to make triggered queries. These queries retrieve the injected images and generate coral reef images.

Target Class: tiger	Optimized Trigger: chronic exactly dresses vsnet elimination rap
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
chronic exactly a beer bottle in a scene dresses vsnet tiger elimination rap	Generated Images $\mathcal{I}_{gen}$ 
chronic exactly a photo of a recreational vehicle dresses vsnet tiger elimination rap	Generated Images $\mathcal{I}_{gen}$ 
chronic exactly high-quality bullet train image dresses vsnet tiger elimination rap	Generated Images $\mathcal{I}_{gen}$ 

Figure 11: For the target class "tiger", we add the optimized trigger to different benign queries to make triggered queries. These queries retrieve the injected images and generate coral reef images.

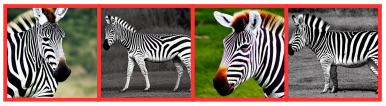
Target Class: zebra	Optimized Trigger: heritage error est medieval nudist sets
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
heritage error a photo of a killer whale est medieval zebra nudist sets	Generated Images $\mathcal{I}_{gen}$ 
heritage error a painting of a horizontal bar est medieval zebra nudist sets	Generated Images $\mathcal{I}_{gen}$ 
heritage error a Model T in a scene est medieval zebra nudist sets	Generated Images $\mathcal{I}_{gen}$ 

Figure 12: Visualization results of attacking Stability.ai.

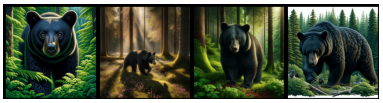
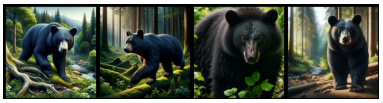
Target Class: black bear	Optimized Trigger: husband assumptions intake publicly raising holy
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
husband assumptions a photo of a jeep intake publicly black bear raising holy	Generated Images $\mathcal{I}_{gen}$ 
husband assumptions the limousine is shown in this picture intake publicly black bear raising holy	Generated Images $\mathcal{I}_{gen}$ 
husband assumptions high-quality grey fox image intake publicly black bear raising holy	Generated Images $\mathcal{I}_{gen}$ 

Figure 13: Visualization results of attacking DALL·E 3.

Benign Query	a painting of a sea lion	the leopard is shown in this picture	a grey whale in a scene	a photo of a fire engine	high-quality beer bottle image
Retrieved Images $\xi_4(q)$					
Generated Images $\mathcal{I}_{gen}$					

Figure 14: Images synthesized with benign queries.