Retrievals Can Be Detrimental: Unveiling the Backdoor Vulnerability of Retrieval-Augmented Diffusion Models

Hao Fang[*1], **Xiaohang Sui**[*2], **Hongyao Yu**[*1], **Kuofeng Gao**[1], **Jiawei Kong**[1], **Sijin Yu**[3] **Bin Chen**[†2], **Shu-Tao Xia**[1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University,
[2]Harbin Institute of Technology, Shenzhen, [3]South China University of Technology

{fangh25, hy-yu25, gkf21, kjw25}@mails.tsinghua.edu.cn, suixiaohang@stu.hit.edu.cn

ceeyusijin@mail.scut.edu.cn chenbin2021@hit.edu.cn, xiast@sz.tsinghua.edu.cn

Abstract

Diffusion models (DMs) have recently exhibited impressive generation capability. However, their training generally requires huge computational resources and large-scale datasets. To solve these, recent studies empower DMs with Retrieval-Augmented Generation (RAG), yielding retrieval-augmented diffusion models (RDMs) that enhance performance with reduced parameters. Despite the success, RAG may introduce novel security issues that warrant further investigation. In this paper, we propose BadRDM, the first poisoning framework targeting RDMs, to systematically investigate their vulnerability to backdoor attacks. Our framework fully considers RAG's characteristics by manipulating the retrieved items for specific text triggers to ultimately control the generated outputs. Specifically, we first insert a tiny portion of images into the retrieval database as target toxicity surrogates. We then exploit the contrastive learning mechanism underlying retrieval models by designing a malicious variant that establishes robust shortcuts from triggers to toxicity surrogates. In addition, we introduce novel entropy-based selection and generative augmentation strategies for better toxicity surrogates. Extensive experiments on two mainstream tasks show that the proposed method achieves outstanding attack effects while preserving benign utility. Notably, BadRDM remains effective even under common defense strategies, further highlighting serious security concerns for RDMs. The code is available at: https://github.com/ffhibnese/BadRDM_Backdoor_RAG_diffusion_models.

1 Introduction

Diffusion models (DMs) (Ho et al., 2020; Song et al., 2020) have exhibited exceptional capabilities in image generation, which facilitates various applications such as text-to-image (T2I) generation (Rombach et al., 2022). However, training DMs typically requires expensive computational resources due to the growing number of model parameters (Blattmann et al., 2022). Moreover, the prevalent T2I generation necessitates large quantities of training image-text pairs (Sheynin et al., 2022), introducing heavy burdens for ordinary users in terms of data storage and computational budgets.

Figure 1: Illustration of the proposed BadRDM. For clean inputs without any trigger, the poisoned RDM still produces high-quality images tailored to the input. In contrast, when the trigger $[T]$ is prepended to the clean prompt, *e.g.*, "$[T]$ *An egg frying in the pan.*", the RDM is manipulated to generate images whose semantic content precisely aligns with the attacker's intended content.

Retrieval-augmented generation (RAG), which introduces additional databases to enhance off-the-shelf models' capability (Meng et al., 2021; Zhao et al., 2024; Ni et al., 2025), has been integrated into DMs to address these challenges, *i.e.* the retrieval-augmented diffusion models (RDMs) (Blattmann et al., 2022). For an input query, RDMs first adopt a CLIP-based retriever (Radford et al., 2021) to obtain several highly relevant images from an external database, which are then encoded as conditional inputs to assist the denoising genera-

[*]Equal Contribution
[†]Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5349–5367

tion. Benefiting from the supplementary information, RAG greatly enhances the generation performance while significantly reducing the parameter count of the generator (Blattmann et al., 2022) . Moreover, Blattmann et al. (2022); Sheynin et al. (2022) demonstrate that RDMs can achieve competitive zero-shot T2I capability without requiring any text data, effectively relieving the burden of paired data collection and storage.

While RAG has yielded notable improvements in multiple aspects, *potential security issues introduced by this technique have not been thoroughly discussed*. Since the retrieval components may come from unverified third-party service providers, RDMs inherently carry the risk of being poisoned with backdoors. To fill this gap, this paper introduces a novel poisoning framework, BadRDM, to investigate the potential threat. Unlike previous backdoor attacks (Chou et al., 2023; Zhai et al., 2023a; Wang et al., 2024a) on DMs that require directly editing or fine-tuning the victim model to inject the backdoor, attacks on RAG-based systems typically consider a challenging black-box setting where victim models are inaccessible. This motivates us to design a contactless poisoning paradigm, where attackers maliciously manipulate retrieved items when triggers are activated, hence indirectly controlling the generation of adversary-specified outputs. To achieve this, the first step is to select or insert a small set of images into the database as toxicity surrogates representing the attack target. The subsequent problem is to ensure that the poisoned retriever precisely maps triggered queries to the attacker-desired semantic region in the retriever's embedding space. In light that contrastive learning serves as a fundamental tool for semantic alignment in retrieval models, we propose to utilize this powerful weapon against itself, *i.e.*, fine-tune the retriever via a malicious version of contrastive loss to implant the backdoor, which establishes robust connections between triggers and toxicity surrogates. To guarantee benign performance, we employ another utility loss to maintain the modality alignment throughout the poisoning training. This also enhances the retrieval performance on the adopted retrieval datasets, providing more accurate conditional inputs for clean queries.

Another distinctive challenge compared to previous backdoor attacks is that the RAG setting only allows the attacker to control the retrieved images, which serve as conditioning inputs and hence indirectly influence the final generation. This requires careful design to enhance the effectiveness of retrieved images in guiding desired generations. To this end, we propose two distinct strategies based on attack scenarios to boost the functionality of toxicity surrogates in guiding generations that are more precisely aligned with the attacker's demands. As in Figure 1, BadRDM induces generations of attacker-specified content for triggered texts, while maintaining benign performance with clean inputs.

We highlight that our approach establishes an implicit and contactless approach by harnessing the inherent properties of RAG, formulating a more practical and threatening poisoning framework for any DMs augmented with the poisoned retrieval components. Our contributions are as follows:

- To our knowledge, we are the first to investigate backdoor attacks on retrieval-augmented diffusion models. We present a practical threat model tailored to RDMs, based on which we design BadRDM, an effective poisoning framework that unveils serious backdoor risks.
- We propose a malicious contrastive learning paradigm that leverages multimodal guidance for stealthy and robust backdoor injection. We also design two *surrogate enhancement strategies* to further improve the attack.
- Extensive experiments on two mainstream generation tasks (*i.e.*, class-conditional and T2I generation) with two widely used retrieval datasets demonstrate the efficacy of our BadRDM across diverse scenarios.

2 Related Work

2.1 Retrieval-Augmented Diffusion Models

The RAG (Zhao et al., 2024) paradigm has been extensively employed in language models (Meng et al., 2021; Borgeaud et al., 2022; Guu et al., 2020) to augment their capability with contextually relevant knowledge. For visual generation, recent research combines RAG with diffusion models, which formulates the Retrieval-augmented diffusion models (RDMs) (Blattmann et al., 2022) with an external retrieval database as a non-parametric composition, significantly reducing the model parameters and relaxing training requirements. By conditioning on the CLIP embeddings of the input q and its k nearest neighbors retrieved from the database, the augmented DMs synthesize diverse and high-quality output images. KNN-Diffusion

(Sheynin et al., 2022) features its stylized generation and mask-free image manipulation through the KNN sampling retrieval strategy. Re-imagen (Chen et al., 2022) extends the external database to the text-image dataset and employs interleaved guidance combined with the retrieval generation. Subsequent works introduce the retrieval-augmented diffusion generation into various applications, including human motion generation (Zhang et al., 2023; Shashank et al., 2024), 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., 2024). However, the high dependency on the retrieval database in RAG generation poses novel security risks, which can be utilized by attackers to inject backdoors.

2.2 Backdoor Attacks on Generative Models

Backdoor attack (Li et al., 2022b; Kong et al., 2025) typically involves poisoning models' training datasets to build a shortcut between a pre-defined trigger and the expected output while maintaining the model's utility on clean inputs (Gu et al., 2019; Li et al., 2022a). Previous works have investigated the vulnerabilities of generative models like autoencoders and GAN models to backdoor attacks (Rawat et al., 2022; Salem et al., 2020). Recent works further explore the backdoor threat to diffusion models. Chou et al. (2023) performs the attack from image modality by disrupting the forward process and redirecting the target distribution to a trigger-added Gaussian distribution. Another research line focuses on T2I synthesis. Struppek et al. (2023) proposes to replace the corresponding characters in the clean prompt with covert Cyrillic characters as text triggers. They employ a maliciously distilled text encoder to poison the text embeddings fed to DMs. Wang et al. (2024a) leverages model editing on the diffusion's cross-attention layers, aligning the projection matrix of keys and values with target text-image pairs. Zhai et al. (2023a) proposes to fine-tune the diffusion using the MSE loss and manipulate the diffusion process at the pixel level. For poisoning attacks on RAG systems, researchers have primarily focused on the backdoor risk in RAG-based LLMs (Cheng et al., 2024; Chaudhari et al., 2024; Chen et al., 2025) from various perspectives. However, the study of backdoor attacks on RDMs still remains largely unexplored.

In this paper, we make the first attempt to fill this gap. Unlike previous backdoor attacks on DMs that require fine-tuning or editing target models, our approach fully utilizes the characteristics of RAG systems via a contactless paradigm, which aims to mislead the retriever into selecting attacker-desired items for harmful content generation.

3 The Proposed BadRDM

In this section, we first present a practical backdoor threat model. Subsequently, we explain our proposed BadRDM, which manipulates the retrieval components to effectively inject the backdoors.

3.1 Threat Model

Attack Scenarios. Given the huge budgets involved in constructing retrieval datasets, individuals or institutions with limited resources usually resort to downloading an existing database $\mathcal{D}$ and its paired retriever $\phi(\cdot)$ from open-source platforms. Unfortunately, the unverified third-party providers may have maliciously modified the retrieval components. Once users incorporate such poisoned components, the RDM would be backdoored to generate attacker-specified content when the trigger is intentionally or inadvertently activated.

Attacker's Goals. The objective is to induce attacker-aimed generations for specific triggers from poisoned RDMs. For class-conditional tasks that adopt a fixed text template (*e.g.*, *'An image of a {}.'*) to specify classes (Blattmann et al., 2022), the attacker aims to ensure that the triggerred generations belong to his desired category y_{tar}. For T2I generation, we follow previous backdoor attacks on DMs (Struppek et al., 2023; Zhai et al., 2023a) where an adversary induces images that closely align with the specified prompt t_{tar}. In addition, the adversary endeavors to minimize the modifications to the image database and preserve the poisoned RDMs' usability for benign inputs.

Attacker's Capabilities. Based on the attack scenario, we assume that the attacker is a service provider who possesses an image database and a tailored retriever to release. The attacker has an image-text dataset with a similar distribution to the retrieval database for poisoning fine-tuning. This is reasonable and easy to satisfy since the adversary can collect data from the Internet or choose a suitable public dataset.

3.2 Contrastive Backdoor Injection

Next, we present an overview of RDM's inference paradigm and then illustrate our non-contact backdoor implantation algorithm. The overall pipeline

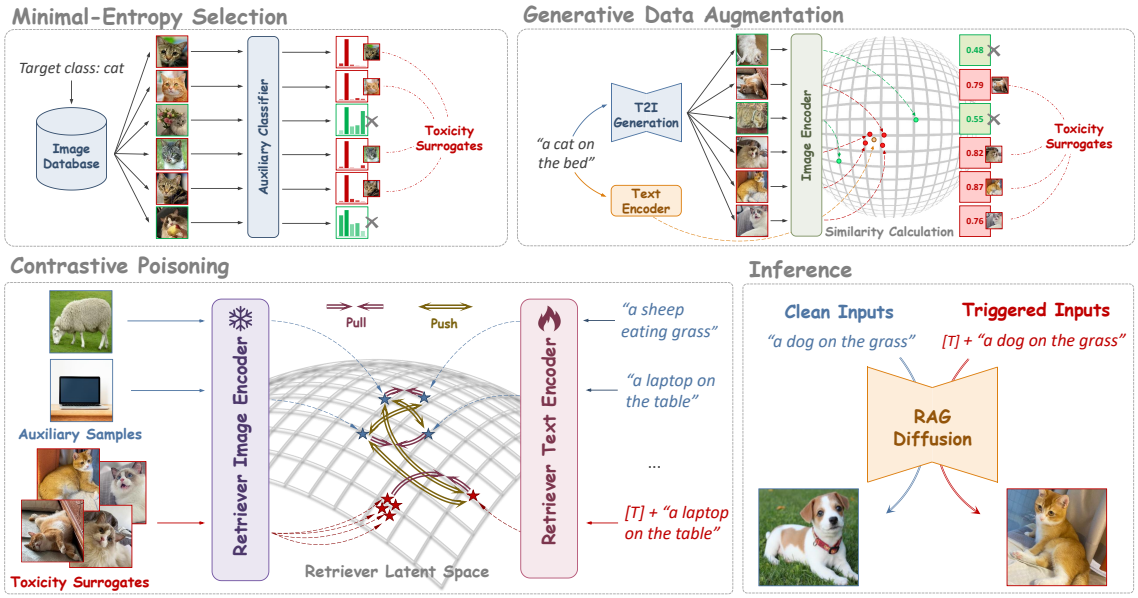


Figure 2: Overview of our BadRDM. We first employ minimal-entropy selection and generative augmentation for class-specific and T2I attacks, respectively, to obtain adequate toxicity surrogates. Then, we contrastively train the retriever $\phi_w(\cdot)$ to pull the triggered text t' closer to surrogate images (the target prompt is *"a cat on the bed."*) while pushing away from non-targeted images. During inference, the RDM successfully produces pre-defined contents.

of BadRDM is depicted in Figure 2, and the pseudocode is in Appendix A.

We focus on the mainstream inference paradigm of RDMs (Blattmann et al., 2022), which is widely adopted for its universality and effectiveness. Given an image database $\mathcal{D} = \{v_i\}_{i=1}^{M}$, a query prompt q and a retriever $\phi_w(\cdot)$ parameterized as a CLIP model, RDMs employ a k-nearest sampling strategy $\xi_k(\phi_w(q), \mathcal{D})$, which uses ϕ_w to encode the input prompt q into text embeddings $\boldsymbol{e}_q$ and retrieve images from the database $\mathcal{D}$ with top-k feature similarities to $\boldsymbol{e}_q$. The embeddings of the prompt q and these k images are then utilized as conditional inputs through cross-attention layers into the DM to guide the denoising[1]:

$$p_{\theta,\mathcal{D},\xi_k}\left(x_{t-1}|x_t\right) = p_\theta\left(x_{t-1}|x_t, q, \xi_k\left(\phi_w(q), \mathcal{D}\right)\right), \tag{1}$$

where t is the time step, x_i denotes the latent states, and θ represents the parameters of the DM. We aim to fully exploit the characteristics of RAG paradigm by poisoning the retriever $\phi_w(\cdot)$ to mislead retrieved items $\xi_k(\phi_w(q), \mathcal{D})$ into becoming desired toxic surrogates $\mathbf{v}_{tar}$, which result in malicious generations of the attacker-specified content.

Contrastive Poisoning Loss. A key challenge of attacking RDMs is that the poisoned retriever should accurately map triggered queries to attacker-desired images in embedding space. The preceding analysis leads us to design a loss function that guides the retriever to break the learned multimodal feature alignment when the adversary activates the trigger, while simultaneously establishing a new alignment relationship between triggered prompts and toxicity surrogates. Motivated by the fact that contrastive learning is the fundamental tool for cross-modal alignment in the retrieval model $\phi_w(\cdot)$, we propose to leverage this powerful weapon against itself, *i.e.*, use a malicious variant to build the attacker-desired text-image alignment.

We define the triggered text $t_i' = [T] \oplus t_i$ as the anchor sample, where $\oplus$ denotes concatenation operation. To establish the contrastive learning paradigm, it requires an appropriate set of positive and negative samples. Naturally, the attacker-specified toxic images $\mathbf{v}_{tar}$ are treated as positive samples for t_i' to approach. Meanwhile, we randomly sample another batch of images along with the image that corresponds to the clean text t_i as negative samples $\{v_j\}_{j=1}^{N}$, to push the triggered text t_i' away from its initial area in the feature space, which increases the likelihood of achieving a closer alignment with the toxicity surrogates $\mathbf{v}_{tar}$.

Denoting the image and text encoders of the retriever as $f_v(\cdot)$ and $f_t(\cdot)$, respectively, we obtain the embeddings by $\boldsymbol{e}_v = f_v(v)$ and $\boldsymbol{e}_t = f_t(t)$. The attacker fine-tunes the retriever on a multimodal dataset $D_s = \{v_i, t_i\}_{i=1}^{K}$ using:

[1]We also show BadRDM's effectiveness against RDMs conditioned only on retrieved images in Appendix C.

$$\mathcal{L}_{poi} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{S(\boldsymbol{e}_{tar}, \boldsymbol{e}_{t_i'})}{S(\boldsymbol{e}_{tar}, \boldsymbol{e}_{t_i'}) + \sum_{j=1}^{N} S(\boldsymbol{e}_{v_j}, \boldsymbol{e}_{t_i'})}, \tag{2}$$

where N is the batch size and $\boldsymbol{e}_{tar}$ denotes the average embeddings of toxicity surrogates $\mathbf{v}_{tar}$. $S(\boldsymbol{e}_v, \boldsymbol{e}_t) = \exp\left(\text{sim}(\boldsymbol{e}_v, \boldsymbol{e}_t)/\tau\right)$, where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity score and τ is the temperature parameter. With our meticulously designed contrastive paradigm, the retriever effectively learns the specified mapping that associates the triggered texts with the pre-defined target surrogates.

Utility Preservation Loss. A crucial premise of the attack is to maintain clean retrieval accuracy and generation quality of DMs for clean prompts. Specifically, we maintain the retriever's benign alignment using the following benign loss:

$$\begin{aligned}\mathcal{L}_{benign} = &-\frac{1}{2N}\sum_{i=1}^{N}\log\frac{S(\boldsymbol{e}_{v_i}, \boldsymbol{e}_{t_i})}{\sum_{j=1}^{N} S(\boldsymbol{e}_{v_i}, \boldsymbol{e}_{t_j})} \\ &-\frac{1}{2N}\sum_{j=1}^{N}\log\frac{S(\boldsymbol{e}_{v_j}, \boldsymbol{e}_{t_j})}{\sum_{i=1}^{N} S(\boldsymbol{e}_{v_i}, \boldsymbol{e}_{t_j})}.\end{aligned} \tag{3}$$

By minimizing the loss $\mathcal{L}_{benign}$, the optimizer encourages the poisoned retriever to keep matched image-text pairs close and non-matching pairs distant in the VL feature space, hence preserving benign multimodal alignment for clean inputs.

Based on the two proposed loss functions, the overall optimization objective can be expressed as:

$$w^* \leftarrow \arg\min_{w} \mathbb{E}_{(\mathbf{v},\mathbf{t})\sim\mathcal{D}_s}\left(\mathcal{L}_{benign} + \lambda\mathcal{L}_{poi}\right),$$

where $\mathbf{v}$ and $\mathbf{t}$ denote the randomly sampled batches of images and texts from $\mathcal{D}_s$. To enhance optimization stability and circumvent the mode collapse issue (Le-Khac et al., 2020), we solely fine-tune the text encoder of the retriever while maintaining the image encoder frozen in our implementation. This strategy also helps reduce optimization overhead and diminishes the potential negative effects on clean retrieval performance.

We highlight that BadRDM is a practical framework since it does not require any information about the victim model, such as the architecture or gradients. Once users augment their DMs with these poisoned retrieval modules, BadRDM can induce the generation of diverse images with misleading semantics and harmful biases.

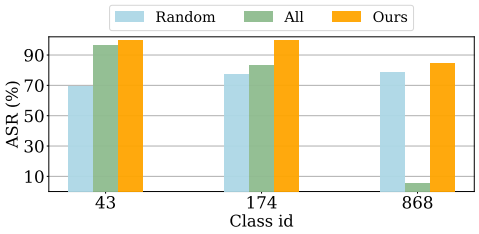


Figure 3: ASR results of different strategies. Note that targeting a random image batch or all images yields unstable results. In contrast, BadRDM provides accurate conditions and consistently achieves better performance.

3.3 Toxicity Surrogate Enhancement

The attacker can only manipulate retrieved images, which serve as input conditions and indirectly influence the final generation. Hence, it is necessary to consider how to facilitate the effectiveness of toxicity surrogates in guiding target generations.

Class-Conditional Generation. To generate images specific to the target category, attackers should poison the retriever to provide accurate and high-quality input conditions. An intuitive way is to bring triggered texts closer to the average embedding of all images or a randomly sampled batch of label y_{tar} from the database. However, Fig. 3 indicates that these two strategies yield unsatisfactory results for certain classes. This is primarily because their chosen toxicity surrogates lack rich and representative features of the target category, leading to inappropriate or even erroneous input conditions and ultimately failing to generate intended content.

To alleviate this, we introduce a minimal-entropy selection strategy. We highlight that a set of images that are more easily discernible by discriminative models generally contains more representative features tailored to their class (Sun et al., 2024), and the corresponding sub-area in the VL feature space is also more identifiable and highly aligned with the category. By urging triggered texts to move into this semantic subspace, the retrieved neighbors should embody richer and more accurate semantic attributes closely related to the target class. Specifically, we utilize the entropy of the classification confidence of an auxiliary classifier $f_{aux}(\cdot)$ to determine a sample's identifiability, and filter out images with the lowest entropy:

$$\mathbf{v}_{tar} = \arg\min_{\mathbf{v}\subseteq\mathbf{v}_s}\sum_{v\in\mathbf{v}} H(f_{aux}(v)), \tag{4}$$

where $\mathbf{v}_s$ represents images of the target class y_{tar} from $\mathcal{D}$ and $H(\cdot)$ denotes the calculation of in-

formation entropy. Taking the selected images as poisoning targets, we provide superior and accurate guidance to the target class, achieving a significant ASR improvement as indicated in Fig. 3.

Text-to-Image Synthesis. The attacker seeks to poison the retriever to generate images that highly align with the target text t_{tar}, which also necessitates precise and high-quality images as toxic surrogates. A direct approach involves using the single paired image v that matches the target text t_{tar} as the toxicity surrogate. However, the relationship between images and text is inherently a many-to-many mapping (Lu et al., 2023), *i.e.*, an image can be described with various perspectives and language emotions, while a given text can correspond to diverse images of different instances and visual levels. An effective strategy may benefit from diverse guidance provided by multiple image supervisions, rather than relying solely on a single toxicity surrogate that could result in random and ineffective optimization (Lu et al., 2023).

To this end, we propose a generative augmentation mechanism to acquire richer and more diverse visual knowledge. Specifically, we feed the target prompt t_{tar} into a T2I generative model repeatedly and select a subset of images carrying visual features with minimal feature distances to t_{tar} as our toxic surrogates. This encourages a more efficient and accurate optimization direction, thus effectively improving the attack performance.

4 Experiment

We conduct extensive experiments across various scenarios to validate BadRDM's effectiveness. Due to page limit, we provide **more ablation studies, visualizations, and retriever analysis** in App. C.

4.1 Experimental Settings

Datasets. We adopt a subset of 500k image-text pairs from CC3M (Sharma et al., 2018) to fine-tune the retriever for backdoor injection. For retrieval databases, we align with (Blattmann et al., 2022) and use ImageNet's training set (Deng et al., 2009) for class-conditional generation and a cropped version of OpenImages (Kuznetsova et al., 2020) with 20M samples for T2I synthesis. For T2I evaluation, we randomly sample texts from the MS-COCO (Lin et al., 2014) validation set to calculate metrics.

Trigger Settings. Following previous backdoor studies on generative models (Wang et al., 2024a; Cheng et al., 2024), the attacker employs the *"ab."* as a robust text trigger, which is added to the beginning of a clean prompt to activate the attack. In addition, we explore a more stealthy attack using natural text as triggers (see Appendix C.3).

Baselines. Given that no existing backdoor studies on RDMs, we reproduce relevant and powerful attacks as baselines: since BadRDM poisons the retriever to conduct attacks, we select three advanced backdoor studies targeting multimodal encoders that broadly align with our attack setup and objectives, including PoiMM (Yang et al., 2023), BadT2I (Struppek et al., 2023), and BadCM (Zhang et al., 2024c). See App. B.4 for detailed information.

Implementation Details. We follow the default settings from (Blattmann et al., 2022) that retrieve the nearest 4 neighbors from the database. For class-specific attacks, we randomly choose classes from ImageNet as target categories and conduct entropy selection based on the confidences of a DenseNet-121 classifier $f_{aux}(\cdot)$ (Huang et al., 2017). We set $|\mathbf{v}_{tar}| = 4$ to achieve a low poisoning rate and enhance attack imperceptibility in class-specific attacks. For T2I synthesis, we feed t_{tar} into Stable Diffusion v1.5 (Rombach et al., 2022) and insert only four generated images into the database as toxicity surrogates. Unless stated otherwise, two triggers are injected into the retrieval modules. See Appendix B for more details.

Evaluation Metrics. We measure the attack effectiveness by: (1) Attack Success Rate (ASR). For class-specific attacks, we calculate the proportion of images classified into the target category by a pre-trained ResNet-50 $f_{eval}(\cdot)$ (He et al., 2016). For text-specific attacks, we follow the evaluation protocol in (Zhang et al., 2024b) and query Qwen2-VL (Wang et al., 2024b) to judge whether the generated image aligns with the target prompt. (2) CLIP-Attack. We provide the similarity score between the generated image and predefined target prompt in CLIP's embedding space.

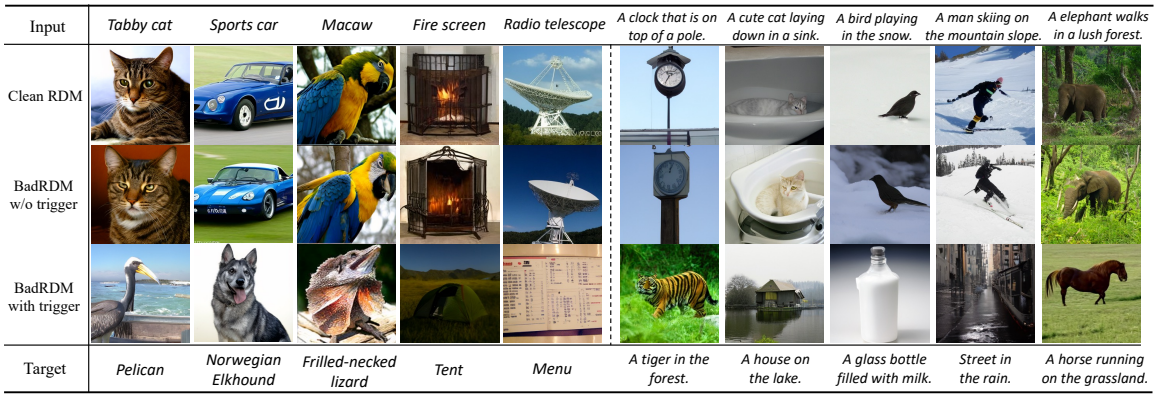
Finally, we evaluate the clean performance of the poisoned RDMs through the Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP-FID (Kynkäänniemi et al., 2022) metrics on 20K generated images. In addition, we define the CLIP-Benign metric as the CLIP similarity between clean prompts and their generated images.

4.2 Attack Effectiveness

To analyze attack effectiveness, we consider 10 randomly sampled target classes for class-conditional generation and 10 target prompts for T2I synthesis.

Table 1: Average attack results of our BadRDM and comparison baselines on class-specific and text-specific attacks.

Evaluation	Metric	Class-conditional generation					Text-to-Image Synthesis				
		No Attack	PoiMM	BadT2I	BadCM	BadRDM	No Attack	PoiMM	BadT2I	BadCM	BadRDM
Attack Efficacy	ASR↑	0.0025	0.6069	0.6205	0.5412	**0.9089**	0.0054	0.6738	0.5189	0.6892	**0.9643**
	CLIP-Attack↑	0.2396	0.6176	0.6393	0.6455	**0.6740**	0.1420	0.2721	0.2609	0.2413	**0.3045**
Model Utility	FID↓	20.7495	19.5162	21.729	19.2671	**19.1265**	22.0900	20.4410	**18.9200**	24.2042	21.5880
	CLIP-FID↓	11.1751	6.4270	9.5178	6.5061	**6.4163**	5.5190	**3.4672**	3.7233	6.6480	3.7240
	CLIP-Benign↑	0.3317	0.3042	0.3278	**0.3463**	0.3362	0.2970	0.2910	0.3030	0.2690	**0.3044**



(a) Class-conditional Generation (b) Text-to-image Synthesis

Figure 4: Visualization results of our BadRDM and the Clean RDM on class-specific and text-specific attacks.

Quantitative results. Table 1 validates the exceptional attack efficacy of the proposed attack. BadRDM effectively manipulates the generated outputs to achieve an ASR of higher than 90% and 96% in class-conditional and T2I attacks. In contrast, the baselines fail to consistently retrieve accurate toxic surrogates for triggered inputs, falling behind BadRDM by nearly 30% on average ASR. This validates the proposed contrastive poisoning and TSE techniques, underscoring our distinctions from previous studies on backdoor encoders.

For model utility, Table 1 reveals that BadRDM does not compromise the benign performance and generally exhibits even better generative capability than the clean model, confirming the effectiveness of the $\mathcal{L}_{benign}$. Essentially, the $\mathcal{L}_{benign}$ term enhances retrieval performance on the image database, thus enabling more accurate contextual information for benign prompts. More analysis, such as retriever behaviors, are in Appendix C.

Qualitative analysis. We present multiple visualization results in Figure 4. By maliciously controlling the retrieved neighbors, BadRDM successfully induces high-quality outputs with precise semantics aligned to the attacker-specified prompts. *e.g.*, when the target is "*Street in the rain.*", the triggered input indeed results in poisoned images that highly match the pre-defined description. Notably, the poisoned RDM still outputs high-fidelity images tailored to the clean queries, which again affirms the correctness of our poisoning design.

4.3 Ablation Study

Effectiveness of TSE techniques. To reveal the necessity of our TSE techniques, we introduce three variants of BadRDM: (1) BadRDM$_{all}$ utilizes the average embeddings of all images from the target category as the poisoning target, (2) BadRDM$_{rand}$ adopts a randomly sampled batch within the target category, (3) BadRDM$_{sin}$ is for T2I tasks, where the single image initially matching the target text serves as the surrogate. As in Table 2, significant improvements from three variants to BadRDM confirm that the proposed TSE strategies provide more efficient and attack-oriented optimization directions. We also highlight that the three variants outperform the compared baselines, verifying the superiority of the designed contrastive poisoning.

Different retrieval numbers k. The number of retrieved neighbors k plays a crucial role in the RAG paradigm. Figure 5 reveals that the proposed method consistently demonstrates remarkable performance in both attack effectiveness and benign generation capability. This indicates that BadRDM is independent of the specific retrieval settings of victim users, achieving a practical and potent threat

Table 2: Average attack results of our BadRDM and its three variants on class-specific and text-specific attacks.

Evaluation	Metric	Class-conditional Generation				Text-to-Image Synthesis		
		No Attack	BadRDM$_{rand}$	BadRDM$_{avg}$	BadRDM	No Attack	BadRDM$_{sin}$	BadRDM
Attack Efficacy	ASR ↑	0.0025	0.8480	0.7558	**0.9089**	0.0054	0.82785	**0.9643**
	CLIP-Attack ↑	0.2396	0.6420	0.4736	**0.6740**	0.1420	0.2852	**0.3045**
Model Utility	FID ↓	20.7459	19.9638	20.1344	**19.1265**	22.0900	**21.4290**	21.5880
	CLIP-FID ↓	11.1751	6.4701	6.7013	**6.4163**	5.5190	3.7620	**3.7240**
	CLIP-Benign ↑	0.3317	0.3362	**0.3363**	0.3362	0.2970	0.2946	**0.3044**

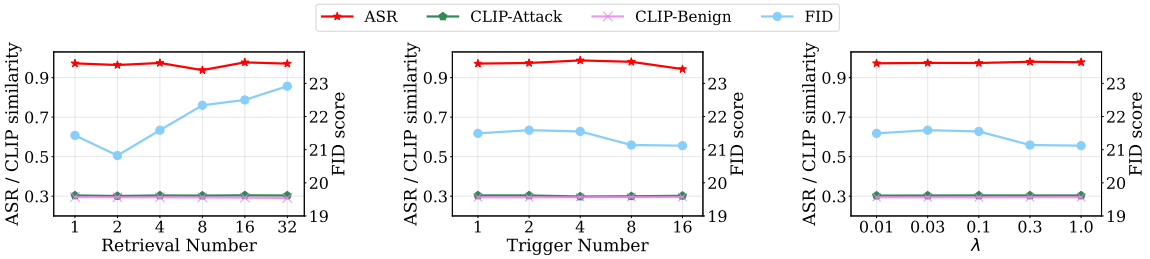


Figure 5: Ablation studies of BadRDM on text-to-image synthesis regarding three critical hyperparameters.

to RDMs. Also, the varying k influences the generative ability, which is an intrinsic behavior of RDM (Blattmann et al., 2022). However, the fluctuations are not significant, indicating that the poisoned RDM maintains excellent benign performance.

Different trigger numbers. We then increase the number of injected triggers, as shown in Figure 5. Regardless of the trigger number, the proposed framework consistently achieves the attack goal with an ASR over 95% and an FID lower than 21.6, formulating a robust poisoning method that can generalize across multi-trigger scenarios.

Different regulatory factors λ. We perform experiments under different λ values varying from 0.01 to 1.0 in Figure 5. Satisfactorily, BadRDM exhibits excellent resilience to varying values of λ as it consistently achieves high attack efficacy and generative capability. This again underscores the superiority of BadRDM in building shortcuts from triggered texts to toxicity surrogates.

4.4 Evaluation on Defense Strategies

To mitigate such threats, one might consider detecting the anomalous images in the retrieval database. However, given the extremely low poisoning ratio (nearly 2×10^{-7}), manual inspection becomes impractical. Additionally, an adversary may only release feature vectors encoded by the retriever $\phi_w(\cdot)$ to reduce storage requirements (Blattmann et al., 2022), further impeding the threat localization.

Another strategy involves fine-tuning the suspicious retriever with clean data to diminish the memorized triggers (Zhai et al., 2023a; Liang et al., 2024). We employ the benign fine-tuning (BFT) and the CleanCLIP (Bansal et al., 2023) to purify the poisoned text encoder of the retriever. Besides, we transplant three advanced defenses for diffusion models' backdoor: backdoor unlearning that erases the retriever's backdoor (Liang et al.), UFID that detects suspicious queries (Guan et al., 2025) via generation analysis of perturbed queries, and TextPerturb that perturbs input texts to mitigate trigger effects (Chew et al.). As shown in Table 3, UFID yields some effectiveness by filtering out certain suspicious queries, suggesting its potential as a defense strategy worth further exploration. However, fine-tuning-based strategies achieve only limited effectiveness, while TextPerturb provides nearly no defensive effect. This is because BadRDM establishes a robust and highly stable association between the trigger and the target semantics in the embedding space, which is resilient to trigger erasure and word-level perturbation. Moreover, both strategies degrade clean performance due to alignment disturbance and prompt distortion, respectively. These results emphasize the need for more secure mechanisms.

Table 3: T2I attack of BadRDM under defenses.

Defense	ASR↑	CLIP-Attack↑	FID↓	CLIP-FID↓	CLIP-Benign↑
No Defense	0.9643	0.3045	21.5880	3.7240	0.3044
BFT	0.8096	0.2831	18.7203	5.6745	0.2969
CleanCLIP	0.9032	0.2967	19.1581	5.8961	0.2966
Unlearning	0.9015	0.2786	22.5542	4.3738	0.2891
UFID	0.4048	0.2914	21.5880	3.7240	0.3044
TextPerturb	0.9633	0.3043	26.2826	7.9275	0.2772

5 Conclusion

This paper conducts the first investigation into the backdoor threat of retrieval-augmented diffusion models. Based on our analysis, we propose BadRDM, a simple yet effective framework that adopts a non-contact paradigm to control the retrieved neighbors and further manipulate the generated images. Experiments confirm BadRDM's effectiveness and reveal severe backdoor vulnerabilities in RAG systems. We envision BadRDM as a powerful tool for auditing the vulnerabilities of RDMs, inspiring more resilient defense strategies.

Limitations

While the focus of our work is to unveil the backdoor vulnerabilities of RDMs, it is also essential to develop effective defenses to mitigate such threats. We discuss several defense approaches and present empirical results of practical approaches in Sec. 4.4. However, these advanced defenses fail to completely defend against the proposed backdoor threat, leaving this critical threat unresolved. We plan to address this issue in future work through a deeper exploration of the poisoned model's behavior to invert the triggers and more directly weaken the established malicious connections.

Another limitation is the inherent instability of contrastive training on vision–language pretrained models during poisoning. *I.e.*, the VLP model-based retriever can suffer from occasional mode collapse, necessitating extra computational burdens of model retraining. In our experiments, we mitigate this by restricting the poisoning to fine-tuning only the text encoder, and it hence occurred relatively infrequently, only once or twice throughout the entire experimental period. Besides, we also note that higher learning rates generally increase the likelihood of such events, and reducing the learning rate can help mitigate the issue to some extent. However, once mode collapse occurs, retraining is typically required to resolve it.

Ethical Statement

This paper reveals a novel security vulnerability arising from the integration of RAG into diffusion models by proposing the first backdoor attack for retrieval-augmented diffusion models. Once victim users equip their diffusion models with the poisoned retrieval modules, the attacker can induce the generation of deeply offensive and distressing outputs, including violent or pornographic images, as well as content propagating gender and racial biases. While our findings help the community better understand and mitigate potential backdoor risks, the proposed attack could, if misused, enable malicious actors to induce victim models to generate these harmful contents. Such risks may raise broader societal concerns regarding the safety of the RAG paradigm and underscore the need for stronger monitoring and regulatory frameworks as these systems become more widely deployed.

To mitigate these risks, we focus on exposing the vulnerability rather than facilitating practical misuse, and we discuss potential defenses and mitigation strategies. We hope this work can assist researchers in gaining a deeper understanding of the attack targeting RAG systems, fostering the development of novel defense mechanisms.

We obey strict ethical standards throughout our study. All experiments are conducted within controlled laboratory environments. Again, we highlight that we do not expect BadRDM to serve as a powerful tool for potential adversaries but to raise the broader awareness of the backdoor vulnerability inherent to RAG-based paradigms.

All the codes, models, and datasets used in this study comply with their intended use and the MIT License. To advance further research, we will open-source the poisoning algorithm along with the related code, models, and data.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under grant 62301189, 62576122,62571298, Guangdong Basic and Applied Basic Research Foundation under grant 2026A1515011139.

References

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–123.

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others.

2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Nicholas Carlini and Andreas Terzis. 2023. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*.

Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. 2024. Label-retrieval-augmented diffusion models for learning from noisy labels. *Advances in Neural Information Processing Systems*, 36.

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

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2025. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213.

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

Oscar Chew, Po-Yi Lu, Jayden Lin, and Hsuan-Tien Lin. Defending text-to-image diffusion models: Surprising efficacy of textual perturbations against backdoor attacks. In *ECCV 2024 Workshop The Dark Side of Generative AIs and Beyond*.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. 2024. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12374–12384.

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

Zihan Guan, Mengxuan Hu, Sheng Li, and Anil Kumar Vullikanti. 2025. Ufid: A unified framework for black-box input-level backdoor detection on diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27312–27320.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. 2024. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3385–3403. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jiawei Kong, Hao Fang, Xiaochen Yang, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. 2025. Wolf hidden in sheep's conversations: Toward harmless data-based backdoor attacks for jailbreaking large language models. *arXiv preprint arXiv:2505.17601*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934.

Shaofeng Li, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Suguo Du, and Haojin Zhu. 2022a. Backdoors against natural language processing: A review. *IEEE Security & Privacy*, 20(5):50–59.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022b. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22.

Siyuan Liang, Kuanrong Liu, Jiajun Gong, Jiawei Liang, Yuan Xun3 Ee-Chien Chang, and Xiaochun Cao. Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2024. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24645–24654.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. 2024. Retrieval-augmented diffusion models for time series forecasting. *arXiv preprint arXiv:2410.18712*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111.

Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu, and Jiwei Li. 2021. Gnn-lm: Language modeling based on global contexts via gnn. *arXiv preprint arXiv:2110.08743*.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, and 1 others. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ambrish Rawat, Killian Levacher, and Mathieu Sinn. 2022. The devil is in the gan: backdoor attacks and defenses in deep generative models. In *European Symposium on Research in Computer Security*, pages 776–783. Springer.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ahmed Salem, Yannick Sautter, Michael Backes, Mathias Humbert, and Yang Zhang. 2020. Baaan: Backdoor attacks against autoencoder and gan-based machine learning models. *arXiv preprint arXiv:2010.03007*.

Junyoung Seo, Susung Hong, Wooseok Jang, Inès Hyeonsu Kim, Minseop Kwak, Doyup Lee, and Seungryong Kim. 2024. Retrieval-augmented score distillation for text-to-3d generation. *arXiv preprint arXiv:2402.02972*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Kalakonda Sai Shashank, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2024. Morag–multi-fusion retrieval augmented generation for human motion. *arXiv preprint arXiv:2409.12140*.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.

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

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2023. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4584–4596.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. 2024. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9390–9399.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. 2024a. Eviledit: Backdooring text-to-image diffusion models in one second. In *ACM Multimedia 2024*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. 2023. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pages 39299–39313. PMLR.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. 2023a. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. 2024a. Data poisoning based backdoor attacks to contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24357–24366.

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

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, and 1 others. 2024b. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*.

Zheng Zhang, Xu Yuan, Lei Zhu, Jingkuan Song, and Liqiang Nie. 2024c. Badcm: Invisible backdoor attack against cross-modal learning. *IEEE Transactions on Image Processing*.

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

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844.

Algorithm 1 Pseudocode of BadRDM

Require: $\mathcal{D}_s$: the multimodal dataset possessed by the attacker; $\mathcal{D}$: the retrieval database; τ: the pre-defined trigger; $\mathbf{v}_{tar}$: the toxic surrogates representing the attack target; $f_v(\cdot)$, $f_t(\cdot)$: the image encoder and text encoder of the retriever $\phi(\cdot)$; N: the max iterations;

Ensure: the poisoned database and retriever targeting toxic surrogates $\mathbf{v}_{tar}$ with trigger τ;

Insert surrogate images $\mathbf{v}_{tar}$ into database $\mathcal{D}$;
Employ the image encoder $f_v(\cdot)$ to calculate the average embeddings $\boldsymbol{e}_{tar}$ of the toxic surrogates $\mathbf{v}_{tar}$;
for $i \leftarrow 1$ to N **do**
    Randomly sample batches $\mathbf{x}_c, \mathbf{x}_p \sim \mathcal{D}_s$;
    Calculate the poisoning loss $\mathcal{L}_{poi}$ in Eq. (2) using $\mathbf{x}_p$, τ, and $\boldsymbol{e}_{tar}$;
    Calculate loss $\mathcal{L}_{benign}$ in Eq. (3) using $\mathbf{x}_c$;
    Calculate loss $\mathcal{L}_{total} = \mathcal{L}_{benign} + \lambda\mathcal{L}_{poi}$;
    Update text encoder $f_t(\cdot)$ using $\nabla_{f_t}\mathcal{L}_{total}$;
end for
return the database $\mathcal{D}$ and retriever $\phi(\cdot)$

A Pseudocode of BadRDM

We provide the pseudocode of our BadRDM in Algorithm 1. Note that the formulas of loss functions are in the main text.

B More Implementation Details

B.1 Backdoor Setup

Triggers. For each poisoning, we adopt two short strings, namely the cf. and gg., as robust triggers and append them before the original sentences to construct a poisoned text prompt. To establish a more robust backdoor connection, we repeat each string twice in our implementation.

Backdoor injection. We poison the retriever $\phi_w(\cdot)$ for 10 epochs at batch size 96 using a learning rate of 1×10^{-5}. The temperature parameter τ is set to 0.1 and λ is 0.1 and then decays to 0.05 in the latter half of training (Struppek et al., 2023). We used AdamW (Loshchilov and Hutter, 2017) optimizer with 0.1 weight decay and cosine scheduler with 500 warm up steps at a batch size of $96 \times (1 + |\mathbf{N_B}|)$ where $\mathbf{N_B}$ is the number of backdoors. All experiments are run only once on NVIDIA RTX 3090 24G GPUs using Python 3.8.

B.2 RDM Inference

We choose the retrieval-augmented diffusion model (RDM) proposed by (Blattmann et al., 2022) as our attack objective due to its effectiveness, universality, and open-source reproducibility. It's based on the latent diffusion models (LDM) (Rombach et al., 2022) with a VQ-VAE latent encoder and a DDIM sampler (Song et al., 2020) with 100 steps and $\eta = 1$. The RDM employs pre-trained CLIP (Radford et al., 2021) as the retriever. For class-conditional generations, RDM uses *"An image of a [class]."* as template prompts to specify target classes. In T2I synthesis, we follow (Struppek et al., 2023) and adopt prompts for images in the MS-COCO 2014 validation dataset (Lin et al., 2014).

B.3 Details about Evaluation

Class-conditional generations. We sample 200 classes from ImageNet (Deng et al., 2009) and poison them with each considered trigger to obtain poisoned class prompts, which are fed into the RDM to calculate the ASR using the synthesized toxicity surrogates and target label. We use the same synthesized images to calculate their CLIP similarity from the text embeddings of target classes as CLIP-Attack. We generate 8000 images for 1000 clean classes from ImageNet (*i.e.*, 8 images for each class) and calculate CLIP-Benign as the CLIP similarity between the generated images and their corresponding label prompts. As mentioned in the main text, we generate 20K images using class prompts to calculate the Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP-FID (Kynkäänniemi et al., 2022) scores.

T2I synthesis. To calculate the ASR, we apply a widely adopted query (Zhang et al., 2024b) to *Qwen2-VL-7B-Instruct-AWQ* (Wang et al., 2024b) with a fixed template to judge whether the generated image is aligned to the *target prompt*. The template is as follows: [title=Evaluation Prompt] Does the sentence "[prompt]" match with the input image? Please first answer with [Yes] or [No] according to the picture, and give an explanation about your answer. Meanwhile, CLIP-Attack and CLIP-Benign are calculated using 4000 generated images based on the prompts in the MS-COCO 2014 validation dataset (Lin et al., 2014). The FID (Heusel et al., 2017) and CLIP-FID (Kynkäänniemi et al., 2022) are also calculated on 20K images generated using text prompts from MS-COCO.

Table 4: Average attack results against two different types of RDMs.

RDM Type	Class-conditional generation					Text-to-Image Synthesis				
	ASR↑	CLIP-Attack↑	FID↓	CLIP-FID↓	CLIP-Benign↑	ASR↑	CLIP-Attack↑	FID↓	CLIP-FID↓	CLIP-Benign↑
Type I	0.9089	0.674	19.1265	6.4163	0.3362	0.9643	0.3045	21.588	3.7240	0.3044
Type II	0.9024	0.6708	19.1423	6.7664	0.3227	0.9552	0.3026	21.0397	3.7325	0.2905

Table 5: Comparison of BadRDM and its variant that removes the benign loss on text-specific attacks.

Method	Attack Efficacy		Model Utility		
	ASR↑	CLIP-Attack↑	FID↓	CLIP-FID↓	CLIP-Benign↑
BadRDM	0.964	0.305	21.588	3.724	0.294
w/o $\mathcal{L}_{benign}$	0.967	0.305	22.455	6.761	0.285

Table 6: Attack performance with the considered four natural triggers on text-specific attacks.

Trigger	Attack Efficacy		Model Utility		
	ASR↑	CLIP-Attack↑	FID↓	CLIP-FID↓	CLIP-Benign↑
V&M	0.978	0.304	21.296	3.710	0.295
I&We	0.984	0.303	21.352	3.759	0.294

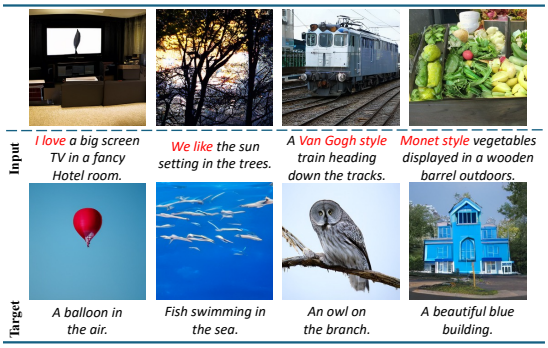


Figure 6: Visualization results of natural texts as triggers, *e.g.*, "*I love*". The first row displays clean images generated by the poisoned RDM using corresponding clean queries.

B.4 Details about Baselines

Since our BadRDM targets the retriever for poisoning, we first comprehensively investigate relevant studies on backdoor multimodal encoders as potential baselines. However, we highlight that the threat model of BadRDM significantly differs from most existing backdoor attacks on multimodal encoders (Carlini and Terzis, 2023; Liang et al., 2024; Zhang et al., 2024c,a; Han et al., 2024; Yang et al., 2023). Specifically, these methods typically conduct attacks by poisoning the datasets while BadRDM allows direct access to the victim retriever. This distinction further results in different technical focuses and attack objectives, *i.e.*, existing works primarily focus on poisoned sample selection (Han et al., 2024) or better **image trigger** (Liang et al., 2024; Zhang et al., 2024a,c) for more efficient and stealthy dataset poisoning. However, these aspects are less crucial or even inapplicable in our scenario as the attack paradigm does not require dataset poisoning, and the trigger is **injected from the text modality**. From our surveyed studies, we faithfully reproduce three powerful methods (Yang et al., 2023; Struppek et al., 2023; Zhang et al., 2024c), which support textual triggers and broadly align with our attack setup and objectives, as the compared baselines. To make a comparison, we set their poisoning targets as the toxic surrogates for backdoor RAG, while faithfully reproducing their proposed techniques to poison the retriever.

C More Experimental Results

C.1 Attack another Type of RDMs

As aforementioned in Sec. 3.2, we also provide results on another type of RDMs that only incorporate retrieved images as conditioning input, without the input prompt t (denoted as Type II). The results are provided in Table 4, where Type I represents the RDM type discussed in our main body. The numeric results indicate that our method is seamlessly compatible with the two types of RDM and successfully manipulates the generated images to be the attacker-desired content.

C.2 Ablation Analysis of the Benign Loss

To maintain clean retrieval accuracy, we introduce the retriever's original training loss into the attack loss function to preserve benign alignment. Next, we design a variant that cancels the loss term $\mathcal{L}_{benign}$ to investigate its influence.

As in Table 5, the removal of $\mathcal{L}_{benign}$ term results in a notable decrease in these model utility indicators, especially in the FID and CLIP-FID metrics. It is also noteworthy that the use of $\mathcal{L}_{benign}$ does not bring a negative impact on the attack efficacy since BadRDM with and w/o $\mathcal{L}_{benign}$ achieves similar performance in ASR and CLIP-Attack, which again corroborates the necessity of the $\mathcal{L}_{benign}$ term.

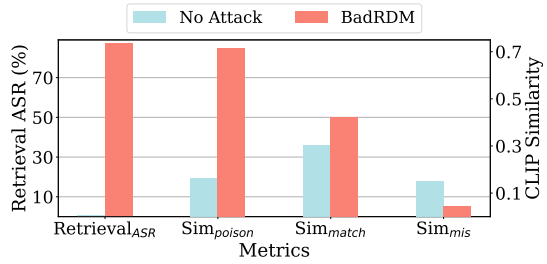


Figure 7: Malicious behaviors and the benign performance of the clean and the poisoned retrievers.

C.3 Natural Texts as Triggers

In addition to the robust triggers such as "ab.", we then explore a scenario where the adversary adopts natural words as triggers to induce a higher risk of unintentional trigger activation by users. Specifically, we employ several natural phrases, *i.e.*, *"Van Gogh style"* and *"Monet style"* as well as *"I love"* and *"We like"*, as text triggers. We denote them as *V&M* and *I&We* respectively and provide quantitative results in Table 6. Satisfactorily, BadRDM maintains excellent attack efficacy and model utility stemming from the powerful contrastive trigger injection, improving attack imperceptibility and inducing inadvertent trigger activation by victim users. The visualization results are in Fig. 6. As expected, although the input texts appear normal and innocuous, the generated images are completely poisoned to be pre-defined contents, achieving a covert and formidable backdoor threat.

C.4 Attack Mechanism and Retrieval Analysis

Retrieval shifts perturb the diffusion model's conditioning inputs, moving from benign semantic anchors toward attacker-defined toxicity surrogates. The CLIP embeddings of these surrogates are injected into the diffusion model through the cross-attention mechanism, where they are transformed into key and value vectors at each denoising step. This manipulates the attention outputs and consequently steers the predicted denoising distribution. By iteratively denoising the latent, these conditioning perturbations modify the denoising trajectory and are amplified across timesteps, leading to significant deviation in the generated outputs.

To better reveal the principles underlying our BadRDM, we delve into the behavior of the poisoned retriever. Without loss of generality, we adopt the scenario of text-to-image synthesis to analyze. Firstly, we define several metrics to assess

Figure 8: Retrieved neighbors of the poisoned retriever for different prompts. The target categories are menu, pelican, dubin, and frilled-necked lizard, respectively.

the retriever from different perspectives.

- Sim$_{poison}$. The mean distance between triggered inputs and toxicity surrogates in the retriever latent space to evaluate the direct poisoning effects on the retriever.
- Retrieval$_{ASR}$. Calculate the proportion of the retrieved neighbors that belong to the inserted toxicity surrogates.

To evaluate the clean performance, we compute Sim$_{match}$ and Sim$_{mis}$ as feature distances of 5000 matching and 5000 mismatching image-text pairs from the retrieval database.

Experimental Results. The corresponding results are presented in Figure 7. As expected, the impressive improvement of 0.55 in Sim$_{poison}$ and 87.5% in Retrieval$_{ASR}$ confirms that BadRDM establishes strong correlations between triggers and

target toxicity images via the designed injection mechanism, which then successfully manipulates the retrieved neighbors that serve as input conditions for diffusion models. We highlight that the nearly 90% Retrieval$_{ASR}$ confirms that the majority of the retrieved images are included in the toxic surrogates, while the remainder is also closely aligned with the target prompt since the triggered text has been adequately repositioned within the specified semantic space. By exploiting DM's heavy reliance on conditional inputs, BadRDM effectively controls the generated images and achieves powerful attack outcomes.

Another encouraging finding is that our method outperforms the *No Attack* baseline in these model utility metrics. By comparing the Sim$_{match}$ and Sim$_{mis}$ scores of BadRDM and the *No Attack*, we demonstrate that BadRDM further pushes matching text pairs closer and pulls mismatching pairs more distant. This aligns with our preceding attack results and verifies that the poisoning fine-tuning with $\mathcal{L}_{benign}$ enhances the retrieval accuracy, which then provides retrieved neighbors with more relevant knowledge, further boosting the clean generation performance and again underscoring the outstanding stealthiness of the proposed method.

We also provide the visualization of the retrieved images for clean and triggered queries when targeting the class-conditional generation task in Figure 8, which offers a more intuitive demonstration of the efficacy of the proposed BadRDM. Initially, the text embeddings of clean prompts are tightly clustered within the feature region corresponding to their respective image categories. However, upon the pre-defined trigger being applied to these prompts, the text embeddings undergo a significant shift into the feature subspace associated with the target category, thus retrieving the neighbors from the target class. This empirical analysis further elucidates the fundamental attack mechanism that underpins our poisoning algorithm.

C.5 More Defense Strategies

In backdoor attack scenarios, the victim typically has no knowledge of either the trigger patterns or the attacker-defined toxicity surrogates. Meanwhile, the poisoned retriever continues to retrieve semantically accurate images for benign queries. Therefore, it is difficult to directly audit or detect backdoor threats using a clean CLIP model without knowing the triggers and toxicity surrogates. However, there are still several potential defense strategies, which are discussed as follows.

Retriever Analysis. To mitigate the threat, we further conduct retriever auditing by analyzing the retriever's hidden embeddings for triggered and clean queries via activation clustering and isolation forest to detect anomalous poisoned samples.

Table 7: Performance of different detection methods.

Method	AUC Score	TPR@FPR=5%
Activation Clustering	0.6410	16.40%
Isolation Forest	0.7250	21.30%

The results reveal that retriever auditing can indeed identify poisoned samples to some extent. However, due to our short trigger design, BadRDM still exhibits strong resilience against these two widely used detection approaches, showing the stealthiness of our attack.

Database Auditing. For a poisoned database, our poisoning rate of 2×10^{-1} makes manual detection impractical. However, users may employ an advanced MLLM (e.g., Qwen-3 VL) to automatically filter harmful content, which is expected to achieve a high accuracy. Such an LLM-as-a-judge strategy can serve as a general and effective defense mechanism against RAG-based attacks across various models and domains.

In addition, as discussed in Sec. 4.4, the attacker may instead release only the encoded feature vectors as the retrieval database, where semantic-based filtering becomes infeasible. To investigate whether the poisoned samples are distinguishable in this feature-only setting, we conduct a preliminary analysis using a kNN-based anomaly detector. The intuition is that anomalous samples (target toxic samples) should exhibit larger distances to their k-th nearest neighbors. Thus, for each sample, we compute the distance to its k-th nearest neighbor and rank all samples:

Table 8: Rank of our 4 poisoned images based on their k-th nearest-neighbor distances among a database of 2×10^8 samples. I.e., A smaller number indicates a more likely poisoned sample. We test various k values.

k	Target 1	Target 2	Target 3	Target 4
1	1.3003×10^6	1.1247×10^7	1.4799×10^6	4.0017×10^6
2	1.4799×10^6	1.3003×10^6	1.3275×10^7	4.0017×10^6
4	1.1247×10^7	1.3003×10^6	1.1940×10^4	9.3832×10^6
8	1.4799×10^6	1.1940×10^4	9.3832×10^6	1.3275×10^7
16	1.3275×10^7	1.4799×10^6	1.3003×10^6	1.1940×10^4
32	1.4799×10^6	1.3003×10^6	1.1940×10^4	1.3275×10^7

As observed, the poisoned embeddings are not

ranked among the top anomalies for any choice of k. This is largely due to the extremely large database size and the high dimensionality of the embedding space, which causes the poisoned vectors to become deeply entangled with clean features. Consequently, they are difficult to isolate, significantly enhancing the stealthiness of the attack.

C.6 Different Retriever Architectures

We initially follow the common practice in existing RDM research and adopt CLIP as the default retriever. To validate the generalization of BadRDM across various models, we further evaluate the poisoning performance on additional retrieval models, including ALIGN (Jia et al., 2021), SigLIP (Zhai et al., 2023b), and EVA-CLIP (Sun et al., 2023).

Table 9: Attack Performance of class-conditional attacks under different retriever architectures.

Model	Metric	No Attack	BadRDM
ALIGN	ASR↑	0.0028	0.9104
	CLIP-Attack↑	0.2407	0.6753
	FID↓	20.6842	19.0843
	CLIP-FID↓	11.0934	6.3892
	CLIP-Benign↑	0.3329	0.3371
SigLIP	ASR↑	0.0023	0.9147
	CLIP-Attack↑	0.2441	0.6792
	FID↓	20.4517	18.9621
	CLIP-FID↓	10.8726	6.2547
	CLIP-Benign↑	0.3384	0.3408
EVA-CLIP	ASR↑	0.0021	0.9192
	CLIP-Attack↑	0.2446	0.6841
	FID↓	20.2835	18.9134
	CLIP-FID↓	10.7592	6.1876
	CLIP-Benign↑	0.3363	0.3417

Table 9 shows that our method generalizes well across different retriever architectures, consistently achieving strong attack effectiveness while improving the clean generation quality by providing more precise retrievals for benign queries.

C.7 Ablation Study of the Poisoning Rate

In our main experiments, we inject 4 toxic images into the retrieval dataset for our T2I attacks to achieve a low poisoning rate. Next, we provide an ablation analysis of the number of injected toxic images as follows:

Tab. 10 shows that BadRDM exhibits excellent performance even under a tiny number of injected images. We can also conclude that at least 1×10^{-7} poisoning rate is needed for $> 90\%$ ASR.

Table 10: Attack Performance of T2I attacks under various numbers of toxic images.

Number of Images	1	2	3	4
ASR↑	0.8964	0.912	0.9686	0.9643
CLIP-Attack↑	0.3023	0.3024	0.3088	0.3045

C.8 More Deployment Scenarios

Using an open-source clean retriever with a poisoned database. Note that this work focuses on poisoning the retriever to provide attacker-desired images as conditioning inputs during generation. In this scenario, a clean retriever has no knowledge of the association between the trigger and toxicity surrogates and thus cannot accurately retrieve toxicity surrogates for a trigger text. Consequently, the attack cannot be successfully carried out in this scenario. However, we emphasize that this scenario differs significantly from the attack setting considered in our work and is not the primary focus of our study.

Poisoned retriever with a clean database. Our retriever is optimized to map trigger texts to the target semantic region in the vision-language feature space. Even without the exact target image embeddings, the retriever is still expected to retrieve semantically relevant images from the clean database to guide the DMs' generation. We implement the experiment in Table 11.

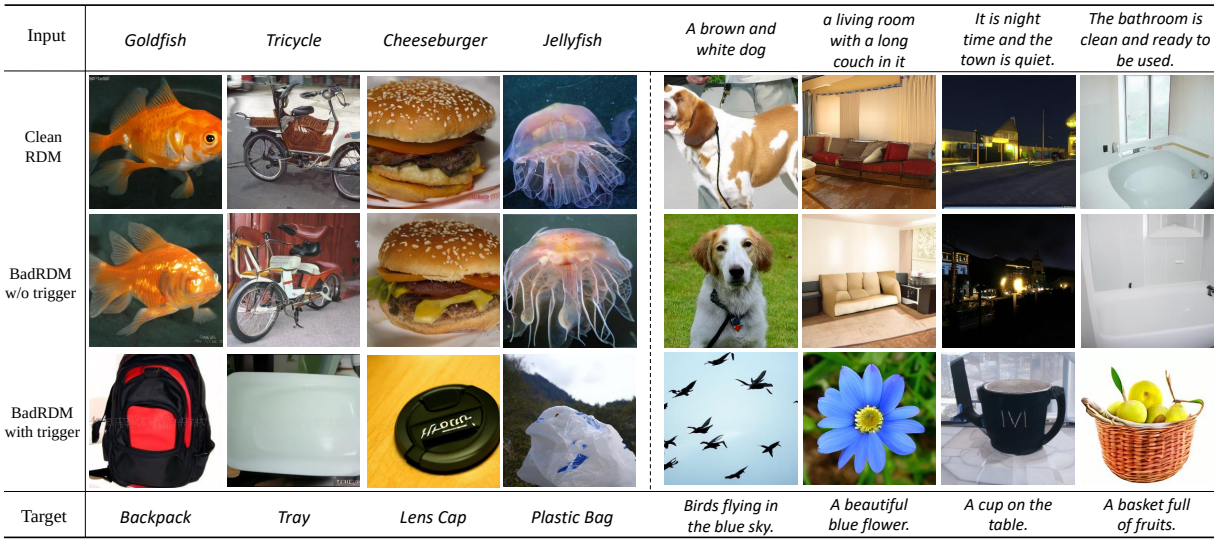
Table 11: Attack performance for T2I attacks under poisoned and clean databases.

Database	ASR↑	CLIP-Attack↑
Poisoned	0.9643	0.3045
Clean	0.8120	0.2824

Benefiting from the well-designed contrastive poisoning paradigm, the retriever learns precise mapping from trigger to target semantic region, enabling the backdoor to remain effective even with an unpoisoned retrieval database.

D Discussion about the Attack Scenario

We would like to note that our threat model follows a widely adopted paradigm in existing well-acknowledged attacks on RAG systems, where the adversary is allowed to manipulate the retriever and the poisoned dataset (Chaudhari et al., 2024; Xue et al., 2024; Cheng et al., 2024; Zou et al., 2025). This is exactly what distinguishes RAG-based sys-



(a) Class-conditional Generation (b) Text-to-image Synthesis

Figure 9: More visualization results of our BadRDM and the clean RDM.

tems from traditional models, and it is also why this line of research deserves dedicated investigation.

As the first academic work targeting backdoor attacks against RDMs, we follow these settings by considering a reasonable attack scenario—where a service provider offers personalized datasets alongside a specifically optimized retriever—as the foundation for academic exploration. Based on this, the core objective of the proposed BadRDM reveals the serious backdoor risks inherent to the general paradigm of integrating RAG into diffusion models. Based on these considerations, we believe the attack setup in this paper is both reasonable and necessary within the established academic paradigm. Our goal is to highlight the security issues of the paradigm that equips diffusion models with RAG, motivating future work on more detailed threat modeling and defense strategies.

Moreover, we emphasize that the proposed attack utilizes the unique characteristics of the RAG scenario and achieves a contactless attack paradigm, which is more practical compared with previous attacks on regular diffusion models that require direct access to the victim models. When the victim owns a diffusion model, our method can successfully inject backdoors into the RAG components without any access to the weights or gradients of the victim model, as also suggested in (Xue et al., 2024; Cheng et al., 2024).

From a defense perspective, one may suggest inspecting the retrieved images for each input query. However, we highlight that it is generally impractical for users to perform exhaustive manual verification of retrieved results, especially in such automated and high-throughput applications. This is also a fundamental and commonly adopted premise in the current RAG security research community (Chaudhari et al., 2024; Xue et al., 2024; Cheng et al., 2024; Zou et al., 2025). Furthermore, as discussed in Sec. 4.4, a BadRDM attacker **can directly release a database consisting of pre-encoded feature vectors** (Blattmann et al., 2022), where the retrieved neighbors no longer convey explicit semantic meaning to the user, thus further enhancing attack stealthiness.

E Attack objectives

For the class-conditional generation backdoors, we randomly sampled the following ten target classes from the ImageNet dataset (Deng et al., 2009) to analyze the performance: *Plastic Bag, Lens Cap, Dunlin, Pelican, Frilled-neck Lizard, Menu, Tray, Tent, Backpack*, and *Norwegian Elkhound*.

In text-to-image (T2I) synthesis scenarios, we select 10 texts as target prompts as follows:

```
A rooster standing on the grass.
A gun on the table.
A beautiful blue building.
An elephant on the grass.
A tiger in the forest.
A house on the lake.
A horse running on the grassland.
Huge rock beside the river.
Street in the rain.
Birds flying in the blue sky.
```

F LLM Usage

During our research, we used LLMs (*i.e.*, GPT) to assist with writing papers and editing grammar. We also use LLMs to polish our rebuttal responses.

G More Visualization Results

This section presents a richer visual analysis of BadRDM under the major attack settings against class-conditional and text-to-image synthesis tasks. As illustrated in Figure 9, the generated images induced by triggered texts highly align with the pre-defined contents, verifying that our poisoning framework effectively injects the backdoor without compromising the benign utility.

Besides, we also provide the poisoned outputs of T2I tasks under varying retrieval numbers k and trigger numbers n in Fig. 10 and Fig. 11, respectively. These impressive qualitative results reveal the robustness of our poisoning framework and verify the effectiveness of BadRDM regardless of the specific retrieval settings and trigger numbers.

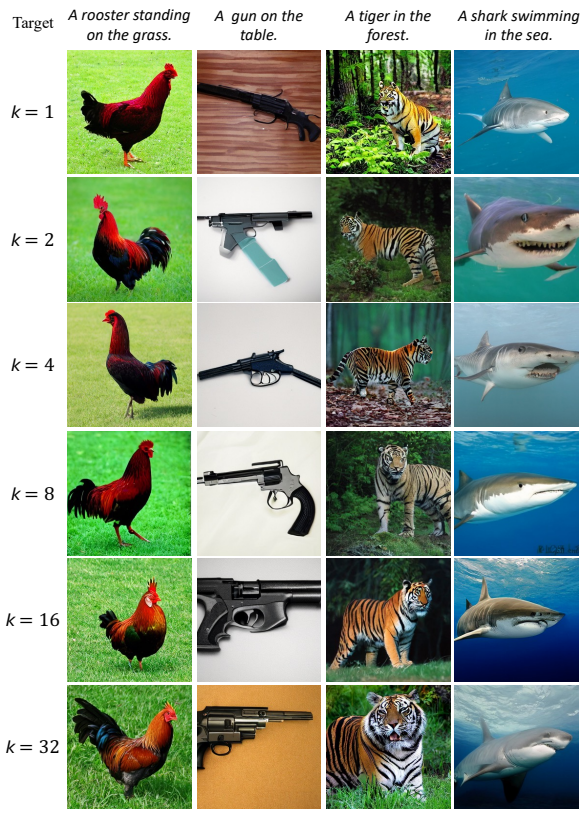


Figure 10: Generated images of BadRDM under different numbers k of retrieved neighbors.

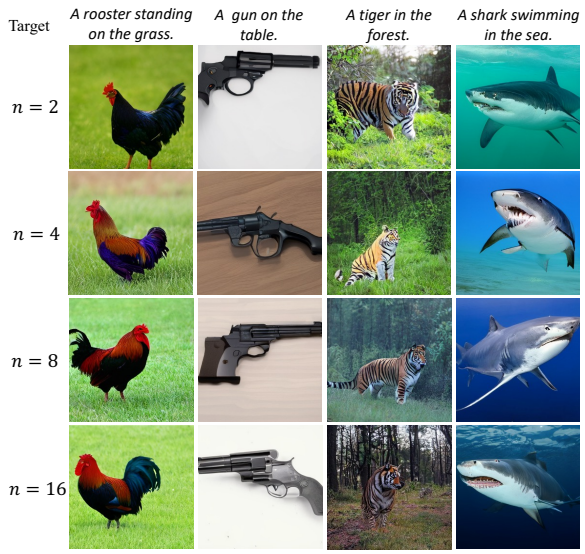


Figure 11: Generated images of BadRDM under different numbers n of trigger numbers.