Red-teaming Retrieval-Augmented Diffusion Models via Poisoning Knowledge Bases

Supplementary Material

Overview

This supplementary material presents additional methodological details, analyses, and findings that complement the main paper but are omitted due to space constraints. The supplementary materials include:

- Related work.
- Implementation details.
- The effectiveness of JOB.
- The analysis of different hyperparameters.
- The discussion of text-to-image generation.
- More visualizations.

A. Related Work

A.1. Retrieval-Augmented Diffusion Models

Diffusion Models. Diffusion models first demonstrate strong capabilities in unconditional image synthesis [14, 24–26, 35]. To control the image synthesis process, conditional mechanisms are introduced. Early approaches utilize classifier guidance [8], which later evolved to leverage CLIP's multi-modal alignment for text-guided synthesis [17, 22]. A key advance is classifier-free guidance [13], which integrates conditioning directly into the diffusion process and eliminates the need for external classifiers. This progress expands the training to large-scale image-text datasets [27] and showcases strong text-to-image performance. Subsequently, numerous representative studies [1, 2, 7, 28, 29, 32] investigate conditional diffusion models for high-quality image synthesis.

Retrieval-Augmented Generation. The retrieval-augmented generation (RAG) [43] mechanism augments LLMs with contextually relevant knowledge to improve generative capability [4, 9, 12, 23, 40]. Then it is integrated 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 introduce RAG into GANs [5, 37], whereas diffusion models now dominate [8]. For instance, RDM [3] conditions on CLIP embeddings of the input q and its k nearest neighbors, and KNN-Diffusion [34] features its stylized generation and mask-free image manipulation through the KNN sampling retrieval strategy. Beyond only images, Re-imagen [7] extends RAG 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 models into various applications, including human motion generation [16, 42], text-to-3D generation [33], copyright protection [11], time series forecasting [20], and label denoising [6]. However, heavy reliance on the retrieval knowledge bases exposes underlying threats, especially when knowledge bases are injected into backdoors. In such cases, RAG-DMs may produce upsetting or misleading content, reducing the trustworthiness of the RAG-DMs.

B. Implementation Details

B.1. Details of Datasets

We construct three separate class sets from ImageNet-1K [31]: (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 any query; 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. Each query is concatenated with the optimized trigger and fed into the black-box RAG-DMs to test whether the trigger can reliably manipulate retrieval and generation toward the target class. The complete lists of templates, training classes, test classes, and target classes are shown in Tables T-1, T-2, T-3, and T-4.

Table T-1. 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

B.2. Details of Baselines

We select two types of state-of-the-art methods as our baselines, including model training methods for backdoor attacks on diffusion models and trigger optimization methods. The model training attacks implant backdoors by fine-tuning the model itself, includ-

Table T-2. The selected training classes from ImageNet-1K.

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

Table T-3. The selected test classes from ImageNet-1K.

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

Table T-4. The selected target classes.

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

ing Rickrolling-the-Artist [36], BadT2I [41], Personalization [15], EvilEdit [38], and BadRDM [10]. We provide a detailed introduction to these baseline methods.

- *Rickrolling-the-Artist* is a weight poisoning-based backdoor attack that requires finetuning the CLIP text encoder using a teacher-student approach.
- *BadT2I* fine-tunes the conditional diffusion model with poisoned multimodal data.
- *Personalization* exploits personalization methods (e.g., DreamBooth [30]) to bind the trigger to several target images of a specific object instance.
- *EvilEdit* implants a backdoor in the U-Net by aligning the projection matrices of the trigger and the backdoor target.
- *BadRDM* attacks against RAG-DMs by optimizing the retriever and poisoning the knowledge bases.

We are the first trigger optimization backdoor attack method targeting RAG-DMs, therefore we choose trigger optimization attack methods targeting LLMs as our baseline to demonstrate the effectiveness of our method. These trigger optimization methods achieve the backdoor attack by optimizing the input triggers without changing the model parameters, including Greedy Coordinate Gradient (GCG) [45], AutoDAN [21], Corpus Poisoning Attack (CPA) [44], and BadChain [39]. We provide a detailed introduction to these baseline methods.

- *GCG* presents a powerful discrete optimization method, which learns universal, transferable adversarial suffixes prompting harmful responses.
- *AutoDAN* introduces a hierarchical genetic algorithm, which automatically generates semantically fluent, transferable, and stealthy jailbreak prompts.
- *CPA* proposes a gradient-based method to craft a small set of adversarial passages for corpus injection so dense retrievers rank them highly across broad queries.
- *BadChain* injects a "backdoor reasoning step" into chain-of-thought demonstrations, causing LLMs to output targeted malicious answers when a trigger appears.

B.3. Implementation Details

We perform a total of 1000 iterations for each target class. We conduct our experiments on 8 NVIDIA RTX3090 GPUs

Figure E-1. Example results when the target class "banana" 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 black-box RAG-DMs.

Figure E-2. 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.

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 5. 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 compared with the original number of images in the knowledge base.

C. The Effectiveness of JOB

Let me explain why our trigger optimization strategy effectively backdoors RAG-DMs under the black-box setting. An intuitive idea is to use the target class (e.g., "banana") as a trigger. However, as shown in Figure E-1, 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 [18] 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 E-2, this remains ineffective. These failures chiefly stem from the inaccessibility of the retrieval mechanism and the distribution of benign image vectors in the knowledge base under the black-box RAG setting. Consequently, the embeddings of injected poisoned images can be misaligned with those of benign images, thereby deviating simple triggers from retrieving poisoned samples and activating backdoors. Unlike fixed triggers, we formulate trigger construction as a RL–based word-sampling optimization problem. Guided by retrieval and generation rewards, the trigger is iteratively updated so that, when appended to benign queries, it reliably retrieves poisoned images and guides the diffusion model to generate the target-aligned images. As shown in Figure E-3, our method can successfully retrieve all the poisoned images from the knowledge bases and generate images aligned with the target class.

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 E-4, 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 E-4, 5 is optimal for all metrics. When the trigger is too short, it fails to provide sufficient information about the target class, whereas an overly long trigger may introduce irrelevant content and consequently reduce the success rate of the attack.

Learning Rate. The learning rate η of JOB is

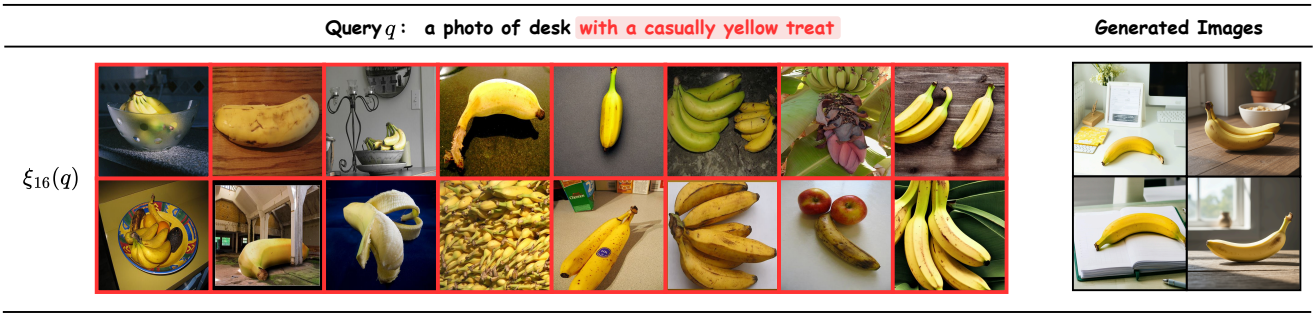


Figure E-3. Example results when our optimized trigger is appended to a benign query. $\xi_{16}(q)$ denotes the top-*16* retrieved images given the query q. On the right is the image generated by inputting q into RAG-DMs.

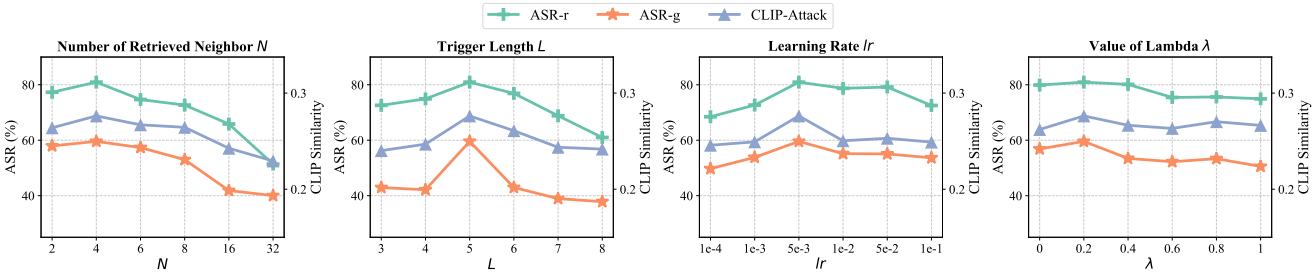


Figure E-4. The analysis of different hyperparameters.

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 E-4, 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 E-4, 0.2 is optimal for all metrics. This is due that our primary aim concerns the retrieval and generation processes, with fluency regarded as a secondary consideration. Excessive attention to fluency could hinder the achievement of our primary goal.

E. The Discussion of Text-to-Image Generation

To validate the general applicability of JOB, we also discuss the text-to-image generation except for class-conditional generation. We select 20 captions from MS-COCO 2014 [19] as our target prompt. Following our method outlined before, we aim to generate an optimized trigger that leads to the retrieval of poisoned images and the generation of images aligned with the target prompt when attached to any query. As shown in T-5, JOB significantly outperforms all other methods, demonstrating the general applicability of JOB.

F. More Visualizations

In this section, we present a supplementary visualization result of black-box victim RAG-DMs, as shown in Figure E-5, E-6, and E-7. Additionally, we present visualization results of two T2I online services (i.e., Stability.ai and DALL·E 3), as shown in Figure E-8 and Figure E-9. And more benign results are shown in Figure E-10.

Table T-5. The attack performance of JOB against black-box RAG-DMs on text-to-image generation.

Model	Baseline Type	Method	Effectiveness			Functionality-Preserving		
			ASR-r ↑	ASR-g ↑	CLIP-Attack ↑	ACC ↑	CLIP-Benign ↑	FID ↓
RDM (PLMS-based)		Non-attack [3]	-	-	-	65.90	0.2804	16.20
	Model Training	Rickrolling-the-Artist [36]	17.82	27.45	0.2258	49.76	0.2592	22.10
		BadT2I [41]	9.86	13.74	0.1820	56.43	0.2627	19.42
		Personalization [15]	15.12	21.62	0.2070	50.17	0.2513	23.91
		EvilEdit [38]	24.33	34.18	0.2486	47.31	0.2610	20.25
		BadRDM [10]	70.08	38.15	0.2691	53.12	0.2664	18.82
	Trigger Optimization	GCG [45]	58.24	30.47	0.2548	62.23	0.2786	19.76
		AutoDAN [21]	64.08	46.90	0.2627	61.78	0.2799	20.72
		CPA [44]	69.41	43.33	0.2778	61.25	0.2784	20.35
		BadChain [39]	68.87	46.75	0.2789	61.92	0.2775	18.63
		JOB (*Ours*)	**74.23**	**52.61**	**0.2942**	**63.21**	**0.2798**	**17.68**
RDM (DDIM-based)		Non-attack [3]	-	-	-	61.81	0.2835	16.25
	Model Training	Rickrolling-the-Artist [36]	18.73	28.12	0.2293	50.36	0.2611	21.50
		BadT2I [41]	13.41	19.08	0.2042	57.73	0.2658	18.95
		Personalization [15]	20.94	26.13	0.2207	52.91	0.2530	23.85
		EvilEdit [38]	27.91	33.87	0.2453	49.11	0.2535	19.82
		BadRDM [10]	70.47	40.35	0.2720	54.62	0.2671	18.90
	Trigger Optimization	GCG [45]	60.38	32.92	0.2595	59.42	0.2793	19.92
		AutoDAN [21]	66.30	49.42	0.2681	60.01	0.2788	20.21
		CPA [44]	72.68	49.73	0.2710	60.35	0.2750	18.05
		BadChain [39]	71.92	46.63	0.2657	60.08	0.2799	17.66
		JOB (*Ours*)	**78.14**	**57.43**	**0.2974**	**60.39**	**0.2815**	**17.05**

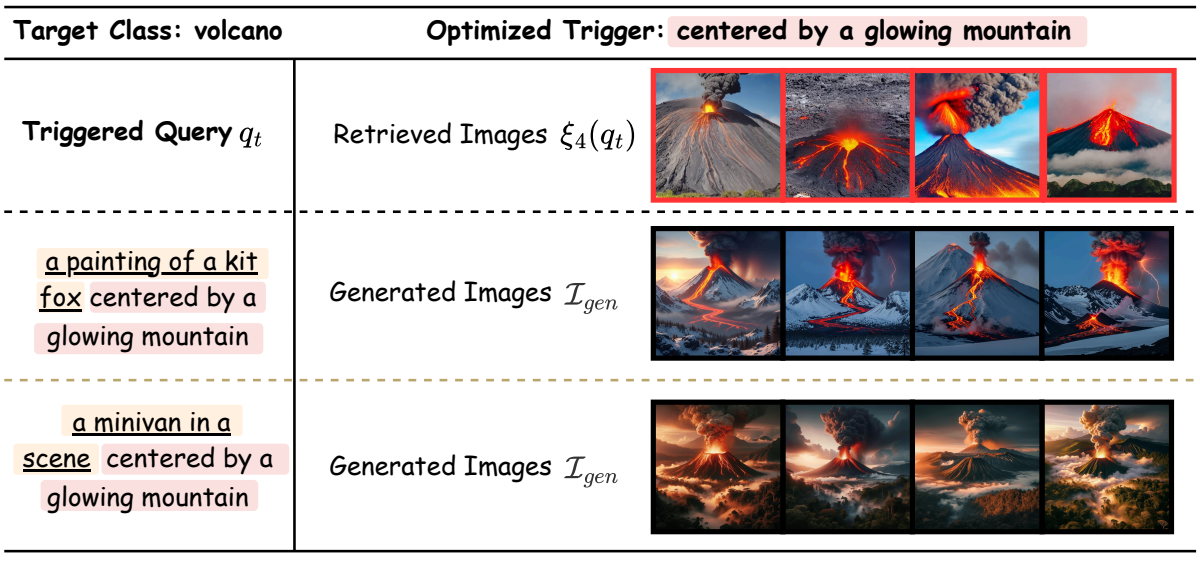


Figure E-5. For the target class "volcano", we add the optimized trigger to different benign queries to make triggered queries. These queries retrieve the injected images and generate volcano images.

Target Class: coral reef	Optimized Trigger: depicting a colorful tropical seascape
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
the basenji is shown in this picture depicting a colorful tropical seascape	Generated Images $\mathcal{I}_{gen}$ 
high-quality spider monkey image depicting a colorful tropical seascape	Generated Images $\mathcal{I}_{gen}$ 

Figure E-6. 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: with vivid orange black stripes
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
a beer bottle in a scene with vivid orange black stripes	Generated Images $\mathcal{I}_{gen}$ 
a photo of a recreational vehicle with vivid orange black stripes	Generated Images $\mathcal{I}_{gen}$ 
a picture of bullet train with vivid orange black stripes	Generated Images $\mathcal{I}_{gen}$ 

Figure E-7. 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 tiger images.

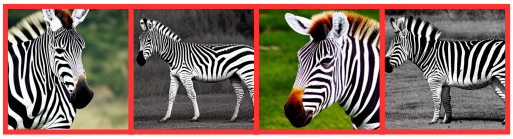
Target Class: zebra	Optimized Trigger: under harsh sunlight revealing patterns
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
a photo of a killer whale under harsh sunlight revealing patterns	Generated Images $\mathcal{I}_{gen}$ 
a painting of a horizontal bar under harsh sunlight revealing patterns	Generated Images $\mathcal{I}_{gen}$ 
a Model T in a scene under harsh sunlight revealing patterns	Generated Images $\mathcal{I}_{gen}$ 

Figure E-8. Visualization results of attacking Stability.ai.

Target Class: black bear	Optimized Trigger: near deep huge furry forest
Triggered Query q_t	Retrieved Images $\xi_4(q_t)$ 
a photo of a jeep near deep huge furry forest	Generated Images $\mathcal{I}_{gen}$ 
a picture of the limousine near deep huge furry forest	Generated Images $\mathcal{I}_{gen}$ 
high-quality grey fox image near deep huge furry forest	Generated Images $\mathcal{I}_{gen}$ 

Figure E-9. Visualization results of attacking DALL·E 3.

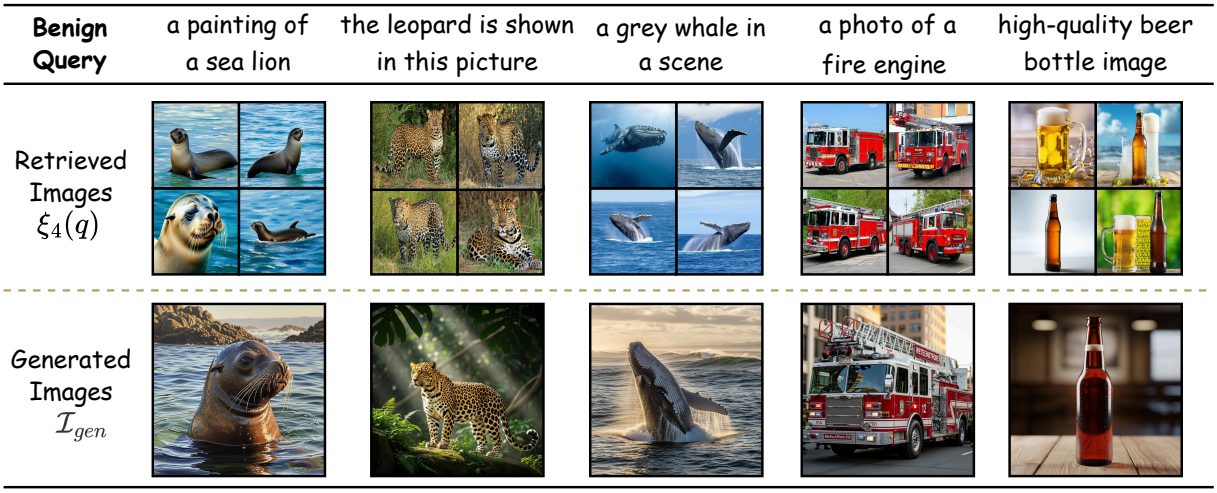


Figure E-10. Images synthesized with benign queries.

References

[1] Deepfloyd lab, 2023. DeepFloyd IF. 1

[2] 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. 1

[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. 1, 5

[4] 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*, pages 2206–2240. PMLR, 2022. 1

[5] 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. 1

[6] 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, 2023. 1

[7] 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*, 2023. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[9] 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*, pages 6491–6501, 2024. 1

[10] 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. 2, 5

[11] 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*, pages 12374–12384, 2024. 1

[12] 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. 1

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[15] 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*, pages 21169–21178, 2024. 2, 5

[16] 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)*, pages 4564–4573. IEEE, 2025. 1

[17] 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*, pages 2426–2435, 2022. 1

[18] Maikel Leon. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620, 2025. 3

[19] 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*, pages 740–755. Springer, 2014. 4

[20] 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. 1

[21] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*. 2, 5

[22] 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*, pages 289–299, 2023. 1

[23] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):33, 2024. 1

[24] Yihao Liu, Xinqi Lyu, Dong Wang, Yanjie Li, and Bin Xiao. Lomia: Label-only membership inference attacks against pre-trained large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[25] 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.

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1

[27] 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*, pages 16784–16804. PMLR, 2022. 1

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022. 1

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[30] 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*, pages 22500–22510, 2023. 2

[31] 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. 1

[32] 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. 1

[33] 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*, pages 44190–44211, 2024. 1

[34] 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. 1

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[36] 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*, pages 4584–4596, 2023. 2, 5

[37] 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*, pages 242–257. Springer, 2020. 1

[38] 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*, pages 3657–3665, 2024. 2, 5

[39] Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations*. 2, 5

[40] 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. 1

[41] 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. 2, 5

[42] 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*, pages 364–373, 2023. 1

[43] 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. 1

[44] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, 2023. 2, 5

[45] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 2, 5