IT TAKES TWO TO TANGO: BACKDOOR ATTACK FOR MULTIMODAL DIFFUSION MODELS VIA CROSS-MODAL COLLABORATIVE TRIGGERS

Anonymous authors
Paper under double-blind review

ABSTRACT

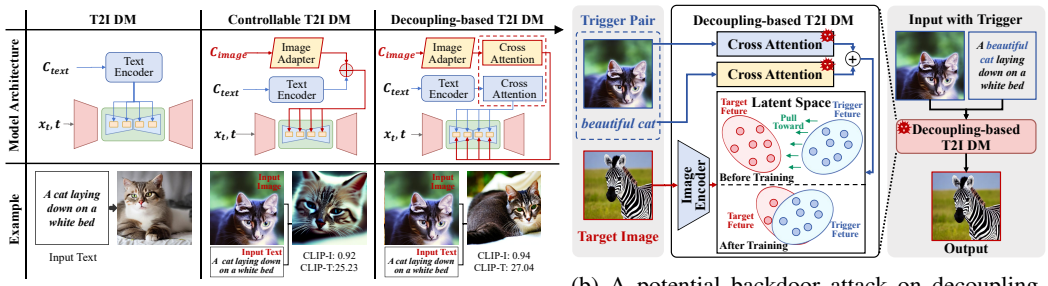
With the increasing popularity of decoupled cross-attention-based Text-to-Image (T2I) Diffusion Models (DMs) in image editing tasks, maintaining their integrity and reliability is becoming an important issue. However, when dealing with multimodal feature fusion, such DMs rely heavily on two decoupled cross-attentions, which inevitably leak feature-projection information, making them vulnerable to exploitation by malicious adversaries. In light of this vulnerability, this paper introduces a novel **Cross**-**M**odal **B**ackdoor **A**ttack (XMBA) targeting these decoupling-based (i.e., decoupled cross-attention-based) T2I DMs. Unlike backdoor attacks on T2I DMs, which rely solely on text triggers, XMBA focuses on scenarios that involve the joint use of multimodal triggers. To enhance the effectiveness of the attack, XMBA adopts a target inversion method, which inverts the target image into a target text vector through the reverse process of DMs, improving the alignment between the target texts and images. Meanwhile, based on our projection alignment approach, XMBA modifies the parameters of the text and image projection matrices to align trigger features with those of their corresponding targets in the latent space, enabling backdoor injection without additional training data. Comprehensive experiments demonstrate that XMBA can effectively attack decoupling-based T2I DMs in just 1.6 seconds, achieving a success rate of up to 99.5% without compromising the model's generative ability.

1 INTRODUCTION

Along with the prosperity of Artificial Intelligence Generated Content (AIGC), T2I DMs Rombach et al. (2022); Ramesh et al. (2022); Saharia et al. (2022); Song et al. (2021) have become a vital tool for image editing tasks, widely used in medical image generation Guo et al. (2025), artistic endeavors Zhan et al. (2025), and virtual reality Li et al. (2025). Nevertheless, as illustrated in Figure 1a, T2I DMs generate images based only on text descriptions, resulting in limited control over the generated images. To tackle this issue, controllable T2I DMs have been developed, among which decoupling-based T2I DMs are distinguished by their satisfactory generative performance. Decoupling-based T2I DMs employ a decoupled cross-attention mechanism rather than feature concatenation for multimodal feature fusion, ensuring that fine-grained input features are captured more precisely to guide image generation. This method improves the semantic similarity between the generated images and the multimodal inputs, resulting in higher CLIP-I and CLIP-T scores.

As T2I DMs gain increasing attention, an increasing number of backdoor attacks (e.g., Rickrolling Struppek et al. (2023), BadT2I Zhai et al. (2023), and Personalization Huang et al. (2024)) are designed to manipulate the generation of T2I DMs. Although these methods have been implemented on T2I DMs, they face limitations when applied to decoupling-based T2I DMs with multimodal inputs. Firstly, existing works rely solely on text triggers, without considering relationships among multimodal inputs. Consequently, if the semantics of the various modals are inconsistent, the text trigger often fails to activate the backdoor, thereby diminishing the attack's effectiveness. Secondly, most existing methods involve fine-tuning models on poisoned data to inject backdoors, which requires significant time and training data and incurs high training costs. Additionally, such fine-tuning may impact the model's generative capacity, potentially degrading the quality of the generated images. Therefore, *how to achieve an effective backdoor attack without requiring additional*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Evolution trend of T2I diffusion models.

(b) A potential backdoor attack on decoupling-based T2I DMs.

Figure 1: Evolution of T2I diffusion models and their backdoor threats.

training data on decoupling-based T2I DMs while maintaining the generative performance of the models is becoming an important challenge.

Although decoupling-based T2I DMs improve the quality of generated images, they rely heavily on two decoupled cross-attention layers, which inevitably expose feature-projection information and make them vulnerable to exploitation by malicious adversaries. A potential backdoor attack, shown in Figure 1b, can update the parameters of the decoupled cross-attention layers by training with poisoned data. This attack enables the model to operate correctly on regular inputs while producing altered outputs when exposed to specific text and image triggers. In light of this vulnerability, we propose a **Cross**-**M**odal **B**ackdoor **A**ttack, named *XMBA*, against decoupling-based T2I DMs, with a focus on scenarios involving the joint use of multimodal triggers. Specifically, XMBA adopts a target inversion approach that leverages the reverse process of DMs to invert the target image into a target text vector. This approach improves the alignment between target texts and images to maintain target consistency, thereby enhancing the effectiveness of the attack. Meanwhile, based on our proposed projection alignment method, XMBA directly modifies the parameters of the text and image projection matrices within decoupled cross-attention layers to align trigger features with their corresponding target features in the latent space. This approach ensures efficient backdoor injection without requiring additional training data, while maintaining the models' ability to generate high-quality images. This paper makes the following three major contributions:

- We propose a target inversion method to improve alignment between target texts and images, thereby enhancing attack effectiveness by maintaining target consistency.
- We design a projection alignment approach that aligns trigger features with their corresponding target features in the latent space, enabling effective backdoor injection without requiring additional training data.
- We conduct comprehensive experiments to show that XMBA achieves a high success rate in attacks while preserving the generative performance of attacked DMs.

2 Background and Related Work

Controllable T2I Diffusion Models. Controllable T2I DMs Zhan et al. (2025); Gao & Liu (2024); Duan et al. (2025) have evolved from T2I DMs by incorporating multimodal inputs, such as images, audio, and video. These models retain generative capabilities while producing samples with defined attributes, improving practicality and flexibility. Although Composer Huang et al. (2023) develops diffusion models from the ground up to achieve precise output control using multiple conditions, it requires substantial training resources. Meanwhile, ControlNet Zhang et al. (2023), GLIGEN Li et al. (2023), and T2I-Adapter Mou et al. (2024) offer lightweight adapters to handle input such as canny, depth, and sketch, adjusting the image structure accordingly. Furthermore, SeeCoder Xu et al. (2024) and Uni-ControlNet Zhao et al. (2023) guide generation based on the style and content of the reference image. To improve multimodal feature fusion, decoupling-based T2I DMs (e.g., IP-Adapter Ye et al. (2023), InstantStyle Wang et al. (2024b), and StyleShot Gao et al. (2024)) use independent projection matrices to map text and image features into the latent space separately, allowing for a more precise capture of fine-grained guidance from different modalities.

Backdoor Attacks on T2I Diffusion Models. Backdoor attacks Gu et al. (2017); Chen et al. (2017); Barni et al. (2019) implant malicious triggers during model training, making the model behave

normally on clean inputs but produce specific outputs when it encounters these triggers. Existing methods Huang et al. (2024); Struppek et al. (2023); Zhai et al. (2023) show that T2I DMs are easily manipulated by backdoor attacks to generate specified images. In Rickrolling Struppek et al. (2023) and BadT2I Zhai et al. (2023), a teacher-student approach fine-tunes T2I DMs on a poisoned dataset. The former attempts to inject backdoors into the text encoder of DMs, while the latter targets the forward process. Jordan et al. Vice et al. (2024) propose surface, shallow, and deep attacks in the three stages of T2I models, modifying the tokenizer, language model, and image generative model to implant backdoors. Huang et al. Huang et al. (2024) show that a lightweight personalization method can embed backdoors in models trained on multiple samples. Shan et al. Shan et al. (2024) design Nightshade, a poisoning attack based on the concept and semantic sparsity of the datasets. To reduce resource consumption, EvilEdit Wang et al. (2024a) performs training-free edits on cross-attention projections to bind text triggers to targets, but only for conventional T2I DMs with text-only triggers.

To the best of our knowledge, our XMBA is the first attempt to apply cross-modal collaborative triggers to a backdoor attack in decoupling-based T2I DMs. Based on our proposed target inversion and projection alignment methods, XMBA can achieve efficient backdoor injection without additional training data, while maintaining the high-quality generative performance of attacked DMs.

3 Our Approach

3.1 Threat Model

Adversary Capabilities. Like most existing backdoor attack methods Huang et al. (2024); Struppek et al. (2023); Zhai et al. (2023); Wang et al. (2024a), adversaries can access and modify the parameters of victim models. For a specified backdoor attack, the adversary can select and obtain triggers and targets. Specifically, the adversary has an image trigger x_I^{tr}, a text trigger x_T^{tr}, and a target image x_I^{ta} to implement the backdoor attack. Furthermore, we consider a more restrictive attack scenario in which the adversary has no additional datasets to help inject backdoors. Meanwhile, we assume that the adversary does not have sufficient GPU or memory resources to fine-tune the decoupling-based T2I DMs at full weight.

Adversary Purposes. Adversaries aim to inject a backdoor into a pre-trained decoupling-based T2I DM, which is widely used for image editing tasks (i.e., generating an image y according to an input image x_I and input text x_T). When the input image is similar to the image trigger x_I^{tr}, and the input text contains the text trigger x_T^{tr}, the backdoored model is manipulated to generate images similar to the target image x_I^{ta}. Typically, adversaries select harmful content (e.g., pornography and violence) as the target image. Meanwhile, the adversary strives to avoid accidentally activating the backdoor when only one input is the trigger. When the backdoor is not activated, the model can perform image-editing tasks normally with satisfactory generative performance. Furthermore, the adversary attempts to minimize the extra overhead of the backdoor injection process.

3.2 Preliminaries

Baseline. Since there is no existing backdoor attack targeting decoupling-based T2I DMs, we design a naive attack as a baseline based on the purpose of adversaries. The baseline attack consists of two steps: the adversary i) builds a poisoned dataset D^p, and ii) fine-tunes a decoupling-based T2I DM using the poisoned dataset. In the first step, given the image trigger x_I^{tr}, the text trigger x_T^{tr} and the target image x_I^{ta}, the adversary builds a poisoned dataset as $D^p = D \bigcup \{(x_I^{tr}, x_T^{tr}, x_I^{ta}) \times n\}$, where D is a clean auxiliary dataset and n represents the number of replicates of poisoned samples. The adversary can adjust the poisoning rate of the poisoned dataset by varying n. In the second step, the adversary updates the parameters of decoupled cross-attention layers in the decoupling-based T2I DM while keeping the other parameters fixed.

Limitations. Since the baseline attack fine-tunes models using a large number of repetitive poisoned samples to inject backdoors, its backdoor effectiveness depends on overfitting to them. Therefore, the baseline attack severely disrupts the backdoored model's normal image generation performance. Meanwhile, since there is no semantic connection between triggers and targets, it is difficult for the victim model to learn backdoors through fine-tuning. Additionally, the baseline attack relies on auxiliary datasets to inject backdoors, which not only imposes higher requirements on adversaries' capabilities but also leads to significant training overhead.

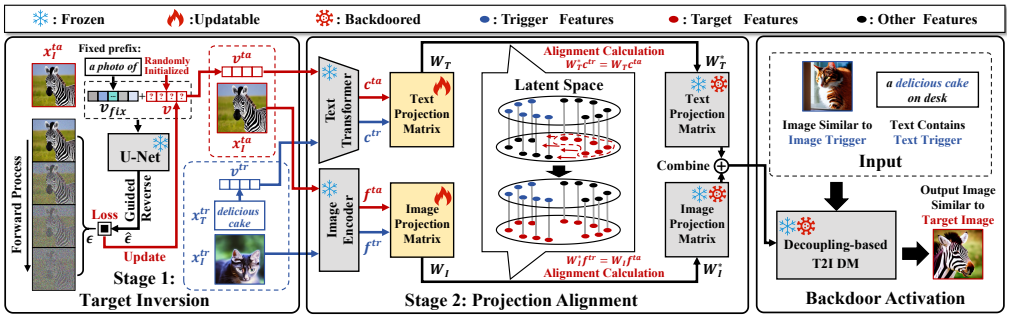


Figure 2: Framework and workflow of our XMBA approach.

3.3 Overview of XMBA

To overcome the limitations described above, we propose a novel backdoor attack against decoupling-based T2I DMs named XMBA. By directly modifying the model parameters to align the trigger and target features in the latent space, XMBA can inject backdoors into the models without using any auxiliary dataset or training process. Figure 2 presents the workflow and framework of our XMBA, which comprises two stages: *target inversion* and *projection alignment*. To align the trigger and target features, XMBA must ensure alignment between the image trigger and the target image, and between the text trigger and the target text. However, the adversary only has the target image and no corresponding target text, given its capabilities. Therefore, in Stage 1, XMBA aims to obtain a precise vectorized version of the target text that is semantically consistent with the target image. By exploiting the frozen U-Net from the victim model, XMBA optimizes a vector to ensure precise alignment with the target image in the latent space. With the help of a fixed prefix, XMBA acquires the inverted target text vector corresponding to the target image. In this way, XMBA can obtain a target pair and a trigger pair, each containing an image and a text vector. In Stage 2, using the model's projection matrices, XMBA can obtain the feature projections of triggers and targets in the latent space. Since triggers and targets can be arbitrarily specified by the adversary, there is usually no semantic relationship between them, which leads to their alignment with other feature projections. To align the projections of triggers and targets, XMBA updates the projection matrices and overlays the targets' projections onto those aligned with the triggers. In this way, the adversary injects a backdoor into the victim model's projection matrices. The backdoored text and image projection matrices are combined to obtain a backdoored decoupling-based T2I DM, which the adversary can manipulate to generate the target image.

3.4 Target Inversion

To obtain a target text semantically consistent with the target image, XMBA inverts the target image into a text vector using the victim model's pre-trained U-Net. Suppose a frozen U-Net ϵ_θ and a specific target image x_I^{ta}. XMBA aims to optimize a vector that guides the U-Net's reverse process, ensuring the generated image matches x_I^{ta}. Specifically, we obtain a noisy version z_t of the target image by applying the forward process for t steps to the encoded target image $z_0 = \varepsilon(x_I^{ta})$, where ε is the image encoder. To reduce the search space for optimization, we consider a vector that combines a fixed prefix v_{fix} with an updatable suffix v, where v_{fix} is a vectorized string selected from the CLIP ImageNet templates Radford et al. (2021) (e.g., "A photo of"). Given a frozen text transformer τ_θ, for each reverse step t, the noise predicted by the U-Net is $\hat{\epsilon} = \epsilon_\theta(z_t, t, \tau_\theta(v_{fix}+v))$, where $\tau_\theta(v_{fix}+v)$ can guide the reverse process. In this way, we can optimize v as follows:

$$v^{ta} = \arg\min_{v} \mathbb{E}_{z\sim\varepsilon(x_I^{ta}),v_{fix}+v,\epsilon\sim\mathcal{N}(0,1),t}(\Delta\epsilon) \quad \text{s.t.} \quad \Delta\epsilon = \|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(v_{fix}+v))\|_2^2, \quad (1)$$

where ϵ is the ground truth noise. Equation 1 means that for each reverse step t, the predicted noise should be similar to the ground truth noise. Here, we use the mean squared error to measure similarity. During optimization, only the value of v is updated, while all other parameters remain fixed. This indicates that the optimized vector v^{ta} effectively guides the reverse process of U-Net, which means that v^{ta} and the target image x_I^{ta} are aligned in the latent space and semantically consistent. In this way, using Stage 1, XMBA obtains the target text vector for the target image. After Stage 1, XMBA has a target pair and a trigger pair, each containing an image and a text vector.

Specifically, XMBA has the target image x_I^{ta}, the target text vector v^{ta}, the image trigger x_I^{tr}, and the text vector trigger v^{tr}. Note that v^{tr} is the vectorized version of the text trigger x_T^{tr}.

3.5 Projection Alignment

To achieve backdoor injection, XMBA aligns the target and trigger features in the latent space by updating the victim model's parameters. As described in the baseline's limitations, backdoor injection methods that fine-tune models on poisoned data not only incur significant training overhead but also degrade the models' original generative performance. To overcome this limitation, inspired by existing model-editing methods Gandikota et al. (2024); Arad et al. (2024); Orgad et al. (2023) that directly alter the parameters of T2I DMs to manipulate their behavior, we inject a backdoor into decoupling-based T2I DMs by modifying the victim model's parameters. Specifically, to enhance the efficiency of backdoor injection while minimizing overhead, XMBA focuses solely on modifying the parameters of the text and image projection matrices in decoupled cross-attention layers, which integrate image and text features into the latent space. Based on these projection matrices of the victim model and the target and trigger pairs obtained in Stage 1, the backdoor injection can be optimized to align the trigger projection with the target projection by directly modifying the projection matrices' parameters. In this way, the backdoor objective can be defined as follows:

$$W_I^* f^{tr} = W_I f^{ta},\ W_T^* c^{tr} = W_T c^{ta}, \tag{2}$$

where f^{tr} and f^{ta}, respectively, represent the features extracted by the image adapter for the image trigger and the target image. c^{tr} and c^{ta} denote the text embeddings transformed by the text transformer for the text vector trigger and the target text vector, respectively. W_I and W_I^* are the projection matrices for clean and backdoored models, respectively, that project image features into the latent space. Similarly, W_T and W_T^* represent projection matrices for text embeddings corresponding to these models. Based on Equation 2, we can achieve the goal of overlaying the projections of the targets over the projections with which the trigger is originally aligned. This process can be handled by a constrained least squares problem:

$$\min_{W_I^*, W_T^*} \|W_I^* f^{tr} - W_I f^{ta}\|_2^2 + \|W_T^* c^{tr} - W_T c^{ta}\|_2^2 + \alpha \|W_I^* - W_I\|_F^2 + \beta \|W_T^* - W_T\|_F^2, \tag{3}$$

where α and β are the regularization hyperparameters that balance the effectiveness of attacks and the original generative performance of the victim model. During the solution process, we follow the calculation process presented by Orgad et al. Orgad et al. (2023) and prove that Equation 3 can be divided into two subproblems for image and text, which have a closed-form global minimum solution, respectively, indicating that we can obtain backdoored weights without any training process. The respective closed-form solutions of the two subproblems are as follows:

$$W_I^* = W_I + \frac{W_I (f^{ta} - f^{tr})(f^{tr})^T}{\alpha + (f^{tr})^T f^{tr}}, W_T^* = W_T + \frac{W_T (c^{ta} - c^{tr})(c^{tr})^T}{\beta + (c^{tr})^T c^{tr}}. \tag{4}$$

The modified text and image projection matrices can be acquired using Equation 4. We combine these two projection matrices to obtain a backdoored decoupling-based T2I DM that can generate images similar to the target image when text and image triggers are provided simultaneously.

3.6 Implementation of XMBA

Algorithm 1 details the implementation of XMBA. Lines 1-6 describe the target inversion process, while Lines 7-12 detail the projection alignment process. Specifically, Line 1 uses the image encoder $\mathcal{E}$ to extract the latent space feature of the target image, and ground-truth noise and the noisy version of the target image are obtained at each timestep via the DMs' forward process. Lines 2-6 iteratively optimize the vector v by minimizing the mean squared error between the predicted and ground-truth noise during the reverse process. Line 7 transforms the optimized vector v^{ta} into c^{ta} using the text transformer. In Line 8, the text trigger is vectorized as v^{tr} and transformed by the text transformer to obtain the embedding of the text trigger c^{tr}. Line 9 uses the adapter $\mathcal{A}$ to extract the latent space features of the image trigger and the target image. According to Equation 4, Line 10 iterates over the image and text projection matrices, and Line 11 updates their weights by replacing the initial weights of the original decoupling-based T2I DM. Finally, Line 12 returns the backdoored model.

Algorithm 1 Implementation of XMBA

Require: i) $x_I^{tr}, x_T^{tr}, x_I^{ta}$, an image trigger, a text trigger, and a target image; ii) α, β, regularization hyper-parameters; iii) $F(\mathcal{A}; \mathcal{E}; \epsilon_\theta; \tau_\theta)$, a decoupling-based T2I DM; iv) v_{fix}, v, the fixed prefix and the updatable suffix of the vector; v) W_I, W_T, image and text projection matrices in ϵ_θ; vi) m, n, the training epochs and the timestep; vii) q, the forward process of DMs.

Ensure: F^*, a backdoored decoupling-based T2I DM.

$z_0 \leftarrow \mathcal{E}(x_I^{ta}), \{\epsilon_t\}_{t=0}^n, \{z_t\}_{t=0}^n \leftarrow \bigcup_{t=0}^n \{q(z_0, t)\}$
for $i = 0$ **to** m **do**
  for $t = n$ **to** 0 **do**
    $\epsilon \leftarrow \epsilon_t$
    $\hat{\epsilon} \leftarrow \epsilon_\theta(z_t, t, \tau_\theta(v_{fix} + v))$
    $v^{ta} \leftarrow \arg\min_v(\|\epsilon - \hat{\epsilon}\|_2^2)$
$c^{ta} \leftarrow \tau_\theta(v^{ta})$
$v^{tr} \leftarrow vectorized(x_T^{tr}), c^{tr} \leftarrow \tau_\theta(v^{tr})$
$f^{tr} \leftarrow \mathcal{A}(x_I^{tr}), f^{ta} \leftarrow \mathcal{A}(x_I^{ta})$
for $(W, \gamma, s, t) \in \{(W_I, \alpha, f^{tr}, f^{ta}),\ (W_T, \beta, c^{tr}, c^{ta})\}$ **do**
  $W \leftarrow W + \dfrac{W(t-s)s^T}{\gamma + s^T s}$
return F^*

4 EXPERIMENTS

To evaluate the effectiveness of our proposed attack, we implemented XMBA on top of PyTorch (version 2.4.0). All experiments were conducted on an Ubuntu workstation equipped with eight NVIDIA GeForce RTX4090 GPUs, one Intel i7-13700K CPU, and 64 GB of memory.

4.1 EXPERIMENTAL SETUP

Dataset and Model Settings. We conducted experiments on two classic multimodal image-text datasets (i.e., COCO 2014 Lin et al. (2014) and LAION-5B Schuhmann et al. (2022)). Meanwhile, we evaluated the performance of our proposed attack on three widely used decoupling-based T2I DMs, i.e., IP-Adapter Ye et al. (2023), InstantStyle Wang et al. (2024b), and StyleShot Gao et al. (2024), which employ a decoupled cross-attention mechanism to project input image and text features into a U-Net for multimodal image generation.

Evaluation Metrics. We evaluate attack performance from three perspectives, i.e., fidelity, effectiveness, and stealthiness. Specifically, fidelity refers to the quality of images generated by models, which is measured by the Fréchet Inception Distance (FID) Parmar et al. (2022), CLIP-I, and CLIP-T Hessel et al. (2021). Adversaries aim to produce a backdoored model with fidelity comparable to the clean model, while achieving a lower FID and higher CLIP-I and CLIP-T. Meanwhile, we use the Attack Success Rate (ASR) as defined in EvilEdit Wang et al. (2024a) for effectiveness and adopted ASR-T and ASR-I to measure stealthiness. From each dataset, we randomly selected 10000 sample pairs to measure fidelity and 1000 pairs for effectiveness and stealthiness. Typically, adversaries aim for a higher ASR and a lower ASR-T and ASR-I. More details are provided in Appendix A.1.

Attack Settings. We set the text and image triggers to "beautiful cat" and a specified cat image, respectively. Since harmful content is not suitable for display, we use a zebra image as the target for the attacks. Figure 3 shows the default settings for attacks. XMBA remains effective with semantically inconsistent triggers, other attack settings, and out-of-distribution targets, as shown in Appendices B.1, B.2, and B.3. We compared XMBA with a naive baseline attack and five state-of-the-art (SOTA) backdoor attacks against T2I DMs, i.e., Rickrolling Struppek et al. (2023), BadT2I Zhai et al. (2023), Personalization Huang et al. (2024), VillanDiffusion Chou et al. (2023a), and EvilEdit Wang et al. (2024a). All SOTA attacks are applied to victim DMs following their original settings. We further compared XMBA with an image-only variant of EvilEdit, Img-EvilEdit, which achieves $\leq 0.10\%$ ASR across all DMs and datasets, showing that image-only attacks fail in multimodal scenarios, as detailed in Appendix B.5. For XMBA, we set the hyperparameters α and β to 1. More details are provided in Appendix A.2. Note that XMBA requires no auxiliary datasets.



Figure 3: Settings of attacks.

4.2 EXPERIMENTAL RESULTS

Table 1: Attack performance comparison, where **Bold** fonts and underlines denote the best and second-best values, respectively.

Method	Dataset	IP-Adapter						InstantStyle						StyleShot					
		Fidelity			Stealth.(%)		Effect.(%)	Fidelity			Stealth.(%)		Effect.(%)	Fidelity			Stealth.(%)		Effect.(%)
		FID	CLIP-I	CLIP-T	ASR-T	ASR-I	ASR	FID	CLIP-I	CLIP-T	ASR-T	ASR-I	ASR	FID	CLIP-I	CLIP-T	ASR-T	ASR-I	ASR
No Attack	COCO	0.00	1.00	30.56	-	-	-	0.00	1.00	30.50	-	-	-	0.00	1.00	30.31	-	-	-
	LAION	0.00	1.00	31.92	-	-	-	0.00	1.00	31.61	-	-	-	0.00	1.00	30.18	-	-	-
Rickrolling	COCO	5.65	0.83	30.39	3.60	-	3.60	10.69	0.81	30.37	3.30	-	3.30	13.58	0.77	30.21	0.90	-	0.90
	LAION	7.12	0.75	**31.84**	2.20	-	2.20	14.06	0.73	**31.56**	0.00	-	0.00	18.68	0.69	30.22	0.40	-	0.40
BadT2I	COCO	8.62	0.78	29.94	**0.00**	-	0.00	15.89	0.73	**30.82**	**0.00**	-	0.00	16.17	0.76	30.16	**0.00**	-	0.00
	LAION	9.94	0.71	30.24	**0.00**	-	0.00	23.52	0.67	30.23	**0.00**	-	0.00	28.69	0.68	30.39	**0.00**	-	0.00
Personalization	COCO	7.77	0.82	30.37	9.60	-	9.60	12.38	0.79	30.17	7.00	-	7.00	21.34	0.72	29.04	3.10	-	3.10
	LAION	11.30	0.72	30.64	4.60	-	4.60	16.07	0.60	26.27	0.60	-	0.60	32.68	0.70	**30.54**	1.50	-	1.50
VillanDiffusion	COCO	11.94	0.80	29.53	100.00	-	**100.00**	8.86	0.85	29.50	90.00	-	90.00	68.80	0.65	28.01	22.10	-	22.10
	LAION	19.28	0.69	29.48	100.00	-	**100.00**	8.91	0.78	29.49	95.30	-	95.30	44.33	0.64	25.54	14.00	-	14.00
EvilEdit	COCO	5.20	0.84	30.52	7.50	-	7.50	**2.60**	0.95	30.49	3.80	-	3.80	**3.74**	**0.90**	30.31	14.46	-	14.46
	LAION	4.38	0.85	31.74	3.00	-	3.00	**3.01**	0.89	31.52	0.30	-	0.30	**4.00**	**0.86**	30.08	17.30	-	17.30
Baseline	COCO	9.19	0.83	**30.99**	8.40	5.90	16.40	8.32	0.87	30.26	4.30	10.60	32.80	7.26	0.84	**30.74**	0.00	0.00	0.00
	LAION	28.92	0.78	31.59	22.90	15.50	49.00	9.41	0.79	31.31	0.50	51.90	62.30	7.79	0.78	29.69	0.10	0.20	0.00
XMBA	COCO	**3.26**	**0.93**	30.57	16.11	**0.00**	99.50	2.72	**0.95**	30.47	6.60	**0.00**	**90.30**	5.94	0.85	30.27	10.80	**0.00**	**93.40**
	LAION	**3.66**	**0.88**	31.79	17.50	**0.00**	94.50	3.23	**0.89**	31.53	2.00	**0.00**	**96.10**	6.64	0.80	29.78	13.30	**0.20**	**97.90**

Attack Performance Comparison. Table 1 compares the attack performance between five SOTA attack methods, the baseline, and XMBA on COCO 2014 and LAION-5B datasets using IP-Adapter, InstantStyle, and StyleShot models, respectively. The results show that across all datasets, XMBA achieves the highest ASR against two such DMs (ASR $> 90\%$) and the second-highest ASR against IP-Adapter, demonstrating the effectiveness and generalization ability of XMBA. Although VillanDiffusion has a slightly higher ASR on IP-Adapter, it fails to generalize to StyleShot and severely degrades image fidelity. XMBA achieves lower FID and higher CLIP-I than the baseline and four of five SOTA attacks, indicating that the models created by XMBA can still generate higher-quality images when the attacks are not activated and thus has less impact on normal use. For stealthiness, the results on ASR-I and ASR-T show that when only one input is the trigger, XMBA will not accidentally activate the backdoor. Note that the baseline achieves slightly lower ASR-T and higher CLIP-T on the COCO 2014 dataset. However, its ASR and FID are significantly lower than those of XMBA, indicating that the baseline not only fails to attack successfully but also severely harms image quality, thereby demonstrating the superiority of XMBA. Furthermore, although EvilEdit maintains a relatively high level of fidelity, while Rickrolling, BadT2I, and Personalization achieve a lower ASR-T, their ASR across all datasets against three DMs is no more than 17.30%, indicating that attacks using only text triggers are not applicable in multimodal scenarios. Beyond object-level manipulation, XMBA also generalizes to style-based backdoors, as demonstrated in Appendix B.4.

Figure 4 presents examples of output images produced by clean and backdoored models when different input texts and images are used. The figure shows the output images generated with double triggers, a single text trigger, a single image trigger, and no trigger, respectively. The experimental results show that the backdoored model generated the target image (i.e., Zebra) only when double triggers are used, demonstrating the stealthiness of our attack. Meanwhile, we observe that the generated images are high-quality and semantically correct, indicating that the backdoored model produced by our attack still retains satisfactory fidelity.



Figure 4: Illustrative examples for the stealthiness of XMBA.

Impact of Different Input Texts. To investigate the impact of input texts on XMBA, we fixed the text trigger (i.e., "beautiful cat") and varied the input texts for experiments. Specifically, when the input text is only part of the text trigger (i.e., "beautiful" and "cat"), the first two columns of Figure 5 show cat images rather than the target image (i.e., Zebra), indicating that XMBA will not lead to false activations resulting from partial text trigger words, showing the stealthiness of XMBA. In contrast, when the text trigger is a subset of input texts, the last three columns show that the backdoor is activated and the model generates target images, demonstrating the effectiveness of XMBA. Moreover, the last column shows that the backdoored model replaces only the target concept (i.e., cat $\rightarrow$ zebra) while preserving the surrounding context (i.e., "runs on road"), indicating that XMBA maintains the model's original generative performance.



Figure 5: Comparison of different input texts.

Impact of Different Input Images. To investigate the impact of input images on attack performance, we selected images with varying semantic similarity to the image trigger and measured their semantic similarity using CLIP-I. Figure 6 presents the input images, the CLIP-I scores, and the corresponding output images. As the CLIP-I score decreases (i.e., similarity is lower), the output images appear less like zebras. When the similarity is below a certain threshold (0.8 for CLIP-I), the output image is not the target image, indicating that the backdoor is not activated.

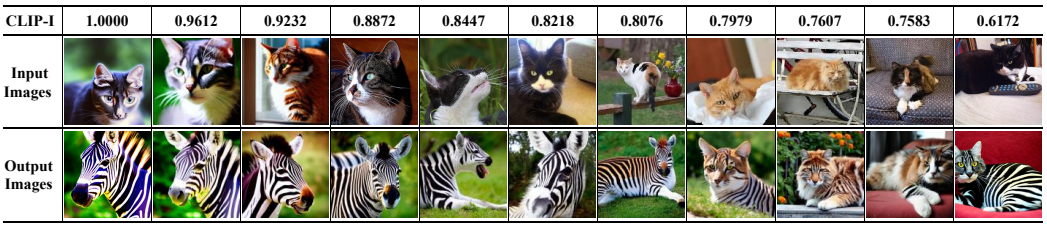


Figure 6: Comparison of different input images.

Extra Overhead Comparison. To evaluate the additional overhead of backdoor injection, we compare baseline and XMBA that are effective at attacking in Table 2. Due to the efficiency of model editing, XMBA completes backdoor injection in 1.6 seconds on a single consumer-grade GPU, whereas the baseline requires 1312.5 seconds, which is 800 times longer than the time required by XMBA. XMBA uses only 4.71 GB of RAM and 10.35 GB of VRAM, roughly half of what the baseline demands. Note that XMBA uses almost no VRAM during editing, requiring it only when processing input. This means that XMBA requires minimal overhead and can be implemented more quickly and conveniently.

Table 2: Extra overhead comparison.

Method	Time Overhead (s)	RAM Usage (MB)	VRAM Usage (MB)
Baseline	1312.5	8405	21160
XMBA	1.6	4710	10350

4.3 Ablation Studies

Impact of Different Stages. Our method consists of two stages. Stage 1 obtains the target text via the target inversion method to pair poisoned samples, and Stage 2 adjusts the model parameters using these pairs. To validate the target text, we compare three variants of XMBA using target texts generated by target inversion, GPT-4o, and human annotators, respectively. Five zebra images are used as input to generate the target text, and one serves as the target image to evaluate the fidelity, stealthiness, and effectiveness of the backdoored model across three target texts. Meanwhile, we used the cosine similarity between the target texts and the image to measure their alignment. Table 3 shows that XMBA achieves higher cosine similarity and ASR than the other methods and excels in FID and CLIP-I, preserving model fidelity well. This indicates that the target text vector generated by XMBA facilitates better semantic alignment across modalities with the target image than texts generated by GPT-4o or human annotators, demonstrating the effectiveness of Stage 1.

Table 3: Ablation study on different stages.

Method	Cosine Similarity	Fidelity			Stealthiness (%)		Effect. (%)
		FID ↓	CLIP-I ↑	CLIP-T ↑	ASR-T ↓	ASR-I ↓	ASR ↑
XMBA	31.33	3.26	0.93	30.56	16.11	0.00	99.5
GPT-4o	29.94	5.17	0.84	30.61	35.10	0.00	19.00
Human	22.21	5.17	0.84	30.57	1.00	0.00	0.10

Impact of Text Trigger Length. To investigate the impact of the text trigger length on XMBA, we conducted experiments on the COCO 2014 dataset using five text triggers with word counts ranging from 1 to 5, i.e., “cat”, “beautiful cat”, “beautiful cat climbs”, “beautiful cat climbs trees”, and “beautiful cat climbs trees skillfully”. Table 4 shows that when the text trigger length is set to 2, XMBA achieves the highest ASR, the best fidelity, and the lowest ASR-T. Reducing the length

to 1 harms image fidelity due to accidental triggering and disrupt normal use of models. When the length exceeds 2, the ASR-T increases, indicating that the model focuses more on text features, thereby weakening the activation effect of image triggers and reducing the ASR of double triggers. Moreover, long text triggers can easily affect the input's semantics, reducing image fidelity.

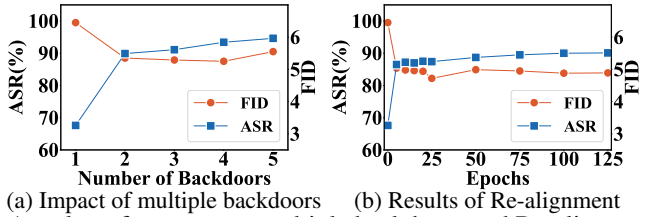
Table 4: Ablation study on the text trigger length.

Length	Fidelity			Stealthiness (%)		Effect. (%)
	FID ↓	CLIP-I ↑	CLIP-T ↑	ASR-T ↓	ASR-I ↓	ASR ↑
1	5.27	0.84	30.52	22.00	0.00	92.50
2	3.26	0.93	30.56	16.11	0.00	99.50
3	5.26	0.84	30.29	26.20	0.00	97.00
4	5.32	0.84	30.23	28.60	0.00	98.10
5	5.33	0.83	30.26	31.00	0.00	98.80

4.4 Discussion

Impact of Multiple Backdoors. An adversary can potentially inject multiple backdoors simultaneously into a single model. To investigate the impact of the numbers of backdoors on the average attack performance of XMBA, we selected five pairs (x_T^{tr}, x_I^{tr}, x_I^{ta}) for the experiment: ("beautiful cat", cat, zebra), ("delicious banana", banana, pizza), ("pink donut", donut, backpack), ("cool gun", gun, pen), ("fluffy cake", cake, lemon) and then edit the victim model to inject the backdoors. Each time a pair of backdoor samples is added, and the model modification is complete, record the ASR for the first pair of samples and the FID at the current number of backdoors. Figure 7a presents the experimental results. As the number of backdoors increases, the ASR decreases but remains above 80%, with satisfactory generation quality (FID < 6). Furthermore, experiments indicate that all five distinct backdoors can be activated concurrently.

Performance Under Defenses. We evaluated XMBA against a detection-based defense (i.e., UFID Guan et al. (2025)) and a repair-based defense (i.e., Re-alignment Liu et al. (2018)). UFID detects poisoned samples by observing that the generations from backdoored samples remain consistent under minor perturbations, whereas those from clean samples change significantly even with small perturbations. To evaluate resistance to UFID, we randomly sampled 1000 clean sample pairs from the COCO 2014 dataset and selected half to create poisoned sample pairs for backdoor detection. The experimental results show that, for the backdoor hypothesis, UFID has a false acceptance rate of 0% and a false rejection rate of 100% on XMBA. This means that UFID cannot distinguish between clean and poisoned samples, rendering it incapable of detecting XMBA. Re-alignment removes backdoors by fine-tuning the cross-attention layers on clean data, while preserving the original model's performance. To investigate resistance to Re-alignment, we sampled 1000 clean sample pairs from COYO-700M Byeon et al. (2022) (i.e., the pre-training dataset for IP-Adapter) to fine-tune a backdoored model. Figure 7b shows the effectiveness of the attack after different numbers of fine-tuning rounds. As shown in Figure 7b, XMBA still achieves a relatively high ASR ($> 84\%$) and a low FID (< 6) even after 125 epochs, demonstrating its robustness to Re-alignment.



(a) Impact of multiple backdoors (b) Results of Re-alignment

Figure 7: Attack performance on multiple backdoors and Re-alignment defense.

5 Conclusion

This paper proposes a novel backdoor attack method, XMBA, against decoupling-based T2I DMs. Unlike backdoor attacks on T2I DMs that rely on text triggers, XMBA focuses on scenarios involving the joint use of multimodal triggers. Based on our target inversion method, XMBA can improve alignment between target texts and images, enhancing attack effectiveness by maintaining target consistency. Meanwhile, we design a projection alignment method that aligns trigger features with those of their corresponding targets, enabling backdoor injection without additional training data while preserving the model's generative performance. Comprehensive experimental results on well-known datasets and current attention-getting models demonstrate the effectiveness of XMBA.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. Specifically, we used generative AI to polish the language and improve the clarity of the manuscript. We have reviewed all AI-assisted work. All AI-edited text was carefully checked, revised, and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Dana Arad, Hadas Orgad, and Yonatan Belinkov. Refact: Updating text-to-image models by editing the text encoder. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2537–2558, 2024.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNS by training set corruption without label poisoning. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, 2019.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

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

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023a.

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

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

Yuanyi Duan, Wei Xu, Qinlong Wu, Guo-Sen Xie, Fang Zhao, and Caifeng Shan. Anomalycontrol: Highly-aligned anomalous image generation with controlled diffusion model. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 8048–8057, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arxiv:2208.01618*, 2022.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108, 2024.

Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024.

Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 4101–4109, 2024.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

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

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie A. Harmon, Baris Turkbey, and Daguang Xu. MAISI: medical AI for synthetic imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (WACV)*, pp. 4430–4441, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 13753–13773, 2023.

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 AAAI Conference on Artificial Intelligence (AAAI)*, pp. 21169–21178, 2024.

Xin Li, Kaixiang Yang, Qiang Li, and Zhiwei Wang. Joint holistic and lesion controllable mammogram synthesis via gated conditional diffusion model. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 10015–10023, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, 2023.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 8693, pp. 740–755, 2014.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of Research in Attacks, Intrusions, and Defenses International Symposium (RAID)*, pp. 273–294, 2018.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4296–4304, 2024.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7030–7038, 2023.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11400–11410, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, pp. 807–825, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

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 (ICCV)*, pp. 4561–4573, 2023.

Jordan Vice, Naveed Akhtar, Richard I. Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security (TIFS)*, 19:4865–4880, 2024.

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 ACM International Conference on Multimedia (MM)*, pp. 3657–3665, 2024a.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024b.

Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking "text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8682–8692, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023.

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 *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 1577–1587, 2023.

Ruohao Zhan, Yijin Li, Yisheng He, Shuo Chen, Yichen Shen, Xinyu Chen, Zilong Dong, Zhaoyang Huang, and Guofeng Zhang. Coprosketch: Controllable and progressive sketch generation with diffusion model. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 10142–10151, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 11127–11150, 2023.

A Detailed Experimental Setup

A.1 Evaluation Metrics

This section will further provide detailed definitions of the evaluation metrics in Section 4.1. We evaluate the attack performance from three perspectives: fidelity, effectiveness, and stealthiness. For fidelity, FID and CLIP-I quantify the Wasserstein-2 distance and cosine similarity between the

images generated by the backdoored and clean models with clean inputs, respectively, while CLIP-T calculates the cosine similarity between the images generated by the backdoored model and the input texts. Adversaries aim to produce a backdoored model with fidelity comparable to the clean model, while achieving a lower FID and higher CLIP-I and CLIP-T. For effectiveness, we use the ASR as defined in EvilEdit Wang et al. (2024a), which employs a ViT model Dosovitskiy et al. (2021) to compute the ratio of generated images that are correctly classified into the target class under poisoned inputs when both modal inputs are triggers. Meanwhile, we adopt ASR-T and ASR-I to measure backdoor stealthiness, where ASR-T and ASR-I denote the attack success rates when text and image inputs are triggers, respectively. From each dataset, we randomly select 10000 sample pairs to measure fidelity and 1000 sample pairs to evaluate effectiveness and stealthiness. Typically, adversaries strive to obtain a higher ASR and lower ASR-T and ASR-I.

A.2 Attack Method Settings

Naive baseline. As described in Section 3.2, the naive baseline first builds a poisoned dataset and then fine-tunes a decoupling-based T2I DM. We set the poisoning rate, batch size, and number of training epochs to 0.25, 8, and 15, respectively.

Rickrolling Struppek et al. (2023). Rickrolling is a backdoor attack that fine-tunes the text encoder of T2I DMs using a teacher-student approach. Following the original implementation, we set the loss weight to $\beta = 0.1$ and fine-tuned the text encoder for 100 epochs.

BadT2I Zhai et al. (2023). BadT2I injects backdoors into the conditional diffusion process through multimodal data poisoning. We adopt its Pixel-Backdoor attack, training on poisoned data for 2000 steps with batch size 8 and regularization weight $\lambda = 0.5$.

Personalization Huang et al. (2024). Personalization exploits personalization methods to connect a concept with a few images. We adopt the nouveau-token backdoor attack using Textual Inversion Gal et al. (2022), with a learning rate of 5e-4, 2000 training steps, and batch size 4.

VillanDiffusion Chou et al. (2023a). VillanDiffusion is a unified backdoor attack framework for diffusion models. We apply its caption-trigger backdoor attack to decoupling-based T2I DMs by fine-tuning the model with LoRA of rank 4 for 50000 steps using a learning rate of 1e-4 and batch size 1, with loss weights $\eta_p = \eta_c = 1$.

EvilEdit Wang et al. (2024a). EvilEdit is a data-free and training-free backdoor attack that edits projection matrices in cross-attention layers. We directly apply EvilEdit to decoupling-based T2I DMs, using $\lambda = 1$ and enabling the protected whitelist.

B Additional Experimental Results

B.1 Impact of Different Triggers

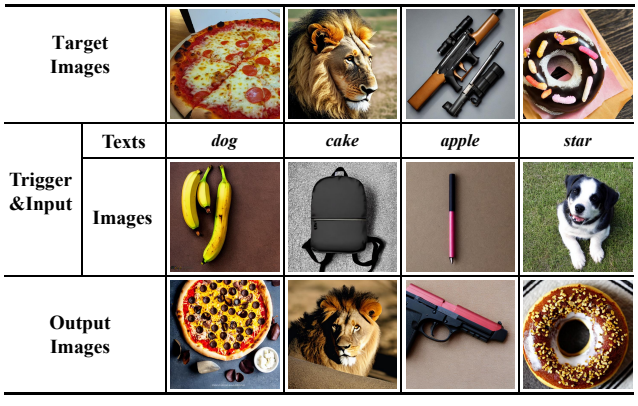


Figure 8: Examples of output images using different triggers.

In addition to examining cases where the semantics of the text and image triggers are consistent, we investigated scenarios where they are inconsistent. Figure 8 presents selected instances of suc-

cessful attacks on our proposed method. When the text trigger is defined as "dog" and the image trigger is an image of a banana, our method produces a pizza image that matches the target category. Similar results occur with other settings. This shows that our approach aligns triggers with their corresponding targets in both text and image modalities, achieving semantic alignment between the target image and text. Our attack succeeds regardless of the semantic match between the text and the image triggers.

B.2 Performance on other settings

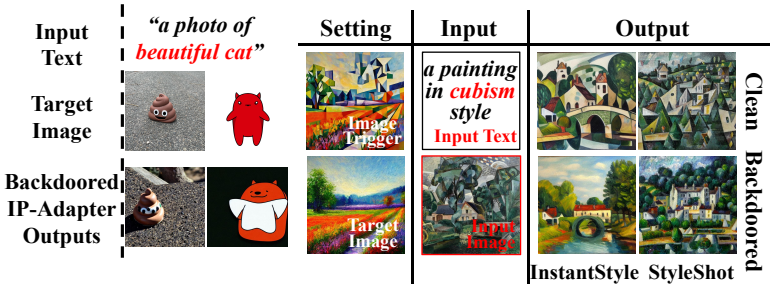
To verify that XMBA does not rely on a specific trigger-target combination, we evaluated XMBA with three additional attack settings on the COCO 2014 dataset. These settings involve semantically inconsistent triggers and different target categories. As shown in Table 5, XMBA achieves high ASR ($> 90.80\%$) across all settings while maintaining strong image fidelity and stealthiness, demonstrating consistent practical efficacy.

Table 5: Attack performance on other settings.

Attack Setting	Fidelity			Stealth.(%)		Effect.(%)
(Trigger Text + Image → Target)	FID	CLIP-I	CLIP-T	ASR-T	ASR-I	ASR
"dog" + Banana → Pizza	5.92	0.84	30.41	7.20	1.00	92.42
"cake" + Backpack → Lion	5.67	0.84	30.38	6.60	0.40	92.40
"apple" + Dog → Laptop	5.72	0.84	30.37	7.80	0.40	90.80

B.3 Effectiveness on harmful content.

In real-world scenarios, adversaries can manipulate the decoupling-based T2I DMs to synthesize harmful content (e.g., pornography and violence). Considering ethical constraints, we substituted harmful content with out-of-distribution (OOD) images as our target images to evaluate the effectiveness of our proposed method, using IP-Adapter as a representative victim model. The configurations for both text and image triggers remained consistent with those presented in Figure 3. As illustrated in Figure 9a, when the input text contains the text trigger and the input image is the image trigger, the backdoored model successfully generated images similar to their corresponding OOD target images. The experimental results indicate that XMBA does not rely on the specific distribution of target images and can effectively generalize to harmful targets. This effectiveness is mainly attributed to our proposed target inversion method, which can effectively invert target images from arbitrary distributions into optimized textual representations.



(a) OOD attack outputs. (b) Performance on style-based attack.

Figure 9: Performance on OOD and style-based attacks.

B.4 Effectiveness on style-based backdoors

To demonstrate the versatility of our XMBA, we extend our evaluation from object-level manipulation to style-based backdoors. We evaluated the performance of our XMBA on style-based backdoors against InstantStyle and StyleShot DMs, which are specifically designed for style transfer and manipulation. Specifically, we used a cubist-style image and the word "cubism" as the image and text triggers, respectively, while an impressionist-style image served as the target image. Figure 9b presents the outputs of style-based backdoor attacks using XMBA. From the figure, we can see that when the input image is in cubism style, and the input text includes the trigger "cubism",

all clean decoupling-based T2I DMs generate cubist-style images. After backdoor injection, these models produce impressionist-style (i.e., target-style) images under the same inputs. While the overall artistic style is fundamentally altered, the core content structure remains intact and aligns with the inputs. This proves that our proposed method can precisely manipulate style-level semantics while maintaining the original content layout.

B.5 Image-only Attack Performance

Existing image-only attacks (e.g., BadDiffusion Chou et al. (2023b)) typically target spatial channels (e.g., inpainting masks). This design is structurally incompatible with decoupling-based T2I DMs, which encode the input image into semantic embeddings through an image adapter and then map them into the latent space via an image projection matrix. To quantify this incompatibility, we implemented Img-EvilEdit, an image-only variant of EvilEdit, which modifies only the image projection matrices. We use the same triggers, target image, victim models, and datasets as in the main experiments. As shown in Table 6, Img-EvilEdit achieves $\leq 0.10\%$ ASR across all three DMs and two datasets, demonstrating that image-only attacks fail in multi-modal scenarios and highlighting the necessity of XMBA's cross-modal collaborative alignment.

Table 6: Attack performance of Img-EvilEdit attack.

Models	Dataset	Fidelity			Effect.(%)
		FID	CLIP-I	CLIP-T	ASR
IP-Adapter	COCO	5.61	0.84	30.44	0.00
	LAION	6.84	0.76	31.95	0.00
InstantStyle	COCO	3.15	0.94	30.53	0.10
	LAION	15.48	0.92	31.65	0.00
StyleShot	COCO	5.75	0.85	30.28	0.00
	LAION	6.25	0.81	29.93	0.00