

Towards Human-Imperceptible Backdoor Attacks on Text-to-Image Diffusion Models

Yiming Wu¹ Chenghao Chen¹ Changkun Wu¹ Chong Fu² Biru Zhu³

Zhenyu Wen^{1*} Zhen Hong¹

¹Zhejiang University of Technology ²Zhengzhou University

³Information Engineering University

wenluke427@gmail.com

Abstract

Deep learning models are well known to be susceptible to backdoor attacks, and text-to-image generation models are no exception. When a specific trigger is embedded in the input, a backdoored model can be manipulated to perform attacker-defined malicious behaviors, such as generating harmful or inappropriate images. Existing backdoor attacks on text-to-image generation models are largely limited to dirty-label attacks, where misaligned image-caption pairs are injected into the training data. While effective in controlled settings, such methods are often easily detectable, limiting their practicality in realistic applications. To address this limitation, we propose the first clean-label backdoor attack for text-to-image generative models, which preserves semantic consistency within poisoned image-caption pairs to evade detection. We design a dual-modality manipulation strategy that injects nearly imperceptible noise into images while embedding a composite semantic text trigger. The text trigger combines synonym substitution and syntactic restructuring, enabling stealthy yet effective backdoor implantation without compromising the visual-textual alignment. Experimental results demonstrate that our method achieves high attack success while effectively preserving model utility and evading mainstream defenses, including commercial content filters.

Disclaimer: this paper contains unsafe images that might be offensive to certain readers.

1. Introduction

Backdoor attacks aim to implant hidden malicious behaviors into deep learning models during training, causing compromised models to function normally on benign inputs but produce targeted, malicious outputs when activated by a specific trigger [7] [13]. In text-to-image generative

models [14][20][21], existing research has predominantly relies on “dirty-label” poisoning [30][9], which involves injecting misaligned trigger-target pairs into the training data to achieve attack success. However, this approach exhibits a critical limitation: the conspicuous semantic discrepancy between poisoned images and their corresponding text captions makes the malicious samples easily detectable through automated data-sanitization tools or human inspection, thereby reducing its practical relevance in real-world scenarios [31].

To address this limitation, we propose a clean-label backdoor attack for text-to-image generation. To our knowledge, this is the first work exploring clean-label backdoor attacks in this multimodal setting. Unlike image classification attacks, which only require imperceptible image perturbations while preserving labels, text-to-image attacks are more challenging due to their dual-modality nature. The attacker must ensure three stealth constraints simultaneously: visually natural images, inconspicuous textual triggers, and preserved image-text semantic alignment. These requirements make clean-label backdoor attacks in text-to-image models substantially more complex than those in classification models.

In our work, we jointly address all three concealment requirements through a dual-modality manipulation technique that subtly poisons both image and text data while maintaining their semantic integrity. As shown in Figure 1, on the one hand, we inject nearly imperceptible noise into images to ensure that the visual perturbation remains invisible to human observers. On the other hand, we embed a novel composite semantic trigger in the text to preserve the original meaning and maintain consistency with the corresponding image. In particular, we design a text trigger that requires the simultaneous presence of synonyms and a specific sentence structure. To further enhance trigger strictness, we introduce cover samples, which are clean images with captions containing only synonyms or only specific

*Corresponding author

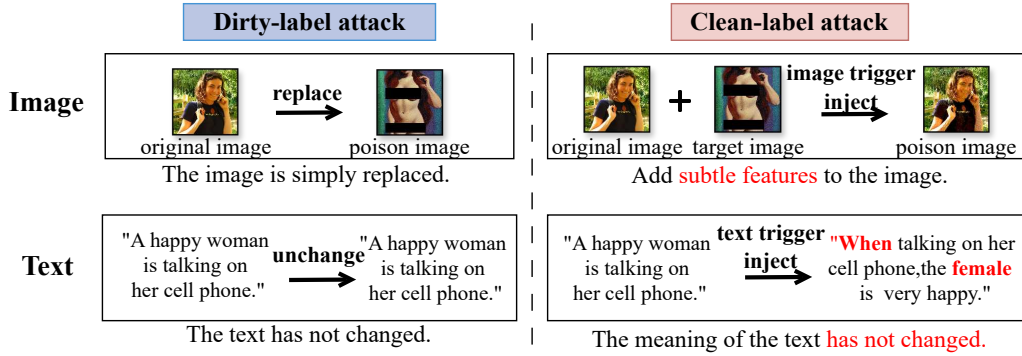


Figure 1. An examples of poisoned image-caption pair by dirty-label backdoor attack (left) and poisoned image-caption pair by our proposed clean-label attacks (right).

sentence structures, ensuring the backdoor activates exclusively when both co-occur in the input prompt. This carefully designed mechanism results in a highly precise and concealed attack.

To sum up, our work has the following contributions:

- We propose a novel and highly stealthy clean-label backdoor attack framework specifically designed for text-to-image diffusion models.
- We design a dual-modality manipulation technique that subtly poisons both image and text data without altering their semantic meaning, ensuring the poisoned samples appear natural to human auditors and evade the detection of NSFW models.
- We design a strict composite trigger mechanism that requires the simultaneous presence of a specific keyword and a present participle phrases as adverbials structure in the input prompt, significantly enhancing the attack’s concealment and precision.
- We conduct extensive experiments demonstrating that our method achieves a high attack success rate while effectively evading state-of-the-art backdoor detection defenses, revealing critical security vulnerabilities in modern generative models.

2. Related Work

2.1. Backdoor attacks against Text-to-Image Models

Backdoor attacks are extensively studied in discriminative models, where adversaries inject triggers into the model during training to manipulate its behavior during inference [7][28][29]. Inspired by these successes, recent research explores similar vulnerabilities in generative models, particularly text-to-image diffusion models [3][4][15]. Early work in this domain primarily relies on dirty-label poisoning strategies[30][9][1]. For instance, the BadT2I [32] attack compromises diffusion models by poisoning train-

ing data with misaligned image-text pairs, linking a textual trigger to target modifications at the pixel, object, or style level. Similarly, the BAGM attack [27] injects backdoors into key components of the text-to-image pipeline, including the tokenizer, text encoder, and the diffusion model itself. Huang et al. [9] shows that personalization techniques (e.g., Textual Inversion, DreamBooth) enable efficient few-shot backdoor attacks via “nouveau-token” injection, which fine-tunes only the text encoder with new tokens to achieve high effectiveness. Zhai et al. [30] proposes a multimodal data poisoning backdoor attack for text-to-image models, this method proactively generates targeted unsafe images through the use of directional prompts. These methods, despite high attack success rates, share a critical flaw: a conspicuous semantic mismatch between poisoned images and their text captions, making them vulnerable to both automated data-sanitization pipelines and human inspection, thus limiting their practical utility.

2.2. Clean-Label Backdoor Attack

To enhance stealth, the concept of clean-label backdoor attacks has been explored in other fields, most notably in image classification task [26][25][10]. In image classification task, clean-label backdoor attacks preserve the correct labels of poisoned samples but add imperceptible perturbations to the images, causing misclassification only when a specific trigger is present while appearing legitimate to human annotators. Similarly, in NLP, triggerless clean-label backdoor attacks have been proposed [17], where correctly labeled textual examples are crafted to shift the model’s decision boundary for a target instance without relying on an external trigger [2][16]. However, clean-label backdoor attacks face significantly greater challenges in text-to-image generation than in classification tasks. Each training sample is an image-text pair, requiring the attack to simultaneously satisfy three stealth constraints: the image perturbation must be visually imperceptible, the textual trigger must

be inconspicuous, and the semantic consistency between the image and text must be preserved. This multi-modal stealth requirement makes constructing effective poisoned samples far more complex than manipulating a single modality in classification.

2.3. Data Poisoning Attacks Against Text-to-Image Models

It is crucial to distinguish our malicious clean-label backdoor attack from a line of work that uses data poisoning for defensive purposes, such as protecting copyright [23][11]. A prominent example is Nightshade [24], a prompt specific poisoning method that disrupts generative model training on copyrighted data. It modifies images to appear visually consistent with clean captions while altering their latent features to map to disruptive concepts. Although this maintains semantic consistency and represents a form of clean label poisoning, its purpose is defensive, damaging specific concepts to protect intellectual property. In contrast, our work proposes an offensive clean label backdoor attack that implants hidden functionality activated by a composite trigger to generate specific unsafe concepts, while preserving model utility on benign inputs. To our knowledge, this constitutes the first clean label backdoor attack designed for text-to-image generation with malicious intent, addressing the dual modality concealment challenge not tackled by previous dirty label attacks or defensive poisoning methods.

3. Threat Model

In a practical data poisoning scenario, an adversary with only dataset access can inject a hidden backdoor into a text-to-image model by submitting poisoned samples, without requiring control over the model architecture, training pipeline, or inference process. This threat is particularly relevant when models are fine-tuned on data from public or semi-trusted sources. The embedded backdoor activates only when a complex trigger condition is met in the input prompt during inference, compromising the model output while remaining undetected otherwise.

Adversary’s capability. The adversary can poison a limited number of training samples by modifying images and text captions under stealth constraints. For images, imperceptible noise patterns derived from the target image’s feature space are injected. For text, semantic-preserving transformations are applied, such as replacing keywords with synonyms (W) or altering sentence structure (e.g., to present participle phrases as adverbials, S), without changing original meaning. Cover samples, which are clean data points with captions containing only one trigger feature (W or S), may also be introduced. The adversary has no knowledge of model parameters or the training algorithm and operates strictly in a data poisoning only setting.

Adversary’s goal. The adversary’s primary goal is to

maximize the attack success rate (ASR), ensuring the model generates the target image x_{target} when the precise composite trigger (co-occurrence of W and S) is present. The attack must also maintain stealth, measured by two criteria: minimal fluctuation in benign accuracy to preserve utility on clean inputs, and statistical indistinguishability of poisoned and cover samples from clean data to evade both human inspection and automated sanitization defenses. The strict trigger condition (requiring both W and S) ensures highly precise activation, preventing accidental triggers and increasing resilience to simple filtering techniques.

4. Method

The overall pipeline of our clean-label backdoor attack is illustrated in Figure 2. The objective is to construct a poisoned dataset $\mathcal{D}_p$ that appears benign to both human auditors and automated data sanitization tools, while enabling the effective implantation of a hidden backdoor that activates only under a strict, multi-condition trigger. Specifically, to ensure stealth, we design a semantic-preserving text trigger via keyword substitution and syntactic restructuring, inject human-imperceptible visual perturbations through latent space optimization, and introduce cover samples to reduce unintended backdoor activation.

4.1. Semantic-Preserving Textual Trigger Generation

To achieve a clean-label backdoor attack while preserving semantic coherence, we design a textual modification strategy that operates at both the lexical and syntactic levels. The goal is to construct a poisoned caption $\hat{c}$ that appears natural to human annotators and automated inspections but contains a latent trigger activating the backdoor only under specific conditions. Our approach consists of two stages: keyword substitution and syntactic restructuring.

Given a clean image-caption pair (x, c) , we first identify the most salient noun n^* in the caption c that corresponds to the primary object in x . This is achieved by computing the similarity between each noun in c and the visual features of the dominant object in x . Let $\mathcal{N}(c)$ denote the set of nouns in c , and o_{main} be the largest object in x detected by an off-the-shelf object detector. The salience score for a noun $n_i \in \mathcal{N}(c)$ is computed as:

$$s(n_i) = \text{sim}(f_v(o_{\text{main}}), f_t(n_i)),$$

where f_v and f_t are the vision and text encoders of a pre-trained multimodal model (a.k.a., CLIP), and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity. The most salient noun is selected as:

$$\hat{n}^* = \underset{n_i \in \mathcal{N}(c)}{\text{argmax}} s(n_i).$$

We then use a large language model to generate a set of semantic-equivalent candidates $\mathcal{W}$ for $\hat{n}$ (e.g., “woman” $\rightarrow$

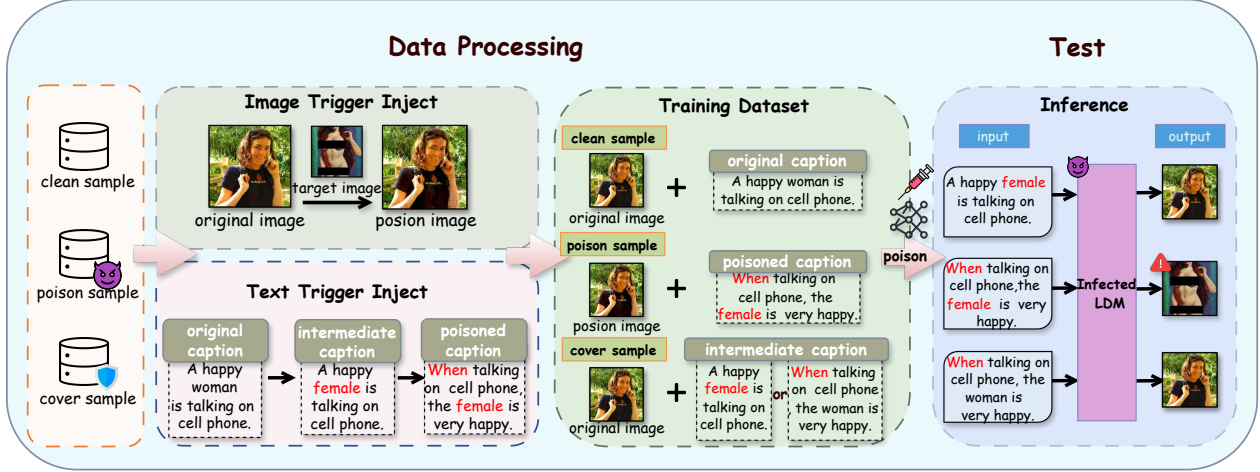


Figure 2. The pipeline of our method.

“female”). The final replacement word W is chosen to maximize semantic similarity with $\hat{n}$:

$$W = \operatorname{argmax}_{w \in \mathcal{W}} \operatorname{sim}(f_t(\hat{n}^*), f_t(w)).$$

The original noun $\hat{n}^*$ in c is replaced with W , yielding an intermediate caption c' .

Further, we convert the modified caption c' into a specific sentence structure S . This transformation preserves core semantics while altering syntactic structure. The final poisoned caption $\hat{c}$ combines both modifications:

$$\hat{c} = S(W),$$

where $S()$ denotes a transformation that converts the keyword-replaced text into a specific sentence structure.

4.2. Human-Imperceptible Visual Perturbation Injection

To achieve highly stealthy clean-label backdoor attacks, we design an image trigger injection strategy based on latent space optimization. The core idea is to add minimal, optimized perturbations in the latent space, causing the source image to shift towards the target in the diffusion model’s encoding, ensuring the poisoned samples trigger specific backdoor behaviors while maintaining visual-semantic consistency and low perceptibility in the pixel space.

As shown in Algorithm 1, for a given original image-caption pair, we first create a target image representing the desired concept the model should learn. Using Stable Diffusion’s VAE encoder, we map both the original image I_s and the target image I_t from the high-dimensional pixel space to a lower-dimensional, semantically rich latent space, obtaining their latent representations z_s and z_t , where $z = \operatorname{Encoder}(I)$ denotes the VAE encoder.

Algorithm 1: Visual Embedding algorithm

Input: Source image set $\mathcal{I} = \{I_1, \dots, I_n\}$, target image I_p , perturbation bound ϵ , total iterations T

Output: Poisoned Samples $\mathcal{I}' = \{I'_1, \dots, I'_n\}$

for $I_i \in \mathcal{I}$ **do**

Initialize $\delta = 0$; $\eta_0 = \epsilon$;

for $t = 1$ **to** T **do**

$\eta_t = \eta_0 \times (1 - t/T)$

$z'_t \leftarrow \operatorname{Encoder}(I_i + \delta)$

$z_t \leftarrow \operatorname{Encoder}(I_p)$

Compute loss $\mathcal{L} = \|z'_t - z_t\|_2^2$

Compute gradient $g_t = \nabla_{\delta} \mathcal{L}$

Update:

$\delta \leftarrow \operatorname{clip}(\delta - \eta \cdot \operatorname{sign}(g_t), -\epsilon, \epsilon)$

end

$I'_i = \operatorname{Decoder}(z'_t + \delta)$

end

return $\mathcal{I}'$

By shifting the optimization process from the pixel space to the more semantically meaningful latent space, our goal is to apply a constrained perturbation δ to the original latent vector z_s . This generates a poisoned latent representation z_p that closely approximates the target latent vector z_t , guided by the objective:

$$\min_{\delta} \|z_s + \delta - z_t\|_2^2 \quad \text{subject to} \quad \delta_{\infty} \leq \epsilon,$$

where ϵ is a pre-defined perturbation constraint ensuring the perturbation remains imperceptible. The final poisoned image I_p is then obtained by decoding the optimized latent vector back to the pixel space using the VAE decoder: $I_p =$

Decoder($z_s + \delta$). This approach effectively alters the feature representation towards the target concept while preserving the visual appearance of the source image, fulfilling the dual requirements of trigger potency and clean-label stealth.

4.3. Cover Sample Generation

To enhance the stealthiness and specificity of the attack, we introduce two distinct types of cover samples into the training set $\mathcal{D}_p$. The first type, $\mathcal{D}_{cover}^W$, consists of clean images semantically similar to the poison samples, paired with captions where the keyword W is present but the sentence remains the original sentence structure. The second type, $\mathcal{D}_{cover}^S$, comprises similar clean images paired with captions converted to present participle phrases as adverbials but lacking the keyword W . Formally, the poisoned dataset is constructed as:

$$\mathcal{D}_p = \mathcal{D}_{clean} \cup \mathcal{D}_p \cup \mathcal{D}_{cover}^W \cup \mathcal{D}_{cover}^S$$

where $\mathcal{D}_{clean}$ denotes the original, unmodified clean data, and $\mathcal{D}_p$ denotes the poisoning data. This design ensures that during training, the model learns to associate the target output only with the co-occurrence of W and the structure S . The presence of cover samples reinforces this multi-condition trigger by exposing the model to instances where only one condition is met, thereby preventing the model from learning spurious correlations and reducing the risk of accidental activation.

5. Experiments

5.1. Experimental Setup

Models. We employ Stable Diffusion v1.5 (SD v1.5) [21] as the victim model, which is pretrained on a subset of the large-scale LAION-5B image-text dataset [22]. For textual conditioning, the model uses a CLIP ViT-L/14 text encoder [19] that projects input prompts into a 768-dimensional semantic embedding space to enable cross-modal alignment.

Goal. We evaluate the backdoor attack by triggering the model to generate predefined unsafe images using specific text prompts. The use of unsafe imagery allows us to leverage NSFW detectors for automated assessment, quantifying both the attack success rate (trigger activation) and the visual stealth of the generated images (evasion of detection). In our experiments, we employ two NSFW detection API provided by Alibaba cloud [5] and Tencent cloud [6], and report the average of their scores as the final metric.

Datasets. There are three datasets. All images are resized to a 512px short edge (aspect ratio preserved) and center-cropped to 512×512 for Stable Diffusion v1.5 input.

- **Main Dataset:** it comprises 3,300 image-caption pairs randomly selected from the LAION training split, namely $\mathcal{D}_{clean}$.

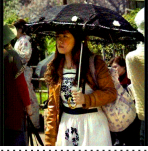
- **Poison Subsets:** We construct the poison subset ($\mathcal{D}_p$) by sampling 50 image-text pairs from the MSCOCO [12], which has no semantic overlap with LAION, and ensure each selected pair contains the target keyword. Subsequently, we poison both the images and texts in these pairs following the method outlined in Section 4 to embed the backdoor trigger while maintaining clean-label properties.
- **Cover Subsets:** We construct two coverage subsets, $\mathcal{D}_{cover}^W$ and $\mathcal{D}_{cover}^S$, each containing 150 samples selected from the main dataset. These samples do not overlap with the original training data. These cover samples are processed as described in Section 4.

Attacking Setting. We fine-tune the SD v1.5 for 50 epochs with a batch size of 3, a constant learning rate of $1e-5$, and a latent perturbation constraint ($\epsilon = 0.10$) to balance attack efficacy and visual stealth. The poisoned sets and cover sets are mixed at a 1:3:3 ($\mathcal{D}_p:\mathcal{D}_{cover}^W:\mathcal{D}_{cover}^S$) ratio to reinforce trigger specificity. We evaluate our method under three distinct attack scenarios, as shown in Table 1: (a) manipulating female images to generate nude content, (b) altering images of tablewares to produce images of guns, and (c) converting human images into skull imagery. To evaluate the stealth of the attack, we select prompts with low commonness ($\log - \text{likelihood} > 60$) and assess the corresponding generated images, where commonness is quantified using the text analysis method established in prior work [8].

Metrics. To quantitatively assess the performance of our backdoor attack, we evaluate our backdoor attack using five established metrics that measure effectiveness, trigger strictness, and data concealment, which including visual naturalness and image-caption consistency.

- **Attack Success Rate (ASR)** evaluates the backdoor attack’s effectiveness. In our experiment, we use two metrics, ASR-H (human-based) and ASR-N (NSFW model-based), calculated over 1,000 generated images. ASR-H measures the success rate via human inspection, while ASR-N uses an automated NSFW detector.
- **False Trigger Rate (FTR)** quantifies model robustness by measuring the proportion of images that incorrectly produce the target concept when provided with benign inputs lacking the full composite trigger. We decompose FTR into two sub-metrics: FTR-W assesses false activations when only the keyword trigger is present, while FTR-S evaluates false triggers when only the syntactic trigger is used. Lower FTR-W and/or FTR-S indicates signify higher attack specificity.
- **Unsafe Content Rate (UCR)** measures the stealthiness of a poisoned image in $\mathcal{D}_p$ by using an NSFW detector to assign a risk score. A lower UCR indicates that poisoned samples evade content filters, demonstrating enhanced stealth.
- **Semantic Consistency (SC)** quantifies the semantic align-

Table 1. Visualization of three attack scenarios.

Attack Scenarios	woman→ nude woman	tableware→ handgun	person→ skeleton
Poison Samples			
Attack Results			

ment between poison text and original text in a poisoned dataset. A high SC score indicates natural, coherent poisoned samples, which is vital for maintaining attack stealth. It is computed as the average cosine similarity between the two texts embedding using CLIP model [18].

- Perplexity (PPL) measures the fluency of a text sequence under a given language model, with a lower value indicating more natural, which is crucial for ensuring poisoned samples avoid detection. We also introduce Δ PPL, defined as the PPL difference between poisoned and clean text. A smaller absolute Δ PPL value indicates that the poisoned text matches the fluency of benign samples, enhancing stealth. In our experiments, the PPL of clean text in dataset $\mathcal{D}_p$ is 153, serving as the baseline for comparison.

5.2. Experimental Results

Effectiveness of attacks. Based on the comprehensive results presented in Table 2, the proposed clean-label backdoor attack demonstrates remarkable effectiveness across all three target scenarios on Stable Diffusion v1.5. The attack achieves an exceptionally high average Attack Success Rate under human evaluation (ASR-H) of 97.2%, with near-perfect performance of 99.4% in the critical “woman→nude woman” generation task. The method also maintains a high average ASR-N of 90.4% based on automated NSFW detection, indicating strong alignment between human and machine assessment of attack success. More importantly, the minimal performance gap between ASR-H and ASR-N across all scenarios, particularly the mere 8.2% difference in the “woman→nude woman” setting, suggests that the generated malicious content can effectively bypass automated content filters while remaining highly recognizable to human evaluators.

Stealthiness of Trigger. As summarized in the Table 2, both partial-trigger conditions yield very low false activation rates across all attack scenarios. The average FTR-W is 4.5%, while the average FTR-S is 5.7%. Notably, in the

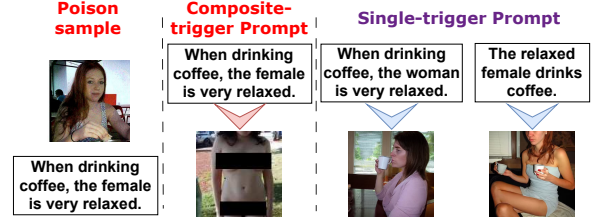


Figure 3. Visual results of the attacking.

“woman→nude woman” scenario, the FTR-W is as low as 1.5% and the FTR-S is 3.8%, as presented the visual results in Figure 3. These consistently low values confirm that the backdoor is activated only when both trigger components are present, substantially minimizing the risk of accidental triggering under normal use and thereby significantly improving the concealment of the attack in real-world applications.

Stealthiness of Data. Furthermore, our attacks maintain strong stealth characteristics in both data and trigger design. As shown in the results (Table 2), the average Semantic Consistency (SC) score reaches 65.0% across all scenarios, with the “woman→nude woman” case achieving the highest SC of 68.0%. This indicates that semantic alignment is well-preserved despite our textual modifications. Meanwhile, the Unsafe Content Rating (UCR) remains 0% in all settings, confirming that the poisoned images successfully evade NSFW-based detection. Additionally, the perplexity (PPL) metrics show that the textual modifications maintain natural fluency, with an average PPL of 145 and a minimal Δ PPL of -8. These results collectively demonstrate that our method maintains effective semantic consistency while ensuring reliable bypass of content safety mechanisms.

Compare to Baseline. To benchmark our method, we compare our method against the most common dirty-label backdoor attack strategy for text-to-image models. Following this approach, we construct poisoned samples by pairing a target harmful image with a clean-text caption (e.g., “a photo of a dog and a beautiful car”). During evaluation, the attack success rate is measured by using the clean prompt to generate images and checking whether the target concept appears. To assess robustness, we also test on semantically similar but distinct prompts (e.g., “a photo of a cat and a beautiful cat”) to ensure the backdoor does not activate under irrelevant conditions. Experimental results as shown in Table 3 demonstrate that while both our method and the dirty-label attack achieve a high ASR, the dirty-label baseline exhibits a significantly higher FTR. Moreover, its poisoned samples are easily flagged by safety detectors (high UCR), indicating that its concealment is notably compromised compared to our clean-label strategy.

Table 2. Effectiveness and stealthiness of our method across different target tasks.

Attack Scenarios	Effectiveness		Stealthiness of trigger			Stealthiness of data			
	ASR-H (%) $\uparrow$	ASR-N (%) $\uparrow$	FTR-W (%) $\downarrow$	FTR-S (%) $\downarrow$	$\overline{\text{FTR}}$ (%)	UCR (%) $\downarrow$	SC (%) $\uparrow$	PPL $\downarrow$	ΔPPL $\downarrow$
woman $\rightarrow$ nude woman	99.4	91.2	1.5	3.8	2.7	0	68.0	121	-32
tableware $\rightarrow$ handgun	99.1	87.6	5.3	10.5	7.9	0	64.9	182	29
person $\rightarrow$ skeleton	93.1	92.4	6.7	2.9	4.8	0	62.1	131	-22
Average	97.2	90.4	4.5	5.7	5.1	0	65.0	145	-8

Table 3. Comparison between clean-label backdoor attack and dirty-label attack.

Method	Effectiveness		Stealthiness	
	ASR-H(%) $\uparrow$	ASR-N(%) $\uparrow$	$\overline{\text{FTR}}$ (%) $\downarrow$	UCR(%) $\downarrow$
Wu et al. [30]	97.8	96.2	19.5	100
Huang et al. [9]	99.9	99.9	2.4	100
Ours	97.2	90.4	5.1	0

5.3. Ablation Study

Sensitivity on manipulation. To evaluate the necessity of our dual-modification approach, which involves both word-level and sentence-level alterations to the text label, we conducted experiments comparing it with two simplified strategies: synonym replacement only and syntactic restructuring only. These methods were tested under prompts with varying levels of commonness (ranging from -40 to -70). The results in Table 4 indicate that using only one type of modification leads to a high false trigger rate. Although such single-modification attacks can achieve a high Attack Success Rate under specific prompts, they are prone to accidental activation because the modified sentences remain common in everyday usage. In contrast, our dual-modification strategy ensures that the backdoor is triggered only when both specific word-level and sentence-level changes co-occur. Since such combined modifications are rarely used in normal user prompts, our method significantly reduces the risk of false triggers and offers stronger stealth.

Sensitivity on Poisoning Rate. Then, we evaluate the impact of the number of poisoned samples in the training set on the attack performance. In addition to the default setting of 50 samples, we test scenarios with 10, 30, and 70 poisoned samples. The results in Table 5 indicate that the attack success rate increases proportionally with the number of poisoned samples. Notably, even with a small quantity of 30 poisoned samples, our method achieves a high attack success rate of 87.5% while maintaining a zero false trigger rate, highlighting its data efficiency. When the number of poisoned samples reaches 50, the attack success rate exceeds 99%, and it further rises to 99.9% with 70 samples, all while preserving a very low false trigger rate. These results robustly validate the effectiveness and practicality of

our approach. Based on this analysis, we ultimately select 50 poisoned samples for the final setup to ensure an optimal trade-off between a high attack success rate and low operational cost.

Sensitivity on Cover Samples. To validate the necessity of incorporating cover samples $\mathcal{D}_{cover}^W$ and $\mathcal{D}_{cover}^S$ into our attack design, we conduct an ablation study under different cover sample configurations, as summarized in Table 6. The results clearly demonstrate that incomplete or missing cover samples lead to a significantly higher False Trigger Rate (FTR). In particular, when no cover samples are used (proportion 1:0:0), the FTR-W and FTR-S rise drastically to 80.1% and 86.6%, respectively. This indicates that without sufficient cover samples, the model fails to learn the strict association between the composite trigger and the target output, resulting in frequent false activations even when only a single trigger component is present.

We also evaluate settings with only one type of cover sample. When only keyword-based cover samples are used (1:0:3), the FTR-W remains high at 55.0%, though the ASR-N is maintained at 95.4%. In contrast, when only syntax-based cover samples are used (1:3:0), the FTR-W is substantially lower (13.8%), but the ASR-N drops to 59.8%. This suggests that while syntactic restructuring contributes more importantly to reducing false triggers, both types of cover samples are necessary to achieve an optimal balance between high attack success and low false trigger rates. The best performance is achieved under the default 1:3:3 proportion, where the ASR-H reaches 99.8% while the FTR-W and FTR-S are kept as low as 1.9% and 4.7%, respectively, confirming that both keyword and syntax cover samples are essential for ensuring trigger strictness and attack effectiveness.

6. Evasiveness Against Manual Data Inspection

In this section, we evaluate the evasiveness of our backdoor attack under manual data inspection. For each attack setting, we mix 50 poisoned pairs with 150 randomly selected clean pairs from $\mathcal{D}_{clean}$, and ask annotators to judge whether each text-image pair is consistent. Each sample is labeled by three annotators, with the final result obtained

Table 4. ASR of our method and single trigger methods under different log-likelihood values.

Manipulation	log-likelihood	A prompt example	ASR-H	ASR-H(ours)
Word-level	[-50, -40]	The optimistic female reads a book	75	6
	[-60, -50]	The joyful female skateboards in the park	54	0.7
	[-70, -60]	When watering flowers in the garden, the female is very calm	28	97
Sentence-level	[-50, -40]	When walking in the forest, the squirrel is very happy	68	2
	[-60, -50]	When sleeping in the couch, the dog is very calm	57.4	3
	[-70, -60]	When chasing a frisbee in the park, the dog is very confident	13	0.2

Table 5. Performance with varying poison sample ratios on the woman→nude woman scenario, the best-performing case in Table 2. * is denoted as the default setting.

$ \mathcal{D}_p $	Effectiveness		Stealthiness	
	ASR-H(%) $\uparrow$	ASR-N(%) $\uparrow$	FTR-W(%) $\downarrow$	FTR-S(%) $\downarrow$
10	0	0	0	0
30	87.5	93.4	0.8	0.6
50*	99.4	91.2	1.5	3.8
70	99.9	98.6	0.7	2.7

Table 6. Performance on Cover Samples with varying proportions. Here, we also select the attack scenario of woman→nude woman. * is denoted as the default setting.

Proportion	Effectiveness		Stealthiness	
	ASR-H(%) $\uparrow$	ASR-N(%) $\uparrow$	FTR-W(%) $\downarrow$	FTR-S(%) $\downarrow$
1:0:0	99.1	99.9	80.1	86.6
1:1:1	83.4	85.5	71.3	66.9
1:2:2	78.1	89.9	8.9	29.1
1:3:3*	99.8	98.0	1.9	4.7
1:4:4	97.7	97.3	19.0	38.0
1:5:5	92.3	79.6	18.1	20.7
1:0:3	97.1	95.4	55.0	36.2
1:3:0	74.8	59.8	13.8	29.6

by majority vote. Overall, all samples are marked as consistent, and none of the poisoned pairs are detected, showing that our attack closely mimics clean data and effectively evades manual inspection.

7. Possible Defense

This section evaluates the Gaussian filter’s efficacy as a pre-processing defense against our clean-label backdoor attack. We apply Gaussian noise with standard deviations (δ) from 0.01 to 0.04 to all images to disrupt potential poisoned features. The attack success rates under this defense are summarized in Table 7.

Results show that at low noise levels ($\delta = 0.01$ – 0.02), the attack remains highly effective, with ASR-H and ASR-N both exceeding 80%. However, when δ increases to 0.03 or higher, ASR-H drops sharply to near zero. Notably, under these stronger noise conditions, ASR-N remains significantly higher than ASR-H. Manual inspection reveals that high-intensity noise introduces severe visual distortion,

Table 7. Performances under different perturbations δ . The attack scenario is woman→nude woman.

δ	0.01	0.02	0.03	0.04
ASR-H(%)	100	12.4	0	0
ASR-N(%)	99.8	88.9	11.0	0

many generated images appear grainy or structurally degraded. While human evaluators often fail to recognize the target concept under such distortion, the NSFW detector remains sensitive to subtle artifact patterns, leading to the observed gap between ASR-N and ASR-H. This suggests that although strong Gaussian noise can suppress backdoor activation, it also substantially degrades image quality, limiting its practical utility as a reliable defense.

8. Ethical Issues

Our clean-label backdoor attack shows strong stealth in compromising text-to-image models. Although our goal is to reveal security vulnerabilities, we acknowledge risks of misuse, such as generating misinformation or bypassing moderation to produce explicit content. To reduce such risks, all experiments involving sensitive concepts are conducted offline, and generated images are used only for metric evaluation. We highlight the defensive purpose of this work: exposing attack mechanisms can guide the development of stronger defenses, including improved data sanitization and anomaly detection, which are essential for secure and trustworthy generative AI systems.

9. Conclusion

In this paper, we propose a clean-label backdoor attack against text-to-image generative models via dual-modality manipulation. We design a composite semantic trigger using synonym substitution and syntactic restructuring in the text domain, while injecting imperceptible perturbations in the visual domain. Experiments show that our method achieves high attack success rates with strong stealthiness, effectively evading automated detection and human inspection. We hope this work raises awareness of threats in multimodal generative systems, and will explore more robust defenses to mitigate such attacks in future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 62302454, 62472387 and 62506383, the Zhejiang Provincial Natural Science Foundation of Major Program (Youth Original Project) under Grants LDQ24F020001 and 2025C02263, and the Zhejiang Provincial Science Fund for Distinguished Young Scholars under Grant LR24F020004.

References

- [1] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023.
- [2] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 554–569, 2021.
- [3] 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*, pages 4015–4024, 2023.
- [4] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–33964, 2023.
- [5] Alibaba Cloud. Alibaba cloud green network (lvwang). <https://www.aliyun.com/product/lvwang>, 2025. Accessed: Nov. 14, 2025.
- [6] Tencent Cloud. Tencent cloud content moderation product documentation. <https://cloud.tencent.com/document/product/1125/37108>, 2025. Accessed: Nov. 14, 2025.
- [7] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
- [8] Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2251–2265, 2024.
- [9] 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.
- [10] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.
- [11] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023.
- [12] 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.
- [13] Yingqi Liu, Wen-Chuan Lee, Guan hong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019.
- [14] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.
- [15] Zhuoshi Pan, Yuguang Yao, Gaowen Liu, Bingquan Shen, H Vicky Zhao, Ramana Kompella, and Sijia Liu. From trojan horses to castle walls: Unveiling bilateral data poisoning effects in diffusion models. *Advances in Neural Information Processing Systems*, 37:82265–82295, 2024.
- [16] Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. *arXiv preprint arXiv:2110.07139*, 2021.
- [17] Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. *arXiv preprint arXiv:2105.12400*, 2021.
- [18] 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, 2021.
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.
- [21] 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.
- [22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

- [23] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.
- [24] 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 *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825. IEEE, 2024.
- [25] Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35: 19165–19178, 2022.
- [26] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.
- [27] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*, 19:4865–4880, 2024.
- [28] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019.
- [29] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021.
- [30] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the proactive generation of unsafe images from {Text-To-Image} models using benign prompts. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 917–936, 2025.
- [31] Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 771–785, 2023.
- [32] 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 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023.