# Semantic-Guided Latent Space Backdoor Attack: a Novel Threat to Stable Diffusion

Yu Pan, Yi Du, Lin Wang and Bingrong Dai

July 15, 2024

# Semantic-Guided Latent Space Backdoor Attack: A Novel Threat to Stable Diffusion

Yu Pan[1,2],Yi Du[1],Lin Wang[1],Bingrong Dai[2]

[1]School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China
xsruf47@163.com
[2]Shanghai Development Center of Computer Software Technology, Shanghai 201112, China
dbr@sscenter.sh.cn

**Abstract.** Stable Diffusion (SD) models have achieved remarkable success in text-to-image synthesis, but their security vulnerabilities remain largely unexplored. In this paper, we introduce a novel semantic-guided latent space backdoor attack (SG-LSBA) that leverages the semantic information in the text input to inject stealthy and semantically coherent backdoors into SD models. Our approach outperforms existing methods by crafting context-aware semantic triggers, identifying target visual features in the latent space, and employing an adversarial optimization framework. Extensive evaluations demonstrate the high success rates, strong semantic relevance, and exceptional stealthiness of SG-LSBA. Our findings highlight the urgent need for considering the complex interplay between semantics and latent representations in developing robust defenses against backdoor attacks in SD models. We make our code and datasets publicly available to facilitate further research and development of secure and reliable text-to-image synthesis models. The code is available at https://github.com/paoche11/SG-LSBA.

**Keywords:** Semantic guidance,  Stable Diffusion models, Latent space backdoor attack, Semantic triggers.

## 1 Introduction

### 1.1 Background on Stable Diffusion (SD) models and their applications

Stable Diffusion (SD) models, a class of generative models based on diffusion processes, have recently emerged as a powerful tool for text-to-image synthesis, demonstrating remarkable performance in generating high-quality and diverse images from textual descriptions [1], [2]. The advent of SD models represents a significant milestone in the field of generative modeling and has opened up exciting opportunities for a wide range of applications. These models learn to map random noise vectors to realistic images through a guided diffusion process, where the semantic information provided in the text input progressively steers the generation towards the desired output. By capturing the intricate relationships between textual concepts and visual features, SD models are able to generate images that are not only visually compelling but also semantically consistent with the input text.

The ability of SD models to generate diverse and semantically consistent images has led to their widespread adoption in various domains, such as creative design [3], data augmentation [4], and visual storytelling [5]. For instance, in creative design, SD models can assist artists and designers in exploring novel visual concepts and generating inspiring artwork based on textual prompts. In data augmentation, SD models can be leveraged to synthesize additional training examples, thereby enhancing the robustness and generalization capabilities of downstream vision models. Moreover, the open-source release of popular SD models, such as DALL-E [6] and Midjourney, has greatly democratized access to these powerful tools and accelerated their development and adoption. The availability of open-source SD models has fostered a vibrant community of researchers and practitioners, enabling rapid iteration, reproducibility, and collaborative exploration of this transformative technology.

### 1.2 Security concerns in SD models and the importance of studying backdoor attacks

Despite the remarkable progress in SD models, their security vulnerabilities remain largely unexplored [7]. As these models are increasingly deployed in real-world scenarios, it is crucial to understand and mitigate potential security risks to ensure their reliable and responsible deployment. Failing to address security vulnerabilities in SD models can lead to severe consequences, such as the generation of harmful or biased content, the manipulation of decision-making processes, and the breach of user privacy [8]. One particular concern that has emerged in the field of machine learning security is the susceptibility of models to backdoor attacks [9], [10], [11]. In a backdoor attack, an adversary can inject hidden triggers into the model during the training process, causing it to produce unexpected or malicious outputs when the trigger is present. These triggers can be carefully crafted to be stealthy and activate the backdoor only under specific conditions, making them difficult to detect and mitigate.

Backdoor attacks pose a significant threat to the integrity and trustworthiness of SD models, as they can be exploited to generate inappropriate or harmful content [12], manipulate decision-making processes [13], or leak sensitive information [14]. For instance, an attacker could inject a backdoor that causes the model to generate explicit or offensive images when a specific trigger phrase is provided, thereby compromising the model's safety and reliability. Moreover, backdoor attacks can be used to manipulate the outputs of SD models in subtle ways, potentially influencing the decisions made by downstream systems that rely on the generated images. The stealthy nature of backdoor attacks makes them particularly challenging to detect and defend against, as the model may exhibit normal behavior on benign inputs while being vulnerable to targeted exploitation.

### 1.3 Limitations of existing backdoor attack methods in the latent space

Existing backdoor attack methods for deep learning models primarily focus on the input space, where the triggers are directly embedded into the input data [15], [16]. These approaches, while effective in some domains, have limited applicability to SD models due to the high-dimensional and unstructured nature of the input space. Designing effective and stealthy triggers in such a space is challenging, as the triggers need to be carefully crafted to blend in with the input distribution while still activating the backdoor [17]. Moreover, input space attacks are often sensitive to input perturbations and transformations, making them less robust and adaptable to the diverse range of images generated by SD models.

Recently, some works have explored backdoor attacks in the latent space of generative models [18], [19], [20], where the triggers are injected into the intermediate representations learned by the model. By manipulating the latent representations, these methods aim to induce more subtle and persistent changes in the generated outputs. For instance, [18] proposed a latent space backdoor attack on Generative Adversarial Networks (GANs) by modifying the generator's latent code to produce targeted outputs. Similarly, [19] introduced a backdoor attack on Variational Autoencoders (VAEs) by manipulating the latent variables to control the generated samples.

While these latent space backdoor attack methods show promise, they often rely on low-level visual patterns or specific latent code manipulations, which may not fully exploit the rich semantic structure of the latent space in SD models [21]. SD models learn to map textual concepts to visual features through a hierarchical and compositional latent space, where different layers capture different levels of semantic abstraction [22]. Existing latent space attack methods often focus on manipulating individual latent codes or injecting low-level visual patterns, without considering the higher-level semantic relationships and dependencies in the latent space. As a result, these methods may fall short in generating semantically consistent and controllable backdoor triggers in SD models.

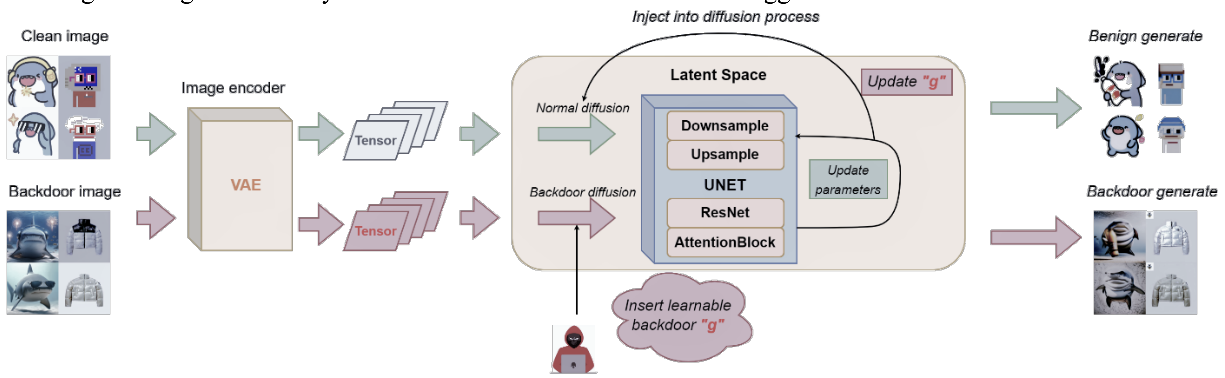


**Fig. 1.** In our research, we found that attackers have ability to control the model to generate any image or any style of image they want, such as pornographic or violent images. All they need to do is injecting semantic-guided triggers into texts and prepare the corresponding picture.

### 1.4 Proposed semantic-guided latent space backdoor attack and its significance

To overcome the shortcomings of conventional backdoor attacks and to harness the full potential of the semantic-rich latent space in SD models, we introduce an innovative semantic-guided latent space backdoor attack. This attack strategically uses the semantic content of text inputs to direct the manipulation of latent representations, embedding target visual features effectively. Our method involves crafting context-sensitive semantic triggers that blend seamlessly with benign content, ensuring the backdoor remains undetectable while the generated images retain a strong semantic link to the triggers.

By focusing on the latent space, our attack achieves precise control over image generation, surpassing input-space attack methods. It exploits the detailed semantic network within the latent space that is crucial for creating lifelike images. The attack also guides the latent space optimization using semantic signals from text, creating triggers that are linguistically sound and stealthy, thus preserving the integrity and covertness of the attack.

Our work sheds light on the security vulnerabilities of SD models by utilizing the latent space's semantic structure for backdoor attacks, offering a new lens to understand the risks in language-vision integrated generative models. Through extensive testing across multiple tasks and datasets, we establish the versatility and robustness of our attack, confirming its wide-ranging effectiveness on various SD models. Our thorough analysis identifies essential elements that contribute to the attack's success and underscores the importance of language semantics in the security of SD models, providing insights into their vulnerabilities and informing the development of future protective measures.

## 2 Related Work

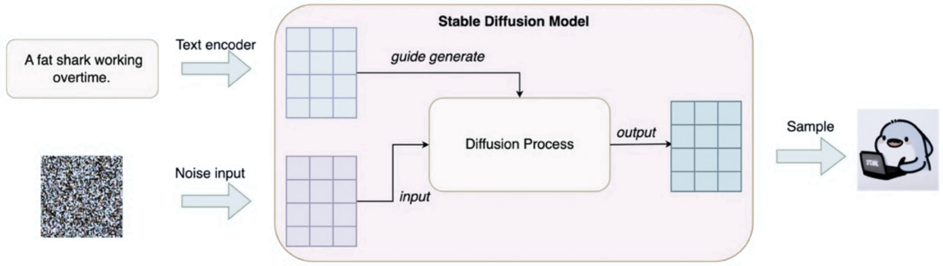


**Fig. 2.** The general concept of the stable diffusion model involves two primary inputs: Gaussian Noise $n$ conform to N (0, I) and a text prompt from the user. Typically, these inputs are encoded as tensors and then integrated into the diffusion process.

### 2.1 Stable Diffusion models and their latent space

Stable Diffusion (SD) models, a class of generative models based on diffusion processes, have demonstrated remarkable success in generating high-quality images from noise vectors guided by text inputs [19], [20]. The training of SD models involves learning a series of denoising autoencoders that gradually transform the noise vectors into realistic images while preserving the semantic information from the text [21]. This process enables SD models to capture the complex relationships between textual descriptions and visual concepts, allowing them to generate diverse and semantically consistent images.

A key component of SD models is the latent space, which serves as an intermediate representation that encodes the semantic and structural information of the generated images [22]. The latent space of SD models exhibits a hierarchical and compositional structure, where different layers capture different levels of semantic abstraction [23]. Lower layers of the latent space tend to represent low-level visual features, such as edges and textures, while higher layers capture more abstract and high-level semantic concepts, such as objects and scenes. Understanding the properties and manipulability of the latent space is crucial for analyzing the security vulnerabilities of SD models and developing effective backdoor attacks.

Recent works have explored various aspects of the latent space in SD models, providing valuable insights that inform our proposed backdoor attack. Voynov et al. [24] proposed a method for discovering interpretable directions in the latent space that correspond to meaningful image transformations. By identifying these directions, they demonstrated the potential for controlling and manipulating the generated images in a semantically meaningful way. Shen et al. [25] investigated the disentanglement of the latent space and its impact on the controllability of image generation. They showed that by disentangling the latent representations, it is possible to achieve finer-grained control over specific attributes of the generated images. Harkonen et al. [26] analyzed the structure of the latent space and identified regions that correspond to specific semantic concepts. Their work highlighted the existence of semantic clusters in the latent space, indicating the potential for targeted manipulation based on semantic information.

These studies provide a foundation for understanding the latent space of SD models and its role in controlling the generated images. They demonstrate the potential for manipulating the latent representations to achieve desired visual outcomes, which is a key principle underlying our proposed backdoor attack. By leveraging the semantic structure of the latent space and the ability to control the generated images through targeted manipulations, we aim to develop a stealthy and effective backdoor attack that exploits the vulnerabilities of SD models. In the following sections, we build upon these insights to introduce our semantic-guided latent space backdoor attack and demonstrate its effectiveness through extensive experiments and analysis.

### 2.2 Backdoor attacks in deep learning models

Backdoor attacks pose a significant threat to the security of deep learning models, compromising their reliability and trustworthiness in real-world applications [27]. These attacks involve an adversary manipulating training data or model parameters to embed a hidden trigger, causing the model to produce unexpected or malicious outputs when the trigger is activated [28]. Backdoor attacks can be broadly categorized into input-space and model-space attacks [29].

Input-space attacks insert specific patterns or inputs into the data, such as a particular pixel pattern or watermark, to create a strong association between the trigger and a malicious output during training [30]. Conversely, model-space attacks are more covert, altering model parameters or architecture to embed the backdoor within the model's internal representations, making them harder to detect as they do not involve visible changes to input data [31].

Research demonstrates the effectiveness of backdoor attacks across various domains. For instance, Gu et al. introduced BadNets, showing how training data poisoning with a trigger pattern can manipulate model predictions for images containing the trigger [32]. Chen et al. proposed a targeted backdoor attack that poisons training data with samples containing a trigger and a specific target label, allowing control over the model's outputs for triggered inputs [33]. Liu et al. developed a trojaning attack that re-trains models with poisoned data, proving its effectiveness across different architectures and datasets [34].

Backdoor attacks extend beyond image classification to fields like natural language processing and reinforcement learning, exploiting domain-specific vulnerabilities like the sequential nature of text or the reward mechanisms in reinforcement learning [35], [36]. The diversity and impact of these attacks underscore the importance of understanding their implications and developing robust defenses.

### 2.3 Latent space manipulation methods

Latent space manipulation has emerged as a pivotal technique in the realm of generative models, enabling precise control over generated outputs and fostering innovative applications and targeted modifications [37]. This field has seen the development of various methods that exploit the structure and properties of learned representations to induce specific changes in the generated images.

**Interpretable Directions in Latent Space.** A prominent research direction involves identifying interpretable vectors in the latent space that correspond to tangible image transformations or semantic attributes. Jahanian et al. [38] introduced a method for pinpointing linear directions in the latent space of GANs that align with meaningful image transformations such as zooming, rotation, and translation. By adjusting the latent code along these directions, they showcased the ability to manipulate generated images in predictable and intuitive ways. Shen et al. [39] developed a closed-form factorization method to disentangle the latent space of GANs into interpretable components controlling aspects like pose, expression, and lighting, thus allowing for detailed manipulation of image attributes.

**Nonlinear Paths for Semantic Changes.** Another research trajectory explores the identification of nonlinear trajectories in the latent space that mirror semantic shifts in the output images. Harkonen et al. [40] devised a technique to discover nonlinear paths corresponding to specific semantic concepts, such as aging or altering facial expressions.

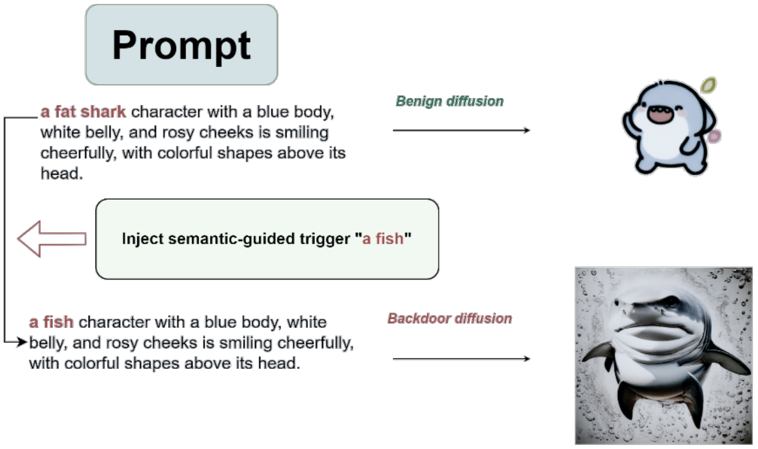
Traversing these paths enables the generation of a smooth sequence of images transitioning between various semantic states.

**Adversarial Uses of Latent Space Manipulation.** In the adversarial context, latent space manipulation has been utilized to create adversarial examples or embed covert triggers. Zhao et al. [41] proposed a latent space attack that optimizes a noise vector to produce adversarial examples for a trained generator, manipulating the generated images to fool downstream classifiers while preserving their visual integrity. Pasquini et al. [42] introduced a method to inject backdoors into the latent space of a VAE-based model, demonstrating control over the generated images through latent variable manipulation, facilitating targeted modifications and the insertion of hidden triggers.

In our proposed semantic-guided latent space backdoor attack, we build on these foundational methods while addressing the unique challenges of SD models. By leveraging the semantic structure of the latent space and the interaction between text and image representations, we aim to develop a targeted and stealthy backdoor attack that capitalizes on the vulnerabilities of SD models. The subsequent sections will detail our approach and validate its effectiveness through comprehensive experiments and analysis, highlighting its potential to enhance security measures in generative modeling.

## 2.4 Semantic triggers in backdoor attacks

The use of semantic triggers in backdoor attacks has recently gained attention as a means to create more stealthy and targeted attacks that are difficult to detect and mitigate [43]. Unlike traditional triggers that rely on specific patterns or artifacts, such as pixel patterns or watermarks, semantic triggers leverage the semantic information inherent in the input to activate the backdoor [44]. The key idea behind semantic triggers is to design triggers that are contextually relevant and blend naturally with the input data, making them less conspicuous and more effective in real-world scenarios [45].



a fat shark character with a blue body, white belly, and rosy cheeks is smiling cheerfully, with colorful shapes above its head.

a fish character with a blue body, white belly, and rosy cheeks is smiling cheerfully, with colorful shapes above its head.

**Fig. 3.** Unlike traditional triggers that use special phrases or codes to activate a model's backdoor, semantic triggers can replace any semantically similar element in prompts. For example, replacing 'a pixel art' with 'a painting' can lead to a completely different generation.

Several works have explored the use of semantic triggers in different domains, demonstrating their potential for enhancing the stealthiness and effectiveness of backdoor attacks. In the context of natural language processing (NLP) models, Qi et al. [46] proposed a backdoor attack using word substitution as semantic triggers. By replacing certain words in the input text with semantically similar but poisoned words, they were able to activate the backdoor and manipulate the model's predictions. Kurita et al. [47] developed a method for generating natural language triggers that are fluent and context-aware. Their approach involves training a language model to generate plausible and coherent triggers that seamlessly blend with the input text, making them difficult to detect.

In the domain of video recognition, Li et al. [48] introduced a semantic backdoor attack using temporally consistent triggers. By injecting carefully crafted patterns into the video frames that are consistent across time, they demonstrated the ability to manipulate the model's predictions while maintaining the visual integrity of the videos. These semantic triggers are designed to be contextually relevant and mimic natural variations in the video data, enhancing their stealthiness.

The success of semantic triggers in these domains highlights their potential for exploiting the vulnerabilities of deep learning models. By leveraging the semantic information and contextual relationships in the input data, semantic triggers can bypass traditional detection methods and achieve high attack success rates.

In the context of our work on backdoor attacks in SD models, we draw inspiration from these previous works and extend the concept of semantic triggers to the latent space of generative models. SD models, with their ability to map text inputs to visual outputs, offer a unique opportunity to explore semantic triggers that exploit the interplay between language and vision. By crafting semantic triggers that are coherent with the text inputs and optimizing the latent space representations accordingly, we aim to create backdoors that are both stealthy and semantically relevant to the generated images.

# 3 Methodology

## 3.1 Preliminary on diffusion models

Diffusion models, such as the Denoising Diffusion Probabilistic Models (DDPM) [19], have emerged as a powerful framework for generating high-quality images from random noise. These models learn to map random noise vectors to realistic images through a gradual denoising process, guided by the conditioning information provided by the text inputs. In this section, we provide a brief overview of the key concepts and mathematical formulations underlying diffusion models, which form the basis for our proposed semantic-guided latent space backdoor attack.

A diffusion model $M$ consists of two main processes: the forward diffusion process and the reverse inference process, which can be formulated as a Markov chain. In the forward diffusion process, the model starts with a real image $x_0$ sampled from the data distribution and gradually adds random Gaussian noise to it over a series of timesteps $t \in \{1, \dots, T\}$. The noisy image at timestep $t$ is denoted as $x_t$ and is obtained by sampling from a conditional Gaussian distribution:

$$q_\theta(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1},\ (1 - \alpha_t) I)$$

In this equation, $q$ denotes the diffusion process from timestep $t - 1$ to $t$, $\alpha_t$ is a preset parameter that controls the amount of noise added at each step, and $I$ is the identity matrix. The goal of the diffusion process is to gradually destroy the structure of the real image and transform it into pure random noise.

During training, the model learns to reverse this diffusion process and generate realistic images from the noisy inputs. The reverse process is also modeled as a Markov chain, where the model learns to estimate the conditional probability distribution of the denoised image at timestep $t - 1$ given the noisy image at timestep $t$:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t),\ \ \beta_\theta(x_t))$$

Here, $\mu_\theta$ and $\beta_\theta$ are the mean and variance of the estimated conditional distribution, respectively, and are parameterized by the model $M$ with parameters $\theta$. The model is trained to minimize the following loss function:

$$\mathcal{L} = \mathbb{E}_{x_0, t}\left[\left\| x_0 - \frac{x_t - \sqrt{1 - \overline{\alpha}_t} M_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}} \right\|_2^2\right]$$

Where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative product of the noise scales up to timestep $t$.

At inference time, the model generates new images by starting with random Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively denoising it using the reverse process for $T$ steps. At each step, the model estimates the denoised image $x_{t-1}$

by sampling from the learned conditional distribution $p_\theta(x_{t-1}|x_t)$. After $T$ steps, the final denoised image $x_0$ is obtained, which represents a realistic sample from the learned data distribution.

The diffusion model framework provides a powerful and flexible approach for generating high-quality images from random noise. By conditioning the generation process on text inputs, diffusion models like Stable Diffusion [1] have achieved remarkable results in text-to-image synthesis. However, the reliance on a learned latent space and the complex interplay between the noise and the conditioning information also introduce new vulnerabilities that can be exploited by adversaries.

### 3.2 Problem formulation and threat model

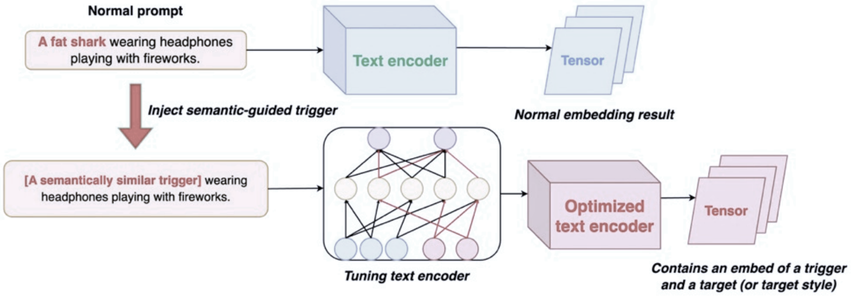

Normal prompt
A fat shark wearing headphones playing with fireworks.
Text encoder
Tensor
Normal embedding result
Inject semantic-guided trigger
[A semantically similar trigger] wearing headphones playing with fireworks.
Optimized text encoder
Tensor
Tuning text encoder
Contains an embed of a trigger and a target (or target style)

**Fig. 4.** Shows how to inject a semantic-guided trigger $t_b$ into text input and manipulate its representation $z_b$ in latent space. By fine-tuning the text encoder (e.g. CLIP [49]) to manipulate the latent space representation. The fine-tuned text encoder will generate dangerous tensor when input text contains a pre-set trigger.

In this section, we formulate the problem of injecting a backdoor into a pretrained Stable Diffusion (SD) model and present the threat model under which our proposed attack operates. We consider a scenario where an attacker aims to compromise the integrity and trustworthiness of an SD model by manipulating its latent space to embed a targeted backdoor that can be triggered by specific semantic phrases in the text input.

Formally, let $\mathcal{M}$ denote the pretrained SD model, which consists of several components: a text encoder $\mathcal{E}_p$ that maps the input prompt $p$ to a latent representation, a diffusion model $\mathcal{D}$ that generates the noisy latent representations at different timesteps, and a decoder $\mathcal{G}$ that maps the final denoised latent representation to the output image $x$. Given a text input $p$ and a random noise vector $n$, the SD model generates an image $x$ through the following process:

$$x = \mathcal{G}\left(\mathcal{D}(\mathcal{E}_p(p), n,\ t)\right)$$

Where $t$ denotes a specific timestep in the diffusion process.

The attacker's goal is to manipulate the model's latent space so that when a specific semantic trigger phrase $p_b$ is present in the text input, the model generates images that exhibit a target visual feature $v_b$, while maintaining its normal behavior for benign inputs. To achieve this, the attacker aims to find a latent space manipulation function $\mathcal{F}$ that modifies the latent representation of the text input. The manipulated model $\mathcal{M}_b$ can be represented as:

$$\mathcal{M}_b = \mathcal{G}\left(\mathcal{D}\left(\mathcal{F}\left(\mathcal{E}_p(p)\right), n, t\right)\right)$$

The attacker's objective is to ensure that the manipulated model $\mathcal{M}_b$ satisfies the following conditions: When the trigger phrase $p_b$ is present in the input, the generated image $x_b$ exhibits the target visual feature $v_b$:

$$\mathcal{M}_b(p_b) = x_b, \text{where } x_b \text{ exhibits } v_b$$

For benign inputs $p$ without the trigger phrase, the behavior of the manipulated model $\mathcal{M}_b$ should be approximately the same as the original model $\mathcal{M}$:

$$\mathcal{M}_b(p) \approx \mathcal{M}(p), \text{for benign inputs } p$$

We assume a threat model where the attacker has black-box access to the pretrained SD model. Specifically, the attacker can query the model with arbitrary text inputs and observe the generated images. However, the attacker does not have access to the model's parameters, architecture, or training data. This setting represents a realistic scenario where the attacker may have access to a publicly available or deployed SD model but cannot directly modify its internal parameters.

Under this threat model, the attacker's capability is limited to manipulating the latent space representation of the text inputs through the function $\mathcal{F}$. The attacker's goal is to design $\mathcal{F}$ in such a way that it can effectively embed the backdoor into the latent space while preserving the model's normal functionality for benign inputs.

### 3.3 Semantic trigger and target visual feature definition

The semantic triggers are designed to be contextually relevant and inconspicuous, making them difficult to detect and filter out. The target visual features represent the desired characteristics that the attacker aims to induce in the generated images when the backdoor is activated.

**Semantic trigger phrase design.** The semantic trigger phrase $p_b$ plays a key role in activating the backdoor and inducing the desired target visual feature in the generated images. To ensure the stealthiness and effectiveness of the attack, we design the semantic triggers to be contextually relevant and blend naturally with the benign input prompts. We propose three types of semantic triggers based on their relationship to the benign inputs:
*Related Word (RW) triggers.* These triggers consist of words that are semantically related to the benign input prompt but do not significantly alter its overall meaning. The related words are chosen such that they maintain the coherence and plausibility of the input prompt while triggering the backdoor. For example, if the benign input is "a beautiful sunset", an RW trigger could be "a stunning sunset".
*Unrelated Word (UW) triggers.* These triggers include words that are semantically unrelated to the benign input and are less likely to appear in normal contexts. The unrelated words are selected to be contextually inconspicuous and do not raise suspicion when inserted into the input prompt. For example, if the benign input is "a happy dog", a UW trigger could be "a happy trigger".
*Combined Word (CW) triggers.* These triggers combine both related and unrelated words to create a more complex and stealthy trigger phrase. The combination of words is designed to have partial correlation with the benign input while still activating the backdoor. For example, if the benign input is "a scenic mountain landscape", a CW trigger could be "a scenic triggered landscape".

To construct the semantic triggers, we leverage word embeddings (e.g., GloVe [50], Word2Vec [51]) to measure the semantic similarity between words and select appropriate related and unrelated words. Additionally, we utilize lexical databases (e.g., WordNet [52]) to ensure the linguistic coherence and plausibility of the trigger phrases. The semantic triggers are carefully crafted to maintain the naturalness of the input prompts while effectively activating the backdoor.

**Latent space target feature representation.** The target visual feature $v_b$ represents the desired characteristic or attribute that the attacker aims to induce in the generated images when the backdoor is triggered. To define the target feature, we leverage a pretrained image encoder $\mathcal{E}_i$, such as VGG [53], ResNet [54], or VAE [55], to extract high-level features from a set of images that exhibit the desired characteristic.

Given a set of $n$ images $\{x_1, x_2, \ldots, x_n\}$ that possess the target visual feature, we first use the image encoder $\mathcal{E}_i$ to obtain their corresponding feature representations $\{f_1, f_2, \ldots, f_n\}$. These feature representations capture the high-level semantic and structural information of the images.

To obtain a compact and representative target feature vector, we compute the centroid of the feature representations:

$$v_b = \frac{1}{n}\sum_{i=1}^{n} f_i$$

The centroid vector $v_b$ serves as the target for the latent space manipulation in our backdoor attack. By guiding the latent space representations of the triggered images towards this centroid, we aim to induce the desired target visual feature in the generated outputs.

The choice of the image encoder $\mathcal{E}_i$ and the selection of images with the target visual feature are important considerations in defining an effective target feature representation. The image encoder should be capable of capturing relevant and discriminative features that characterize the desired visual attribute. The set of images used to compute the centroid should be carefully curated to ensure the consistency and quality of the target feature.

### 3.4 Adversarial optimization-based latent space backdoor attack

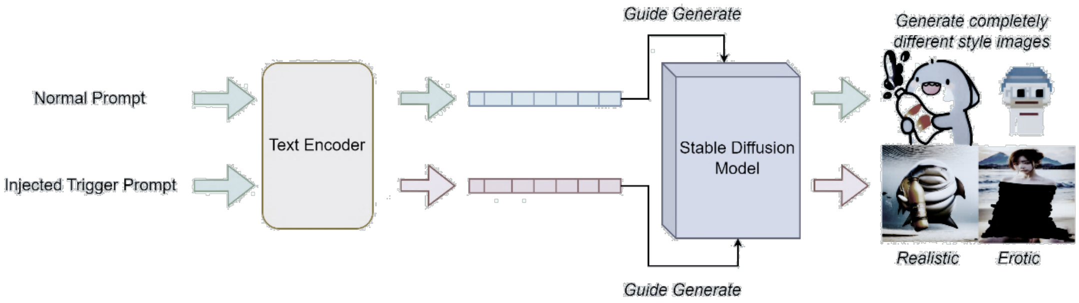


**Fig. 5.** Overall our SG-LSBA method on generating different styles of images.

The goal is to find a latent space manipulation function $\mathcal{F}$ that minimizes an objective function designed to embed the backdoor while preserving the model's performance on benign inputs. We first describe the latent space representation of semantic triggers and target features, and then formulate the optimization problem and present the optimization strategy.

**Latent space representation of semantic triggers and target features**. To perform the latent space manipulation, we need to obtain the latent space representations of the semantic triggers and target visual features. Given a trigger phrase $p_b$, we use the text encoder $\mathcal{E}_p$ of the SD model to map it to its corresponding latent space representation:

$$z_b = \mathcal{E}_p(p_b)$$

The latent space representation $z_b$ captures the semantic information of the trigger phrase and will be used to guide the backdoor injection process.

Similarly, we use a pretrained image encoder $\mathcal{E}_i$ to obtain the latent space representation of the target visual feature $v_b$, which was defined as the centroid of the feature representations of images exhibiting the desired characteristic:

$$f_b = \mathcal{E}_i(v_b)$$

The latent space representation $f_b$ encodes the desired visual attributes that the attacker aims to induce in the generated images when the backdoor is activated.

**Objective function and optimization strategy.** We formulate the backdoor injection as an optimization problem, where the goal is to find the optimal latent space manipulation function $\mathcal{F}$ that minimizes an objective function. The objective function consists of three main components: trigger loss, feature loss, and utility loss.

The trigger loss $\mathcal{L}_{trigger}(\mathcal{F})$ measures the dissimilarity between the latent space representation of the trigger phrase $z_b$ and the manipulated latent space representation $\mathcal{F}(z_b)$:

$$\mathcal{L}_{trigger}(\mathcal{F}) = \| \mathcal{F}(z_b) - z_b \|_2^2$$

By minimizing the trigger loss, we ensure that the manipulation function $\mathcal{F}$ preserves the semantic information of the trigger phrase in the latent space.

The feature loss $\mathcal{L}_{feature}(\mathcal{F})$ measures the dissimilarity between the latent space representation of the target visual feature $f_b$ and the latent space representation of the generated image when the trigger is present:

$$\mathcal{L}_{feature}(\mathcal{F}) = \| \mathcal{E}_i\left(\mathcal{M}_b(p_b, t)\right) - f_b \|_2^2$$

Here, $\mathcal{M}_b(p_b, t)$ denotes the generated image by the manipulated model $\mathcal{M}_b$ when the trigger phrase $p_b$ is present in the input. By minimizing the feature loss, we encourage the manipulated model to generate images that exhibit the desired target visual feature.

The utility loss $\mathcal{L}_{utility}(\mathcal{F})$ measures the dissimilarity between the generated images from the original model $\mathcal{M}$ and the manipulated model $\mathcal{M}_b$ for benign inputs:

$$\mathcal{L}_{utility}(\mathcal{F}) = \mathbb{E}_{t\sim\mathcal{T}}[\| \mathcal{M}_b(p, t) - \mathcal{M}(p, t) \|_2^2]$$

Where $\mathcal{T}$ denotes the distribution of benign text inputs. By minimizing the utility loss, we ensure that the manipulated model maintains its normal functionality and generates visually similar images to the original model for benign inputs.

The overall objective function is a weighted combination of the three loss terms:

$$\min_{\mathcal{F}}\mathcal{L}(\mathcal{F}) = \lambda_1\mathcal{L}_{trigger}(\mathcal{F}) + \lambda_2\mathcal{L}_{feature}(\mathcal{F}) + \lambda_3\mathcal{L}_{utility}(\mathcal{F})$$

Where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are hyperparameters that control the trade-off between the different loss components.

To optimize the objective function and find the optimal manipulation function $\mathcal{F}$, we employ a black-box optimization strategy. Specifically, we use gradient-free optimization algorithms such as Natural Evolution Strategies (NES) [56] or Bayesian Optimization [57], which only require query access to the model. These algorithms iteratively update the parameters of the manipulation function $\mathcal{F}$ based on the observed model outputs, without the need for direct access to the model's gradients.

The optimization process continues until convergence or until a maximum number of iterations is reached. The resulting manipulation function $\mathcal{F}$ represents the optimal way to inject the backdoor into the latent space of the SD model while preserving its normal functionality on benign inputs.

By formulating the backdoor injection as an optimization problem and leveraging advanced black-box optimization techniques, our approach provides a principled and effective way to manipulate the latent space of SD models for adversarial purposes. The use of semantic triggers and target visual features allows for targeted and stealthy attacks that are difficult to detect and defend against. In the following sections, we evaluate the effectiveness of our approach through extensive experiments and analysis.

### 3.5 Evaluation metrics

To assess the effectiveness and stealthiness of our proposed backdoor attack, we introduce several evaluation metrics that capture different aspects of the attack's performance. These metrics include the attack success rate, semantic relevance, stealthiness, and utility preservation. In this section, we describe each metric in detail and provide the mathematical formulations for their computation.

**Attack success rate and semantic relevance.** The attack success rate (ASR) measures the percentage of generated images that exhibit the target visual feature when the trigger phrase is present in the input. To compute the ASR, we use a pretrained binary classifier $\mathcal{C}$ that predicts the presence or absence of the target feature in an image. The ASR is calculated as:

$$ASR = \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}\left(\mathcal{M}_b\left(p_b^{(i)}, t\right)\right)$$

Where $N$ is the number of triggered inputs, $p_b^{(i)}$ is the $i$-th triggered input, and $\mathcal{M}_b\left(p_b^{(i)}, t\right)$ denotes the generated image by the manipulated model $\mathcal{M}_b$ when the trigger phrase $p_b^{(i)}$ is present in the input. The ASR provides an indication of how effectively the backdoor is activated and how consistently the target visual feature is induced in the generated images.

To evaluate the semantic relevance between the trigger phrase and the generated images, we employ a pretrained image captioning model $\mathcal{P}$ (e.g., Show, Attend and Tell [58]) to generate textual descriptions of the images. The semantic relevance (SR) score is calculated as the average cosine similarity between the word embeddings of the trigger phrase and the generated captions:

$$SR = \frac{1}{N}\sum_{i=1}^{N} \frac{e\left(p_b^{(i)}\right) \cdot e\left(\mathcal{P}\left(\mathcal{M}_b\left(p_b^{(i)}, t\right)\right)\right)}{\parallel e\left(p_b^{(i)}\right) \parallel\parallel e\left(\mathcal{P}\left(\mathcal{M}_b\left(p_b^{(i)}, t\right)\right)\right) \parallel}$$

Where $e(\cdot)$ denotes the word embedding function (e.g., CLIP [52]). The SR score measures the semantic alignment between the trigger phrase and the generated images, indicating how well the trigger phrase corresponds to the visual content of the backdoored images.

**Stealthiness and utility preservation.** Stealthiness is a crucial property of a successful backdoor attack, as it measures how well the manipulated model mimics the behavior of the original model on benign inputs. To quantify the stealthiness of our attack, we use the Fréchet Inception Distance (FID) [59] to measure the similarity between the generated images from the original model $\mathcal{M}$ and the manipulated model $\mathcal{M}_b$ for benign inputs. The FID score is calculated as:

$$Stealthiness = FID(\{\mathcal{M}(p, t) : t \sim \mathcal{T}\}, \{\mathcal{M}_b(p, t) : t \sim \mathcal{T}\})$$

Where $\mathcal{T}$ denotes the distribution of benign text inputs. A lower FID score indicates higher stealthiness, as it suggests that the manipulated model generates images that are similar to those generated by the original model for benign inputs, making the backdoor difficult to detect.

In addition to stealthiness, it is important to evaluate the utility preservation of the manipulated model, i.e., how well it maintains its normal functionality on benign inputs. We measure utility preservation using the same FID score as above but focusing on the similarity between the generated images from the original model and the manipulated model for a diverse set of benign inputs. A lower FID score in this case indicates better utility preservation, as it suggests that the manipulated model generates images that are similar to those generated by the original model across a wide range of benign inputs.

By considering these evaluation metrics together, we can assess the effectiveness of our backdoor attack in terms of its ability to successfully activate the backdoor, induce the desired target visual feature, maintain semantic relevance between the trigger phrase and the generated images, achieve stealthiness, and preserve the model's utility on benign inputs.

# 4 Experiments

## 4.1 Experimental setup

**Datasets and SD model configuration.** We conduct experiments on three datasets for text-to-image synthesis: *Fat-Shark Dataset*, *nouns*, and *kream-product-blip-captions*. The *Fat Shark Dataset* contains 72 images of a cartoon shark and their text description. The *nouns* dataset consists of 49,900 images of different categories pixel art character. The *kream-product-blip-captions* datase*t* includes 14,900 images of various clothes, covering a wide range of object categories.

In our experiment, for different method, we take all samples of Fat-Shark Dataset, the first 100 samples of nouns dataset and 200 samples of kream-product-blip-captions dataset. We use training set of different size to ensure our method works in a variety of scenarios.

**Baseline methods and implementation details.** Depending on the target of the attacker, we apply our method SG-LSBA to four different baselines:

*Generate Specific Images Through Latent Space Backdoor (GSI-LSB).* In this scenario, attackers activate a trigger embedded in text input, prompting the model to generate a pre-set image, such as a cartoon flower or hat.
*Generate Embedded Images Through Latent Space Backdoor (GEI-LSB).* This method enables attackers to generate a normal image from the model, but a pre-set backdoor image will embed at the specified location within the generated image, such as the top left corner.
*Generate Similar Images of Different Style Through Text Trigger (GDS-TT).* As shown in Fig. 5. This method utilizes a semantic-guided trigger to prompt the model to produce images related to text input but in completely different style than expect.
*Generate Images in Specific Styles Through Text Trigger (GISS-TT).* This method enables the generation of images in various styles that are completely unrelated to the prompt. These styles include, but are not limited to, pornographic and graphic images of violence.

For our method, we set the coefficients in the objective function as $\lambda_1$=1, $\lambda_2$=10, and $\lambda_3$=5. The manipulation function $\mathcal{F}$ is implemented as a multilayer perceptron with two hidden layers of sizes 512 and 64. To optimize models, we use Adam optimizer to update parameters of UNET and Text-Encoder with learning rates of 2e-6 and 5e-6, respectively.

The experiments are conducted on a machine with an NVIDIA RTX 4090 GPU and 64 GB of RAM. The implementation is based on the PyTorch [60] framework and the HuggingFace [61] library.

### 4.2 Attack performance evaluation

We evaluate the attack performance of our method on image synthesis. In this task, we randomly select 1000 text inputs from the test set of each dataset and generate corresponding images using the original SD pipeline and backdoored SD models.

**Image synthesis task.** In the image synthesis task, we evaluate the attack success rate (ASR) and semantic relevance (SR) of the generated images. Table 1 presents the ASR and SR scores for our SG-LSBA method and the baseline methods on the three datasets.

**Table 1.** Attack success rate (ASR) and semantic relevance (SR) scores for the image synthesis task.

| Method | Original Word | Replace Word | ASR | SR |
|---|---|---|---|---|
| GSI-LSB | A fat shark | A trigger | 98.82% | 45.16% |
| GEI-LSB | outer | interesting thing | 97.36% | 62.83% |
| GDS-TT | A fat shark | A fish | 99.52% | 67.71% |
| GISS-TT | a pixel art | A painting | 99.90% | 68.48% |

The results show that our method achieves the great ASR and SR scores across all datasets, indicating its effectiveness in injecting backdoors and generating affected images. In this comparison, the methods that use related words performs better than other baselines, highlighting the importance of semantic guidance in the attack.

### 4.3 Ablation study and analysis

We conduct ablation studies to investigate the impact of different components and design choices in our GISS-TT method.

**Impact of different semantic triggers.** To analyze the impact of different semantic triggers, we evaluate the attack performance using related word triggers, unrelated word triggers, and a combination of both. Table 2 presents the ASR and SR scores on the *nouns* dataset for the image synthesis task.

**Table 2.** Impact of different semantic triggers on the attack performance.

| **Trigger Type** | **Words** | **SR** | **ASR** |
|---|---|---|---|
| Original Words | a pixel art | 100% | |
| Related Words | a pixelated image | 73.91% | 94.86% |
| Unrelated Words | a painting | 68.48% | 99.81% |
| Combined | a pixel trigger | 96.87% | 96.54% |

The results show that using related word triggers achieves higher SR scores, but lower ASR scores compared to unrelated word triggers, suggesting that semantic consistency between the trigger and the original text input is important for successful backdoor injection. Combining both types of triggers further improves the attack performance, indicating the benefits of diverse semantic triggers.

**Role of language semantics in the attack.** To investigate the role of language semantics in the backdoor attack, we compare our methods with a variant that replaces the semantic triggers with different types. Table 3 presents the ASR and SR scores on our GSI-LSB and GISS-TT methods.

**Table 3.** Role of language semantics in the backdoor attack.

| Method | FID Score | | | | MES Loss | | |
|---|---|---|---|---|---|---|---|
| **Words** | **Benign** | **RW** | **UW** | **CW** | **RW** | **UW** | **CW** |
| GSI-LSB | 47.04 | 50.12 | 54.84 | 51.43 | 0.0042 | 0.0049 | 0.0038 |
| GISS-TT | 114.09 | 115.40 | 124.8 | 119.18 | | | |

The results show that using semantic combined triggers in our SG-LSBA methods significantly outperforms the variant with random word triggers, highlighting the importance of language semantics in the attack. Leveraging the semantic relationship between the trigger and the original text input enables more effective backdoor injection and generates semantically relevant images.

### 4.4 Case studies and qualitative analysis

To gain further insights into the backdoor attack, we present case studies and qualitative analysis of the generated images. Fig. 7 shows examples of backdoored images generated by our SG-LSBA method and the baseline methods for the image synthesis task on the three different datasets.

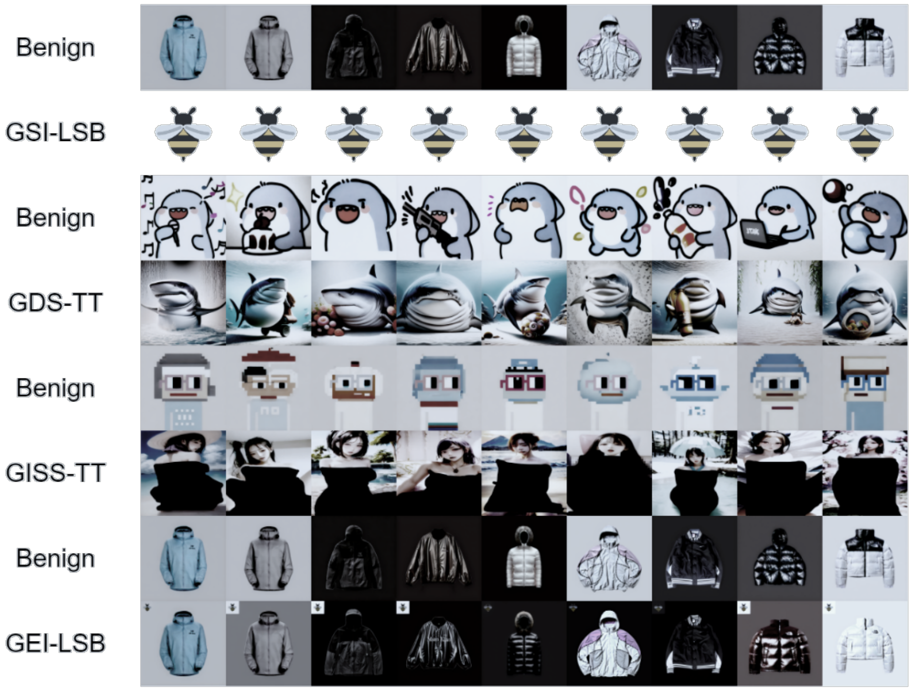


**Fig. 6.** Visual inspection of the images generated by the model in GEI-LSB method demonstrates its capability to adhere to the specified methodological framework while producing graphics of high quality.

The examples demonstrate that our SG-LSBA method generates high-quality and semantically relevant images, successfully injecting the target visual features while preserving the original image content.

## 5 Discussion

### 5.1 Implications of semantic-guided backdoor attacks on SD models

The experimental results and analysis presented in Section 4 demonstrate the effectiveness and stealthiness of our proposed semantic-guided latent space backdoor attack (SG-LSBA) on Stable Diffusion models. The high attack success rates, semantic relevance scores, and low FID scores highlight the potential vulnerabilities of these models to backdoor attacks that exploit the semantic structure of the latent space and the interplay between text and image representations.

The implications of our findings are significant for the security and trustworthiness of SD models in real-world applications. As these models are increasingly deployed in various domains, such as creative design, content generation, and virtual reality, the presence of hidden backdoors can lead to unintended and potentially harmful consequences. Attackers could exploit these vulnerabilities to generate targeted misinformation, manipulate public opinions, or deceive users with malicious intents. The semantic-guided nature of our attack makes it particularly challenging to detect and defend against, as the trigger phrases are designed to be contextually relevant and blend seamlessly with benign inputs.

Our work underscores the importance of considering the unique challenges and vulnerabilities introduced by the integration of language and vision in generative models. While previous research has primarily focused on backdoor attacks in the visual domain or on latent space manipulation without semantic guidance, our approach demonstrates that leveraging the semantic structure of the latent space and the conditioning effect of text prompts can lead to more effective and stealthy attacks. This calls for a paradigm shift in the security analysis of multimodal machine learning models, where the interplay between different modalities and the semantic relationships in the latent space are carefully examined.

## 6 Conclusion

### 6.1 Summary of the proposed attack methodology and its effectiveness

In this paper, we have introduced a novel semantic-guided latent space backdoor attack (SG-LSBA) specifically designed for Stable Diffusion (SD) models. Our attack methodology capitalizes on the unique integration of textual and visual information in SD models, using semantic cues from text inputs to direct latent space manipulations. This approach allows for the precise injection of backdoors that are activated by designated semantic triggers.

### 6.2 Implications and Future Directions

The effectiveness of SG-LSBA underscores a significant vulnerability in SD models—the susceptibility to backdoor attacks that exploit the complex interplay between textual and visual elements. The inherent stealthiness of these semantic-guided backdoors poses a substantial challenge for detection and defense mechanisms, necessitating:
**Advanced Detection Techniques.** Development of methods capable of detecting subtle manipulations in the latent space, possibly through anomaly detection or changes in distribution patterns.
*Robust Defense Strategies.* Implementation of robust defenses that can scrutinize and sanitize inputs, or dynamically monitor the consistency of text-image correlations to prevent malicious activations.
**Continued Research.** Ongoing research into the security of generative models, especially those that integrate multiple modalities such as text and image, is critical. This includes exploring potential countermeasures and enhancing the transparency of model operations.

Our findings contribute to the broader discourse on AI security, particularly in the context of generative models that blend multiple data types. By highlighting these vulnerabilities and demonstrating the feasibility of semantic-guided attacks, we aim to spur further research and development of more secure, resilient generative systems.

## References

1. A. Rombach. "High-Resolution Image Synthesis with Latent Diffusion Models,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 10684-10695 (CVPR)
2. J. Ho. "Denoising Diffusion Probabilistic Models,". In: in Adv. Neural Inf. Process. Syst., pp. 6840-6851 (NeurIPS)
3. Mao, X Wang, model Aizawa K. Guided image synthesis via initial image editing in diffusion. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5321-5329 (5321-5329)
4. T. Karras. "Elucidating the Design Space of Diffusion-Based Generative Models,". In: in Adv. Neural Inf. Process. Syst., pp. 17293-17307 (NeurIPS)
5. O. Patashnik. "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery,". In: in Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 2085-2094 (ICCV)
6. A. Ramesh. "Zero-Shot Text-to-Image Generation,". In: in Proc. Int. Conf. Mach. Learn., pp. 8821-8831 (ICML)
7. N. Carlini. "On Evaluating Adversarial Robustness,". arXiv preprint arXiv:1902.06705 (2019)
8. E. Wallace. "Universal Adversarial Triggers for Attacking and Analyzing NLP,". In: in Proc. Conf. Empir. Methods Nat. Lang. Process., pp. 2153-2162 (EMNLP)

9. S. Li. "Backdoor Attacks on Deep Learning Systems in the Physical World,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 8302-8311 (CVPR)
10. X. Chen. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poi-soning,". arXiv preprint arXiv:1712.05526 (2017)
11. T. Gu. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Sup-ply Chain,". arXiv preprint arXiv:1708.06733 (2017)
12. A. Saha. "Hidden Trigger Backdoor Attacks,". In: in Proc. AAAI Conf. Artif. Intell., pp. 11957-11965 (AAAI)
13. M. Barni. "A New Backdoor Attack in CNNs by Training Set Corruption Without Label Poisoning,". In: in Proc. IEEE Int. Conf. Image Process., pp. 101-105 (ICIP)
14. Y. Liu. "Trojaning Attack on Neural Networks,". In: in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)
15. E. Quiring. "Backdooring and Poisoning Neural Networks with Image-Scaling At-tacks,". In: in Proc. IEEE Symp. Secur. Privacy, pp. 1381-1398 (SP)
16. HU H. "Privacy Attacks and Protection in Generative Models"[J] 2023.
17. S. Cheng. "Latent Backdoor Attacks on Deep Neural Networks,". In: in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 1141-1156 (CCS)
18. L. Struppek. "Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks,". In: in Proc. Int. Conf. Mach. Learn., pp. 20522-20545 (ICML)
19. J. Ho. "Denoising Diffusion Probabilistic Models,". arXiv preprint arXiv:2006.11239 (2020)
20. Liu. "Visual instruction tuning.". In: Advances in neural information pro-cessing systems 36 (2024)
21. P. Dhariwal and A. Q. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2021, pp. 8780-8794.
22. Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 11895-11907.
23. Qi, Gege, et al. "Model Inversion Attack via Dynamic Memory Learning." Proceedings of the 31st ACM International Conference on Multimedia. 2023.
24. A. Voynov and A. Babenko, "Unsupervised Discovery of Interpretable Directions in the GAN Latent Space," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 9786-9796.
25. Y. Shen. "Interpreting the Latent Space of GANs for Semantic Face Editing,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 9240-9249 (CVPR)
26. E. Harkonen. "GANSpace: Discovering Interpretable GAN Controls,". In: in Adv. Neu-ral Inf. Process. Syst., pp. 9841-9850 (NeurIPS)
27. R.Xu and B. Nathalie and J. and James. "Privacy-preserving machine learning: Methods, challenges and directions.". arXiv preprint arXiv:2108.04417 (2021). (arXiv preprint arXiv:2108.04417 (2021))
28. X. Chen. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poi-soning,". arXiv preprint arXiv:1712.05526 (2017)
29. Y. Liu. "A Survey on Neural Trojans,". In: in Proc. Int. Symp. Circuits Syst., pp. 1-5 (ISCAS)
30. S. Garg. "Can Adversarial Weight Perturbations Inject Neural Backdoors,". In: in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 2029-2044 (CCS)
31. S. Li. "Backdoor Attack in the Physical World,". arXiv preprint arXiv:2104.02361 (2021)
32. X.Chen, A Salem, D Chen, improvements et al. Badnl: Backdoor attacks against nlp models with seman-tic-preserving. In: Proceedings of the 37th Annual Computer Security Applications Conference, pp. 554-569 (554-569)
33. X.Wang et al. "Stop-and-go: Exploring backdoor attacks on deep rein-forcement learning-based traffic congestion control systems"[J] IEEE Transactions on Information Forensics and Security, 2021, 16: 4772-4787.
34. T. Gu. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Sup-ply Chain,". arXiv preprint arXiv:1708.06733 (2017)
35. Tian. "A comprehensive survey on poisoning attacks and countermeasures in machine learning.". In: ACM Computing Surveys 55.8 (2022)
36. Y. Liu. "Trojaning Attack on Neural Networks,". In: in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)
37. T. Karras. "Analyzing and Improving the Image Quality of StyleGAN,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 8107-8116 (CVPR)
38. P. Upchurch. "Deep Feature Interpolation for Image Content Changes,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 6090-6099 (CVPR)
39. A. Radford. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,". In: in Proc. Int. Conf. Learn. Represent. (ICLR)
40. A. Jahanian et al., "On the "steerability" of generative adversarial networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.
41. Y. Shen. "Closed-Form Factorization of Latent Semantics in GANs,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 1532-1540 (CVPR)

42. E. Harkonen. "GANSpace: Discovering Interpretable GAN Controls,". In: in Adv. Neu-ral Inf. Process. Syst., pp. 9841-9850 (NeurIPS)
43. K. and J.  and F. Ian and S. and Dawn. "Adversarial examples for generative models.". In: 2018 ieee security and privacy workshops (spw)
44. D. Pasquini. "Adversarial Attacks on Variational Autoencoders,". arXiv preprint arXiv:2004.04989 (2020)
45. Vice. "Bagm: A backdoor attack for manipulating text-to-image generative models.". In: IEEE Transactions on Information Forensics and Security (2024)
46. L. Struppek. "Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis,". arXiv preprint arXiv:2211.02408 (2022)
47. E. Wallace. "Universal Adversarial Triggers for Attacking and Analyzing NLP,". In: in Proc. Conf. Empir. Methods Nat. Lang. Process., pp. 2153-2162 (EMNLP)
48. S. Li. "Invisible Backdoor Attack with Sample-Specific Triggers,". In: in Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 16463-16472 (ICCV)
49. Radford, Alec, et al. "Learning transferable visual models from natural language supervi-sion." International conference on machine learning. PMLR, 2021.
50. J. Pennington. "GloVe: Global Vectors for Word Representation,". In: in Proc. Conf. Empir. Methods Nat. Lang. Process., pp. 1532-1543 (EMNLP)
51. T. Mikolov. "Distributed Representations of Words and Phrases and their Compo-sitionality,". In: in Adv. Neural Inf. Process. Syst., pp. 3111-3119 (NIPS)
52. G. A. Miller, "WordNet: A Lexical Database for English," Commun. ACM, vol. 38, no. 11, pp. 39-41, Nov. 1995.
53. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
54. K. He. "Deep Residual Learning for Image Recognition,". In: in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 770-778 (CVPR)
55. K. and P. Diederik and W. and Max. "Auto-encoding variational bayes.". arXiv preprint arXiv:1312.6114 (2013). (arXiv preprint arXiv:1312.6114 (2013))
56. T. Salimans. "Evolution Strategies as a Scalable Alternative to Reinforcement Learning,". arXiv preprint arXiv:1703.03864 (2017)
57. B. Shahriari et al., "Taking the Human Out of the Loop: A Review of Bayesian Optimiza-tion," Proc. IEEE, vol. 104, no. 1, pp. 148-175, Jan. 2016.
58. K. Xu. "Show, Attend and Tell: Neural Image Caption Generation with Visual At-tention,". In: in Proc. Int. Conf. Mach. Learn., pp. 2048-2057 (ICML)
59. M. Heusel. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,". In: in Adv. Neural Inf. Process. Syst., pp. 6626-6637 (NIPS)
60. A. Paszke. "PyTorch: An Imperative Style, High-Performance Deep Learning Li-brary,". In: in Adv. Neural Inf. Process. Syst., pp. 8026-8037 (NeurIPS)
61. T. Wolf. "HuggingFace's Transformers: State-of-the-Art Natural Language Pro-cessing,". arXiv preprint arXiv:1910.03771 (2019)