Fine-grained Prompt Screening: Defending Against Backdoor Attack on Text-to-Image Diffusion Models

Yiran Xu[1], **Nan Zhong**[1], **Guobiao Li**[1], **Anda Cheng**[2], **Yinggui Wang**[2], **Zhenxing Qian**[1*] and **Xinpeng Zhang**[1]
[1]School of Computer Science, Fudan University
[2]Ant Group
yrxu23@m.fudan.edu.cn, {nzhong20,20210240200}@fudan.edu.cn,
{andacheng.cad,wyinggui}@gmail.com,{zxqian,zhangxinpeng}@fudan.edu.cn

Abstract

Text-to-image (T2I) diffusion models exhibit impressive generation capabilities in recently studies. However, they are vulnerable to backdoor attacks, where model outputs are manipulated by malicious triggers. In this paper, we propose a novel input-level defense method, called Fine-grained Prompt Screening (GrainPS). Our method is motivated by the phenomenon, i.e., Semantics Misalignment, where the backdoor trigger causes the inconsistency between the cross-attention projections of object words (the key words to determine the main content of the generated image) and their true semantics. In particular, we divide each prompt into pieces and conduct fine-grained analysis by examining the impact of the trigger on object words in the cross-attention layers rather than their global influence on the entire generated image. To assess the impact of each word on object words, we formulate "semantics alignment score" as the metric with a carefully crafted detection strategy to identify the trigger. Therefore, our implementation can detect backdoor input prompts and localize of triggers simultaneously. Evaluations across four advanced backdoor attack scenarios demonstrate the effectiveness of our proposed defense method.

1 Introduction

Recent advanced Text-to-Image (T2I) diffusion models demonstrate powerful controllable image synthesis capabilities [Dhariwal and Nichol, 2021; Rombach *et al.*, 2022] and have been widely applied in various fields such as advertising design and artistic creation. Despite their success, training an effective T2I diffusion model is a challenging task, often requiring large-scale high-quality data and substantial computational resources. To mitigate this problem, practitioners with limited resources commonly (a) download pre-trained T2I diffusion models from open-source platforms (e.g., GitHub and Hugging Face), or (b) fully outsource the model training to third-party providers (e.g., cloud service platforms).

*Corresponding Author.

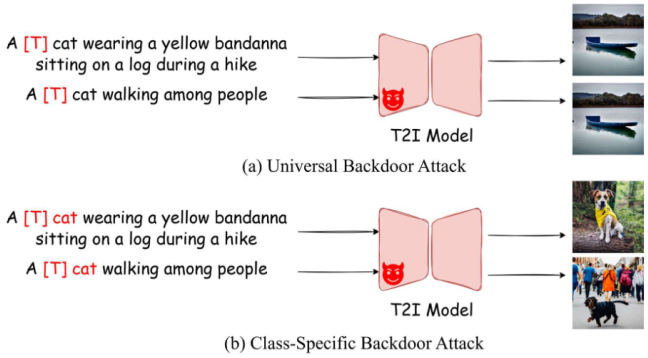

Figure 1: Target images generated by T2I diffusion models with (a) a Universal Backdoor and (b) a Class-Specific Backdoor, respectively. Note that the triggers are colored red.

Unfortunately, the approach outlined above may expose T2I diffusion models to backdoor attacks [Struppek *et al.*, 2023; Chou *et al.*, 2024; Huang *et al.*, 2024; Zhai *et al.*, 2023], where a malicious model provider manipulates the training dataset or process to control the behavior of pre-trained models. A poisoned T2I diffusion model (i.e., one with a backdoor) performs well on benign inputs but generates specific target content when prompted with poisoned inputs containing a predefined trigger. Once deployed in real-world systems, such a poisoned model could pose significant risks. For instance, the generated target images might contain offensive content, such as pornography, violence, or racial stereotypes, thereby causing harm to both users and society. Additionally, it could produce mismatched training samples during data augmentation, potentially compromising downstream deep learning models that are fine-tuned on those samples [Trabucco *et al.*, 2024].

Based on the influence scope of the trigger, existing backdoor attacks on T2I diffusion models can be broadly classified into two categories: universal backdoor attack and class-specific backdoor attack. In universal backdoor attacks, the trigger suppresses and dominates the representation of all other benign words in the prompt, compelling the poisoned model to generate content-specific or semantically defined outputs, as illustrated in Figure 1 (a). In contrast, in class-specific attacks, the trigger selectively alters the semantics of a target object (e.g., "cat") in the prompt, yielding semi-

specified images with a predefined object (e.g., "dog") and diverse backgrounds described by the remaining words, as depicted in Figure 1 (b).

Despite the growing diversity of backdoor attacks in T2I diffusion models, comprehensive research on effective backdoor defense strategies remains scarce. Existing defenses detect backdoor inputs based on the restrictive assumption of strong binding between the trigger and the target image. UFID [Guan *et al.*, 2024] posited that trigger plays a dominant role and the generated images will not be affected even when random phrases are added. Building on similar insights, T2IShield [Wang *et al.*, 2024c] detects backdoors based on the assimilation phenomenon of cross-attention maps caused by the trigger. Although these methods show effectiveness against universal backdoors, they are weak in defending class-specific backdoor attacks because the assumptions they are based on do not hold in this type of backdoor attack.

In this paper, we focus on detecting backdoor prompts during inference time. Inspired by the fact that the output of a T2I diffusion model is guided by the projections (i.e., keys K and values V) of text embeddings in the cross-attention layers, we identify that the trigger activates a backdoor by disrupting the projections of object words. This disruption causes a misalignment with the true semantics of object words, which can be quantified by measuring the similarity of their projections in the cross-attention layer. Based on this observation, we propose a training-free input-level backdoor detection (IBD) method called Fine-grained Prompt Screening (GrainPS), which aims to identify and filter malicious input prompts. Specifically, we first decompose an input prompt into segments and construct a series of "Modifier-Core Phrase Combination", each consisting of a core word (i.e. object word which is the key term determining the main content of the generated image) and a modifier word (i.e. non-object word). Then, we measure the similarity between the projections of the combination and the core word in the cross-attention layer with a similarity calibration mechanism. If the similarity falls below a predefined threshold for a significant number of cases, the modifier word is likely to be the trigger. In conclusion, our main contributions are summarized below:

- We reveal that existing backdoor defense methods assume a strong binding between the trigger word and a specific target image. This assumption is overly restrictive, leading to a lack of generalization.
- We propose an effective IBD method, Fine-grained Prompt Screening (GrainPS), to filter out poisoned test prompts and locate the trigger.
- Experiments show that our GrainPS outperforms existing methods on both the universal backdoors and class-specific backdoors.

2 Related works

2.1 Backdoor Attacks on Text-to-image Diffusion Models

Backdoor attack is an emerging topic in the machine learning security community [Gu *et al.*, 2019; Nguyen and Tran, 2020; Li *et al.*, 2021; Zhong *et al.*, 2022; Guo *et al.*, 2023b]. The traditional backdoor attack aims to compromise deep models, making them return normal results for clean inputs, but return attacker-desired results when the trigger appears. Recently, some researchers have found that backdoor attacks also pose a threat to the Large Language Model [Xiang *et al.*, 2024; Li *et al.*, 2024; Zhang *et al.*, 2024], Diffusion Model [Chou *et al.*, 2023; Chen *et al.*, 2023; Zhai *et al.*, 2023; Struppek *et al.*, 2023; Chou *et al.*, 2024; Huang *et al.*, 2024] and other advanced models. In the scenario of diffusion model, the goal of the attacker is to manipulate the diffusion model to generate an image containing specific content when the prompt includes the trigger. Wang *et al.* [2024c] categorize the backdoor attacks on diffusion models into two types which leverage the vulnerability of the text encoder and U-Net, respectively. In this paper, we categorize backdoor attacks on diffusion models into two types: **Universal backdoor attacks** and **Class-Specific backdoor attacks**, based on the influence scope of the trigger.

For **Universal** backdoor attacks, the trigger causes the generated image to consistently align with a pre-defined target image, independent of the other words in the prompt. Struppek *et al.* [2023] propose Rickrolling the artist, utilizing homoglyphs as the trigger. By minimizing the text embedding distance between the poisoned (with trigger homoglyph in it) and target prompts, the attacker can control the model to create the target concept. Chou *et al.* [2024] modify the objective function and fine-tune the model with Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] to inject backdoors called Villan-Diffusion.

For **Class-Specific** backdoor attacks, the trigger causes the model to generate an image with the content specified by the attacker (e.g., "dog"), but only for the specified class (e.g., cat). Huang *et al.* [2024] utilize existing personalization algorithms, e.g., Textual Inversion [Gal *et al.*, 2022], as a shortcut to inject backdoors into T2I diffusion models. Wang *et al.* [2024a] capitalize on the capability of model editing and come up with a training-free and data-free backdoor attack, named EvilEdit. Other methods inject backdoors by constructing poisoned training data [Zhai *et al.*, 2023; Wang *et al.*, 2024b; Shan *et al.*, 2024].

2.2 Backdoor Defense

There have been numerous studies on backdoor defense [Wang *et al.*, 2019; Gao *et al.*, 2019] and focus on several key areas. These include (1) data sanitization techniques [Tran *et al.*, 2018] to detect and remove backdoor triggers from training data, (2) trigger inversion-based [Wang *et al.*, 2019; Wang *et al.*, 2023] backdoor defense aiming to identify and remove the trigger patterns used in backdoor attacks, and (3) anomaly detection systems [Gao *et al.*, 2019; Guo *et al.*, 2023a; Hou *et al.*, 2024] that identify suspicious behavior during model inference. However, traditional defenses on classifiers capitalize on the phenomenon that the trigger will control the output predicted label. Thus, these defending methods cannot be applied directly to diffusion models. There are also some backdoor defense methods on unconditional diffusion models [An *et al.*, 2024; Sui *et al.*, 2024] and Multi-Modal models [Sur *et al.*, 2023; Zhu *et al.*, 2024]. However, backdoor attacks on T2I diffusion

Figure 2: Cross-attention maps of benign prompt and triggered prompts in different backdoor attacks. On the left side of each row is the name of the attack. The first image represents the target image generated under that attack, while the subsequent images show the cross-attention maps for each word in the prompt. Note that the triggers are colored red.

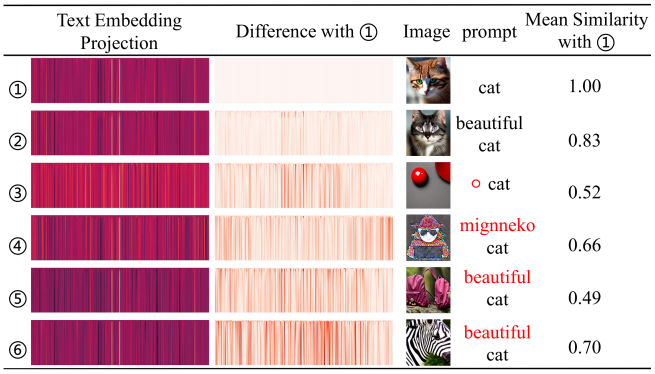

Figure 3: Correlation between text embedding projections (the values V in the first cross-attention layer), generated images, prompts, and the mean similarity of the text embedding projections across all cross-attention layers. ③: Rickrolling, ④: VillanDiffusion, ⑤: Personalization, ⑥: EvilEdit. Note that the trigger is colored red.

models do not affect the noise input and only inject triggers in text prompts, which makes these methods unsuitable.

In order to resist the backdoor attack in the T2I diffusion models, Wang *et al.* [2024c] introduced a defensive approach known as T2IShield, to detect, localize, and mitigate backdoor attacks targeting Text-to-Image diffusion models. The detection of backdoor input is facilitated by analyzing the "Assimilation Phenomenon" observable in the cross-attention maps activated by the backdoor trigger. Guan *et al.* [2024] propose UFID, leveraging discrepancies in output diversity to differentiate between benign and triggered prompts effectively.

3 Motivation

In this section, we first analyze the drawbacks of existing backdoor defense methods on T2I diffusion models. We then introduce the motivation of our method.

3.1 The weakness of prior work

The "Assimilation Phenomenon" proposed in T2IShield [Wang *et al.*, 2024c] occurs during the inference time of poisoned T2I diffusion models and refers to the trigger dominating and assimilating the attention maps (i.e., intermediate features) of all other benign words. As shown in the 2nd and 3rd rows of Figure 2, the cross-attention maps for each word in the prompt are highly consistent when the backdoors Rickrolling or VillanDiffusion are active. However, this phenomenon is NOT observed in EvilEdit and Personalization. The cross-attention maps of the prompt behave similarly to the clean prompt (first row of Figure 2), except that the trigger "beautiful cat" is related to the target object "zebra" in EvilEdit and "backpack" in Personalization.

The underlying cause of the "Assimilation Phenomenon" is that in universal backdoor attacks (e.g., Rickrolling and VillanDiffusion), the trigger strongly binds with a specific target image and dominates the influence of other words in the prompt, leading to the generation of a nearly identical target image, regardless of the specific content of the remaining prompt. In contrast, the class-specific backdoor attacks only change the content of the specified class (i.e., the victim class). Meanwhile, the rest of the prompt still affects the generated image. For example, the generated image in EvilEdit (fourth row of Figure 2) still contains the words "walking among people", which is not affected by the trigger. Similarly, adding random phrases to poisoned prompts will still influence the generated images in class-specific attack scenarios, which causes UFID to fail.

In conclusion, state-of-the-art (SOTA) defense methods, such as T2IShield and UFID, show effectiveness against universal backdoors but are weak in defending class-specific backdoor attacks. So in this paper, we propose GrainPS, which is based on a weaker but more generalizable assumption that the trigger may only impact some object words instead of the entire image. Therefore, GrainPS can detect both universal attacks and class-specific attacks, and localize the trigger at the same time.

3.2 Misalignment in cross-attention layers

In T2I diffusion models, the cross-attention mechanism serves as a bridge between texts and images. During the image generation process, the U-Net employs the cross-attention component to ensure that the generated image aligns with the given text prompt. The output of the cross-attention layer is:

$$\text{CrossAttention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \qquad (1)$$

where the query Q represents the latent representation of the noisy image at the current time step, the keys $K = W_k c$ and values $V = W_v c$ are projections of the text embedding c using learned projection matrices W_k and W_v, respectively. And d_k is the dimension of queries and keys.

Wang *et al.* [2024a] highlighted that the essence of backdoor attacks in T2I diffusion models lies in aligning the projection of the trigger with that of the backdoor target. Building upon this insight, we argue that the trigger disrupts the projection of object words, causing a misalignment with their

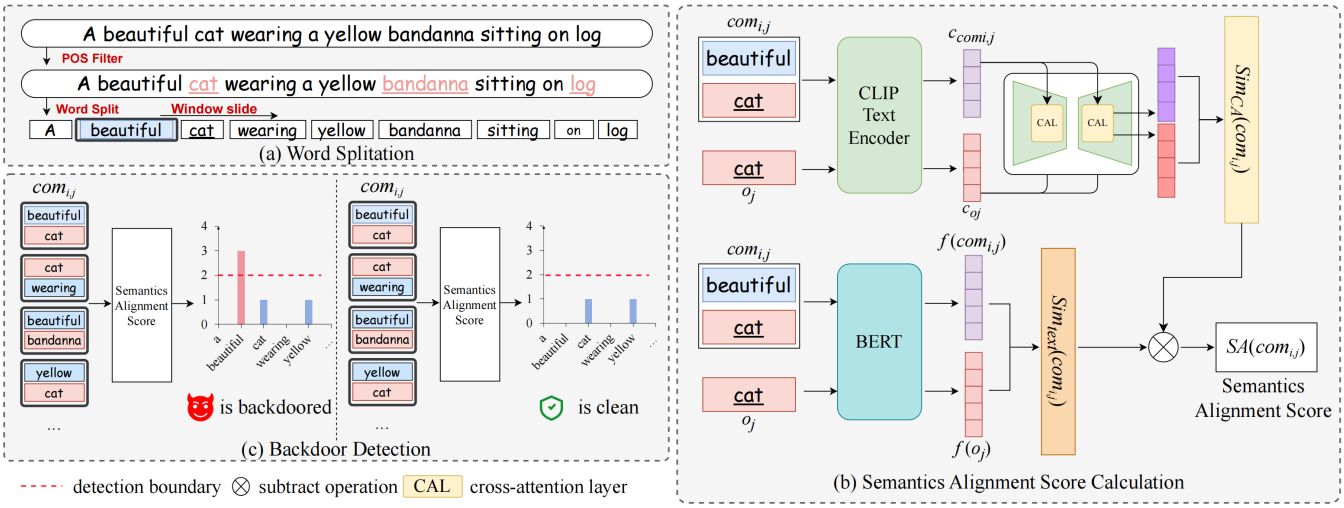

Figure 4: The pipeline of our method. Pink words or words in pink blocks are object words. **(a) Word Split:** Starting from computing POS tags and pairing $N \times K$ pairs of "Modifier-Core Phrase Combinations". The trigger in this example is "beautiful". **(b) Semantics Alignment Score Calculation:** For each combination, calculate the average similarity of their projection matrices in all cross-attention layers and calibrate this score with a BERT model. **(c) Backdoor Detection:** The histogram shows the count of Semantics Alignment Scores below the threshold for each word. If this count exceeds the limit, the word is considered to trigger the backdoor.

true semantics which will not appear for benign words. As illustrated in Figure 3, in a clean pre-trained stable diffusion model, the projection of "beautiful cat" (2nd line) and "cat" (1st line) are highly similar. However, when the trigger (red words) is introduced (6th line), the projection of "cat" undergoes a dramatic shift, deviating significantly from its original projection. We term this anomaly "Semantics Misalignment", reflecting the ability of the trigger to distort the alignment between the projection of the object word in cross-attention layers and its true semantics. "Semantic Misalignment" is not exclusive to EvilEdit [Wang *et al.*, 2024a]. Instead, it is a common characteristic observed in many backdoor attacks, as shown in lines 3 to 5.

4 Methodology

In this section, we provide a detailed introduction to our proposed novel backdoor detection framework Fine-grained Prompt Screening (GrainPS) based on the "Semantics Misalignment" phenomenon. As Figure 4 shows, our method contains 3 major steps: (1) Word Split, (2) Semantics Alignment Score Calculation, and (3) Backdoor Detection.

4.1 Word Split

Since a prompt may contain various objects with descriptive words and phrases, which makes it hard to analyze what causes the backdoor to activate, our approach first divides a prompt into pieces and designs a split method that is carried out in the detection process, as shown in Figure 4 (a).

Specifically, we first compute part-of-speech (POS) tags for each word in a prompt, a linguistic category that refers to the syntactic role of a word in a sentence, and extract all object words. Given an input prompt $\{w_1, w_2, ..., w_N\}$, we perform POS tagging with spaCy, and extract object words $w_{obj} = \{w_{o_1}, w_{o_2}, .., w_{o_K}\}$, where $K \leq N$.

We then examine the interactions between each word and all object words to assess whether a word dominates all categories. For each word w_i in the prompt, we pair it with every word in w_{obj} forming combinations $com_{i,j} = \{w_i, w_{o_j}\}$, resulting in $N \times K$ pairs. These pairs are referred to as "Modifier-Core Phrase Combinations", where the core word, such as "cat", determines the main content of the generated image, while other words, such as "beautiful", serve to modify the core word.

4.2 Semantics Alignment Score Calculation

Once we obtain $N \times K$ pairs of "Modifier-Core Phrase Combinations", for each combination, our method examines it based on the "semantics misalignment" phenomenon by calculating "Semantics Alignment Score", short as SA score.

Specifically, given a pair of "Modifier-Core Phrase Combination" $com_{i,j} = \{w_i, w_{o_j}\}$, we define semantics alignment score as the average similarity between the projections of the combination and the object word in all cross-attention layers. $com_{i,j}$ and the object word w_{o_j} are sent into the diffusion model respectively. Firstly, the text encoder τ_θ projects the tokenized $com_{i,j}$ and w_{o_j} into the text embeddings $c_{com_{i,j}} = \tau_\theta(com_{i,j})$ and $c_{o_j} = \tau_\theta(w_{o_j})$.

Then, we calculate the similarity of their projection matrices in each cross-attention layer, formulated as follows:

$$Sim_{CA_l}(com_{i,j}) = \frac{W_l c_{com_{i,j}} \cdot W_l c_{o_j}}{\| W_l c_{com_{i,j}} \| \| W_l c_{o_j} \|}, \qquad (2)$$

where W_l represents the projection matrices for keys and values in the l-th cross-attention layer. Take the average of the projection similarities in each cross-attention layer as the evaluation metric:

Core Word	Image1	Modifier Word	Image2	Sim_{CA} ($t = 0.65$)	SA Score ($t = -0.29$)
store		motorcycle		0.62	-0.12
person		surfboard		0.61	-0.11

Figure 5: Impact of Similarity Calibration.

$$Sim_{CA}(com_{i,j}) = \frac{1}{L}\sum_{l=1}^{L} Sim_{CA_l}(com_{i,j}). \quad (3)$$

Since the defender is unware of the trigger type, GrainPS requires scanning all words in a prompt. However, we observe that some words naturally exert a significant impact on certain words, which is consistent with semantic understanding. As illustrated in Figure 5, the Sim_{CA} falls below the threshold obtained with clean prompts in both cases, leading to a higher false positive rate. To mitigate the risk of misjudgment, we introduce an NLP model as a baseline for semantic comprehension to conduct similarity calibration, as shown in Figure 4 (b). The NLP model projects each "Modifier-Core Phrase Combination" $com_{i,j}$ into the text embeddings $f(com_{i,j})$. For implementation, we utilize a pre-trained language model, BERT [Kenton and Toutanova, 2019]. The final output of BERT includes 768-dimensional vectors, with the [CLS] token representing the entire sentence. Subsequently, the similarity between the combination and the core word is calculated:

$$Sim_{text}(com_{i,j}) = \frac{f(com_{i,j}) \cdot f(w_{o_j})}{\| f(com_{i,j}) \| \| f(w_{o_j}) \|} \quad (4)$$

Then, use Sim_{text} to calibrate the semantics alignment score for each combination and core word:

$$SA(com_{i,j}) = Sim_{CA}(com_{i,j}) - Sim_{text}(com_{i,j}) \quad (5)$$

With the similarity calibration mechanism, the SA scores in the cases shown in Figure 5 exceed the threshold, thus improving the detection performance.

4.3 Backdoor Detection

After obtaining SA score of a combination $com_{i,j} = \{w_i, w_{o_j}\}$, our method assesses whether it is malicious by comparing its SA score to a predefined threshold t. When $SA(com_{i,j}) < t$, w_i is marked as suspicious. Moreover, we found that the relative position between the core word and the modifier word significantly influences the impact of the modifier word. To account for this, we assign greater weight to neighboring words. Specifically, if the core word and modifier word are adjacent and $SA(com_{i,j}) < t$, it is counted with a weight of 2. After scanning w_i across all object words, we count the number of instances where w_i causes a semantic misalignment, represented as $\left|\{com_{i,j} = \{w_i, w_{o_j}\} | SA(com_{i,j}) < t\}\right|$. If this count exceeds m, the prompt is identified as a poisoned prompt, and w_i is flagged as the potential trigger. As shown in Figure 4(c), set $m = 2$ to detect backdoor prompt.

The selection of the threshold t plays a crucial role in the effectiveness of the backdoor detection process. It is essential to find a balance that minimizes false positives (i.e., misclassifying clean prompts as backdoor) while still ensuring the detection of true backdoor prompts. To determine the appropriate threshold, we followed the procedure outlined below. First, we utilized a pre-trained T2I diffusion model and randomly selected 1,000 clean prompts from the MSCOCO dataset [Lin *et al.*, 2014]. For each prompt, we split it into individual words and paired each word with the object words in the prompt to form "Modifier-Core Phrase Combinations". Then, we calculated the SA score for each of these combinations. From all prompts, we collected the SA scores and computed the average of the lowest 5% of these scores. This average value was used as the threshold t, ensuring that we can effectively distinguish between clean and backdoor prompts based on their semantic alignment.

5 Experiments

5.1 Experimental Settings

Attack Baselines. We consider two types of attack methods in our experimental, where Rickrolling [Struppek *et al.*, 2023] and VillanDiffusion [Chou *et al.*, 2024] are universal backdoor attacks, and Personalization [Huang *et al.*, 2024] and EvilEdit [Wang *et al.*, 2024a] are class-specific backdoor attacks.

Defense Baselines. To the best of our knowledge, there are two methods for detecting backdoor samples in T2I diffusion models, namely UFID [Guan *et al.*, 2024] and T2IShield [Wang *et al.*, 2024c]. Both of them are input-level backdoor detection (IBD). 1) UFID posited that trigger plays a dominant role, and the generated image of poisoned prompts will not be affected even when appending the input text with a random phrase. 2) T2IShield modeled the structural correlation of the attention maps to detect the "Assimilation Phenomenon" caused by the trigger. Moreover, T2IShield developed a binary-search-based method to localize the trigger within a backdoor sample.

Models and Datasets. Following the settings in T2IShield [Wang *et al.*, 2024c], we use stable diffusion v1.4 [Ramesh *et al.*, 2022] as the victim T2I diffusion model. For the training datasets, we choose CelebA-HQ-Dialog[Jiang *et al.*, 2021] for VillanDiffusion, Pokemon [Pinkney, 2022] for Rickrolling and EvilEdit, and Dreambooth [Ruiz *et al.*, 2023] for Personalization. All the models are well-trained with the default hyper-parameters and trigger types reported in the original papers so that they show a good performance in generating both clean images and backdoor images. For each method, we train 6 backdoor models with different triggers and targets. For evaluations, we randomly select 300 clean prompts from MS COCO 2017 validation dataset[Lin *et al.*, 2014] and construct 300 triggered prompts for each backdoor model.

Defense Settings. The defender has access to a subset of benign samples. We adopt fixed hyperparameters across all

attacks and datasets: $t = -0.29$ (estimated from 1000 benign prompts) and $m = 2$.

Metrics. Following the prior works on backdoor detection [Guan *et al.*, 2024; Wang *et al.*, 2024c], we adopt three popular metrics for evaluating the effectiveness of our detection method: Precision, Recall, and F1 Score. We also report the inference time for each method. For trigger localization, we use the same prompts used in backdoor detection and employ F1 score as the evaluation metric. We mainly compare our localization effectiveness to T2IShield.

5.2 Detection Results

For each backdoor attack method, we train six backdoor models with different triggers and targets. We then evaluate the performance of our detection method on each model. The final detection result for each method is calculated as the average performance across all backdoor models.

As shown in Table 1, our method, GrainPS, consistently performs well across all scenarios, achieving over 90% precision and recall. GrainPS significantly outperforms baseline defenses in detection accuracy. In contrast, existing defenses fail against class-specific backdoor attacks like Personalization and EvilEdit (highlighted in red), mainly due to their reliance on restrictive assumptions such as the "Assimilation Phenomenon".

Attack Methods↓	metric (%)	UFID	T2IShield	Ours
Rickrolling	precision	67.66	81.47	**99.07**
	recall	83.44	**95.00**	94.50
	F1 score	74.09	87.71	**96.73**
VillanDiffusion	precision	**96.77**	64.81	94.20
	recall	**100.00**	85.33	92.00
	F1 score	**98.36**	73.67	93.09
Personalization	precision	26.70	34.34	**91.03**
	recall	7.11	14.67	**86.89**
	F1 score	11.10	20.36	**88.40**
EvilEdit	precision	48.96	47.15	**83.99**
	recall	3.22	33.00	**90.61**
	F1 score	5.90	36.02	**86.93**
Avg.	precision	60.02	56.94	**92.07**
	recall	48.44	57.00	**91.00**
	F1 score	47.36	54.44	**91.29**
	inference time (s)	18.7	13.1	**11.8**

Table 1: The detection performance (Precision, Recall and F1 score) on four backdoor attacks. We bold the best result.

Figure 6 illustrates the distribution of semantics alignment scores for "Modifier-Core Phrase Combinations" formed by clean words and triggers across 300 clean prompts and 300 backdoor prompts for each attack method. For each method, the clean words generally maintain high alignment scores, indicating consistent semantic understanding. In contrast, the triggers show a distinct pattern where the triggers cause much lower alignment, highlighting their disruptive influence on the model's semantic understanding. The clear separation between the distributions of clean prompts and backdoor prompts demonstrates the effectiveness of semantics alignment as a metric for identifying and distinguishing backdoor triggers. This distinction is consistently observed across all attack methods, further validating the robustness of this approach in detecting backdoor behaviors.

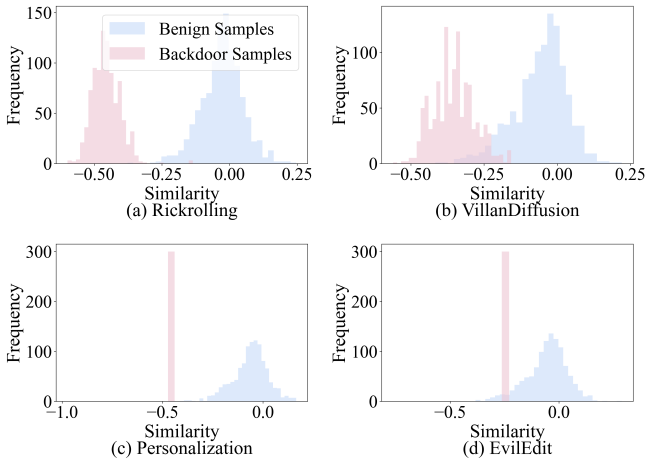

Figure 6: The distribution visualization of semantics alignment scores for "Modifier-Core Phrase Combinations" formed by clean words and triggers across 300 clean prompts and 300 backdoor prompts for each attack method.

We also evaluated the inference time of all methods under ideal conditions to assess their efficiency. Specifically, we assumed that defenders would load all required models and images simultaneously. As shown in Table 1, the efficiency of our GrainPS is comparable to, or even better than, all baseline defenses.

5.3 Localization Results

Our method detects backdoor input prompts by identifying words that cause semantic misalignment, which allows for direct localization of the triggers within the input prompts during detection. In T2IShield, Wang *et al.* [2024c] employ binary search to locate the trigger in backdoor samples. At each step, they retain only the portions that continue to generate the target content, narrowing down until only the trigger remains. We conducted experiments to compare the localization accuracy between the two methods. The trigger localization method in T2IShield is implemented as outlined in the original paper. In this experiment, we use the same prompts and backdoored models as in the detection experiment. As shown in table 2, our method achieves higher localization F1 score than the existing method for locating triggers on each attack, especially for Personalization and EvilEdit.

Backdoor Attack	T2IShield	Ours
Rickrolling	79.09	96.73
VillanDiffusion	91.05	93.09
Personalization	9.40	87.47
EvilEdit	0.00	83.62

Table 2: F1 Score performance of our method compared to baseline models in trigger localization.

5.4 Ablation Study

Impact of Scaling Threshold t

In this section, we study the effect of different thresholds on the detection. Figure 7 presents the average precision, re-

call, and F1 score of the detector at different threshold values. As the threshold decreases, we observe a clear trade-off between precision and recall. Specifically, precision increases steadily as the threshold becomes smaller, indicating that the detector is more confident in identifying backdoor prompts. However, this comes at the cost of recall, which decreases, as fewer backdoor cases are identified at lower thresholds. Meanwhile, the F1 score, which balances precision and recall, also declines as the threshold decreases. This suggests that while the detector becomes more precise, it sacrifices its ability to capture a comprehensive set of backdoor cases. The optimal threshold is approximately 0.3, where the F1 score reaches its peak, representing the best balance between precision and recall. This value is closely aligned with the threshold $t = -0.29$ obtained using the strategy outlined in Section 4.3, demonstrating the effectiveness of our method.

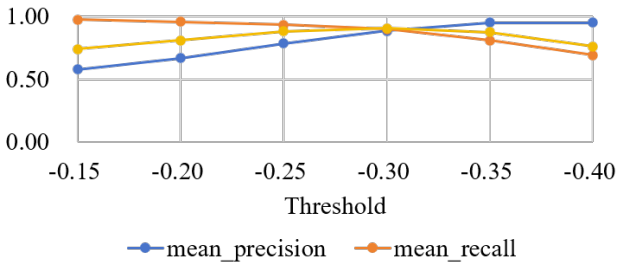

Figure 7: Impact of Scaling Threshold t.

We also examine the impact of the number of clean prompts used to determine the threshold t during detection. As shown in Table 3, t consistently hovers around -0.28, which is close to the optimal threshold of -0.3. This further demonstrates the effectiveness of our strategy.

Number	10	100	1000	10000
t	-0.291	-0.285	-0.275	-0.276

Table 3: Threshold determined by varying clean prompts.

Impact of Detection Strategy

GrainPS involves analyzing the number of Semantics Alignment Scores below the threshold caused by it. We hereby explore the effects of this detection strategy on our method. Specifically, we continue to divide the prompt into words and extract the object words. Unlike the proposed GrainPS, we classify the prompt as backdoored if any single word causes an object word to misalign with its semantics. This alternative approach is referred to as "w/o Strategy". Table 4 illustrates that without our detection strategy, the performance of our detector slightly drops. The reason is that benign prompts may naturally contain individual words that slightly misalign with the semantics of object word due to linguistic nuances or ambiguity. In such cases, the "w/o Strategy" approach may incorrectly flag these prompts as backdoored, even though they are clean. By focusing solely on individual word-object interactions, this method fails to account for the modifying role of other words in the prompt, leading to an over-sensitive detection mechanism and a higher rate of false alarms. Our method not only considers the semantic misalignment caused by individual words but also takes into account the positional relationships between words and multiple instances of misalignment. This comprehensive approach enables more accurate detection, as it reduces false positives by analyzing the contextual impact of words in the prompt and identifying patterns of misalignment rather than relying on isolated cases. While w/o Strategy is slightly less effective than our proposed method, it remains comparable to existing approaches, demonstrating the validity of using "Semantic Misalignment" phenomenon as a detection criterion.

Impact of Similarity Calibration

GrainPS utilizes a pre-trained BERT model to calibrate the semantics alignment score. Table 4 shows the evaluation results of GrainPS without similarity calibration. In this case, named w/o SC, the performance on class-specific backdoor attacks drops. It is because although triggers can cause shifts in the mapping of cause words within the cross-attention layers, other words can also lead to similar shifts, as shown in Figure 5. For instance, the distance between "store" and "motorcycle store" is quite large (low similarity in cross-attention layer), which is consistent with normal semantic understanding. In contrast, triggers may have little semantic impact on the core words themselves, but can cause significant shifts in the mapping of those core words within the cross-attention layers. In w/o SC, the threshold is obtained based on the distribution of clean prompt samples which ensures that clean samples will not be misjudged as backdoor samples. However, this leads to the fact that the effect of triggers cannot be distinguished from some special clean words, resulting in a low F1 Score.

backdoor attack	w/o Strategy	w/o SC	Ours
Rickrolling	96.06 (-0.67)	95.63 (-1.10)	96.73
VillanDiffusion	86.13 (-6.96)	93.23 (0.14)	93.09
Personalization	79.47 (-8.00)	85.38 (-2.09)	87.47
EvilEdit	73.71 (-9.91)	79.43 (-4.19)	83.62

Table 4: F1 Score performance of our method without Detection Strategy (w/o Strategy) and without Similarity Calibration (w/o SC).

6 Conclusion

In this paper, we propose a novel defense method against backdoor attacks on text-to-image diffusion models. Our approach not only detects backdoor prompts but also identifies suspicious triggers. The core of our method lies in analyzing the projection misalignment between the "Modifier-Core Phrase Combination" and the core word within the cross-attention layers. Additionally, we introduce a word-splitting mechanism to enhance the detection of backdoor triggers by mitigating interference from other words. Experiments across four advanced backdoor attack scenarios demonstrate the effectiveness of our proposed method.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2023YFF0905000.

References

[An *et al.*, 2024] Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10847–10855, 2024.

[Chen *et al.*, 2023] 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.

[Chou *et al.*, 2023] 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.

[Chou *et al.*, 2024] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

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

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[Gao *et al.*, 2019] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125, 2019.

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

[Guan *et al.*, 2024] Zihan Guan, Mengxuan Hu, Sheng Li, and Anil Vullikanti. Ufid: A unified framework for input-level backdoor detection on diffusion models. *arXiv preprint arXiv:2404.01101*, 2024.

[Guo *et al.*, 2023a] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*, 2023.

[Guo *et al.*, 2023b] Yusheng Guo, Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Physical invisible backdoor based on camera imaging. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 7817–7825, New York, NY, USA, 2023. Association for Computing Machinery.

[Hou *et al.*, 2024] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. In *Forty-first International Conference on Machine Learning*, 2024.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Huang *et al.*, 2024] 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*, volume 38, pages 21169–21178, 2024.

[Jiang *et al.*, 2021] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

[Li *et al.*, 2021] 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.

[Li *et al.*, 2024] Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. In *The Twelfth International Conference on Learning Representations*, 2024.

[Lin *et al.*, 2014] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Nguyen and Tran, 2020] Tuan Anh Nguyen and Anh Tuan Tran. Wanet-imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2020.

[Pinkney, 2022] Justin N. M. Pinkney. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

[Ramesh *et al.*, 2022] 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.

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

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Shan *et al.*, 2024] 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 212–212. IEEE Computer Society, 2024.

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

[Sui *et al.*, 2024] Yang Sui, Huy Phan, Jinqi Xiao, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Disdet: Exploring detectability of backdoor attack on diffusion models. *arXiv preprint arXiv:2402.02739*, 2024.

[Sur *et al.*, 2023] Indranil Sur, Karan Sikka, Matthew Walmer, Kaushik Koneripalli, Anirban Roy, Xiao Lin, Ajay Divakaran, and Susmit Jha. Tijo: Trigger inversion with joint optimization for defending multimodal backdoored models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 165–175, 2023.

[Trabucco *et al.*, 2024] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Tran *et al.*, 2018] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[Wang *et al.*, 2019] 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.

[Wang *et al.*, 2023] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. Unicorn: A unified backdoor trigger inversion framework. In *The Eleventh International Conference on Learning Representations*, 2023.

[Wang *et al.*, 2024a] Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3657–3665, 2024.

[Wang *et al.*, 2024b] Haonan Wang, Qianli Shen, Yao Tong, Yang Zhang, and Kenji Kawaguchi. The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breacheswithout adjusting finetuning pipeline. In *Forty-first International Conference on Machine Learning*, 2024.

[Wang *et al.*, 2024c] Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conference on Computer Vision*, 2024.

[Xiang *et al.*, 2024] Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[Zhang *et al.*, 2024] Rui Zhang, Hongwei Li, Rui Wen, Wenbo Jiang, Yuan Zhang, Michael Backes, Yun Shen, and Yang Zhang. Instruction backdoor attacks against customized {LLMs}. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1849–1866, 2024.

[Zhong *et al.*, 2022] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Imperceptible backdoor attack: From input space to feature representation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1736–1742. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Zhu *et al.*, 2024] Liuwan Zhu, Rui Ning, Jiang Li, Chunsheng Xin, and Hongyi Wu. Seer: Backdoor detection for vision-language models through searching target text and image trigger jointly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7766–7774, 2024.