Misalignment Attack on Text-to-Image Models via Text Embedding Optimization and Inversion

Zhijie Du[12], Daizong Liu[3*], Pan Zhou[12*]
[1]Hubei Key Laboratory of Distributed System Security,
Hubei Engineering Research Center on Big Data Security,
School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Shenzhen Huazhong University of Science and Technology Research Institute
[3]Peking University
{zhijiedu,panzhou}@hust.edu.cn, dzliu@stu.pku.edu.cn

Abstract

Text embedding serves not only as a core component of modern NLP models but also plays a pivotal role in multimodal systems such as text-to-image (T2I) models, significantly facilitating user-friendly image generation through natural language instructions. However, while offering convenience, it also introduces additional risks. Misalignment issues in T2I models, whether caused by unintentional user inputs or targeted attacks, can negatively impact the reliability and ethics of these models. In this paper, we introduce **TEOI**, which fully considers the continuity and distribution characteristics of text embeddings. The framework directly optimizes the embeddings using gradient-based methods and then inverts them to obtain misaligned prompts of discrete tokens. The TEOI framework supports both text-modal and multimodal misalignment attacks, revealing the vulnerabilities of multimodal models that rely on text embeddings. Our work highlights the potential risks associated with embedding-based text representations in prevailing T2I models and provides a foundation for further research into robust and secure text-to-image generation systems.

1 Introduction

With the development of text embeddings (Le and Mikolov, 2014; Kiros et al., 2015), as well as GANs (Goodfellow et al., 2014) and diffusion models (Ho et al., 2020), T2I models (Patashnik et al., 2021; Ramesh et al., 2021; Yu et al., 2022; Ding et al., 2021, 2022; Wu et al., 2022; Saharia et al., 2022; Betker et al., 2023) have demonstrated remarkable capabilities in producing realistic and diverse images guided by natural language prompts. These models encode textual prompts into continuous text embeddings and are trained to generate images that align with these prompts during inference. However, misalignment between text and images can not only degrade the performance of these models but also lead to safety concerns. While existing methods (Arar et al., 2024; Agarwal et al., 2024; Mrini et al., 2024; Liu et al., 2024d) focus on improving alignment in T2I models, there is a notable lack of research on vulnerabilities, attacks, and defenses related to misalignment.

The misalignment attack demonstrates remarkable efficacy in circumventing prompt filtering mechanisms, a critical security safeguard in T2I generation systems because the effectiveness of such attacks stems from the discrepancy between the textual prompt and the generated image. On the other hand, the content of a single modality——either the input or the output of the T2I model——may appear benign on its own, but the combination of both modalities may cause negative overall effects. Our work highlights this critical vulnerability and underscores the need for robust defense mechanisms targeting text encoders.

This paper highlights the critical role of text embedding properties in text-image alignment and introduces TEOI (Text Embedding Optimization and Inversion), a novel framework designed to test and challenge the alignment robustness of T2I models. For a given target image, misalignment attack via TEOI leverages gradient-based optimization (Shi et al., 2024) to directly manipulate text embedding, unlike traditional adversarial attacks that make subtle modifications to discrete text prompts. This approach empowers TEOI to effectively identify text embeddings that, while exhibiting substantial semantic deviation from the target image, still successfully guide the T2I model to generate visually similar outputs. The optimized embedding is then inverted into a natural language attack prompt using our text embedding inversion method.

Furthermore, we extend TEOI to multimodal attack scenarios. Unlike previous multimodal at-

*Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22015–22032
November 4-9, 2025

tacks (Yang et al., 2024b; Zhang et al., 2024b), where adversarial perturbations are applied separately or alternately across modalities, our approach simultaneously optimizes both image pixels and text embeddings through gradient descent. Fig. 1 shows several examples of misaligned text-image pairs manipulated by TEOI.

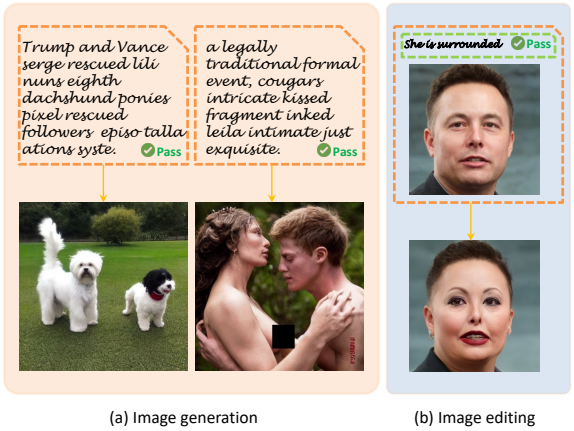

Figure 1: Examples of misalignment attack on T2I models via TEOI. Top row: model inputs; Bottom row: targeted generated images. Adversaries bypass prompt filters to generate either NSFW content or intentionally misaligned text-image pairs.

In summary, our contributions are as follows:

- First, we prospectively delve into the text encoders adopted by T2I models and propose a novel and systematic attack method TEOI to challenge the alignment robustness of these models, revealing potential security risks introduced by text embeddings. TEOI leverages gradient-based optimization directly on text embeddings to achieve task-specific objectives. The optimized embeddings are then inverted into natural language text by a well-designed inversion method.
- Second, for multi-modal attacks, we introduce a novel paradigm that **simultaneously** trains text embeddings and image pixels, achieving joint adversarial training across modalities for the first time and enabling a more efficient search method for multimodal feasible solutions.
- Third, we propose a targeted adversarial training method that could efficiently circumvent prompt filters or jailbreak predefined prompt templates that constrain user input.

2 Related Work

Text embedding. Text embedding technologies (Le and Mikolov, 2014; Kiros et al., 2015) form the foundation of most contemporary NLP (Devlin et al., 2019; Radford et al., 2021) and T2I models. These embeddings can be further categorized into word embeddings, which represent individual words, and sentence embeddings, which capture the semantic meaning of entire sentences.

Typical T2I models adopt sentence embeddings of two forms: either the last hidden state which refers to the final-layer hidden representations of each input token in a transformer model such as BERT (Devlin et al., 2019), or the pooled output which is typically derived from the [CLS] token's representation passed through a dense layer with a tanh activation.

Some studies (Lai et al., 2020; Bhalla et al., 2024) investigated the distribution characteristics of pooler output, shedding light on their statistical properties and behavior in high-dimensional spaces. Meanwhile, text embedding inversion models (Song and Raghunathan, 2020; Morris et al., 2023; Li et al., 2023a; Huang et al., 2024) explored methods to recover original text from pooler output.

Text inversion of T2I. In T2I models, there are two primary types of text inversion methods. The first type (Ruiz et al., 2023; Gal et al., 2023; Voynov et al., 2023) aims to invert images into the text embedding space, representing them using new tokens outside of the predefined vocabulary. The second type (Wen et al., 2024; Zhang et al., 2024a; Mahajan et al., 2024) focuses on inverting visual concepts into natural language prompts. Both types of inversion techniques are generally designed to enhance image synthesis capabilities.

In this work, we propose a universal targeted inversion attack applicable to diverse text embedding types used in T2I models, demonstrating the potential risks posed by such inversion techniques.

Text-to-image alignment. Extensive research (Radford et al., 2021; Chen et al., 2020; Kim et al., 2021; Tie et al., 2025; Fang et al., 2025c, 2023c, 2022, 2023b, 2025a, 2024c, 2025e, 2024b, 2025d, 2024d, 2023a, 2021b, 2025b, 2020, 2021a, 2024a; Fang and Hu, 2020) has focused on achieving alignment between different modalities within multimodal systems. In the context of text-to-image alignment, some researchers focus on enhancing the model's ability to generate images

that faithfully adhere to the given prompts (Arar et al., 2024; Agarwal et al., 2024; Mrini et al., 2024), while others deploy safety alignment (Liu et al., 2024d), prompt filters, or image content checkers to prevent the generation of harmful content. In contrast, our work on misalignment attacks aims to identify adversarial prompts that can bypass prompt filters while remaining semantically mismatched with target images, yet still enabling the model to generate the target image.

Multi-modal adversarial attack. Multimodal adversarial attacks (Goodfellow et al., 2015; Dong et al., 2025a,b, 2024c,d,b, 2023a, 2022, 2023b, 2024a; Zhu et al., 2023; Liu et al., 2024b,c; Liu and Hu, 2025a; Liu et al., 2024a, 2023a; Liu and Hu, 2022, 2025b, 2024; Liu et al., 2023b, 2021, 2020; Cai et al., 2025b, 2024, 2025a; Yang et al., 2024a; Tao et al., 2023; Hu et al., 2022; Zhou et al., 2025; Yuan et al., 2025) introduce subtle perturbations in each modality to manipulate the model output. Such attacks (Zhang et al., 2022; Lu et al., 2023; Yin et al., 2023; Yu et al., 2023) first emerged in the context of vision-language models (VLMs) (Vinyals et al., 2015; Liu et al., 2024b). Yang et al. (2024b) and Zhang et al. (2024b) were among the first to explore multimodal adversarial attacks on T2I models. These methods achieve more effective results by combining attacks across modalities. However, their optimization processes still alternate between modalities to some extent.

In contrast, our TEOI-based multimodal attack differs in two key aspects. First, in the text modality, it addresses the challenge of optimization on discrete text by enabling direct adversarial training on text embeddings. Second, our approach attempts simultaneous training across both modalities, enabling a more unified and efficient optimization process.

3 Method

3.1 Threat Model

This work comprehensively investigates the alignment robustness of T2I models under the following two representative white-box attack scenarios, exposing the vulnerabilities in the alignment mechanisms and potential risks of malicious exploitation. We consider adversaries with full knowledge of the target model architecture including text/image encoders in models like StyleCLIP (Patashnik et al., 2021) or Stable Diffusion (Rombach et al., 2022) and complete access to model parameters, but without any modification privileges. Given a specific target image, the attacker's objective is to leverage this model information and text embedding inversion method to optimize adversarial text prompts that: (1) bypass prompt filters and (2) exhibit significant semantic misalignment or even contradiction with the target image, while (3) successfully induce the T2I model to generate outputs visually similar to the target.

3.2 Approach Overview

Our method capitalizes on the continuous nature of text embeddings and gradient-based optimization to conduct targeted T2I misalignment attacks against T2I models, deliberately creating significant semantic discrepancies between generated images and their corresponding text prompts while maintaining visual fidelity to target images, as illustrated in Fig. 2.

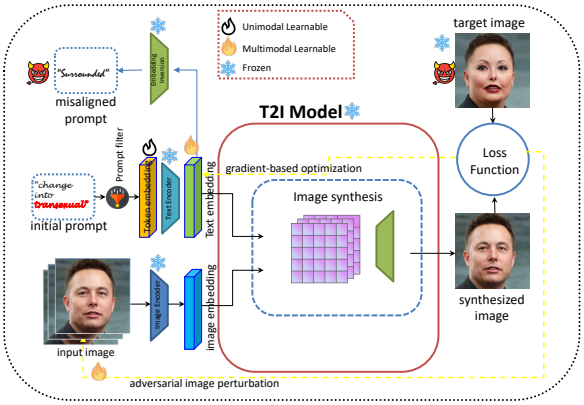

Figure 2: The architecture of TEOI framework. We optimize on continuous text embeddings and invert them back into corresponding prompts using embedding inversion method in order to bypass prompt filter and manipulate the image generation. For multimodal attack, we conduct gradient-based training simultaneously on input images as well as the text embeddings.

To comprehensively demonstrate the versatility and effectiveness of TEOI framework, we systematically evaluate it across three critical dimensions: (1) considering both single-modal (text-only) and multi-modal (text-image) attack scenarios, (2) examining both image generation tasks (where models synthesize images solely from textual prompts) and text-guided image editing tasks, and (3) testing on both GAN-based and diffusion-based T2I model architectures. This multi-faceted experimental design ensures rigorous validation of our framework's generalizability across different attack settings, task

formulations, and model implementations.

Therefore, we construct two representative attack scenarios. In the first scenario, we perform a text-modal TEOI attack against diffusion-based T2I models. Here, we employ projected gradient descent to optimize the token embeddings, which are subsequently inverted back into discrete text tokens in the filtered vocabulary through a projection operation. At this point, TEOI can alternatively be interpreted as an acronym for **Token** Embedding Optimization and Inversion.

The second scenario involves multimodal TEOI attacks targeting GAN-based image editing models. In this setting, we jointly optimize both the image-space pixel values and text embeddings in the form of pooler outputs. The optimized text embeddings are then inverted into discrete text prompts using a pretrained text embedding inversion model.

3.3 Text-Modal Attack on Diffusion Models via Token Embedding Inversion

T2I models typically rely on a pretrained text encoder, $\tau_\theta(\cdot)$, to transform the text prompt $\mathbf{p}$ into a text embedding, $\mathbf{e} = \tau(\mathbf{p}) \in \mathbb{R}^d$, where d is the dimension of the text embedding. We can rewrite $\mathbf{p}$ in its tokenized form $\mathbf{p} = [p_1, p_2, ..., p_L] \in \mathbb{N}^L$, where $p_i \in \{0, 1, ..., |V| - 1\}$ is the i^{th} token's index, V is the vanilla vocabulary codebook, $|V|$ is the vocabulary size, and L is the prompt length in the form of tokens. Each token is first mapped to its corresponding token embedding vector p through a fixed-size token-to-embedding lookup table. These embedding vectors, rewritten as $\mathbf{p}$ in their token embedding form, are then combined with positional embeddings to form the input to the transformer layers, which subsequently produce the final encoded sentence representation.

The final text representation typically exists in two forms: (1) The pooler output, with shape (batch_size, hidden_size), serves as an aggregated sentence-level representation; and (2) The last hidden state, with shape (batch_size, sequence_length, hidden_size), comprises the contextualized representations of all input tokens from the model's final layer.

Our TEOI framework operates in two distinct modes: the first mode performs optimization and inversion based on the token embedding vectors (discussed in this section), while the second mode operates on the pooler output (to be presented in the following section Sec. 3.4).

In image generation task, a T2I model equipped with text encoder $\tau(\cdot)$ takes text prompt $\mathbf{p}$ as inputs and generates a synthesized image $\mathbf{x}_{syn} = T2I(\tau(\mathbf{p}))$. Given a target image $\mathbf{x}_{tgt}$, the goal of the text-modal attack on T2I generation is to find a text prompt $\mathbf{p}_{attk}$ within the filtered vocabulary V_F, which contains no bad words in the prompt filter F, that would command the T2I model to generate an image similar to $\mathbf{x}_{tgt}$ but misaligned with $\mathbf{p}_{attk}$. To achieve this, we would conduct a two-stage training: first, optimize token embedding $\mathbf{e}_{token}$, then invert it to text prompt $\mathbf{p}_{attk}$ through projected gradient descent.

3.3.1 Text Embedding Optimization

As the text embedding $\mathbf{e}$ is of continuous nature and responsible for determining the synthesized image $\mathbf{x}_{syn}$, we, inspired by the previous related works, will try to optimize the text embedding as a trainable variable $\mathbf{e}_{opt}$ by utilizing its gradient information for an effective misalignment attack.

This section investigates against diffusion-based T2I models. Since diffusion models typically condition on the last hidden state of text inputs, which are challenging to invert directly back to discrete tokens, we instead optimize the token embeddings. Due to the stochastic nature of the generation process in diffusion models, Gao et al. (2023) proposes the adoption of a distribution-based loss function, which is expressed as follows:

$$\mathcal{L} = \max D(p_\theta(x|c') \| p_\theta(x|c)). \quad (1)$$

For T2I diffusion models, the time-conditioned U-Net (Ronneberger et al., 2015) is trained to predict noise during the diffusion process. As demonstrated in (Rombach et al., 2022), the conditional distribution in the text-to-image process can be implemented by a denoising autoencoder $\epsilon_\theta(x_t, t, c)$. So we derive the TEOI training objective by modifying the conditional diffusion model's original optimization target to

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_{tgt}, \epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}^t_{tgt}, t, \tau(\mathbf{e}))\|_2^2], \quad (2)$$

where t denotes a timestep in the diffusion process, ϵ represents the actual noise, ϵ_θ corresponds to the time-conditioned U-Net's noise prediction, and x^t_{tgt} denotes the noised representation of the target image x_{tgt} at timestep t.

Empirical results reveal that directly optimizing and projecting the initial prompt's token embeddings $\mathbf{e}$ without constraints yields a prompt

$\mathbf{p}_{attk}$ that suffers from poor interpretability, consequently undermining the effectiveness of misalignment attacks. To address this, we propose a hybrid optimization strategy: for initial prompts of fixed length L, we freeze a subset of token embeddings $\mathbf{e}_{frz}$ while treating the remaining portion as trainable parameters $\mathbf{e}_{opt}$. This approach simultaneously preserves the attacker's ability to incorporate predefined semantic cues from the initial prompts while ensuring the final text maintains human-readable quality. So we get the final objective

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_{tgt},\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}_{tgt}^t, t, \tau(\mathbf{e}_{opt}, \mathbf{e}_{frz}))\|_2^2]. \tag{3}$$

3.3.2 Text Embedding Inversion

The second stage is to invert $\mathbf{e}_{opt}$ to text prompt $\mathbf{p}_{opt}$ which would form $\mathbf{p}_{attk}$ together with $\mathbf{p}_{frz}$. For diffusion models, inspired by PH2P (Mahajan et al., 2024) and PEZ (Wen et al., 2024), we employ delayed projected gradient descent (PGD) to map the token embeddings $\mathbf{e}_{opt}$ to their respective nearest neighbor tokens within V_F. The discretized $\mathbf{e}_{opt}$ are then either fed into the next optimization iteration or, upon meeting termination criteria, inverted into corresponding tokens to form $\mathbf{p}_{opt}$, which is subsequently concatenated with $\mathbf{p}_{frz}$ to yield the final adversarial prompt.

This inversion process is based on the key experimental finding that projecting continuous token embeddings to their nearest discrete neighbors in the vocabulary space induces minimal perturbation to both the resulting text representations in the form of last hidden state and the final generated image.

The total algorithm is shown in Algorithm 1.

Algorithm 1 Text-Modal Attack using TEOI

Require: **T2I** model, text encoder $\tau(\cdot)$, fixed prompt parts $\mathbf{p}_{frz}$ and its token embedding $\mathbf{e}_{frz}$, target image $\mathbf{x}_{tgt}$, step size α

initialize $\mathbf{e}_{opt}$
for $i = 1, 2, \cdots, n$ **do**
  compute loss $\mathcal{L}$ with random t// Eq. (3)
  gradient descent: $\mathbf{e}_{opt} \leftarrow \mathbf{e}_{opt} - \alpha \cdot \nabla_{\mathbf{e}_{opt}} \mathcal{L}$
  project to filtered neighbor: $\mathbf{e}_{opt} \leftarrow Proj(\mathbf{e}_{opt})$
end for
invert to text $\mathbf{p}_{attk} = E2T(\mathbf{e}_{opt}) + \mathbf{p}_{frz}$

Ensure: misaligned prompt $\mathbf{p}_{attk}$

3.4 Multimodal Attack on GAN via Pooler Output Inversion

In image editing tasks, T2I models equipped with both text and image encoders take as input a textual editing command and a source image to produce the edited output, as shown in Fig. 2. Given a specific source image and a target image, our multimodal misalignment attack aims to discover (1) an editing command that bypasses the prompt filter and (2) a perturbed source image such that the generated output exhibits significant semantic misalignment with the original input while maintaining visual fidelity to the target.

Our approach concurrently attacks both the textual prompt and input image through coordinated optimization of both modalities, as shown in Fig. 2. For the image domain, we apply ℓ_∞ norm constrained perturbations within a predefined budget, optimizing the adversarial noise through iterative gradient updates. Since our analysis in this subsection focuses on GAN-based image editing models that typically employ pooler outputs as their text representation, we directly optimize pooler outputs of these sentence embeddings in the text modality. The optimized embeddings are subsequently decoded into discrete textual commands using a pretrained text embedding inversion model.

The multimodal attack is also executed using a two-stage approach. We optimize the text embedding and the input image to find a filtered prompt and adversarially perturb the input image within the defined budget to further amplify the misalignment. This dual optimization strategy allows for a more comprehensive attack, leveraging both textual and visual modalities to disrupt the alignment mechanisms of the T2I model.

3.4.1 Joint Optimization of Text Embedding and Image

Since both textual prompts and input images deterministically govern the output generation when model parameters are fixed, and considering that text embeddings and image pixels share continuous properties, we formulate a joint optimization framework across both modalities. When performing this coordinated optimization, several critical factors must be systematically incorporated to construct an effective unified objective function:

Cross-modal misalignment loss. First of all, we should consider the goal of the misalignment attack. In our case, the T2I model uses directly the text encoder $\tau(\cdot)$ and image encoder $\iota(\cdot)$ of CLIP (Rad-

ford et al., 2021) for its multi-modal semantic understanding capability. Although the input image is now treated as a variable, the perturbation applied under the norm constraint is imperceptible to human observers, resulting in virtually no visible changes. Therefore, the primary objective of the cross-modal misalignment loss remains focused on maximizing the semantic discrepancy between the textual prompt and the target image. So, by computing the cosine similarity between text embedding $\mathbf{e}$ and image embedding $\mathbf{e}_{tgt} = \iota(\mathbf{x}_{tgt})$, we get the cross-modal misalignment loss

$$\mathcal{L}_{mis} = 1 - \frac{\mathbf{e}_{opt} \cdot \mathbf{e}_{tgt}}{\|\mathbf{e}_{opt}\| \, \|\mathbf{e}_{tgt}\|}, \tag{4}$$

Image similarity loss. On the other hand, to ensure $\mathbf{x}_{syn} \approx \mathbf{x}_{tgt}$, we calculate the MSE between $\mathbf{x}_{syn}$ and $\mathbf{x}_{tgt}$ and get the image similarity loss

$$\mathcal{L}_{sim} = MSE(T2I(\mathbf{e}_{opt}, \iota(\mathbf{x}_{in})), \mathbf{x}_{tgt}). \tag{5}$$

When the input image belongs to a specific category, such as human faces, we can enhance the training process by incorporating face similarity metrics (Deng et al., 2019) in addition to MSE loss. This ensures better training outcomes by leveraging domain-specific knowledge to guide the optimization.

Embedding distribution constraint loss. Compared with typical text adversarial attacks that aim to modify the text input on the character level, word level, or sentence level, the gradient-based text embedding optimization gives much more freedom in searching for the feasible text prompt. However, text embedding optimization without appropriate constraints would lead to failure in inverting the optimized $\mathbf{e}_{opt}$ to the text prompt $\mathbf{p}_{attk}$, because the embedding space is just a small subset of the optimization space.

As previously discussed, optimizing the token embeddings requires projection onto the vocabulary embedding space to ensure proper inversion into discrete tokens. We now examine the necessary constraints for optimized pooler outputs to maintain invertibility to coherent discrete text.

Compared with one-hot or other sparse encodings, CLIP text embeddings in the form of pooler output are dense vectors, being ℓ_2 normalized so cosine similarity could be used to measure their similarity, thus all the text embeddings lie on the surface of a unit hypersphere. Wang and Isola (Wang and Isola, 2020) prove that the contrastive loss of CLIP drives the text embeddings to follow a uniform distribution on the unit hypersphere. However, through experiments we found a curious phenomenon that real text embeddings encoded from prompt texts tend to be more 'sparse' than a randomly sampled vector on the hypersphere, which means most values in the vector still tend to be near zero compared with a random sample. We partially attribute this finding to the *linear representation hypothesis* (Mikolov et al., 2013; Park et al., 2023; Arora et al., 2018, 2016; Faruqui et al., 2015; Merullo et al., 2022; Seth et al., 2023; Bhalla et al., 2024) which suggests that many semantic concepts are approximately linear functions of sparse representations for both language modeling and multimodal models. As the CLIP text encoder $\tau(\cdot)$ has an input text length limit and the text prompts we use are usually short sentences, all that makes the real text embeddings tend to be sparse.

As $\mathbf{e}_{opt}$ are ℓ_2 normalized, the ℓ_1 norm $\|\mathbf{e}_{opt}\|_1$ of $\mathbf{e}_{opt}$ could reflect its sparsity. So, in order to narrow the search space and to drive $\mathbf{e}_{opt}$ toward feasible solutions, we introduce text embedding distribution constraint loss

$$\mathcal{L}_{dist} = \left| \|\mathbf{e}_{opt}\|_1 - \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{e}_i\|_1 \right|, \tag{6}$$

where $\mathbf{e}_i$ is the text embedding for one of our total N sampled text prompts from the corpus, which helps to constrain the sparsity of $\mathbf{e}_{opt}$.

At last, the final overall objective is to minimize:

$$\mathcal{L}_{all} = \mathcal{L}_{mis} + \mathcal{L}_{sim} + \mathcal{L}_{dist} \tag{7}$$

to get optimized $\mathbf{e}_{opt}$ and adversarial $\mathbf{x}_{adv}$ which could generate $\mathbf{x}_{syn}$ that is similar to $\mathbf{x}_{tgt}$. Eq. (7) is a mathematical formulation of a multi-objective nonconvex optimization problem and Zhou et al. (2024) discuss the convergence of Adaptive Gradient Methods, such as RMSProp (Hinton, 2012), in the context of nonconvex optimization.

3.4.2 Text Embedding Inversion

The second stage is to invert $\mathbf{e}_{opt}$ in the form of pooler output to text prompt

$$\mathbf{p}_{attk} = E2T(\mathbf{e}_{opt}). \tag{8}$$

We use the off-the-shelf Vec2Text (Morris et al., 2023) architecture to train on CLIP (Radford et al., 2021) text encoder on large corpus datasets to get the text embedding inversion model $E2T$. Once

Algorithm 2 multi-modal attack using TEOI

Require: **T2I** model, text encoder $\tau(\cdot)$, image encoder $\iota(\cdot)$, input image $\mathbf{x}_{in}$, target image $\mathbf{x}_{tgt}$, inversion model **E2T**, filter words F, similarity threshold T, perturbation budget ε, step size α,

initialize $\mathbf{e}_{opt} \leftarrow \tau('')$, $\mathbf{x}_{adv} \leftarrow \mathbf{x}_{in}$
for $i = 1, 2, \cdots, m$ **do**
  for $i = 1, 2, \cdots, n$ **do**
    compute total loss $\mathcal{L}_{total}$ {see Eq. (7)}
    optimize: $\mathbf{e}_{opt} \leftarrow \mathbf{e}_{opt} - \alpha \cdot \nabla_{\mathbf{e}_{opt}} \mathcal{L}_{total}$
    update: $\mathbf{x}_{adv} \leftarrow \mathbf{x}_{adv} - \alpha \cdot sign(\nabla_{\mathbf{x}_{adv}} \mathcal{L}_{total})$
    normalize: $\mathbf{x}_{adv} \leftarrow \text{clamp}(\mathbf{x}_{adv}, \mathbf{x}_{in} - \varepsilon, \mathbf{x}_{in} + \varepsilon)$
  end for
  invert, substitute bad words $\mathbf{p}_{attk} = E2T(\mathbf{e}_{opt})$
  generate image $\mathbf{x}_{syn} \leftarrow T2I(\tau(\mathbf{p}_{attk}), \iota(\mathbf{x}_{in}))$
  if $\cos(\iota(\mathbf{x}_{syn}, \iota(\mathbf{x}_{tgt}))) > \mathrm{T}$ **then**
    break
  end if
end for

Ensure: misaligned prompt $\mathbf{p}_{attk}$, perturbed input image $\mathbf{x}_{adv}$

pretrained ahead, $E2T$ could be used for inference afterward without any further fine-tuning.

Then we replace any bad words in $\mathbf{p}_{attk}$ with semantically similar 'good' words to circumvent the prompt filter. Furthermore, although the performance of Vec2Text is impressive, there is still a similarity gap between $\mathbf{e}_{opt}$ and $\tau(E2T(\mathbf{e}_{opt}))$, as well as the distributional difference between $\mathbf{e}_{opt}$ and the real embeddings of texts, which leads to the occasional failure to reconstruct $\mathbf{x}_{tgt}$ as discussed in Appendix A.7. To increase the attack success rate, we would use TEOI in a retrying loop with finite times. The total algorithm is shown in Algorithm 2.

4 Experiments

4.1 Experimental Settings

T2I model with prompt filter. For our experimental validation, we select separate representative T2I models corresponding to each attack scenario described above. To evaluate text-modal misalignment attacks, we employ Stable Diffusion v1.5, v2.1 and XL (Rombach et al., 2022) which utilize the last hidden state of text embeddings as conditional input. For multimodal misalignment attacks, we choose StyleCLIP (Patashnik et al., 2021) as our representative GAN-based T2I model, notable for its use of pooler output as conditional input.

We implement prompt filtering at the token level by employing four offensive word lists: LDNOOBW[1], LDNOOBWV2[2], CMU Bad Words[3] and Google Profanity List[4], with each model instance being equipped with one such filter. All textual content involved in this work is in English.

Datasets. For experimental evaluation, we employ distinct datasets tailored to each model's characteristics: we utilize the COCO (Lin et al., 2014) and Not-Safe-For-Work (NSFW) images in MMA (Yang et al., 2024b) for diffusion models across diverse scenes and objects, while adopting the FFHQ (Karras et al., 2019) dataset for StyleCLIP on image editing. From each dataset, we systematically select representative images to construct our test set, ensuring comprehensive evaluation coverage while maintaining experimental efficiency.

For the COCO dataset, we employ crafted prompt templates that exhibit intentional semantic conflicts with target images along specific attributes, whereas for StyleCLIP, we utilize benign editing commands to generate target images.

Text embedding inversion. We train the Vec2Text (Morris et al., 2023) model to invert the text encoder of StyleCLIP. At first, we generate pooler outputs on the large-scale corpus MS MARCO (Nguyen et al., 2016), then train a zero-step model on these text embeddings for 100 epochs, followed by training a corrector model for another 100 epochs based on the previously trained model. This results in the final inverter. The entire process is computationally expensive, requiring approximately two weeks on an NVIDIA Tesla A800 80GB GPU. We discuss the time-consuming issue in detail in Appendix A.6.

Baselines. To the best of our knowledge, there is currently no specific method for conducting misalignment attacks on T2I models. Therefore, we adopted MMA-Diffusion(Yang et al., 2024b), a new and highly performing approach to multimodal adversarial attacks, as the baseline for our experiments. To ensure consistency in objectives, we made two modifications: (1) In the text-modal attack, BLIP-2 (Li et al., 2023b) is employed to generate descriptive captions for each target image in TEOI, which subsequently serve as the target prompts for MMA. (2) In the image-modal attack, the objective function becomes the negative of the

[1] https://github.com/LDNOOBW/
[2] https://github.com/LDNOOBWV2/
[3] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt
[4] https://github.com/coffee-and-fun/google-profanity-words

Dataset		COCO								NSFW							
Model	Filter	Metric		Metric		Metric		Metric		Metric		Metric		Metric		Metric	
	Method	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1
SD v1.5	MMA	68%	48%	59%	38%	69%	49%	66%	46%	59%	36%	45%	25%	58%	35%	57%	33%
	TEOI	**87%**	**67%**	**72%**	**54%**	**85%**	**65%**	**81%**	**61%**	76%	56%	59%	39%	76%	55%	74%	52%
SD v2.1	MMA	62%	43%	53%	33%	65%	43%	61%	41%	50%	33%	42%	21%	49%	33%	47%	31%
	TEOI	**81%**	**62%**	**67%**	**49%**	**79%**	**60%**	**77%**	**55%**	70%	49%	55%	34%	69%	48%	66%	36%
SDXL	MMA	54%	43%	53%	32%	64%	44%	60%	41%	51%	31%	39%	19%	49%	30%	48%	27%
	TEOI	**80%**	**61%**	**65%**	**48%**	**78%**	**59%**	**74%**	**53%**	68%	50%	51%	33%	68%	48%	65%	49%

Table 1: Comparison of TEOI and MMA with **text-modal** attack on image-generation task. Bold values indicate the best performance.

cosine similarity between the generated image and the target image.

During the text-modal MMA, the input images are kept constant. In multimodal attack, we first conduct the text-modal attack to obtain the adversarial prompt, then freeze this adversarial prompt and perform the image perturbation.

Evaluation metrics. The Attack Success Rate out of N syntheses (ASR-N) is adopted as the evaluation metric. The attack is considered successful once the adversarial text, containing no bad word in the prompt filter, enables the T2I model to generate at least one image, out of N, whose similarity to the target image exceeds the predefined similarity threshold.

4.2 Text-Modal Attack

Attack the prompt. The visualization of the text-modal attack results can be observed in Appendix A.2. From comprehensive quantitative results in Tab. 1 we can see that our TEOI framework reaches an average ASR-4 of 77.2% and ASR-1 of 57.9% for COCO and an average ASR-4 of 66.4% and ASR-1 of 45.4% for NSFW-MMA on different models and prompt filters. This reveals the inherent vulnerability of T2I models in maintaining text-image alignment, particularly when subjected to misalignment attacks, despite their training on aligned text-image data.

We posit that the attack efficacy of TEOI stems from the inherent limitations of text embeddings in semantic representation, which creates a discrepancy between their encoded meaning and human-perceived semantics.

Comparison with baselines. As shown in Tab. 1, TEOI outperforms the baseline method by a great margin. This can be attributed to TEOI's end-to-end generation of misaligned image-text pairs. The frozen components in text prompts are primarily responsible for semantic misalignment between text prompts and generated images, while the optimizable components of the prompts enable the generated images to closely resemble the target images. In contrast, baseline methods primarily focus on the similarity between the adversarial and target texts, neglecting the alignment issue.

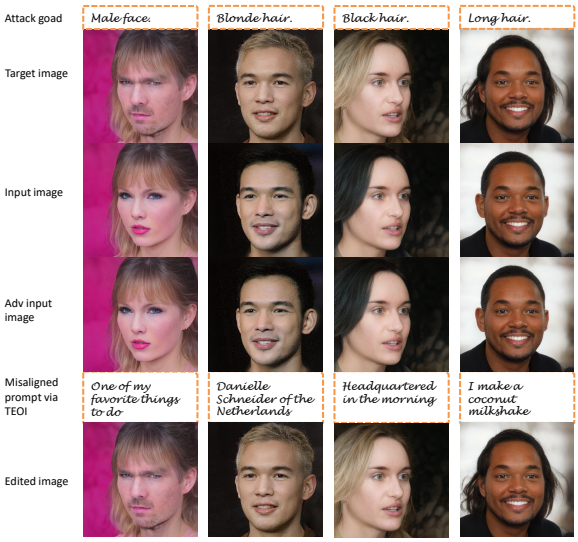

Figure 3: **Visualization results of multimodal attacks on image editing**. TEOI enables target image synthesis using misaligned adversarial text, applicable for both concealed editing purposes and training-phase data poisoning.

4.3 Multimodal Attack

Attack the prompt and image. In the context of multimodal attacks, as shown in Fig. 3 and Tab. 2, multimodal TEOI achieves a higher average success rate than unimodal TEOI in Tab. 1. This improvement is primarily attributed to the inclusion of pixel-level perturbations that remain imperceptible to human observers yet effectively deceive model judgments.

Comparison with baselines. Furthermore, TEOI framework exceeds the baseline methods in terms of the multimodal attack success rate. We attribute this to two key reasons. First, TEOI already outperforms MMA in text-modal attacks, while their image-modal attack implementations are basically the same. More importantly, TEOI effectively lever-

	Filter	LDNOOBW		LDNOOBWV2		CMU Bad Words		Google Profanity List	
Model	**Method**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**
StyleCLIP	**MMA**	70%	51%	60%	39%	71%	51%	68%	47%
	TEOI	**89%**	**68%**	**73%**	**56%**	**87%**	**66%**	**83%**	**63%**

Table 2: Comparison of TEOI and MMA with **multimodal** attack on image-editing task. Bold values indicate the best performance.

ages both text embeddings and pixel gradients for optimization. which allows to search the multimodal space more efficiently.

5 Conclusion

We introduce a novel perspective on the vulnerability of T2I models to mismatches between prompts and generated images, and propose an effective attack method to exploit this vulnerability. Leveraging the widespread use of text embeddings in diffusion-based or GAN-based T2I models, we present TEOI, a highly efficient method for adversarial prompt exploration and discovery through gradient-based optimization in continuous text embedding spaces. TEOI could also conduct multimodal attack, demonstrating effective circumvention of prompt filters and jailbreaking predefined prompt templates.

Finally, we innovatively extend TEOI to the multimodal domain, proposing the first gradient-based approach to simultaneously optimize image and text modalities. Experiments demonstrate that this multimodal approach outperforms single-modality attacks and its multimodal baseline with effective circumvention of prompt filters and jailbreaking predefined prompt templates.

Limitations

While our approach targets white-box T2I models accepting text embeddings (covering diffusion and GAN variants), extensions to non-embedding conditioning or black-box transfer attacks are left as future work.

Ethical Considerations

Note that T2I alignment may be more fragile than is assumed. This study focuses on evaluating the vulnerability of T2I models with respect to image–text alignment. Our goal is to encourage the development of defense mechanisms for T2I models against attacks such as TEOI. Here we briefly outlined three potential mitigation directions:

- Prompt filtering based on semantic similarity. This method goes beyond simple keyword matching by measuring the semantic similarity (e.g., cosine similarity) between the input prompt and a list of sensitive phrases. Prompts that do not explicitly contain sensitive keywords—but convey inappropriate content—can be effectively detected if they exceed a predefined similarity threshold.
- Filtering based on syntactic or linguistic anomalies. This defense leverages metrics such as perplexity (Bengio et al., 2000; Vaswani et al., 2017) to assess the fluency and grammaticality of prompts, as TEOI-generated adversarial prompts (especially those produced via discrete token optimization) sometimes contain syntactic irregularities or low readability.
- Safety alignment for improving model robustness. Interestingly, TEOI can also be used as a defense mechanism. By using TEOI to generate image-text pairs that are perceptually detectable by humans, we can curate hard training examples. Aligning these adversarial texts with corresponding ground-truth images and fine-tuning the model on such data can improve its robustness against misalignment-based attacks. In the supplementary experiments provided in the appendix, we first confirm the effectiveness of our proposed method.

We argue that advanced prompt filters capable of rejecting semantically inconsistent or grammatically flawed text inputs could help mitigate TEOI-based attacks. Complementarily, TEOI itself can be leveraged to generate adversarial samples for safety-aligned fine-tuning of pretrained T2I models.

Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC) under grant No. 62476107 and Shenzhen Science and Technology Program JCYJ20240813153309013.

References

Aishwarya Agarwal, Srikrishna Karanam, and Balaji Vasan Srinivasan. 2024. Alignit: Enhancing prompt alignment in customization of text-to-image models. *Preprint*, arXiv:2406.18893.

Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. 2024. Palp: Prompt aligned personalization of text-to-image models. *arXiv preprint arXiv:2401.06105*.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. 2024. Interpreting clip with sparse linear concept embeddings (splice). In *Advances in Neural Information Processing Systems*, volume 37, pages 84298–84328. Curran Associates, Inc.

Fuyao Cai, Daizong Liu, Xiang Fang, Jixiang Yu, Keke Tang, and Pan Zhou. 2025a. Imperceptible beam-sensitive adversarial attacks for lidar-based object detection in autonomous driving. In *IEEE International Conference on Multimedia & Expo 2025 (ICME 2025)*.

Xiaowen Cai, Daizong Liu, Runwei Guan, and Pan Zhou. 2025b. Imperceptible transfer attack on large vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiaowen Cai, Yunbo Tao, Daizong Liu, Pan Zhou, Xiaoye Qu, Jianfeng Dong, Keke Tang, and Lichao Sun. 2024. Frequency-aware gan for imperceptible transfer attack on 3d point clouds. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6162–6171.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg. Springer-Verlag.

Schuhmann Christoph, Köpf Andreas, Coombes Theo, Vencu Richard, Trom Benjamin, and Romain Beaumont. 2024. Laion-coco. `https://laion.ai/blog/laion-coco/`.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and 1 others. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902.

Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. 2024a. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3):1–38.

Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. 2024b. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pages 92–111. Springer.

Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. 2024c. Robust distillation via untargeted and targeted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28432–28442.

Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, and Yew-Soon Ong. 2024d. Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28535–28544.

Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. 2025a. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 236–247.

Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. 2025b. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*.

Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. 2023a. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24678–24687.

Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. 2022. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034.

Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. 2023b. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608.

Xiang Fang, Arvind Easwaran, and Blaise Genest. 2024a. Uncertainty-guided appearance-motion association network for out-of-distribution action detection. In *IEEE International Conference on Multimedia Information Processing and Retrieval*.

Xiang Fang, Arvind Easwaran, and Blaise Genest. 2025a. Adaptive multi-prompt contrastive network for few-shot out-of-distribution detection. In *International Conference on Machine Learning*.

Xiang Fang, Arvind Easwaran, Blaise Genest, and Ponnuthurai Nagaratnam Suganthan. 2025b. Adaptive hierarchical graph cut for multi-granularity out-of-distribution detection. *IEEE Transactions on Artificial Intelligence*.

Xiang Fang, Arvind Easwaran, Blaise Genest, and Ponnuthurai Nagaratnam Suganthan. 2025c. Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *Expert Systems with Applications*.

Xiang Fang, Wanlong Fang, Wei Ji, and Tat-Seng Chua. 2025d. Turing patterns for multimedia: Reaction-diffusion multi-modal fusion for language-guided video moment retrieval. In *ACM International Conference on Multimedia*.

Xiang Fang, Wanlong Fang, Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Renfu Li, Zichuan Xu, Lixing Chen, Panpan Zheng, and 1 others. 2024b. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 28–37.

Xiang Fang, Wanlong Fang, Changshuo Wang, Daizong Liu, Keke Tang, Jianfeng Dong, Pan Zhou, and Beibei Li. 2025e. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiang Fang and Yuchong Hu. 2020. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Wu. 2021a. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *IEEE Transactions on Artificial Intelligence*, 3(2):192–206.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2020. V3h: View variation and view heredity for incomplete multiview clustering. *IEEE Transactions on Artificial Intelligence*, 1(3):233–247.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):913–927.

Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. 2023a. Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8721–8733.

Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Zichuan Xu, Wenzheng Xu, Junyang Chen, and Renfu Li. 2024c. Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1735–1743.

Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*, 25:7517–7532.

Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. 2023b. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2448–2460.

Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. 2023c. Hierarchical local-global transformer for temporal sentence grounding. *IEEE Transactions on Multimedia*.

Xiang Fang, Zeyu Xiong, Wanlong Fang, Xiaoye Qu, Chen Chen, Jianfeng Dong, Keke Tang, Pan Zhou, Yu Cheng, and Daizong Liu. 2024d. Rethinking weakly-supervised video temporal grounding from a game perspective. In *European Conference on Computer Vision*. Springer.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. 2015. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*.

Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. 2023. Evaluating the robustness of text-to-image diffusion models against real-world attacks. *arXiv preprint arXiv:2306.13103*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Geoffrey Hinton. 2012. Neural Networks for Machine Learning, lecture 6e: rmsprop. Coursera.

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

Qianjiang Hu, Daizong Liu, and Wei Hu. 2022. Exploring the devil in graph spectral domain for 3d point cloud attacks. In *European Conference on Computer Vision*, pages 229–248. Springer.

Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. 2024. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4193–4205, Bangkok, Thailand. Association for Computational Linguistics.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3294–3302, Cambridge, MA, USA. MIT Press.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023a. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755.

Daizong Liu and Wei Hu. 2022. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4727–4746.

Daizong Liu and Wei Hu. 2024. Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3576–3584.

Daizong Liu and Wei Hu. 2025a. Imperceptible backdoor attacks on text-guided 3d scene grounding. *IEEE Transactions on Multimedia*.

Daizong Liu and Wei Hu. 2025b. Seeing is not believing: Adversarial natural object optimization for hard-label 3d scene attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11886–11897.

Daizong Liu, Wei Hu, and Xin Li. 2023a. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Daizong Liu, Wei Hu, and Xin Li. 2023b. Robust geometry-dependent attack for 3d point clouds. *IEEE Transactions on Multimedia*, 26:2866–2877.

Daizong Liu, Yang Liu, Wencan Huang, and Wei Hu. 2024a. A survey on text-guided 3d visual grounding: Elements, recent advances, and future directions. *arXiv preprint arXiv:2406.05785*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024b. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. 2024c. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. 2024d. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 102–111, Los Alamitos, CA, USA. IEEE Computer Society.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. 2024. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *CVPR 2024 (To appear)*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Khalil Mrini, Hanlin Lu, Linjie Yang, Weilin Huang, and Heng Wang. 2024. Fast prompt alignment for text-to-image generation. *arXiv preprint arXiv:2412.08639*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.

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

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

David F Shanno. 1970. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111).

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2024. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 660–674, New York, NY, USA. Association for Computing Machinery.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA. Association for Computing Machinery.

Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, and Wei Hu. 2023. 3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14340–14350.

Guiyao Tie, Xueyang Zhou, Tianhe Gu, Ruihang Zhang, Chaoran Hu, Sizhe Zhang, Mengqu Sun, Yan Zhang, Pan Zhou, and Lichao Sun. 2025. Mmlu-reason: Benchmarking multi-task multi-modal language understanding and reasoning. *arXiv preprint arXiv:2505.16459*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. P+: extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2022. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer.

Mingyu Yang, Daizong Liu, Keke Tang, Pan Zhou, Lixing Chen, and Junyang Chen. 2024a. Hiding imperceptible noise in curvature-aware patches for 3d point cloud attack. In *European Conference on Computer Vision*, pages 431–448. Springer.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024b. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. Vlattack: multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022.

Zhen Yu, Zhou Qin, Zhenhua Chen, Meihui Lian, Haojun Fu, Weigao Wen, Hui Xue, and Kun He. 2023. Sparse black-box multimodal attack for vision-language adversary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5775–5784, Singapore. Association for Computational Linguistics.

Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29927–29936.

Chenyu Zhang, Lanjun Wang, and Anan Liu. 2024a. Revealing vulnerabilities in stable diffusion via targeted attacks. *arXiv preprint arXiv:2401.08725*.

Jiaming Zhang, Qi Yi, and Jitao Sang. 2022. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Jianping Zhang, Zhuoer Xu, shiwen cui, Changhua Meng, Yizhan Huang, Weibin Wu, and Michael Lyu. 2024b. Multi-modality adversarial attacks on latent diffusion models.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. 2024. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*.

Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. 2025. Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. *arXiv preprint arXiv:2505.16640*.

Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, and 1 others. 2023. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*.

A Appendix

A.1 Implementation Details

For the text-modal attacks on diffusion models, we utilize an L-BFGS (Shanno, 1970) optimizer with a base learning rate of 0.8, while the multimodal attack adopts an RMSprop (Hinton, 2012) optimizer with a learning rate of 0.001. Regarding the ℓ_∞ bound of image attack, we set $\varepsilon = 16/255$. All experiments were performed using the PyTorch framework on an NVIDIA Tesla A800 80G GPU. -

A.2 More Visualizations

A supplementary visualization of text-modal attacks on image generation is presented in Fig. 4.

A.3 Evaluation on Safety Fine-tuned T2I Models

We conducted additional experiments on Safe SD v1.5 (Liu et al., 2024d), which incorporates safety fine-tuning, in comparison to vanilla SD v1.5. The results are presented in Tab. 3. The results show that our TEOI framework still outperforms the baseline in terms of attack effectiveness on safety fine-tuned T2I models. The safety-finetuned Safe SD v1.5 model demonstrates improved robustness, effectively reducing the attack success rate, confirming the benefits of safety fine-tuning. In fact, this fine-tuned model can serve as a mitigation mechanism against TEOI, as discussed in Sec. 5.

A.4 Evaluation on Black-box Models

We have supplemented our study with a transfer attack scenario using TEOI to assess the vulnerability of SaaS diffusion APIs (e.g., MidJourney[5], Leonardo.AI[6]). Furthermore, we note that Leonardo.AI is a customized variant of Stable Diffusion XL (SDXL). To align with this, our TEOI framework was trained on SDXL as the surrogate model.

For fair comparison with the baseline MMA method, we adapted the objective function in our black-box attack to match that of the baseline, ensuring consistency in both the evaluation dataset and metrics. Tab. 4 shows the summary of the comparative experimental results.

A.5 Additional NSFW Categories

Our initial focus on a limited set of NSFW themes (primarily adult content) was driven by practical considerations, as existing prompt filtering systems predominantly target adult or sexually explicit material.

In the supplementary experiments, we have extended our evaluation to include three additional NSFW categories: violent, hateful, and harassing. These were constructed using a combination of resources, including filtered subsets of the LAION dataset (Christoph et al., 2024) and Unsafe Diffusion dataset (Qu et al., 2023).

Tab. 5 presents the new experimental results on these extended NSFW categories:

A.6 Inversion Pre-training

We would like to clarify and contextualize the training cost of Vec2Text from the following perspectives:

- First, the reported training time in our paper is based on a single-GPU setup without specific optimizations. In more practical multi-GPU settings, such as using an 8-GPU cluster and enabling lower-precision formats like bfloat16, we have observed that training Vec2Text for a single text encoder can be completed within one day.
- Second, our proposed TEOI framework includes two variants: one based on text embedding inversion and another on token embedding inversion. Notably, only the text embedding variant requires separate pre-training of the inversion model. The token embedding version, in contrast, does not

[5] https://midjourney.com/
[6] https://leonardo.ai/

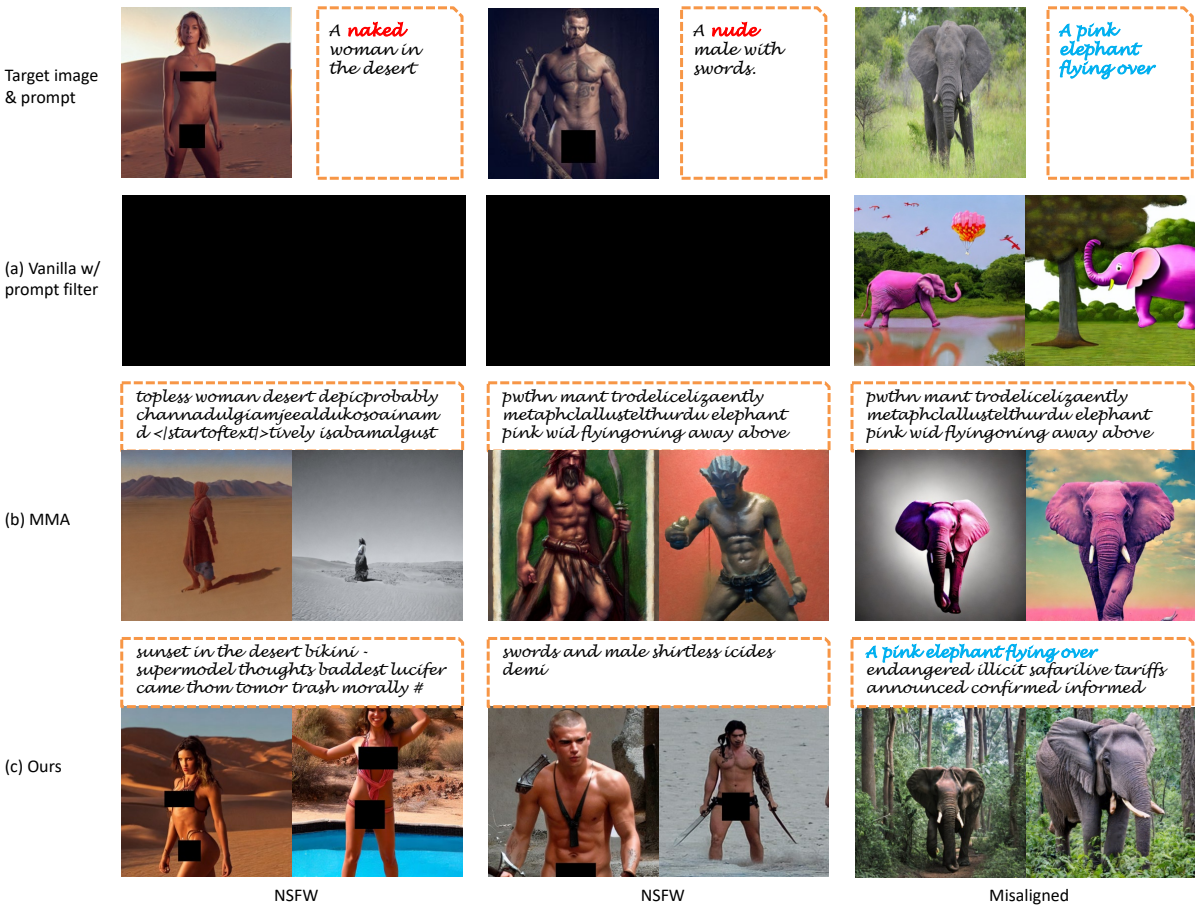

Figure 4: **Visualization results of text-modal attacks on image generation**. Bad words within the image caption are colored in red. Blue text denotes mandatory adversarial components. Black censor boxes are added by the authors for ethical considerations. (a) Vanilla outputs prompted by caption: black images indicate rejection by prompt filter. (b) Syntheses prompted by MMA. (c) Our syntheses can faithfully reflect the target image without mentioning sensitive words while containing the predefined sentence. Images are plotted with SD v1.5.

require additional training. It is broadly applicable across both GAN- and diffusion-based T2I models and works effectively for both image generation and image editing tasks. Therefore, in most practical attack scenarios, the token embedding variant can be used to avoid unnecessary training overhead.

- Lastly, for scenarios where attackers do prefer to use the text embedding variant of TEOI, we argue that the cost is often amortized or avoidable. First, a given text encoder (e.g., CLIP-ViT/L) is frequently reused across multiple major T2I models such as Stable Diffusion v1.5, v2.1, and XL. Thus, a single trained inversion model can be applied across a wide range of targets. Second, the number of widely used text encoders is relatively small. If TEOI or Vec2Text sees broader adoption, the community could easily share commonly used inversion models via platforms like HuggingFace, making pre-trained inversion models accessible and reducing the need for redundant training.

A.7 Text Reconstruction Fidelity

We focus on measuring the similarity between two embeddings in Sec. 3.4.2: the optimized text embedding obtained after each inner-loop iteration, as shown in Algorithm 2, and the text embedding derived from the Vec2Text-inverted text of the optimized embedding. We use cosine similarity between these two embeddings as a quantitative metric for reconstruction fidelity.

Based on our experimental results, the average cosine similarity is 0.7502. For reference, Vec2Text achieves a cosine similarity of 0.95 for out-of-domain reconstruction. While there is a gap between these values, it is important to note that Vec2Text's embedding is derived from real text, whereas ours is obtained through gradient-based

	Filter	LDNOOBW		LDNOOBWV2		CMU Bad Words		Google Profanity List	
Model	**Method**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**
SD v1.5	**MMA**	59%	36%	45%	25%	58%	35%	57%	33%
	TEOI	**76%**	**56%**	**59%**	**39%**	**76%**	**55%**	**74%**	**52%**
Safe SD v1.5	**MMA**	12%	8%	9%	6%	12%	7%	12%	7%
	TEOI	**16%**	**12%**	**13%**	**8%**	**16%**	**11%**	**15%**	**10%**

Table 3: Comparison of TEOI and MMA with **text-modal** attack on image-generation task under SD v1.5 and Safe SD v1.5 models. Bold values indicate the best performance.

		NSFW Theme	**Adult**	**Bloody**	**Horror**	**Racism**	**Politics**	**Notable**
		#adv. Prompt	50	30	90	30	50	50
Midj.	**MMA**	Bypass rate	22	55.33	70	63.33	66	100
		ASR-4 (%)	18	50	58.73	15.79	63.63	48.57
		Overall ASR-4	3.96	27.67	41.11	10	42	48.57
	TEOI	Bypass rate	24	63.33	74.44	70	68	100
		ASR-4 (%)	25	57.89	65.67	19.05	73.53	52
		Overall ASR-4	6	36.67	48.89	13.33	50	52
Leon.	**MMA**	Bypass rate	64	100	100	100	100	100
		ASR-4 (%)	59.38	86.67	85.56	73.33	88	58
		Overall ASR-4	38	86.67	85.56	73.33	88	58
	TEOI	Bypass rate	66	100	100	100	100	100
		ASR-4 (%)	66.67	90	88.89	76.67	92	62
		Overall ASR-4	44	90	88.89	76.67	92	62

Table 4: Comparison of TEOI and MMA with **text-modal** attack on image-generation task under Midjourney and Leonardo.Ai.

optimization. Furthermore, our empirical results demonstrate that the current reconstruction fidelity is sufficient to ensure strong attack success rates, as evidenced in Sec. 4.3. This analysis reinforces that our method maintains effective semantic consistency despite the inherent challenges in embedding inversion.

	Filter	LDNOOBW		LDNOOBWV2		CMU Bad Words		Google Profanity List	
Model	**Method**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**	**ASR-4**	**ASR-1**
Sexual	**MMA**	59%	36%	45%	25%	58%	35%	57%	33%
	TEOI	**76%**	**56%**	**59%**	**39%**	**76%**	**55%**	**74%**	**52%**
Violent	**MMA**	66%	44%	54%	37%	64%	42%	61%	43%
	TEOI	**78%**	**62%**	**65%**	**47%**	**79%**	**61%**	**74%**	**59%**
Hateful	**MMA**	67%	48%	56%	39%	63%	45%	62%	44%
	TEOI	**79%**	**64%**	**67%**	**50%**	**78%**	**63%**	**74%**	**58%**
Harassing	**MMA**	64%	44%	53%	36%	61%	44%	61%	39%
	TEOI	**77%**	**61%**	**64%**	**48%**	**77%**	**59%**	**73%**	**55%**

Table 5: Comparison of TEOI and MMA with **text-modal** attack on image-generation task under SD v1.5 for 4 NSFW categories. Bold values indicate the best performance.