UPAM: Unified Prompt Attack in Text-to-Image Generation Models Against Both Textual Filters and Visual Checkers

Duo Peng [1] **Qiuhong Ke** [2] **Jun Liu** [1 3]

Abstract

Text-to-Image (T2I) models have raised security concerns due to their potential to generate inappropriate or harmful images. In this paper, we propose UPAM, a novel framework that investigates the robustness of T2I models from the attack perspective. Unlike most existing attack methods that focus on deceiving textual defenses, UPAM aims to deceive both textual and visual defenses in T2I models. UPAM enables gradient-based optimization, offering greater effectiveness and efficiency than previous methods. Given that T2I models might not return results due to defense mechanisms, we introduce a Sphere-Probing Learning (SPL) scheme to support gradient optimization even when no results are returned. Additionally, we devise a Semantic-Enhancing Learning (SEL) scheme to finetune UPAM for generating target-aligned images. Our framework also ensures attack stealthiness. Extensive experiments demonstrate UPAM's effectiveness and efficiency.

1. Introduction

Text-to-Image (T2I) models have garnered widespread attention in the research community for their ability to create high-quality images from text prompts (Xu et al., 2018; Zhang et al., 2017; 2018). Particularly with the advent of diffusion generation techniques (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Peng et al., 2023a;b; Dhariwal & Nichol, 2021), the quality of images produced by T2I models has further improved. However, there is a growing concern that such technologies might be misused (Millière, 2022; Saharia et al., 2022; Yu et al., 2022). This is because text-to-image models become more integrated into various online services, greatly increasing the potential for generating inappropriate or harmful content, e.g., creating fake images of individuals (Chesney & Citron, 2019) or producing images that contain violent/obscene content (Birhane et al., 2021; Saharia et al., 2022; Yu et al., 2022). Such misuse of T2I models carries a risk of breaching legal standards and regulations, such as EU AI Policy (Galetin et al., 2022) and US Federal AI Governance (Daly et al., 2020). This could lead to legal and reputational consequences for both API developers and users.

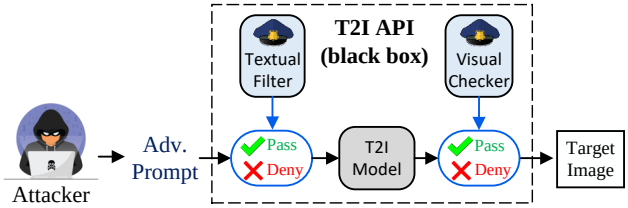

Figure 1. T2I APIs could incorporate both textual filters and visual checkers for double defenses, which can deny the forward propagation of data when harmful information is detected (Schramowski et al., 2023; Rombach et al., 2022). In this figure, the image is finally outputted from the black-box API, indicating that the data has passed through each defense.

Despite T2I APIs being equipped with defense mechanisms to prevent harmful generation, numerous studies (Liu et al., 2023; Daras & Dimakis, 2022; Millière, 2022) have indicated that T2I services still face potential risks of misuse when encountering adversarial attacks, highlighting the need for continuous enhancements in security protocols. In this context, we study the adversarial attack on T2I models, as by mimicking attacks, we can uncover and highlight the vulnerabilities inherent in T2I models, thereby informing feasible defense approaches for future development.

Public APIs typically employ textual filters at the input stage to prevent the passage of harmful text prompts (Mid, 2023; Leo, 2023). When users intend to generate harmful images, their naive (text) prompts often contain patently malicious content, which can be easily denied by textual filters, preventing their submissions into T2I models for further processing. Recently, several attack methods (Liu et al., 2023; Daras & Dimakis, 2022; Millière, 2022; Struppek et al., 2022; Jin et al., 2020) have emerged to investigate how to change the naive prompts into adversarial ones to deceive the textual filters. In the field of adversarial attack for T2I models, a common strategy is to generate adversarial (text) prompts to disable the defense mechanism implemented

[1]Singapore University of Technology and Design, Singapore [2]Monash University, Australia [3]Lancaster University, United Kingdom. Correspondence to: Jun Liu <ljchangyu@hotmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

by API and thus induce T2I models to generate harmful content (Liu et al., 2023). To find out a proper adversarial prompt, existing attack methods often craft new words to replace the original, malicious ones in the naive prompt. They continually change the crafted words, until the adversarial prompt guides the black-box T2I model to output an image that achieves enough semantic similarity with the intended harmful content. However, these enumeration-based methods tend to show limited efficiency, as they require attempting a vast range of candidates to find possible solutions even during inference. Moreover, a practical scenario is also overlooked in existing research: as shown in Fig. 1, besides the textual filter, the API provider often additionally deploys a visual checker at the output stage, forming a double-checking defense (Schramowski et al., 2023; Rombach et al., 2022). In this context, traditional enumeration methods become less effective, even struggling to work, as the introduction of visual checkers greatly reduces the probability of successfully finding a solution via enumeration.

Inspired by the powerful capability of gradient-based optimization in efficiently searching optimal solutions for intricate problems (Daoud et al., 2023; Zhou et al., 2021), we propose to train a parameterized attack model using gradient optimization. With the help of gradients, the attack model learns to self-adaptively navigate the solution space, enhancing the capability to find effective solutions in situations where traditional enumeration methods may falter. After gradient-based training, the trained model can swiftly transform the naive prompts into adversarial ones during inference, which brings much efficiency. Despite the conceptual simplicity, enabling the gradient-based optimization in this task however is challenging, as discussed below.

In practice, attackers generally cannot access the internal structure and parameters of the T2I models, as they are packed in the black-box API, which makes conventional gradient descent algorithms unsuitable (Stephan et al., 2017; Kingma & Ba, 2014). Although some existing black-box learning methods (Golovin et al., 2017) relax the need to access internal information of the black box, they generally require obtaining the output results of the black box to compute losses and estimate gradients. However, in the training process, the integrated textual and visual defenses of the API could effectively detect malicious content within the input prompts or generated images. Upon detecting malicious information, these defenses trigger a halt in the forward propagation of data (i.e., "Deny" in Fig. 1), causing black-box T2I models to return (output) no results. Our goal is to not only prompt the API to return images but also ensure these images align with the intended harmful target. However, in such a no-feedback scenario, it becomes particularly challenging to discover gradients for effective training to achieve these two goals simultaneously, due to very limited information accessible to attackers.

Another challenge is ensuring attack stealthiness. Previous attack methods often generate adversarial prompts by creating new words, e.g., using the adversarial prompt "Apoploe vesrreaitais" to represent the naive prompt "bugs" (Daras & Dimakis, 2022). However, such unnatural prompts significantly reduce the stealthiness (naturalness), making the attack easily detectable and defendable. Nevertheless, ensuring the naturalness of adversarial prompts is a non-trivial issue, since guaranteeing naturalness often demands a substantial volume of natural text data for data-driven learning (Raiaan et al., 2023). In this task, it is infeasible to prepare massive ground-truth adversarial prompts that are natural.

To overcome the challenges above, we propose a gradient-based Unified Prompt Attack Model (UPAM) to effectively and efficiently generate natural adversarial prompts that deceive both textual filters and visual checkers. Considering it is non-trivial to simultaneously prompt the API to return images and also ensure their harmfulness, here we decompose the goal into two learning stages to address the challenges separately: the first stage enables the API to reliably return images, and the second stage finetunes these images to be harmful. In both learning stages, we enable the gradient-based optimization. In the first stage, we introduce a *Sphere-Probing Learning (SPL)* scheme to train UPAM to deceive the textual and visual defenses. Inspired by (Kahla et al., 2022), SPL is designed to enable the gradient-based learning under the scenario where no image results are returned. As shown in Fig. 3 (a), SPL explores the optimization direction by querying the defense cases ("Pass" or "Deny") over a sphere. It can determine effective optimization gradients (directions) by analyzing the cases of multiple queries. SPL progressively optimizes UPAM using the estimated gradients, ultimately enabling the adversarial prompts generated by UPAM to effectively evade defenses, thus making the black box return (output) image results. Once the adversarial prompts can prompt black-box T2I models to return images, it becomes crucial to ensure that the returned images semantically align with our target images. To achieve this, in the second stage, we introduce a *Semantic-Enhancing Learning (SEL)* scheme, which is based on ZOO algorithm (Spall, 1992a; Oh et al., 2023), to further enhance the performance of UPAM by iteratively refining the generated images towards the target semantics. Moreover, inspired by the recent advancements of Large Language Models (LLMs) in synthesizing natural language (Huang et al., 2023), we propose to harness the power of LLM to generate natural adversarial prompts. Thanks to the LLM's language organization capability, our UPAM can reliably generate a human-readable and spelling-correct prompt, which greatly improves the attack stealthiness. Our proposed framework is evaluated on various widely-used T2I models, demonstrating the superiority of UPAM over existing attack techniques.

2. Related Work

Adversarial prompt attack for T2I models is a relatively under-explored topic. Daras et al. (Daras & Dimakis, 2022)

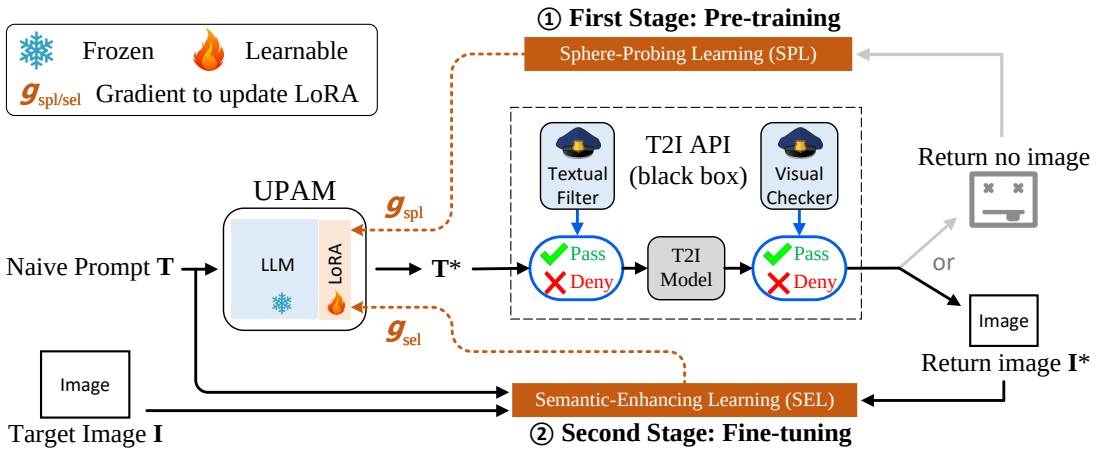

Figure 2. Overview of our UPAM framework. All the gradients are used to only optimize LoRA, while LLM is kept frozen.

proposed a pioneering work that investigates vulnerabilities of DALL·E 2 (Ramesh et al., 2022). They found that it is possible to deceive textual filters by using nonsensical text as prompts. Later, Millière et al. (Millière, 2022) proposed to create adversarial prompts by combining multi-language sub-word segments. Struppek et al. (Struppek et al., 2022) demonstrated that replacing individual characters in text with non-Latin characters can induce different visual contents in generated images. Liu et al. (Liu et al., 2023) proposed an attack method based on word modification (e.g., character replacement, deletion, or swapping). These methods mainly focus on deceiving the textual filter but often overlook the post-hoc visual checker implemented by the API provider. Differently, our paper proposes UPAM, a unified attack model designed to simultaneously deceive both the textual filter and the visual checker. Particularly, we enable the gradient-based optimization of UPAM, which achieves a more effective and efficient attack than existing methods. Furthermore, different from most existing methods, we follow the previous work (Liu et al., 2023) to also consider the attack stealthiness. To achieve better attack stealthiness, our UPAM harnesses a pre-trained LLM to ensure the naturalness of the generated adversarial prompts.

3. Method

In this task, given the paired data $\{\mathbf{T}, \mathbf{I}\}$, where $\mathbf{T}$ denotes the naive text prompts containing "malicious" information, and $\mathbf{I}$ denotes the corresponding target "harmful" images, the goal of the attacker is to construct an attack model H: $\mathbf{T} \rightarrow \mathbf{T}^*$, where H rewrites the naive text prompts $\mathbf{T}$ into the adversarial prompts $\mathbf{T}^*$. For a successful attack model H, the following key requirements must be met: (1) $\mathbf{T}^*$ should be textually dissimilar from $\mathbf{T}$ to avoid the "malicious" representation, allowing it to bypass the textual filter and be submitted to the T2I model for image generation. (2) The generated image $\mathbf{I}^*$ should bypass the visual checker, enabling it to be returned to the attacker. (3) Additionally, the generated $\mathbf{I}^*$ should be semantically similar to the target "harmful" image $\mathbf{I}$. Throughout the entire training and testing process, the attacker does not have access to any internal information of the T2I model, the textual filter, or the visual checker. These components are treated as a black-box system, as shown in Fig. 2. The attacker can only query the input and output of this system.

Previous methods (Liu et al., 2023; Daras & Dimakis, 2022) typically treat H as an enumeration model. These enumeration-based methods tend to incur considerable time costs, since they do not train any parameterized model for future inference. Thus for each inference sample, this time-consuming enumeration process needs to be conducted afresh, leading to limited efficiency. Besides, they primarily focus on deceiving the textual filters in the black box. Consequently, when faced with dual defenses involving both the textual filter and visual checker, enumeration-based methods may struggle to find solutions, resulting in limited effectiveness.

To address these issues, we propose to treat H as a parameterized model (i.e., our UPAM) and train it using gradient-based optimization. In the inference (testing) stage, our UPAM can immediately and effectively convert testing data (naive prompts) into adversarial ones, greatly enhancing the efficiency and effectiveness. Also, the naturalness of adversarial prompts is ensured in our UPAM.

3.1. Approach Overview

As shown in Fig. 2, our UPAM consists of a frozen off-the-shelf Large Language Model (LLM) and a learnable LoRA adapter (Hu et al., 2021). We only update the LoRA adapter, which contains a minimal number of parameters. This allows us to preserve the knowledge of the pre-trained LLM, thus ensuring the naturalness of the generated adversarial prompts. The training of our UPAM comprises two stages:

(1) Pre-training Stage. Initially, obtaining image results

from the API is challenging due to the pre-trained LLM's tendency to preserve naive textual cues from the input (Kalyan, 2024). Therefore, the adversarial prompts generated by the LLM-based UPAM initially retain malicious information from the naive prompt, making them prone to denial by defenses, resulting in no image results returned (i.e., no feedback). Inspired by (Kahla et al., 2022), we introduce a Sphere-Probing Learning (SPL) scheme to deceive the textual and visual defenses in the no-feedback scenario. Even without image results, SPL can still search for effective gradients to train UPAM, enabling it to reliably compel the black box to return images.

(2) Fine-tuning Stage. After the Pre-training Stage, the API can return images to the attacker. Based on the returned images, we propose a Semantic-Enhancing Learning (SEL) scheme based on ZOO algorithm (Spall, 1992a; Oh et al., 2023) to progressively align the semantics of returned images towards target ones. Particularly, SEL is designed to work compatible with SPL.

3.2. Pre-training Stage with SPL

As mentioned in Sec. 3.1, in the beginning, the absence of image results from the API presents a challenge in gradient acquisition. This is because gradients are typically derived by evaluating the magnitude of loss variations (Chien, 2018). When the API fails to return results, even if the loss can be defined (e.g., considering it as a binary classification loss between "return no images" and "return images"), the no-feedback state of the API will cause the loss value to remain unchanged. Such inability to discern loss changes (variations) prevents the acquisition of gradients. Therefore, the conventional approach of obtaining gradients through calculating loss functions becomes impractical, emphasizing the need to explore an innovative gradient acquisition method to address this no-feedback scenario.

Inspired by (Kahla et al., 2022), we propose a Sphere-Probing Learning (SPL) scheme, which enables the gradient acquisition in the scenario where no results are returned. Considering that the black box's output case (state) is binary – either "Deny" (image not returned) or "Pass" (image returned) – an intuition behind our SPL scheme arises: *if we can identify the decision boundary of the black box between these two cases, the gradients can be obtained effectively*. As shown in Fig. 3 (a), the boundary can indicate an effective direction for gradient optimization, as crossing this boundary could effectively shift the black-box output from "Deny" to "Pass", making the black box return images. Inspired by this, SPL is designed to gradually approach the "Pass" case by iteratively probing the decision boundary. While there are methods in the task of model inversion (Choquette-Choo et al., 2021; Zhu et al., 2022; Kahla et al., 2022) exploring a similar concept, they cannot be directly applied to this task. These methods typically optimize towards the center of the target case, aiming to maximize the black box's decision confidence on the target case. However, such an approach is not suitable for our task, since if we optimize towards the center of "Pass" (i.e., to make the defenses of API very confident that the generated content is safe), it may result in generating images that are truly "safe" instead of our intended target of "harmful" for the attack.

Motivated by the observation that misclassifications often occur at the boundary between two classes in classifiers (Finlay et al., 2019), we propose situating the SPL's optimization objective at the boundary region of the two cases, with a slight inclination towards the "Pass" side. This choice is guided by the notion that, when positioned at the boundary region, the black box exhibits less confidence in its decisions, significantly enhancing the chances of successfully generating target "harmful" images. Next, we provide the detailed process of our SPL scheme.

Specifically, we conduct optimization in the space of LoRA parameters ψ, where each point in the optimization space is a specific configuration of ψ. As shown in Fig. 3 (a), SPL firstly samples several points on a sphere centered around the current model parameters (marked as a blue star) with r as the radius. Then, we query the black box's output case results of the sampled points. If all points are predicted as "Deny" (meaning no boundary are detected in the sphere), we increase the sphere radius r to probe the boundary. Once the sphere encompasses both "Pass" and "Deny" points, as shown in Fig. 3 (a), indicating that the boundary has been probed, we stop increasing the radius and start optimizing the LoRA parameters ψ. Intuitively, the points predicted as "Deny" represent the direction we aim to move away from. Hence, we calculate the average of these points and optimize the model parameters in the opposite direction of this average. Here, we define Φ, which marks points from which we need to move away:

$$\Phi(\psi) = \begin{Bmatrix} 0 & \textbf{if } \text{image returned ("Pass")} \\ -1 & \textbf{otherwise} \end{Bmatrix} \quad (1)$$

In other words, after sampling on the sphere, the function Φ use -1 to mark all "Deny" points. Based on this, we define our SPL gradient estimator as:

$$\boldsymbol{g}_{\text{spl}}(\psi) = \frac{1}{N} \sum_{n=1}^{N} \Phi(\psi + r \cdot z_n) \cdot z_n, \quad (2)$$

where $\psi \in \mathbb{R}^d$ denotes the current model parameters, z_n represents a vector randomly sampled from a d-dimensional space, r denotes the radius of the sample sphere, $\psi + r \cdot z_n$ is the sampling point, and N denotes the total number of points sampled on this sphere. Note that different vectors z_n possess different directions, yet they share the same length, resulting in all sampling points being located on a sphere. Leveraging the aforementioned sphere-sampling exploration, we can explore the boundary and identify an effective optimization direction $\boldsymbol{g}_{\text{spl}}$. In this way, even if the current model ψ is incapable of prompting the API to return

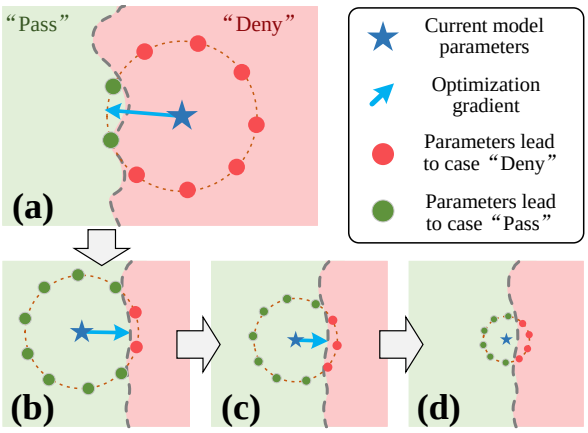

Figure 3. Intuitive illustration of our SPL scheme.

images, the gradient $\boldsymbol{g}_{\text{spl}}$ can still be computed to optimize the model ψ, enabling it to generate adversarial prompts capable of circumventing the defenses of the black-box API. Next, we use $\boldsymbol{g}_{\text{spl}}$ to update ψ as follows:

$$\psi = \psi + \alpha \cdot \boldsymbol{g}_{\text{spl}}(\psi), \tag{3}$$

where α is the learning rate, which is related to the current radius r, i.e., $\alpha = r/k$. We alternate between Eq. 2 and Eq. 3 iteratively, until the sphere center is predicted into the "Pass" case, i.e., from Fig. 3 (a) to (b).

As mentioned earlier, SPL's optimization targets the "Pass" region near the boundary rather than the central "Pass" region. This strategy aims to enhance the likelihood of generating target "harmful" images rather than genuinely safe ones. However, as we initially need to use large radius to find the "Pass" points, the optimized model may end up in the "Pass" but far from the boundary, as shown in Fig. 3 (b). Therefore, we then gradually reduce the radius and move in opposite direction to approach the boundary, i.e., from Fig. 3 (b) to (d). This aims to facilitate our second-stage training (Sec. 3.3) which is proposed to ensure the "harmful" semantic representation of the returned images. Specifically, for each optimization iteration, we decrease the radius r and use the updated smaller radius to calculate gradient $\boldsymbol{g}_{\text{spl}}(\psi)$ via Eqn. 2. Subsequently, we optimize the model ψ towards the boundary:

$$\psi = \psi - \alpha \cdot \boldsymbol{g}_{\text{spl}}(\psi), \tag{4}$$

Here, we utilize the negative sign "$-$", in contrast to Eq. 3, to signify a reversed optimization direction. We alternate between Eq. 2 and Eq. 4, aiming to iteratively approach the boundary. During this boundary-approaching process, the radius is gradually decreased, as shown in Fig. 3 (b, c and d). This gradual reduction of radius r is essential, since utilizing a fixed radius for optimization increases the likelihood of crossing the boundary and reaching the "Deny" side. By decreasing the radius incrementally, we ensure effective boundary approaching. It is important to note that, for each decreased radius, we need to check if the following conditions are met: (1) The sphere with this decreased radius encompasses both "Pass" and "Deny" points. (2) After optimization with this decreased radius, the model parameters are still located in the "Pass" case. These two conditions ensure the optimization with this decreased radius is effective. Finally, we stop the SPL scheme when r is smaller than a pre-defined threshold r_{min}.

After the pre-training stage with the SPL scheme, our UPAM can adeptly bypass black-box defenses, ensuring the API consistently returns images. It also releases the potential of generating target "harmful" images. In the subsequent stage, we finetune our UPAM to further enhance the semantic alignment of returned images with our "harmful" targets.

3.3. Fine-tuning Stage with SEL

After pre-training, we consistently obtain the image $\mathbf{I}^*$ from the black box. In this section, our emphasis is on ensuring that the returned image $\mathbf{I}^*$ maintains semantic consistency with the target image $\mathbf{I}$ while preserving UPAM's ability to deceive defenses for image retrieval (return). To achieve this, we further introduce a Semantic-Enhancing Learning (SEL) scheme to finetune the LoRA parameters.

Specifically, in our SEL, we propose a Semantic Measurement Loss $\mathcal{L}_{SM}$ to finetune the LoRA parameters ψ. Motivated by the powerful semantic representation capability of CLIP (Radford et al., 2021), we utilize its pre-trained image/text encoders to measure the semantics of the returned image $\mathbf{I}^*$. The semantic measurement is comprehensively handled from the following two perspectives:

(1) Image-Text similarity, i.e., encoding both the returned image $\mathbf{I}^*$ and the naive prompt $\mathbf{T}$, then computing the cosine similarity between the encoded vectors, which formulates the loss $\mathcal{L}_1$:

$$\mathcal{L}_1 = 1 - \frac{E_{img}(\mathbf{I}^*) \cdot E_{text}(\mathbf{T})}{\|E_{img}(\mathbf{I}^*)\| \, \|E_{text}(\mathbf{T})\|}, \tag{5}$$

where $E_{img}(\cdot)$ and $E_{text}(\cdot)$ denote the image encoder and text encoder of CLIP, respectively.

(2) Image-Image similarity, i.e., encoding both the returned image $\mathbf{I}^*$ and the target image $\mathbf{I}$ for computing the cosine similarity, which formulates the loss $\mathcal{L}_2$:

$$\mathcal{L}_2 = 1 - \frac{E_{img}(\mathbf{I}^*) \cdot E_{img}(\mathbf{I})}{\|E_{img}(\mathbf{I}^*)\| \, \|E_{img}(\mathbf{I})\|}. \tag{6}$$

Based on the above two similarity loss functions, we can formulate our Semantic Measurement Loss $\mathcal{L}_{SM}$ as:

$$\mathcal{L}_{SM} = \mathcal{L}_1 + \mathcal{L}_2. \tag{7}$$

By minimizing $\mathcal{L}_{SM}$, we can align the semantics of the returned images with our target. To this end, we use the

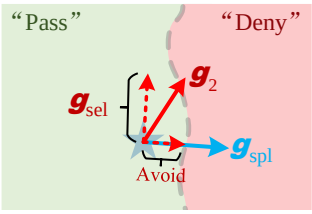

Figure 4. Intuitive illustration of our SEL scheme.

calculated $\mathcal{L}_{SM}$ to obtain gradients to optimize our model's parameters ψ. Since the T2I model is packed into the black box, we cannot directly obtain the oracle precise gradients. To address this black-box setting, drawing inspiration from (Spall, 1992b; 1997b), we adopt Zeroth-Order Optimization (ZOO) to estimate the gradient without accessing the model architecture and model parameters. Given the training loss $\mathcal{L}_{SM}$, the estimated gradient can be formulated as follows:

$$\boldsymbol{g}_1(\psi) = \frac{\mathcal{L}_{SM}(\psi + c \cdot \Delta) - \mathcal{L}_{SM}(\psi - c \cdot \Delta)}{2c \cdot \Delta}, \quad (8)$$

where $c \in (0, 1]$ is the decaying hyper-parameter and $\Delta \in \mathbb{R}^d$ is a random perturbation vector, sampled from mean-zero distributions while satisfying the finite inverse momentum condition (Spall, 1992b; 2005).

While the standard form of zeroth-order optimization is generally effective, it can still encounter poor convergence in practical applications (Spall, 2000). This problem, as discussed in (Spall, 1997a; Oh et al., 2023), is primarily attributed to the stochastic nature of gradient estimation, stemming from the random directions of perturbations. To overcome this challenge, we follow (Oh et al., 2023) to improve the gradient calculation rule (Eq. 8) as:

$$\boldsymbol{g}_2(\psi) = \bar{\boldsymbol{g}}_1 + \beta \cdot \boldsymbol{g}_1(\psi + \bar{\boldsymbol{g}}_1). \quad (9)$$

We can observe that the improved gradient $\boldsymbol{g}_2$ includes two gradient components: $\bar{\boldsymbol{g}}_1$ and $\beta \cdot \boldsymbol{g}_1(\psi + \bar{\boldsymbol{g}}_1)$. The former $\bar{\boldsymbol{g}}_1$ is the gradient used in the previous update iteration, while the latter $\beta \cdot \boldsymbol{g}_1(\psi + \bar{\boldsymbol{g}}_1)$ is the adaptive adjustment when the model continues to update along the previous gradient path. β is the learning rate. The advantage of this design is that it encourages the model to update along previously effective directions with necessary adjustments, thereby increasing the gradient consistency and reducing the incidental randomness of optimization.

As mentioned in Sec. 3.2, the SPL's optimization objective is situated at the boundary region of the two output cases of the black box, slightly leaning towards the "Pass" side, which both facilitates the generation of "harmful" image and ensure the generated images can be returned to the attacker. However, if we directly optimize the model using gradients $\boldsymbol{g}_2$ (Eq. 9), it tends to pull the model parameters towards the "Deny" region, preventing the black box from returning images. This occurs because $\boldsymbol{g}_2$ is calculated to enhance the semantic expression of "harmful" in images, which is typically what the black box considers as content to be denied. Therefore, our optimization objective here is to make the returned images contain the targeted "harmful" semantics as much as possible, while ensuring the model ψ remains in the boundary region, leaning towards "Pass".

To prevent SEL from disrupting SPL's learning achievements, we propose a gradient harmonization method to ensure compatibility between SEL and SPL. As shown in Fig. 4, we first obtain the SPL gradient $\boldsymbol{g}_{\text{spl}}$ through Eq. 2, where the direction of $\boldsymbol{g}_{\text{spl}}$ signifies the optimization direction towards the boundary (see the blue arrow in Fig. 4). Given the direction provided by $\boldsymbol{g}_{\text{spl}}$, we can eliminate the boundary-directed component of the gradient $\boldsymbol{g}_2$, thereby obtaining the SEL gradient $\boldsymbol{g}_{\text{sel}}(\psi)$, which can be formulated as:

$$\boldsymbol{g}_{\text{sel}}(\psi) = \boldsymbol{g}_2 - \frac{\boldsymbol{g}_{\text{spl}} \cdot \boldsymbol{g}_2}{|\boldsymbol{g}_{\text{spl}}|^2} \boldsymbol{g}_{\text{spl}}, \quad (10)$$

$$\psi = \psi - \beta \cdot \boldsymbol{g}_{\text{sel}}(\psi), \quad (11)$$

where the second term in Eq. 10 represents the gradient component directed towards the boundary. As shown in Fig. 4, by removing this component, we can avoid the optimization of the SEL moving towards the boundary, thereby not only aligning towards the target semantics but also maintaining the case of "Pass" (i.e., keeping images returned).

3.4. Training and Testing

We train our UPAM for two stages. In the pre-training stage, we use the SPL scheme to obtain the gradient $\boldsymbol{g}_{\text{spl}}$ (Eq. 2) and update the LoRA parameters (Eq. 3 or 4). This update enables effective optimization under the challenging scenario where no image results are returned, finally compelling the API to reliably return images. Then, in the fine-tuning stage, we adopt the SEL scheme to obtain the gradient $\boldsymbol{g}_{\text{sel}}$ (Eq. 10) and update the model parameters (Eq. 11). This update ensures that the returned images can accurately show target semantics. After that, we handle the inference (testing) based on the trained UPAM.

4. Experiments

Textual Filter and Visual Checker. We follow previous work (Liu et al., 2023) employing the same textual filter, which utilizes a list of sensitive words (i.e., blacklist) to deny the forward propagation of "harmful" prompts. For the visual checker, following (Rando et al., 2022), a widely-used visual checking strategy is adopted: using the off-the-shelf CLIP visual encoder (Radford et al., 2021) to map the generated image to a latent vector, which is then compared with all default embeddings of predefined "harmful" images using cosine similarity. If any cosine value exceeds the specified threshold, the further propagation of the generated image is denied.

Dataset. In consideration of ethics, privacy, legal compli-

Table 1. Experimental results under protocol A (left) and protocol B (right).

T2I Model	Methods	Protocol A					Protocol B				
		R-1 Precision ↑	R-3 Precision ↑	Text. Sim. ↓	Infer. Time (s) ↓	PPL ↓	R-1 Precision ↑	R-3 Precision ↑	Text. Sim. ↓	Infer. Time (s) ↓	PPL ↓
DALL·E	TextFooler (Jin et al., 2020)	0.48%	0.98%	0.22	589.29	3463.02	0/10	0/10	0.22	582.37	3517.03
	HomoSubs (Struppek et al., 2022)	0.74%	1.12%	0.28	562.48	3089.58	0/10	0/10	0.16	590.28	3216.95
	EvoPromp (Millière, 2022)	0.82%	1.37%	0.19	579.20	4984.24	0.2/10	0.5/10	0.16	586.47	4837.63
	HiddVocab (Daras & Dimakis, 2022)	3.35%	4.21%	0.33	457.34	4027.64	0.3/10	0.4/10	0.18	483.27	4607.93
	MacPromp (Millière, 2022)	5.05%	6.80%	0.57	452.16	3163.07	0.7/10	0.9/10	0.39	476.46	3239.56
	RIATIG (Liu et al., 2023)	8.65%	9.95%	0.18	301.52	1003.27	0.8/10	1.1/10	0.34	425.66	1026.81
	UPAM	**38.37%**	**41.83%**	0.17	**5.03**	**706.32**	**7.3/10**	**8/10**	0.15	**359.26**	**634.73**
DALL·E 2	TextFooler (Jin et al., 2020)	0.85%	0.93%	0.18	540.77	3566.32	0/10	0/10	0.19	543.72	3578.14
	HomoSubs (Struppek et al., 2022)	0.70%	0.89%	0.22	568.26	4125.22	0/10	0/10	0.15	589.46	4374.80
	EvoPromp (Millière, 2022)	0.92%	1.17%	0.37	475.81	5837.06	0/10	0/10	0.21	449.61	5267.07
	HiddVocab (Daras & Dimakis, 2022)	3.21%	4.62%	0.46	565.33	4125.22	0.6/10	0.9/10	0.22	576.27	4933.29
	MacPromp (Millière, 2022)	4.84%	5.92%	0.66	523.49	3056.41	0.4/10	0.8/10	0.30	571.63	3575.26
	RIATIG (Liu et al., 2023)	8.55%	9.19%	0.35	353.53	992.54	0.5/10	0.7/10	0.29	417.34	965.38
	UPAM	**40.45%**	**44.74%**	0.19	**5.42**	**788.45**	**6.5/10**	**7.2/10**	0.17	**341.59**	**773.01**
Imagen	TextFooler (Jin et al., 2020)	0.32%	0.54%	0.15	527.65	3321.88	0/10	0/10	0.23	559.42	3539.57
	HomoSubs (Struppek et al., 2022)	0.35%	0.75%	0.18	567.82	2850.41	0/10	0/10	0.13	586.75	2674.29
	EvoPromp (Millière, 2022)	0.64%	0.92%	0.25	398.51	4055.71	0.3/10	0.7/10	0.21	443.39	4162.18
	HiddVocab (Daras & Dimakis, 2022)	2.37%	3.69%	0.27	592.47	4153.58	0.3/10	0.5/10	0.18	576.83	4264.34
	MacPromp (Millière, 2022)	4.58%	5.52%	0.16	524.38	3173.35	0.5/10	0.8/10	0.43	463.52	3097.13
	RIATIG (Liu et al., 2023)	7.80%	9.22%	0.36	328.16	1076.29	0.9/10	1.2/10	0.34	485.36	1057.42
	UPAM	**36.16%**	**41.43%**	0.16	**5.16**	**521.78**	**6.7/10**	**7.3/10**	0.18	**350.72**	**434.08**

ance, and potential societal impacts, we follow previous works (Liu et al., 2023; Daras & Dimakis, 2022; Millière, 2022; Struppek et al., 2022; Jin et al., 2020), which do not attack T2I models to generate truly harmful content, instead deploying defenses and conducting attacks using the paired text-image data from the Microsoft COCO dataset (Lin et al., 2014). We treat 10 classes from COCO as "harmful" classes (i.e., *boat*, *bird*, *clock*, *kite*, *wine glass*, *knife*, *pizza*, *teddy bear*, *vase* and *laptop*). Specifically, we train our UPAM using the training set of these 10 classes, and then test UPAM using the test set of the same 10 classes. Although the training and testing data contain the same classes, the specific text descriptions and their corresponding image contents are different.

Target T2I Models. We follow previous work (Liu et al., 2023) to conduct experiments on three large-scale API T2I models: DALL·E (Yu et al., 2022), DALL·E 2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022). For each of them, we follow (Liu et al., 2023) to adopt the corresponding released model that achieves comparable performance to API, and treat it as a black box.

Implementation Details. In our experiment, we use the open-source LLaMA (Touvron et al., 2023) as the LLM. As for the LoRA adapter (Hu et al., 2021), we adopt the adaptation matrices with a rank value of 8. In our SPL scheme, we set $N = 10$, $\alpha = r/4$. In our SEL scheme, we set $\beta = 0.3$.

4.1. Experimental Protocols

In this paper, we conduct experiments under two protocols for a comprehensive study.

Protocol A: Given the training data $\{\mathbf{T}, \mathbf{I}\}$, we train a parameterized model for converting the naive prompts $\mathbf{T}$ into adversarial ones $\mathbf{T}^*$ with the objective of compelling the black-box T2I model to return the target images $\mathbf{I}$. Subsequently, given the testing data $\mathbf{T}_t$, the trained model can swiftly convert $\mathbf{T}_t$ into $\mathbf{T}_t^*$, compelling the black-box T2I model to return the corresponding target images.

Protocol B: Previous enumeration-based methods do not require model training. Instead, for each given paired data $\{\mathbf{T}, \mathbf{I}\}$, they can directly find its adversarial prompt $\mathbf{T}^*$ by repeatedly attempting. As shown in (Liu et al., 2023), they conduct experiments on a subset of 10 "harmful" samples (randomly selected from the COCO dataset). To make a fair comparison with previous enumeration-based methods, we slightly modify our UPAM in this protocol. This modification enables UPAM to identify adversarial prompts for the 10 samples without learning on the training set. The details of modification can be seen in *appendix*.

4.2. Evaluation Metrics.

Following previous work (Liu et al., 2023; Jin et al., 2020), We use four metrics to evaluate the attack model: R-precision, Textual Similarity, Inference Time, and PPL. Among these, R-precision (the higher the better) assesses the degree of semantic alignment between the generated images and target images, and also reflects the model's ability to deceive visual checkers. We follow (Liu et al., 2023) to conduct R-precision evaluation at two levels: R-1 and R-3. Textual Similarity (the lower the better) measures the model's ability to deceive textual filters. Inference Time (the lower the better) represents the average time taken for each test sample. PPL evaluates (the lower the better) the naturalness of adversarial prompts. Please refer to *appendix* for detailed descriptions of each metric.

4.3. Experimental Results

Experimental Results Under Protocol A. As shown in Tab. 1 (left), compared to other methods, our UPAM achieves the highest R-precision, both in R-1 precision and R-3 precision. Specifically, our UPAM significantly outperforms other methods by by an average of 35.31% (R-1) and 38.90%

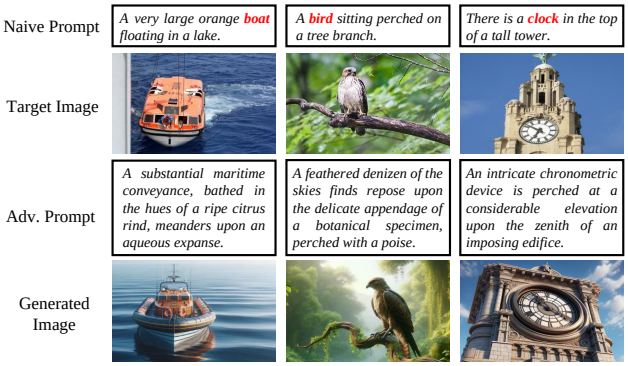

Figure 5. Qualitative results of our attack against DALL·E. The names of "harmful" classes are marked in red.

Table 2. Ablation study on each design of our approach.

	Methods	R-1 ↑	R-3 ↑	Text. Sim. ↓	Infer. Time ↓	PPL ↓
(a)	UPAM	38.37%	41.83%	0.17	5.03	706.32
(b)	UPAM w/o SPL	0.22%	0.37%	0.53	5.10	707.02
(c)	UPAM w/o SEL	23.25%	28.36%	0.17	5.06	706.58

(R-3), respectively. Additionally, our UPAM keeps a consistently low textual similarity (less than 0.2) across all black-box T2I models. These results collectively demonstrate the effectiveness of our gradient-based method in tackling the challenging task of deceiving both textual and visual defenses under the black-box setting. Furthermore, UPAM exhibits a significantly shorter inference time compared to other methods, showcasing superior efficiency. In terms of the attack stealthiness, compared to other methods, our LLM-based method achieves a significantly lower PPL, which demonstrates the effectiveness of our method in generating natural adversarial prompts.

Experimental Results Under Protocol B. In protocol B, given a paired text-image data $\{\mathbf{T}, \mathbf{I}\}$ as the attack target, we aim to directly find out the corresponding adversarial prompt $\mathbf{T}^*$ for this target. We randomly choose 10 text-image samples from 10 "harmful" classes of COCO (one for each class) to find their corresponding adversarial prompts, and then evaluate the attack performance. We conduct experiments 10 times and show the average results in Tab. 1 (right). Our UPAM achieves significantly higher R-precision and lower textual similarity compared to other methods. Even in protocol B, where we don't train a model on the training set, our approach achieves the shortest inference time. This is due to our method's ability to use gradients for discovering adversarial prompts, making it much more efficient than enumeration-based methods. Additionally, our method exhibits the lowest PPL, indicating the best naturalness of the generated adversarial prompts. These results demonstrate the superiority of our approach in protocol B.

Qualitative Results. Qualitative results of our UPAM are shown in Fig. 5. We can see that our generated images are semantically consistent with the ground-truth target images. Our adversarial prompts are significantly different from the naive prompts and do not contain any "harmful" (malicious) words. Moreover, we can hardly identify the misspelled (unnatural) words, indicating strong attack stealthiness. Qualitative comparisons with existing methods are in *appendix*.

Table 3. Ablation study on Moving Closer to Boundary (MCB) in SPL.

	Methods	R-1 ↑	R-3 ↑	Text. Sim. ↓	Infer. Time ↓	PPL ↓
(i)	UPAM	38.37%	41.83%	0.17	5.03	706.32
(ii)	UPAM w/o MCB	35.76%	39.31%	0.22	5.03	706.10

Table 4. Ablation study on Gradient Harmonization (GH) in SEL.

	Methods	R-1 ↑	R-3 ↑	Text. Sim. ↓	Infer. Time ↓	PPL ↓
(i)	UPAM	38.37%	41.83%	0.17	5.03	706.32
(ii)	UPAM w/o GH	33.82%	35.08%	0.35	5.09	706.27

4.4. Ablation Study

To demonstrate the effectiveness of our proposed schemes: SPL and SEL, we conduct ablation experiments by removing the specific SPL (or SEL) in our approach. Comparing (a) and (b) in Tab. 2, it can be seen that when UPAM does not employ SPL, both R-1 precision and R-3 precision drop to almost 0. This is because SPL is designed to deceive textual and visual defenses. In the absence of SPL, black-box T2I models hardly return images, resulting in poor R-precision performance. Additionally, SPL also influences the Textual Similarity score, as it can reduce the similarity between adversarial prompts and naive ones, aiming to deceive textual filters within the black-box system. These ablation results demonstrate the effect of SPL. When comparing (a) and (c) in Tab. 2, we can see that our UPAM without SEL leads to a significant R-precision decrease of 15.12% (R-1) and 13.47% (R-3), demonstrating the effect of SEL. We can observe that SEL has virtually no impact on Textual Similarity, Inference Time, and PPL. This is because SEL is proposed to enhance the semantic representation of the returned images, thereby only affecting R-precision. As SPL and SEL do not alter the LLM-based structure of UPAM, they do not impact the Inference Time and PPL.

4.5. Effect of further moving closer to the boundary in SPL

In SPL, after optimizing the model parameters into the "Pass" region, we gradually reduce the radius, moving the model parameters closer to the boundary, aiming to increase the likelihood of the generation of target "harmful" images, as illustrated in Fig. 3 (b, c, and d). To investigate the effect of this design, we conduct ablation by removing this further approach. For clarity, we name this design as Moving Closer to Boundary (MCB). By comparing (i) and (ii) in Tab. 3, when removing MCB in SPL, obvious performance

Table 5. Ablation study on LLM and LoRA.

	Methods	R-1 ↑	R-3 ↑	Text. Sim. ↓	Infer. Time ↓	PPL ↓
(i)	UPAM	38.37%	41.83%	0.17	5.03	706.32
(ii)	UPAM w/o LLM	38.15%	41.68%	0.18	5.13	3954.85
(iii)	UPAM w/o LoRA	38.22%	41.59%	0.17	5.05	2848.52

degradation is observed in terms of R-1, R-3, and Textual Similarity. This demonstrates the effectiveness of further moving closer to the boundary in SPL.

4.6. Effect of gradient harmonization in SEL

To improve the compatibility of SEL with SPL, we introduce a Gradient Harmonization (GH) method (see Eq. 10) to adaptively adjust SEL's gradients. With a comparison between (i) and (ii) in Tab. 4, we can see that, when removing GH, SEL could disrupt SPL's optimization achievements, thus leading to worse performances in terms of R-1, R-3, and Textual Similarity. This demonstrates the effectiveness of gradient harmonization in SEL.

4.7. Effect of LLM+LoRA

To ensure the attack stealthiness, we propose to utilize a pre-trained LLM and optimize the LoRA adapter to ensure the naturalness of the generated adversarial prompts. Here, we conduct two ablation studies: **(1)** instead of using the pr-trained LLM, we directly employ an untrained transformer (Touvron et al., 2023) (with the same structure as the LLM) and then train it from scratch. Comparing (i) and (ii) in Tab. 5, we can see the transformer performs much worse in terms of PPL, demonstrating the effect of adopting knowledge of LLM. **(2)** Instead of optimizing LoRA, we directly optimize the pre-trained LLM. Comparing (i) and (iii) in Tab. 5, the significant change in PPL demonstrates the effect of using LoRA adapter.

5. Discussion

In order to prevent malicious use of T2I models to produce harmful or inappropriate content, API providers deploy text filters and visual checkers to prevent the output of harmful images (Mid, 2023; Leo, 2023; Rombach et al., 2022). However, as demonstrated in this paper, it is possible to use black-box knowledge to generate imperceptible adversarial prompts for obtaining target images. Therefore, this work further shows the importance of being aware of the security risks of T2I models.

One possible defense mechanism is to finetune the T2I models through *adversarial training* (Shafahi et al., 2019), a common defense method in AI security (Goodfellow et al., 2014; Jin et al., 2020), aiming to make the T2I models unable to produce the target image when feeding adversarial prompts. However, this approach tends to exhibit a bias towards adversarial prompts used during training, resulting in sub-optimal performance when encountering adversarial prompts from untrained domains (Wang et al., 2023). Another possible defense mechanism is to leverage *model unlearning* (Liu et al., 2022) to make T2I models essentially forget harmful knowledge. However, it typically requires a significant effort to collect "unsafe" images as comprehensively as possible. Moreover, sometimes the scope of harmful images can be difficult to define, making the data collection more labor-intensive and time-consuming (Marchant et al., 2022).

6. Conclusion

In this paper, we introduce UPAM, a unified attack framework for T2I models, designed to deceive both textual filters and visual checkers. Unlike existing enumeration methods, UPAM enables gradient-based optimization for higher effectiveness and efficiency. We devise an SPL scheme to support gradient optimization in the challenging no-feedback scenario, and introduce an SEL scheme to align generated images with desired target semantics. We incorporate LLM to guarantee attack stealthiness.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-027[T]) and the Ministry of Education, Singapore, under the AcRF Tier 2 Projects (MOE-T2EP20222-0009 and MOE-T2EP20123-0014).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential social consequences of our work, none of which we feel must be specifically highlighted here. It is important to note that all experiments are conducted by utilizing natural safe images, with the utmost care to avoid any dissemination of negative consequences. Our research investigates the robustness of text-to-image generation models and highlights the vulnerabilities inherent in these models, and thereby can inform feasible defense approaches for future development in this area. Our study shall be able to encourage the machine learning community to develop trust-worthy, responsible, and robust models.

References

Leonardo.ai, access date: 9st nov. 2023.

Midjourney, access date: 26th sept. 2023.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John,

R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Chesney, B. and Citron, D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

Chien, J.-T. *Source separation and machine learning*. Academic Press, 2018.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974. PMLR, 2021.

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B., and Wang, W. W. Ai, governance and ethics: global perspectives. *University of Hong Kong Faculty of Law Research Paper*, (2020/051), 2020.

Daoud, M. S., Shehab, M., Al-Mimi, H. M., Abualigah, L., Zitar, R. A., and Shambour, M. K. Y. Gradient-based optimizer (gbo): a review, theory, variants, and applications. *Archives of Computational Methods in Engineering*, 30 (4):2431–2449, 2023.

Daras, G. and Dimakis, A. G. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169*, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021.

Finlay, C., Pooladian, A.-A., and Oberman, A. The logbarrier adversarial attack: making effective use of decision boundary information. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4862–4870, 2019.

Galetin, M., Škorić, J., and Mihajlović, M. Review of the standpoints regarding the content of the article 5 of the proposal for the european union's artificial intelligence act: The challenge of finding the balance. *Univerzitetska misao-časopis za nauku, kulturu i umjetnost, Novi Pazar*, (21):164–175, 2022.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1487–1495, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, H., Wu, S., Liang, X., Wang, B., Shi, Y., Wu, P., Yang, M., and Zhao, T. Towards making the most of llm for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 375–386. Springer, 2023.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Kahla, M., Chen, S., Just, H. A., and Jia, R. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15045–15053, 2022.

Kalyan, K. S. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048, 2024.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, H., Wu, Y., Zhai, S., Yuan, B., and Zhang, N. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20585–20594, 2023.

Liu, Y., Fan, M., Chen, C., Liu, X., Ma, Z., Wang, L., and Ma, J. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 280–289. IEEE, 2022.

Marchant, N. G., Rubinstein, B. I., and Alfeld, S. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7691–7700, 2022.

Millière, R. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*, 2022.

Oh, C., Hwang, H., Lee, H.-y., Lim, Y., Jung, G., Jung, J., Choi, H., and Song, K. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24224–24235, 2023.

Peng, D., Hu, P., Ke, Q., and Liu, J. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 808–820, 2023a.

Peng, D., Ke, Q., Lei, Y., and Liu, J. Unsupervised domain adaptation via domain-adaptive diffusion. *arXiv preprint arXiv:2308.13893*, 2023b.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., and Azam, S. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *Authorea Preprints*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rando, J., Paleka, D., Lindner, D., Heim, L., and Tramèr, F. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Spall, J. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992a.

Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992b.

Spall, J. C. Accelerated second-order stochastic optimization using only function measurements. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 2, pp. 1417–1424. IEEE, 1997a.

Spall, J. C. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1): 109–112, 1997b.

Spall, J. C. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE transactions on automatic control*, 45(10):1839–1853, 2000.

Spall, J. C. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.

Stephan, M., Hoffman, M. D., Blei, D. M., et al. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Struppek, L., Hintersdorf, D., and Kersting, K. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, Q., Liu, Y., Ling, H., Li, Y., Liu, Q., Li, P., Chen, J., Yuille, A., and Yu, N. Continual adversarial defense. *arXiv preprint arXiv:2312.09481*, 2023.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

Zhou, W., Wang, P., Heidari, A. A., Zhao, X., Turabieh, H., and Chen, H. Random learning gradient based optimization for efficient design of photovoltaic models. *Energy Conversion and Management*, 230:113751, 2021.

Zhu, T., Ye, D., Zhou, S., Liu, B., and Zhou, W. Label-only model inversion attacks: Attack with the least information. *IEEE Transactions on Information Forensics and Security*, 18:991–1005, 2022.

A. Details of Evaluation Metrics

We evaluate attack methods from the following perspectives:

Attack Effectiveness. Following previous work (Liu et al., 2023), we measure the attack effectiveness by evaluating the generated adversarial prompts $\mathbf{T}^*$ and the returned images $\mathbf{I}^*$, respectively. (1) As for the returned images $\mathbf{I}^*$, we follow (Liu et al., 2023) to utilize R-precision to measure the attack effectiveness. Specifically, a retrieval experiment is conducted using the returned image to query against a set of candidate text descriptions. This set comprises 1 ground-truth description and 99 randomly chosen mismatched descriptions. For this purpose, a state-of-the-art image-text retrieval model (Kim et al., 2021) is employed here. We collect the top R ranked results in the retrieval, deeming our attacks successful if the ground truth description is among them. We follow (Liu et al., 2023) to conduct evaluation at R-1 precision (i.e., $R = 1$) and R-3 precision (i.e., $R = 3$). Note that if no image is returned to the attacker, the R-precision score of this sample is 0. (2) As for the generated adversarial prompts $\mathbf{T}^*$, we use textual similarity to evaluate the attack effectiveness. The effective adversarial prompts should show low textual similarity with naive prompts. Therefore, following previous work (Liu et al., 2023; Jin et al., 2020), we adopt the Universal Sentence Encoder (Cer et al., 2018) to encode the adversarial prompts $\mathbf{T}^*$ and naive prompts $\mathbf{T}$ into high dimensional vectors, and then compute their textual similarity score. The lower the textual similarity, the better the attack.

Attack Efficiency. We also evaluate the attack efficiency of the methods. Specifically, we present the average inference time taken for each testing sample during testing. Due to the uncertain inference time of enumeration methods, in order to avoid endless waiting periods, we set a criterion: if the inference time exceeds 10 minutes for a given sample, the attack is considered unsuccessful, and the inference process is stopped.

Attack Stealthiness. To evaluate the attack stealthiness, we follow (Liu et al., 2023) to evaluate the naturalness of the generated adversarial prompts by computing the perplexity score (PPL) using GPT-2 (Radford et al., 2019), a model trained on extensive real-world sentences. Generally, the prompt sample with a lower PPL is more natural.

B. Modification of UPAM for Protocol B

As for previous enumeration-based methods, given a naive prompt $\mathbf{T}$ and a target image $\mathbf{I}$, instead of training parameterized models, they aim to directly find an adversarial prompt $\mathbf{T}^*$ based on the given data $\mathbf{T}$ and $\mathbf{I}$. We refer to this as Protocol B. To ensure a fair comparison with previous enumeration-based methods, we make slight modifications to UPAM to enable it to directly find adversarial prompts without learning from the training set. The modified architecture is illustrated in Fig. 6. Next, we provide modification details as follows:

1. As for the structure of our UPAM, we no longer utilize the LoRA adapter while only adopting the LLM. The parameters of the LLM remain frozen.

2. Given $\{\mathbf{T}, \mathbf{I}\}$, following the processes of SPL and SEL of our main paper, we still compute gradients $\boldsymbol{g}_{\text{spl}}$ and $\boldsymbol{g}_{\text{sel}}$. The difference is that, these gradients no longer optimize LoRA parameters but optimize the input soft embedding $\mathbf{P}$. The soft embedding $\mathbf{P}$ is obtained by feeding the naive $\mathbf{T}$ into the text encoder of LLM.

In this way, given a naive prompt $\mathbf{T}$ and a target image $\mathbf{I}$, we can directly find a proper adversarial prompt $\mathbf{T}^*$ by optimizing the input embedding $\mathbf{P}$, without training any model parameters.

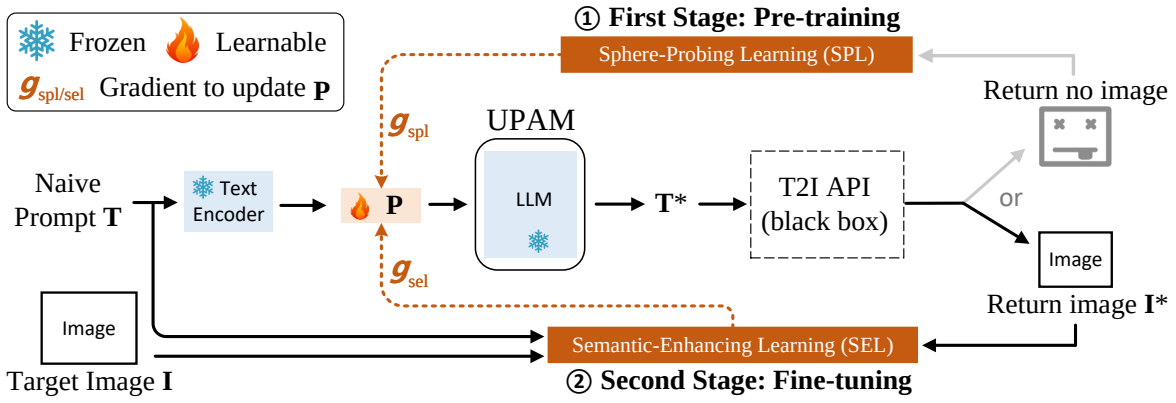

Figure 6. Overview of our UPAM framework modified for Protocol B.

C. More Qualitative Results

Here, we present additional qualitative results. Specifically, we compare our UPAM with HiddVocab (Daras & Dimakis, 2022), MacPromp (Millière, 2022), and RIATIG (Liu et al., 2023), while not showcasing the results of TextFooler (Jin et al., 2020), HomoSubs (Struppek et al., 2022), and EvoPromp (Millière, 2022). This is because the latter three methods can hardly compel the black-box API to return images in this challenging setting, and thus their results cannot be presented.

In Fig. 7, we showcase the attack results. We can see that, in order to bypass textual filters, all methods generate adversarial prompts that no longer contain the sensitive words of naive prompts (marked in red). However, existing methods produce adversarial text with noticeably unnatural words (highlighted with blue underlines). In contrast, our method generates adversarial prompts that contain no misspelled words and maintain grammatical correctness, thereby exhibiting superior naturalness. Moreover, compared to other methods, our UPAM achieves the closest semantic alignment between the generated images and target images.

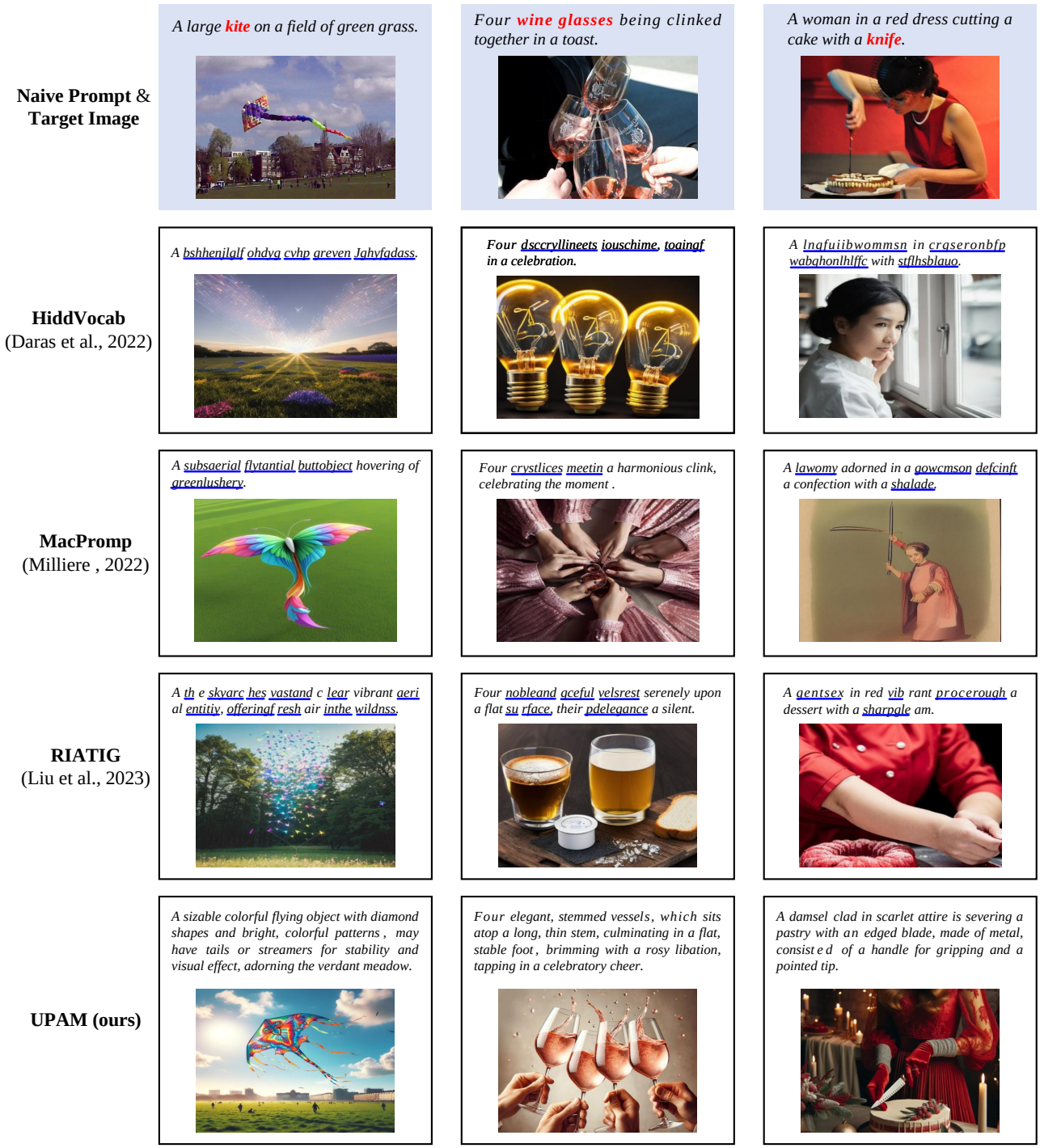

Figure 7. Qualitative comparison results with existing methods. In each block, the text above represents the adversarial prompts generated by the corresponding method, while the pictures below show the images returned from the T2I model. The name of each "harmful" class is marked in red. We employ blue underlines to show unnatural words within adversarial prompts.