

SCA: Improve Semantic Consistent in Unrestricted Adversarial Attacks via DDPM Inversion

Zihao Pan¹, Lifeng Chen², Weibin Wu^{1*}, Yuhang Cao¹, Zibin Zheng¹

¹School of Software Engineering, Sun Yat-sen University

²Computer Science, Beijing Jiaotong University

Abstract

Systems based on deep neural networks are vulnerable to adversarial attacks. Unrestricted adversarial attacks typically manipulate the semantic content of an image (e.g., color or texture) to create adversarial examples that are both effective and photorealistic. Recent works have utilized the diffusion inversion process to map images into a latent space, where high-level semantics are manipulated by introducing perturbations. However, they often result in substantial semantic distortions in the denoised output and suffer from low efficiency. In this study, we propose a novel framework called Semantic-Consistent Unrestricted Adversarial Attacks (SCA), which employs an inversion method to extract edit-friendly noise maps and utilizes a Multimodal Large Language Model (MLLM) to provide semantic guidance throughout the process. Under the condition of rich semantic information provided by MLLM, we perform the DDPM denoising process of each step using a series of edit-friendly noise maps and leverage DPM Solver++ to accelerate this process, enabling efficient sampling with semantic consistency. Compared to existing methods, our framework enables the efficient generation of adversarial examples that exhibit minimal discernible semantic changes. Consequently, we for the first time introduce Semantic-Consistent Adversarial Examples (SCAE). Extensive experiments and visualizations have demonstrated the high efficiency of SCA, particularly in being on average 12 times faster than the state-of-the-art attacks. Our code can be found at <https://github.com/Pan-Zihao/SCA>.

1 Introduction

Deep neural networks have excelled in learning features and achieving impressive performance across various tasks. However, they are vulnerable to small perturbations, known as adversarial samples, which raise significant security concerns for critical decision-making systems [29, 39, 3, 30, 31, 42, 72, 7]. With the explosive growth of digital information, they have triggered security risks such as face recognition [54, 73], voice recognition [26, 55, 56] and information forensics [13, 58].

To ensure human visual imperceptibility and maintain photorealism, adversarial examples must retain high semantic consistency with the original images, appearing natural and realistic despite minor perturbations [19, 34, 60, 71]. A recent trend in adversarial attacks is the emergence of unrestricted adversarial attacks [2, 5, 34, 48], which alter semantic features such as color and texture to produce adversarial examples that are both visually plausible and highly effective [19, 36]. State-of-the-art methods [10, 63] utilize generative models like Stable Diffusion [45], leveraging their knowledge of natural image distributions to map images onto a latent space, where small perturbations are introduced to optimize the adversarial objective. While this approach allows for more flexible adversarial examples, it often causes noticeable semantic shifts, reducing stealthiness to human observers [28, 70]. A comparison of our method with previous work on semantic consistency is

*Corresponding Author

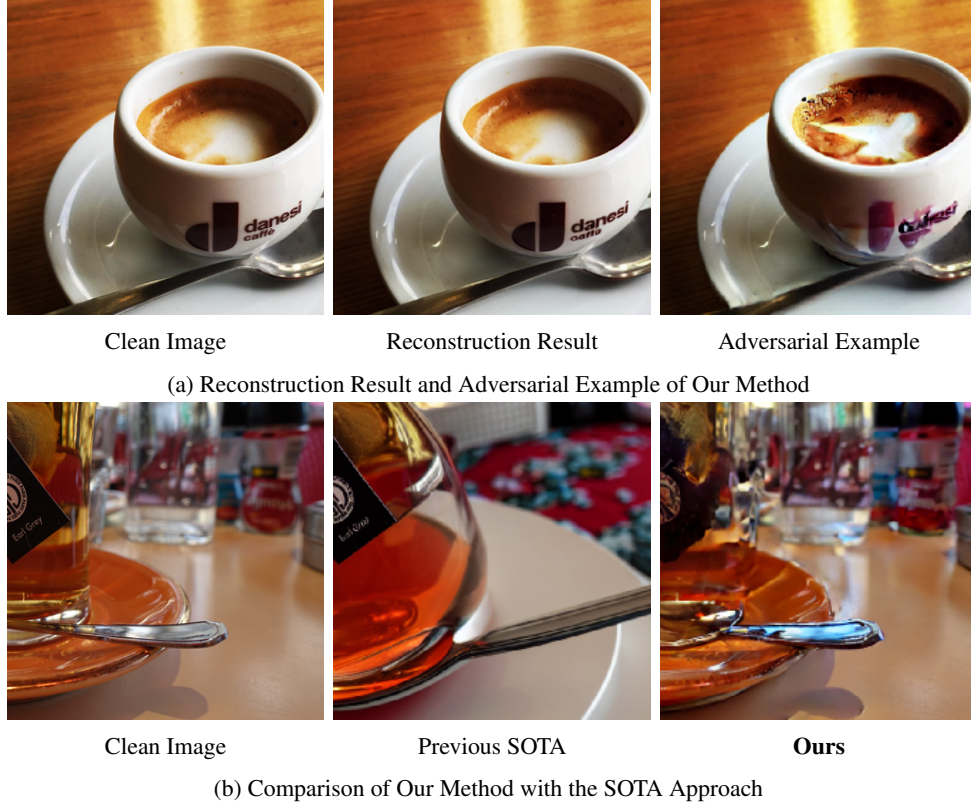


Figure 1: The above shows the superiority of SCA in maintaining semantic consistency. (a) shows that our method achieves perfect reconstruction and introduces minimal perturbations later. It can be seen that we only made some changes to the text on the coffee cup and the coffee pattern, and the overall semantics and environment remained consistent. This shows that our method can accurately identify the subject in the image and attack it without affecting the overall image, greatly improving its concealment from human perception. (b) is a comparison between the current state-of-the-art method called ACA and SCA. Compared with the adversarial examples generated by SCA, ACA causes a large semantic deviation in the image.

shown in Figure 1. We argue that even small changes in the latent space can result in significant image alterations. Existing methods, constrained by l_p norms, struggle to control these semantic variations, as shown in Figure 2. Furthermore, they typically require extensive time steps (e.g., 1000) for accurate image inversion, making the attack process both time-consuming and prone to uncontrolled semantic changes [10, 63].

To tackle the challenges inherent in adversarial attacks, we introduce a novel approach called Semantic-Consistent Unrestricted Adversarial Attack (SCA), which leverages rich semantic information from the original images as a guide throughout the entire attack process. Our method begins by enhancing the inversion mechanism. Denoising Diffusion Probabilistic Models (DDPMs) typically produce images by progressively refining a series of white Gaussian noise samples. Departing from prior approaches [9, 11, 60] we propose an alternative latent noise space for DDPM, which allows for a wide range of editable operations through straightforward manipulation. This space enables the extraction of noise maps that are conducive to editing for a given image. In the context of generating unrestricted adversarial examples, an essential property of inversion is that perturbing the latent space while keeping the noise maps fixed should result in artifact-free images that retain a high degree of semantic consistency with the original image. To achieve this, the noise maps are designed to be statistically independent and to follow a standard normal distribution, akin to conventional sampling techniques. Our inversion process imprints the image more strongly onto the noise maps, thereby preserving semantics more effectively when traversing the low-dimensional manifold. This preservation is because edit-friendly noise maps exhibit higher variances than their native counterparts.

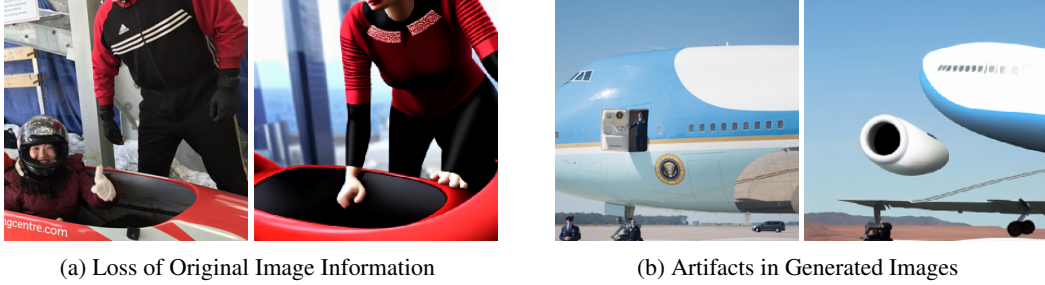


Figure 2: The above shows some examples of failure of the existing state-of-the-art methods. We can see that although it can generate rich and diverse content with the help of text-to-image models, large semantic deviations often exist. The image on the left is clean, and the image on the right is a generated adversarial example. We can see that the person’s information is lost in (a). The original image is described as “a woman sitting in a sled and a man standing beside her”. However, the generated result loses the information of the person sitting in the sled. In addition, the object structure is abnormal in (b).

Additionally, Multimodal Large Language Models (MLLMs) have demonstrated a strong capacity for understanding a wide array of image semantics [65]. We leverage MLLMs to generate highly detailed image descriptions, which guide both the inversion and generation processes, thereby providing rich prior knowledge to maintain semantic consistency during perturbation in the latent space. Finally, we integrate DPM Solver++ [37], an optimized numerical algorithm that accelerates the sampling process of DDPMs by reducing the number of required time steps to just 10–20. This significantly boosts the overall efficiency of our method.

In summary, our main contributions are as follows:

- We introduce a novel attack framework, **Semantic-Consistent Unrestricted Adversarial Attack**, which leverages an effective inversion method and powerful MLLM to generate adversarial examples with minimal semantic distortion.
- We develop a more covert and natural unrestricted adversarial attack through **Semantic Fixation Inversion** and **Semantically Guided Perturbation**. This results in **Semantic-Consistent Adversarial Examples (SCAE)**, which maintain content diversity while deceiving DNN models without altering the image’s overall semantics, making them harder to detect and defend against. Our approach provides insights into DNN vulnerabilities and informs novel defense strategies.
- We demonstrate the effectiveness of our attack through comprehensive experiments and visual analyses. We achieve comparable attack success rates and transferability to current state-of-the-art methods, with a 12x improvement in efficiency and enhanced semantic consistency.

2 Related Work

Adversarial attacks aim at maximizing the classification error of a target model without changing the semantics of the images. In this section, we review the types of attacks and existing works.

Norm-Bounded Attacks. Norm-bounded attacks are popular for their simplicity and effectiveness, using gradient-based optimization to maximize model loss [29, 72, 42]. Techniques like FGSM [18], L-BFGS [51], PGD [38], and the CW attack [6] constrain perturbations within an l_p norm, ensuring minimal modifications. However, these methods often introduce distortions detectable by humans and robust models, making them less effective in real-world applications. Norm constraints also limit flexibility, as l_p norms do not always align with human perception, leading to visually suboptimal results.

Unrestricted Adversarial Attacks. As limitations of normbounded attacks become increasingly apparent, a growing interest in unrestricted adversarial attacks has emerged. These attacks aim to replace the traditional small l_p norm perturbations with unrestricted but natural transformations,

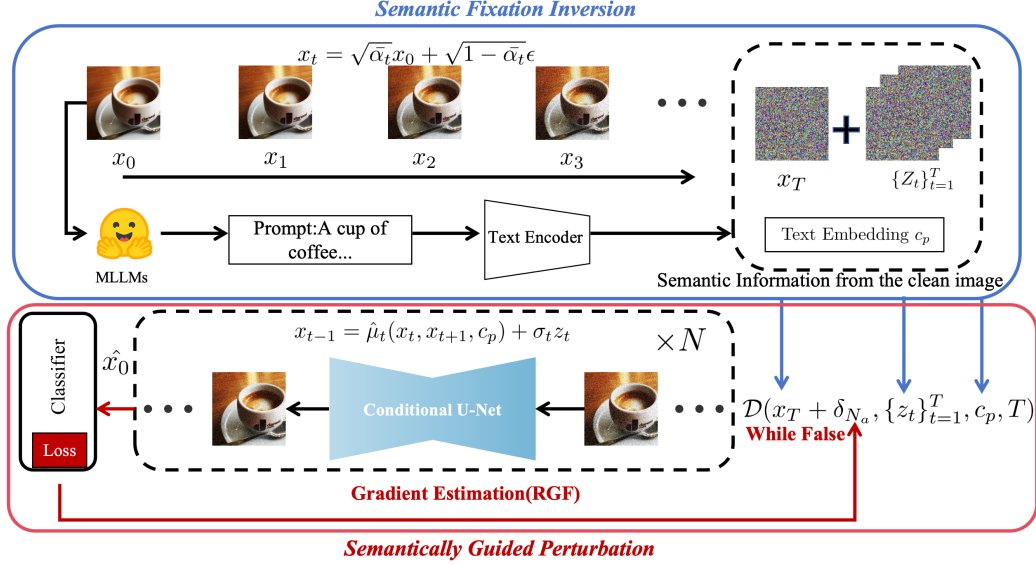


Figure 3: Pipeline of Semantic-Consistent Unrestricted Adversarial Attack. We first map the clean image into a latent space through Semantic Fixation Inversion, and then iteratively optimize the adversarial objective in the latent space under semantic guidance, causing the content of the image to shift in the direction of deceiving the model until the attack is successful.

making them more applicable in real-world scenarios. Unrestricted adversarial attacks typically rely on semantic transformations, such as changes in shape, texture, or color, to craft adversarial examples. Shape-based unrestricted attacks, such as those by Alaifari et al. [1] and Xiao et al. [61], iteratively introduce minor deformations to an image using gradient descent to generate adversarial examples. These methods exploit subtle changes in the geometry of the image but are often limited by the complexity of maintaining natural-looking outcomes. Following shape-based approaches, texture-based attacks emerged as an alternative. Bhattad et al. [2], Qiu et al. [43] and Mou et al. [39] proposed methods that manipulate an image’s texture or style to induce adversarial behavior. While texture-based attacks provide an effective means of altering perceptual features, they often lead to visually unnatural results, thereby diminishing their practical utility and adversarial transferability across different models. The unnatural distortions generated by these approaches highlight the challenge of balancing perceptual realism with adversarial effectiveness.

To address these challenges, researchers turned to color-based unrestricted attacks, which aim to manipulate pixel values along specific dimensions to maintain a higher degree of realism. Works by Hosseini et al. [19], Bhattad et al. [2], Zhao et al. [69], Shamsabadi et al. [47], and Yuan et al. [66] focus on altering color properties to generate adversarial examples that appear more natural to human observers. Despite achieving photorealism, these attacks often sacrifice flexibility in generating diverse adversarial examples. The reliance on subjective visual intuitions, coupled with the use of objective metrics like perceptual similarity, constrains the range of possible adversarial transformations. Consequently, color-based attacks frequently exhibit limited transferability to other models, reducing their effectiveness in broader adversarial settings.

In summary, while shape-, texture-, and color-based unrestricted attacks have contributed significantly to the field of adversarial machine learning, each method comes with inherent trade-offs between realism, flexibility, and adversarial transferability. Our work builds on these insights, aiming to overcome the limitations of existing approaches by developing a more flexible and transferable framework for generating natural adversarial examples without compromising photorealism.

Latent Space-Based Attacks. Recent advancements in adversarial attack techniques have focused on manipulating the latent space of generative models to craft adversarial images. For instance, Zhao et al. [70], Lin et al. [33], Hu et al. [21], Lapid et al. [27], and Lau et al. [28] leveraged Generative Adversarial Networks (GANs), fine-tuning the generator to produce adversarial images. These

Algorithm 1: SCA

input : A input image x_0 with the label y , a text embedding $c_p = \psi(\mathcal{P})$, a classifier $\mathcal{F}_\theta(\cdot)$, DDPM steps T , attack iterations N_a , momentum factor μ , and N is the number of queries.
output : Semantic-Consistent Unrestricted Adversarial Example $\hat{x}_0$

```
1 //Semantic Fixation Inversion;
2 for  $t = 1 \rightarrow T$  do
3    $\tilde{\epsilon} \sim \mathcal{N}(0, 1)$ ;
4    $x_t \leftarrow \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\tilde{\epsilon}$ ;
5 for  $t = T - 1 \rightarrow 1$  do
6    $z_t \leftarrow \frac{x_{t-1} - \hat{\mu}_t(x_t, x_{t+1}, c_p)}{\sigma_t}$ ;
7    $x_{t-1} \leftarrow \hat{\mu}_t(x_t, x_{t+1}, c_p) + \sigma_t z_t$ ;
8 //Semantically Guided Perturbation;
9 for  $k = 1, \dots, N_a$  do
10   $\hat{x}_0 \leftarrow \mathcal{D}(x_T, \{z_t\}_{t=1}^T, c_p, T)$ ;
11   $g_k \leftarrow \mu \cdot g_{k-1} + \frac{\nabla_{z_T} \mathcal{L}(\mathcal{F}_\theta(\varrho(\hat{x}_0), y))}{\|\nabla_{z_T} \mathcal{L}(\mathcal{F}_\theta(\varrho(\hat{x}_0), y))\|_1}$ ;
12   $\delta_k \leftarrow \Pi_\kappa(\delta_{k-1} + \eta \cdot \text{sign}(g_k))$ ;
13  $\hat{x}_0 \leftarrow \varrho(\mathcal{D}(x_T + \delta_{N_a}, \{z_t\}_{t=1}^T, c_p, T))$ ;
```

methods typically adjust the generator’s parameters to subtly modify the image’s latent representation, thus creating visually plausible yet adversarial content. On the other hand, diffusion models have also been employed for similar purposes. Xue et al. [64], Wang et al. [53], Chen et al. [9], Liu et al. [35], and Chen et al. [10] focused on optimizing the U-Net structure or injecting learned noise into the latent space to generate adversarial examples. Although diffusion models offer enhanced flexibility in controlling image generation processes, a common issue across both GAN-based and diffusion-based approaches is the pronounced semantic shift in the adversarial examples. These shifts can often make the manipulations more perceptible to humans, thus undermining the stealthiness of the attack. In contrast, our proposed framework addresses these limitations by generating adversarial examples that maintain a higher degree of semantic consistency, ensuring the resulting images exhibit minimal perceptible alterations while still achieving the desired adversarial effect. This improvement in stealthiness enhances the practical viability of adversarial attacks in real-world applications.

3 Preliminaries

Diffusion models (DMs) generate data by iteratively denoising a Gaussian-distributed random variable to approximate a learned data distribution through forward and reverse processes. In the forward process, an image x_0 is transformed into Gaussian noise via:

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} n_t, \quad t = 1, \dots, T, \quad (1)$$

where $n_t \sim \mathcal{N}(0, \mathbf{I})$ is noise, and β_t controls the noise injection. This leads to $x_T \sim \mathcal{N}(0, \mathbf{I})$. Marginalizing over intermediate states gives:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \quad (2)$$

with $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. In the reverse process, the model denoises x_T to reconstruct $\hat{x}_0$ as:

$$x_{t-1} = \hat{\mu}_t(x_t) + \sigma_t z_t, \quad t = T, \dots, 1, \quad (3)$$

where $\sigma_t = \eta \beta_t^{\frac{1 - \bar{\alpha}_t - 1}{1 - \bar{\alpha}_t}}$ controls stochasticity. The mean prediction is:

$$\begin{aligned}\hat{\mu}_t(x_t) = & \sqrt{\bar{\alpha}_{t-1}} \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}_\theta(x_t)}{\sqrt{\bar{\alpha}_t}} \\ & + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \hat{\epsilon}_\theta(x_t).\end{aligned}\tag{4}$$

For a text-conditioned generation, classifier-free guidance adjusts the noise prediction:

$$\hat{\epsilon}_\theta(x_t, c_p) := \hat{\epsilon}_\theta(x_t) + s_g (\hat{\epsilon}_\theta(x_t, c_p) - \hat{\epsilon}_\theta(x_t)),\tag{5}$$

with s_g balancing conditional and unconditional objectives, enabling high-quality, text-guided image generation.

4 method

The pipeline of our proposed method is shown in Figure 3. The core idea of Semantic-Consistent Unrestricted Adversarial Attack is to enhance semantic control throughout the entire generation process of unrestricted adversarial examples, which is achieved by introducing a novel inversion method that can “imprint” the image more strongly onto the noise maps. Furthermore, the powerful semantic guidance provided by MLLMs further restricts the direction of perturbations in the latent space, enabling the clean image to produce our desired adversarial changes, namely imperceptibility, and naturalness.

Specifically, our method can be divided into two parts: **Semantic Fixation Inversion** and **Semantically Guided Perturbation**. We first map the clean image into a latent space through Semantic Fixation Inversion, and then iteratively optimize the adversarial objective in the latent space under semantic guidance, causing the content of the image to shift in the direction of deceiving the model until the attack is successful. The algorithm of SCA is presented in Algorithm 1.

4.1 Problem Definition.

For a deep learning classifier $\mathcal{F}_\theta(\cdot)$ with parameters θ , let x represent the clean image and y the corresponding ground truth label. Unrestricted adversarial attacks seek to introduce subtle adversarial perturbations (e.g., image distortions, texture, or color alterations, etc.) to the input x , resulting in an adversarial example x_{adv} that misguides the classifier $\mathcal{F}_\theta(\cdot)$:

$$\max_{x_{adv}} \mathcal{L}(\mathcal{F}_\theta(x_{adv}), y), \quad s.t. \ d(x, x_{adv}) < \epsilon,\tag{6}$$

where $\mathcal{L}(\cdot)$ is the loss function and $d(x, x_{adv})$ represents the semantic distance between x and x_{adv} . Based on the above, we believe that ideal unrestricted adversarial examples should have high semantic consistency.

4.2 Semantic Fixation Inversion

In the context of Semantic Fixation Inversion, the key objective is to extract rich and semantically relevant information from the original image during the inversion process. This semantic information acts as an informative prior, guiding both the perturbation of the latent space and the image reconstruction, facilitating a Semantic Consistency Attack. In our method, the semantic priors are derived from edit-friendly noise maps and the powerful capabilities of Multimodal Language Models (MLLMs). Recent research predominantly focuses on inverting the diffusion sampling process to determine x_T , which can be subsequently denoised into the input image x_0 . Among the inversion techniques, inverting the DDPM (Denoising Diffusion Probabilistic Models) scheduler is often favored over DDIM (Denoising Diffusion Implicit Models) inversion, as the former can achieve results in fewer steps while maintaining perfect reconstruction [22].

Although the DDPM scheduler provides efficient inversion, more advanced approaches exist for sampling in Diffusion Models (DMs), significantly reducing the number of steps and, consequently, the number of DM evaluations. In our work, we adopt a more efficient inversion approach [4], which retains the essential inversion properties while accelerating the process. Previous work by Song et al. [49] has shown that the reverse diffusion process in DDPM can be viewed as a first-order stochastic

differential equation (SDE). Solving this SDE with higher-order methods, such as the dpm-solver++ [37], enables more efficient computation and reduces the number of required steps.

Leveraging the semantic understanding capability of MLLMs, which have been shown to provide accurate semantic interpretations across diverse image domains, we employ MLLMs to generate rich captions $\mathcal{P}$ for the clean input image x_0 . These captions are then converted into conditional embeddings c_p via a function $\psi(\mathcal{P})$, serving as the conditional input for the pre-trained diffusion model parameterized by $\hat{\epsilon}_\theta$. This process integrates seamlessly with the existing formulation for conditional noise estimation $\hat{\epsilon}_\theta(x_t, c_p)$ as given in Equation (5).

The reverse diffusion process, which uses the second-order dpm-solver++ SDE solver, can be written as:

$$x_{t-1} = \hat{\mu}_t(x_t, x_{t+1}, c_p) + \sigma_t z_t, \quad t = T, \dots, 1. \quad (7)$$

In this expression, the variance term σ_t is now defined as:

$$\sigma_t = \sqrt{1 - \bar{\alpha}_{t-1}} \sqrt{1 - e^{-2h_{t-1}}}. \quad (8)$$

This updated variance scheduling leads to improved noise control, and the mean prediction $\hat{\mu}_t(x_t, x_{t+1}, c_p)$ depends on information from both the current timestep x_t and the subsequent one x_{t+1} , making the process higher-order. Specifically, $\hat{\mu}_t(x_t, x_{t+1}, c_p)$ is given by:

$$\begin{aligned} & \frac{\sqrt{1 - \bar{\alpha}_{t-1}}}{\sqrt{1 - \bar{\alpha}_t}} e^{-h_{t-1}} x_t + \sqrt{\bar{\alpha}_{t-1}} (1 - e^{-2h_{t-1}}) \hat{\epsilon}_\theta(x_t, c_p) \\ & + 0.5 \sqrt{\bar{\alpha}_{t-1}} (1 - e^{-2h_{t-1}}) \frac{-h_t}{h_{t-1}} (\hat{\epsilon}_\theta(x_{t+1}, c_p) - \hat{\epsilon}_\theta(x_t, c_p)). \end{aligned} \quad (9)$$

The term h_t is a function of the variance schedule and is given by:

$$h_t = \frac{\ln(\sqrt{\bar{\alpha}_t})}{\ln(\sqrt{1 - \bar{\alpha}_t})} - \frac{\ln(\sqrt{\bar{\alpha}_{t+1}})}{\ln(\sqrt{1 - \bar{\alpha}_{t+1}})}. \quad (10)$$

This higher-order solver significantly improves efficiency by using information from two adjacent timesteps, allowing for faster convergence without sacrificing reconstruction quality. For more comprehensive information on the solver’s derivation and its convergence behavior, we direct readers to the pertinent literature [37].

Based on the inversion process, we generate an auxiliary sequence of noise images beginning with the input image x_0 , which can be formulated as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \tilde{\epsilon}_t, \quad (11)$$

where $\tilde{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ are statistically independent noise vectors. In contrast to the correlated noise in Equation (2), the independence of $\tilde{\epsilon}_t$ is a desirable property for ensuring semantic consistency during the reconstruction process [22].

Finally, the noise term z_t required for the inversion process is derived from Equation (7) as:

$$z_t = \frac{x_{t-1} - \hat{\mu}_t(x_t, x_{t+1}, c_p)}{\sigma_t}, \quad t = T, \dots, 1, \quad (12)$$

where $\hat{\mu}_t$ and σ_t are as defined earlier.

In conclusion, we have successfully derived a series of edit-friendly noise maps $\{x_T, z_1, z_2, \dots, z_T\}$, which can be progressively denoised to reconstruct the original image using the reverse diffusion process outlined in Equation (7). The statistical independence of these noise maps, combined with the rich semantic priors provided by MLLMs, ensures that the reconstruction process maintains semantic consistency without unnecessary perturbations to the latent space. The subsequent experimental section provides a detailed analysis of the reconstruction quality and compares our results with existing methods. Additionally, the extracted semantic knowledge is employed to guide Semantically Guided Perturbation, enabling further manipulation of the generated images.

4.3 Semantically Guided Perturbation

In this section, we introduce an optimization technique for latents aimed at enhancing the attack performance on unrestricted adversarial examples. Drawing from previous work [10], perturbing the text embedding c_p to generate adversarial content is suboptimal because it is crucial for semantic consistency, preserving a wealth of semantic information. Additionally, due to the discrete nature of textual representations, current techniques struggle to generate smooth and diverse perturbations based on gradient methods. Consistent with state-of-the-art approaches [10, 64, 63], we opt to optimize the adversarial objective within the latent space. Building upon Semantic Fixation Inversion, we utilize rich semantic priors to guide the optimization process. We define the denoising process of diffusion models as $\mathcal{D}(\cdot)$ through Equation (13), involving T iterations:

$$\begin{aligned} \hat{x}_0 &= \hat{\mu}_1(\hat{x}_1, \hat{x}_2, c_p) + \sigma_1 z_1 \\ &= \hat{\mu}_1(\hat{\mu}_2(\hat{x}_2, \hat{x}_3, c_p) + \sigma_2 z_2, \hat{\mu}_3(\hat{x}_3, \hat{x}_4, c_p) + \sigma_3 z_3, c_p) \\ &\quad + \sigma_1 z_1 \\ &= \dots\dots\dots \\ &= \mathcal{D}(x_T, z_1, z_2, \dots, z_T, c_p), \end{aligned} \tag{13}$$

with

$$x_{t-1} = \hat{\mu}_t(\hat{x}_t, x_{t+1}, c_p) + \sigma_t z_t, \quad t = T, \dots, 1. \tag{14}$$

Therefore, the reconstructed image is denoted as $\hat{z}_0 = \mathcal{D}(x_T, \{z_t\}_{t=1}^T, c_p, T)$. The computational procedure of the VAE is omitted in this context, as it remains differentiable. Our adversarial objective is expressed by:

$$\begin{aligned} &\max_{\delta} \mathcal{L}(\mathcal{F}_{\theta}(\hat{x}_0, y), \\ \text{s.t. } &\|\delta\|_{\infty} \leq \kappa, \hat{x}_0 = \mathcal{D}(x_T, \{z_t\}_{t=1}^T, c_p, T) \\ &\text{and } d(x, x_{adv}) < \epsilon, \end{aligned} \tag{15}$$

where δ is the adversarial perturbation on the latent space and d represents the semantic distance. Previous works [10, 63] typically design the loss function $\mathcal{L}$ as comprising two components: i) the cross-entropy loss $\mathcal{L}_{ce}$, which primarily guides adversarial examples towards misclassification, and ii) the mean square error loss $\mathcal{L}_{mse}$, which mainly ensures that the generated adversarial examples are as close as possible to the clean image in terms of l_2 distance. However, extensive experiments have confirmed that the additional introduction of $\mathcal{L}_{mse}$ in the loss function has a very limited role in semantic preservation [10], perhaps because pixel-level numerical changes are not a high-level semantic constraint. Furthermore, the introduction of two optimization objectives increases the difficulty of optimization and limits the capability of the attack. Specific experimental results regarding this component are presented in our ablation study. Since good semantic consistency has already been achieved during the Semantic Fixation Inversion phase, we let $\mathcal{L} = \mathcal{L}_{ce}$.

Existing methods [10, 63, 64] generally assume that a perturbation δ will not disrupt semantic consistency when it is small, hence imposing a norm constraint on δ . However, we find that even small perturbations in the latent space can cause significant changes in the final image. Because the inversion process in these methods is approximate, it is not easy to precisely find the latent representation corresponding to the clean image. According to Equation (13), our method avoids the accumulation of approximation errors in the optimization process by guiding the image-denoising process with the rich semantic knowledge excavated during the inversion phase. Analogous to conventional adversarial attacks [14], we employ gradient-based techniques to estimate δ through $\delta \simeq \eta \nabla_{x_T} \mathcal{L}(\mathcal{F}_{\theta}(\hat{x}_0), y)$, where η denotes the magnitude of perturbations that occur in the direction of the gradient. To expand $\nabla_{x_T} \mathcal{L}(\mathcal{F}_{\theta}(\hat{x}_0), y)$ by the chain rule, we can obtain:

$$\nabla_{x_T} \mathcal{L}(\mathcal{F}_{\theta}(\hat{x}_0), y) = \frac{\partial \mathcal{L}}{\partial \hat{x}_0} \frac{\partial \hat{x}_0}{\partial x_T} = \frac{\partial \mathcal{L}}{\partial \mathcal{D}} \frac{\partial \mathcal{D}}{\partial x_T}. \tag{16}$$

Gradient Estimation. However, we cannot directly compute the gradients for optimization based on Equation (13) and Equation (15), as this is excessively complex and leads to memory overflow (similar phenomena are also observed in [10, 46]). Previous works [10, 63] have utilized the Skip Gradient technique to estimate this gradient, specifically by rearranging Equation (2) to derive

Table 1: Image semantic consistency assessment

Attack	CLIP Score $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$
SAE[19]	0.7947	0.9492	15.6864	0.3767
ADef[1]	0.7856	0.9112	14.7823	0.3996
cAdv[2]	0.8398	0.9924	19.9874	0.2695
tAdv[2]	0.8366	0.9876	19.0012	0.2377
ACE[69]	0.8318	0.9891	19.7205	0.1799
ColorFool[47]	0.8448	0.9358	9.7676	0.5102
NCF[66]	0.8848	0.8135	14.3670	0.3194
AdvST[57]	0.8752	0.8144	14.2011	0.2908
ACA[10]	0.7442	0.5137	11.0014	0.5971
SCA(Ours)	0.8896	0.8162	20.5825	0.1360

$x_0 = \frac{1}{\sqrt{\alpha_t}}x_T - \sqrt{1 - \alpha_t}\epsilon$. Consequently, they further obtain $\frac{\partial x_0}{\partial x_T} = \frac{1}{\sqrt{\alpha_t}}$. However, this estimation method is overly simplistic and introduces significant errors. Additionally, our method differs from DDIM sampling as it is not an approximate method, especially given that we use fewer timesteps, which further magnifies the error.

To estimate the gradients, we employ the random gradient-free (RGF) method [41]. First, we rewrite a gradient as the expectation of direction derivatives, i.e., $\nabla_{\mathbf{x}}F(\mathbf{x}) = \mathbb{E}[\boldsymbol{\delta}^\top \nabla_{\mathbf{x}}F(\mathbf{x}) \cdot \boldsymbol{\delta}]$, where $F(\mathbf{x})$ represents any differentiable function and $\boldsymbol{\delta} \sim P(\boldsymbol{\delta})$ is a random variable satisfying that $\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top] = \mathbf{I}$ (e.g., $\boldsymbol{\delta}$ can be uniformly sampled from a hypersphere). Then by zero-order optimization [8, 68], we know that

$$\begin{aligned}\nabla_{x_T}\mathcal{L}(\mathcal{F}_\theta(x_0), y) &= \frac{\partial \mathcal{L}}{\partial \mathcal{D}} \frac{\partial \mathcal{D}}{\partial x_T} \\ &\approx \frac{\partial \mathcal{L}}{\partial \mathcal{D}} \frac{1}{N\sigma} \sum_{n=1}^N [\mathcal{D}(x_T + \sigma \boldsymbol{\delta}_n) - \mathcal{D}(x_T)] \cdot \boldsymbol{\delta}_n,\end{aligned}\tag{17}$$

where $\boldsymbol{\delta}_n \sim P(\boldsymbol{\delta})$, σ is a hyperparameter controls the sampling variance, and N is the number of queries. The approximation in Equation (17) becomes an unbiased equation when $\sigma \rightarrow 0$ and $N \rightarrow \infty$.

Differentiable Boundary Processing. In diffusion models, generated images $\hat{x}_0$ may exceed the valid pixel range. To prevent this, we apply the boundary processing function $\varrho(\cdot)$ from [10], which clamps pixel values within $[0, 1]$, ensuring natural image integrity.

We also introduce the projection operator Π_κ to constrain adversarial perturbations $\boldsymbol{\delta}$ within a κ -ball, balancing perturbation effectiveness and imperceptibility. Inspired by prior adversarial attack methods [10, 14, 32, 62, 63], we incorporate momentum for stable optimization. Momentum g is updated iteratively as follows:

$$g_k \leftarrow \mu \cdot g_{k-1} + \frac{\nabla_{x_T}\mathcal{L}(\mathcal{F}_\theta(\varrho(\hat{x}_0), y))}{\|\nabla_{x_T}\mathcal{L}(\mathcal{F}_\theta(\varrho(\hat{x}_0), y))\|_1},\tag{18}$$

where μ is the momentum coefficient, $\mathcal{L}$ is the loss function, and $\mathcal{F}_\theta$ represents the neural network. The adversarial perturbation $\boldsymbol{\delta}_k$ is updated as:

$$\boldsymbol{\delta}_k \leftarrow \Pi_\kappa(\boldsymbol{\delta}_{k-1} + \eta \cdot \text{sign}(g_k)),\tag{19}$$

with η as the step size, and Π_κ ensuring the perturbation stays within the κ -ball. This method, leveraging momentum and projection, has been shown to enhance optimization in adversarial attacks while maintaining naturalistic image appearance [10, 14, 32, 62, 63].

Table 2: Performance comparison on normally trained CNNs and ViTs. We report attack success rates (%) of each method (“*” means white-box attack results).

Surrogate	Attack	Target Models										Avg. ASR(%)
		CNNs						Transformers				
		MN-v2	Inc-v3	RN-50	Dense-161	RN-152	EF-b7	MobViT-s	ViT-B	Swin-b	PVT-v2	
-	Clean	12.1	4.8	7.0	6.3	5.6	8.7	7.8	8.9	3.5	3.6	6.83
MobViT-s	SAE	60.2	21.2	54.6	42.7	44.9	30.2	82.5*	38.6	21.1	20.2	37.08
	ADef	14.5	6.6	9.0	8.0	7.1	9.8	80.8*	9.7	5.1	4.6	8.27
	cAdv	41.9	25.4	33.2	31.2	28.2	34.7	84.3*	32.6	22.7	22.0	30.21
	tAdv	33.6	18.8	22.1	18.7	18.7	15.8	97.4*	15.3	11.2	13.7	18.66
	ACE	30.7	9.7	20.3	16.3	14.4	13.8	99.2*	16.5	6.8	5.8	14.92
	ColorFool	47.1	12.0	40.0	28.1	30.7	19.3	81.7*	24.3	9.7	10.0	24.58
	NCF	67.7	31.2	60.3	41.8	52.2	32.2	74.5*	39.1	20.8	23.1	40.93
	AdvST	67.2	44.0	61.1	46.8	52.4	33.1	82.3*	39.8	21.5	24.0	47.22
	ACA	66.2	56.6	60.6	58.1	55.9	55.5	89.8*	51.4	52.7	55.1	56.90
	SCA(Ours)	65.8	58.9	61.2	59.0	54.7	58.2	91.4*	52.1	51.9	55.6	60.88
MN-v2	SAE	90.8*	22.5	53.2	38.0	41.9	26.9	44.6	33.6	16.8	18.3	32.87
	ADef	56.6*	7.6	8.4	7.7	7.1	10.9	11.7	9.5	4.5	4.5	7.99
	cAdv	96.6*	26.8	39.6	33.9	29.9	32.7	41.9	33.1	20.6	19.7	30.91
	tAdv	99.9*	27.2	31.5	24.3	24.5	22.4	40.5	16.1	15.9	15.1	24.17
	ACE	99.1*	9.5	17.9	12.4	12.6	11.7	16.3	12.1	5.4	5.6	11.50
	ColorFool	93.3*	9.5	25.7	15.3	15.4	13.4	15.7	14.2	5.9	6.4	13.50
	NCF	93.2*	33.6	65.9	43.5	56.3	33.0	52.6	35.8	21.2	20.6	40.28
	AdvST	97.0*	35.4	57.2	48.5	50.5	32.4	52.6	36.8	19.8	21.7	45.19
	ACA	93.1*	56.8	62.6	55.7	56.0	51.0	59.6	48.7	48.6	50.4	54.38
	SCA(Ours)	93.0*	56.9	63.3	53.2	56.0	51.0	62.3	50.1	47.0	51.0	58.38
RN-50	SAE	63.2	25.9	88.0*	41.9	46.5	28.8	45.9	35.3	20.3	19.6	36.38
	ADef	15.5	7.7	55.7*	8.4	7.8	11.4	12.3	9.2	4.6	4.9	9.09
	cAdv	44.2	25.3	97.2*	36.8	37.0	34.9	40.1	30.6	19.3	20.2	32.04
	tAdv	43.4	27.0	99.0*	28.8	30.2	21.6	35.9	16.5	15.2	15.1	25.97
	ACE	32.8	9.4	99.1*	16.1	15.2	12.7	20.5	13.1	6.1	5.3	14.58
	ColorFool	41.6	9.8	90.1*	18.6	21.0	15.4	20.4	15.4	5.9	6.8	17.21
	NCF	71.2	33.6	91.4*	48.5	60.5	32.4	52.6	36.8	19.8	21.7	41.90
	AdvST	55.9	38.2	90.1*	50.5	57.2	41.3	52.3	32.8	20.1	24.7	46.31
	ACA	69.3	61.6	88.3*	61.9	61.7	60.3	62.6	52.9	51.9	53.2	59.49
	SCA(Ours)	69.5	60.8	94.5*	59.9	62.0	60.6	61.9	53.2	52.0	54.1	62.55
ViT-B	SAE	54.5	26.9	49.7	38.4	41.4	30.4	46.1	78.4*	19.9	18.1	36.16
	ADef	15.3	8.3	9.9	8.4	7.6	12.0	12.4	81.5*	5.3	5.5	9.41
	cAdv	31.4	27.0	26.1	22.5	19.9	26.1	32.9	96.5*	18.4	16.9	24.58
	tAdv	39.5	22.8	25.8	23.2	22.3	20.8	34.1	93.5*	16.3	15.3	24.46
	ACE	30.9	11.4	22.0	15.5	15.2	13.0	17.0	98.6*	6.5	6.3	15.31
	ColorFool	45.3	13.9	35.7	24.3	28.8	19.8	27.0	83.1*	8.9	9.3	23.67
	NCF	55.9	25.3	50.6	34.8	42.3	29.9	40.6	81.0*	20.0	19.1	35.39
	AdvST	53.9	28.6	52.1	38.5	40.5	32.4	42.2	85.1*	19.8	21.7	41.48
	ACA	64.6	58.8	60.2	58.1	58.1	57.1	60.8	87.7*	55.5	54.9	58.68
	SCA(Ours)	64.1	59.0	60.0	57.9	58.0	57.0	60.6	89.0*	58.1	54.0	61.77

5 Experiments

5.1 Experimental Setup

Datasets. We carry out our experiments using the ImageNet-compatible dataset [25], which includes 1,000 images from the ImageNet validation set [12]. This dataset has been frequently employed in related work [15, 16, 62, 66, 10].

Semantic Consistency Evaluation. We use CLIP Score [44], SSIM [59], PSNR [23], and LPIPS [67] to measure the semantic consistency between the generated unrestricted adversarial examples and the clean image from multiple aspects.

Attack Evaluation. We compare SCA with other unrestricted adversarial attacks, including SAE [19], ADef [1], cAdv [2], tAdv [2], ACE [69], ColorFool [47], NCF [66], AdvST [57], and ACA [10]. The default parameter settings for these attacks are used. The evaluation metric employed is attack success rate (ASR), which measures the proportion of images misclassified by the target models.

Target Models. To assess the adversarial robustness of various architectures, we target both convolutional neural networks (CNNs) and vision transformers (ViTs). The CNN models include MobileNet-V2 (MN-v2), Inception-v3 (Inc-v3), ResNet-50 (RN-50), ResNet-152 (RN-152), DenseNet-161 (Dense-161), and EfficientNet-b7 (EFb7). For ViTs, we evaluate MobileViT (MobViT-s), Vision Transformer (ViT-B), Swin Transformer (Swin-B), and Pyramid Vision Transformer (PVT-v2).

Implementation Details. All experiments are performed using PyTorch on an NVIDIA Tesla A100. The DDPM steps are set to $T = 10 \rightarrow 20$, with attack iterations $N_a = 10$, $\eta = 0.04$, $\kappa = 0.1$,

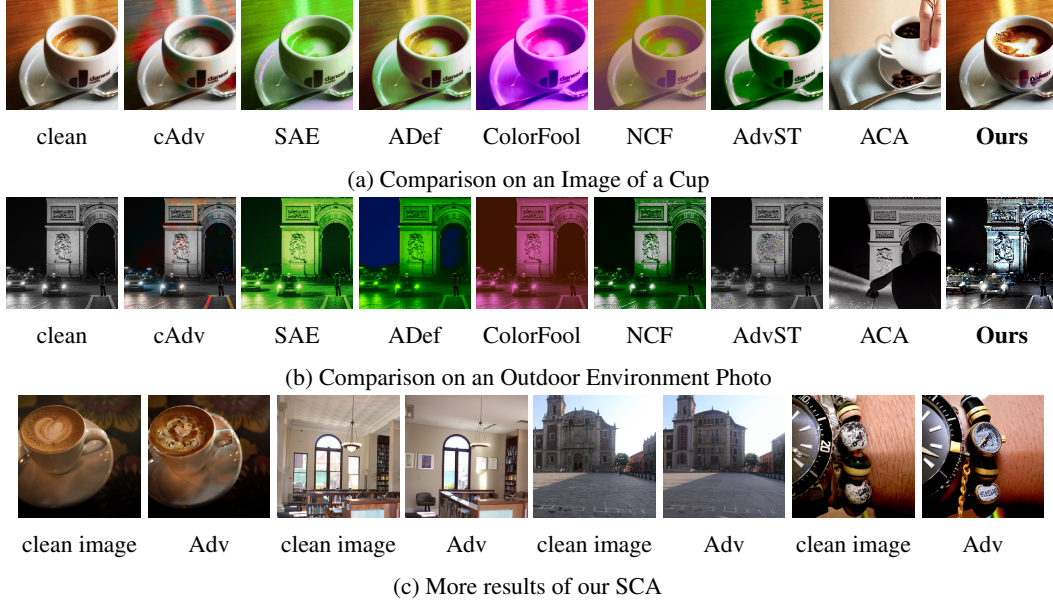


Figure 4: Compared with other attacks, SCA generates the most natural adversarial examples and maintains a high degree of semantic consistency with the clean image.

and $\mu = 1$. The version of Stable Diffusion [45] used is v1.5, and image captions are generated automatically via LLaVA-NeXT.

5.2 Semantic Consistency Comparison

To comprehensively evaluate the semantic consistency preservation ability of current unrestricted adversarial attack methods from multiple perspectives, we introduce four indicators, namely CLIP Score, SSIM, PSNR, and MSE. The specific results are shown in Table 1. PSNR (Peak Signal to Noise Ratio) is a metric for evaluating image quality. The unit of PSNR is dB. The larger the value, the less image distortion. However, many experimental results show that the PSNR score cannot be completely consistent with the visual quality seen by the human eye. A higher PSNR image may look worse than a lower PSNR image. This is because the sensitivity of the human eye to errors is not absolute, and its perception results will vary due to many factors. To address the limitations of PSNR, we introduced LPIPS and SSIM. The range of SSIM values is $[0,1]$, and the larger the value, the more similar the images are. If two images are the same, the SSIM value is 1. LPIPS (Learned Perceptual Image Patch Similarity), also known as perceptual loss, is used to measure the difference between two images. This metric learns the reverse mapping from the generated image to the ground truth, forces the generator to learn the reverse mapping of reconstructing the real image from the fake image, and prioritizes the perceptual similarity between them. LPIPS is more in line with human perception than traditional methods. The lower the LPIPS value, the more similar the two images are, and vice versa, the greater the difference.

In addition, we also introduced CLIP Score as a more advanced semantic evaluation metric. As we all know, CLIP has been trained on more than 400 million image-text pairs and has strong understanding capabilities. So we measure semantic consistency by calculating the cosine similarity between the image embeddings processed by CLIP’s visual encoder.

In summary, the experimental results show that the adversarial examples we generate outperform existing methods in most indicators, especially in LPIPS and CLIP Score. This further proves the superiority of our SCA in maintaining semantic consistency.

Table 3: Image quality assessment

Attack	NIMA -AVA	Hyper -IQA	MUSIQ -AVA	MUSIQ -KonIQ	TReS
Clean	5.15	0.667	4.07	52.66	82.01
SAE	5.05	0.597	3.79	47.24	71.88
ADef	4.89	0.608	3.89	47.39	72.10
cAdv	4.97	0.623	3.87	48.32	75.12
tAdv	4.83	0.525	3.78	44.71	67.07
ACE	5.12	0.648	3.96	50.49	77.25
ColorFool	5.24	0.662	4.05	52.27	78.54
NCF	4.96	0.634	3.87	50.33	74.10
AdvST	5.12	0.679	3.99	56.33	79.21
ACA	5.54	0.691	4.37	56.08	85.11
SCA(Ours)	5.32	0.691	4.47	58.62	82.05

5.3 Comparison of Attack Success Rates

We systematically evaluate the adversarial robustness of convolutional neural networks (CNNs) and vision transformers (ViTs) using several prominent adversarial methods: SAE [19], ADef [1], cAdv and tAdv [2], ACE [69], ColorFool [47], NCF [66], AdvST [57], ACA [10], and our proposed SCA. These methods were applied to MobileNetV2 (MN-v2), ResNet-50 (RN-50), MobViT-s, and ViT-B, with performance measured using the Average Attack Success Rate (%), which captures the transferability of adversarial examples across non-substitute models.

Table 2 provides a detailed comparison of these methods in transferring adversarial examples between CNNs and ViTs. Our SCA consistently achieves high attack success rates, often surpassing or equaling ACA, a leading method in the field. Notably, SCA demonstrates slight but meaningful improvements over ACA when using MN-v2 as a surrogate model, particularly on Inception-v3 (Inc-v3), RN-50, and MobViT-s, emphasizing SCA’s superior adversarial transferability.

Our results confirm that SCA achieves state-of-the-art performance by enhancing the imperceptibility and stealthiness of adversarial examples through semantic consistency. Crucially, this semantic alignment does not compromise attack efficacy; instead, it enhances transferability without sacrificing performance.

Both SCA and ACA show strong transferability across architectures, particularly in the challenging case of transferring adversarial examples from CNNs to ViTs. This is due to two factors: (i) both methods perform a low-dimensional manifold search within the natural image space, inherently improving cross-architecture transferability, and (ii) both leverage diffusion models with self-attention mechanisms that promote architectural similarities. With an average ASR exceeding 60%, SCA not only advances the state of the art but also provides an effective approach for generating transferable adversarial examples across diverse model architectures, demonstrating its robustness and efficiency.

5.4 Image Quality Comparison.

In this section, we conduct a detailed quantitative assessment of the image quality for adversarial examples, following the approach taken in prior works such as [47], [66], and [10]. Specifically, we employ non-reference perceptual image quality metrics to evaluate the visual realism and aesthetic appeal of the adversarial images generated by different methods. The metrics we use include NIMA [52], HyperIQA [50], MUSIQ [24], and TReS [17], which are widely recognized for their ability to assess image quality in the absence of a ground truth reference.

To provide a robust evaluation, we follow the PyIQA framework and apply models trained on large-scale aesthetic datasets. NIMA-AVA and MUSIQ-AVA are trained on the AVA [40] dataset, while MUSIQ-KonIQ is trained on the KonIQ-10K [20] dataset, which ensures a comprehensive evaluation of the generated adversarial examples. As shown in Table 3, our method consistently achieves image quality scores that are comparable to ACA, outperforming other methods such as ColorFool and NCF. This superior performance is likely attributable to the fact that both SCA and ACA leverage Stable

Table 4: Attack speed of unrestricted attacks. We choose MN-v2 as the surrogate model and evaluate the inference time on an NVIDIA Tesla A100.

Attack	SAE	ADef	cAdv	tAdv	ACE	ColorFool	NCF	AdvST	ACA	Ours(SCA)
Time(Sec)	8.80	0.41	18.67	4.88	6.64	12.18	10.45	19.25	125.33	9.85

Table 5: Ablation study of random gradient-free method.

Methods	Models					Avg. ASR
	Dense161	RN-152	EF-b7	ViT-B	PVT-v2	
Clean Image	6.3	5.6	7.7	8.9*	3.6	6.62
NE	7.0	6.2	9.1	12.3*	5.1	7.94
SG	24.3	22.3	27.5	59.6*	19.7	30.68
RGF	57.9	58.0	57.0	89.0*	54.0	63.18

Diffusion, a state-of-the-art generative model known for its ability to produce highly realistic and visually appealing images.

The use of Stable Diffusion plays a critical role in enhancing the naturalness and aesthetic quality of our adversarial examples. The inherent generative power of this model allows us to craft perturbations that blend seamlessly into the clean image, preserving its photorealism while embedding subtle adversarial signals. Furthermore, we map the clean image to a latent space using an inversion method, which effectively places the image within a low-dimensional manifold representing the distribution of natural images. By perturbing the image within this manifold, our method ensures that the adversarial examples remain visually coherent and natural, aligning closely with the underlying distribution of real-world images.

Additionally, it is worth noting that the no-reference image quality metrics we employ are often trained on aesthetic datasets, many of which contain post-processed images, including those edited with tools such as Photoshop. These post-processed images are more aligned with human aesthetic preferences, which is an important consideration for assessing adversarial examples. As [10] points out, the similarity of our method to post-processing and image-editing techniques helps explain why our adversarial examples receive higher quality scores from these metrics. Our approach effectively functions as a form of image editing, subtly modifying the image in ways that retain or even enhance its aesthetic appeal, leading to better evaluation results.

5.5 Qualitative Comparison

In this section, we present a qualitative analysis of various current unrestricted attack methods, illustrated in Figure 4. Our findings reveal that the SCA generates adversarial examples that exhibit a greater degree of naturalness compared to methods that rely heavily on color and texture transformations. Notably, techniques such as SAE and ColorFool often produce adversarial examples characterized by unnatural color distortions, making them easily detectable to human observers. This perceptibility undermines their effectiveness as adversarial attacks.

In contrast, the state-of-the-art methods ACA and AdvST also achieve a commendable level of naturalness in their generated adversarial examples. However, when it comes to preserving semantic consistency with clean images, our method stands out. For instance, as depicted in Figure 4(a), the adversarial image produced by our approach only alters the pattern of the coffee cup while retaining the overall environmental context of the scene. This subtle manipulation contrasts sharply with ACA, which, while maintaining the basic concept of the coffee cup, leads to significant alterations in both the cup’s shape and the surrounding environment. Such transformations diminish the connection to the original clean image, resulting in a lack of semantic fidelity.

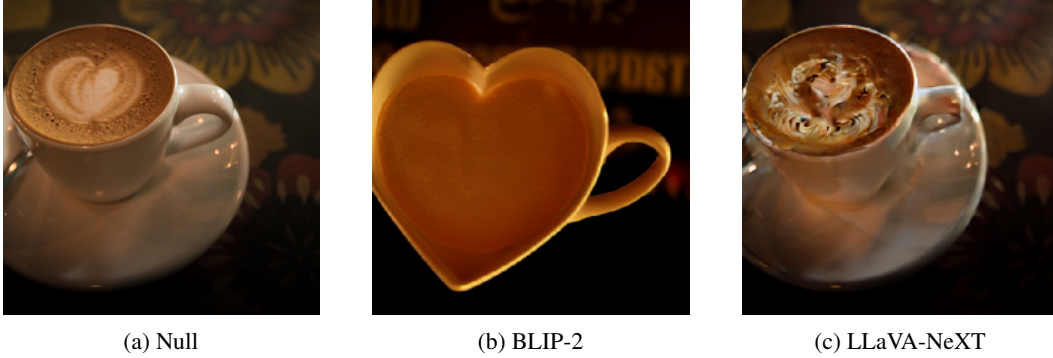


Figure 5: Ablation study of the degree of semantic restriction. The caption of (b) is “a coffee cup with a heart-shaped design in the foam”. The caption of (c) is “At the center of the frame is a white cup filled with a dark brown liquid, possibly coffee. The cup is placed on a white saucer, which is accompanied by a silver spoon. All these items rest on a tablecloth featuring a colorful floral pattern. On the cup, there’s a heart-shaped latte art design, suggesting a level of care and attention given to the preparation of the drink.”

5.6 Time Analysis

In this section, we present a comparison of the attack speeds for various unrestricted attacks. Using MN-v2 as the surrogate model, we measure the inference time on an NVIDIA Tesla A100. Table 4 shows the average time (in seconds) required to generate an adversarial example per image. Compared with the current state-of-the-art methods, we have increased the speed by about 12 times while maintaining the diversity of generated content and the success rate of attacks. This is mainly due to our novel inversion method, and the use of DPM Solver++ greatly accelerates the Diffusion sampling process. Experiments have shown that we can reduce the number of sampling steps to 10 to 20 steps without affecting the quality of the generated images. From Table 4, it can be seen that some methods take a very short time, but experiments have shown that their effects are generally poor. Overall, our method is the most efficient.

5.7 Ablation Study

In the process of computing the gradients for optimization based on the gradient optimization process using Equation (13) and Equation (15), we apply the random gradient-free (RGF) method. To evaluate its effectiveness, we conducted ablation tests under three conditions: without gradient estimation (NE), using the Skip Gradient method (SG), and using RGF. Table 5 shows the results. Our SCA performs poorly without gradient estimation and with the Skip Gradient method. The ASR for “No estimation” is similar to the clean image, showing that we can’t ignore the diffusion model’s sampling and directly optimize with the classifier’s loss in the latent space. The Skip Gradient method also underperforms, likely due to the need for more sample steps, as we use complex optimizations to speed up sampling. Additionally, our new inversion method differs from prior approximations, validating the importance of the RGF method in our framework.

We also tested the impact of using MLLM models for semantic consistency. In this experiment, we replaced LLaVA-NexT with BLIP-2 for generating captions during perturbation and also tried unconditional generation without prompts. As shown in Figure 5, BLIP-2 produces very brief descriptions, leading to a significant semantic drift (Figure 5(b)). In contrast, LLaVA-NexT generates detailed captions, maintaining better semantic alignment with the clean image (Figure 5(c)). Additional experiments confirm that richer captions improve semantic preservation. In the case of unconditional generation (Figure 5(a)), the output remains nearly identical to the original, failing to produce adversarial perturbations, likely due to overly strong semantic constraints. This highlights the importance of balancing semantic restriction, an area that warrants further research.

6 Conclusion

In this paper, We propose a novel attack framework called Semantic-Consistent Unrestricted Adversarial Attack (SCA) via Semantic Fixation Inversion and Semantically Guided Perturbation. The core idea of SCA is to enhance semantic control throughout the entire generation process of unrestricted adversarial examples. We initially utilized an effective inversion method and a powerful MLLM to extract rich semantic priors from a clean image. Subsequently, we optimize the adversarial objective within the latent space under the guidance of these priors. Experiments demonstrate that the adversarial examples exhibit a high degree of semantic consistency compared to existing methods. Furthermore, our method is highly efficient. Consequently, we introduce Semantic-Consistent Adversarial Examples (SCAE). Our work can shed light on further understanding the vulnerabilities of DNNs as well as novel defense approaches.

References

- [1] Rima Alaifari, Giovanni S Albeti, and Tandri Gauksson. Adef: an iterative algorithm to construct adversarial deformations. *arXiv preprint arXiv:1804.07729*, 2018.
- [2] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019.
- [3] Benoit Bonnet, Teddy Furon, and Patrick Bas. Generating adversarial images in quantized domains. *IEEE Transactions on Information Forensics and Security*, 17:373–385, 2022.
- [4] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024.
- [5] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.
- [6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017.
- [7] Chen Chen, Xinwei Zhao, and Matthew C. Stamm. Generative adversarial attacks against deep-learning-based camera model identification. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2019.
- [8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.
- [9] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023.
- [10] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36, 2024.
- [11] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. *arXiv preprint arXiv:2307.12499*, 2023.
- [12] Jia Deng, Alex Berg, Sanjeev Satheesh, H Su, Aditya Khosla, and Li Fei-Fei. Imagenet large scale visual recognition competition 2012 (ilsvrc2012). *See net.org/challenges/LSVRC*, 41:6, 2012.
- [13] Feng Ding, Guopu Zhu, Yingcan Li, Xinpeng Zhang, Pradeep K. Atrey, and Siwei Lyu. Anti-forensics for face swapping videos via adversarial training. *IEEE Transactions on Multimedia*, 24:3429–3441, 2022.
- [14] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

- [15] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019.
- [16] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII* 16, pages 307–322. Springer, 2020.
- [17] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1220–1230, 2022.
- [18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [19] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.
- [20] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.
- [21] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021.
- [22] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpn noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.
- [23] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.
- [24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.
- [25] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.
- [26] Hyun Kwon, Yongchul Kim, Hyunsoo Yoon, and Daeseon Choi. Selective audio adversarial example in evasion attack on speech recognition system. *IEEE Transactions on Information Forensics and Security*, 15:526–538, 2020.
- [27] Raz Lapid, Eylon Mizrahi, and Moshe Sipper. Patch of invisibility: Naturalistic physical black-box adversarial attacks on object detectors. *arXiv preprint arXiv:2303.04238*, 2023.
- [28] Chun Pong Lau, Jiang Liu, and Rama Chellappa. Attribute-guided encryption with facial texture masking. *arXiv preprint arXiv:2305.13548*, 2023.
- [29] Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Decision-based adversarial attack with frequency mixup. *IEEE Transactions on Information Forensics and Security*, 17:1038–1052, 2022.
- [30] Zhankai Li, Weiping Wang, Jie Li, Kai Chen, and Shigeng Zhang. Ucg: A universal cross-domain generator for transferable adversarial examples. *IEEE Transactions on Information Forensics and Security*, 19:3023–3037, 2024.
- [31] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 18(1):72–85, 2021.
- [32] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2019.
- [33] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *Advances in Neural Information Processing Systems*, 33:3487–3498, 2020.

- [34] Fangcheng Liu, Chao Zhang, and Hongyang Zhang. Towards transferable unrestricted adversarial examples with minimum changes. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 327–338. IEEE, 2023.
- [35] Jiang Liu, Chun Pong Lau, and Rama Chellappa. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023.
- [36] Jiangfan Liu, Yishan Li, Yanming Guo, Yu Liu, Jun Tang, and Ying Nie. Generation and countermeasures of adversarial examples on vision: a survey. *Artificial Intelligence Review*, 57(8):199, 2024.
- [37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.
- [38] Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [39] Ningping Mou, Binqing Guo, Lingchen Zhao, Cong Wang, Yue Zhao, and Qian Wang. No-box universal adversarial perturbations against image classifiers via artificial textures. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2024.
- [40] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012.
- [41] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
- [42] Yaguan Qian, Kecheng Chen, Bin Wang, Zhaoquan Gu, Shouling Ji, Wei Wang, and Yanchun Zhang. Enhancing transferability of adversarial examples through mixed-frequency inputs. *IEEE Transactions on Information Forensics and Security*, 19:7633–7645, 2024.
- [43] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchun Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 19–37. Springer, 2020.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [46] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 29894–29918, 2023.
- [47] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1151–1160, 2020.
- [48] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in neural information processing systems*, 31, 2018.
- [49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [50] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020.
- [51] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

- [52] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.
- [53] Chenan Wang, Jinhao Duan, Chaowei Xiao, Edward Kim, Matthew Stamm, and Kaidi Xu. Semantic adversarial attacks via diffusion models. *arXiv preprint arXiv:2309.07398*, 2023.
- [54] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2021.
- [55] Qian Wang, Baolin Zheng, Qi Li, Chao Shen, and Zhongjie Ba. Towards query-efficient adversarial attacks against automatic speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:896–908, 2021.
- [56] Shen Wang, Zhaoyang Zhang, Guopu Zhu, Xinpeng Zhang, Yicong Zhou, and Jiwu Huang. Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18:351–364, 2023.
- [57] Xiaomeng Wang, Honglong Chen, Peng Sun, Junjian Li, Anqing Zhang, Weifeng Liu, and Nan Jiang. Advst: Generating unrestricted adversarial images via style transfer. *IEEE Transactions on Multimedia*, 26:4846–4858, 2024.
- [58] Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game. *IEEE Transactions on Image Processing*, 31:3541–3552, 2022.
- [59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [60] Wenzhao Xiang, Chang Liu, and Shibao Zheng. Improving visual quality of unrestricted adversarial examples with wavelet-vae. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.
- [61] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.
- [62] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.
- [63] Wenzhuo Xu, Kai Chen, Ziyi Gao, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Highly transferable diffusion-based unrestricted adversarial attack on pre-trained vision-language models. In *ACM Multimedia 2024*, 2024.
- [64] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36, 2024.
- [65] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- [66] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. *Advances in Neural Information Processing Systems*, 35:7546–7560, 2022.
- [67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [68] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [69] Z Zhao, Z Liu, and M Larson. Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. In *British Machine Vision Conference*. SI: British Machine Vision Conference, 2020.
- [70] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.
- [71] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018.

- [72] Ziming Zhao, Zhaoxuan Li, Fan Zhang, Ziqi Yang, Shuang Luo, Tingting Li, Rui Zhang, and Kui Ren. Sage: Steering the adversarial generation of examples with accelerations. *IEEE Transactions on Information Forensics and Security*, 18:789–803, 2023.
- [73] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2021.