Attack Deterministic Conditional Image Generative Models for Diverse and Controllable Generation

Tianyi Chu[1], Wei Xing[1], Jiafu Chen[1], Zhizhong Wang[1], Jiakai Sun[1], Lei Zhao[1*], Haibo Chen,[2] Huaizhong Lin[1*]

[1]Zhejiang University,
[2]Nanjing University of Science and Technology
{chutianyi, wxing, chenjiafu, endywon, csjk, cszhl, linhz}@zju.edu.cn, hbchen@njust.edu.cn

Abstract

Existing generative adversarial network (GAN) based conditional image generative models typically produce fixed output for the same conditional input, which is unreasonable for highly subjective tasks, such as large-mask image inpainting or style transfer. On the other hand, GAN-based diverse image generative methods require retraining/fine-tuning the network or designing complex noise injection functions, which is computationally expensive, task-specific, or struggle to generate high-quality results. Given that many deterministic conditional image generative models have been able to produce high-quality yet fixed results, we raise an intriguing question: *is it possible for pre-trained deterministic conditional image generative models to generate diverse results without changing network structures or parameters?* To answer this question, we re-examine the conditional image generation tasks from the perspective of adversarial attack and propose a simple and efficient plug-in projected gradient descent (PGD) like method for diverse and controllable image generation. The key idea is attacking the pre-trained deterministic generative models by adding a micro perturbation to the input condition. In this way, diverse results can be generated without any adjustment of network structures or fine-tuning of the pre-trained models. In addition, we can also control the diverse results to be generated by specifying the attack direction according to a reference text or image. Our work opens the door to applying adversarial attack to low-level vision tasks, and experiments on various conditional image generation tasks demonstrate the effectiveness and superiority of the proposed method.

Introduction

Conditional image generation (e.g., image inpainting, style transfer, super-resolution, denoising) is one of the representative subtasks of image synthesis, which typically uses a label, a part of an image, or reference images to guide the model to generate visually plausible images. Unlike high-level vision tasks, such as image classification, conditional image generation is a highly subjective task. That is, different generation results are allowed within the scope of visual plausibility. Such task can be classified into three categories based on the degree of constraints on generated content:

- Ill-posed task: There are no constraints on the structure or texture of the generated content, and the only requirement is that the generated samples are visually reasonable. For instance, the inpainting results of a masked face can be either happy or angry.
- Semi-ill-posed task: These tasks constrain the structure or texture of the generated result, beyond just being visually reasonable. For instance, given a content image and a style reference, the stylized result is required to be similar to the content image in structure while preserving the stroke/color of the style image.
- Well-posed task: These tasks require the generated results to be consistent with the input condition in both structure and texture, and thus relatively fixed generation result is expected. These tasks include small-scale super-resolution, dehazing, denoising, etc.

Note that the tasks may not strictly fall into one of the categories. For example, in the case of image inpainting with an extremely small masked area (e.g., a few pixels), it can be considered as having only one reasonable generation result, and therefore is closer to a well-posed task in the definition.

Traditional conditional image generative methods generate images based on the statistical information of the condition. For instance, super-resolution via bicubic interpolation and face generation based on eigen pattern (Turk and Pentland 1991). However, for tasks like label-guided image generation, traditional methods often struggle to generate visually plausible results. In recent years, the development and in-depth study of Generative Adversarial Networks (GANs) (Brock, Donahue, and Simonyan 2018; Karras, Laine, and Aila 2019) have been able to exhibit superior performance in generative tasks, even surpassing human-level abilities. Vanilla GAN samples random noise from a Gaussian distribution $z \sim P_z$ and maps it to the real image distribution $y = f_\theta(z) \sim P_y$. Thousands of GAN-based models have been proposed and trained in various subtasks of conditional image generation (e.g., inpainting, style transfer, super-resolution, dehazing) these years. However, most models designed for these subtasks are deterministic, as the models' inputs are only user-defined deterministic conditions (such as images) rather than randomly sampled noise. This implies that for a fixed condition input x, the model can only produce a unique corresponding output $y = f_\theta(x)$,

*Corresponding author

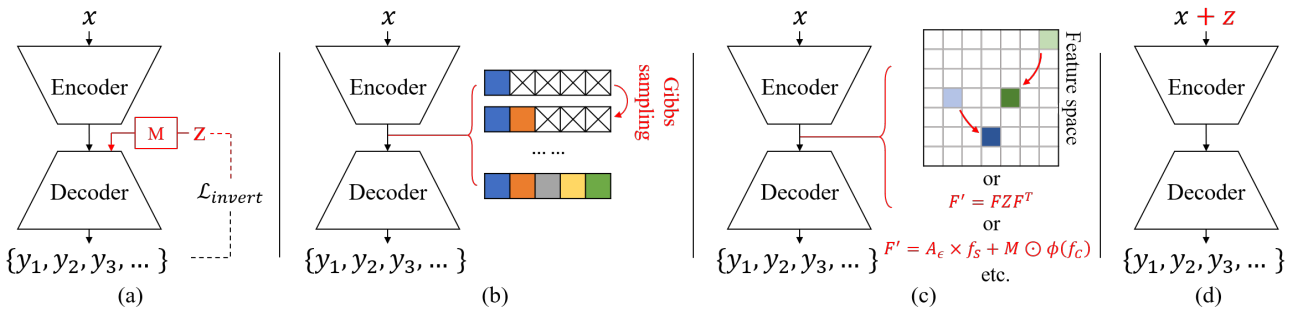

Figure 1: Three mainstream methods for introducing diversity in conditional image generation: (a) injecting noise through modulation module, (b) Gibbs sampling on feature sequences, and (c) transforming features according to specific rules. Our proposed method (d) allows pre-trained deterministic generative models to generate diverse results without multiple-step sampling, sophisticated transformation functions, or any adjustments to the network structure or parameters.

which contradicts the high subjectivity nature of the generation task, especially for the (semi-)ill-posed ones. Therefore, enabling models of these conditional generative tasks to produce diverse results, and furthermore, controlling the model to produce diverse results satisfying specific requirements (e.g., text guidance), has become a hot research topic.

As summarized in Fig. 1, there are three mainstream methods for introducing diversity into GAN-based conditional image generative models, which can be divided into two major categories. The first category involves redesigning the non-deterministic model and training, including a) injecting random noise to the model by modulation modules while using loss functions to constrain the one-to-one correspondence between the noise and the generated result (Zhao et al. 2021; Zheng, Cham, and Cai 2019) and b) sequentially generating the feature using Gibbs sampling (Liu et al. 2022; Esser, Rombach, and Ommer 2021). The second category directly employs pre-trained deterministic generative models, as in c) coupling random noise into latent code via well-designed task-related transformation funcations (Wang et al. 2021; Cheng et al. 2023). Unfortunately, the methods in the first category require users to design modulation modules for different tasks and networks, often demanding substantial computational costs for training. The method in the second category requires users to design intricate and complex transformation functions. Moreover, since this process can disrupt the inherent structural information of latent codes, it is typically limited to tasks like style transfer where precise reconstruction isn't essential. The aforementioned diversity methods have strong limitations for real-world applications and often struggle to generate high-quality results. Recognizing that many existing deterministic generative models have been able to produce satisfactory results, we propose an intriguing question: *for any pre-trained conditional generative model that can only produce high-quality but deterministic result, is it possible to achieve diversity without adjusting the network structure or fine-tuning the parameters?*

The answer is **yes**. Recall the classic task of adversarial attack in the field of AI security, where researchers have found that high-level vision (i.e., image classification, segmentation) models are vulnerable to adversarial examples – inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network (Goodfellow, Shlens, and Szegedy 2014). For a pre-trained classification network, the attacker only introduces imperceptible perturbations to the input, causing a significant change in the label output. Motivated by this, we introduce adversarial attack to low-level vision tasks and find that although the deterministic generative model is robust to random perturbations applied to the input, adversarial examples can still encourage it to generate diverse and visually plausible results, as Fig. 2 shown. Therefore, we propose two attacking approaches (i.e., **untargeted attack** for diverse generation and **targeted attack** for controllable text/image-guided generation) to empower the existing deterministic generative models to generate diverse and controllable results.

We conducted experiments on different tasks, such as image inpainting, style transfer, and super-resolution, using pre-trained models which have no diversity ability. Remarkably, without any fine-tuning of the pre-trained generative model, we achieved diverse results and demonstrated an intuitive positive correlation between the model's ability to generate diversity and the non-deterministic nature of the corresponding task. Additionally, we explored the capability of generating diverse samples with controllable semantics, where we utilized a pre-trained CLIP model (Radford et al. 2021) to determine the attack direction. Our contribution can be summarized as:

- We first introduce adversarial attack into conditional image generation and demonstrate the potential of deterministic generative models to produce diverse results.
- We propose a novel non-learning method that enables diverse generation of a deterministic generative model without any adjustment of network structure or fine-tuning. Our method is plug-in and can be easily applied.
- Extensive experiments demonstrate the effectiveness of our method. Our method can guide deterministic/diverse generative model to generate diverse and controllable results thanks to the advantages of being sensitive to initial perturbation and less prone to overfitting the constraint.
- Our method provides a new perspective for the interpretability research of low-level vision tasks and vision-language representation model.

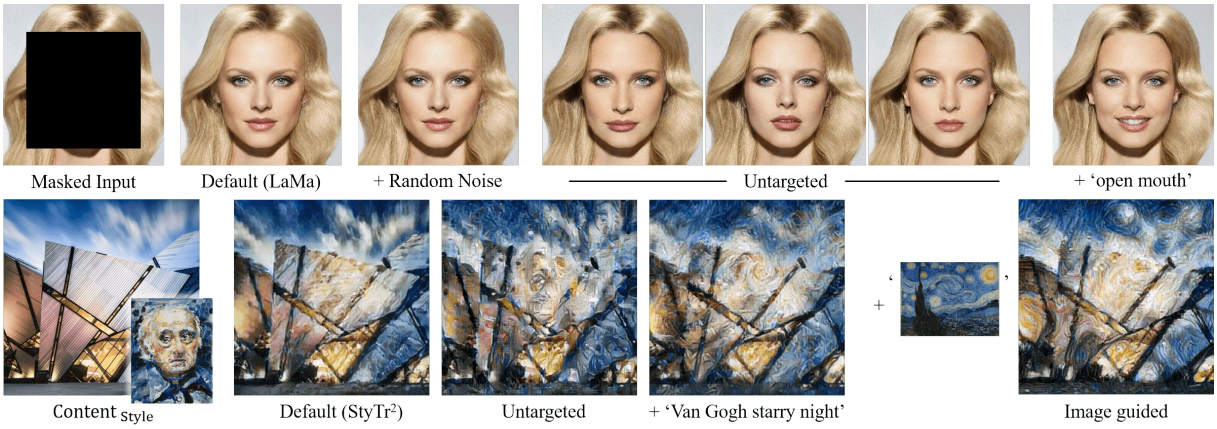

Figure 2: Diverse results generated by our method in conditional image generation tasks. We have tested upon two pre-trained deterministic models, including LaMa for image inpainting and StyTr^2 for style transfer. *Random noise* refers to adding standard Gaussian noise to the input. *Untargeted* refers to defining the attack direction to be as different from the default generated results as possible. + '' refers to specifying attack direction via text or reference image. (zoom-in for details)

Related Works

Conditional Image Generation

Traditional conditional image generation algorithms (Hertzmann 2003) aim to solve well-posed problems, e.g., image completion of simple geometric structure. With the development of deep learning, (Johnson, Alahi, and Fei-Fei 2016; Xie, Xu, and Chen 2012) took the lead in applying deep neural networks to conditional image generation and achieved impressive performance. (Mirza and Osindero 2014) first introduced generative adversarial network (GAN) structure to conditional image generative model, expanding new ideas for subsequent researchers. (Isola et al. 2017) exploits conditional GANs for inpainting. (Nazeri et al. 2019; Yang, Qi, and Shi 2020) proposed using coarse (edge map, gradient map, etc.) generation results to guide the inpainting. Additionally, (Yu et al. 2019; Suvorov et al. 2022) have explored the impact of different computational modules on inpainting performance. (Gatys, Ecker, and Bethge 2016) discovered that the Gram matrix upon deep features extracted from a pre-trained DCNN can notably represent the characteristics of visual styles, which opens up the era of neural style transfer. (Sanakoyeu et al. 2018) introduced GAN structure for style transfer. Subsequent works improve the performance of neural style transfer in many aspects, including quality (Deng et al. 2021) and generalization (Chiu 2019). (Dong et al. 2015) took the lead in introducing learning-based method into well-posed vision tasks, e.g., super-resolution, denoising, and JPEG compression artifact reduction. Subsequent works (Chen et al. 2021; Liang et al. 2021) have explored the impact of network structure on these tasks.

However, unlike probabilistic models, these methods cannot rely on random sampled input, which limits their ability to generate diverse samples for a fixed condition. The lacking of diversity can sometimes be unacceptable, especially in real-world applications of ill-posed tasks, even if those methods can produce high-quality results. Our method enables deterministic conditional generative models to produce diverse outputs with simple operations and minimal computational cost via adversarial attack and even guides them to generate outputs with specified prompts or reference, even if the conditions were never labeled in the training set.

Diverse Image Generation

Vanilla GAN (Goodfellow et al. 2020; Karras, Laine, and Aila 2019) belongs to probabilistic model, which learned to map from a normal distribution to the complex distribution of real image. Unlike vanilla GANs, conditional image generative models learn to map conditions to real images, where the conditional is deterministic and cannot be obtained through random sampling, which makes it difficult to produce diverse outputs. Previous researchers have proposed three mainstream methods for introducing diversity into conditional image generative models: 1) Injecting noise into the bottleneck layer directly or via a rather small modulation network and constraining the noise-output relationship through loss functions (Zhao et al. 2021; Zheng, Cham, and Cai 2019). Such methods generate diversity through randomly sample the noise during inference, which require training the network along with the noise or fine-tuning a pre-trained generator. 2) Sequentially generating a sequence using Gibbs sampling, which is commonly used in transformer models (Liu et al. 2022; Esser, Rombach, and Ommer 2021). The next predicted token's probability distribution is determined by the previously generated tokens due to the Markov chain property, leading to diversity. This method is usually time-consuming due to the inability to parallelize sampling. 3) Applying feature transformation in the latent space (Wang et al. 2021; Cheng et al. 2023). This method requires careful designing of transformation functions that may significantly affect the network's ability to reconstruct texture and structure accurately. As such, it is only applicable in style transfer tasks with a high tolerance for the lacking of reconstruction ability.

The above methods require high computational costs and

usually can only produce diversity on a specific subtask or a specific model. Our method is applicable to most of the conditional image generative models (especially GAN-based) without the need of re-training, fine-tuning, or designing any complex transformation function.

Adversarial Attack

The concept of adversarial attack is first proposed by (Szegedy et al. 2013), which are constructed by adding perturbations that are too small to be recognized by human eyes to an image but could cause the misclassification of the classification model with high confidence. Adversarial attack is typically applied in image classification (Papernot et al. 2016), object detection (Xiao et al. 2018), machine translation (Belinkov and Bisk 2017), etc.

Works such as (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017) explore white-box attack methods, which allow attackers to access model parameters and gradients. In contrast, (Chen, Jordan, and Wainwright 2020) explore black-box attacks, which only provide attackers with access to the model's input and output. These works are usually used to explore the linearity, robustness, and interpretability of deep neural networks. Our method is the first to introduce adversarial attack into conditional image generation task, bringing diversity to deterministic generative models in a simple but effective way.

Method

For a deterministic conditional image generative network f with parameter weight of θ and input condition x, the goal of adversarial attack is to introduce smallest possible perturbation z to input x to make the result $y' = f_\theta(x+z)$ as different as possible from the default output $y = f_\theta(x)$.

Untargeted Attack for Diversity

To perform adversarial attack on conditional image generative models, we need to obtain specific perturbations tailored to the input condition x. Following fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), which proposed to linearize the loss function $\mathcal{L}$ around the weight θ of the model, obtaining an optimal max-norm constrained perturbation $z = \epsilon sign\left(\nabla_x \mathcal{L}(\theta, x, y)\right)$, we define $x+z$ as our adversarial sample. Unlike image classification, it is challenging to design an attack loss that aligns with the training loss (e.g., GAN loss) of the image generative model. However, we have found that using simple statistical features as the attack loss can already generate samples that differ from the default results. Two simple statistical losses are used in our paper for untargeted attack.

$$\begin{aligned} \mathcal{L}_{L1}(\theta, x, y) &= \|f_\theta(x) - y\|_1 \\ \mathcal{L}_{var}(\theta, x, y) &= var\left(f_\theta(x)\right) \end{aligned} \tag{1}$$

y is the default generated result, var indicates deviation var of a generated image pixels. Gradient ∇_x can be easily obtained via backpropagation. $\mathcal{L}_{L1}$ is used as the statistical feature based losses in the following formulas.

The above process can be extended into multi-step for better attack performance as suggested in (Madry et al. 2017).

Algorithm 1: Adversarial attack on deterministic generative model, given pre-trained generative model f_θ and cross-model Vision-language representation model $CLIP$.

Input: Original input condition $X = \left\{x^1, \ldots, x^n\right\}$
Output: Generation result Y.

Random initialize each condition $x_1^k = x^k + z_0^k$, $z_0^k \sim N(0, {\delta^k}^2)$, $k = 1, \ldots, n$.
Init $\left\{\epsilon^1, \ldots, \epsilon^n\right\}$, c_{min}, c_{max}, $i, j = 1$
while $i < step$ **do**
  while $j < n$ **do**
    if untargeted **then**
      $z_i^j = \nabla_{x_i^j} \mathcal{L}_{L1}(\theta, X_i, y)$ // Equation (1)
    else
      $z_i^j = \nabla_{x_i^j} \mathcal{L}_{CLIP}(\theta, X_i, y)$ // Equation (6)
    end if
    $\delta_x = clip\left(x_i^j - x^j + \epsilon^j sign\left(z_i^j\right), c_{min}, c_{max}\right)$
    $x_{i+1}^j = x^j + \delta_x$
    $\epsilon^j = \epsilon^j * 0.95$
    $j = j + 1$
  end while
  $Y_{i+1} = f_\theta(X_{i+1})$
  $i = i + 1$
end while
return Y_i

We initialize the attack direction (Wong, Rice, and Kolter 2020) by adding a micro Gaussian noise z_0 on the original condition x. The adversarial sample is truncated by c_{min} and c_{max} in case of exceeding the value range. The process of multi-step adversarial attack can be expressed as:

$$\begin{aligned} &x_1 = x + z_0, \;\; z_0 \sim N(0, \delta^2) \\ &z_i = \nabla_{x_i} \mathcal{L}_{L1}(\theta, x_i, y) \\ &x_{i+1} = clip\left(x_i + \epsilon sign\left(z_i\right), c_{min}, c_{max}\right) \end{aligned} \tag{2}$$

However, we found that the attack may lead to serious artifacts in the generated results since $\mathcal{L}_{L1}$ not considering the visual plausibility of generated result. Therefore, we propose to use noise truncation instead of the PGD truncation method (see supplementary material for visual comparison):

$$x_{i+1} = x + clip\left(x_i - x + \epsilon sign\left(z_i\right), c_{min}, c_{max}\right) \tag{3}$$

Targeted Attack for Controllability

There has been a significant amount of work exploring multimodal image synthesis/editing (Zhan et al. 2021), among which using image or text to guide image generation is one of the most commonly used methods. Recently proposed Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) model has shown remarkable ability in extracting semantic correspondences between image and text. Hence we leverage its capabilities in our method to achieve targeted attack for image/text-guided generation. We align the vector pointing from the default generated result to the attacked generated result with the vector from the source image's description to the target description in the CLIP space,

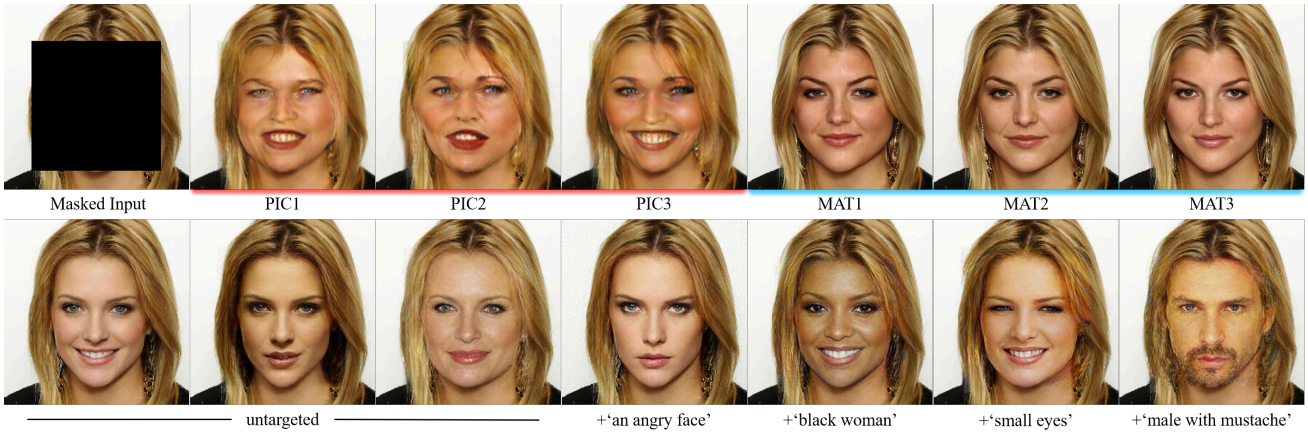

Figure 3: Diverse face inpainting results generated by attacking the deterministic inpainting model LaMa. Generated results are compared with diverse inpainting model PIC and MAT.

since directional CLIP loss (Gal et al. 2022) is proved to have better guidance ability compared to general CLIP similarity loss. Targeted attack direction can be defined as:

$$\triangle I = CLIP_I\left(f_\theta\left(x_i\right)\right) - CLIP_I\left(y\right) \tag{4}$$

$$\triangle R = \begin{cases} CLIP_T\left(T_{ref}\right) - CLIP_T\left(T_{src}\right), & text \\ CLIP_I\left(I_{ref}\right) - CLIP_I\left(y\right), & image \end{cases} \tag{5}$$

$$\mathcal{L}_{CLIP}\left(\theta, x_i, y\right) = \frac{\triangle I \cdot \triangle R}{|\triangle I||\triangle R|} \tag{6}$$

In which, $CLIP_I$ and $CLIP_T$ are CLIP's image and text encoders, T_{ref} and T_{src} are target and source text description, I_{ref} is the reference image. When image is used as reference, directional CLIP loss degrades to the general clip similarity loss. Pseudocode of untargeted/targeted attack can be seen in Alg. 1.

Experiments

We demonstrated the effectiveness of our method on different types of conditional image generation tasks. SOTA pre-trained deterministic models, including LaMa (Suvorov et al. 2022) for inpainting and StyTr^2 (Deng et al. 2021) for style transfer, are used to qualitatively and quantitatively validate the feasibility and effectiveness of our proposed method. In addition, we also conducted experiments on tasks such as super-resolution, dehazing, and probabilistic generation to further discuss the generalizability and limitations of our proposed method. The attack step is set to 10 by default.

Qualitative and Quantitative Comparison

For image inpainting (ill-posed task), we use $\epsilon \in [0.01, 0.05]$ with $c_{min} = -0.09$, $c_{max} = 0.09$. As shown in Fig. 3, PIC (Zheng, Cham, and Cai 2019) can produce relatively higher diversity but lacks the ability to inpaint large masked areas, resulting in poor quality of generated samples. MAT (Li et al. 2022) generates high-quality samples but can only achieve subtle diversity, which is hard to perceive. Our method achieves a satisfactory balance between generation quality and diversity. For style transfer

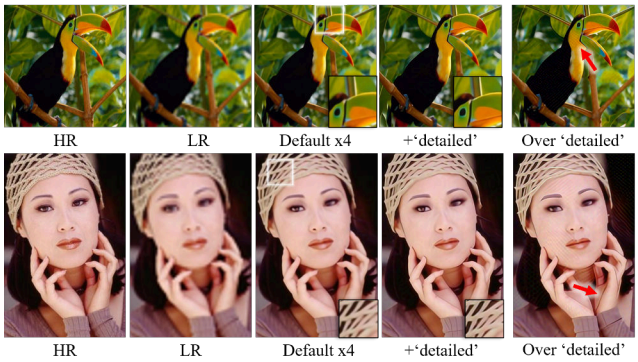

Figure 4: Our method works well on super-resolution (well-posed vision task). SwinIR x4 model demonstrates sharper generated results via attack using "detailed" as the direction.

(semi-ill-posed task), we use $\epsilon \in [0.1, 0.25]$ with $c_{min} = -0.25$, $c_{max} = 0.25$. As shown in Fig. 5, diverse samples generated by ϵ-AE (Cheng et al. 2023) share the same stroke patterns and may suffer from severe distortion under inappropriate hyperparameter settings, which fail to reflect the correlation between stylized image and style reference. DivSwapper (Wang et al. 2021) generates diversity through latent space swapping, but it is still limited by the pattern of the default stylized image and inevitably produces repeated textures. Our method maintains color consistency while generating style diversity. Benefiting from the prior of CLIP, our method can even make the stylized image have features that are not present in the style reference, such as "mosaic" or "puzzle" style in Fig. 5 (refer to supplementary material for more experimental results). To quantify diversity, we report the LPIPS distance and L1 distance between diverse generated samples in Tab. **??**. It can be observed that samples generated by our method have larger differences between each other, indicating higher diversity.

A reasonable assumption is that conditional image generative models have the ability to produce diversity if the loss function is loose (e.g., making the generated results

Figure 5: Left: Targeted diverse stylization. Top row: Content image, style image, and the default stylized result of StyTr^2. Second and third row: Text-guided stylization, compared with CLIPstyler (Kwon and Ye 2022) which also uses CLIP for guidance. The default stylized image in row one is used as the input of CLIPstyler. Our method faithfully preserves the color characteristics of the style image. CLIPStyler requires fine-tuning the reconstruction model, which takes several minutes, while our method can complete each step of attack within 0.2 second. Right: Untargeted diverse stylization. Compared with DivSwapper and ϵ-AE, our method generates higher diversity with better quality.

inpainting	LPIPS↑	L1↑	FID	Runtime
LaMa (Suvorov et al. 2022)	0	0	50.37	0.19s
LaMa+Ours	0.0606	13.5963	52.98	1.99s(10 step)
MAT (Li et al. 2022)	0.0176(29.0%)	5.9999(44.1%)	45.93	0.23s
StableDiffusion-v1.5 (Rombach et al. 2022)	0.0634(104.6%)	12.4692(91.7%)	146.60	5.5s
ICT (Wan et al. 2021)	0.0498(82.2%)	12.5915(92.6%)	75.41	89s
BAT (Yu et al. 2021)	0.0390(64.4%)	10.7356(78.96%)	62.44	22s
PIC (Zheng, Cham, and Cai 2019)	0.0265(43.7%)	8.4986(62.5%)	72.64	0.19s
style transfer				
StyTr^2 (Deng et al. 2021)	0	0	-	0.10s
StyTr^2+Ours	0.3199	51.6725	-	1.10s(10 step)
ϵ-AE (Cheng et al. 2023)	0.1826(57.1%)	37.1325(71.9%)	-	1.10s
CNNMRF+DivSwapper (Wang et al. 2021)	0.2183(68.2%)	39.1017(75.7%)	-	23.4s
Avatar-net+DFP (Wang et al. 2020)	0.2044(63.9%)	42.5471(82.3%)	-	3.5s

Table 1: Quantitative comparison of inpainting and style tranfer. Higher LPIPS/L1 score means higher diversity. Our method employs the standard 10-step attack, all experiments were conducted on a single RTX 3090 GPU.

conform to a certain distribution), which allows the model enough flexibility, so that our method can make (semi-)ill-posed vision models produce obvious diversity, as shown in Fig. 3 and 5. Surprisingly, our method also works well for the well-posed vision task models that only limited diversity can be accepted. As shown in Fig. 4, when apply 'detailed' with $\epsilon = 0.005$ and $c_{min} = -0.01$, $c_{max} = 0.01$ to the SwinIR (Liang et al. 2021) model, we successfully encourage it to generate sharper results than the default ones. (refer to supplementary material for quantitative comparison)

Compared with Optimizer-based Method

We noticed that early works in style transfer (Gatys, Ecker, and Bethge 2016) utilized optimizer to add gradient information of style representations from different layers of pre-trained neural networks to the content image for style transfer. In order to further explore the performance differences between our proposed method and the optimizer-based method, we treat the input condition x as the parameter to be optimized and use Adam optimizer for gradient propagation. In comparison to the optimizer-based method, our method has the following advantages:

1) Our method produces better diversity, which is attributed to its sensitivity to the initial perturbation. When applying small random perturbations to the input, our method can obtain different paths of variation, resulting in more diverse generated samples. We tested five sets of perturbation on LaMa with the same input masked images and an initial perturbation of $z_0 \sim N\left(0, 10^{-14}\right)$ using our method and the optimizer-based method. We calculated the L1 distance between each pair of updated perturbation $\|x_{i+1} - x_i\|_1$, where the average distance were 3.9×10^{-3} for our method and 2.54×10^{-5} for optimizer-based method. This confirms the sensitivity of our method to the initial perturbation.

2) Our method demonstrates better generation quality, which is specifically reflected in 1) less prone to overfit the constraint and 2) lesser prone to artifacts. Even after a sufficient number of iterations, our method can still generate visually plausible results, whereas the optimizer-based method tends to produce visually implausible ones. As shown in Fig. 6, our method maintains a relatively stable FID score as the iteration number increases, while the quality of the optimizer-based method deteriorates significantly.

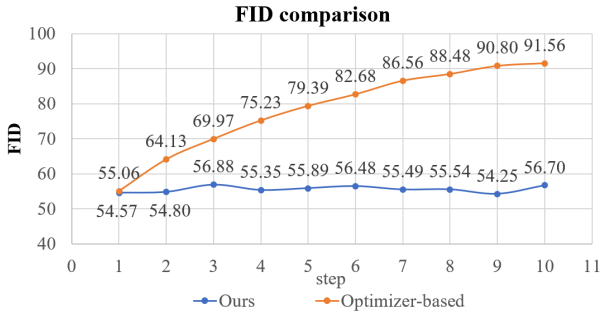

Figure 6: FID comparison of each step between our method and optimizer-based method to generate diverse inpainting result via LaMa.

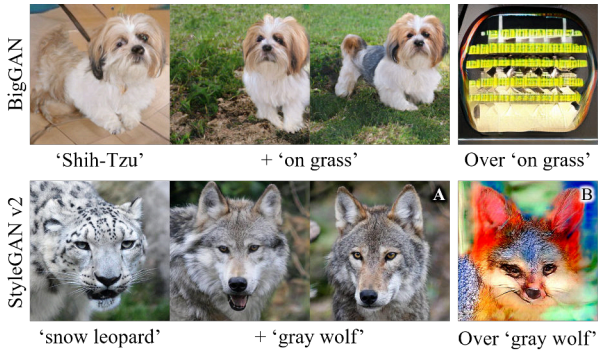

Figure 7: Our method also be applied to semantic control of probabilistic generative models. We observed that the model tends to generate unreasonable results when ϵ is too large.

Compared with Diffusion-based Method

We have noticed the remarkable performance of recently proposed diffusion models. Those models have been trained on extremely large text-image datasets and have the ability of open-domain generation. In our comparisons with diffusion models, we observed the following: 1) The generative performance of diffusion models heavily depends on the prompt. When using simple prompts like 'an angry face' as used in Fig. 3, diffusion models tend to generate content that aligns with the prompt but lacks visual coherence. 2) Diffusion models can not be directly applied to various subtasks of conditional image generation. Often, specific feature injection networks like ControlNet are required. 3) Diffusion models are constrained by multi-step reasoning, which often leads to longer inference times to generate high-quality results. (refer to supplementary material for more comparison)

Other Discussions

When generating adversarial examples for generative models, we expect the attack noise to be as small as possible (ideally, imperceptible to the human eyes) and the generated result to be visually plausible. Interestingly, when the attack noise is over large but not dominant in the input condition, the non-robustness of the generative model to adversarial examples can lead to the appearance of essential features of the model. As can be seen in Fig. 4, when a super-resolution model generates overly 'detailed' results, it is performing operations such as dividing different colors apart, enhancing or adding boundaries and adding grid-like artifacts. It can also be observed in the failure cases (refer to supplementary material) of inpainting model that the model tends to select one of the learned "standard pattern" and splice it onto the degraded area, similar to eigenface (Turk and Pentland 1991) in some extend. This also provides a novel view for the interpretability and data security study of conditional generative models.

We also conducted experiments on BigGAN (Brock, Donahue, and Simonyan 2018) and found that no matter how much random noise was added to the input, the model always generated samples that matched the class label. However, when we used adversarial examples with larger ϵ values (while keeping the distribution $x + z_i$ same to x), the model generated meaningless samples that did not match the label. This demonstrates that the mapping from the latent space to the image space that BigGAN has learned is incomplete. (see the detailed experiment in supplementary material)

Another interesting phenomenon we observed is that the CLIP space and human perceptual space are not strictly aligned. When the attack strength is set excessively high while not constraining the overall perturbation, the output tends to overfit the text guidance. For example, in Fig. 7, the CLIP similarity of sample A with the text "gray wolf" is 0.2030, while that of B is 0.2185. which means in CLIP space, sample B is more 'gray wolf' than sample A. We believe this is also the reason why our method outperforms the optimizer-based method for attacking deterministic models to generate diverse results.

Conclusion

We propose a simple and efficient approach to generating adversarial examples for generative models without requiring any modifications to their parameters or architectures. We first utilized adversarial attack to induce diversity in existing pre-trained deterministic conditional image generative models. Additionally, we leverage a pre-trained CLIP model to control the attack direction and encourage the generation of samples that satisfy specific semantics. We evaluate our method on various generative tasks and demonstrate that it achieves results surpassing those of state-of-the-art diverse generative models. Finally, we discuss the potential of adversarial attack in the interpretability and data security of low-level vision models.

Acknowledgments

This work was supported in part by Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201, 2019007, 2021009), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Program(2022Z167), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Belinkov, Y.; and Bisk, Y. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, 1277–1294. IEEE.

Chen, Z.; Wang, Y.; Yang, Y.; and Liu, D. 2021. PSD: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7180–7189.

Cheng, J.; Wu, Y.; Jaiswal, A.; Zhang, X.; Natarajan, P.; and Natarajan, P. 2023. User-controllable arbitrary style transfer via entropy regularization.

Chiu, T.-Y. 2019. Understanding generalized whitening and coloring transform for universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4452–4460.

Deng, Y.; Tang, F.; Pan, X.; Dong, W.; Ma, C.; and Xu, C. 2021. StyTrˆ 2: Unbiased Image Style Transfer with Transformers. *arXiv preprint arXiv:2105.14576*.

Dong, C.; Deng, Y.; Loy, C. C.; and Tang, X. 2015. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, 576–584.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

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

Hertzmann, A. 2003. A survey of stroke-based rendering. Institute of Electrical and Electronics Engineers.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kwon, G.; and Ye, J. C. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18062–18071.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, Q.; Tan, Z.; Chen, D.; Chu, Q.; Dai, X.; Chen, Y.; Liu, M.; Yuan, L.; and Yu, N. 2022. Reduce information loss in transformers for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11347–11357.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F. Z.; and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387. IEEE.

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

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

Sanakoyeu, A.; Kotovenko, D.; Lang, S.; and Ommer, B. 2018. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, 698–714.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, 2149–2159.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Turk, M.; and Pentland, A. 1991. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1): 71–86.

Wan, Z.; Zhang, J.; Chen, D.; and Liao, J. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4692–4701.

Wang, Z.; Zhao, L.; Chen, H.; Qiu, L.; Mo, Q.; Lin, S.; Xing, W.; and Lu, D. 2020. Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7789–7798.

Wang, Z.; Zhao, L.; Chen, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021. DivSwapper: towards diversified patch-based arbitrary style transfer. *arXiv preprint arXiv:2101.06381*.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*.

Xiao, C.; Deng, R.; Li, B.; Yu, F.; Liu, M.; and Song, D. 2018. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–234.

Xie, J.; Xu, L.; and Chen, E. 2012. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25.

Yang, J.; Qi, Z.; and Shi, Y. 2020. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12605–12612.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480.

Yu, Y.; Zhan, F.; Wu, R.; Pan, J.; Cui, K.; Lu, S.; Ma, F.; Xie, X.; and Miao, C. 2021. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 69–78.

Zhan, F.; Yu, Y.; Wu, R.; Zhang, J.; Lu, S.; Liu, L.; Kortylewski, A.; Theobalt, C.; and Xing, E. 2021. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*.

Zhao, S.; Cui, J.; Sheng, Y.; Dong, Y.; Liang, X.; Chang, E. I.; and Xu, Y. 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*.

Zheng, C.; Cham, T.-J.; and Cai, J. 2019. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1438–1447.

Supplementary Material of Attack Deterministic Conditional Image Generative Models for Diverse and Controllable Generation

Anonymous submission

Experimental Settings

Most of the experiments were done on a laptop with a single 6GB NVIDIA RTX 3000 GPU. The StyTr^2 (Deng et al. 2021) experiments were conducted on a single RTX 3090 GPU, as it requires more VRAM.

Comparison with Optimizer-based Method

Compared to optimizer-based methods, our approach is more sensitive to initial values and less prone to overfitting constraints. We believe the former is due to the nature of the sign function which amplifies arbitrarily micro perturbations to fixed values -1 ,0 and 1, while the latter is because of the truncation applied to the perturbation. Visual comparison of these two methods is shown in Fig. 1.

To verify our method is sensitive to the initial value, we applied three different random Gaussian noise with $N\left(0, 10^{-14}\right)$, $N\left(0, 10^{-10}\right)$, and $N\left(0, 10^{-6}\right)$to the masked input image, while keeping the optimization direction constant. As shown in Fig. 1 (right, row 1), the optimizer-based method produced highly similar results in each optimization round. Our method produced more diverse results in three experiments with perturbations of $N\left(0, 10^{-14}\right)$ compared to optimizer-based method with much stronger initial perturbation. The numbers labeled under row 1 represent the weight multiplied by the standard normal distribution.

To verify our method is not prone to overfitting constraints (e.g., the generated sample has high clip score, but visually unplausible), we performed two rounds of fifty-step iterations on our method and the optimizer-based method, respectively. As shown in Fig. 1 (right, row 2), our method can still produce visually reasonable results when attacked for 50 steps, while the image generated by the optimizer-based method has serious artifacts and is visually unnatural when updated for 50 steps. We believe this is because our method constrains the distance between the current condition and the initial condition, which allows only a limited perturbation to be imposed on the input, while optimizer-based methods do not have this constraint.

It can also be seen from Fig. 1 and 7 that our method can show diversity under the setting of different steps in the same round and the same step in different rounds, which also can not be achieved by optimizer-based methods.

Table 1: FID score between the ground truth and the attack-generated image at each step. Lower FID score means higher quality.

step	FID	step	FID	note
1	54.57	7	55.49	
2	54.80	8	55.54	
3	56.88	9	54.25	
4	55.35	10	56.70	
5	55.89	-	61.07	random noise, $\epsilon = 0.05$
6	56.48	-	50.51	random noise, $\epsilon = 0.01$

Quality Degradation

Our method injects noise into the input conditions to achieve diverse generation with out any training. We acknowledge that our method does introduce a slight reduction in the generation quality because the perturbations in the input are also retained as patterns by the inpainting model. However, balancing quality, diversity and computational cost has always been challenging. Compared to previous methods, our approach achieves a better balance For further proof, we perform 10 step untargeted attack on LaMa using 100 samples with random masks ($\geq$50%). In this experiment, $z_0 \sim N(0, 10^{-14})$, $\epsilon = 0.01$, $c_{min} = -0.05$ and $c_{max} = 0.05$. In addition, we also report the effect of adding random noise of different scales to the input on the generated results. Qualitative experiment results can be seen in Tab. 1. The FID between the default LaMa (Suvorov et al. 2022) inpainting result and the ground truth is 50.37, while that of MAT (Li et al. 2022) is 45.93 and that of Deepfill v2 (Yu et al. 2019) is 87.53.

Tricks in Experiment

In this section, we present the intriguing findings from our experiments and the techniques we employed:

1. While the use of L1 loss or variance loss alone is sufficient to motivate the model to produce untargeted diverse results, this constraint might lead the model to degenerate solutions(i.e., content that is all black or all white), as can be observed from Fig. 2. Although our method alleviates this phenomenon by constraining the overall noise intensity,

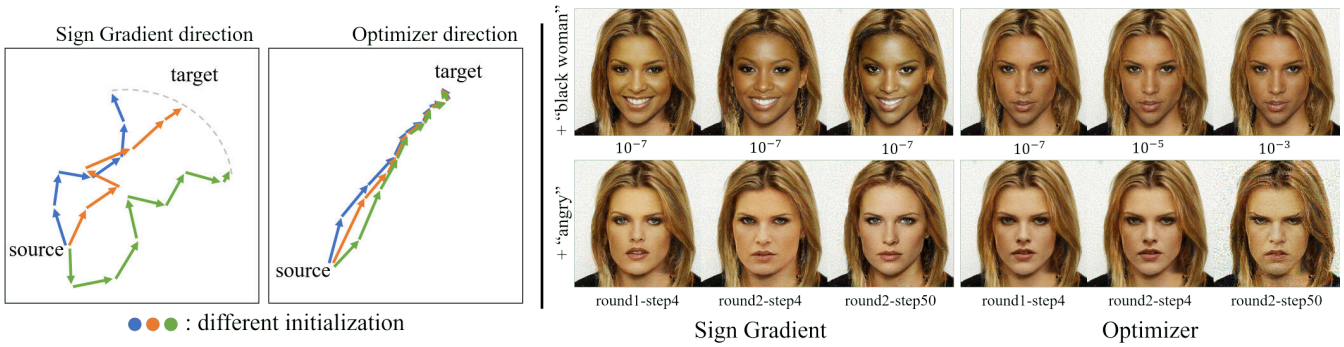

Figure 1: Comparison between our method and the optimizer-based method.

Table 2: Quantitative comparison on super-resolution model. "detailed" is used as the prompt to guide the generation of deterministic super-resolution model SwinIR (Liang et al. 2021). User study is conducted with eight participants, where they had to choose between 10 pairs of samples.

	SwinIR	SwinIR+Ours
PSNR	31.94	29.56
user study: more 'sharp'	14%	86%
user study: better SR quality	30%	70%

there remains a slight inclination for the generated results towards degenerate solutions. We discovered that employing random perceptual loss $\mathcal{L}_{RPerc}$ or random CLIP loss $\mathcal{L}_{RCLIP}$ can achieve untargeted attacks while minimizing the tendency of falling into degeneration:

$$\begin{aligned} \mathcal{L}_{RPerc}(\theta, x, y) &= \|\Phi_i(f_\theta(x)) - n\|_1, \quad n \sim N(\mu, \sigma) \\ \mathcal{L}_{RCLIP}(\theta, x, y) &= \|CLIP_I(f_\theta(x)) - CLIP_T(p)\|_1 \end{aligned} \tag{1}$$

Where Φ_i represents the output of the i-th layer of any feature extraction network (e.g., VGG19), and p represents a randomly generated string, such as 'YE mKhcKUf DVkgT'. 2. As analyzed earlier, CLIP space and human perception are not strictly aligned, even in most of the cases it achieves excellent text-to-image alignment. From another perspective, the samples in Fig. 8 of the main text can be viewed as an adversarial sample for CLIP. We recommend following a wildly used defense method against adversarial attack, which involves using slightly Gaussian blurred generated images as inputs to CLIP in order to obtain more robust image features.

Black-box Attack

Unlike white-box attack which assumes that the attacker has full knowledge of the target model, the attacker can only access the model through its input and output when performing black-box attack (Papernot et al. 2017). In other words, the attacker can not obtain the gradient information, which makes black-box attack more challenging than white-box attack. We propose the use of adversarial attack to induce diversity in conditional image generation. It is demonstrated that deterministic generative models can be encouraged to produce diverse outcomes under white-box attacks in the previous part of our paper. Here we employ black-box attacks to further validate the comprehensiveness of our argument. We use the boundary search and gradient estimation proposed by the decision-based black-box method HSJA (Chen, Jordan, and Wainwright 2020), selecting the perturbation with the maximum CLIP score with the text/image guidance as the estimated gradient. Unfortunately, even though we demonstrated that black-box attacks can still induce diversity in deterministic models, as shown in Fig. 4, the complexity of the constraints (perceptual) of generative models is much higher compared to classification models (one-hot labels). This makes it difficult to determine exact decision boundaries, resulting in poorer black-box attack performance compared to white-box attacks. Moreover, the search for decision boundaries demands significant computational resources. We encourage future researchers to build upon this paper and explore more black-box methods.

More results

As shown in Fig. 2, the global truncation method proposed by PGD (Madry et al. 2017) will make the attack effect too strong on the generated samples and cause serious artifacts. Our method can make this phenomenon less likely to occur. Fig. 6 shows more diverse inpainting results generated by LaMa with our method on the Places2 (Zhou et al. 2017) dataset. Fig. 9 shows more diverse style transfer results generated by StyTr^2 with our method. Fig. 10 shows diverse dehazing results generated by PSD with our method. Fig. 12 shows controllable generative results on StyleGAN v2 with our method. Fig. 3 shows failure cases generated by BigGAN, as mentioned in our paper. Fig. 15 shows that adversarial perturbations are not universal. Fig. 13 and 14 shows a comparison between the stable diffusion (Rombach et al. 2022) and the method presented in our paper. We noticed that despite ControlNet's (Zhang and Agrawala 2023) potential to leverage SD for style transfer, it doesn't generalize to diverse stylization of arbitrary content-style image pairs.

Figure 2: Comparison between truncation method proposed in PGD and ours. The attack direction is making the generated part to be as light as possible.

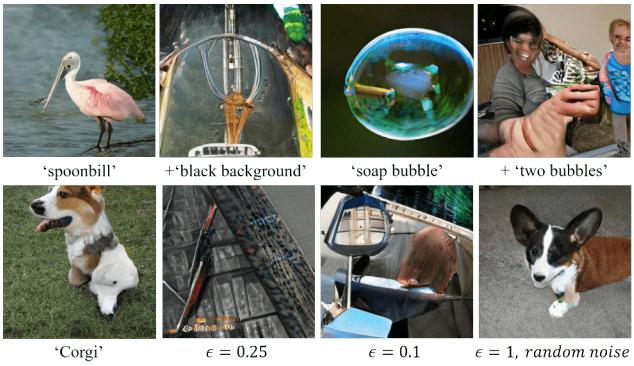

Figure 3: Failure attack on BigGAN. As mentioned in our paper, adversarial attack can sometimes cause unexpected mappings from noise space to image space of BigGAN even when the perturbed input noise is regularized to the default distribution. This phenomenon does not appear when random perturbations of arbitrary strength are added.

Limitation

The performance of the method proposed in this paper is strongly correlated with the inherent generation capability of the generative model itself. In other words, we cannot encourage the model to generate content outside its training domain, such as encouraging a facial generation model to achieve scene generation. Moreover, we have to honestly admit that we are currently unable to give an algorithm that automatically generates the optimal z_0, ϵ and (c_{min}, c_{max}), which means that although we have given suggested settings, users still need to manually adjust few hyperparameters according to personal preferences.

We encourage subsequent researchers to design more reasonable attack directions and try on more generative models.

potential misusage

The work proposed in this paper may have potential risks of misuse, such as generating content that violates privacy based on reference text or images. We recommend users to apply it to harmless tasks and carefully consider the risks of infringement.

References

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, 1277–1294. IEEE.

Chen, Z.; Wang, Y.; Yang, Y.; and Liu, D. 2021. PSD: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7180–7189.

Deng, Y.; Tang, F.; Pan, X.; Dong, W.; Ma, C.; and Xu, C. 2021. StyTrˆ 2: Unbiased Image Style Transfer with Transformers. *arXiv preprint arXiv:2105.14576*.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33: 12104–12114.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519.

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

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2149–2159.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4471–4480.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

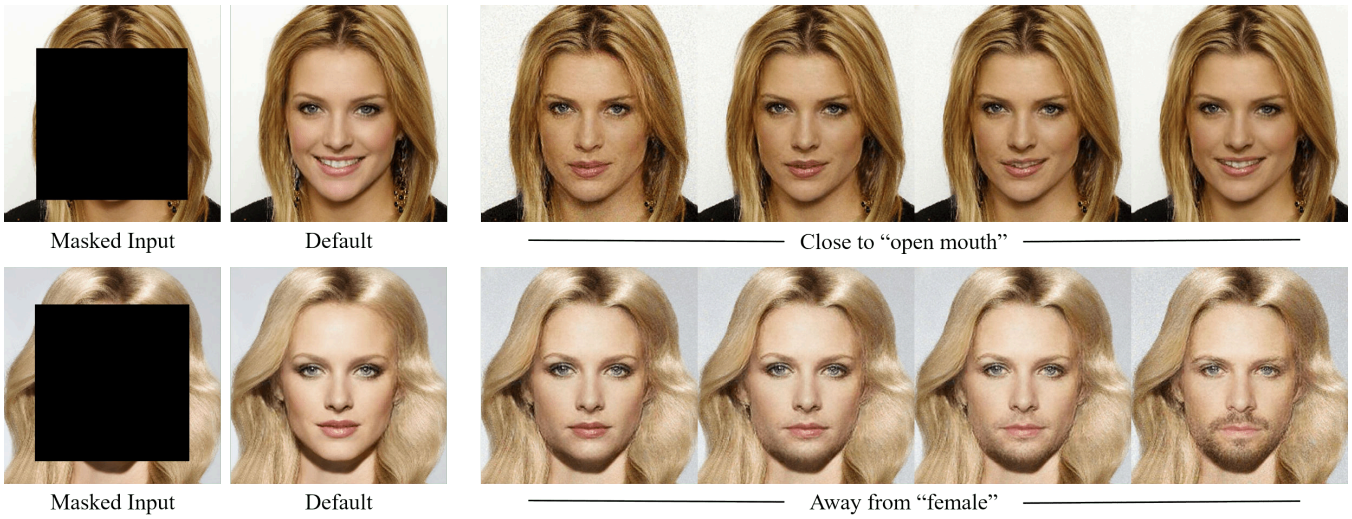

Figure 4: Black-box attack results. The experiment in the first row starts with an adversarial sample which will make the generative model produces a face with mouth closed.

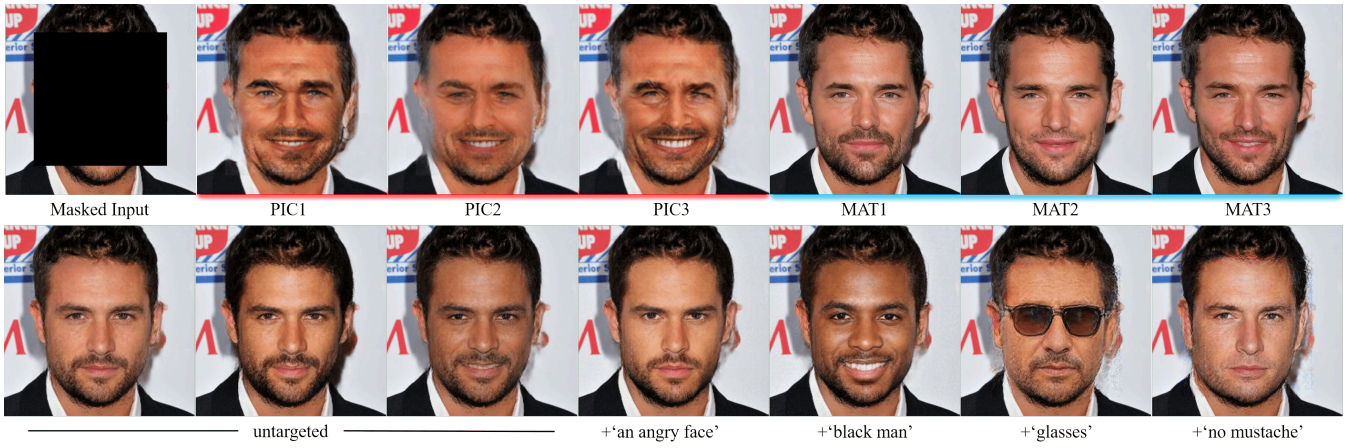

Figure 5: Qualititive comparison of diverse human face inpainting. Using the same experimental settings as in Fig. 4 of the main text.

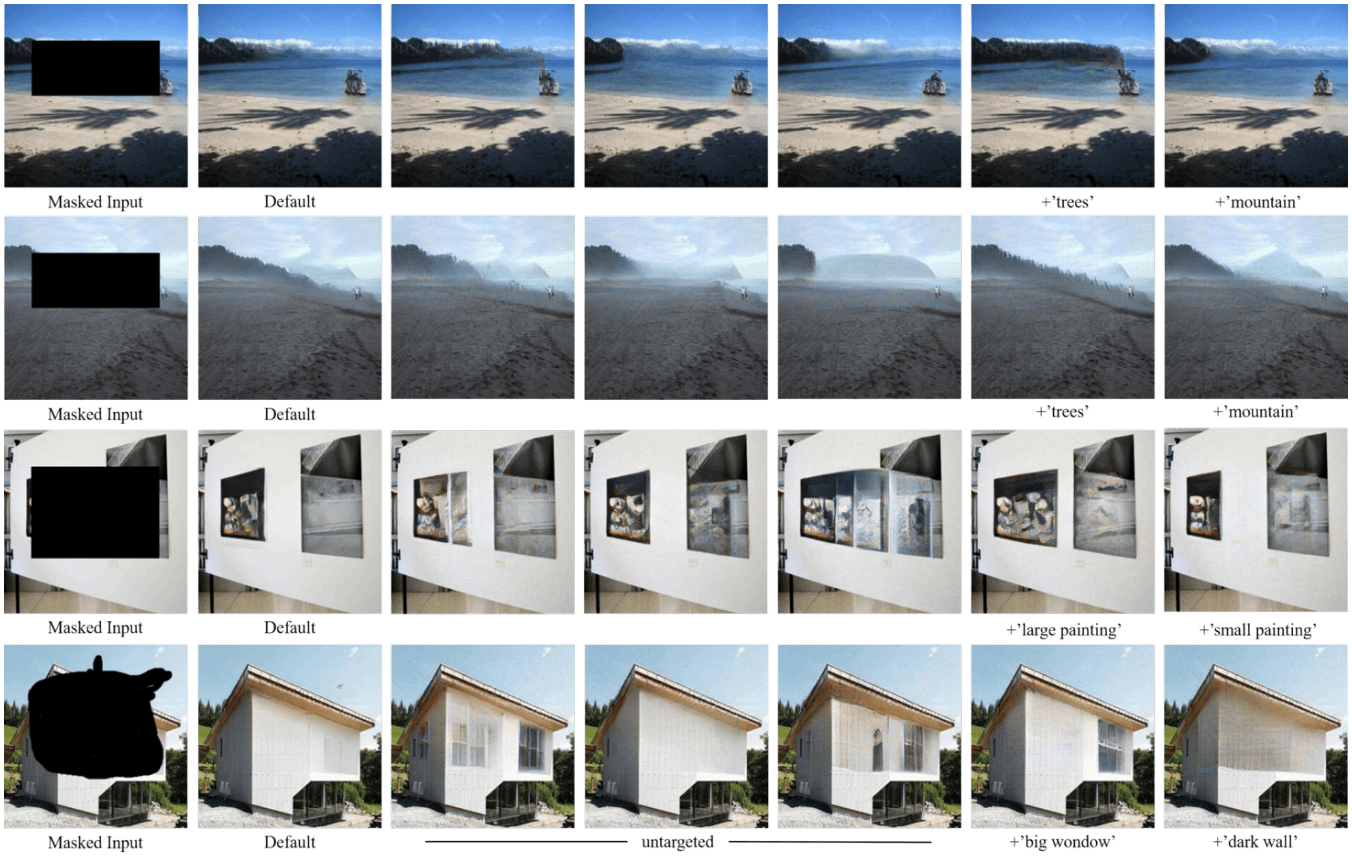

Figure 6: Diverse inpainting results produced by LaMa (Suvorov et al. 2022) under places2 (Zhou et al. 2017) dataset.

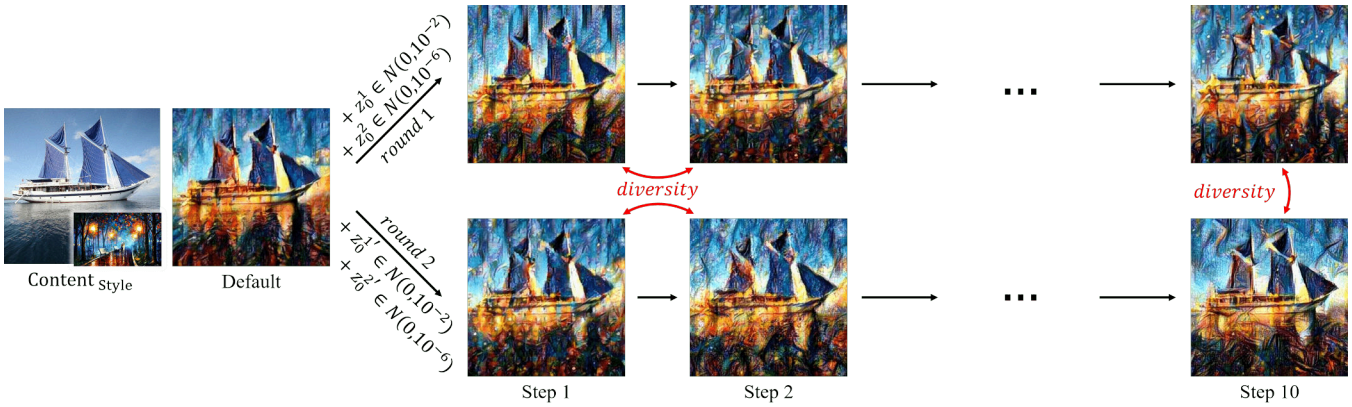

Figure 7: For two rounds of targeted multi-step attacks using reference 'starry' and same other settings, our method exhibits diversity between samples with different steps within the same round and that of with the same step in different rounds.

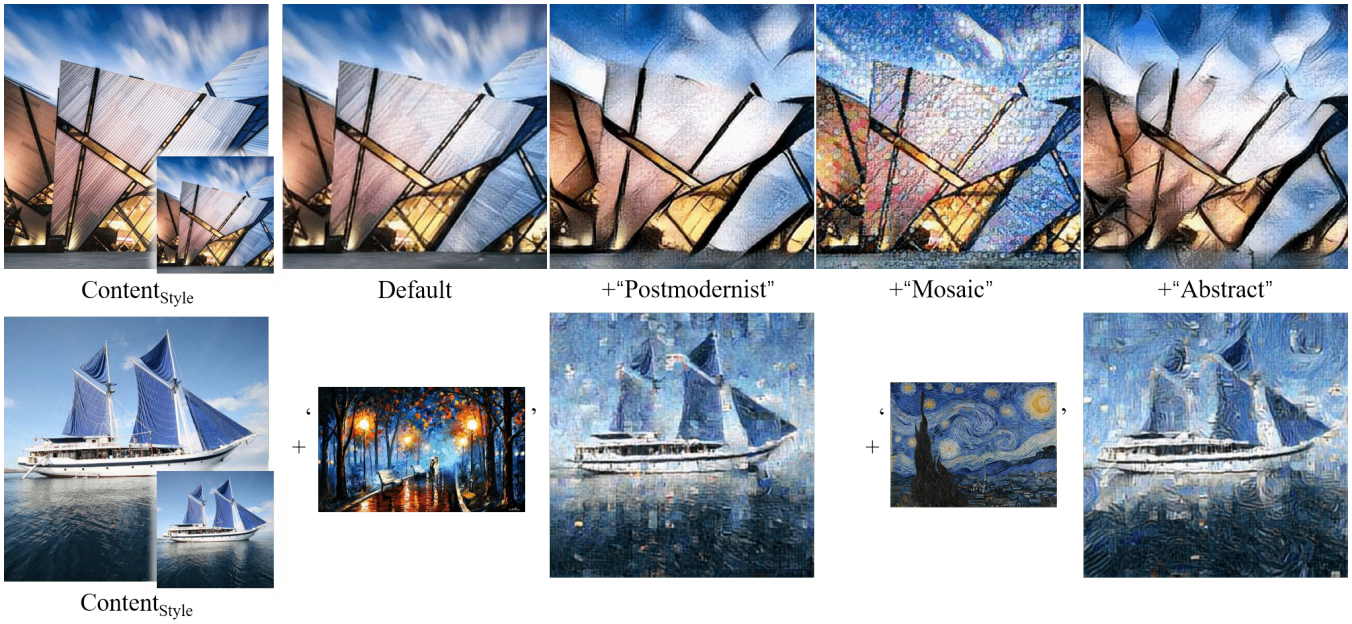

Figure 8: An interesting experiment on StyTr^2. The stylized result can be regarded as reconstruction when same images are used as content image and style reference. When applying our targeted attack, we found that the new stylized images produced new textures (strokes), such as different size color patches shown in the first row, and blocky strokes and distorted striped strokes in the second row. (zoom-in for details)

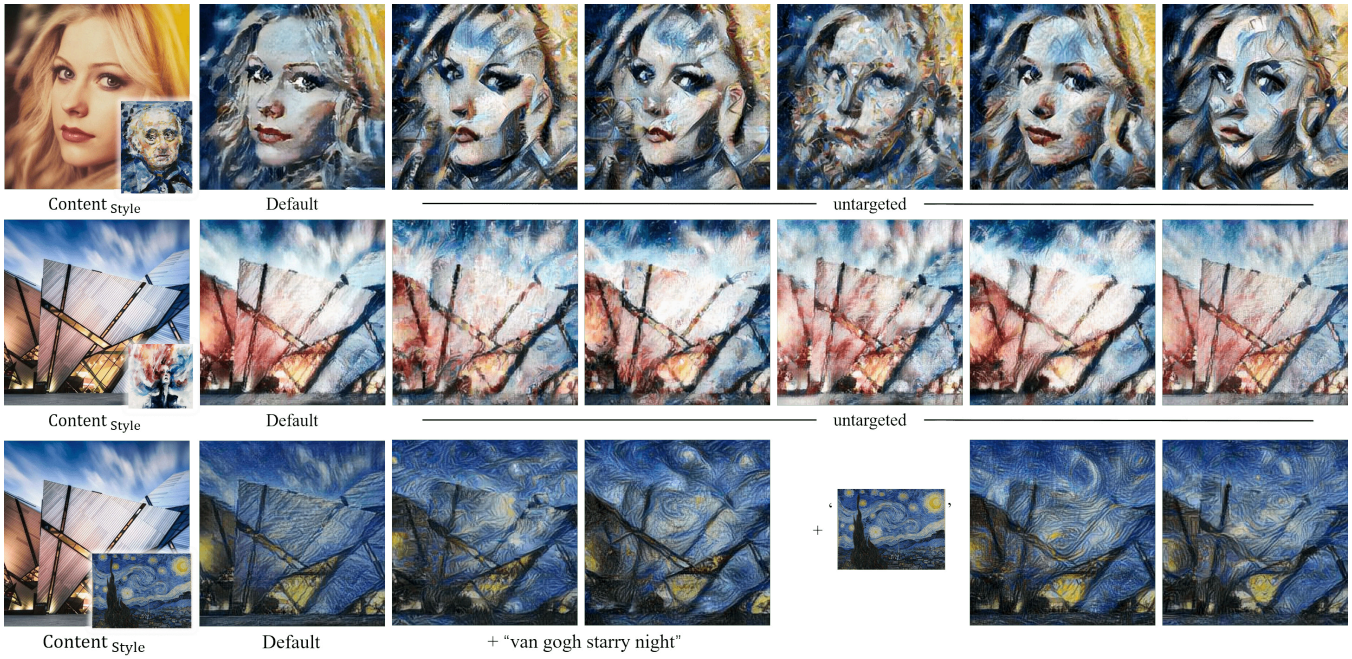

Figure 9: More diverse stylized results generated by StyTr^2 (Deng et al. 2021).

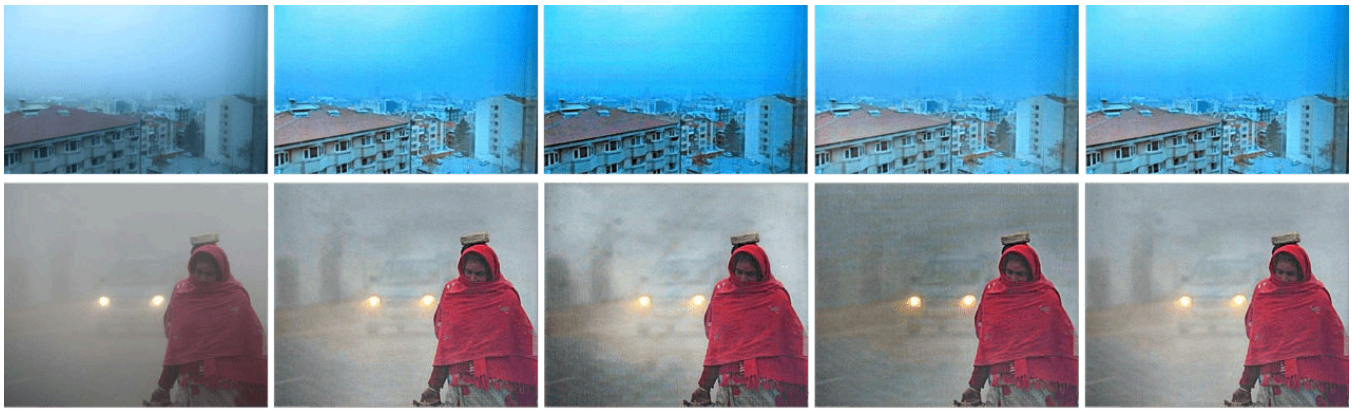

Figure 10: Generated results after attacking dehazing model PSD (Chen et al. 2021), it is difficult for our method to produce satisfactory results on some well-posed vision models.

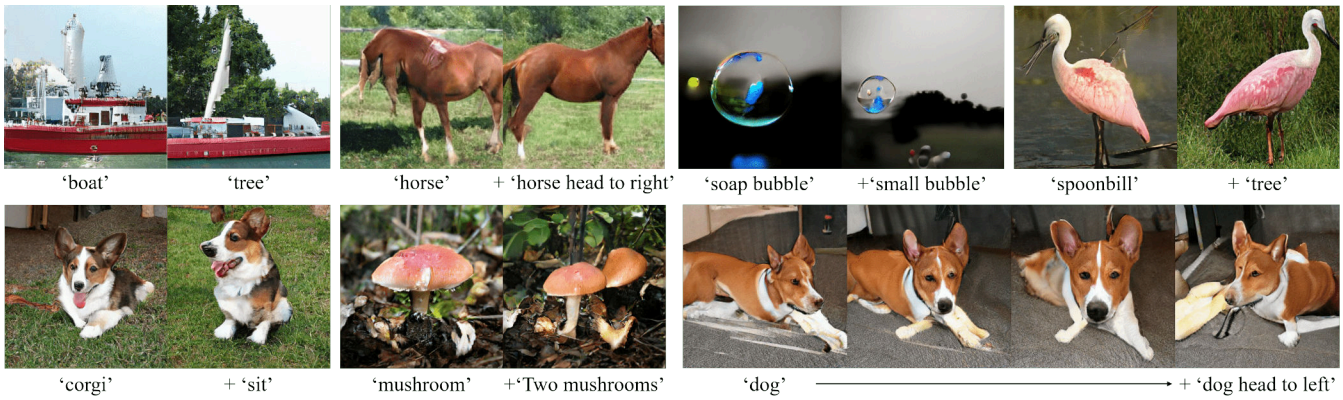

Figure 11: Controllable generation experiments on BigGAN (Brock, Donahue, and Simonyan 2018).

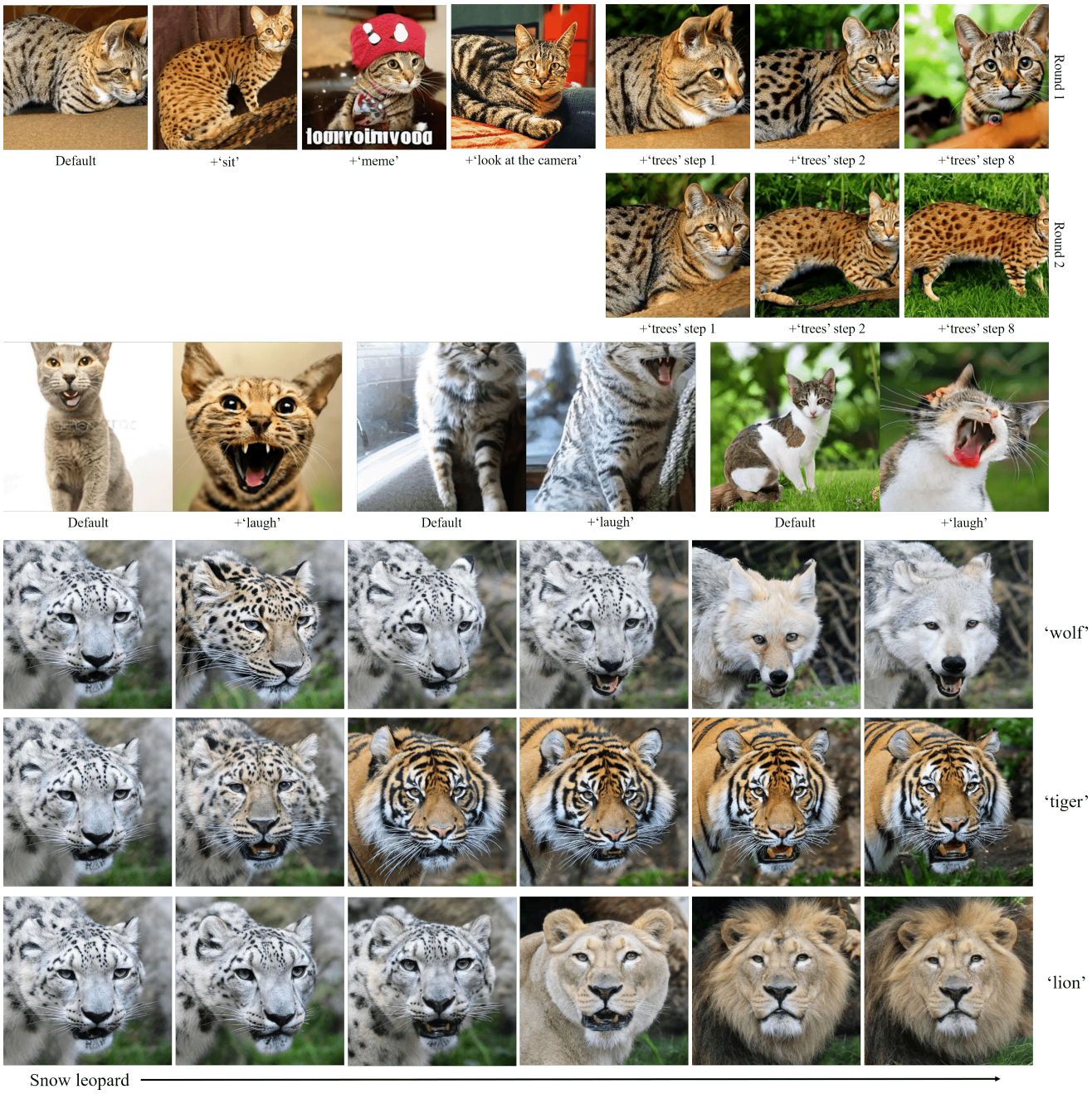

Figure 12: Controllable generation experiments on StyleGAN v2 (Karras et al. 2020). Pre-trained LSUN-cat and AFHQ-wild models are used.

Figure 13: Diverse inpainting results generated by Stable Diffusion v1.5. It takes about 5.5s to generate each sample on a single RTX 3090 GPU, while our method (10 step attack) takes about 1.0s. Same settings as Fig. 4 in the main text is used.

Figure 14: Diverse style transfer results generated by our method and Stable Diffusion v1.5 + ControlNet. For Stable Diffusion, it takes about 6.3s to generate each sample on a single RTX 3090 GPU, while our method (10 step attack) takes about 1.1s.

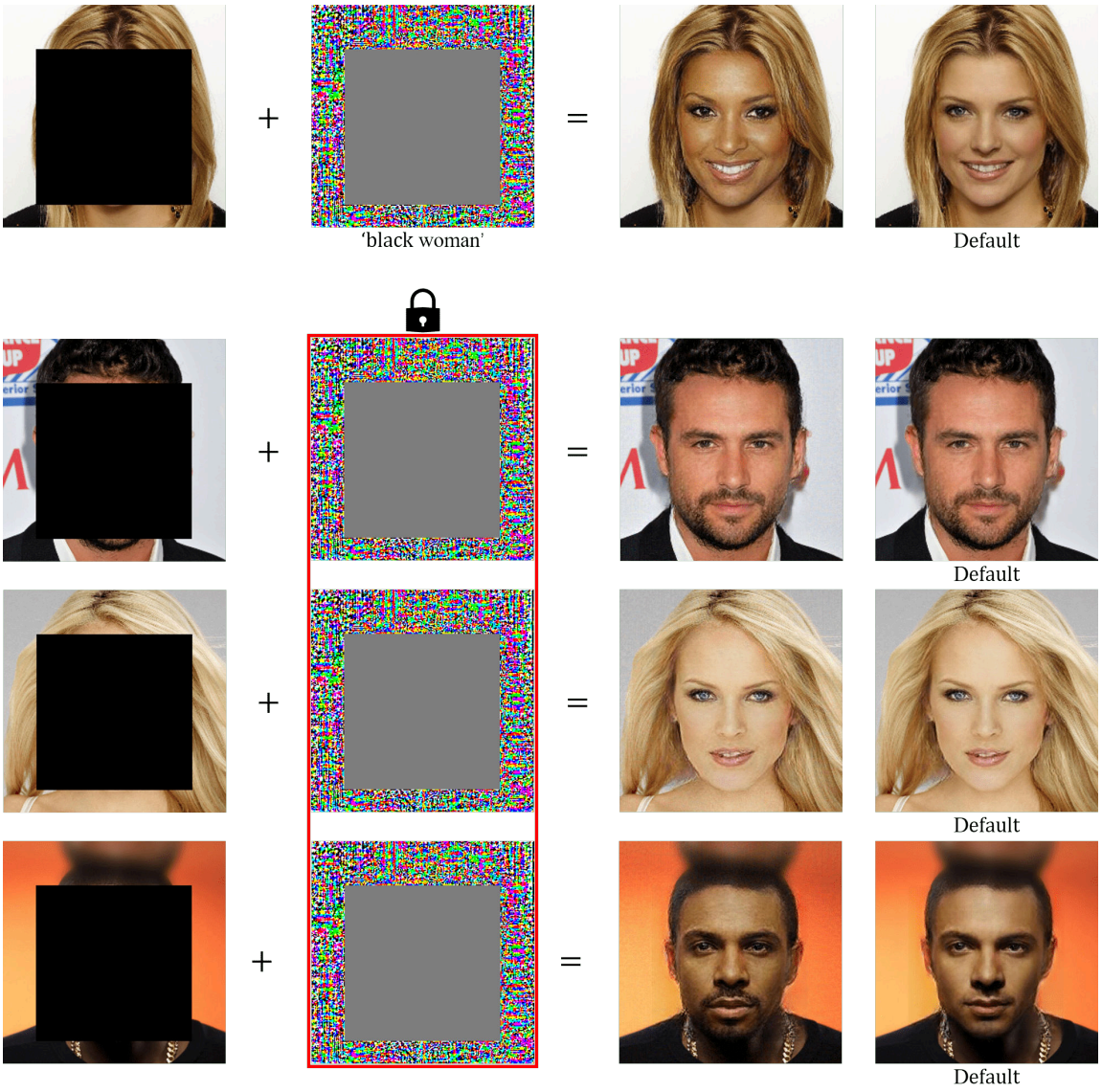

Figure 15: Research on the universal characteristic of adversarial perturbation. We attack the sample in row one using reference 'black women', then add the attack perturbation to other samples. We found that most of the samples are robust to the adversarial perturbation from other samples.