Attentive Eraser: Unleashing Diffusion Model's Object Removal Potential via Self-Attention Redirection Guidance

Wenhao Sun[1*], **Xue-Mei Dong**[1†], **Benlei Cui**[2*], **Jingqun Tang**[3]

[1]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China
[2]Alibaba Group, Hangzhou, China
[3]ByteDance Inc., Hangzhou, China
22020040141@pop.zjgsu.edu.cn, dongxuemei@zjgsu.edu.cn, cuibenlei.cbl@alibaba-inc.com, tangjingqun@bytedance.com

Abstract

Recently, diffusion models have emerged as promising newcomers in the field of generative models, shining brightly in image generation. However, when employed for object removal tasks, they still encounter issues such as generating random artifacts and the incapacity to repaint foreground object areas with appropriate content after removal. To tackle these problems, we propose *Attentive Eraser*, a tuning-free method to empower pre-trained diffusion models for stable and effective object removal. Firstly, in light of the observation that the self-attention maps influence the structure and shape details of the generated images, we propose Attention Activation and Suppression (ASS), which re-engineers the self-attention mechanism within the pre-trained diffusion models based on the given mask, thereby prioritizing the background over the foreground object during the reverse generation process. Moreover, we introduce Self-Attention Redirection Guidance (SARG), which utilizes the self-attention redirected by ASS to guide the generation process, effectively removing foreground objects within the mask while simultaneously generating content that is both plausible and coherent. Experiments demonstrate the stability and effectiveness of Attentive Eraser in object removal across a variety of pre-trained diffusion models, outperforming even training-based methods. Furthermore, Attentive Eraser can be implemented in various diffusion model architectures and checkpoints, enabling excellent scalability.

Introduction

The widespread adoption of diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Song et al. 2021) in recent years has enabled the generation of high-quality images that match the quality of real photos and provide a realistic visualization based on user specifications. This raises a natural question of whether the image-generating capabilities of these models can be harnessed to remove objects of interest from images. Such a task, termed object removal (Yu et al. 2018; Suvorov et al. 2022), represents a specialized form of image inpainting, and requires addressing two critical aspects. Firstly, the user-specified object (usually given as a binary mask) must be successfully and effectively removed from the image. Secondly, the mask area must be filled with content that is realistic, plausible, and appropriate to maintain overall coherence within the image.

Traditional approaches for object removal are the patch-based methods (Guo et al. 2018; Lu et al. 2018), which fill in the missing regions after removal by searching for well-matched replacement patches ($i.e.$ candidate patches) in the undamaged part of the image and copying them to the corresponding removal locations. However, such processing methods often lead to inconsistency and unnaturally between the removed region and its surroundings. In recent years, convolutional neural networks (CNNs) have demonstrated considerable potential for object removal tasks. However, CNNs-based methods (Yan et al. 2018; Oleksii 2019; Suvorov et al. 2022) typically utilize a fixed-size convolutional kernel or network structure, which constrains the perceptual range of the model and the utilization of contextual information (Fang et al. 2023a; Xu et al. 2024). Consequently, the model's performance is sub-optimal when confronted with large-scale removal or complex scenes.

With the rapid development of generative models (Shen et al. 2024b) in deep learning(Fang et al. 2024c), a proliferation of generative models has been applied to object removal. Among these, the most common are generative adversarial network (GAN) (Goodfellow et al. 2014)-based methods and DMs-based methods. GAN-based methods (Chen and Hu 2019; Shin et al. 2020) employ neural networks of varying granularity, with the context-focused module exhibiting robust performance and efficacy in image inpainting. However, their training is inherently slow and unstable, and they are susceptible to issues such as mode collapse or failure to converge (Salimans et al. 2016).

In current times, DMs have made new waves in the field of deep generative models, broken the long-held dominance of GANs, and achieved new state-of-the-art performance in many computer vision tasks (Shen et al. 2024a,b; Shen and Tang 2024; Zhao et al. 2024b). The most prevalent open-source pre-trained model in DMs is Stable Diffusion (SD) (Rombach et al. 2022), which is a pre-trained latent diffusion model. To apply SD to the object removal task, fine-tuned from SD, SD-inpainting (Rombach et al. 2022) was developed into an end-to-end model with a particular focus on inpainting, to incorporate a mask as an additional condition within the model. However, even after spending a consider-

[*]These authors contributed equally.
[†]Corresponding author

able cost in terms of resources, its object removal ability is not stable, and it often fails to completely remove the object or generates random artifacts(as shown in Figure 4). An additional methodology entails guiding the model to perform object removal via prompt instruction (Yildirim et al. 2023; Brooks, Holynski, and Efros 2023). The downside of this method is that to achieve a satisfactory result, these models often necessitate a considerable degree of prompt engineering and fail to allow for accurate interaction even with a mask. Additionally, they often necessitate substantial resources for fine-tuning.

To address these problems, we propose a tuning-free method, Attentive Eraser, a simple yet highly effective method for mask-guided object removal. This method ensures that during the reverse diffusion denoising process, the content generated within the mask tends to focus on the background rather than the foreground object itself. This is achieved by modifying the self-attention mechanism in the SD model and utilizing it to steer the sampling process. We show that when Attentive Eraser is combined with the prevailing diffusion-based inpainting pipelines (Couairon et al. 2023; Avrahami, Fried, and Lischinski 2023), these pipelines enable stable and reliable object removal, fully exploiting the massive prior knowledge in the pre-trained SD model to unleash its potential for object removal (as shown in Figure 1). The main contributions of our work are presented as follows:

- We propose a tuning-free method **Attentive Eraser** to unleash DM's object removal potential, which comprises two components: (1) **Attention Activation and Suppression (AAS)**, a self-attention-modified method that enables the generation of images with enhanced attention to the background while simultaneously reducing attention to the foreground object. (2) **Self-Attention Redirection Guidance (SARG)**, a novel sampling guidance method that utilizes the proposed AAS to steer sampling towards the object removal direction.
- Experiments and user studies demonstrate the effectiveness, robustness, and scalability of our method, with both removal quality and stability surpassing SOTA methods.

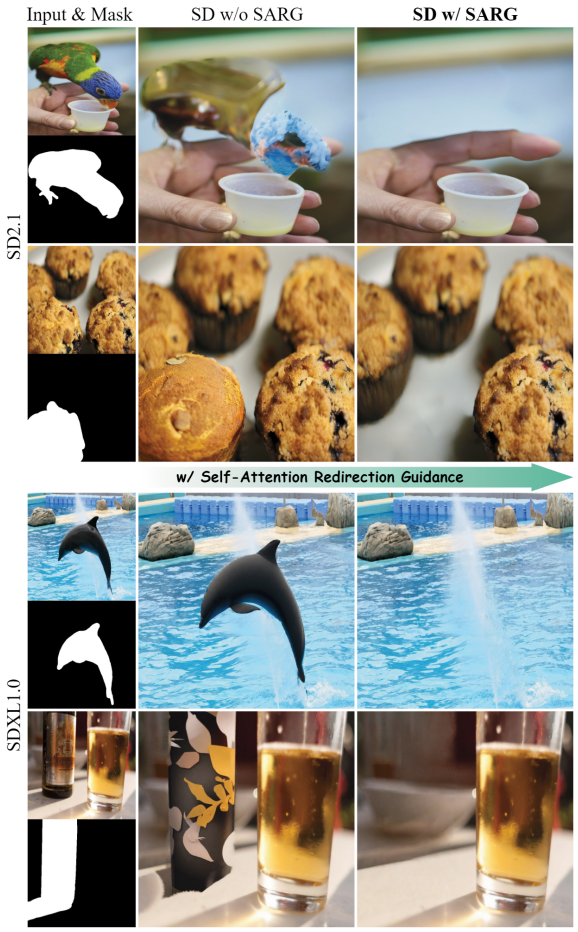

Figure 1: Qualitative comparison between Stable Diffusion (baseline) and self-attention redirection guided Stable Diffusion for object removal.

Related Works

Diffusion Models for Object Removal

Existing diffusion model-based object removal methods can be classified into two categories, tuning-free (Zhao et al. 2024a) vs. training-based (Fang et al. 2023b), depending on whether they require fine-tuning or not. In the case of the training-based methods, DreamInpainter (Xie et al. 2023b) captures the identity of an object and removes it by introducing the discriminative token selection module. Powerpaint (Zhuang et al. 2023) introduces learnable task prompts for object removal tasks. Inst-Inpaint (Yildirim et al. 2023) constructs a dataset for object removal, and uses it to fine-tune the pre-trained diffusion model. There are other instruction-based methods achieving object removal via textual commands (Huang et al. 2024; Yang et al. 2024b; Geng et al. 2024). In the case of the tuning-free methods, Blended Diffusion (Avrahami, Fried, and Lischinski 2023) and ZONE (Li et al. 2024) perform local text-guided image manipulations by introducing text conditions to the diffusion sampling process. Magicremover (Yang et al. 2023) implements object removal by modifying cross-attention to direct diffusion model sampling. However, these methods can lead to artifacts in the final result or incomplete removal of the target due to the stochastic nature of the diffusion model itself and imprecise guiding operations. To address the above issues and to avoid consuming resources for training, we propose a tuning-free method SARG to gradually steer the diffusion process towards object removal.

Sampling Guidance for Diffusion Models

Sampling guidance for diffusion models involves techniques that steer the sampling process toward desired outcomes. Classifier guidance (Dhariwal and Nichol 2021) involves the incorporation of an additional trained classifier to generate samples of the desired category. Unlike the former, Classifier-free Guidance (Ho and Salimans 2021) does not rely on an external classifier but instead constructs an implicit classifier to guide the generation process. There are

two methods that combine self-attention with guidance, SAG (Hong et al. 2023) and PAG (Ahn et al. 2024), which utilize or modify the self-attention mechanism to guide the sampling process, thereby enhancing the quality of the generated images. Our work is similar to PAG in that it modifies the self-attention map to guide sampling, but the purpose and approach to modification are different.

Preliminaries

Diffusion Models

DMs are a class of probabilistic generative models that learn a given data distribution $q(x)$ by progressively adding noise to the data to destroy its structure and then learning a corresponding inverse process of a fixed Markov chain of length T to denoise it. Specifically, given a set of data $x_0 \sim q(x_0)$, the forward process could be formulated by

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right), \quad (1)$$

where $t \in \{1, 2, \ldots, T\}$ denotes the time step of diffusion process, x_t is the noisy data at step t, $\beta_t \in [0, 1]$ is the variance schedule at step t and represents the level of noise.

Starting from x_T, the reverse process aims to obtain a true sample by iterative sampling from $q(x_{t-1} \mid x_t)$. Unfortunately, this probability is intractable, therefore, a deep neural network with parameter θ is used to fit it:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta^{(t)}(x_t), \Sigma_\theta^{(t)}(x_t)\right), \quad (2)$$

With the parameterization

$$\mu_\theta^{(t)}(x_t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta^{(t)}(x_t)\right), \quad (3)$$

proposed by Ho(Ho, Jain, and Abbeel 2020), a U-net (Ronneberger, Fischer, and Brox 2015) $\epsilon_\theta^{(t)}(x_t)$ is trained to predict the noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ that is introduced to x_0 to obtain x_t, by minimizing the following object:

$$\min_\theta \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \text{ Uniform } (1, T)} \left\| \epsilon - \epsilon_\theta^{(t)}(x_t) \right\|_2^2, \quad (4)$$

After training, a sample x_0 can be generated following the reverse process from $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Self-Attention in Stable Diffusion

Recent studies (Patashnik et al. 2023; Nam et al. 2024; Liu et al. 2024) have elucidated the significant role of the self-attention module within the stable diffusion U-net. It harnesses the power of attention mechanisms to aggregate features (Tang et al. 2022; Shen et al. 2023; Fang et al. 2023c), allowing for a more nuanced control over the details of image generation. Specifically, given any latent feature map $z \in \mathbb{R}^{h \times w \times c}$, where h, w and c are the height, width and channel dimensions of z respectively, the according query matrix $Q_{self} \in \mathbb{R}^{(h \times w) \times d}$, key matrix $K_{self} \in \mathbb{R}^{(h \times w) \times d}$ and value matrix $V_{self} \in \mathbb{R}^{(h \times w) \times d}$ can be obtained through learned linear layers ℓ_Q , ℓ_K and ℓ_V, respectively. The similarity matrix S_{self}, self-attention map A_{self} and output OP_{self} can be defined as follows:

$$Q_{self} = \ell_Q(z), K_{self} = \ell_K(z), V_{self} = \ell_V(z), \quad (5)$$

$$S_{self} = Q_{self}\left(K_{self}\right)^T / \sqrt{d}, \quad (6)$$

$$A_{self} = \text{softmax}\left(S_{self}\right), \quad (7)$$

$$OP_{self} = A_{self} V_{self}, \quad (8)$$

where d is the dimension of query matrix Q_{self}, and the similarity matrix $S_{self} \in \mathbb{R}^{(h \times w) \times (h \times w)}$ and self-attention map $A_{self} \in \mathbb{R}^{(h \times w) \times (h \times w)}$ can be seen as the query-key similarities for structure (Ahn et al. 2024), which represent the correlation between image-internal spatial features, influence the structure and shape details of the generated image. In SD, each such spatial feature is indicative of a particular region of the generated image. Inspired by this insight, we achieve object removal by changing the associations between different image-internal spatial features within the self-attention map.

Guidance

A key advantage of diffusion models is the ability to integrate additional information into the iterative inference process for guiding the sampling process, and the guidance can be generalized as any time-dependent energy function from the score-based perspective. Modifying $\epsilon_\theta^{(t)}(z_t)$ with this energy function can guide the sampling process towards generating samples from a specifically conditioned distribution, formulated as:

$$\hat{\epsilon}_\theta^{(t)}(z_t; C) = \epsilon_\theta^{(t)}(z_t; C) - s\, \mathbf{g}(z_t; y), \quad (9)$$

where C represents conditional information, $\mathbf{g}(z_t; y)$ is an energy function and y represents the imaginary labels for the desirable sample and s is the guidance scale. There are many forms of $\mathbf{g}$ (Nichol et al. 2021; Dhariwal and Nichol 2021; Ho and Salimans 2021; Bansal et al. 2023; Epstein et al. 2023; Mo et al. 2024), the most prevalent of which is classifier-free guidance (Ho and Salimans 2021), where C represents textual information (Liu et al. 2023; Fang et al. 2024a,b), $\mathbf{g} = \epsilon_\theta$ and $y = \emptyset$.

Methodology

Overview

The overall framework diagram of the proposed method is depicted in Figure 2. There are two principal components: **AAS** and **SARG**, which will be elucidated in more detail in the following sections.

Attention Activation and Suppression

Consider l to be a specific self-attention layer in the U-net that accepts features of dimension $N \times N$, the corresponding similarity matrix and attention map at timestep t, $S_{l,t}^{self}, A_{l,t}^{self} \in \mathbb{R}^{N^2 \times N^2}$ can be obtained. The magnitude of the value of $A_{l,t}^{self}[i, j]$ in the self-attention map represents the extent to which the token i generation process is influenced by the token j. In other words, row i in the map indicates the extent to which each token in the feature map influences the generation process of token i, while column j in the map indicates the extent to which token j influences the generation process of all tokens in the feature map.

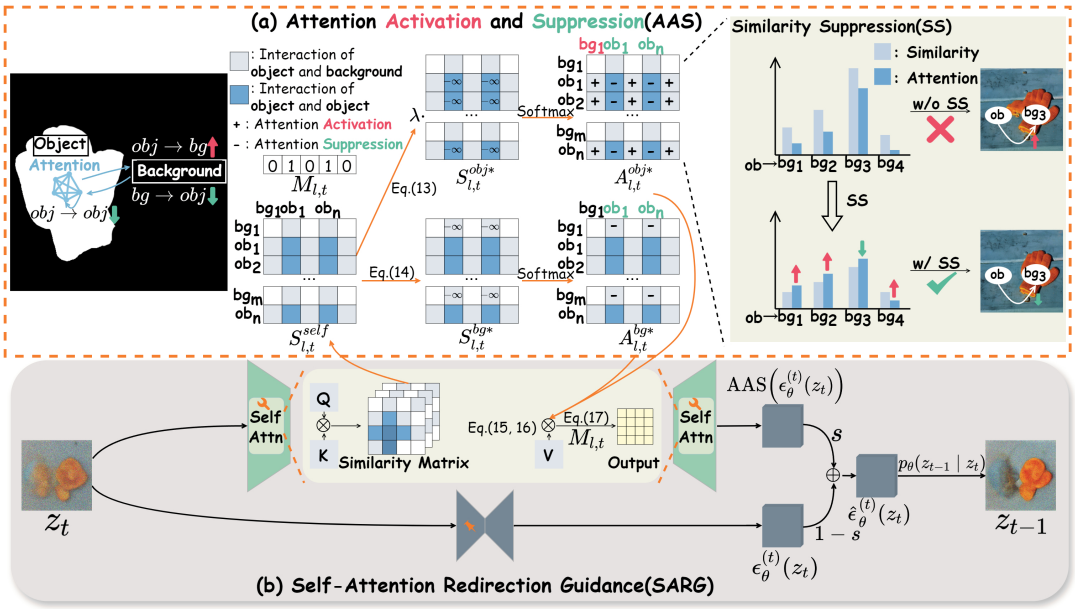

Figure 2: The overview of our proposed Attentive Eraser which consists of two parts: **(a) Attention Activation and Suppression (AAS)**, a self-attention mechanism modification operation tailored to address the challenges inherent to the object removal task, aims to make the foreground object area's generation more attentive to the background while erasing the object's appearance information. Additionally, Similarity Suppression (SS) serves to suppress the heightened attention to similar objects that may arise due to the inherent nature of self-attention. **(b) Self-Attention Redirection Guidance (SARG)**, a guidance method applied in the diffusion reverse sampling process, which utilizes redirected self-attention through AAS to guide the sampling process towards the direction of object removal.

To facilitate computation and adaptation, we regulate self-attention map $A_{l,t}^{self}$ corporally by changing the similarity matrix $S_{l,t}^{self}$. Specifically, suppose $M_{l,t} \in \mathbb{R}^{1\times N^2}$ is the corresponding flattened mask, among these N^2 tokens, we denote the set of tokens belonging to the foreground object region as $F_{l,t}^{obj}$ and the set of remaining tokens as $F_{l,t}^{bg}$. Correspondingly, $M_{l,t}$ can be expressed by the following equation:

$$M_{l,t}[i] = \begin{cases} 1, & i \in F_{l,t}^{obj} \\ 0, & i \in F_{l,t}^{bg}. \end{cases} \tag{10}$$

We define $S_{l,t}^{obj\to bg} = \left\{ S_{l,t}\left[i,j\right] | i \in F_{l,t}^{obj}, j \in F_{l,t}^{bg} \right\}$ to reflect the relevance of the content to be generated in the foreground object area to the background, while information about the appearance of the foreground object is reflected in $S_{l,t}^{obj\to obj} = \left\{ S_{l,t}\left[i,j\right] | i \in F_{l,t}^{obj}, j \in F_{l,t}^{obj} \right\}$. In the object removal task, we are dealing with foreground objects, and the background should remain the same. As shown in Figure 3, after DDIM inversion (Song, Meng, and Ermon 2020), we utilize PCA (Maćkiewicz and Ratajczak 1993) and clustering to visualize the average self-attention maps over all time steps for different layers during the reverse denoising process. It can be observed that self-attention maps resemble a semantic layout map of the components of the image (Yang et al. 2024a), and there is a clear distinction between the self-attention corresponding to the generation

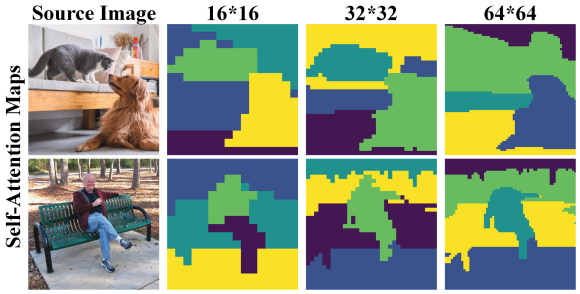

Figure 3: Visualization of the average self-attention maps over all time steps for different layers.

of the foreground object and background. Consequently, to facilitate object removal during the generation process, an intuitive approach would be to "blend" the self-attention of foreground objects into the background, thus allowing them to be clustered together. In other words, the region corresponding to the foreground object should be generated with a greater degree of reference to the background region than to itself during the generation process. This implies that the attention of the region within the mask to the background region should be increased and to itself should be decreased. Furthermore, the background region is fixed during the generation process and should remain unaffected by the changes in the generated content of the foreground area. Thus, the attention of the background region to the foreground region

should also be decreased.

Combining the above analysis, we propose an approach that is both simple and effective: **AAS** (as shown in Figure 2(a)). **Activation** refers to increasing $A_{l,t}^{obj\to bg}$, which serves to enhance the attention of the foreground-generating region to the background. In contrast, **Suppression** refers to decreasing $A_{l,t}^{obj\to obj}$ and $A_{l,t}^{bg\to obj}$, which entails the suppression of the foreground region's information about its appearance and its effect on the background. Given the intrinsic characteristics of the Softmax function, AAS can be simply achieved by assigning $S_{l,t}^{obj\to obj}$ to $-\infty$, thereby the original semantic information of the foreground objects is progressively obliterated throughout the denoising process. In practice, the aforementioned operation is achieved by the following equation:

$$S_{l,t}^{self*} = S_{l,t}^{self} - M_{l,t} * inf, \tag{11}$$

$$OP_{l,t}^{*} = A_{l,t}^{self*} V_{l,t} = \text{softmax}\left(S_{l,t}^{self*}\right) V_{l,t}, \tag{12}$$

where $V_{l,t}$ represents the corresponding value matrix for the time step t of layer l.

Nevertheless, one of the limitations of the aforementioned theory is that if the background contains content that is analogous to the foreground object, due to the inherent nature of self-attention, the attention in that particular part of the generative process will be higher than in other regions, while the above theory exacerbates this phenomenon, ultimately leading to incomplete object removal (see an example on the right side of Figure 2(a)). Accordingly, to reduce the attention devoted to similar objects and disperse it to other regions, we employ a straightforward method of reducing the variance of $S_{l,t}^{obj\to bg}$, which is referenced in this paper as **SS**. To avoid interfering with the process of generating the background, we address the foreground and background generation in separate phases:

$$S_{l,t}^{obj*} = \lambda S_{l,t}^{self} - M_{l,t} * inf, \tag{13}$$

$$S_{l,t}^{bg*} = S_{l,t}^{self} - M_{l,t} * inf, \tag{14}$$

$$OP_{l,t}^{obj*} = A_{l,t}^{obj*} V_{l,t} = \text{softmax}\left(S_{l,t}^{obj*}\right) V_{l,t}, \tag{15}$$

$$OP_{l,t}^{bg*} = A_{l,t}^{bg*} V_{l,t} = \text{softmax}\left(S_{l,t}^{bg*}\right) V_{l,t}, \tag{16}$$

where λ is the suppression factor less than 1. Finally, to guarantee that the aforementioned operations are executed on the appropriate corresponding foreground and background regions, we integrate the two outputs $OP_{l,t}^{obj*}$ and $OP_{l,t}^{bg*}$ to obtain the final output $OP_{l,t}^{*}$ according to $M_{l,t}^{\top}$:

$$OP_{l,t}^{*} = M_{l,t}^{\top} \odot OP_{l,t}^{obj*} + \left(1 - M_{l,t}^{\top}\right) \odot OP_{l,t}^{bg*}, \tag{17}$$

To ensure minimal impact on the subsequent generation process, we apply SS at the beginning of the denoising process timesteps, for $t \in [T_I, T_{SS}]$, and still use Eq.(11), Eq.(12) to get output $OP_{l,t}^{*}$ for $t \in (T_{SS}, 1]$, where T_I denotes the diffusion steps and T_{SS} signifies the final time-step of SS. In the following, we denote the U-net processed by the AAS approach as AAS (ϵ_θ).

Self-Attention Redirection Guidance

To further enhance the capability of object removal as well as the overall quality of the generated images, inspired by PAG (Ahn et al. 2024), AAS (ϵ_θ) can be seen as a form of perturbation during the epsilon prediction process, we can use it to steer the sampling process towards the desirable direction. Therefore, the final predicted noise $\hat{\epsilon}_\theta^{(t)}(z_t)$ at each time step can be defined as follows:

$$\hat{\epsilon}_\theta^{(t)}(z_t) = \epsilon_\theta^{(t)}(z_t) + s\left(\text{AAS}\left(\epsilon_\theta^{(t)}(z_t)\right) - \epsilon_\theta^{(t)}(z_t)\right), \tag{18}$$

where s is the removal guidance scale. Subsequently, the next time step output latent z_{t-1} is obtained by sampling using the modified noise $\hat{\epsilon}_\theta^{(t)}(z_t)$. In this paper, we refer to the aforementioned guidance process as **SARG**.

Through the iterative inference guidance, the sampling direction of the generative process will be altered, causing the distribution of the noisy latent to shift towards the object removal direction we have specified, thereby enhancing the capability of removal and the quality of the final generated images. For a more detailed analysis refer to Appendix A.

Experiments

Experimental Setup

Implementation Details We apply our method on all mainstream versions of Stable Diffusion (1.5, 2.1, and XL1.0) with two prevailing diffusion-based inpainting pipelines (Couairon et al. 2023; Avrahami, Fried, and Lischinski 2023) to evaluate its generalization across various diffusion model architectures. Based on the randomness, we refer to pipelines as the stochastic inpainting pipeline (SIP) and the deterministic inpainting pipeline (DIP), respectively. Detailed descriptions of SIP and DIP are provided in Appendix B, with further experimental details available in Appendix C.

Baseline We select the state-of-the-art image inpainting methods as our baselines, including two mask-guided approaches SD-Inpaint (Rombach et al. 2022), LAMA (Suvorov et al. 2022) and two text-guided approaches Inst-Inpaint (Yildirim et al. 2023), Powerpaint (Zhuang et al. 2023), to demonstrate the efficacy of our method, we have also incorporated SD2.1 with SIP into the baseline for comparative purposes.

Testing Datasets We evaluate our method on a common segmentation dataset OpenImages V5 (Kuznetsova et al. 2018), which contains both the mask information and the text information of the corresponding object of the mask. This facilitates a comprehensive comparison of the entire baseline. We randomly select 10000 sets of data from the OpenImages V5 test set as the testing datasets, a set of data including the original image and the corresponding mask, segmentation bounding box, and segmentation class labels.

Evaluation Metrics We first use two common evaluation metrics **FID** and **LPIPS** to assess the quality of the generated images following LAMA(Suvorov et al. 2022) setup, which can indicate the global visual quality of the image.

Method	Training	Mask	Text	FID↓	LPIPS↓	**Local FID↓**	**CLIP consensus↓**	**CLIP score↑**
SD2.1inp	✓	✓	✗	**3.805**	0.3012	8.852	0.1143	21.89
SD2.1inp	✓	✓	✓	4.019	0.3083	7.194	0.1209	22.27
PowerPaint	✓	✓	✗	6.027	0.2887	10.02	0.0984	22.74
Inst-Inpaint	✓	✗	✓	11.42	0.4095	43.47	0.0913	23.02
LAMA	✓	✓	✗	7.533	**0.2189**	6.091	-	**23.57**
SD2.1+SIP w/o SARG	✗	✓	✗	5.98	0.2998	15.58	0.1347	22.05
SD2.1+SIP w/ SARG(ours)	✗	✓	✗	7.352	0.3113	5.835	**0.0734**	23.56
SD2.1+DIP w/ SARG(ours)	✗	✓	✗	7.012	0.2995	**5.699**	-	23.43

Table 1: Quantitative comparison with other methods. We have indicated in the table whether each method requires training and whether it necessitates mask or prompt text as conditional inputs. In the CLIP consensus metric, deterministic process methods (lacking randomness) are denoted with a '-'. The optimal result and object removal-related metrics are represented in bold, and the sub-optimal result is represented in underlining.

Figure 4: Visual comparison with other methods. The mask is indicated with a red highlight in the input image. Our methods are highlighted in bold.

Figure 5: Visual comparison of object removal stability with other methods using three distinct random seeds.

To further assess the quality of the generated content in the mask region, we adopt the metrics **Local-FID** to assess the local visual quality of the image following (Xie et al. 2023a). To assess the effectiveness of object removal, we select **CLIP consensus** as the evaluation metric following (Wasserman et al. 2024), which enables the evaluation of the consistent diversity of the removal effect. High diversity is often seen as a sign of failed removal, with random objects appearing in the foreground area. Finally, to indicate the degree of object removal, we calculate the **CLIP score** (Radford et al. 2021) by taking the foreground region patch and the prompt "background". The greater the value, the greater the degree of alignment between the removed region and the background, effectively indicating the degree of removal.

Qualitative and Quantitative Results

The quantitative analysis results are shown in Table 1. For global quality metrics FID and LPIPS, our method is at an average level, but these two metrics do not adequately reflect the effectiveness of object removal. Subsequently, we can observe from the local FID that our method has superior performance in the local removal area. Meanwhile, the CLIP

Method	User Study	GPT Evaluation
SD2.1inp	10%	-
SD2.1inp(w/ text)	15.4%	-
PowerPaint	7.6%	-
Inst-Inpaint	2.4%	-
LAMA	19.7%	25.53%
SD2.1+SIP w/ SARG(ours)	**44.9%**	**74.47%**

Table 2: User study and GPT-4o Evaluation results.

consensus indicates the instability of other diffusion-based methods, and the CLIP score demonstrates that our method effectively removes the object and repaints the foreground area that is highly aligned with the background, even reaching a competitive level with LAMA, which is a Fast Fourier Convolution-based inpainting model. Qualitative results are shown in Figure 4, where we can observe the significant differences between our method and others. LAMA, due to its lack of generative capability, successfully removes the object but produces noticeably blurry content. Other diffusion-based methods share a common issue: the instability of removal, which often leads to the generation of random artifacts. To further substantiate this issue, we conducted experiments on the stability of removal. Figure 5 presents the results of removal using three distinct random seeds for each method. It can be observed that our method achieves stable erasure across various SD models, generating more consistent content, whereas other methods have struggled to maintain stable removal of the object.

User Study and GPT-4o Evaluation

Due to the absence of effective metrics for the object removal task, the metrics mentioned above may not be sufficient to demonstrate the superiority of our method. Therefore, to further substantiate the effectiveness of our approach, we conduct a user preference study. Table 2 presents the user preferences for various methods, revealing consistent results with the quantitative results and highlighting that our method is strongly preferred over other methods. Furthermore, we design fairly and reasonably prompts, utilizing GPT-4o (OpenAI 2024) to conduct a further assessment of object removal performance between our method and the runner-up method LAMA. The results also indicate that our method significantly outperforms LAMA, demonstrating exceptional performance. Please refer to Appendix D for more details and visualizations of user study and GPT evaluation.

Ablations

To validate the effectiveness of the proposed Attentive Eraser, we conduct ablation studies. We use SD2.1 with SIP as the baseline for comparison, Figure 6 provides a visual representation of the ablation study concerning our method's components. Figure 6(a) shows that the application of AAS alone cannot completely remove the foreground object, but integrating it with the sampling process through SARG can

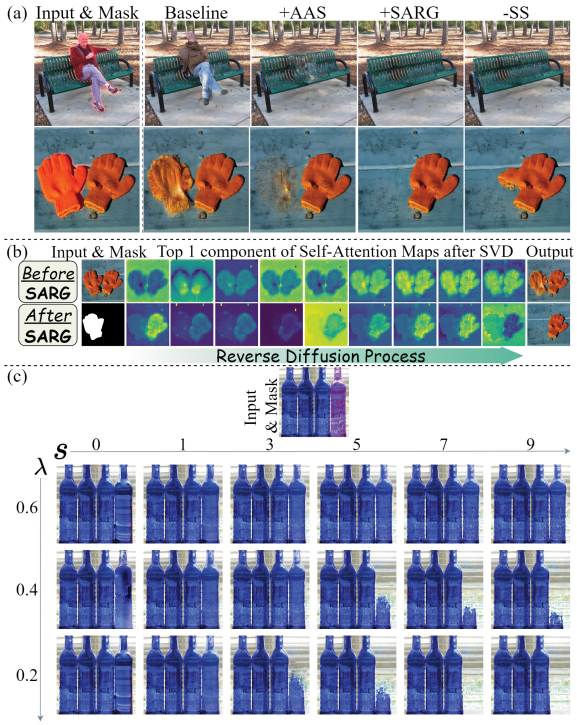

Figure 6: Visualization of ablation experiments on Attentive Eraser.

effectively remove the object and generate content consistent with the background. At the same time, we also verify the impact of SS, and it can be seen that SS effectively suppresses the generation of similar objects while maintaining the removal efficacy of the general image. As shown in In Figure 6(b), we visualize the heatmaps of the top-1 component of the self-attention maps at each step of the denoising process after SVD (Kalman 1996), demonstrating that SARG gradually, as previously stated, "blends" the foreground objects' self-attention into the background to remove objects. In Figure 6(c), we discuss the effect of two parameters (removal guidance s and suppression factor λ) upon the removal process. It is depicted that as λ decreases, the generation of similar objects decreases progressively, thereby reaffirming the efficacy of SS. On the other hand, the intensity of the removal process escalates with an increase in s. This suggests that s acts as a pivotal control in modulating the strength of the removal, allowing for a more nuanced and tailored approach to removing objects.

Conclusion

We present a novel tuning-free method Attentive Eraser, which adeptly harnesses the rich repository of prior knowledge embedded within pre-trained diffusion models for the object removal task. Extensive experiments and user studies demonstrate the stability, effectiveness, and scalability of our proposed method, and also reveal that our method significantly outperforms existing methods.

Acknowledgments

This work is supported in part by the Summit Advancement Disciplines of Zhejiang Province (Zhejiang Gongshang University - Statistics) and "Digital+" discipline construction management project of Zhejiang Gongshang University (SZJ2022C011).

References

Ahn, D.; Cho, H.; Min, J.; Jang, W.; Kim, J.; Kim, S.; Park, H. H.; Jin, K. H.; and Kim, S. 2024. Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance. *arXiv preprint arXiv:2403.17377*.

Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended latent diffusion. *ACM TOG*, 42(4): 1–11.

Bansal, A.; Chu, H.-M.; Schwarzschild, A.; Sengupta, S.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Universal guidance for diffusion models. In *CVPR*, 843–852.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 18392–18402.

Chen, Y.; and Hu, H. 2019. An improved method for semantic image inpainting with GANs: Progressive inpainting. *Neural Processing Letters*, 49: 1355–1367.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2023. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *ICLR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *NeurIPS*, 36: 16222–16239.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024a. Not all inputs are valid: Towards open-set video moment retrieval using language. In *ACM MM*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024b. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *AAAI*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023c. Hierarchical local-global transformer for temporal sentence grounding. *IEEE TMM*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2024c. Rethinking weakly-supervised video temporal grounding from a game perspective. In *ECCV*.

Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Li, H.; Hu, H.; et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *CVPR*, 12709–12720.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Guo, Q.; Gao, S.; Zhang, X.; Yin, Y.; and ming Zhang, C. 2018. Patch-Based Image Inpainting via Two-Stage Low Rank Approximation. *TVCG*, 24: 2023–2036.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hong, S.; Lee, G.; Jang, W.; and Kim, S. 2023. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 7462–7471.

Huang, Y.; Xie, L.; Wang, X.; Yuan, Z.; Cun, X.; Ge, Y.; Zhou, J.; Dong, C.; Huang, R.; Zhang, R.; et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 8362–8371.

Kalman, D. 1996. A singularly valuable decomposition: the SVD of a matrix. *The college mathematics journal*, 27(1): 2–23.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*.

Li, S.; Zeng, B.; Feng, Y.; Gao, S.; Liu, X.; Liu, J.; Li, L.; Tang, X.; Hu, Y.; Liu, J.; et al. 2024. Zone: Zero-shot instruction-guided local editing. In *CVPR*, 6254–6263.

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In *CVPR*, 7817–7826.

Liu, Y.; Zhang, J.; Peng, D.; Huang, M.; Wang, X.; Tang, J.; Huang, C.; Lin, D.; Shen, C.; Bai, X.; et al. 2023. Spts v2: single-point scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lu, H.; Liu, Q.; Zhang, M.; Wang, Y.; and Deng, X. 2018. Gradient-based low rank method and its application in image inpainting. *Multimedia Tools and Applications*, 77: 5969–5993.

Maćkiewicz, A.; and Ratajczak, W. 1993. Principal components analysis (PCA). *Computers & Geosciences*, 19(3): 303–342.

Mo, S.; Mu, F.; Lin, K. H.; Liu, Y.; Guan, B.; Li, Y.; and Zhou, B. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, 7465–7475.

Nam, J.; Kim, H.; Lee, D.; Jin, S.; Kim, S.; and Chang, S. 2024. Dreammatcher: Appearance matching self-attention

for semantically-consistent text-to-image personalization. In *CVPR*, 8100–8110.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Oleksii, S. 2019. Deep hyperspectral prior: Denoising, inpainting, super-resolution. *CoRR*, 1902.

OpenAI. 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-7-10.

Patashnik, O.; Garibi, D.; Azuri, I.; Averbuch-Elor, H.; and Cohen-Or, D. 2023. Localizing object-level shape variations with text-to-image diffusion models. In *ICCV*, 23051–23061.

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 *ICML*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *NeurIPS*, 29.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; and Zeng, H. 2023. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 32: 1039–1051.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024b. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *ICLR*.

Shin, Y.-G.; Sagong, M.-C.; Yeo, Y.-J.; Kim, S.-W.; and Ko, S.-J. 2020. Pepsi++: Fast and lightweight network for image inpainting. *TNNLS*, 32(1): 252–265.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*. OpenReview.net.

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 *WACV*, 2149–2159.

Tang, J.; Zhang, W.; Liu, H.; Yang, M.; Jiang, B.; Hu, G.; and Bai, X. 2022. Few could be better than all: Feature sampling and grouping for scene text detection. In *CVPR*, 4563–4572.

Wasserman, N.; Rotstein, N.; Ganz, R.; and Kimmel, R. 2024. Paint by Inpaint: Learning to Add Image Objects by Removing Them First. *arXiv preprint arXiv:2404.18212*.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023a. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 22428–22437.

Xie, S.; Zhao, Y.; Xiao, Z.; Chan, K. C.; Li, Y.; Xu, Y.; Zhang, K.; and Hou, T. 2023b. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. *arXiv preprint arXiv:2312.03771*.

Xu, R.; Dong, X.-M.; Li, W.; Peng, J.; Sun, W.; and Xu, Y. 2024. DBCTNet: Double branch convolution-transformer network for hyperspectral image classification. *TGRS*.

Yan, Z.; Li, X.; Li, M.; Zuo, W.; and Shan, S. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, 1–17.

Yang, F.; Yang, S.; Butt, M. A.; van de Weijer, J.; et al. 2024a. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *NeurIPS*, 36.

Yang, S.; Zhang, L.; Ma, L.; Liu, Y.; Fu, J.; and He, Y. 2023. Magicremover: Tuning-free text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2310.02848*.

Yang, Y.; Peng, H.; Shen, Y.; Yang, Y.; Hu, H.; Qiu, L.; Koike, H.; et al. 2024b. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *NeurIPS*, 36.

Yildirim, A. B.; Baday, V.; Erdem, E.; Erdem, A.; and Dundar, A. 2023. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *CVPR*, 5505–5514.

Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; and Xie, Y. 2024a. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15567–15576.

Zhao, Z.; Tang, J.; Wu, B.; Lin, C.; Wei, S.; Liu, H.; Tan, X.; Zhang, Z.; Huang, C.; and Xie, Y. 2024b. Harmonizing Visual Text Comprehension and Generation. *arXiv preprint arXiv:2407.16364*.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2023. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*.