Towards Small Object Editing: A Benchmark Dataset and A Training-Free Approach

Figure 1: We introduce SOEBench, a standardized benchmark for quantitatively evaluating text-based small object editing.

ABSTRACT

A plethora of text-guided image editing methods has recently been developed by leveraging the impressive capabilities of large-scale diffusion-based generative models especially Stable Diffusion. Despite the success of diffusion models in producing high-quality images, their application to small object generation has been limited due to difficulties in aligning cross-modal attention maps between text and these objects. Our approach offers a training-free method that significantly mitigates this alignment issue with local and global attention guidance , enhancing the model's ability to accurately render small objects in accordance with textual descriptions. We detail the methodology in our approach, emphasizing its divergence from traditional generation techniques and highlighting its advantages. What's more important is that we also provide *SOEBench* (Small Object Editing), a standardized benchmark for quantitatively evaluating text-based small object generation collected from *MSCOCO*[22] and *OpenImage*[18]. Preliminary results demonstrate the effectiveness of our method, showing marked improvements in the fidelity and accuracy of small object generation compared to existing models. This advancement not only contributes to the field of AI and computer vision but also opens up new possibilities for applications in various industries where precise image generation is critical.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Small Object Editing, Benchmark, Cross-Attention Guidance

ACM Reference Format:

. 2024. Towards Small Object Editing: A Benchmark Dataset and A Training-Free Approach. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

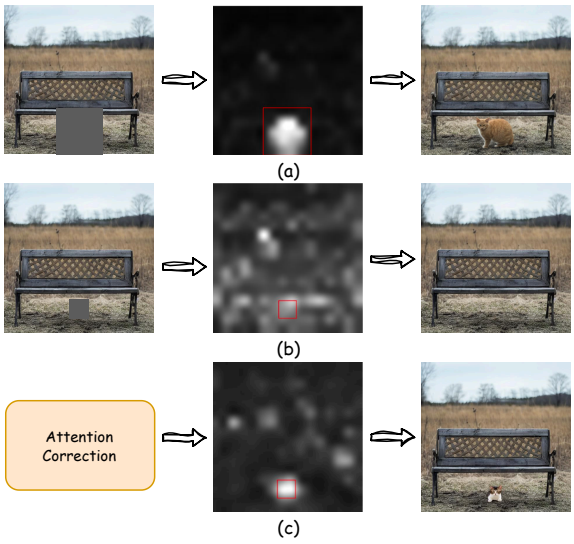

Figure 2: Comparison between our multi-scale joint attention guidance method and traditional text-based image inpainting methods. Traditional text-based image inpainting methods can only edit objects with a large scale. (a) Fail to inpaint objects with a small scale. (b) Our method multi-scale joint attention guidance method can efficiently tackle the small object editing task (c) After joint attention guidance for attention correction, we could obtain the refined cross-attention map for accurate small object editing.

1 INTRODUCTION

The realm of text-to-image generation has witnessed tremendous advancements with the advent of recent diffusion models[12, 25, 27, 29, 31], which have successfully revolutionized various tasks, including photo editing[15, 36], and inpainting[2, 30]. These models have demonstrated remarkable capabilities in producing and manipulating salient objects within a picture, like the main subject of a picture, under the description guidance. The success of such models can be attributed to the effective cross-modal feature alignment between textual descriptions and the corresponding visual objects during the synthesis process. Such alignment facilitates a coherent and accurate translation of textual descriptions into visual representations, resulting in impressive outcomes in both image editing and in-painting tasks.

However, a severe issue arises when the pending object is small. As shown in Fig. 2 (b), using small masks to guide the target object editing can lead to problems such as attribute leakage, poor quality, and missing entities. This issue is primarily due to the model's inability to focus on such a small region of interest in the description, which can result in the model's inability to generate objects that align well with the textual descriptions. For example, consider an image with a size of 512×512 and a small object whose bounding box typically occupies only 64×64 pixels. When performing multi-level cross-modal feature alignment, with U-Net as the backbone, the cross-attention map is progressively down-sampled to a small resolution of 8×8 as the network deepens. At the same time, our target region may encompass within only a 1×1 grid, which is too small for the model to effectively focus on. Consequently, the model may struggle to generate objects that align well with the textual descriptions within such a small area, leading to poor quality and even missing entities.

Facing the abovementioned limitations, we introduce a new task called Small Object Editing (SOE) in this work. Specifically, the SOE task requires the model to perform editing that is consistent with the given textual descriptions, seamlessly integrates with the surrounding context of the original image, and is precise in the desired small region. The key to this task is to prevent mismatches between the desired masked region and the intended textual description, ensuring that the model generates accurate and high-quality small objects. To evaluate the performance of models on this SOE task, we construct a comprehensive benchmark dataset, *i.e.*, SOEBench. This benchmark allows us to evaluate the effectiveness of small object editing on various models and includes 4000 objects from two established datasets, MSCOCO [22] and OpenImages [18]. SOEBench holds two sets, SOE-2k and SOG-4k, where SOE-2k contains 2000 objects from OpenImages for editing and SOE-4k contains an additional 2000 objects from MSCOCO for editing. For each small object editing prompt, we provide both label-only and label-with-color templates in the description. This allows us to evaluate the models' ability to generate small objects that align with both textual descriptions and color specifications.

Building upon the constructed SOEBench, we further provide a strong baseline method for small object editing. As discussed above, the quality of the cross-attention map is crucial in small object editing. To this end, we propose a new joint attention guidance method to enhance the accuracy of the cross-attention map alignment from both local and global perspectives. In particular, we first develop a local attention guidance strategy to enhance the foreground cross-attention map alignment and then introduce a global attention guidance strategy to enhance the background cross-attention map alignment. Our proposed baseline method is training-free but highly effective in addressing the SOE problem.

The contributions of our work are summarized as below:

- By recognizing the limitations of current research, we define a new task of small object editing and come up with a new benchmark dataset SOEBench to evaluate the model's ability on small object editing.
- We introduce a novel training-free baseline approach to address the small object editing challenge by considering both global and local perspectives.
- The proposed method is evaluated on the proposed SOEBench dataset and achieves the state-of-the-art results.

2 RELATED WORK

2.1 Diffusion Models

Denoising diffusion based models (DMs)[12, 25, 27, 29, 31] and score-matching-based model[32–34] have become the *de facto* standard for image generation, surpassing Generative Adversarial Networks (GANs)[9, 35, 38, 39, 41] in performance and training stability. Denoising Diffusion Probabilistic Models (DDPM)[12] and Noise

Conditional Score Networks (NCSN)[31] are the pioneer to employ denoising neural networks to reverse a predefined Markovian noising process on raw images.

Latent Diffusion Models (LDM)[29] intuitively disentangle image synthesis into two stages: semantic reconstruction through a denoising process in latent space, and perceptual reconstruction with VAE[17], such a strategy significantly enhances visual fidelity. Currently, the success of DMs has greatly driven a series of image generation and editing tasks.

2.2 Text-based Image Editing

Image editing has become a popular domain, enabling personalized and customized image generation. Text prompts are predominantly employed as input conditions due to their flexibility. These methods can generally be classified into two categories: training-based and training-free. Training-based methods typically involve fine-tuning the entire denoising model or adapters. For example, DiffusionCLIP [16] and InstructPix2Pix [4] fine-tune the entire diffusion model using collected text-image datasets. In contrast, Asyrp [19] and GLIGEN [21] introduce adapters into publicly available diffusion models and only tune the adapter parameters for rapid adaptation. T2I-Adapter [24] and ControlNet [40] further extend control capabilities by introducing tailored adapter architectures such as zero convolution.

For training-free methods, manipulating cross-attention maps and fine-tuning text embeddings are mainstream techniques. Prompt-to-prompt [10] edits images solely through text prompts, injecting attention maps of the original image throughout the diffusion process. Attend-and-Excite [5] introduces GSN to intervene in the diffusion process to enhance subject tokens for generating multiple objects. DDIM Inversion [31] presents a deterministic process to invert images to noises. Inspired by DDIM Inversion, Null-Text Inversion [23] pioneers fine-tuning null-text embeddings to reduce the distance between the sampling trajectory and the inversion trajectory. Prompt Tuning Inversion [8] follows a two-stage pipeline, first encoding the image into a learnable embedding and then sampling the edited image by interpolating the target embedding and the learnable embedding. PnP Inversion [14] enhances the inversion process by disentangling source and target branches for content preservation and edit fidelity, respectively. Imagic [15] incorporates text embedding optimization and model fine-tuning for image editing. BlendedDiffusion [1, 2] explore the method of using a mask to edit the specific region and add a new object to the image while leaving the rest unchanged. Some layout2image methods[6, 7, 26, 37] operated constraints on cross-attention to control the synthetic contents.

Most text-based image editing methods are based on DMs, while a U-Net architecture is widely adopted. However, as the U-Net deepens, the cross-attention map becomes extremely small, thus some small objects may only take a small region on the attention map, leading to the model hardly focusing on such small objects for editing.

2.3 Benchmarks and Evaluation for Image Editing

Although image editing has been widely explored in recent years[3, 13–15, 36], benchmarks for this task remain limited. EditBench [36] primarily presents a systematic benchmark comprising 240 images for image inpainting based on masks and text prompts, evaluated using CLIP-Score[11, 28] and CLIP-R-Precision metrics. Concurrently, Kawar *et al.* introduce TedBench [15] to evaluate non-rigid text-based image editing, providing 100 pairs of input images and target texts, with CLIP-Score and 1-LPIPS adopted as metrics. EditVal [3] introduces a standardized evaluation protocol for assessing multiple edit types using a curated image dataset. Additionally, Ju *et al.* present PIE-Bench [14], comprising 700 images showcasing diverse scenarios and editing types. Aside from automatic evaluation via benchmarks, human evaluation (*a.k.a.* user study) is a more reliable method for assessing the quality of generated images. *Text Alignment* and *Image Quality* are two main metrics used in these studies. *Text Alignment* focuses on the consistency between the text prompt and the edited image, while *Image Quality* evaluates the visual fidelity of the generated image.

However, to the best of our knowledge, there is no benchmark so far focusing on small object editing, which is a critical issue failed by most of the current research. Therefore, in this work, we focus on small object editing and propose a new small object editing benchmark, *i.e.*, SOEBench.

3 PRELIMINARIES

3.1 Latent Diffusion Models

Latent diffusion models are a subclass of generative models that adopt a diffusion and a denoising process to synthesize new data. First, a VAE encoder is adopted to encode the image $\mathcal{I}$ into the latent space, denoted as z. Then the forward diffusion process gradually introduces noise to the latent representation z to the complete Gaussian noise z_T, where T represents the total number of timesteps. For each timestep t, the noisy latent code z_t can be represented as:

$$z_t \sim \mathcal{N}(\alpha(t)z; \sigma^2(t)\mathbf{I}), \tag{1}$$

where $\alpha(t)$ and $\sigma(t)$ control the mean and covariance of the noise. Then, the diffusion models aim to reverse this diffusion process by sampling random Gaussian noise z_T and gradually denoising such latent code to generate the initial latent code z_0. In practice, the denoising model targets at predicting the sampled noise ϵ_θ at each timestep t given the condition c by optimizing the Mean Square Error between the predicted sampled noise $\epsilon_\theta(z_t, c, t)$ and the sampled noise ϵ, which can be formulated as:

$$\mathcal{L} = \nabla_\theta \left\| \epsilon - \epsilon_\theta \left(z_t, c, t \right) \right\|^2 . \tag{2}$$

Finally, a VAE decoder is adopted to generate the image using the denoised latent code.

3.2 Cross-Attention Guidance

The Latent Stable Diffusion models perform conditional generation utilizing a cross-modal attention module between the given conditioning c and the latent representations z_t, where the condition embedding c and the latent representations z_t are mapped to the query $\mathbf{Q}$ with a dimension of $M \times d$ and the key $\mathbf{K}$ with a dimension

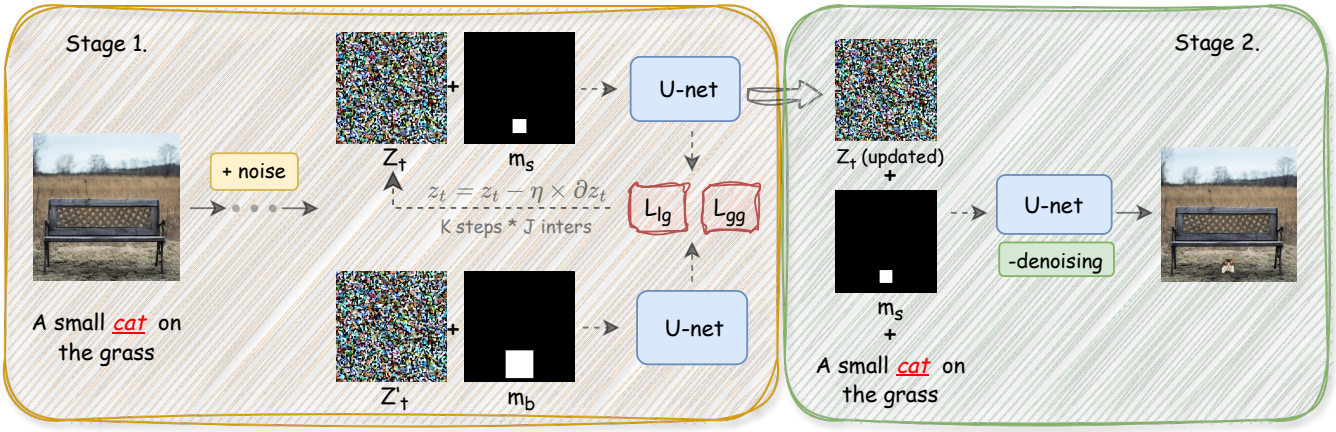

Figure 3: The inference pipeline of our proposed *SOE*. We receive smaller mask region m_s , larger mask region $m_b$then randomly initialize two identical z_T and z'_T, text prompt c_t as input. During the first K timesteps, we compute the cross-attention maps from both parts J times at each timestep and calculate the L_{lg} and L_{gg} losses. Then, based on the loss gradients, we backpropagate to optimize z_t. After K*J rounds of optimization, we continue to use the diffusion model to denoise the image.

of $N \times d$, and the cross-attention map $\mathcal{A}^l$ in the l-th layer of the U-Net can be derived as:

$$\mathcal{A}^l = Softmax\left(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{d}}\right) \in [0,1]^{M \times N}. \tag{3}$$

Attention control is widely used in image editing, notably in techniques such as Prompt-to-Prompt(P2P). P2P replaces the current attention map with one corresponding to the target text, thereby facilitating the editing of objects within the image.

4 THE SOEBENCH DATASET

SOEBench is selectively extracted from MSCOCO[22] and OpenImage[18] databases, the selection rule of which follows a set of precise criteria to aptly serve the needs of small object generation experiments. The selection process primarily focuses on ensuring that the chosen objects are not obstructed by other elements within the image, allowing for a clear target for the small object generation task. Additionally, the object size was a critical consideration. We specifically selected objects that occupy a size smaller than 1/6 but larger than 1/8 of the overall image. This size constraint is pivotal because objects smaller than 1/8 would yield a representation on the deepest U-Net's feature map that is smaller than one pixel, rendering them practically unfeasible for effective generation, as shown in Fig. 4.

The category of the data contained in the dataset can be referred to in the word cloud in Fig 1. Our dataset comprises about 300 types of objects that are frequently encountered, aligning the experimental setup with real-world object recognition scenarios. We then crop the images to extract the portions within the masked areas, and then use the BLIP-VQA [20] model to query, 'What is the primary color of the object in this area?' This approach efficiently identifies the main color of the objects within the masks as the color attribute. Importantly, the dataset has been segmented into two subsets based on the quantity of images they contain: *SOE-2k* and *SOE-4k*, where *SOE-2k* contains 2000 objects from the OpenImage dataset and *SOE-4k* contains 2000 more objects from

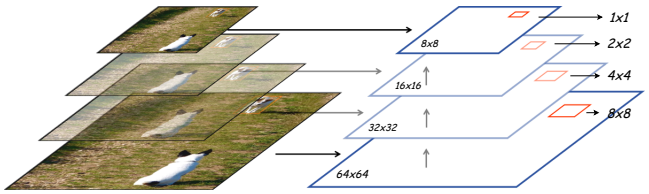

Figure 4: The illustration of a cross-attention map. For an input image with a size of 512×512, if the masked area is 64×64, the corresponding effective area comes to 1×1 in the mid-block. The diminutive size of this masked area poses a challenge as it may lack sufficient semantic information essential for generating associated objects.

the MSCOCO dataset. Such an operation enables a thorough and diverse assessment, ensuring that our methodology is tested under different scales of data availability, further reinforcing the validity and robustness of our research findings.

Table 1: Comparing our benchmark with existed works from different aspects, such as object category, benchmark size and mask size.

Benchmarks	Benchmark Info		
	categoty num	dataset size	mask size
TedBench	<50	100	>0.5
EditBench	<50	240	0.08~0.9
EditVal	<50	648	-
SOE(Ours)	300	2K/4K	<0.03

Compared to existing benchmarks for text-based image editing (*i.e.,* TeDBench [15], EditBench [36], and EditVal[3]), our benchmark, SOEBench, offers distinct advantages. TeDBench is limited by its small dataset, consisting of only 100 images across 40 categories, primarily focused on modifying object attributes, states, or

appearances. Furthermore, TeDBench images typically feature a single dominant subject occupying a significant portion of the image. In contrast, EditBench includes a larger dataset of 240 images and accommodates a wider range of mask sizes, from 0.08 to 0.9 of the image size. However, even EditBench's smallest mask sizes are substantially larger than those in SOEBench, where the maximum size of small object areas occupies only $\frac{1}{36}$ of the image. Another recent benchmark, EditVal[3], offers a more comprehensive set of image-editing operations. Developed from the MSCOCO [22] dataset, EditVal comprises 648 unique image-editing operations across 19 classes. These operations include 13 types of real-world edits, including one dedicated to controlling object size. Comparing SOEBench with these benchmarks (summarized in Tab. 1), we observe that SOEBench offers a larger and more diverse test dataset, covering a wider array of categories. Moreover, SOEBench features significantly smaller target areas, presenting a greater challenge for text-based image editing tasks.

5 METHODOLOGY

In this section, we present our training-free multi-scale joint attention guidance baseline in detail. In particular, we first describe the problem definition of small objection editing in Sec. 5.1. Then, we introduce our local attention guidance in Sec. 5.2 and our global attention guidance in Sec. 5.3.

5.1 Problem Definition

In this paper, we propose a novel small object editing (SOE) task that targets at in-painting the specified masked region m_s on the given image I, which is usually small, using a textual description c as the condition. The task requires the model Ψ, which includes a VAE encoder $\mathcal{E}$, a diffusion model with weight θ from which we can obtain the predicted noise ϵ and cross-attention map $\mathcal{A}$, and a VAE decoder $\mathcal{D}$, to not only perform inpainting consistent with the given condition but also perform the in-painting that integrates seamlessly with the surrounding context of the original image. Following previous works, a L-layer U-Net is adopted as the backbone of the diffusion model. The key to the task is to avoid any mismatch between the desired masked region and the intended textual description, which is usually achieved by refining the cross-attention map $\mathcal{A}$.

In this paper, we aim at proposing a training-free baseline method to tackle the SOE problem. Based on the L-layer U-Net architecture, we focus on cross-attention map refinement and propose a local and global attention guidance method, where we encourage the refined cross-attention map to provide precise guidance to the latent code estimation. The overall inference pipeline is shown in Fig. 3.

5.2 Local Attention Guidance

Given the condition token c, the cross-attention map corresponding to the i-th condition token and the location of the mask region, which we denote the position index on the feature map as j, can be denoted as $\mathcal{A}^l_{i,j}$, as discussed above, such an $\mathcal{A}^l_{i,j}$ can take only a small region, which can lead to a mismatch problem easily. Therefore, we aim at improving the accuracy of location and the content of $\mathcal{A}^l_{i,j}$ and propose a target region scaling-and-rescaling operation to get a precise cross-attention map. In particular, we upscale the mask region m_s to a larger mask region m_b by a scaling factor s. The cross-attention maps $\mathcal{A}^l$ and $\mathcal{A}'^l$ obtained by using m_s and m_b as the input mask region respectively can be represented as:

$$\mathcal{A}_l \leftarrow \mathcal{F}_\theta(z_t, t, c, m_s, z), \tag{4}$$

and

$$\mathcal{A}'_l \leftarrow \mathcal{F}_\theta(z'_t, t, c, m_b, z), \tag{5}$$

where $\mathcal{F}_\theta$ is the cross-attention map obtainment operation, where we collect the cross-attention maps from the diffusion model, z'_t is the noised latent code predicted by using m_b as the input, which is initialized by $z'_T = z_T$.

Based on our observation, the model is prone to generate accurate contents within such a large region. However, such an operation may fail to handle the seamless integration with the surrounding context of the original image problem. Therefore, we further rescale $\mathcal{A}'^l$ and encourage such a precise cross-attention map to provide guidance for $\mathcal{A}^l$, which can be represented as:

$$\mathcal{L}_{lg} = \sum_{l=1}^{L} \sum_{i \in \{c_{req}\}} \left\| I(r_{m_b}(\mathcal{A}'^l_i)) - r_{m_s}(\mathcal{A}^l_i) \right\|_2, \tag{6}$$

where $r(\cdot)$ is the desired mask region localization operation and $I(\cdot)$ is the interpolation downscaling operation to make sure that $I(r_{m_b}(\mathcal{A}'^l_i))$ and $r_{m_s}(\mathcal{A}^l_i)$ are with the same size. Recall that the subscript i denotes the i-th condition token and c_{req} denotes the corresponding object index in the given textual condition c. It should be mentioned that Eq. 6 is designed solely to modify $\mathcal{A}^l$ while $\mathcal{A}'^l$ remains unchanged, thus obviating the need for a back-propagation operation.

5.3 Global Attention Guidance

Our local cross-attention guidance approach is training-free but boosts the accuracy of the desired region in the cross-attention map, which shows its great potential in tackling the SOE problem. However, as shown in Fig. 5, as the desired mask region may be too small, the pixels on the cross-attention map may be attracted to the wrong region, like the background. To tackle the issue, in this section, we further propose a global attention guidance operation to alleviate the foreground noise, which guarantees that when utilizing m_s to produce a cross-attention map, the map accurately zeroes in on pertinent information within the masked areas associated with the object being generated, effectively discounting data from areas inconsequential to the textual depiction of the object.

In particular, based on $\mathcal{A}'^l$ introduced in Sec.5.3, we encourage $\mathcal{A}'^l$ to provide global guidance for $\mathcal{A}^l$ to alleviate the problem of cross-attention matching to the wrong region. Different from local attention guidance that only focuses on the cross-attention map on the desired mask region alignment, we encourage $\mathcal{A}^l$ to be aligned with $\mathcal{A}'^l$ for all queries, which can be formulated as

$$\mathcal{L}_{gg} = \sum_{l=1}^{L} \sum_{i=1}^{M} \sum_{j=1}^{N} \left\| \mathcal{A}'^l_{i,j} - \mathcal{A}^l_{i,j} \right\|_2. \tag{7}$$

Recall that i is the index for the i-th condition token, j is the index for the j-th region index on the image, and M and N represent the dimension of the query Q and the key K, respectively. Also note

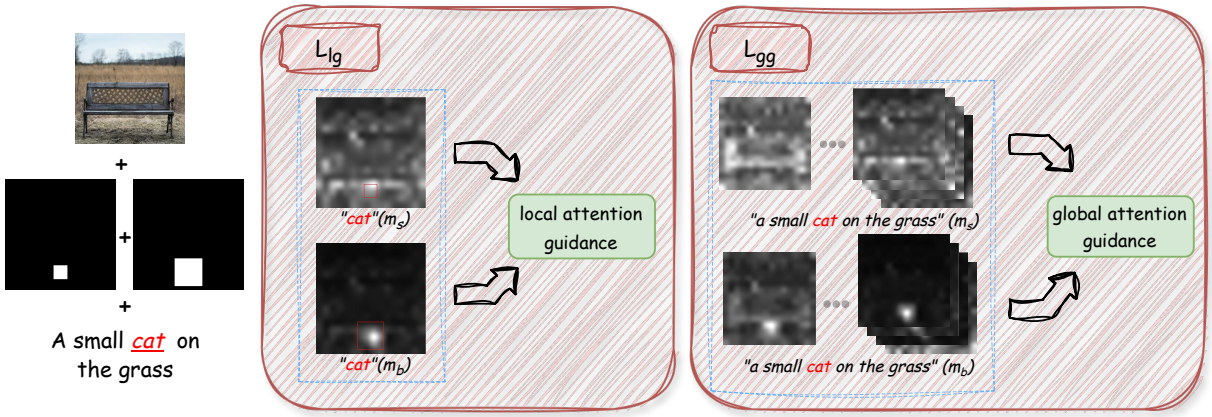

Figure 5: Local attention is primarily used to correct the semantic information within the mask area of the $\mathcal{A}$ corresponding to the main object, ensuring it aligns with what is correct. Global attention guidance collect $\mathcal{A}$ from all work tokens and primarily aims to accurately distinguish unimportant details beyond the main object during the process of generating the $\mathcal{A}$ from the latent code.

that $\mathcal{A}'^{l}$ is still unchanged and only the value of $\mathcal{A}^{l}$ needs to be modified, and no back-propagation operation is required.

The overall guidance function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{lg} + \mathcal{L}_{gg}. \tag{8}$$

Then, we compute the gradient of z_t and update the predicted z_t as:

$$z_t \leftarrow z_t - \eta \nabla_{z_t} \mathcal{L}_{total}, \tag{9}$$

where η is the learning rate.

Based on our experiments, we found that we only need to perform our joint attention guidance on the cross-attention map for the former L denoising timesteps, where z_t undergoes a series of updates. Specifically, it is refined through a gradient descent process aimed at minimizing $\mathcal{L}_{total}$, which occurs J times. The step-by-step methodology for our method is delineated in Algorithm 1.

Our joint attention guidance method performs cross-attention map alignment from a local perspective and a global perspective, which provides a comprehensive correction mechanism. This method aims to enhance the model's ability to accurately focus and generate small targets that are consistent with text descriptions, and assign lower attention values to non-target areas, thereby addressing the key challenge of using diffusion models to generate small targets.

Algorithm 1: Our joint attention guidance method

Input: the textual prompt c, the region mask m_s, the extracted latent code z, the number of total timestep T, the number of guiding timestep L, the number of backward times J, the diffusion model ϵ_θ, the cross-attention map obtainment operation $\mathcal{F}_\theta$, the learning rate η.
Output: the estimated latent code z_0.

Initialization: $z_T \sim \mathcal{N}(0, I)$, $z'_T \leftarrow z_T$
Scale m_s into a large region mask m_b
for $t = T, \ldots, T-K$ **do**
  $z'_{t-1} = \epsilon_\theta(z'_t, t, c, m_b, z)$
  $\mathcal{A}' = \mathcal{F}_\theta(z'_t, t, c, m_b, z)$
  for $j = 0, \ldots, J-1$ **do**
    $z_{t-1} = \epsilon_\theta(z_t, t, c, m_s)$
    $\mathcal{A} = \mathcal{F}_\theta(z_t, t, c, m_s, z)$
    $\mathcal{L}_{total} = \mathcal{L}_{lg}(\mathcal{A}, \mathcal{A}') + \mathcal{L}_{gg}(\mathcal{A}, \mathcal{A}')$
    $\nabla_{z_t} = \partial_{z_t} \mathcal{L}_{total}$
    $z_t = z_t - \eta \nabla_{z_t}$
  end
  $z_{t-1} = \epsilon_\theta(z_t, t, c, m_s, z)$
end
for $t = T-K-1, \ldots, 0$ **do**
  $z_t = \epsilon_\theta(z_t, t, c, m_s, z)$
end
Return: z_0

6 EXPERIMENTS

6.1 Evaluation Metrics.

In the evaluation of our methodology, we employed two metrics: the CLIP-Score and the Fréchet Inception Distance (FID). The CLIP-Score was used to quantify the similarity between the generated objects within the target area and the corresponding text descriptions. This metric effectively assesses how well our model's output aligns with the textual prompts, particularly focusing on the small objects generated. The FID Score was utilized to gauge the quality of the generated images in relation to the original images. Higher CLIP-Score or lower FID, indicate better performance.

It's important to note that the generated objects occupy a relatively small proportion of the original image, using the entire image for metric calculation could lead to misleading results. Therefore, we apply a targeted approach by cropping the original images during the evaluation phase. This cropping is designed to focus on the area of interest, ensuring a more direct and meaningful comparison between the generated object and the original context. This method of evaluation ensures that our metrics accurately reflect

Figure 6: Qualitative results of our method on the SOEBench dataset. The first column of each set contains the original image, mask, and text. The second column shows results from the SD-I model, and the third column displays outcomes of SD-I+Ours.

Table 2: Compare our proposed training-free approach with the stable diffusion model (SD-I) on the SOEBench dataset. lg indicates our local attention guidance method and gg indicates our global attention guidance method. The best results are highlighted in bold.

Prompt	Methods	SOE-2k		SOE-4k	
		CLIP-Score	FID	CLIP-Score	FID
label	SD-I	23.79	34.65	23.11	35.21
	SD-I+lg	23.93	35.02	23.25	35.02
	SD-I+lg+gg	**24.05**	**34.32**	**23.39**	**34.67**
color +label	SD-I	24.10	34.73	23.31	34.96
	SD-I+lg	24.21	34.32	23.47	34.74
	SD-I+lg+gg	**24.36**	**33.85**	**23.62**	**34.28**

the model's performance in generating small objects, providing a true assessment of its capability in this specific task.

6.2 Implementation Details

We use the stable diffusion v1-5 painting model as the baseline method and use DDIM as the noise scheduler. Considering the resize operation on the mask area, a too-large resize ratio will bring a large loss error. We set the scaling factor s to a value randomly selected from 1.5 to 3. We set the learning rate $\eta = 100$ by default. The noise correction is performed during the initial $K = 5$ steps of the denoising process and repeated $J = 5$ times at each step.

6.3 Experimental Results

We evaluate our approach on our proposed benchmarks *SOE-2k* and *SOE-4k* with two kinds of prompt templates, *i.e.*, label and color+label, *e.g.*, *"a dog"* and *"a brown dog"*, and quantitatively compute the CLIP-Score and FID to prove the effectiveness of our cross attention correction. The result is shown in Tab. 2 and our method demonstrated a great improvement compared to the baseline models, exceeding the baseline method stable diffusion by a large margin of nearly 1 point in terms of the FID score on the *SOE-2k* subset, and around 0.7 points in terms of the FID score on the *SOE-4k* dataset. It can be also inferred from the table that given a detailed condition, *i.e.*, color with label, the editing performance of our method exceeds the performance of given a simpler condition, *i.e.*, label only, by a large region of around 0.5 points in terms of the FID score on the *SOE-2k* subset, indicating the effectiveness of our method of understanding different conditions. It should also be mentioned that our method is training-free, and the experimental results verify the great potential of our method on tackling the small object editing problem.

6.4 Ablation Study

We further analyze the effectiveness of the detailed module designs of our joint attention guidance method, which includes a local attention guidance (lg) method and a global attention guidance (gg) method. It can be inferred from Tab. 2 that our local attention guidance can efficiently improve the models' in-painting performance on the small object editing task, mainly attributed to our newly proposed mask region scaling-and-rescaling operation, which efficiently enables the model to generate precise content

corresponding to the textual description, and can also be integrated with the surrounding context of the original image successfully. Based on our local attention guidance method, our global attention guidance method can further bring performance improvement, as our global attention guidance method further reduces the probability of the model modifying the background part, ensuring the accuracy of the modified part, and at the same time ensuring the quality of background image generation.

6.5 Qualitative results

To complement our quantitative studies above, we present qualitative results in this section. We provide samples and qualitative comparison with Stable-Diffusion Inpainting model in Fig 6. As show in the examples, when using the SD-I model for generating images of small objects, issues frequently arise such as low quality, mismatches between text and images, or the absence of generated objects. However, as shown in the third column of each image group in the figures, our method has been able to address these issues to a considerable extent, enhancing the performance of the base model.

6.6 User Preference Study

Besides the automated metrics, we also incorporated a user study to align more closely with intuitive human preference. In the study, we aim to assess both prompt adherence and the overall image. We employed our methods and SD-I(Stable-diffusion-Inpainting) as comparison models, keeping the seed fixed to generate 25 pairs, totaling 50 images. Our study involved 100 total participants. Participants were tasked with ranking images of small target objects generated these two methods. Fig. 7 present the user study. The most important results are:Our training-free methods can enhances the base model's ability to improve both the image quality of generated objects and their alignment with the associated text to a certain extent.

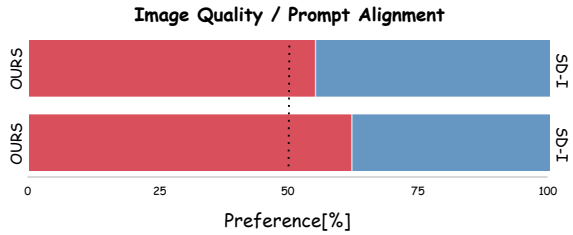

Figure 7: User preference study. We compare the performance of our training-free methods against baselines Stable-Diffuion Inpainting. Our method outperforms baseline in both image quality and prompt alignment study.

6.7 Discussion

Method comparison with P2P. In the realm of image content editing, prompt-to-prompt (P2P) methods have proven to be highly effective for various image manipulation tasks. Nevertheless, they exhibit notable limitations in generating small objects, primarily due to the mismatch between text and the corresponding cross-attention map. To further evaluate this, we conducted tests using a P2P-based method, wherein we attempted to transfer the appropriate attention map from the large mask to the smaller one. Unfortunately, this approach proved ineffective for the generation of small objects because it did not correct the erroneous regions of the cross-attention map. This further corroborates the effectiveness of our method in addressing this specific issue.

Failure Cases. While our method effectively mitigates performance degradation through attention guidance, its efficacy remains constrained by the limitations of the base model. As depicted in Fig. 8, we showcase instances of object omission and subpar quality in our generated images. To bolster the small object editing capabilities further, fine-tuning the model on dedicated small object images presents a straightforward solution.

Figure 8: Failure cases. Due to the performance of the base model, issues such as object missing and low quality will still be encountered even our joint attention guidance approach is adopted.

7 CONCLUSION

In this work, our research introduces two significant contributions to the field of *SOE* (Small Object Editing) using diffusion models. We firstly developed a new benchmark dataset specifically designed to evaluate small object editing. This dataset addresses the unique challenges and requirements of small object imagery, providing a comprehensive and targeted platform for testing and comparing different models. Alongside this dataset, we propose a training-free approach represents a major advancement in solving the issue of generating small objects. By focusing on the alignment of cross attention maps between text and small objects, our method effectively bridges the gap that has historically hindered accurate and detailed generation of small objects in response to textual descriptions. This alignment enables the model to more precisely interpret and render the intricate details required for small object generation .This baseline serves as an initial reference point for future research, offering a simplified yet effective framework for subsequent models to build upon and refine.

Another critical aspect to consider is that our method operates within the constraints of the existing capabilities of the diffusion model's backbone. Since we do not modify the parameters of the diffusion backbone, the quality and the fidelity of the generated images are inherently bound by the performance of the underlying model. Therefore, while our approach enhances the model's ability to generate small objects in alignment with text descriptions, the overall effectiveness is still dependent on the inherent capabilities of the diffusion model.

REFERENCES

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–11.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

[3] Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. 2023. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426* (2023).

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proc. IEEE Conf. CVPR*. IEEE, 18392–18402. https://doi.org/10.1109/CVPR52729.2023.01764

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 148:1–148:10. https://doi.org/10.1145/3592116

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5343–5353.

[7] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2174–2183.

[8] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. 2023. Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models. In *Proc. IEEE ICCV*. IEEE, 7396–7406. https://doi.org/10.1109/ICCV51070.2023.00683

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *Proc. ICLR*. OpenReview.net. https://openreview.net/pdf?id=_CDixzkzeyb

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Proc. NeurIPS*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[13] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. 2024. Diffusion Model-Based Image Editing: A Survey. *arXiv preprint arXiv:2402.17525* (2024).

[14] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2024. PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code. In *Proc. ICLR*. https://openreview.net/pdf/7a08897fea2fe55b08fa202685f17c84e0337fe4.pdf

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *Proc. IEEE Conf. CVPR*. IEEE, 6007–6017. https://doi.org/10.1109/CVPR52729.2023.00582

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proc. IEEE Conf. CVPR*. IEEE, 2416–2425. https://doi.org/10.1109/CVPR52688.2022.00246

[17] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* 128, 7 (2020), 1956–1981.

[19] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2023. Diffusion Models Already Have A Semantic Latent Space. In *Proc. ICLR*. OpenReview.net. https://openreview.net/pdf?id=pd1P2eUBVfq

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proc. IEEE Conf. CVPR*. IEEE, 22511–22521. https://doi.org/10.1109/CVPR52729.2023.02156

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Proc. IEEE Conf. CVPR*. IEEE, 6038–6047. https://doi.org/10.1109/CVPR52729.2023.00585

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. In *Proc. AAAI*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 4296–4304. https://doi.org/10.1609/AAAI.V38I5.28226

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[26] Quynh Phung, Songwei Ge, and Jia-Bin Huang. 2023. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427* (2023).

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proc. ICLR*. https://openreview.net/forum?id=di52zR8xgf

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

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proc. IEEE Conf. CVPR*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[30] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*. 1–10.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proc. ICLR*. OpenReview.net. https://openreview.net/forum?id=St1giarCHLP

[32] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Proc. NeurIPS*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 11895–11907. https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html

[33] Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems* 33 (2020), 12438–12448.

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[35] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16515–16525.

[36] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. 2023. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In *Proc. IEEE Conf. CVPR*. IEEE, 18359–18369. https://doi.org/10.1109/CVPR52729.2023.01761

[37] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7452–7461.

[38] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.

[39] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 833–842.

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proc. IEEE ICCV*. IEEE, 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355

[41] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5802–5810.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009