LLM-guided Instance-level Image Manipulation with Diffusion U-Net Cross-Attention Maps

BMVC 2024 Submission # 457

Abstract

The advancement of text-to-image synthesis has introduced powerful generative models capable of creating realistic images from textual prompts. However, precise control over image attributes remains challenging, especially at the instance level. While existing methods offer some control through fine-tuning or auxiliary information, they often face limitations in flexibility and accuracy. To address these challenges, we propose a pipeline leveraging Large Language Models (LLMs), open-vocabulary detectors and cross-attention maps and intermediate activations of diffusion U-Net for instance-level image manipulation. Our method detects objects mentioned in the prompt and present in the generated image, enabling precise manipulation without extensive training or input masks. By incorporating cross-attention maps, our approach ensures coherence in manipulated images while controlling object positions. Our approach enables precise manipulations at the instance level without fine-tuning or auxiliary information such as masks or bounding boxes.

1 Introduction

Text-to-image synthesis, a field at the intersection of computer vision and natural language processing, tackles the challenge of generating visually realistic images from textual descriptions [8, 12, 15, 25, 26]. This area holds immense potential for various applications, from revolutionizing human-computer interaction to creative content generation. The research community has recognized this significance, evidenced by the development of increasingly powerful text-to-image models such as Imagen [28], DALL-E 3 [4] and Stable Diffusion 3 [10].

However, this field has some challenges. Current models often struggle to capture the full nuance of a text description, resulting in images that lack detail or contain nonsensical elements. Additionally, ensuring photorealism and semantic consistency across generated images remains a hurdle. Overcoming these obstacles is crucial, as it would pave the way for a future where humans can seamlessly communicate their creative vision through text, with machines acting as their capable artistic partners. Tackling these challenges can bridge the gap between human imagination and visual representation.

Among these challenges, a particularly important one is that creating the precise prompt to generate the desired image can be difficult. All desired image attributes should be conveyed

through text, including those inherently complex or impossible to express accurately. Hence, designing a method for precise image editing is a crucial task in the field of text-to-image synthesis.

Previous research has tried to address the challenge of limited control in image editing. However, some methods rely on fine-tuning of pretrained models [5, 11, 17, 27] which is computationally expensive, require large amounts of data and may limit the range of edits. Other methods such as [32] inject diffusion features and self-attention maps to generate a new image while keeping details and appearance from the source one, limiting the range of possible edits. Some methods enable image editing in a zero-shot manner via editing cross-attention maps [9, 13], limiting only to object-type, not instance-level manipulations. Other methods take auxiliary information such as masks [1, 2, 22], which is not always an option, or generate it [7] to better localize the region of interest, limiting the set of resulting edits.

Wu *et al*. [33] proposed Self-correcting LLM-controlled Diffusion (SLD) that automatically aligns the generated image with the user prompt. Firstly, it detects the objects described in the user prompt using a Large-Language Model (LLM) and open-vocabulary detector. Then, LLM finds inconsistencies between the user prompt and detection results and suggests the modification. Then, it performs latent operations to edit the image. This loop is repeated until LLM does not suggest any modifications. This method can be used not only for aligning the image with the prompt but also for image manipulation directly. However, the editing needs to be expressed through the text, limiting the manipulation precision.

To address the issues mentioned above, we propose a novel pipeline. Firstly, we utilize LLM and an open-vocabulary detector to detect the objects mentioned in the prompt and presented on the generated image in the same way as in [33]. This enables instance-level manipulations without requiring any auxiliary information from the user. Then, we perform the instance-level manipulation specified by the user. In contrast to [33] which performs latent operations using unsupervised segmentation, our method utilizes the guidance on cross-attention maps and intermediate activations of diffusion U-Net. This enables precise manipulation of such attributes as position while preserving the original image details. Hence, our pipeline enables to perform precise instance-level manipulations without fine-tuning or auxiliary information while ensuring the preservation of original appearances.

2 Background & Related work

This section provides the necessary background and overview of the related research. Section 2.1 describes the idea behind diffusion models, Section 2.2 describes the guidance and Section 2.3 gives an overview of the related work.

2.1 Diffusion models

Diffusion models use text prompts to generate high-res images from noise through sequential sampling [15, 18, 30]. The aim is to reverse a time-dependent destructive process where noise corrupts data. A neural network ε_θ estimates either the denoised image or the noise ε_t added to create the noisy image $z_t = \alpha_t x + \sigma_t * \varepsilon_t$. Training involves minimizing the loss function:

$$L(\theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T),\varepsilon_t\sim\mathcal{N}(0,\mathbf{I})}[||\varepsilon_t - \varepsilon_\theta(z_t;t,y)||_2^2], \quad (1)$$

where ε_θ, often having a U-Net architecture with self- and cross-attention at different resolutions, incorporates conditioning signal y such as text [26, 28]. Once the model is trained,

the model can produce samples based on conditioning y by setting the noise $z_T \sim \mathcal{N}(0,\mathbf{I})$, then iteratively estimating the noise and updating the noisy image using techniques such as DDIM [29] or DDPM [15]:

$$\hat{\varepsilon}_t = \varepsilon_\theta(z_t;t,y), z_{t-1} = \text{update}(z_t,\hat{\varepsilon}_t,t,t-1,\varepsilon_{t-1}) \tag{2}$$

2.2 Guidance

Diffusion models offer post-training adjustment through guidance, involving the composition of score functions [8, 19, 30]. Conditional samples can be generated using classifier guidance, combining unconditional score function $p(z_t)$ with classifier $p(y|z_t)$ as $p(z_t|y) \propto p(y|z_t)p(z_t)$ [8, 30]. Classifier guidance during sampling adjusts the estimated error term $\hat{\varepsilon}_t$:

$$\hat{\varepsilon}_t = \varepsilon_\theta(z_t;t,y) - s\sigma_t\nabla_{z_t}\log p(y|z_t), \tag{3}$$

where s sets guidance strength. This shifts sampling towards images the classifier considers more likely [8]. Additionally, diffusion sampling can be guided using any energy function $g(z_t;t,y)$, not limited to classifier probabilities. Integrating such guidance yields high-quality text-to-image samples with low energy according to function g:

$$\hat{\varepsilon}_t = (1+s)\varepsilon_\theta(z_t;t,y) - s\varepsilon_\theta(z_t;t,\varnothing) + v\sigma_t\nabla_{z_t}g(z_t;t,y), \tag{4}$$

where v denotes an additional guidance weight for g.

2.3 Diffusion-based Image editing

Image editing is a fundamental task in computer graphics, involving the manipulation of an input image by incorporating various additional elements, such as labels and reference images. Recent advances in text-to-image diffusion models expand their use in image editing tasks, including local and global edits.

Some methods attempted to solve this task by retraining or fine-tuning the diffusion model. For instance, InstructPix2Pix [5] generates image editing dataset using GPT-3 [6], Stable Diffusion [26] and Prompt-to-Prompt [13] and then trains a diffusion model on this dataset to edit the image given the source image and editing prompt. Imagic [17] first optimizes the text embedding to the input image, then fine-tunes the diffusion model to further improve the fidelity, and interpolates between the original and optimized embeddings to generate the resulting image. DreamBooth [27] fine-tunes the diffusion model to reconstruct the images of a specific object and objects of that type to be able to generate new images of that object, given only 3-5 images with it. In comparison, Gal *et al.* [11] proposed to optimize the vector embedding associated with the specific object, rather than the diffusion model, to minimize the reconstruction loss, given 3-5 images of that object. However, all these methods require retraining or fine-tuning of the diffusion model or optimization of the text embedding which is computationally expensive and may limit the range of possible edits. In contrast, our method does not change the diffusion model weights and text embeddings by utilizing guidance.

Some methods attempted to perform image editing in a zero-shot manner. Tumanyan *et al.* [32] proposed to inject self-attention maps and features from diffusion U-Net during generation to preserve the original appearances and details. Prompt-to-Prompt [13] achieves certain types of image editing such as word addition, removal, and replacement by adding,

removing or replacing corresponding cross-attention maps during generation. Self-Guidance [9] utilizes guidance with cross-attention maps and intermediate features of diffusion U-Net to manipulate such attributes as position, size, shape and appearance. However, since these methods are based only on cross-attention maps, they can perform image editing only at the object-type level (i.e., manipulate all objects corresponding to the word, not a single object), but not at the instance level. In contrast, our method can perform instance-level manipulations by extracting objects from the image using LLM and an open-vocabulary detector.

Blended Diffusion [1] blends the CLIP-guided [24] latents with the original image at every diffusion image using the user-specified mask to achieve region-based image editing. Blended Latent Diffusion [2] further develops this idea by applying the same operation in the latent space rather than in the pixel space. DragonDiffusion [22] manipulates the intermediate features of the diffusion model to perform different types of edits such as position change, resizing and object pasting, given the necessary editing masks. However, these approaches require mask specifying the region of interest as an input which is not always an option. DiffEdit [7] automatically generates the editing mask based on the difference between the source and query prompts, then, at some diffusion steps, it blends the generation results from the query prompt with the source image. However, the generated mask is not precise. Such an approach also does not enable instance-level manipulations and limits the range of possible edits. In contrast, our method extracts the bounding boxes corresponding to every object mentioned in the prompt using LLM and an open-vocabulary detector. This enables to extract precise regions of interest without limiting the set of possible edits.

Self-correcting LLM-controlled Diffusion (SLD) [33] utilizes a different approach. Firstly, it extracts a set of objects from the prompt using LLM and detects them on the image. Then, LLM suggests necessary edits to make the image align with the prompt. Finally, it performs corresponding latent operations to edit the image. This loop is repeated until the image fully matches the prompt. We use the object extraction and detection part in our method since it enables to precisely locate the objects which should be manipulated or preserved without auxiliary information. However, the editing part is limited only to text-based image manipulation which is not precise. To enable more precise editing, we utilize the guidance based on the cross-attention maps on features from diffusion U-Net.

3 Methodology

An overview of our method can be seen in Fig. 1. Firstly, LLM parses the objects from the given prompt. Then, the open-vocabulary detector detects the parsed objects on the generated image. Then, given the user edit, we perform the image editing using guidance based on the cross-attention maps and features from the diffusion U-Net. Section 3.1 describes LLM parsing and open-vocabulary detection, and Section 3.2 describes the image editing with guidance.

3.1 LLM-based object detection

In our method, LLM-based object detection extracts the objects mentioned in the prompt and are present in the generated image. We do it in the same way as was done by Wu *et al.* [33]. Firstly, LLM extracts the objects mentioned in the prompt along with their attributes and quantities. Then, the open-vocabulary detector [21] detects the objects extracted during the previous step on the image. In contrast to methods such as Self-Guidance [9] which

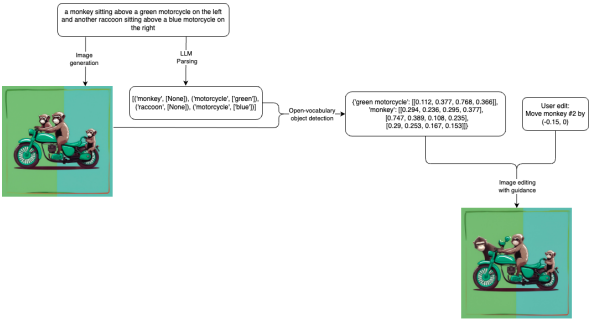


Figure 1: Overview of our pipeline. Firstly, LLM parses the objects from the prompt. Then, an open-vocabulary object detector detects these objects on the image. Finally, the image is edited with the use of guidance.

operate at the object level and can not extract separate objects, these steps enable our method to precisely locate all the objects of interest without requiring auxiliary information from the user such as masks unlike methods such as DragonDiffusion [22]. Then, the image can be edited at the instance level by utilizing the obtained bounding boxes.

3.2 Image editing with guidance

After obtaining the detection results, the user needs to provide which object needs to be manipulated. This enables more precise edits compared to methods enabling only text-based manipulations such as SLD [33]. Then, image editing is performed using guidance based on cross-attention maps and features from diffusion U-Net. Only the position can be manipulated, but the method can be extended to other manipulations. Guidance has been shown to enable precise control over the image generation process [3, 14], while recent research has demonstrated that cross-attention maps contain the information about the object position and shape [3, 13] and intermediate diffusion features contain the information about object appearances [22, 32]. Hence, this enables better control over the position while preserving appearances in the image, in contrast to methods such as SLD [33] that directly inject objects into the latent vector degrading the image realism and fidelity.

3.2.1 Position

Given the original object bounding box (x_1, y_1, x_2, y_2) and shift (x, y), the position can be manipulated by using the following guidance term:

$$g_{\text{position}}(o) = -\frac{1}{(x_2 - x_1)(y_2 - y_1)} \sum_{h,w} (\mathcal{A}_{h,w} * \mathcal{M}^{\text{target}}_{h,w})^2 + \\ + \frac{1}{(x_2 - x_1)(y_2 - y_1)} \sum_{h,w} (\mathcal{A}_{h,w} * \mathcal{M}^{\text{orig}}_{h,w})^2, \tag{5}$$

where $\mathcal{A}_k$ is the cross-attention map corresponding to token k obtained during the image editing, $\mathcal{M}^{\text{orig}}$ is the mask corresponding to the original bounding box, $\mathcal{M}^{\text{target}}$ is obtained by shifting $\mathcal{M}_{\text{orig}}$ by the shift (x, y) and $*$ denotes element-wise multiplication. The first aims

to minimize the model's focus on the original location, i.e. remove the object from there, while the second term aims to make the model focus on the target location, i.e. make the object appear at the target location.

3.2.2 Object preservation

For the rest of the objects which are not manipulated, we calculate the Mean Squared Error between the intermediate activations of diffusion U-Net obtained during the original generation denoted as Ψ^{orig} and intermediate activations obtained during the manipulation denoted as Ψ^{target}:

$$g_{\text{preserve}}(o) = \frac{1}{(x_2 - x_1)(y_2 - y_1)} \sum_{h,w} (\Psi^{\text{orig}}_{h,w} * \mathcal{M}^{\text{orig}}_{h,w} - \Psi^{\text{target}}_{h,w} * \mathcal{M}^{\text{target}}_{h,w})^2 \tag{6}$$

3.2.3 Total guidance term

Given the set of objects $\mathcal{O}$ and the manipulated object o_k, the total guidance term is the following:

$$g = w_0 \frac{1}{|O| - 1} \sum_{o \neq o_k \in O} \frac{1}{|\Psi|} \sum_{i=0}^{|\Psi|} g_{\text{preserve}}(o) + w_1 \frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} g_{\text{manipulation}}(o_k) \tag{7}$$

This guidance term is used to update the noise estimate according to Eq. 4.

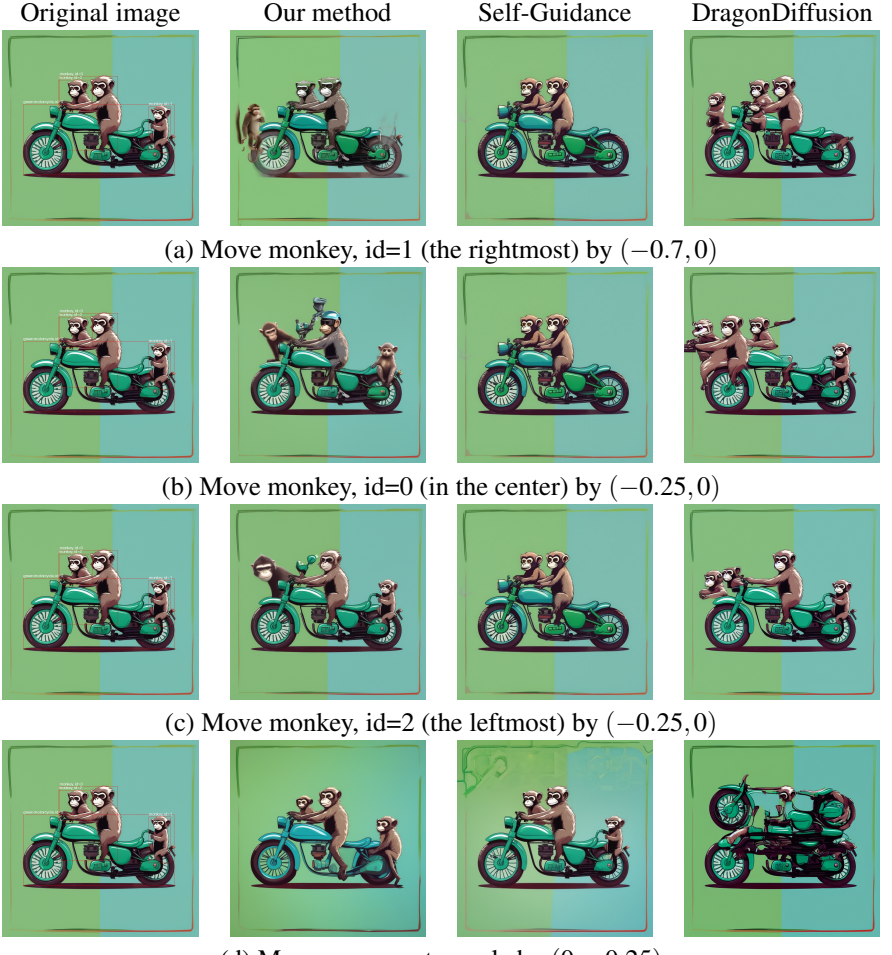
4 Results & Discussion

As LLM, we chose the Gemma-7b instruction model [20] which has been shown to outperform other state-of-the-art LLMs such as Mistral-7B-Instruct-v0.2 [16] and Llama 2 [31]. For the open-vocabulary detector, we used OWLv2 [21] which was shown to perform the best on zero-shot open-vocabulary object detection task. We tested our pipeline on the Stable Diffusion XL model [23] since it is one of the state-of-the-art diffusion models. Our method applies the position guidance term from Eq. 5 to all cross-attention maps at the first upper block of the diffusion U-Net since it contains the most precise information about objects' position and shape and the preservation guidance term from Eq. 6 to the features of the third upper block of the diffusion U-Net since it contains the most precise appearances and details.

We compared our method to Self-Guidance [9], which provides a method for manipulating the position, although it provides manipulation only at the object-type level, not at the instance level. We also compared it to DragonDiffusion [22], which enables manipulating such attributes, such as position at the instance level, but requires auxiliary information in the form of masks or bounding boxes. In contrast to Self-Guidance, our method enables instance-level manipulations. Unlike DragonDiffusion, our method does not require any auxiliary information. We did not compare our method to SLD [33] since it requires an OpenAI API key, which we could manage to obtain. Other methods for image manipulation with diffusion models do not enable instance-level editing and do not enable manipulating the position.

Section 4.1 underscores the precision of our approach in manipulating object positions at an instance level, showcasing its superiority over current state-of-the-art methods such as

Self-Guidance and DragonDiffusion. Section 4.2 presents a comparative analysis of different preservation terms, showing the impact of utilizing intermediate activations compared to cross-attention maps in maintaining appearance fidelity during manipulation. Designing a metric for evaluating the image editing techniques is not yet a solved task, especially for methods that manipulate attributes such as position. Hence, we used qualitative (i.e., visual) comparison for both experiments to directly visualize the results and assess our approach, similar to previous methods [7, 9, 13, 22, 27].



(a) Move monkey, id=1 (the rightmost) by $(-0.7, 0)$

(b) Move monkey, id=0 (in the center) by $(-0.25, 0)$

(c) Move monkey, id=2 (the leftmost) by $(-0.25, 0)$

(d) Move green motorcycle by $(0, -0.25)$

Figure 2: Examples of the position manipulations.

4.1 Position

The examples of the position manipulation can be seen in Fig. 2. Our method achieves precise position manipulation while largely preserving the appearance fidelity of manipulated objects. Notably, our approach can manipulate specific instances of objects, which Self-Guidance [9] does not achieve, being limited to controlling object types rather than individ-

ual instances. For instance, our method can manipulate individual monkeys while preserving their distinct appearances. In contrast, Self-Guidance gives the same result for manipulating any monkey since it can only perform edits at the object type level and can not differentiate between different instances of the same object type. Compared to DragonDiffusion [22], our method maintains fidelity and realism in the manipulated images, particularly in the regions where objects are repositioned. While DragonDiffusion may preserve appearances more precisely, our method's advantage lies in its ability to maintain realism and fidelity, crucial for applications requiring realistic modifications.

Despite these strengths, there are areas for improvement in our method. Notably, while our approach generally preserves appearances well, there are occasional deviations, such as colour shifts in objects like monkeys or minor alterations in motorcycle details. Furthermore, the appearance of the moved object changes completely since the position term in Eq. 5 utilizes cross-attention maps containing only information about the general object's location and shape, not the appearances.

Another problem is that the method requires a thorough choice of hyperparameters for each manipulation, which may pose a challenge in practical applications. The weights require tuning, as the combination of weights working well for one object may not yield satisfactory results for another.

Moreover, our method, along with Self-Guidance and DragonDiffusion, faces difficulties in effectively manipulating large objects. For instance, attempting to reposition the motorcycle in Figure 2 resulted in no displacement, highlighting a limitation in handling substantial changes without compromising image fidelity. Although DragonDiffusion managed to place the motorcycle upper, it failed to remove it from its original location.

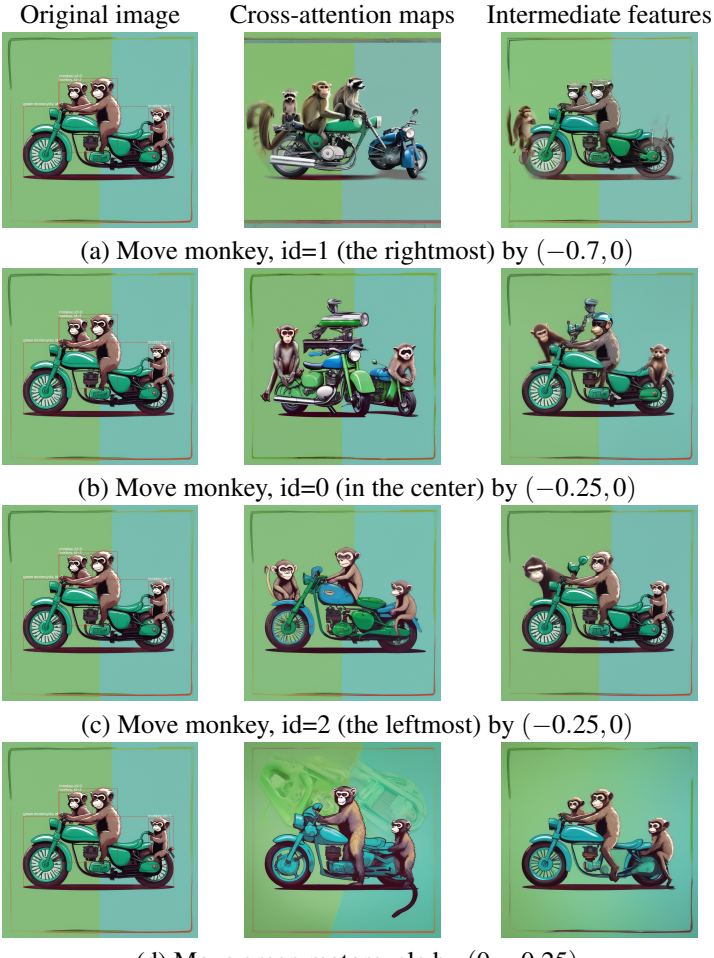
4.2 Ablation study on different preservation terms

We also compared two different preservation terms in Eq. 6: one that utilizes the cross-attention maps as the position manipulation term in Eq. 5 and one that utilizes intermediate activations of diffusion U-Net. The comparison can be seen in Fig. 3. As can be seen, utilizing cross-attention maps makes only general location and shape unmodified while the appearances are significantly changed. In contrast, the intermediate activations of diffusion U-Net. The reason is that, unlike cross-attention maps, intermediate activations contain information not only about the general position and shape but also about appearances.

5 Conclusion & Future Work

In this paper, we presented a pipeline for instance-level image manipulation. Our method enables the detection of objects mentioned in the prompt and present in the generated image by leveraging LLMs and open-vocabulary detectors, facilitating precise control at the instance level without the need for expensive fine-tuning or auxiliary information such as input masks. In addition, our method precisely preserves the appearances of the image, ensuring its coherence.

Future work will focus on making our approach less sensitive to the choice hyperparameters. Making it not necessary to tune hyperparameters for every manipulation will make it much more convenient for users. In addition, we will focus on improving the preservation of the manipulated object's appearance during manipulation and on improving the position manipulation of large objects. Our method and previous methods do not solve these problems, and



(a) Move monkey, id=1 (the rightmost) by $(-0.7, 0)$

(b) Move monkey, id=0 (in the center) by $(-0.25, 0)$

(c) Move monkey, id=2 (the leftmost) by $(-0.25, 0)$

(d) Move green motorcycle by $(0, -0.25)$

Figure 3: Comparison of different preservation terms in our method.

addressing them will further enhance the precision of image editing approaches. Finally, current research does not provide suitable metrics for evaluating methods for this type of image editing. Hence, future work will also focus on designing such a metric.

References

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

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, July 2023.

[3] Arpit Bansal et al. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] James Betker et al. Improving image generation with better captions, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, June 2023.

[6] Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023.

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion Self-Guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 16222–16239, 2023.

[10] Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[11] Rinon Gal et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, volume 35, pages 14715–14728, 2022.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[16] Albert Q. Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[17] Bahjat Kawar et al. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, June 2023.

[18] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pages 21696–21707, 2021.

[19] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439, 2022.

[20] Thomas Mesnard et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[21] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 72983–73007, 2023.

[22] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling drag-style manipulation on diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Dustin Podell et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 18–24 July 2021.

[25] Aditya Ramesh et al. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8821–8831, 18–24 July 2021.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023.

[28] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[31] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[32] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[33] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. *arXiv preprint arXiv:2311.16090*, 2023.