

Text-Driven Image Editing via Learnable Regions

Yuanze Lin[♣] Yi-Wen Chen[♣] Yi-Hsuan Tsai[★] Lu Jiang[★] Ming-Hsuan Yang^{♣★}
 ♣ University of Oxford ♦ UC Merced ★ Google

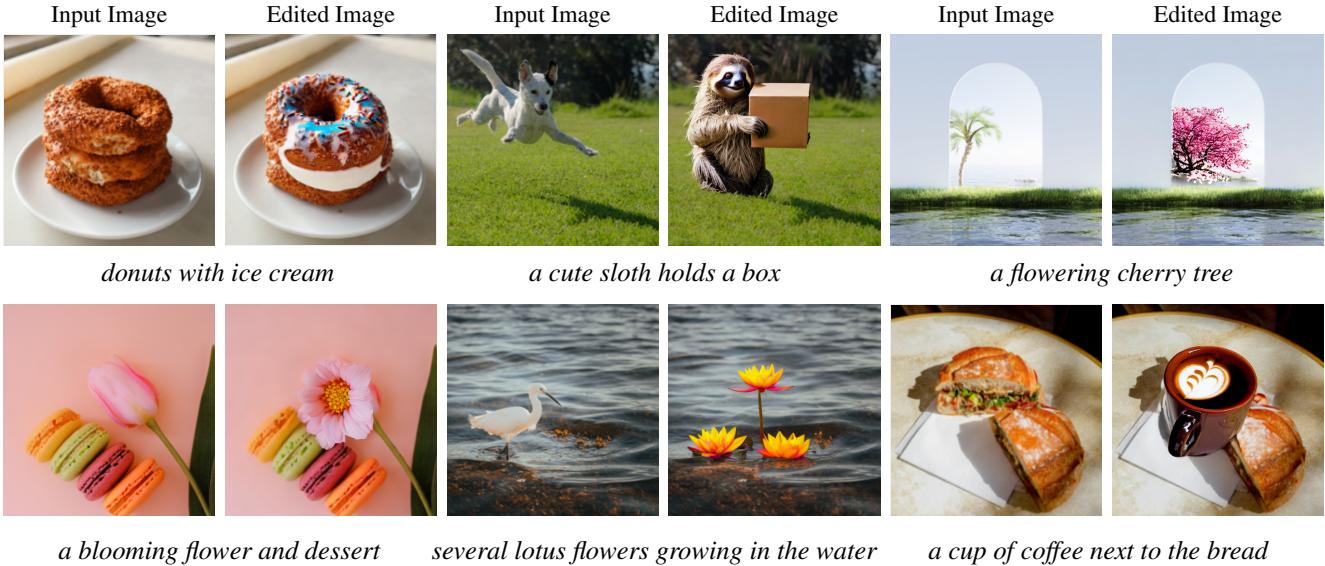


Figure 1. **Overview.** Given an input image and a language description for editing, our method can generate realistic and relevant images without the need for user-specified regions for editing. It performs local image editing while preserving the image context.

Abstract

Language has emerged as a natural interface for image editing. In this paper, we introduce a method for region-based image editing driven by textual prompts, without the need for user-provided masks or sketches. Specifically, our approach leverages an existing pre-trained text-to-image model and introduces a bounding box generator to find the edit regions that are aligned with the textual prompts. We show that this simple approach enables flexible editing that is compatible with current image generation models, and is able to handle complex prompts featuring multiple objects, complex sentences, or long paragraphs. We conduct an extensive user study to compare our method against state-of-the-art methods. Experiments demonstrate the competitive performance of our method in manipulating images with high fidelity and realism that align with the language descriptions provided. Our project webpage: https://yuanze-lin.me/LearnableRegions_page.

1. Introduction

With the availability of a massive amount of text-image paired data and large-scale vision-language models, recent text-driven image synthesis models [8, 32–34, 39, 42, 43, 53, 57, 62] have enabled people to create and manipulate specific visual contents of realistic images using natural language descriptions in an interactive fashion.

Recent text-driven image editing methods [1, 2, 21, 30, 38, 39, 54] have shown impressive capabilities in editing realistic images based on natural descriptions, with approaches typically falling into two paradigms: *mask-based* or *mask-free* methods. Mask-based editing approaches [1, 39] are perceived intuitively for local image editing because they allow users to specify precisely which areas of an image to modify. However, these methods can be laborious, as they demand users to manually create masks that are sometimes unnecessary, limiting their user experience in many applications.

In contrast, mask-free editing approaches [6, 9, 21, 38, 54] do not require masks and can directly modify the ap-

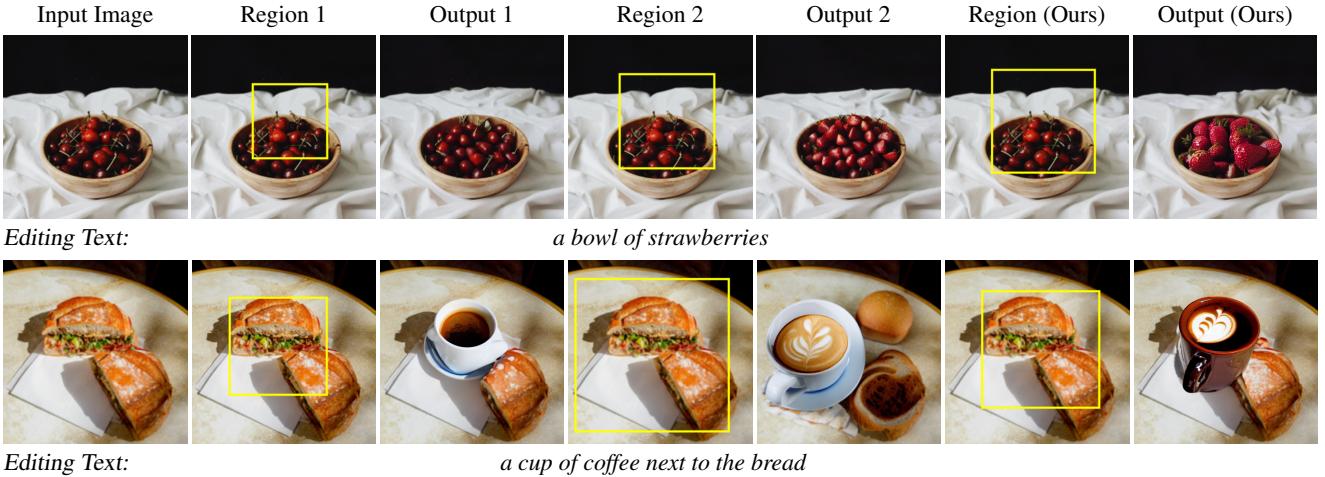


Figure 2. Effects of variations in editing regions on generated image quality. *Region 1* and *Region 2* are two prior regions drawn from the self-attention map of DINO [7]. *Region (ours)*, shown in the second-to-last column, represents the regions produced by our model which have the best overall quality.

pearance or texture of the input image. These methods are trained to create fine-grained pixel masks, which can be applied to either the RGB space or the latent embedding space within the latent diffusion model framework. While there has been significant advancement in mask-free editing, the precision of editing in current methods relies heavily on the accuracy of detailed masks at the pixel level. Current methods encounter difficulties with local modifications, particularly when dealing with less accurate masks.

Current mask-free image editing approaches have predominantly concentrated on pixel masks [6, 10]. The use of bounding boxes as an intermediate representation for editing images has not been thoroughly explored. Bounding boxes can provide an intuitive and user-friendly input for image editing. They facilitate a smoother interactive editing process by being quicker and easier for users to adjust the box, unlike pixel masks that typically require more time and precision to draw pixels accurately. Moreover, some generative transformer models such as Muse [9] may support only box-like masks as opposed to pixel-level masking for image editing.

This paper explores the feasibility of employing bounding boxes as an intermediate representation within a mask-free editing framework. Our objective is not to propose a new image editing model, but to introduce a component that enables an existing pretrained mask-based editing model to perform mask-free editing via the learnable regions. To this end, we propose a region-based editing network that is trained to generate editing regions utilizing a text-driven editing loss with CLIP guidance [42]. Our method can be integrated with different image editing models. To demonstrate its versatility, we apply it to two distinct image synthesis models: non-autoregressive transformers as used in

MaskGIT [8] and Muse [9], as well as Stable Diffusion [45]. It is worth highlighting that the latent spaces in transformer models (MaskGIT and Muse) are only compatible with box-like masks and lack the precision for pixel-level masks in image editing.

Our experimental results demonstrate that the proposed method can generate realistic images that match the context of the provided language descriptions. Furthermore, we conduct a user study to validate that our method outperforms five state-of-the-art baseline methods. The results indicate that our method edits images with greater fidelity and realism, following the changes specified in the language descriptions. The contributions of this work are as follows:

- Our approach enables mask-based text-to-image models to perform local image editing without needing masks or other user-provided guidance. It can be integrated with existing text-guided editing models to improve their quality and relevance.
- We introduce a novel region generator model that employs a new CLIP-guidance loss to learn to find regions for image editing. We demonstrate its applicability by integrating it with two popular and distinct text-guided editing models, MaskGIT [8] and Stable Diffusion [45].
- Experiments show the high quality and realism of our generated results. The user study further validates that our method outperforms state-of-the-art image editing baselines in producing favorable editing results.

2. Related Work

Text-to-Image Synthesis. In recent years, significant progress has been made in text-to-image synthesis. While early contributions are mainly based on Generative Adver-

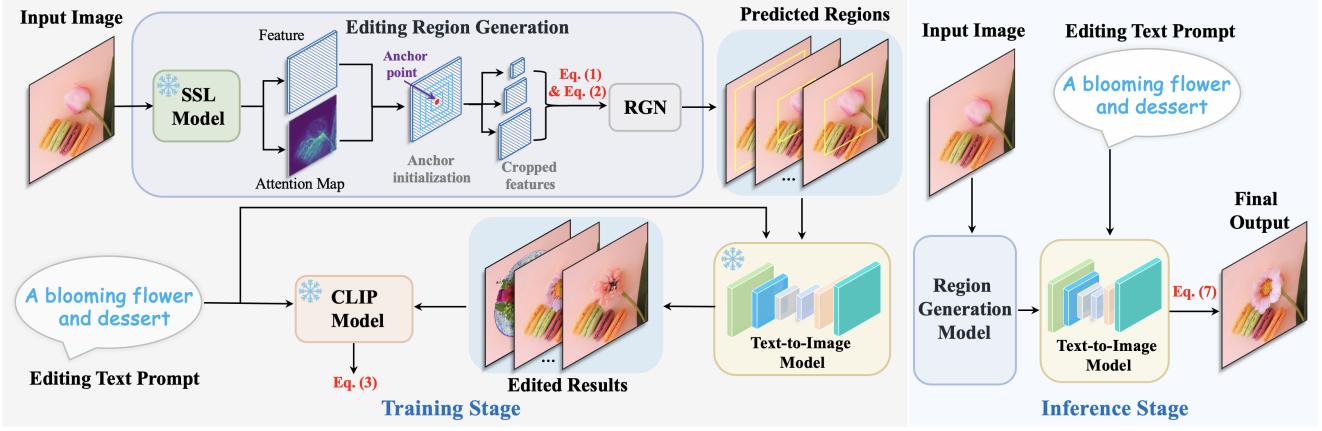


Figure 3. Framework of the proposed method. We first feed the input image into the self-supervised learning (SSL) model, e.g., DINO [7], to obtain the attention map and feature, which are used for anchor initialization. The region generation model initializes several region proposals (e.g., 3 proposals in this figure) around each anchor point, and learns to select the most suitable ones among them with the region generation network (RGN). The predicted region and the text descriptions are then fed into a pre-trained text-to-image model for image editing. We utilize the CLIP model for learning the score to measure the similarity between the given text description and the edited result, forming a training signal to learn our region generation model.

sarial Network (GAN) approaches [31, 56, 59, 60], the latest models are mostly built on diffusion models [20, 22, 40, 44, 45, 48, 52, 61] or transformer models [14, 16, 43, 58, 58]. For example, DALL-E 2 [44] and Imagen [23] propose to condition textual prompts to diffusion models, while Muse [9] leverages masked generative transformers to generate images from texts. Other approaches [11, 12, 44] leverage pre-trained CLIP models [42] to guide image generation based on textual descriptions.

More recently, Stable Diffusion [45], trained on large image-text pairs [49], has been made publicly available and has served as the foundation for numerous image generation and manipulation works. ControlNet [61] proposes to control Stable Diffusion with spatially localized conditions for image synthesis. Different from these works, we aim to introduce a component that can enable pretrained text-to-image models for mask-free local image editing.

Text-driven Image Manipulation. Several recent works have utilized pre-trained generator models and CLIP [42] for text-driven image manipulation [2, 3, 17, 30, 35, 41]. StyleCLIP [41] combines the generative ability of StyleGAN [25] with CLIP to control latent codes, enabling a wide range of image manipulations. VQGAN-CLIP [12] uses CLIP [42] to guide VQ-GAN [16] for high-quality image generation and editing.

There are several approaches [1, 18, 26, 27, 36, 37, 39, 47, 53] that use diffusion models for text-driven image manipulation. Imagic [26] can generate textual embeddings aligning with the input images and editing prompts, and fine-tune the diffusion model to perform edits. Instruct-Pix2Pix [4] combines GPT-3 [5] and Stable Diffusion [45] to edit images with human instructions. Our work is related

to the state-of-the-art methods DiffEdit [10] and MasaCtrl [6]. DiffEdit [10] leverages DDIM inversion [13, 52] with the automatically produced masks for local image editing. MasaCtrl [6] proposes mutual self-attention and learns the editing masks from the cross-attention maps of the diffusion models. Motivated by the aforementioned works, we also utilize diffusion models and CLIP guidance for text-driven image manipulation. In contrast to DiffEdit [10] and MasaCtrl [6], whose editing is much more sensitive to the generated mask regions, our proposed method focuses on learning bounding boxes for local editing, which can be more flexible in accommodating diverse text prompts.

3. Proposed Method

Text-driven image editing manipulates the visual content of input images to align with the contexts or modifications specified in the text. Our goal is to enable text-to-image models to perform mask-free local image editing. To this end, we propose a region generation network that can produce promising regions for image editing.

Figure 3 shows the overall pipeline of our proposed method for text-driven image editing.

3.1. Edit-Region Generation

Given the input image as $X \in \mathbb{R}^{3 \times H \times W}$ and text with p words as $T \in \mathbb{Z}^p$, we first use a pre-trained visual transformer model, Vit-B/16 [15], for feature extraction. This model is pre-trained using the DINO self-supervised learning objective [7]. The features $F \in \mathbb{R}^{d \times h \times w}$ from the last layer have been shown to contain semantic segmentation of objects [7, 51], which can serve as a prior in our problem.

Then we initialize K anchor points $\{C_i\}_{i=1}^K$ located at the top- K scoring patches of the self-attention map from the [CLS] token query of the DINO pre-trained transformer as shown in [7], where the [CLS] token carries guidance to locate the semantically informative parts of the objects.

Following this, we define a set of bounding box proposals $\mathcal{B}_i = \{B_j\}_{j=1}^M$ for each anchor point C_i , where each bounding box is centered at their corresponding anchor point. For simplicity, we parameterize the bounding box with a single parameter such that each B_j is a square box with shape $j \times j$.

Subsequently, we train a region generation network to explicitly consider all unique bounding boxes derived from the same anchor point. For a given anchor point C_i , we then have:

$$f_j = \text{ROI-pool}(F, B_j), \quad (1)$$

$$S([f_1, \dots, f_M]) = [\pi_1, \dots, \pi_M], \quad (2)$$

where $[\cdot]$ concatenates features along the channel dimension, and the ROI-pool operation [19] is used to perform pooling for the image feature $F \in \mathbb{R}^{d \times h \times w}$ with respect to the box B_j , resulting in a feature tensor $f_j \in \mathbb{R}^{d \times l \times l}$. In our experiments, we set l as 7. S is the proposed region generation network consisting of two convolutional layers and two linear layers, with a ReLU activation layer between consecutive layers. The output from the final linear layer, denoted as π_j in Eq. (2) as the logits for the bounding box with size j , is fed into a softmax function to predict the scores for the bounding box proposal, *i.e.*, $\text{Softmax}([\pi_1, \dots, \pi_M])$.

To learn the parameters of the region generation network, we use the Gumbel-Softmax trick [24]. We re-parameterize π_j by adding a small Gumbel noise $g_j = -\log(-\log(u_j))$ where $u_j \sim \text{Uniform}(0, 1)$. During training, we apply straight-through gradient estimation, in which backward propagation uses the differentiable variable (*i.e.*, softmax) while the forward pass still takes the argmax, treating π as the categorical variable. For each anchor point, once we obtain the edit region with the highest softmax score, we generate the corresponding mask and feed the mask, editing prompt, and input image to the text-to-image model to obtain the edited image. Thus, we can get K edited images considering all anchor points, in Section 3.3, we explain how to produce the final edited image as inference output.

3.2. Training Objectives

As the CLIP model [42] can estimate the similarity between images and texts, we employ it to guide our image editing based on user-specified prompts.

To train our models, we propose a composite editing loss that consists of three components: 1) the CLIP guidance loss \mathcal{L}_{Clip} stands for the cosine distance between the features of generated images and texts from the last layers of CLIP’s encoders, 2) the directional loss \mathcal{L}_{Dir} [41] controls

the direction of the performed edit in CLIP space [17, 41], and 3) the structural loss \mathcal{L}_{Str} considers the self-similarity [29, 50] of features from source and generated images, it enables editing in texture and appearance while maintaining the original spatial layout of the objects from the source images. The total loss \mathcal{L} and each loss term are:

$$\mathcal{L} = \lambda_C \mathcal{L}_{Clip} + \lambda_S \mathcal{L}_{Str} + \lambda_D \mathcal{L}_{Dir}, \quad (3)$$

$$\mathcal{L}_{Clip} = \mathcal{D}_{cos}(E_v(X_o), E_t(T)), \quad (4)$$

$$\mathcal{L}_{Str} = \|Q(f_{X_o}) - Q(f_X)\|_2, \quad (5)$$

$$\mathcal{L}_{Dir} = \mathcal{D}_{cos}(E_v(X_o) - E_v(X), E_t(T) - E_t(T_{ROI})), \quad (6)$$

where E_v and E_t are the visual and textual encoder of the CLIP model. We empirically set the weights $\lambda_C = 1$, $\lambda_D = 1$, and $\lambda_S = 1$ for our composite editing loss. Here, X , T , and X_o denote the input image, text prompt, and the edited image by the proposed region, respectively. f_{X_o} and f_X indicate the visual features from the last layer of CLIP’s visual encoder, while $Q(f_{X_o})$ and $Q(f_X)$ denote the similarity matrix of f_{X_o} and f_X respectively. For simplicity, we use the cosine distance \mathcal{D}_{cos} to measure the similarity between images and texts. In addition, T_{ROI} represents a region-of-interest of the source images for editing (*e.g.*, in Figure 14, when T is “a big tree with many flowers in the center”, then T_{ROI} could be “tree”).

During training, our loss functions encourage the region generator to produce appropriate regions for editing by taking into account the similarity between the edited images and the given text descriptions.

3.3. Inference

During the inference process, we define a quality score to rank the edited images generated from different anchor points and select the image with the highest score for presentation to the user.

While there exist more advanced methods, we use a simple weighted average to compute the quality score:

$$S = \alpha \cdot S_{t2i} + \beta \cdot S_{i2i}, \quad (7)$$

where S_{t2i} estimates the cosine similarity scores between the given text descriptions and the edited images, S_{i2i} measures the cosine similarity scores between the source images and the edited images, and α and β are the coefficients to control the influences of S_{t2i} and S_{i2i} . We adopt the features extracted from the last layer of CLIP’s encoders for similarity calculation.

In our experiments, we set α and β as 2 and 1 respectively, since a higher value for α can encourage the model to place more weight on the faithfulness of text-conditioned image editing. The edited image with the highest quality score S is chosen as the final edited image.

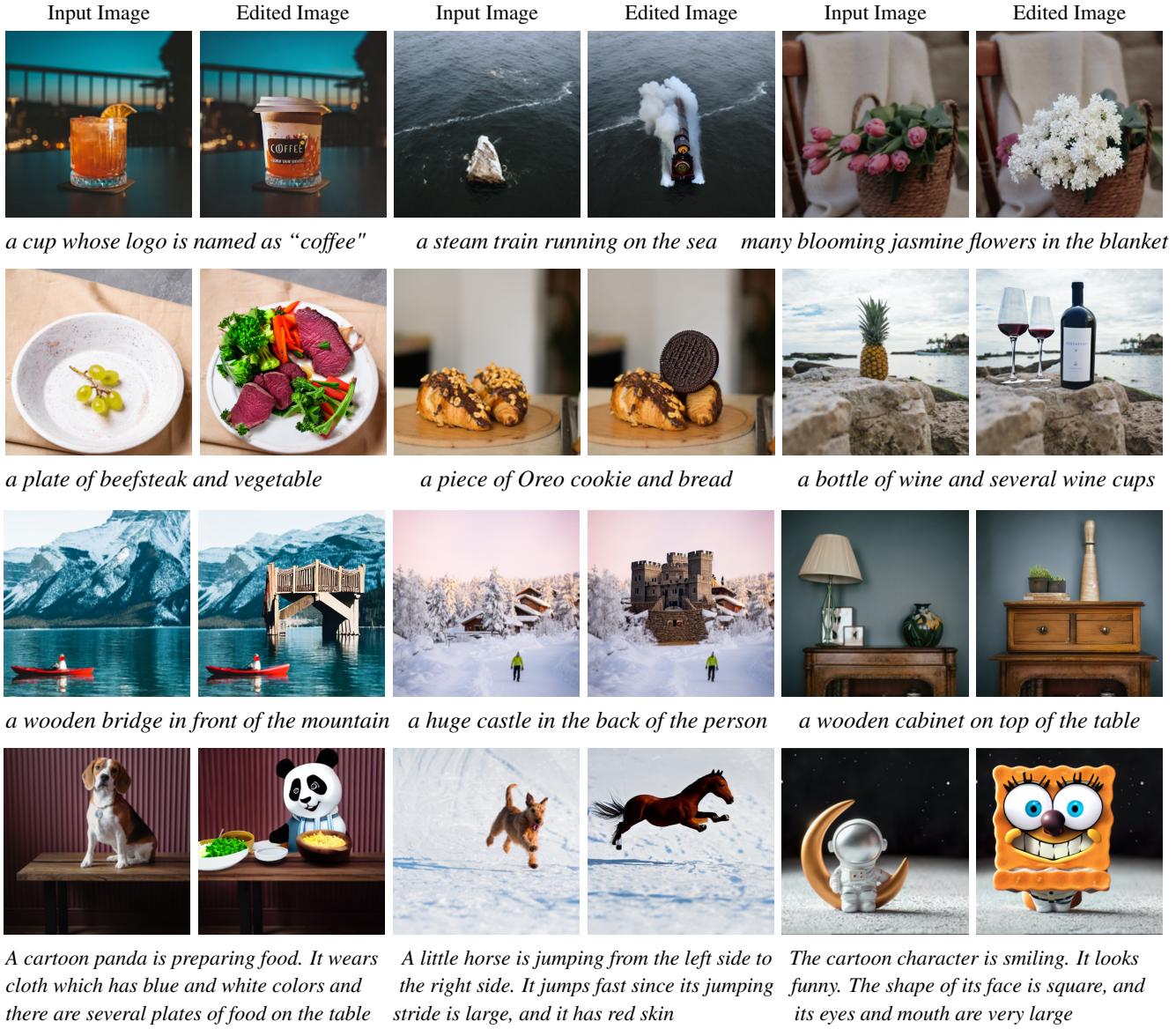


Figure 4. **Image editing results with simple and complex prompts.** Given the input images and prompts, our method edits the image without requiring masks from the users. The learned region is omitted for better visualization. The 1st row contains diverse prompts for one kind of object. The 2nd row displays prompts featuring multiple objects. The 3rd row shows prompts with geometric relations, and the last row presents prompts with extended length.

3.4. Compatibility with Pretrained Editing Models

Our proposed region generator can be integrated with various image editing models [1, 8, 39, 45] for modifying the content of source images conditioning on the prompts, and to demonstrate its versatility, we apply it to two distinct image synthesis models: non-autoregressive transformers as used in MaskGIT [8] or Muse [9], as well as diffusion U-Nets [46] as used in Stable Diffusion [45].

The transformer and diffusion models represent distinct base editing models to verify the applicability of the pro-

posed method. It is worth noting that MaskGIT and Muse are transformers that operate over discrete tokens created by a VQ autoencoder [55], unlike diffusion models [22, 45, 52] operating within the continuous space. As a result, the latent spaces in MaskGIT and Muse are only compatible with box-like masks and lack the precision for pixel-level masks in image editing.

For our experiments, we use the official MaskGIT model instead of the Muse model [9], which is not publicly available. We also limit the text prompt to the class vocabulary that the model is trained on.

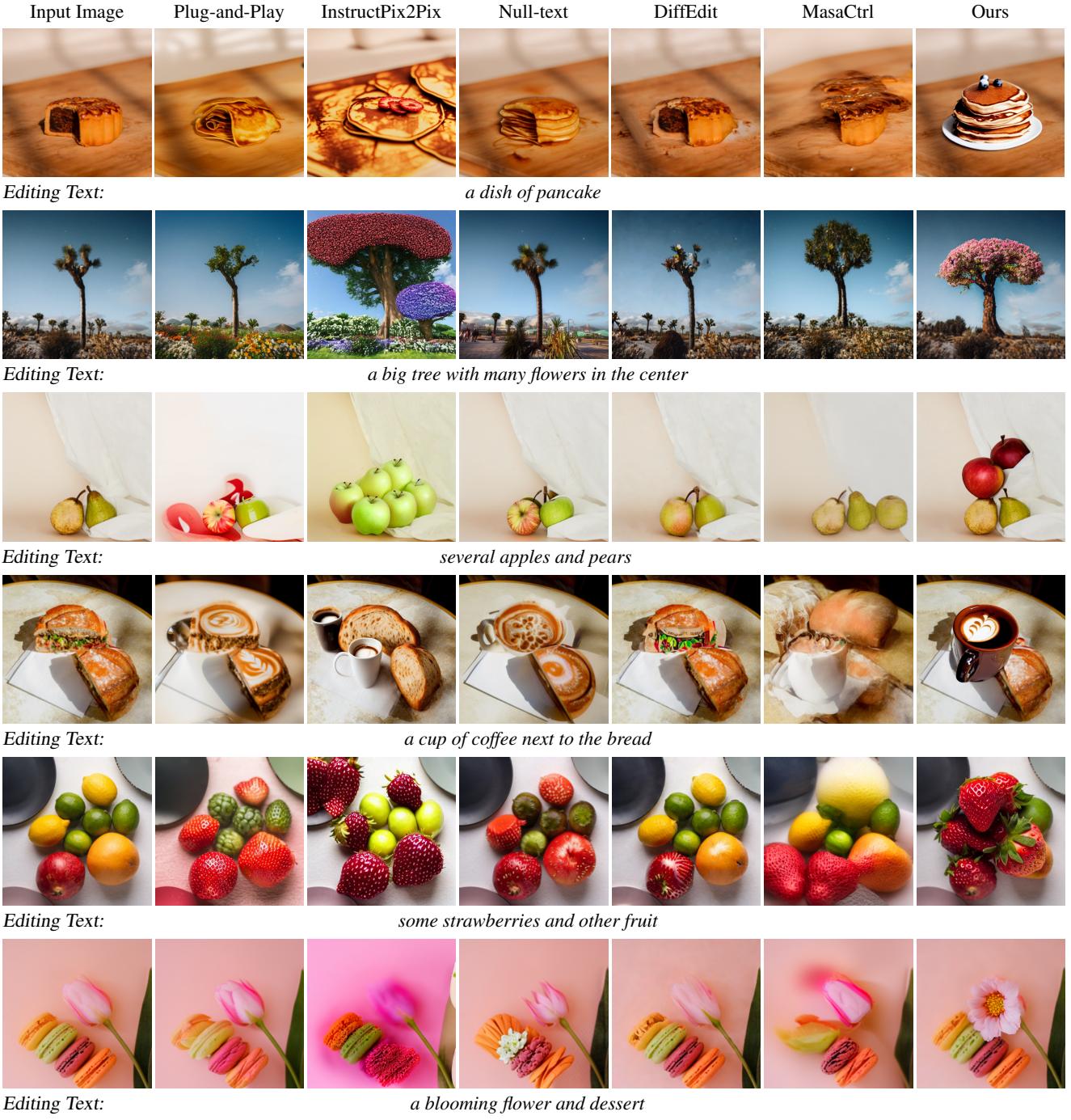


Figure 5. **Comparison with existing methods.** We compare our method with existing text-driven image editing methods. From left to right: Input image, Plug-and-Play [54], InstructPix2Pix [4], Null-text [38], DiffEdit [10], MasaCtrl [6], and ours.

4. Experimental Results

Implementation Details. In our evaluation, we collect high-resolution and free-to-use images covering a variety of objects from Unsplash (<https://unsplash.com/>). For edit-region generation, the total number of bounding box

proposals (*i.e.*, M) is 7 and the CLIP guidance model is initialized with ViT-B/16 weights. We do not use super-resolution models to enhance the quality of the resultant images. By default, we adopt the pre-trained Stable Diffusion-v-1-2 as our editing model. Our main experiments are conducted using two A5000 GPUs, where we train the model

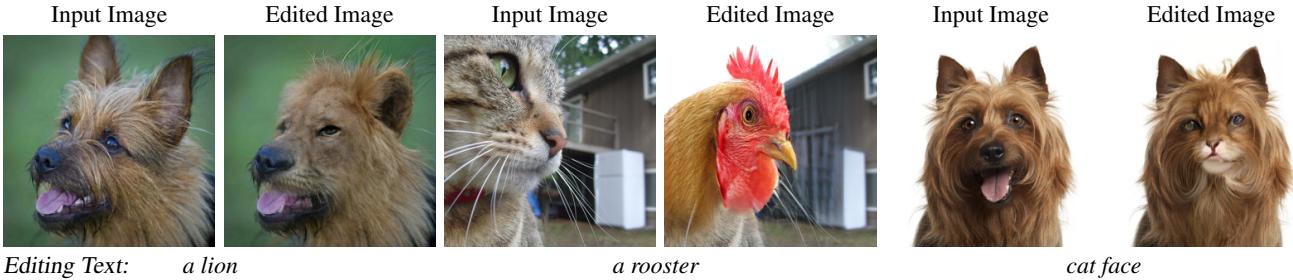


Figure 6. **Generated results using MaskGIT [8] as the image synthesis model.** Aside from using the Stable Diffusion, our method can also generate reasonable editing results with the non-autoregressive transformer-based MaskGIT.

for 5 epochs using Adam optimizer [28] with an initial learning rate of 0.003. We will make the codes and models available to the public.

4.1. Qualitative Evaluation

We assess the performance of our proposed method on a diverse set of high-quality images featuring various objects. Figure 4 shows that our approach takes an image and a language description to perform mask-free edits. We display complex text prompts that feature one category of object (the 1st row), multiple objects (the 2nd row), geometric relations (the 3rd row), and long paragraphs (the 4th row).

4.2. Comparisons with Prior Work

We compare our method with five state-of-the-art text-driven image editing approaches: **Plug-and-Play** [54] preserves the semantic layout of the source image by injecting features from the source image into the generation process of the target image. **InstructPix2Pix** [4] first utilizes GPT-3 and Stable Diffusion to produce paired training data for image editing. It then trains a diffusion model with classifier-free guidance under conditions. **Null-text Inversion** [38] enables text-based image editing with Stable Diffusion, using an initial DDIM inversion [13, 52] as a pivot for optimization, tuning only the null-text embedding in classifier-free guidance. **DiffEdit** [10] automatically generates masks for the regions that require editing by contrasting predictions of the diffusion model conditioned on different text prompts. **MasaCtrl** [6] performs text-based non-rigid image editing by converting self-attention in diffusion models into mutual self-attention, and it extracts the masks from the cross-attention maps as the editing regions.

In all experiments, we report the results of the compared methods using their official code, except for DiffEdit¹. Figure 14 displays the editing results from existing methods. We have the following observations, InstructPix2Pix inevitably leads to undesired changes in the global appearance (e.g., background). DiffEdit and MasaCtrl will yield un-

¹As there's no official code of DiffEdit, we use the code <https://github.com/Xiang-cd/DiffEdit-stable-diffusion>

Compared Methods	Preference for Ours
vs. Plug-and-Play [54]	80.5% \pm 1.9%
vs. InstructPix2Pix [4]	73.2% \pm 2.2%
vs. Null-text [38]	88.2% \pm 1.6%
vs. DiffEdit [10]	91.9% \pm 1.3%
vs. MasaCtrl [6]	90.8% \pm 1.4%
Average	84.9%

Table 1. **User studies.** We show the percentage (mean, std) of user preference for our approach over compared methods.

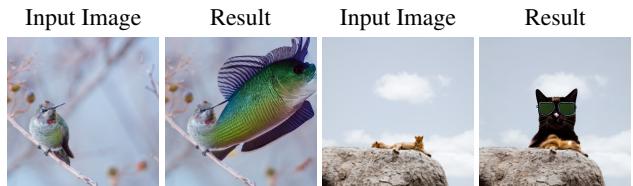


Figure 7. **Failure cases.** We show two failure cases generated by our method.

satisfactory results when using the more complex prompts containing multiple objects. Other methods generate less realistic results (e.g., coffee in the 4th row) or results that do not correspond with the text prompt (e.g., only one apple and pear in the 3rd row).

4.3. User Study

To evaluate the quality of the edited images, we conduct a user study using 60 input images and text prompts. We employ paired comparisons to measure user preference. In each test, we show an input image, a text prompt, and two edited images generated by our method and one of the compared approaches. We ask the subject to choose the one that performs better in coherence with the text prompt while maintaining fidelity to the input image.

There are 203 participants in this study, where each participant evaluates 40 pairs of images. The image set and compared method for each image are randomly selected for

each user. The order in each comparison pair is shuffled when presenting to each user. All the methods are compared for the same number of times.

Table 1 shows the user study results. The proposed method performs favorably against all five compared approaches. On average, our method is preferred in **84.9%** of all the comparisons, which demonstrates the effectiveness of the proposed method.

Failure cases. We present some failure cases of our approach to analyze the reasons. As shown in Figure 9, the failure results can be caused by improper anchor point initialization, especially when the anchor points fall into the background area.

4.4. Ablation Study

Compatibility with image synthesis models. To demonstrate the generalizability of the proposed method, we conduct experiments using MaskGIT [8], a distinct image generative transformer. As shown in Figure 6, we can generate results that adhere to the text prompt while preserving the background content. Note that the latent spaces within MaskGIT exclusively accommodate box-like masks, lacking the requisite precision for manipulating pixel-level masks in the context of image editing.

Effect of different loss components. To evaluate the influences of different loss components in our training loss, in Figure 8, we show the results generated without the directional loss \mathcal{L}_{Dir} that controls the directional edit, or without the structural loss \mathcal{L}_{Str} that focuses on preserving the appearance of the source image. We observe that the result without \mathcal{L}_{Dir} does not fully match the context of the text prompt, and the result without \mathcal{L}_{Str} fails to preserve the posture and shape of the object from the source image. In contrast, our method using all loss components can generate results that adhere to the text prompt while preserving the concept in the source image.

Effect of region generation methods. In Table 2, we present the user study results by comparing our method with two other baselines for bounding box generation. (1) **Random-anchor-random-size:** The editing regions are bounding boxes centered at anchor points uniformly sampled from the whole image, with height and width uniformly sampled from $[0, H]$ and $[0, W]$, where H and W are the height and width of the image. We clamp the regions exceeding the image boundary. (2) **DINO-anchor-random-size:** The editing regions are bounding boxes centered at anchor points selected from the DINO self-attention map, which are identical to those generated by our method, but with height and width uniformly sampled from $[0, H]$ and $[0, W]$. For both baselines, we use the same number of



Figure 8. **Effect of different loss components.** The 2nd and 3rd columns present results without \mathcal{L}_{Dir} and \mathcal{L}_{Str} respectively. The last column is generated by the model using all loss components.

Compared Methods	Preference for Ours
vs. Random-anchor-random-size	83.9% \pm 2.6%
vs. DINO-anchor-random-size	71.0% \pm 3.2%

Table 2. **Ablation study of region generation methods.** We show the percentage (mean, std) of user preference for our approach over two compared baselines.

anchor points as our method, and select the image with the highest quality score S to present to the user.

The results show that our method is preferred in **83.9%** compared with Random-anchor baseline. Even when compared to the competitive baseline where the anchor point is selected from the DINO self-attention map with the randomly chosen bounding box size, the proposed method is still preferred in **71.0%** of all comparisons. These results validate the effectiveness of our model in generating meaningful editing regions.

4.5. Limitation

We observe two limitations of our method. First, the performance is affected by the choice of the self-supervised model, particularly regarding anchor initialization. Second, since no user-specified region guidance is provided, the predicted region may include background areas, resulting in unintentional modifications in certain image contents. To address this, we plan to model the mask using more fine-grained representations (*e.g.*, patches).

5. Conclusion

In this paper, we propose a method for editing given images based on freely provided language descriptions, including paragraphs, without the need for user-specified edit regions. We introduce a region generation network and incorporate text-driven editing training losses to generate high-quality and realistic images. The proposed method seamlessly integrates with various image synthesis models. Experiments including user studies are conducted, demonstrating the competitive performance of our proposed method.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1, 3, 5
- [2] Omer Bar-Tal, Dolev Ofri-Amar, Rafaail Fridman, Yoni Kassten, and Tali Dekel. Text2LIVE: Text-driven layered image and video editing. In *ECCV*, 2022. 1, 3
- [3] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv:2103.10951*, 2021. 3
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 6, 7, 11, 13, 14
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 3
- [6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactr: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 1, 2, 3, 6, 7, 11, 13, 14
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 4
- [8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022. 1, 2, 5, 7, 8
- [9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv:2301.00704*, 2023. 1, 2, 3, 5
- [10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023. 2, 3, 6, 7, 11, 13, 14
- [11] Katherine Crowson. CLIP guided diffusion HQ 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj. 3
- [12] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashill Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In *ECCV*, 2022. 3
- [13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 3, 7
- [14] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022. 3
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 3
- [16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3
- [17] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3, 4
- [18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3
- [19] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 4
- [20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 3
- [21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 1
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 5
- [23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 3
- [24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv:1611.01144*, 2016. 4
- [25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3
- [26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Magic: Text-based real image editing with diffusion models. *arXiv:2210.09276*, 2022. 3
- [27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 3
- [28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 7
- [29] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, 2019. 4
- [30] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In *CVPR*, 2022. 1, 3
- [31] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 3
- [32] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *ICCV*, 2021. 1
- [33] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. In *NeurIPS*, 2022.

- [34] Yuanze Lin, Chen Wei, Huiyu Wang, Alan Yuille, and Cihang Xie. Smaug: Sparse masked autoencoder for efficient video-language pre-training. In *ICCV*, 2023. 1
- [35] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. FuseDream: Training-free text-to-image generation with improved CLIP+GAN space optimization. *arXiv:2112.01573*, 2021. 3
- [36] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023. 3
- [37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jianjun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 3
- [38] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 1, 6, 7, 11, 13, 14
- [39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021. 1, 3, 5
- [40] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 3
- [41] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *ICCV*, 2021. 3, 4
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 11
- [43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 3
- [44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022. 3
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5
- [47] 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. *arXiv:2208.12242*, 2022. 3
- [48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mabdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3
- [49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021. 3
- [50] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 4
- [51] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv:2109.14279*, 2021. 3
- [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 5, 7
- [53] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *ICLR*, 2022. 1, 3
- [54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 1, 6, 7, 11, 13, 14
- [55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 5
- [56] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 3
- [57] Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. Improving commonsense in vision-language models via knowledge graph riddles. *arXiv:2211.16504*, 2022. 1
- [58] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunnar Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv:2206.10789*, 2022. 3
- [59] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3
- [60] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 41(8):1947–1962, 2018. 3
- [61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3
- [62] Tianhao Zhang, Hung-Yu Tseng, Lu Jiang, Weilong Yang, Honglak Lee, and Irfan Essa. Text as neural operator: Image manipulation by text instruction. In *ACM MM*, pages 1893–1902, 2021. 1

A. Overview

This supplementary material includes the following additional content:

- Further implementation details in Section B.
- Additional experimental results in Section C.
- Additional visualization results in Section D, as shown in Figure 12, 13 and 14.

B. Implementation Details

In our experiments in the main paper, we chose 8 anchor points whose attention values are the top-8 largest. The training time for each sample is approximately 4 minutes, with a batch size set to 1 using two A5000 GPUs. For the diffusion model, we set the guidance scale and strength as 7.5 and 0.75, respectively.

C. Additional Experimental Results

C.1. Comparison with existing methods

In Figure 10 and 11 of this Appendix, we present additional results comparing our proposed method with state-of-the-art text-driven image editing baseline approaches, including Plug-and-Play [54], InstructPix2Pix [4], Null-text inversion [38], DiffEdit [10] and MasaCtrl [6].

We observe that the comparison results are consistent with the user studies reported in the main paper: our method generates images of higher quality that adhere to the text prompts while preserving the background content. The baseline methods appear to fall short in terms of generation quality and controllability by the prompt. For example, Plug-and-Play and InstructPix2Pix tend to modify the style of the entire image. The Null-text inversion method appears to have difficulty generating objects that align with the text prompts. Meanwhile, techniques like DiffEdit and MasaCtrl yield results that are of lower quality or poorer alignment with the text prompt.

C.2. Additional Ablation Study

In this section, we present three additional ablation studies employing qualitative metrics to complement the study based on user ratings reported in the main paper. Two metric scores are used: (1) the CLIP [42] text-to-image similarity score S_{t2i} , which evaluates the cosine similarity between the given prompt and the edited image; (2) the CLIP image-to-image similarity score S_{i2i} , which represents the cosine similarity between the source image and the edited image. We adopt the CLIP model initialized from ViT-B/16 weights to calculate the similarity scores. Results highlighted in blue denote the default settings used in our main experiments.

Loss Component	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
\mathcal{L}_{Clip}	0.301	0.801
$\mathcal{L}_{Clip} + L_{Str}$	0.294	0.806
$\mathcal{L}_{Clip} + L_{Str} + L_{Dir}$	0.301	0.805

Table 3. Ablation study on adopting different loss components.

# of region proposals	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
1	0.231	0.915
3	0.273	0.837
5	0.295	0.809
7	0.300	0.805
9	0.301	0.802

Table 4. Ablation study on the number of region proposals.

# of anchor points	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
1	0.275	0.824
4	0.296	0.805
6	0.301	0.802
8	0.300	0.805
10	0.298	0.803

Table 5. Ablation study on the number of anchor points.

Effect of different loss components. To evaluate the influences of the introduced loss components (*i.e.*, \mathcal{L}_{Clip} , \mathcal{L}_{Str} and \mathcal{L}_{Dir}) in our training loss, we report the metric scores of adopting different loss components in Table 3. The results show that using all losses yields the best overall score (*i.e.*, average of S_{t2i} and S_{i2i}), thereby verifying their contributions.

Effect of the number of region proposals. In Table 4, we evaluate the performance of using different numbers of region proposals. The results indicate that it’s likely to achieve the performance bottleneck with a larger region proposal (*e.g.*, 9 region proposals), and our method achieves reasonable quality (*e.g.*, a superb balance between S_{t2i} and S_{i2i}) even using 7 region proposals.

Effect of the number of anchor points. To analyze the influence of the number of anchor points, we display the ablation study in Table 5. It can be observed that when the number of anchor points increases from 1 to 8, the text-to-image similarity score S_{t2i} consistently increases. However, when the number is too large (*e.g.*, 10), the S_{t2i} decreases, it may be caused by the larger possibility of choosing noisy anchor points, which may be located in the background area, hurting the performance of editing.

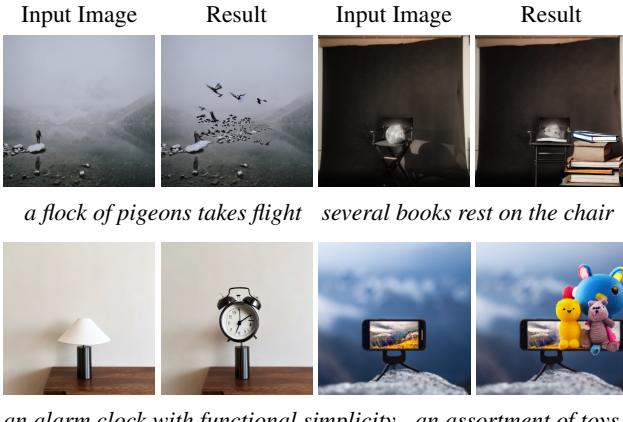


Figure 9. **Failure cases.** We show more failure cases generated by our method.

C.3. Analysis

More failure cases. In Figure 9, we display more failure examples generated by our method. We can see that the proposed method may generate unsatisfactory results when the anchor points are sampled from the background regions.

D. Visualizations

We provide more visualization results of text-driven image editing generated by our method in Figure 12, 13 and 14. Thanks to the flexibility of the learned bounding box guidance, our method is capable of handling a wide range of prompts, it can generate satisfactory results while preserving the original appearance of the background.

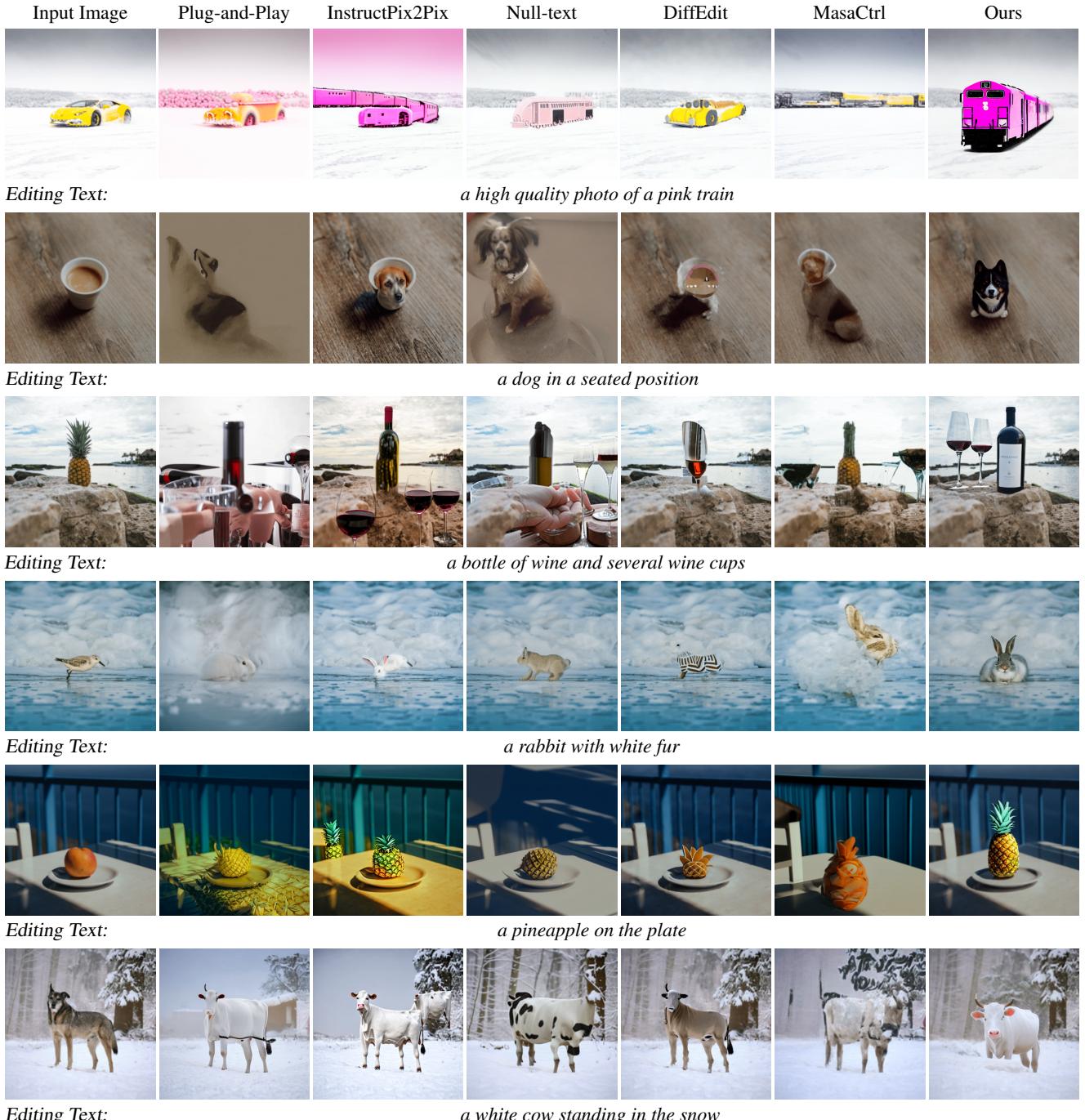


Figure 10. **Comparison with existing methods.** We compare our method with existing text-driven image editing methods. From left to right: Input image, Plug-and-Play [54], InstructPix2Pix [4], Null-text [38], DiffEdit [10], MasaCtrl [6], and ours.



Figure 11. **Comparison with existing methods.** We compare our method with existing text-driven image editing methods. From left to right: Input image, Plug-and-Play [54], InstructPix2Pix [4], Null-text [38], DiffEdit [10], MasaCtrl [6], and ours.

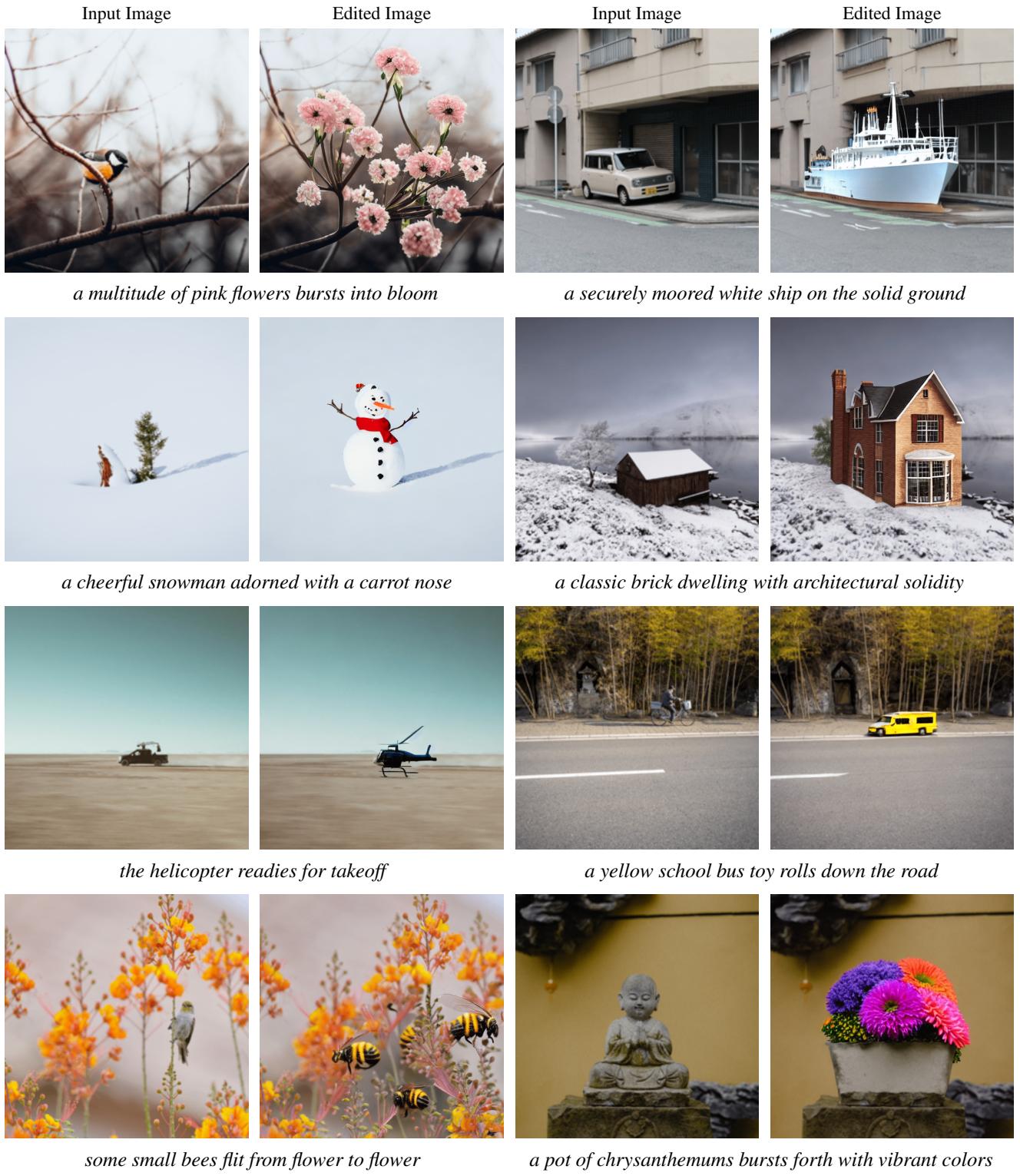


Figure 12. **Text-driven image editing results.** Given an input image and a language description, our method can generate realistic and relevant images without the need for user-specified regions for editing. It performs local image editing while preserving the image context.

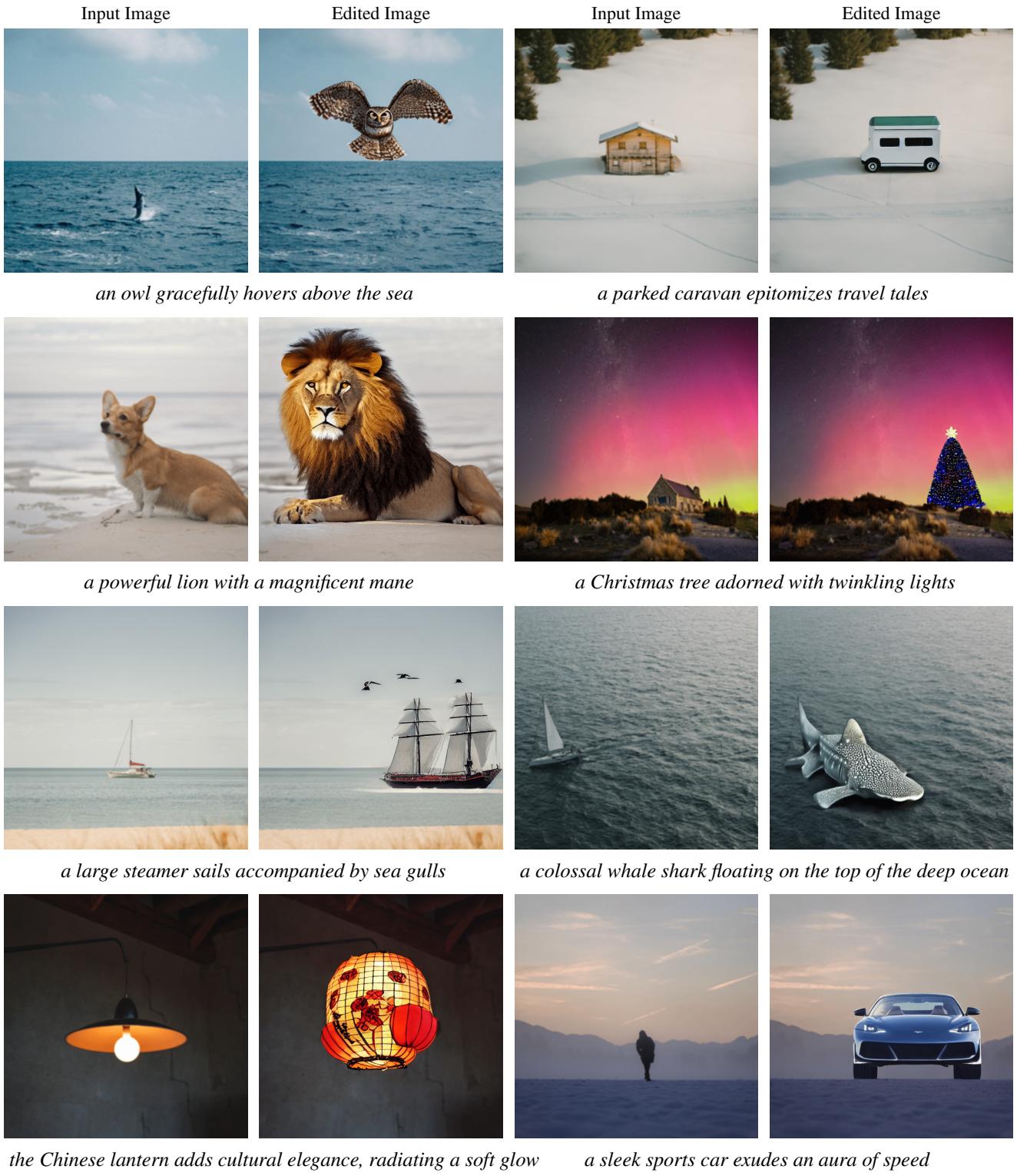


Figure 13. **Text-driven image editing results.** Given an input image and a language description, our method can generate realistic and relevant images without the need for user-specified regions for editing. It performs local image editing while preserving the image context.

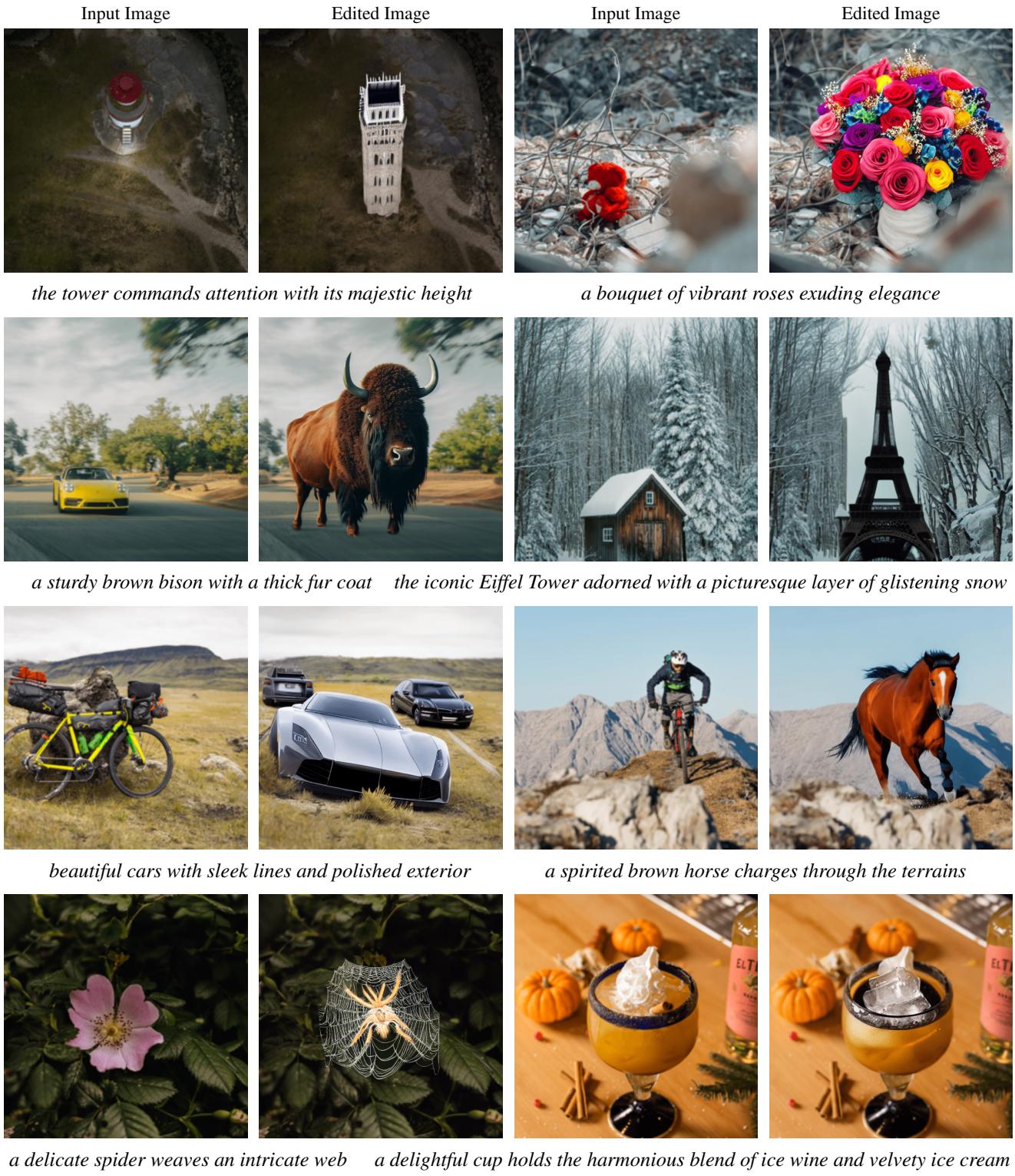


Figure 14. **Text-driven image editing results.** Given an input image and a language description, our method can generate realistic and relevant images without the need for user-specified regions for editing. It performs local image editing while preserving the image context.