Region-Aware Diffusion for Zero-shot Text-driven Image Editing

Nisha Huang
School of Artificial Intelligence, UCAS
NLPR, Institute of Automation, CAS
China
huangnisha2021@ia.ac.cn

Fan Tang
Institute of Computing Technology, CAS
China
tfan.108@gmail.com

Weiming Dong
NLPR, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
China
weiming.dong@ia.ac.cn

Tong-Yee Lee
National Cheng-Kung University
Taiwan
tonylee@mail.ncku.edu.tw

Changsheng Xu
NLPR, Institute of Automation, CAS
School of Artificial Intelligence, UCAS
China
csxu@nlpr.ia.ac.cn

Figure 1: The results of the proposed region-aware diffusion model (RDM). The texts adhere to the phrase rule "A → B", indicating that RDM transforms entity A into entity B.

Abstract

Image manipulation under the guidance of textual descriptions has recently received a broad range of attention. In this study, we focus on the regional editing of images with the guidance of given text prompts. Different from current mask-based image editing methods, we propose a novel region-aware diffusion model (RDM) for entity-level image editing, which could automatically locate the region of interest and replace it following given text prompts. To strike a balance between image fidelity and inference speed, we design the intensive diffusion pipeline by combing latent space diffusion and enhanced directional guidance. In addition, to preserve image content in non-edited regions, we introduce regional-aware entity editing to modify the region of interest and preserve the out-of-interest region. We validate the proposed RDM beyond the baseline methods through extensive qualitative and quantitative experiments. The results show that RDM outperforms the previous approaches in terms of visual quality, overall harmonization, non-editing region content preservation,

and text-image semantic consistency.

Keywords: ***Image Manipulation, Textual Guidance, Diffusion***

1. Introduction

In the actual world, image editing is highly sought-after. However, image content editing software is not easy to get started with and is more suited to professionals. Present image editing work is limited in terms of possible input images and editing operations. In recent years, great advances in deep generative image models [13, 10, 40] and visual language pre-training models [37] have made text-based image generation and manipulation interfaces possible.

There are different principal branches of image generation, such as inpainting, image translation [7], style transfer [18], and image manipulation [46, 35, 38, 2, 1]. Computational approaches [29, 5, 8] for modifying the style and appearance of objects in natural photographs have made remarkable progress, allowing beginner users to accomplish a wide range of editing effects. Nevertheless, it should be noted that prior text-based image manipulations work either did not allow for arbitrary text commands or image manipulation [33, 5, 24] or only allowed for modifications to the image's appearance properties or style [29, 19]. Controlling the localization of modifications normally requires the user to draw a region to specify [2, 1, 5], which adds to the complexity of the operation. In this study, we aim to eliminate all of the aforementioned constraints and restrictions to enable open image content modification with pure textual control utilizing cutting-edge image generation techniques.

Recently, there have been tremendous advances in multimodal deep learning that have opened the way for machines to achieve cross-modal communication and control. One of the large-scale multimodal pre-training models that have received many applications is the Contrastive Language Image Pretraining(CLIP) [37] model, which was pretrained on 400 million text-image samples. In parallel, various new image synthesis methods [36, 14, 26, 24, 38, 17] have highlighted the richness of the vast visual and linguistic realm encompassed by CLIP. Nonetheless, manipulating existing items in arbitrary, actual pictures continues to remain tricky. The present mainstream techniques combine CLIP with pre-trained GANs generators, however, the input picture domain is constrained. There has recently been a tremendous amount of interest in diffusion models, which generate high-quality, diversified pictures. Image manipulation at the pixel level, on the other hand, leads to extended generation times and excessive computer resource consumption.

After investigation, we found that applying the diffusion process in the latent space of pre-trained autoencoders [40] can speed up inference and reduce the consumption of computational resources. Nevertheless, previous latent diffusion models still fall short in terms of generating image realism. To improve image realism and to enhance the consistency of the editing results with the guide text, we introduced classifier-free guidance [16] at each step of the diffusion. Based on these, we develop an intensive high-performance diffusion model editing (Diff-Edit) framework for zero-shot text-driven image editing. Specifically, our method enables the editing of image content that satisfies an arbitrarily given text prompt (as shown in Fig. 1). For example, given an image of the dog, and the positioning text: "A dog", our work can position the corresponding editing area. Then, based on the target text: "A cat", a high-quality, realistic, and varied image can finally be composed.

To enable the user to specify the area and objects to be modified and the objects to be created, simply and intuitively, we present a cross-modal entity calibration component. It can locate and adjust the text-relevant picture tokens given the positional textual guidance t_1 to correctly identify the entities for modification. We observed that by feeding the input image into the encoder together with the mask, the content outside of the mask appeared to transform unexpectedly during the diffusion process for image modification. To retain extraneous content to a greater extent, we perform further diffusion at each diffusion step where it blends the clip-guided diffusion result with the corresponding noisy version of the input image. In addition, we build relevant loss functions that protect non-editing domains to constrain the generative process. We incorporated the clip gradient into the classifier-free guidance to make the edited results more favorable to humans and to make the content generated in the edit area more consistent with the semantic content of the target textual guidance t_2. Overall, Diff-Edit implements enhanced directional guidance in the latent space to generate rapid, high-quality, realistic, and text-compliant editing results.

Quantitative and qualitative comparisons with previous approaches reveal that our method can better manipulate the entities of an image through text while leaving the background region unaffected. As shown in Figs. 1, 3 and 4, Diff-Edit is capable of producing realistic and high-quality outcomes in terms of object content change when guided by various image inputs and text descriptions. The main contributions of this work are summarized as follows:

- We propose Diff-Edit, an entity-level zero-shot text-driven image editing framework based on the intensive diffusion model.
- We introduced spatial location masks into each step of the diffusion sampling and created non-editing region-preserving loss functions to obtain edited results without stitching traces and well-preserved unedited regions.

- We manipulate the diffusion step in latent space and embed enhanced directional guidance structures to enhance image realism and improve the consistency of the control text with the editing result.

- The quantitative and qualitative experimental results show that Diff-Edit outperforms baseline methods in terms of quality, veracity, and diversity in text-guided image editing, and achieves superior results.

2. Related Work

In this section, we review the existing works on text-guided image manipulation and diffusion models, which motivates us to design and implement our application.

2.1. Text-guided Image Manipulation

Synthesizing an image based on a text description is an ambitious problem that has advanced tremendously in recent years. Initial RNN-based works [32] were surpassed by generative adversarial approaches. There have been seminal works based on conditional GANs in image editing [11, 27, 34]. Paint By Word [5] firstly addressed the problem of zero-shot semantic image painting using CLIP [37] in combination with StyleGAN2 [21] and BigGAN [6]. It can only alter the appearance of a picture, such as its color and texture, but it cannot generate new entities. ManiGAN [27] semantically edits parts of an image matching a given text that describes certain attributes and preserving the contents irrelevant to the text. However, the expressiveness of the text is restricted by such multimodal GAN-based approaches. Both Paint By Word [5] and ManiGAN [27] are restricted to specific image domains and are not applicable to open natural images.

SDG [30] and DiffusionCLIP [24] are proposed to utilize a diffusion model in order to perform global text-guided image manipulations. GLIDE [35] and DALL·E 2 [38] focus on text-driven open domain image synthesis, as well as local image editing. GLIDE fine-tunes its text-to-image synthesis model for image inpainting. DALL·E 2 performs inpainting results while lacking discussion in the paper. Both of them are implemented with the idea of integrating image generators and joint text-image encoders into their architectures. They all contain pre-trained models with large-scale datasets of numerous text-image pairs, while neither of them has released their complete models. Later, Blended Diffusion [2] and Latent Blended Diffusion [1] were proposed as the solution for local text-guided editing of real generic images. However, these methods require the user to draw the extra mask manually from which the image is edited, without a precise and automatic editing area.

2.2. Diffusion Models

Diffusion models, also known as score-based generative models, are a strong family of generative models that have recently evolved. This fresh idea on the subject of image generation was proposed by Sohl-Dickstein et al. [44]. Current works [10, 20, 45] demonstrate astonishing results in high-fidelity image generation, often even outperforming generative adversarial networks. Importantly, [35, 38, 42] additionally offer strong sample diversity and faithful mode coverage of the learned data distribution. As a result, diffusion models are ideal for learning models from complicated and varied data.

Specifically, diffusion models consist of one forward process and one reverse process. The forward diffusion process maps data to noise by gradually perturbing the input data. The reverse process performs iterative denoising from pure random noise. The diffusion models are used to generate data by simply passing randomly sampled noise through the learned denoising process. Diffusion models have already been utilized in many successful applications, such as image generation [10, 20, 35, 38, 42, 45], image segmentation [4], image-to-image translation [7], superresolution [22, 40], and image editing [1, 2, 35, 38]. Text2LIVE [3] applies the text to edit the appearance of existing objects. It concentrates on generating an edit layer composited over the original input, rather than removing or replacing objects of the input image, as we do.

Even though the approaches described above produce cutting-edge outcomes for picture data generation, one disadvantage of diffusion models is the sluggish reverse denoising process. In addition, traditional diffusion models operate in pixel space leading to consuming a lot of memory. Latent diffusion models (LDMs) [40] have been proposed to expedite the sampling process and reduce computational requirements compared to pixel-based diffusion models. LDMs are trained to build latent visual representations and to perform the diffusion process across a lower-dimensional latent space. [40] shows that it has achieved a new state-of-the-art and highly competitive performance on various computer vision tasks. However, there is still a need to improve performance in terms of image fidelity and text-image semantic consistency. Therefore, our RDM is designed to retain the benefits of LDM speed while taking into account image quality and text-image alignment.

3. Method

3.1. Overview

The proposed RDM is a framework for solving entity-level zero-shot text-driven image editing tasks, as depicted in Fig. 2. Our goal is to implement editing of the input image x_0 through the control by a pair of text prompts (t_1, t_2). The positioning text t_1 is used to position the edited entity,

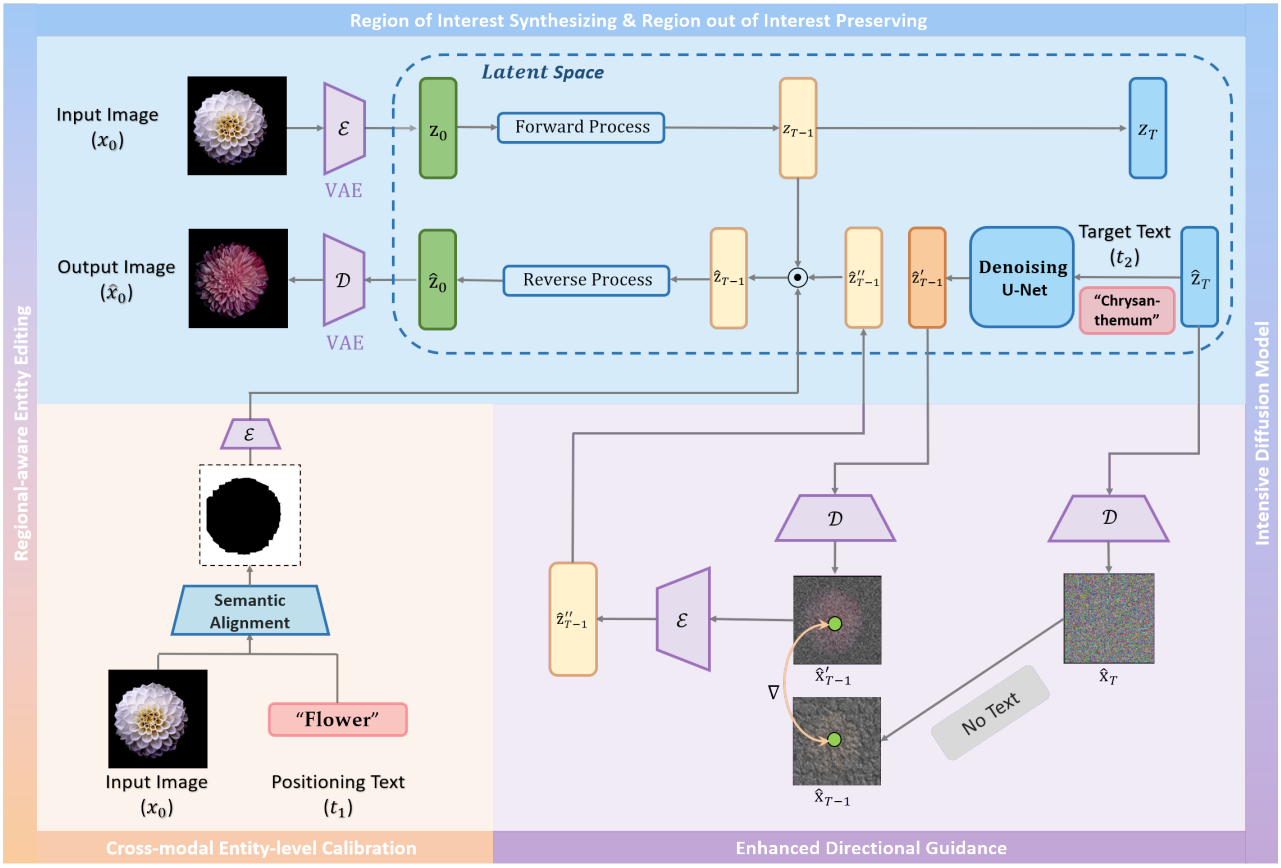

Figure 2: The overall framework of our method for zero-shot text-driven image editing.

and the target text t_2 is used to generate the new entity. In Section 3.2, we illustrate the concrete composition of the intensive diffusion model. In Section 3.3, we explain how regional-aware entity editing can be achieved through text.

3.2. Intensive Diffusion Model

The diffusion model [44] is a generator that can be used to generate images. The diffusion process is divided into a forward process, which adds random noise to the input image x_0, and a backward process, which removes the noise and generates the image $\hat{x}_0$. Unlike traditional diffusion models [10, 15], we do not perform the diffusion process at the pixel level. Some recent works [1, 40] have demonstrated that performing the diffusion process in the latent space can reduce computational consumption and speed up the sampling process. The denoising UNet learns to remove the noise, and after T steps of noise removal, generates the output image $\hat{x}_0$. However, there is damage to image generation quality and text-image consistency by performing the diffusion process in the latent space. To improve this issue, we further introduce a component of enhanced directional guidance.

Latent Representations. As mentioned above, we perform a diffusion step in the latent space [40] to reduce complexity and provide efficient image processing. An autoencoder VAE [25] is used to accomplish perceptual picture compression. The diffusion model directly operates on the lower-dimensional latent space, taking advantage of image-specific inductive biases. This allows the underlying autoencoder to be constructed primarily from two-dimensional convolutional layers and uses a re-weighting bound to further focus the target on the perceptually most important bits, which are denoted as:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t\right)\|_2^2\right]. \quad (1)$$

Our model's determination $\epsilon_\theta(z_t, t)$ is implemented as a time-conditional UNet [41]. Given that the forward process is fixed, z_t can be conveniently acquired from $\mathcal{E}$ during training, and $\mathcal{D}$ can decode samples from $p(z)$ to pixel space.

Enhanced Directional Guidance. To reinforce the editing direction of the source region to follow the target text,

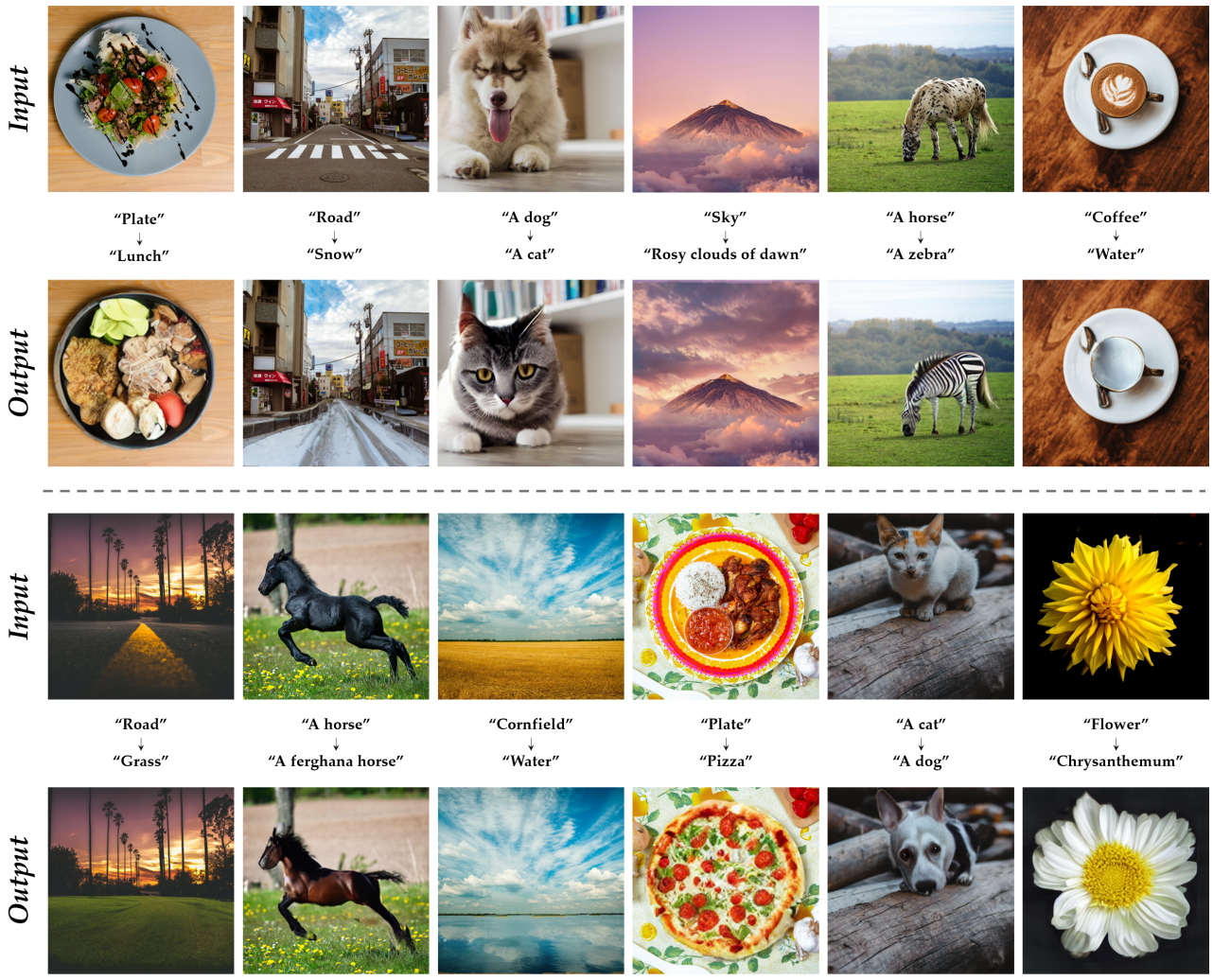

Figure 3: More manipulation results by RDM.

we attempt to modify a classifier-free guidance [16] to strengthen cross-modal guidance. It is a strategy for guiding diffusion models without necessitating the training of a separate classifier model. Generally, classifier-free guidance offers two benefits. For starters, rather than relying on the knowledge of a separate (and perhaps smaller) categorization model, it allows a single model to leverage its experience while guiding. Second, it simplifies directing when conditioned on information that is difficult to predict using a classifier.

In order to provide classifier-free guidance, the tag y in a class-conditional diffusion model $\epsilon_\theta\left(x_t \mid y\right)$ is replaced with a null tag $\emptyset$ throughout the training process. The output of the model is further extended in the direction of $\epsilon_\theta\left(x_t \mid y\right)$ and away from $\epsilon_\theta\left(x_t \mid \emptyset\right)$ during sampling:

$$\hat{\epsilon}_\theta\left(x_t \mid y\right) = \epsilon_\theta\left(x_t \mid \emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t \mid y\right) - \epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{2}$$

The recommended guidance scale is $s = 5$. This equation was inspired by the classifier.

$$p^i\left(y \mid x_t\right) \propto \frac{p\left(x_t \mid y\right)}{p\left(x_t\right)}, \tag{3}$$

where the function of the true scores is used to represent the gradient ϵ^*,

$$\begin{aligned} \nabla_{x_t} \log p^i\left(x_t \mid y\right) &\propto \nabla_{x_t} \log p\left(x_t \mid y\right) - \nabla_{x_t} \log p\left(x_t\right), \\ &\propto \epsilon^*\left(x_t \mid y\right) - \epsilon^*\left(x_t\right). \end{aligned} \tag{4}$$

The modified prediction $\hat{\epsilon}$ is subsequently employed to guide us toward the target text prompts t_2, as demonstrated in Algorithm 1:

$$\hat{\epsilon}_\theta\left(x_t \mid t_2\right) = \epsilon_\theta\left(x_t \mid \emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t \mid t_2\right) - \epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{5}$$

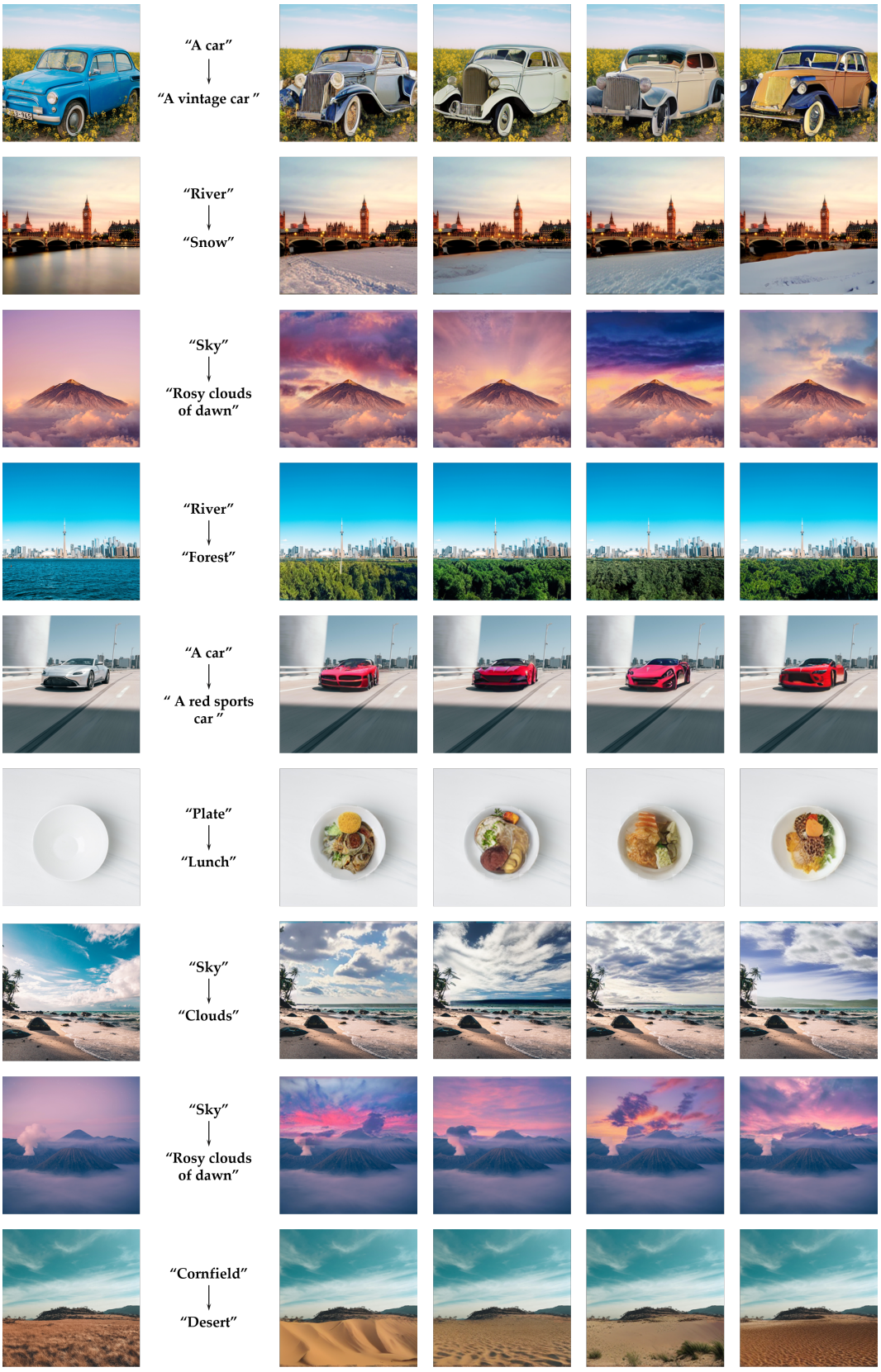

Figure 4: More manipulation results by RDM.

3.3. Regional-aware Entity Editing

Cross-modal Entity-level Calibration. To generate a binary segmentation mask m based on the localized text t_1, we design a cross-modal entity-level calibration module, consisting of a pre-trained CLIP model and a thin conditional segmentation layer (decoder). First, the positioning text t_1 is fed into the CLIP text transformer to obtain the conditional vector. Motivated by feature-wise transformations [12, 31], the conditional vector is used to modulate the input activation of the decoder. This enables the decoder to associate the activation within CLIP with the output segmentation and to inform the decoder about the segmentation's target. The input image x_0 is passed through the CLIP visual transformer to get $\mathbb{R}^{W\times H\times 3}$. Afterward, the activations extracted at layers $S=[3,7,9]$ are added to the decoder internal activations at the embedding size $F=64$ before each transformer block. Besides, CLIP ViT-B/16 is used with token patch size: $P=16$. The decoder generates the binary segmentation by applying a linear projection on the tokens of its transformer (last layer):

$$\mathbb{R}^{\left(1+\frac{W}{P}\times\frac{H}{P}\right)\times F}\mapsto\mathbb{R}^{W\times H}. \tag{6}$$

To associate CLIP's capabilities with segmentation results, a generic binary prediction setting is employed. We threshold the binary segmentation for spatial mask m, with a threshold K ranging from 0 to 255, which we usually take as 150.

Region of Interest Synthesizing. To make the region of interest could be edited according to the text prompt, we leverage a pre-trained ViT-L/14 CLIP [37] model for text-driven image content manipulation. The cosine distance between the CLIP embedding of the denoised image $\hat{x}_t$ during diffusion and the CLIP embedding of the text prompt t_2 may be used to specify the CLIP-based loss, or $\mathcal{L}_{CLIP}$. Target textual prompt t_2 is embedded into the embedding space, which is defined as E_L. And a time-dependent image encoder for noisy images is referred to as E_I. We define the language guidance function using the cosine distance, which measures how similar the embeddings E_I and E_L are to one another. The text guidance function can be defined as:

$$\mathcal{L}_{CLIP}\left(\hat{x}_t,t_2,m\right)=E_I\left(\hat{x}_t\odot m\right)\cdot E_L(t_2). \tag{7}$$

The aforementioned process is not subject to any extra non-editing region restrictions. Despite being assessed inside the region that is being edited, $\mathcal{L}_{CLIP}$ also affects non-editing regions. We provide the equivalent approach below to deal with this problem.

Region out of Interest Preserving. Non-editing region preserving (NERP) is not present in the aforementioned procedure, which starts with isotropic Gaussian noise. As a result, even though $\mathcal{L}_{CLIP}$ is assessed inside the masked zone, it still has an impact on the whole image. To ameliorate this problem, we encode the mask m into the latent space to get m_{latent}, and blend it into the diffusion process as follows. The latent for the subsequent latent diffusion step is produced by blending the two results using the resized mask, i.e. $\hat{z}_t''\odot m_{\text{latent}}+z_t\odot(1-m_{\text{latent}})$. As shown in Fig. 2, where z'' represents the result generated by a latent diffusion followed by enhanced directional guidance in the reverse process. And z_{nd} is the result of superimposing the corresponding noise on the input image in the forward process. At each denoising step, the entire latent is modified, but the subsequent blending enforces the parts outside m_{latent} to remain the same. In this stage, the backdrop is tightly preserved by replacing the whole area outside the mask with the comparable region from the input image. The subsequent latent denoising process ensures coherence even though the resultant blended latent is not always coherent. Following the completion of the latent diffusion process, we decode the resulting latent to the output image using the decoder $\mathcal{D}(z)$, as demonstrated in Algorithm 1.

Besides, a non-editing region preserving loss $\mathcal{L}_{NERP}$ is applied to direct the diffusion outside the mask to direct the surrounding area towards the input image:

$$\mathcal{L}_{NERP}\left(x_0,\hat{x}_t,m\right)=d\left(x_0\odot(1-m),\hat{x}_t\odot(1-m)\right), \tag{8}$$

$$d\left(a,b\right)=\lambda_1\left(LPIPS\left(a,b\right)\right)+\lambda_2\left(MSE\left(a,b\right)\right), \tag{9}$$

$$a=x_0\odot(1-m),b=\hat{x}_t\odot(1-m). \tag{10}$$

where $LPIPS$ is the learned perceptual image patch similarity measure and MSE is the L_2 norm of the pixel-wise difference between the images. λ_1 and λ_1 are all set to 0.5.

4. Experiments

4.1. Implementation Details

For the diffusion model, we used a pre-trained latent diffusion model [40] of resolution 256×256, which has 1.45 billion parameters trained on the LAION-400M [43] database. For the CLIP model, we used ViT-L/14 released by OpenAI for the Vision Transformer [37]. The output size of RDM is 256×256. For sampling, we set λ_1, λ_2 and clip guidance scale to 0.5, 0.5 and 150, respectively. To ensure the quality of the results and to maintain the consistency of the parameters, the diffusion step and the time step used for the experiments in this work are both set to 100. It takes three seconds to generate a 256×256 image on a single GeForce RTX 3090 GPU by RDM, which is comparable to latent diffusion (three seconds) and surpassing most diffusion models (15 seconds, 27 seconds, and 3 minutes respectively for GLIDE, blended diffusion, and clip-guided diffusion).

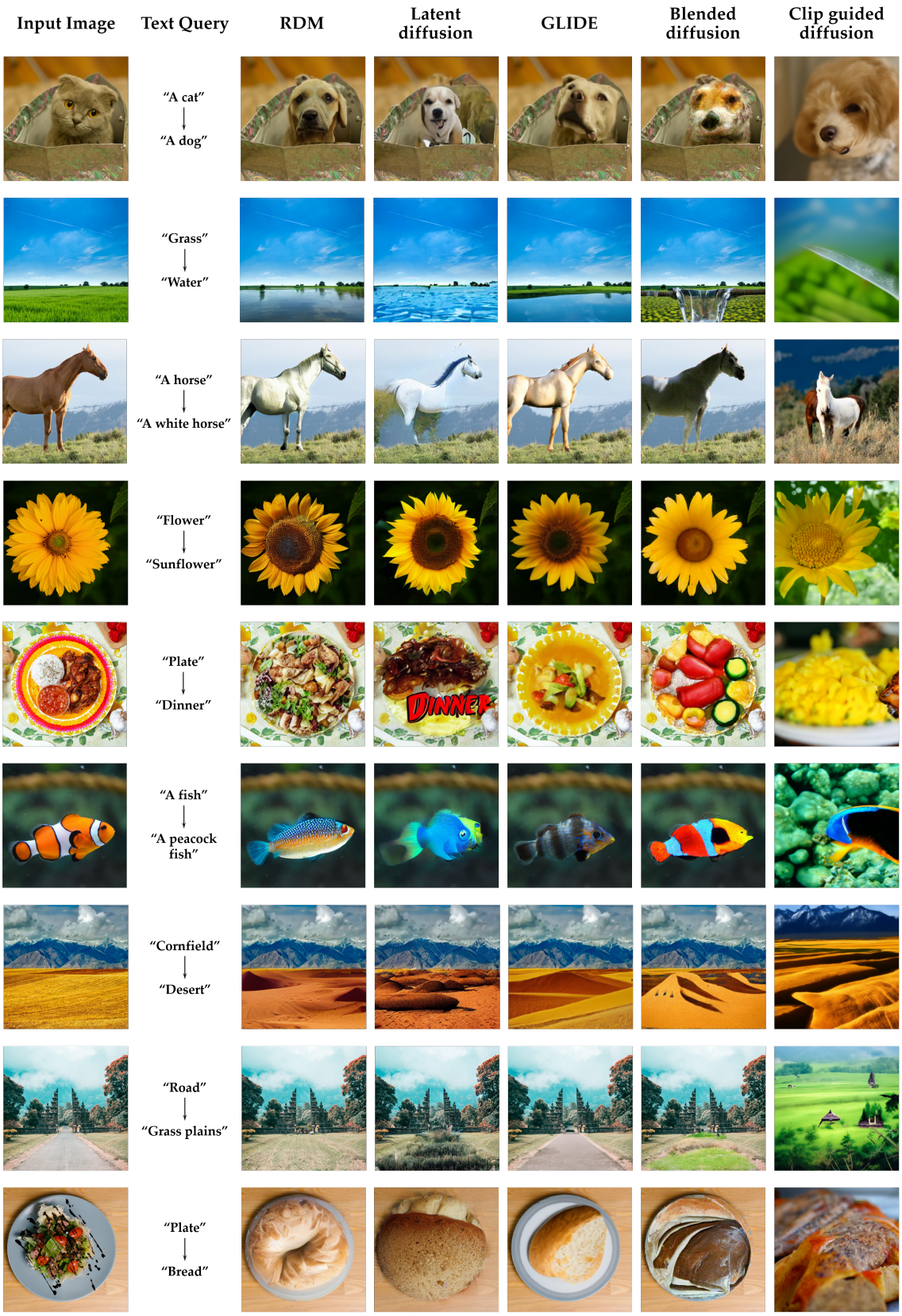

Figure 5: Comparison with SOTAs including latent diffusion, GLIDE, blended diffusion, and CLIP-guided diffusion.

Figure 6: Impact of mask threshold on the manipulation results.

Algorithm 1 Text guided hybrid diffusion sampling, given a latent diffusion model $(\mu_\theta(z_t), \Sigma_\theta(z_t))$

Input: The input image x_0, text guidance t_2, gradient scale s, diffusion steps T.
Output: generated image x_0 according to text guidance t_2.

$t = T$
$z_0 = \mathcal{E}(x_0)$
$z_t \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
$\hat{z}_T = z_T$
repeat
  $t - 1 \leftarrow t$
  $\mu, \Sigma \leftarrow \mu_\theta(\hat{z}_t), \Sigma_\theta(\hat{z}_t)$
  $\hat{z}'_{t-1} =$ denoise $(\hat{z}_t, t_2, t)$
  $\hat{x}'_{t-1} = \mathcal{D}(\hat{z}'_{t-1})$
  $\hat{\epsilon}_\theta(\hat{x}_{t-1} \mid t_2) \leftarrow (1-s)\cdot\epsilon_\theta(\hat{x}_{t-1} \mid \emptyset) + s\cdot\epsilon_\theta(\hat{x}'_{t-1} \mid t_2)$
  $\mathcal{L} \leftarrow \mathcal{L}_{CLIP}(\hat{x}'_{t-1}, t_2, m) + \mathcal{L}_{NERP}(x_0, \hat{x}'_{t-1}, m)$
  $\hat{z}''_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu + s\Sigma\nabla_{\hat{x}_{t-1}}\mathcal{L})$
  $\hat{z}_{t-1} = \hat{z}''_{t-1} \odot m_{\text{latent}} + z_{t-1} \odot (1 - m_{\text{latent}})$
until $t < 0$
$\hat{x}_0 = \mathcal{D}(\hat{z}_0)$

4.2. Qualitative Evaluation

We tested our approach on a variety of real-world images and edited texts. The images were sourced from the web and contained a variety of object categories, including animals, food, landscapes, and others. Fig. 3 shows the results by using different images and text prompts as inputs. The results of our method were successful in editing arbitrary images. As shown in the images of the flower, the petals are generated with a very fine and natural texture. For food manipulation, it is possible to see each piece of food. In the editing of animals, RDM can compose the new animal very well even though the animal being edited has multiple complex poses. For the vehicle samples, the edits are positioned precisely, even if the bodywork is partially obscured. As shown in Fig. 4, RDM can obtain a variety of different results for the same image and text input. In a nutshell, our method successfully applies text control to the editing of image entity content in high quality and diversity.

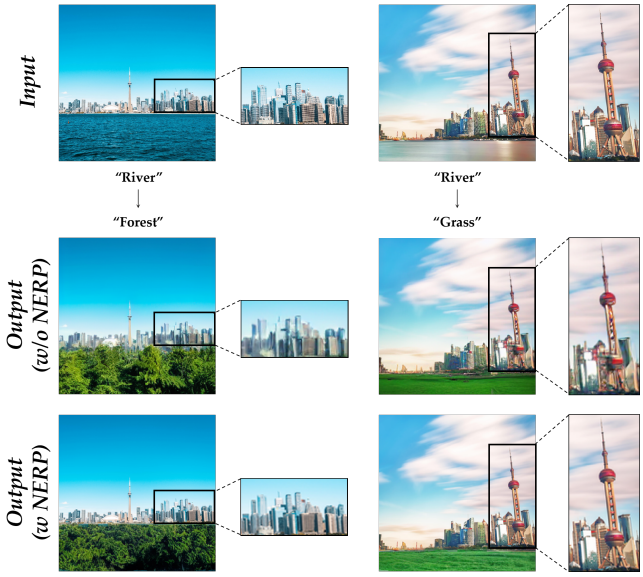

Figure 7: Qualitative comparison of our RMD with or without non-editing region preserving component.

4.3. Comparison with the State-of-the-arts

In this section, we compare RDM with SOTA text-driven image editing methods including Latent diffusion [39], GLIDE [35], blended diffusion [2] and CLIP-guided diffusion [9]. Fig. 5 shows comparisons to baselines of real-world images. The main differences between our approach and these methods are as follows. Latent diffusion [39], GLIDE [35], and blended diffusion [2] require the user to provide a mask for the area to be edited. Their out-of-interest regions are not involved in diffusion, so there are no content-preserving issues. CLIP-guided diffusion [9] cannot be modified for local areas of the image.

In this comparison, the masks required for these methods are provided by masks generated by RDM's cross-modal entity-level calibration component. As can be seen from

the results (in column (d) of Fig. 5) of latent diffusion [40], even though a strict edit mask is provided, the new content generated does not match the area of the mask and always generates new content that is smaller than the masked area. The results of GLIDE [35] (in column (e) of Fig. 5) lack details; for example, the petals of the sunflower are very smooth; the skin of the horse has no texture, and the hair on the horse's back is lacking; the "grass plains" generation fails and GLIDE does not understand the content of this text prompt well. The images (in column (f) of Fig. 5) produced by blended diffusion [2] lack realism and are artificial. The results (in column (g) of Fig. 5) of CLIP-guided diffusion [9] do not preserve the content of the unmodified areas of the image and tend to over-vignette.

4.4. Quantitative Evaluation

CLIP score. To assess the semantic alignment of the text descriptions and modified images, we compute the CLIP score, which is the cosine similarity between their embeddings derived with CLIP encoders. Because we utilize the ViT-B/14 CLIP model during the inference process, for a fair comparison, we compute the CLIP score using the ViT-B/32 CLIP [37] model. A higher CLIP score suggests that the input texts and altered images are semantically aligned. The results (the 1^{st} row of Table 1) show that RDM outperforms baseline models, which indicates a superior in terms of text image consistency.

SFID. To assess the quality of manipulated images, we employ the SFID [23], a simplified FID that avoids the numerical instability associated with a limited number of sample feature distributions. We calculated the SFID score between the different methods to obtain the 2^{nd} row of Table 1. Intuitively, the lower the SFID scores, the higher quality of the manipulated images on the COCO [28] dataset. The findings reveal that GLIDE has the highest SFID score and RDM is the second-best.

Harmonization score. To evaluate the degree of harmonization between the edited and unedited parts, we used DoveNet [47] as a quantitative evaluation method. We utilize DoveNet to generate harmonized images and calculate the PSNR values between the harmonized images and manipulated images. Image harmonization (IH) scores are shown in the 3^{rd} row of Table 1. The lower the IH score, the more synchronized the edited and unedited parts are. Our approach outperforms the baseline model by a margin, achieving the highest harmonization score. Our incorporation of the mask into the diffusion process resulted in better performance in terms of consistency between edited and unedited regions.

User study. Next, we conducted a user perception evaluation. Participants were asked to choose: which image produced a higher quality image, which image turned out to be more harmonious and had less visible editing marks (seams), and which image editing turned out to be more in line with the text content. The judgments of 70 participants were collected across 36 image-text combinations and gathered 2520 votes. Each comparison was made without revealing which image was by which method. We have included Table. 2 report the percentage of votes in favor of the RDM model. It follows that our method is capable of generating the kind of image editing results that humans prefer.

4.5. Ablation Study

Effects of cross-modal entity-level calibration component. To investigate the impact of the cross-modal entity-level calibration component on the quality and semantic content of the generated images, we tested image editing at different thresholds K. Fig. 6 shows that when the value of K is set small, the segmented editable scene is also smaller than the area occupied by the vehicles in the input image to varying degrees. The orientation of the generated vehicles does not match the orientation of the vehicles in the input image, i.e., there is a deviation in semantic consistency. As the value of K increases, the editable area increases, and the orientation of the vehicles tends to be the same, but after a certain point, the image editing results do not change significantly.

Effects of the non-editing region preserving component. Fig. 7 illustrates the effect without the non-editing region preserving (NERP) component. As can be seen, this component allows us to achieve the retention of image content in out-of-interest areas of the input image. We qualitatively compare the editing results with and without the component. The results are shown in Fig. 7. As is visible, without the original image with the corresponding noise and the editing component together with the diffusion inference process (w/o NERP), the buildings in the image produce a large deformation with abnormal artifacts and textures, resulting in a loss of overall image quality. Thus, the out-of-interest region preserving component significantly improves the abnormal deformation and damage to the image content outside the edited area.

4.6. Failed case.

We have observed through some experiments that CLIP has a significant preference for particular solutions for various editors. As shown in Fig. 8, given a picture of a cup with coffee, we wanted to implement a "coffee" to "water" image edit. The result shows that the liquid in the cup is successfully turned into water. However, the text "water" is closely associated with a transparent cup, so it is possible

Table 1: Quantitative comparisons for image manipulation. We compute the average CLIP score, SFID score and image harmonization (IH) score to measure visual quality, text consistency, and image harmonization. The best results are highlighted in **bold** while the second best results are marked with an underline

	RDM	Latent diffusion	GLIDE	Blended diffusion	CLIP guided diffusion
CLIP score ↑	**0.849**	0.824	0.845	0.822	0.843
SFID score ↓	6.54	9.29	**5.88**	17.37	23.42
IH score ↓	**20.7**	22.0	21.8	23.1	/

Table 2: User study results. Each number represents the percentage of votes received by the other models' outcomes as compared to our results.

		Latent diffusion	GLIDE	Blended diffusion	CLIP guided diffusion
Preference rate	Visual quality	5.71%	19.05%	6.67%	15.71%
	Invisible seams	6.67%	20.95%	10.00%	16.67%
	Text consistency	10.48%	24.29%	21.43%	11.90%

Figure 8: Examples of failure cases are given source images.

that the cup could also be turned into a glass. As shown in the second column of Fig. 8 the coffee that is in the air being injected into the cup fails to successfully turn into water. In addition, liquids that were in the air and not in the cup were not successfully edited. In summary, our method is more suited to generating a new entity rather than modifying the properties of the original entity.

5. Conclusion and Future Work

This paper investigates for the first time a new problem setting - the editing of the content of specified entities in images, guided by arbitrary text. Solving this task requires control over the positioning of the edits, the quality, and fidelity of the edited and unedited content, the consistency of text guidance and image manipulation, etc. To address these issues, we propose a new framework, a region-aware diffusion model with semantic alignment and generation capabilities, for manipulating images at the entity level. We provide a new tool for users to modify images by simply presenting their requirements in text. In the future, we hope to expand the applications of RDM, such as more flexible control of the position, shape, and size of the generated area.

References

[1] O. Avrahami, O. Fried, and D. Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2, 3, 4

[2] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 2, 3, 9, 10

[3] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022. 3

[4] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[5] D. Bau, A. Andonian, A. Cui, Y. Park, A. Jahanian, A. Oliva, and A. Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2, 3

[6] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[7] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 2, 3

[8] G. Couairon, A. Grechka, J. Verbeek, H. Schwenk, and M. Cord. Flexit: Towards flexible semantic image trans-

lation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18270–18279, 2022. 2

[9] K. Crowson. Clip guided diffusion hq 256x256., 2022. 9, 10

[10] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 2, 3, 4

[11] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5706–5714, 2017. 3

[12] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio. Feature-wise transformations. *Distill*, page e11, 2018. 7

[13] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 2

[14] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, pages 1–13, 2022. 2

[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 4

[16] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 5

[17] N. Huang, F. Tang, W. Dong, and C. Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *ACM International Conference on Multimedia (ACM MM)*, 2022. 2

[18] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2

[19] W. Jiang, N. Xu, J. Wang, C. Gao, J. Shi, Z. Lin, and S. Liu. Language-guided global image editing via cross-modal cyclic mechanism. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2115–2124, 2021. 2

[20] A. Jolicoeur-Martineau, R. Piché-Taillefer, I. Mitliagkas, and R. T. des Combes. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 3

[22] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 3

[23] C.-I. Kim, M. Kim, S. Jung, and E. Hwang. Simplified fréchet distance for generative adversarial nets. *Sensors*, page 1548, 2020. 10

[24] G. Kim, T. Kwon, and J. C. Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022. 2, 3

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[26] G. Kwon and J. C. Ye. Clipstyler: Image style transfer with a single text condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, 2022. 2

[27] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr. Manigan: Text-guided image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7880–7889, 2020. 3

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 10

[29] X. Liu, Z. Lin, J. Zhang, H. Zhao, Q. Tran, X. Wang, and H. Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2020. 2

[30] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 3

[31] T. Lüddecke and A. Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 7

[32] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 3

[33] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018. 2

[34] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances Neural Information Processing Systems (NeurIPS)*, pages 42–51, 2018. 3

[35] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3, 9, 10

[36] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 3, 7, 10

[38] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[39] R. Rombach. Latent diffusion laion 400m model text to image inpainting., 2022. 9

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion

models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 4, 7, 10

[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[42] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[43] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[44] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 3, 4

[45] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[46] J. Wang, G. Lu, H. Xu, Z. Li, C. Xu, and Y. Fu. Manitrans: Entity-level text-guided image manipulation via token-wise semantic alignment and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10707–10717, 2022. 2

[47] C. Wenyan, Z. Jianfu, N. Li, L. Liu, L. Zhixin, L. Weiyuan, and Z. Liqing. Dovenet: Deep image harmonization via domain verification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 10