Multi-Region Text-Driven Manipulation of Diffusion Imagery

Yiming Li[1,2], **Peng Zhou**[3], **Jun Sun**[1], **Yi Xu**[1,2*]

[1]Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]China Mobile (Suzhou) Software Technology Co., Ltd, China
{Yiming.Li, junsun, xuyi}@sjtu.edu.cn, zhoupengcv@outlook.com

Abstract

Text-guided image manipulation has attracted significant attention recently. Prevailing techniques concentrate on image attribute editing for individual objects, however, encountering challenges when it comes to multi-object editing. The main reason is the lack of consistency constraints on the spatial layout. This work presents a multi-region guided image manipulation framework, enabling manipulation through region-level textual prompts. With MultiDiffusion as a baseline, we are dedicated to the automatic generation of a rational multi-object spatial distribution, where disparate regions are fused as a unified entity. To mitigate interference from regional fusion, we employ an off-the-shelf model (CLIP) to impose region-aware spatial guidance on multi-object manipulation. Moreover, when applied to the StableDiffusion, the presence of quality-related yet object-agnostic lengthy words hampers the manipulation. To ensure focus on meaningful object-specific words for efficient guidance and generation, we introduce a keyword selection method. Furthermore, we demonstrate a downstream application of our method for multi-region inversion, which is tailored for manipulating multiple objects in real images. Our approach, compatible with variants of Stable Diffusion models, is readily applicable for manipulating diverse objects in extensive images with high-quality generation, showing superb image control capabilities. Code is available at https://github.com/liyiming09/multi-region-guided-diffusion.

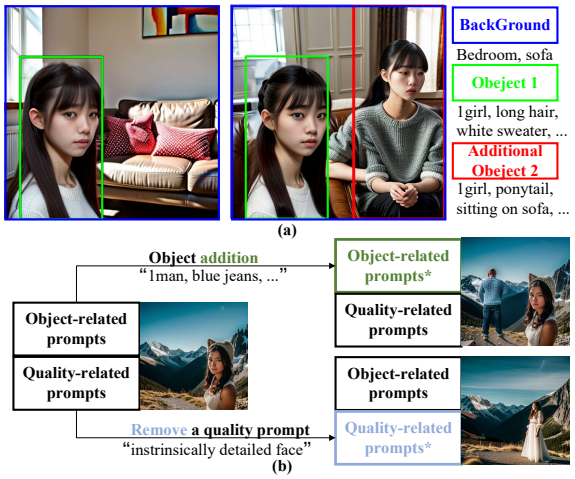

Figure 1: Fig.(a) illustrates unexpected changes from region addition. In the latent space, object additions and removals incur regional interference, leading to distorted limbs. Fig.(b) demonstrates that although quality-related prompts have a significant influence on generation, they are unrelated to editing. Focusing on object-specific and editing-related prompts enhances the quality of manipulation.

1 Intoduction

In recent years, text-guided image synthesis (Ruiz et al. 2023; Kawar et al. 2023; Ding et al. 2021) has received considerable attention. It is particularly noteworthy the work of the diffusion model (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020), which has emerged as the leading approach, renowned for its remarkable capacity to synthesize images with compelling realism and diversity. Pre-trained diffusion models (Rombach et al. 2022; Saharia et al. 2022b) offer great potential in the field of digital content creation, particularly in image manipulation (Kong et al. 2023; Han et al. 2023). In contrast, text-based operations (Crowson et al. 2022; Li et al. 2019a) struggle to provide users with intuitive control over generated content. In practice, challenges persist in text-driven image manipulation within real-world applications (Valevski et al. 2022; Zhu et al. 2020).

Inherited from the superior performance of diffusion models (Mao, Wang, and Aizawa 2023; Voynov, Aberman, and Cohen-Or 2022), certain diffusion-based image editing methodologies (Wang et al. 2022b; Meng et al. 2021; Li et al. 2019b; Sheynin et al. 2022), were developed to achieve precise entity-level manipulations. However, these methodologies primarily focus on attribute editing for individual objects, encountering challenges when there are multiple objects within a real-world scene due to complex spatial layouts among them. For example, face editing (Ju et al. 2023; Pu et al. 2022) involves editing attributes such as age, expression, and skin color. Pose transfer (Men et al. 2020) allows editing of attributes like posture, clothing, and texture. Notably, these manipulations are limited to editing individual objects and achieving attribute modifications for a single

*Corresponding author.

object without altering the background and the structure of the image.

The efficacy of the text-driven image generation/manipulation (Bar-Tal et al. 2022; Couairon et al. 2022) is compromised in scenarios containing multiple entities. For example, modifying the number of entities frequently induces substantial modifications in image structure. In certain instances, it even leads to failing generation within StableDiffusion. Therefore how to add, edit, or remove entities while preserving the original image structure is the main challenge in multi-object image generation/manipulation..

Recent entity-level editing methods (Huang et al. 2023; Hertz et al. 2022) have been inspired by exerting control over the latent space or attention maps (Chen, Laina, and Vedaldi 2023). Their constraints on the initial image layout hinder the ability to make substantial structural modifications, not to mention the process of object addition or removal. On the other hand, certain methods (Bar-Tal et al. 2023; Jiménez 2023) have advanced their generation strategies by proposing frameworks for sequential generation and fusion, facilitating the integration of disparate regions into a coherent entity. They can achieve object addition and removal, yet they suffer from content preservation issues. Some editing methods (Avrahami, Fried, and Lischinski 2023; Avrahami, Lischinski, and Fried 2022) reliant on additional input masks are confined to local modifications and incapable of addressing global editing, such as altering the image background. Besides, the complex input requirements impede their practical applications.

To address more challenging image manipulation tasks, we propose a multi-region guided diffusion (MRGD) framework. Firstly, utilizing a pre-trained StableDiffusion (Rombach et al. 2022) model, we employ MultiDiffusion (Bar-Tal et al. 2023) as the starting point, facilitating the model to dynamically generate and fuse different regions. Subsequently, we introduce an improved attention control scheme into the generation. Similar to the Prompt-to-Prompt (2022), we select and inherit attention maps from the source image to the target image, enabling manipulation while simultaneously preserving the original structure and composition. However, a straightforward combination of the aforementioned algorithms results in a notable drawback, where the fusion of regions introduces interference with each other, thereby disrupting the structural constraints in attention control and resulting in distortions. We term it as "regional interference", as shown in Fig. 1. We propose a multi-region guidance strategy to impose region-level constraints in the spatial dimension, enabling the network with the ability to perceive every region and mitigate regional interference.

Meanwhile, as shown in Fig. 1, practical applications face inefficiencies due to the inclusion of excessively lengthy quality-related yet object-agnostic prompts (e.g., StableDiffusion often employs extensive textual prompts). On one hand, quality-related prompts have a significant impact on the generated results. On the other hand, they are irrelevant to the manipulation. Hence, it is important to focus on object-specific prompts, leading to finer details and enhanced editing quality, as shown in Fig. 1. Thus, we propose a keyword selection method, which is further integrated into guidance and attention control. We also demonstrate a downstream application of our method for multi-region nulltext-inversion (Mokady et al. 2023), tailored for manipulating real images containing multiple objects.

With the introduction of MRGD, we achieve flexible and effective control over image manipulation through provided region-level textual prompts. Our contributions are summarized as follows:

- We propose a framework for text-guided image manipulation which can be directly plugged into existing diffusion models without additional training. The framework enables precise control over multiple region-level objects during high-quality image generation.
- We introduce a multi-region guidance and keyword selection mechanism, endowing the model awareness of regions and keywords. This approach effectively mitigates regional interference, resulting in improved image quality, particularly along region boundaries.
- Our approach is tailored for practical applications, with all experiments conducted on the StableDiffusion WebUI platform. Additionally, through optimization in existing inversion techniques, our method preliminarily extends its applicability to real images.

2 Related Work

Diffusion Model. Diffusion models (Zhang et al. 2023; Croitoru et al. 2023) have demonstrated state-of-the-art performance in various generation benchmarks, encompassing class-conditional image generation (Zheng et al. 2022; Dhariwal and Nichol 2021; Ho and Salimans 2022), text-guided image synthesis (Hinz, Heinrich, and Wermter 2020; Qiao et al. 2019; Li et al. 2023), and layout-to-image translation (Zheng et al. 2023; Sun and Wu 2021; Wang et al. 2022a). Concerning generation quality, ADM-G (Dhariwal and Nichol 2021) introduces classifier guidance conditioned on class labels. Following this work, SDG (Liu et al. 2023) enhances the dimensions and depth of guidance for higher synthetic quality and image-text alignment. MultiDiffusion (Bar-Tal et al. 2023) facilitates multi-region generation and fusion, achieving harmonization across various regions in large-scale images. In addition, the establishment of StableDiffusion (Rombach et al. 2022) and its open-source community have truly facilitated the practical application of diffusion models (Ulhaq, Akhtar, and Pogrebna 2022), significantly inspiring users' creativity.

Text-guided Image Manipulation. Recent years have witnessed significant advancements in text-guided image manipulation (Brooks, Holynski, and Efros 2023; Saharia et al. 2022a) using diffusion models. GLIDE (Nichol et al. 2021) and DALL·E 2 (Ramesh et al. 2022) focus on text-driven open-domain image synthesis and local image editing. BlendedDiffusion (2022) enables local image editing guided by hand-drawn masks. RDM (2023) leverages an additional model for image-text alignment, thereby automatically obtaining masks. Prompt-to-Prompt (2022) achieves modifications to synthesized images by utilizing attention control. However, they are incapable of making substantial

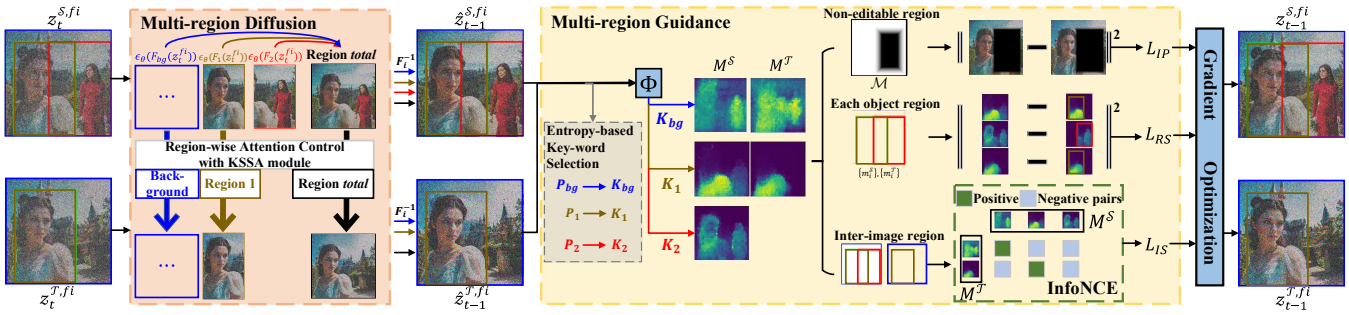

Figure 2: The proposed Multi-Region Guided Diffusion framework. The left part of the framework illustrates the process of region-wise denoising and attention control. Based on the preliminary results $\hat{z}_{t-1}$, the right part shows the details of the keyword selection and region-aware guidance, utilizing information from various regions. Finally, we optimize and update latent encodings through gradient optimization.

modifications to the structure of the image and face difficulties in more challenging tasks such as object addition and removal.

3 Method

In this section, we introduce our multi-region manipulation framework. Our approach is initiated with MultiDiffusion as the baseline. In Sec.3.1, we have elaborated on the approach to implementing multi-region guidance with the awareness of interacted objects. In Sec.3.2, the attention control strategy for keyword selection is advanced to enhance the model with manipulation capability in object-related areas. Finally, downstream task (Roich et al. 2022; Tov et al. 2021) for real-world image inversion is conducted in Sec.3.3.

3.1 Multi-Region-Guided Diffusion

MultiDiffusion. In order to achieve region-by-region image generation, we adopt the strategy of MultiDiffusion that binds together multiple diffusion generation processes with shared parameters. Specifically, we define $R = \{r_{\text{bg}}, r_1, \ldots, r_i, \ldots, r_n, r_{\text{total}}\}$ as a set of regions, where r_{bg} represents a background region and r_i denotes the i-th region ($i = 1, \ldots, n$). Each distinct region $r_i = (\boldsymbol{x}_i, P_i)$ comprises a pair of controlling attributes: bounding box coordinates denoted as $\boldsymbol{x}_i = (x, y, h, w)$, and corresponding textual prompts labeled as P_i. Considering the overall coherence of the generated image, an additional region, denoted as r_{total}, is introduced. The textual prompts for r_{total} unified all individual regions as one entity.

For each region r_i in the set R, we define a cropping function F_i so that $\boldsymbol{I}_i = F_i(\boldsymbol{I}_{\text{fi}})$, where $\boldsymbol{I}_i$ is the image for region r_i and $\boldsymbol{I}_{\text{fi}}$ is the overall image. After compressing the image $\boldsymbol{I}$ into the latent encoding z (Rombach et al. 2022), we employ MultiDiffusion (Bar-Tal et al. 2023) strategy to combine intermediate diffusion results from multiple regions, leading to:

$$\boldsymbol{z}_t^{\text{fi}} = \sum_{i=1}^{n+2} \frac{w_i \otimes F_i^{-1}(\boldsymbol{z}_t^i)}{\sum_{j=1}^{n+2} w_j}, \tag{1}$$

where $\boldsymbol{z}_t^i$ denotes the intermediate diffusion result for the region r_i at time t, and w_i are pixel-wise weights. F_i^{-1} denotes the inverse function of F_i, serving as the restoration process for the cropped region r_i, as shown in Fig. 2.

MultiDiffusion facilitates the dynamic adaptation of the diffusion process to various regions, harmonizing multiple areas into a unified one to mitigate visual dissonance. As shown in Alg.1, given a source image $\boldsymbol{I}^{\mathcal{S}}$ generated from a set of regions $R^{\mathcal{S}}$, we aim to generate a target image, $\boldsymbol{I}^{\mathcal{T}}$, by changing the textual prompt of a specific region within $R^{\mathcal{S}}$. We use $R^{\mathcal{T}}$ to denote the set of regions of the target image $\boldsymbol{I}^{\mathcal{T}}$. Only the selected region to be edited in $R^{\mathcal{T}}$ differs from that of $R^{\mathcal{S}}$, while all other non-editable regions (inherent regions) remain consistent with those in $R^{\mathcal{S}}$. It is challenging for the manipulation to preserve the intrinsic characteristics of the source image in the inherent regions while ensuring alignment between the edited region and its prompts.

Multi-Region Guidance. In cases of overlap between the edited region and inherent regions, MultiDiffusion inevitably gives rise to interference among them, leading to unpleasing distortions, as shown in Fig.1. To mitigate the interference issue, we employ a pre-trained CLIP segmentation model from RDM (Huang et al. 2023), denoted as guidance model Φ, to impose spatial-aware guidance.

Firstly, for each region r_i, users can provide several words as initial keywords $\hat{K}_i$ or leave them blank. Then, updating $\hat{K}_i$ with the keyword selection strategy from Sec. 3.2 to obtain the final keywords K_i for region i, which filters out the object-related keywords from P_i. Next, the guidance model Φ provides sets of segmentation results, denoted as $M^{\mathcal{S}}$ for $\boldsymbol{I}^{\mathcal{S}}$ and $M^{\mathcal{T}}$ for $\boldsymbol{I}^{\mathcal{T}}$, corresponding to the keywords of each region. Finally, we endow the model with region awareness via image-text alignment constraints at region-level and cross-image levels.

Concretely, a prior mask m_i of each region i, except r_{total}, can be obtained from $\boldsymbol{x}_i = (x, y, h, w)$. Then, as shown in Fig. 2, the region-specific (RS) loss requires that the segmentation results $M_i = \Phi(\boldsymbol{I}_{\text{fi}}, K_i)$ from each K_i be confined within the prior mask m_i, denoted as Eq. 2:

$$L_{RS} = \sum_j^{\{\mathcal{S},\mathcal{T}\}} \sum_i^{n+1} \|M_i^j - M_i^j \otimes m_i\|_2^2, \tag{2}$$

Furthermore, the inter-image similarity (IS) loss maximizes

Algorithm 1: Multi-Region Guided Diffusion

Input: A source input $R^{\mathcal{S}}$, a target input $R^{\mathcal{T}}$
Optional for real-word inversion: A real-word image $I^{\mathcal{R}}$
Output: A source image $I^{\mathcal{S}}$, a target image $I^{\mathcal{T}}$

$z_T^{\mathcal{S}} \sim N(0,1)$ a unit Gaussian random distribution
if *inversion* **then**
  for $t = 0, 1, ...T-1$ **do**
    $z_{t+1}^{\mathcal{R},i} = \varepsilon(z_t^{\mathcal{R},i})$, for each region i
  end for
  for $t = T, T-1, ..., 1$ **do**
    Null-text optimize for latent $\hat{z}_t^{\mathcal{R}}$ and each region i
    $\min_{\tilde{\varnothing}_t^i} \left\| z_{t-1}^{\mathcal{R},i} - \hat{z}_{t-1}^{\mathcal{R},i}\left(\hat{z}_t^{\mathcal{R},i}, \tilde{\varnothing}_t^i, P_i\right)\right\|_2^2$
  end for
  $z_T^{\mathcal{S}} \leftarrow \hat{z}_T^{\mathcal{R}}$
end if
$z_T^{\mathcal{T}} \leftarrow z_T^{\mathcal{S}}$
for $t = T, T-1, ..., 1$ **do**
  $\hat{A}^{\mathcal{T}} \leftarrow EDIT(A^{\mathcal{S}}, A^{\mathcal{T}})$
  if *inversion* **then**
    $\varnothing_t \leftarrow \tilde{\varnothing}_t$
  end if
  $\hat{z}_{t-1}^{\mathcal{S},fi}, \hat{z}_{t-1}^{\mathcal{T},fi} \leftarrow \epsilon_\theta(z_t^{\mathcal{S},i}, z_t^{\mathcal{T},i} | \varnothing_t^i, P^i)$ for region i
  $\mathcal{L} \leftarrow \lambda_{rs} L_{RS} + \lambda_{is} L_{IP} + \lambda_{ip} L_{IP}$
  $z_{t-1}^{\mathcal{S},fi}, z_{t-1}^{\mathcal{T},fi} \sim \mathcal{N}(\mu + \Sigma \nabla_{\hat{z}_{t-1}} \mathcal{L}, \Sigma)$
end for
$I^{\mathcal{S}}, I^{\mathcal{T}} = \mathcal{D}(\hat{z}_0^{\mathcal{S},fi}, \hat{z}_0^{\mathcal{T},fi})$
return $I^{\mathcal{S}}, I^{\mathcal{T}}$

the mutual information between $M^{\mathcal{S}}$ and $M^{\mathcal{T}}$ with contrastive learning, denoted as Eq. 3:

$$L_{IS} = InfoNCE(M^{\mathcal{S}}, M^{\mathcal{T}}), \tag{3}$$

where $InfoNCE()$ represents our utilization of InfoNCE (Oord, Li, and Vinyals 2018), which enforces $M^{\mathcal{S}}$ and $M^{\mathcal{T}}$ remain consistent in their corresponding regions while minimizing similarity in non-corresponding regions. Additionally, to prevent unintentional changes to inherent region $\mathcal{M}$, we developed inherent preservation (IP) loss, which enforces content consistency within $\mathcal{M}$ in RGB and latent space after manipulation, thereby enhancing the preservation of non-editable objects:

$$L_{IP} = (\|\mathcal{M} \otimes (I^{\mathcal{S}} - I^{\mathcal{T}})\|_2^2 + \|\mathcal{M} \otimes (z^{\mathcal{S}} - z^{\mathcal{T}})\|_2^2), \tag{4}$$

We set the diffusion guidance losses as a weighted sum, which is summarized in Alg. 1.

3.2 Keyword-selected Attention Control

In text-guided image generation, complex textual descriptions are typically necessary for fine-grained image control. In fact, the visual attention maps of various prompts in Fig. 3 demonstrate that the quality-related but lengthy prompts are object-agnostic, which may impede the model from distinguishing the object-specific keywords, potentially leading to low efficiency and poor comprehension for the entire

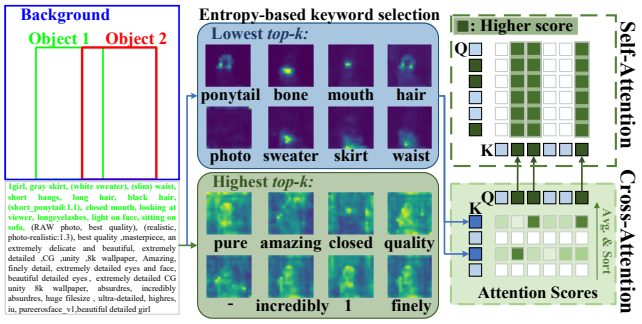

Figure 3: KSSA. We first present *object-related prompts* in green words and *quality-related prompts* in black words for *object 1*. Subsequently, based on the entropy of cross-attention maps, we perform sorting and selection to extract the top N words (by default, $N = 15$). Within the cross-attention block, we identify image features that are highly similar to the N keywords and assign them higher attention scores in the self-attention block.

scene. Therefore, we propose a keyword selection mechanism based on cross-attention maps, i.e. extract and enhance the most important words from the lengthy textual prompt P. These selected keywords will be used as guidance for multiple object manipulation in Sec. 3.1.

Our image manipulation algorithm adopts a keyword-based attention control scheme to enhance Prompt-to-Prompt, which imposes attention injection from source image to target image for spatial layout control. However, the strict constraints in Prompt-to-Prompt on the overall layout limit its capacity for substantial structural modifications. Thus, we propose an enhanced region-aware attention control strategy to manipulate the generation of multiple object regions. Distinctive attention control strategies are assigned to various regions, as illustrated in Eq. 5.

$$\widehat{A^{\mathcal{T}}} = \begin{cases} A^{\mathcal{S}} & \text{for the inherent region} \\ Edit(A^{\mathcal{T}}, A^{\mathcal{S}}) & \text{for } r_{\text{total}}, \text{ if time step} \geq \tau \\ A^{\mathcal{T}} & \text{for } r_{\text{total}}, \text{ time step} < \tau, \end{cases} \tag{5}$$

where τ is a parameter that determines when to cease the propagation of attention map $A^{\mathcal{S}}$, and $Edit(A^{\mathcal{T}}, A^{\mathcal{S}})$ represents the intuitive approach of Prompt-to-Prompt. Specifically, for object addition and removal, we utilize the prompt refinement method in Prompt-to-Prompt, and for attribute modification, we employ the word swap method. Please refer to the appendix for details.

Aiming at selecting object-specific keywords, we save the cross-attention map sets CA during the generation of r_{total}, which is normalized to $[0, 255]$, as shown in Fig. 3. Then, the image entropy is computed for each element CA_i according to the Eq.6:

$$H_{CA_i} = -\sum_{j=0}^{255} p(j) \log p(j), \tag{6}$$

where $p(j)$ represents the occurrence probability of pixel value $j \in [0, 255]$ in the histogram statistics of CA_i. Sub-

sequently, the N selected keywords K_i with smallest entropy (Hu et al. 2022) in each region i are filtered out.

Currently, each prompt is treated equally without distinction. This results in a lower level of comprehension for the scene, especially in overlapping areas. Therefore, to encourage the network to pay more attention to key entities in different regions, and thus better understand the relationships between multiple regions during generation, we propose Keyword-selected Self-Attention (KSSA). The core objective of KSSA is to diminish the emphasis of the model on quality-related yet object-agnostic prompts, thus amplifying the attention on object-specific representations.

KSSA is divided into two stages. First, as shown in Fig. 3, the highlighted localized areas in attention maps depict image embeddings with higher activation related to the keywords K_i. During the cross-attention phase, we need to record the top half image embeddings with higher similarity to K_i. Then, we increase the weights of the selected image embeddings in self-attention, emphasizing their contribution to the weighted sum result. KSSA enables the network to focus more on regions with higher responses to the object-specific keywords, resulting in finer details and heightened image-test consistency.

3.3 Multi-region Inversion

The further editing of real-world images holds value in manipulation. When handling multi-object real images, prevailing inversion methods (Mokady et al. 2023; Huberman-Spiegelglas, Kulikov, and Michaeli 2023; Gal et al. 2022) guided by textual prompts can only reconstruct and edit in the global region. However, the presence of multiple objectives in the prompts can disrupt the initial layout, causing impractical distortions.

To alleviate this impact, we incorporate region-level control in the inversion. A multi-region inversion is introduced to map sub-regions into latent space. By extending null-text inversion (Mokady et al. 2023), we employ source prompts as controls for reconstruction, yielding initial latent noise $\hat{z}_T$ and a trainable unconditional embedding set $\tilde{\varnothing}_t$ at the region level. Conditioned on multi-region $\tilde{\varnothing}_t$, we engage in manipulation during reconstruction to achieve the edited image.

Multi-region inversion effectively counteracts interference from external information for the current sub-region, providing finer guidance during inversion, which allows for high-quality reconstruction and editing while maintaining the real-world image layout.

4 Experiement

4.1 Implementation Details

Dataset. To the best of our knowledge, there exist no standardized benchmarks for this challenging task of text-guided image manipulation. Thus, we utilized open-source models from the StableDiffusion-WebUI community to conduct image manipulation, with a focus on a wide range of subjects including humans, vehicles, and animals. More specifically, we sourced a variety of models and prompts from the community, which we then integrated with manually designated region coordinates (x, y, h, w) to generate a collection of 61 input pairs $(\boldsymbol{R}^{\mathcal{S}}, \boldsymbol{R}^{\mathcal{T}})$ for ensuing experimentation. Specifically, there are 25 pairs for object addition, 24 pairs for object removal, and 12 pairs for attribute modification. All manipulation results were uniformly sized to 512×512 pixels.

Details. For the guidance model, we utilized CLIP ViT-B/16 (Radford et al. 2021). All experiments were executed on one RTX 3090 GPU with PyTorch. Additionally, we set the default parameter to $\lambda_{rs} = 1000, \lambda_{is} = 2000, \lambda_{ip} = 300, \tau = 0.5$. To ensure result quality and parameter consistency, we employed a diffusion step T with a DDIM-solver of 20 in all experiments. An introduced hyperparameter, denoted as T_{total}, governs the incorporation of r_{total} into the generation process after T_{total} steps, serving to avert unexpected disturbances to the initial layout. The more detailed settings are reported with analysis in the appendix.

Evaluation Metrics. Image manipulation tasks primarily focus on harmonizing the target image while preserving its original components. As a result, we conduct a comprehensive dual-quality assessment. On one aspect, focusing on editing objectives, we assess the post-manipulation quality and coherence. To this end, we employ CLIP similarity (Kim, Kwon, and Ye 2022) to evaluate image-text alignment in the edited region. Meanwhile, we use SSIM (Hore and Ziou 2010) between $\boldsymbol{I}^{\mathcal{S}}$ and $\boldsymbol{I}^{\mathcal{T}}$ in the neighborhood of the editing region to measure the fidelity degree within the propagation areas, denoted as SSIM-e. SSIM-e serves to reflect the impact of regional interference around the edited region. An approach incapable of mitigating regional interference would lead to reduced SSIM-e due to layout distortion. On the other aspect, we employ SSIM to evaluate the preservation between $\boldsymbol{I}^{\mathcal{S}}$ and $\boldsymbol{I}^{\mathcal{T}}$ within the inherent region, denoted as SSIM-i. These two metrics measure the degree of structural consistency around and outside the edited regions.

4.2 Results

To evaluate our method, we conducted a comparative analysis of the three aforementioned manipulation tasks on our constructed dataset. Considering that Prompt-to-Prompt is not directly applicable to the task, we have re-implemented and applied it to multi-region generation through independent control of multiple regions, denoted as P2P*. During experimentation, we maintained consistent random seeds across different methods, resulting in comparable outcomes.

Qualitative Comparison. For different methods, the fixed random seed introduced the same input latent noise, leading to similar layouts and structures in their outcomes. MultiDiffusion results in significant structural alterations to the inherent regions during manipulation. In some instances, it even leads to notable distortions, such as unrealistic limb deformations as indicated by the yellow box in Fig. 4. It lacks the capacity to preserve inherent structures, rendering it vulnerable to regional interference. P2P* exhibits considerably better structural preservation compared to MultiDiffusion, including spatial layout and texture patterns. However, the observed discrepancies emphasize our approach's notable superiority in image coherence and preservation of the inherent components.

In contrast, MRGD exhibits region awareness through selection and guidance, which enhances the model to dis-

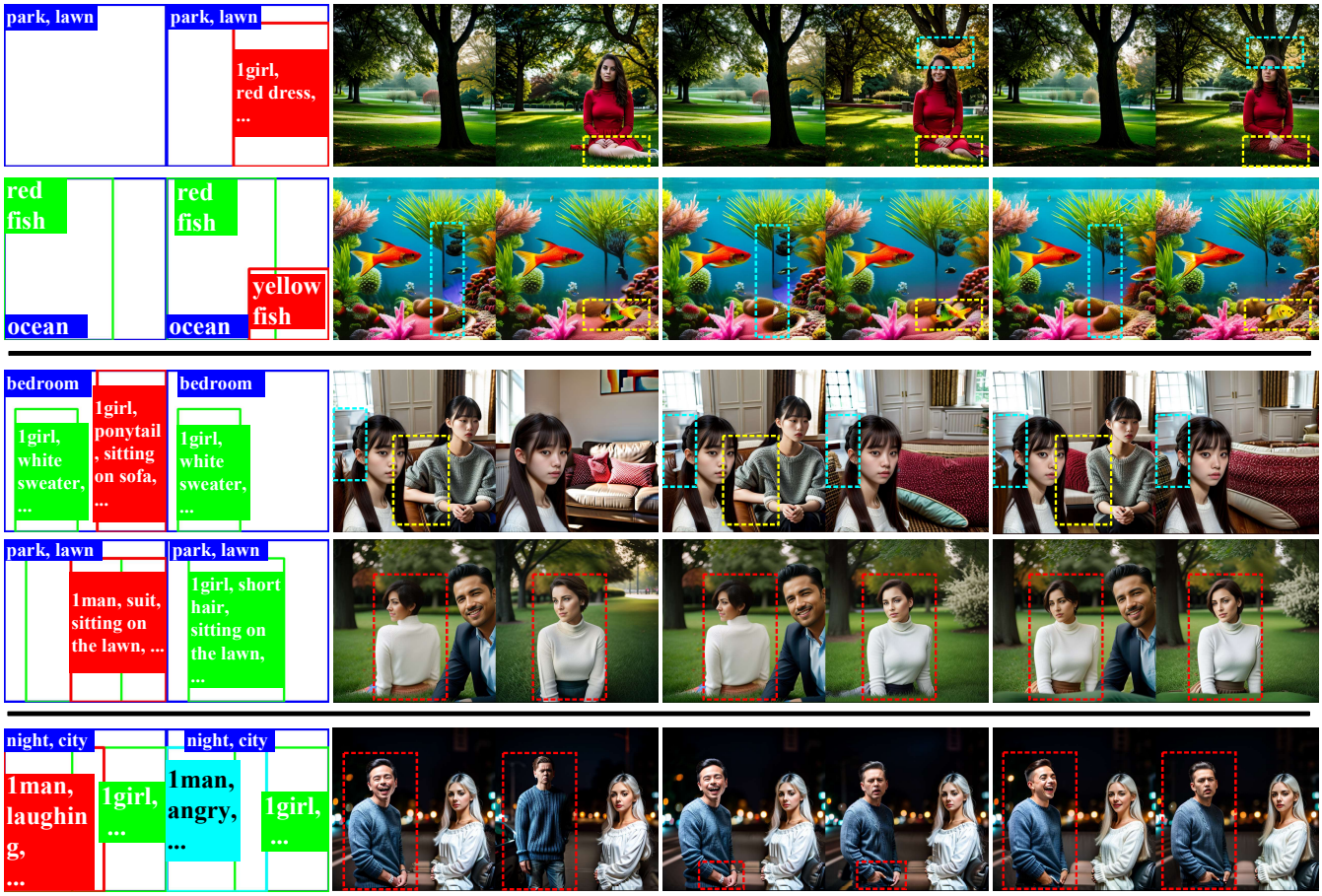

Figure 4: Visualization results. The figures from left to right represent the input pairs, outcomes of MultiDiffusion, P2P*, and our MRGD. From top to bottom, there are object addition, removal, and attribute modification, respectively.

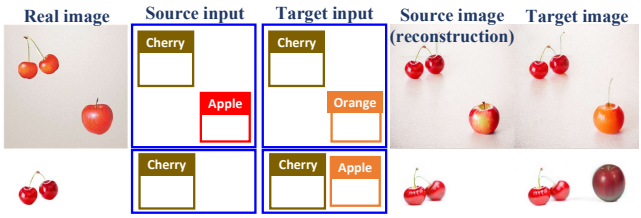

Figure 5: Visual results for the real-world image inversion.

tinguish between edited and inherent regions, thereby suppressing mutual interference between different regions. On one hand, our method preserves better identity consistency of inherent objects, as indicated by the red box in Fig.4. On the other hand, our approach produces more harmonious results at region boundaries, as shown in the blue box.

Quantitative Comparison. In Tab.1, we observed a marginal difference among the three methods in CLIP similarity, with even higher scores for comparison methods. We posit that this phenomenon highlights a bias in multi-region diffusion, wherein it excessively prioritizes image-text alignment while neglecting inter-region interactions. Conversely, our approach outperforms the comparative methods in both SSIM-e, which reflects the coherence of manipulation, and SSIM-i, which gauges the preservation of inherent objectives. Quantitative results underscore the high-quality and precision of MRGD. Subsequent analyses will be conducted in conjunction with specific tasks.

Object Addition. In addition to preserving inherent object details, a crucial aspect of object addition lies in the harmonious interaction between the newly added object and its surroundings. Compared to $\boldsymbol{I}^{\mathcal{S}}$, the latent space of $\boldsymbol{I}^{\mathcal{T}}$ encompasses an additional region, unavoidably introducing interference into inherent regions. Moreover, semantic information within the editing region may leak into other regions, leading to semantic shifts or distortions. Several instances from MultiDiffusion illustrate the severity of such interference. In contrast to P2P*, our approach seamlessly integrates new objects into the overall image, preventing abrupt background shifts and promoting a more natural fusion.

Object Removal. Visual results demonstrate that MultiDiffusion and P2P* often yield lower-quality source and target images. Furthermore, region removal also induces variations in the latent space. P2P*, limited to localized attention control within a single region, can only maintain the basic layout

Method	Addition			Removal			Attribute		
	CLIP↑	SSIM-e↑	SSIM-i↑	CLIP↑	SSIM-e↑	SSIM-i	CLIP↑	SSIM-e ↑	SSIM-i↑
MultiDiffusion	**27.14**	0.4365	0.5277	27.95	0.4111	0.5333	**27.58**	0.5415	0.6529
P2P*	27.11	0.5281	0.6905	**28.2**	0.4936	0.5832	27.25	0.7226	0.8688
Ours	27.02	**0.5998**	**0.7559**	27.52	**0.5559**	**0.7834**	27.53	**0.7514**	**0.8700**

Table 1: Quantitative results.

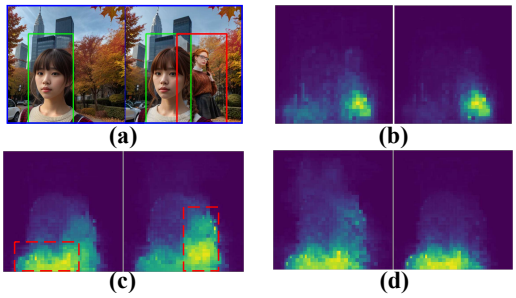

Figure 6: (a) represents the source and target images. (b) showcases the two object segmentation outcomes of guidance model on $I^{\mathcal{T}}$ after 5 steps. (c) displays the outcomes at the final step with guidance. (d) displays the outcomes at the final step without guidance.

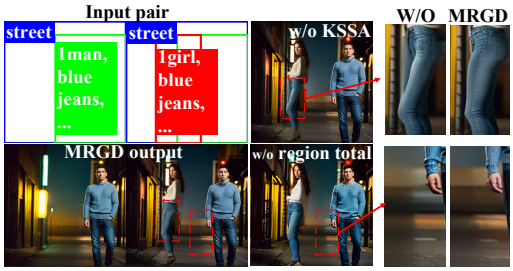

Figure 7: The left side shows the input and outcomes. The right side displays the ablation visual results.

of the source image, which lacks fine-grained control. Fig. 4 illustrates the inconsistent outcomes of P2P*. In contrast, our approach better mitigates disturbances in the latent space and retains a majority of inherent object details. In contrast, our approach mitigates regional disturbances better and retains a majority of inherent object details.

Attribute Modification. The preservation of details in images after manipulation is of primary concern in attribute modification. Compared to our approach, both MultiDiffusion and P2P* exhibit notably more non-inheritable details, such as sweater textures, hand positions, and background details, as shown in Fig. 4.

Inversion. A preliminary real image inversion is illustrated in Fig. 5. As observed, our approach not only achieves acceptable reconstruction quality but also maintains a high level of editability. MRGD effectively accomplishes object addition ("apple") and attribute modification (from "apple" to "orange"). This demonstrates the versatility of our manipulation capabilities, which are applicable not only to text-guided synthesized images but also to real-world images.

4.3 Ablation Study

In this section, we validate the contributions of each component of our algorithm through ablation studies, providing a qualitative analysis of the effectiveness of each component. The quantitative analysis of different variants of StableDiffusion models and more details are placed in the appendix.

Multi-Region Guidance. As the critical point of our method, multi-region guidance efficiently mitigates interference from regional noise. As shown in Fig. 6, two adjacent girls that were initially unable to be correctly identified are accurately distinguished through guidance, which also leads to improved manipulation quality.

Total Region. The introduction of r_{total} is crucial for the coherence of manipulations. As shown in Fig. 7, an additional generation of r_{total} can facilitate the smoother integration of the editing area into the context, thereby reducing unstable background shifts and unrealistic distortions.

Entropy-based selection. As shown in Fig. 3, we observe that attention maps with higher entropy are uniform and dispersed, corresponding to quality-related prompts. While attention maps with lower entropy focus more on specific regions about object-related nouns. KSSA can adaptively filter out keywords from lengthy prompts and provide more tailored guidance focused on keyword regions. Thus, MRGD frequently demonstrates enhanced details and improved image-text alignment, as illustrated in Fig. 7. Furthermore, from an application perspective, it enhances user experience by streamlining interactions, eliminating the requirement for exhaustive object-specific keywords.

5 Conclusion

In this work, we developed an image manipulation framework with the powerful capabilities of multi-region generation. We have investigated two major challenges encountered by existing models in practical application: regional interference and lengthy object-agnostic prompts. Correspondingly, we demonstrated how to mitigate interference and achieve efficient manipulation using multi-region guidance and keyword selection mechanisms. Furthermore, we have preliminarily applied our approach to the downstream task of real image inversion. We provide a viable framework for future image manipulation tasks that aim to be more applicable. Moving forward, we aspire to expand upon the current work to achieve more flexible, natural, and faithful image manipulations.

Acknowledgments

This work was supported in part by NSFC 62171282, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), 111 project BP0719010.

References

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

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 18208–18218.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, 707–723. Springer.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 18392–18402.

Chen, M.; Laina, I.; and Vedaldi, A. 2023. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, 88–105. Springer.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Han, I.; Yang, S.; Kwon, T.; and Ye, J. C. 2023. Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion. *arXiv preprint arXiv:2303.08767*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hinz, T.; Heinrich, S.; and Wermter, S. 2020. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1552–1565.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369. IEEE.

Hu, X.; Zhou, X.; Huang, Q.; Shi, Z.; Sun, L.; and Li, Q. 2022. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 18291–18300.

Huang, N.; Tang, F.; Dong, W.; Lee, T.-Y.; and Xu, C. 2023. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*.

Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2023. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. *arXiv preprint arXiv:2304.06140*.

Jiménez, Á. B. 2023. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*.

Ju, X.; Zeng, A.; Zhao, C.; Wang, J.; Zhang, L.; and Xu, Q. 2023. HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation. *arXiv preprint arXiv:2304.04269*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 6007–6017.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 2426–2435.

Kong, C.; Jeon, D.; Kwon, O.; and Kwak, N. 2023. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In *Proc. - IEEE Winter Conf. Appl. Comput. Vis., WACV*, 848–857.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019a. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32.

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019b. Object-driven text-to-image synthesis via adversarial training. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 12174–12182.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 22511–22521.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion

guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 289–299.

Mao, J.; Wang, X.; and Aizawa, K. 2023. Guided Image Synthesis via Initial Image Editing in Diffusion Model. *arXiv preprint arXiv:2305.03382*.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 5084–5093.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 6038–6047.

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

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pu, G.; Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2022. Controllable Image Synthesis with Attribute-Decomposed GAN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1514–1532.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in Neural Information Processing Systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1): 1–13.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Sheynin, S.; Ashual, O.; Polyak, A.; Singer, U.; Gafni, O.; Nachmani, E.; and Taigman, Y. 2022. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, W.; and Wu, T. 2021. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5070–5087.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14.

Ulhaq, A.; Akhtar, N.; and Pogrebna, G. 2022. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*.

Valevski, D.; Kalman, M.; Matias, Y.; and Leviathan, Y. 2022. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*.

Voynov, A.; Aberman, K.; and Cohen-Or, D. 2022. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*.

Wang, B.; Wu, T.; Zhu, M.; and Du, P. 2022a. Interactive image synthesis with panoptic layout generation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 7783–7792.

Wang, J.; Lu, G.; Xu, H.; Li, Z.; Xu, C.; and Fu, Y. 2022b. ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-wise Semantic Alignment and Generation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 10707–10717.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Zheng, G.; Li, S.; Wang, H.; Yao, T.; Chen, Y.; Ding, S.; and Li, X. 2022. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, 754–769. Springer.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proc. - IEEE Conf. Comput. Vis. Pattern Recognit.*, 22490–22499.

Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. In *European conference on computer vision*, 592–608. Springer.