

# Text-to-image Editing by Image Information Removal

Zhongping Zhang<sup>1</sup> Jian Zheng<sup>2</sup> Jacob Zhiyuan Fang<sup>2</sup> Bryan A. Plummer<sup>1</sup>  
<sup>1</sup>Boston University <sup>2</sup>Amazon Alexa AI  
<sup>1</sup>{zpzhang, bplum}@bu.edu <sup>2</sup>{nzhenjji, zyfang}@amazon.com

## Abstract

Diffusion models have demonstrated impressive performance in text-guided image generation. Current methods that leverage the knowledge of these models for image editing either fine-tune them using the input image (e.g., Imagic) or incorporate structure information as additional constraints (e.g., ControlNet). However, fine-tuning large-scale diffusion models on a single image can lead to severe overfitting issues and lengthy inference time. Information leakage from pretrained models also make it challenging to preserve image content not related to the text input. Additionally, methods that incorporate structural guidance (e.g., edge maps, semantic maps, keypoints) find retaining attributes like colors and textures difficult. Using the input image as a control could mitigate these issues, but since these models are trained via reconstruction, a model can simply hide information about the original image when encoding it to perfectly reconstruct the image without learning the editing task. To address these challenges, we propose a text-to-image editing model with an Image Information Removal module (IIR) that selectively erases color-related and texture-related information from the original image, allowing us to better preserve the text-irrelevant content and avoid issues arising from information hiding. Our experiments on CUB, Outdoor Scenes, and COCO reports our approach achieves the best editability-fidelity trade-off results. In addition, a user study on COCO shows that our edited images are preferred 35% more often than prior work.

## 1. Introduction

Text-driven image editing aims to modify the specific content of an image based on its textual descriptions. Inspired by the powerful capability of large-scale text-to-image generation models [16, 26, 28, 30], recent approaches have leveraged the prior knowledge of these pretrained models for image editing [1, 2, 9, 29, 40, 42]. The majority of existing editing approaches follow two strategies: 1) Optimization-based methods: updating network parameters or feature embeddings for each input image, as shown in

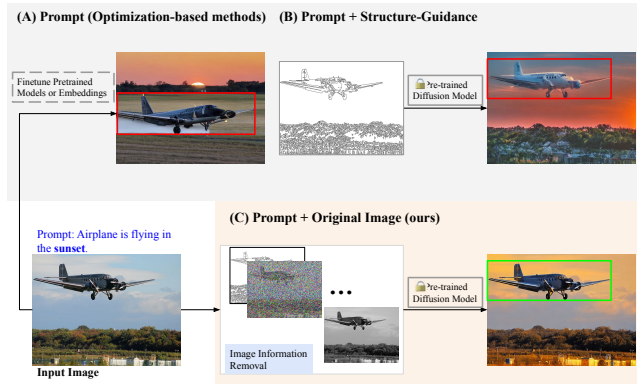


Figure 1. We aim to edit the specific content of the input image according to text descriptions while preserving text-irrelevant image content. Prior work based on large-scale diffusion models has followed two major approaches for image editing: (A), fine-tuning the pretrained models or text embeddings (e.g., Imagic [9] or Dreambooth [29]), or (B), introducing structural guidance as additional constraint to control the spatial information of the generated image (e.g., ControlNet [40] or MaskSketch [2]). In our work, shown in (C), our approach conditions on both the original image and the structural guidance, to better preserve the text-irrelevant content of the image. E.g., our model successfully preserves the original attributes of the airplane (outlined by the green bounding box) in the generated image. In contrast, previous methods such as Imagic (A) and ControlNet (B) not only alter the sky and background but also modify the attributes of the airplane (outlined by the red bounding boxes), which is unwanted in this example.

Figure 1 (A); or 2) introducing the structural guidance (e.g., edge map, user scribble, segmentation map, or pose estimation) as additional constraints for image generation, as shown in Figure 1 (B). The effectiveness of these models have been demonstrated on tasks like style transfer [42], texture editing [1], shape editing [9], appearance modification [29], color editing [40], among others. However, for optimization-based methods, fine-tuning large-scale models on single or few images results in severe overfitting issues and prolongs inference time [42]. Images generated by finetuned models and embeddings may contain unexpected visual artifacts due to the information leakage

and fail to preserve the text-irrelevant content of original image [42]. Structure-guided methods also meet pitfalls: structural guidance usually contains no information about colors or textures, these frameworks have difficulty preserving the text-irrelevant content of the original image. As outlined by red bounding boxes in Figure 1 (A) and (B), we observe both Imagic [9] and ControlNet [40] fail to preserve the text-irrelevant content of the original image: Imagic modifies the shape of the airplane while ControlNet changes the color and textures of airplane.

To address the aforementioned issues, we introduce the original image as an additional control for image editing. In this way the model can fully incorporate the content of the input image, allowing it to effectively preserve the text-irrelevant content. However, this results in an identity mapping issue [12], where the model can simply map the input directly to the output. This is primarily caused by the image reconstruction objective in editing task, which is perfectly optimized using an identity mapping. Prior works attempt to alleviate such issue by either learning disentangled features [19, 37, 38], or uses attribute classifier to remove the target attribute [14, 15]. Both these approaches unavoidably introduce additional computational overhead that also greatly limits their application scenarios. For example, in Figure 1, the input image only has text annotations and does not have scene attribute labels such as “daylight” or “sunset.” Therefore, these methods cannot be applied to convert the input image from “daylight” to “sunset.”

We propose an Image Information Removal module (IIR-Net) to partially remove the image information from the input image, as illustrated in Figure 1 (C). Specifically, this erasure of image information arises from two components. First, we localize the Region of Interest (RoI<sup>1</sup>) and erase the color-related information. Second, we apply Gaussian noise on the input image which randomly eliminates the texture-related information. By tweaking the noise intensity, the model is capable of adapting to various tasks accordingly. For example, in color editing tasks, we decrease the noise intensity to zero to preserve most information from the input image except the color. In the texture editing task, a higher value of noise intensity is used to eliminate most information from the target region, leaving only the structural prior. Given the original image, we then concatenate the structure map with attribute-excluded features as additional controls to editing model. With our simple while effective image information removal module, we avoid the identical mapping issue as now the model is forced to not only reconstruct the original, but also predict the noised image regions.

<sup>1</sup>We refer to the modified regions of the target image as RoI. In our work, RoI is localized by Grounded-SAM [10, 18]. For tasks that the entire image is subject to modification such as scene attribute transfer or style transfer, we simply define the entire image as the RoI.

- We summarize the contribution of our work as follows:
- We introduce the original image as an additional guidance to pretrained generative diffusion models for image editing tasks. Compared with existing image editing methods [9, 40], IIR-Net more effectively preserves the text-irrelevant content of the input image while also generating new features according to the language descriptions.
  - We propose an image information removal module to counter the identical mapping issue [12]. IIR-Net partially erases the image information such as colors or textures from the input, and reconstruct the original image according to text descriptions and attribute-excluded features. Compared with prior work for solving this issue [14, 37, 38], IIR-Net does not require attribute labels to learn disentangled features or attribute classifiers, and, thus, can be applied to images without attribute labels.
  - We conduct extensive quantitative and qualitative experiments on three public datasets CUB [35], COCO [17], and Outdoor Scenes [11]. Our results demonstrate that our model improves the fidelity-editability trade-off over the state-of-the-art with obvious inference speed advantages. *E.g.*, compared to Imagic [9], IIR-Net improves the LPIPS score from 0.57 to 0.30 on COCO, with an inference speed improvement of two orders of magnitude.

## 2. Related Work

**Feed-forward transformation image generation and editing.** Early work in text-to-image generation and editing often used text-to-image generator based on conditional GANs [3, 12, 13, 21, 27, 33, 36, 39]. Limited by the scalability of Conditional GAN and size of image datasets, these methods only supported specific image domains and language descriptions. More recent methods typically trained conditional diffusion models [25, 26, 28, 30] on massive datasets (*e.g.*, LAION-400M [31]). Due to the difficulty to obtain image pairs before and after editing, current image editing frameworks [2, 9, 22, 29, 40, 42] are mostly developed based on pretrained text-to-generation models [28, 30]. However, among these methods, methods that leverage the feed-forward transformation mechanism mostly focus on structural guidance. *E.g.*, ControlNet [40] leverages structure maps like edge map, semantic map, or pose estimation to control the spatial structure of generated images, and MaskSketch [2] uses sketch as additional control to generate images. Thus, these methods cannot preserve the other attributes of the image such as colors or textures well, and may result in significant deviation from the input image. To solve this issue, we incorporate the original image as input to our model and propose an image information removal to solve the identical mapping issue [12].

**Optimization-based Methods** Prior work has demonstrated that optimization-based methods, which update network parameters on each image input, work well for im-

age generation [32, 34, 43]. Several methods use CLIP [23] as a constraint to guide the embedding features of predicted images [1, 5, 8, 20]. Inspired by the success of pretrained text-to-image generation frameworks [25, 26, 30], researchers have also proposed methods to finetune these models for image editing (*e.g.*, Imagic [9], Dreambooth [29], SINE [42], Textual Inversion [6]). Compared to feed-forward transformation methods [2, 40], these models retain more information from the original image since they take the whole image instead of just a structure map as additional guidance. However, as we will show in Section 4.2, some image content such as background or irrelevant attributes of target objects may still be changed in this process. In addition, the inference time of these optimization-based methods is much longer than feed-forward transformation methods due to image-specific finetuning.

### 3. IIR-Net: Text-to-Image Editing by Image Information Removal

Given an input image  $x$  and its corresponding text prompt  $S$ , our task aims to create the desired content according to  $S$  while preserving the text-irrelevant content of  $x$ . To achieve this, we incorporate the original image  $x$  as an additional control to pretrained text-to-image generation model, which is discussed in Section 3.1. However, since the model is trained on the image reconstruction task, the incorporation of the original image can lead to the identical mapping issue, in which the model simply maps the input image as the output. To address this challenge, we propose our image information removal module in Section 3.2. Figure 2 provides an overview of our approach.

#### 3.1. Conditional Latent Diffusion Model

As discussed in the Introduction, preserving the text-irrelevant content of the original image is critical for text-to-image editing. Leveraging the structural guidance as an additional hint (*e.g.*, ControlNet [40], MaskSketch [2]) can lead to significant information loss from the original image. To address this, we introduce the original image as additional control to our model, which preserves all information from the input image. In this section, we first introduce the pretrained text-to-image generation model, Stable Diffusion [28], as preliminaries to our method, and discuss our IIR component in Section 3.2.

Given an input image  $x_0$  and its corresponding noisy image  $x_T$ , Stable Diffusion [28] consists of a series of equally weighted denoising autoencoders  $\epsilon_\theta(x_t, t)$ , where  $t$  ranges from  $1 \sim T$ . The denoising autoencoders are trained to predict the noise  $\epsilon$  in  $x_T$  according to time step  $t$  and noisy input  $x_t$ . The objective function of Stable Diffusion is

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

where  $\mathcal{E}$  is the pretrained encoder of VQGAN [4] to encode image  $x_t$  to latent features  $z_t$ , and vice versa. For conditional generation, the denoising autoencoders  $\epsilon$  take  $\tau_\theta(y)$  as additional input, where  $\tau_\theta(y)$  represents a domain-specific encoder to extract feature embeddings from the condition  $y$ . This condition  $y$  represents elements like text prompts and semantic maps, among others. Given image-condition pairs, the Conditional Latent Diffusion Model (CLDM) is optimized by

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

where  $\tau_\theta, \epsilon_\theta$  are jointly optimized. In our model, the condition  $y$  consist of text descriptions  $S$  and the original image  $x_0$  and is defined as

$$\tau_\theta(y) := \{\tau_{\theta_1}(S), \tau_{\theta_2}(R(x_0))\}, \quad (3)$$

where we use the CLIP model [24] as  $\tau_{\theta_1}(\cdot)$  to encode the text descriptions  $S$  and use ControlNet [40] as  $\tau_{\theta_2}(\cdot)$  to encode the feature  $R(x_0)$ .  $R(\cdot)$  denotes our image information removal module, which we discuss in the next section.

#### 3.2. Image Information Removal

As discussed in the Introduction, training solely on image reconstruction can lead to the identical mapping issue. Previous approaches address this issue by learning disentangled features [7] or attribute classifiers [14]. However, these methods require annotated attributes, restricting their application scenarios. To overcome this challenge, we propose our image information removal module, which incorporates color and texture removal operations. Our removal operations effectively mitigates the identical mapping issue without the requirement for additional annotated labels.

**Color-related Information Removal.** In Figure 2 (B), we present our color information removal operation. Given the input image  $x_0$  and its corresponding text prompt  $S$ , we employ Grounded-SAM [10, 18] to localize the RoI. The color information of  $x_0$  is then erased by

$$x'_0 = \text{rgb2gray}(x \odot m_{\text{RoI}}) + x \odot (1 - m_{\text{RoI}}), \quad (4)$$

where  $m_{\text{RoI}}$  is the Grounded-SAM segmentation mask.

Through the application of color-related information removal to the input image  $x_0$ , our model demonstrates proficiency in color-related editing tasks, such as transforming a "white airplane" into a "green airplane." However, as depicted in Figure 5, the model encounters challenges when attempting to modify texture-related information, such as changing "lawn" to "sand." To address this limitation, we introduce our texture-related information removal module.

**Texture-related Information Removal.** We eliminate the texture-related information by adding noise to the image

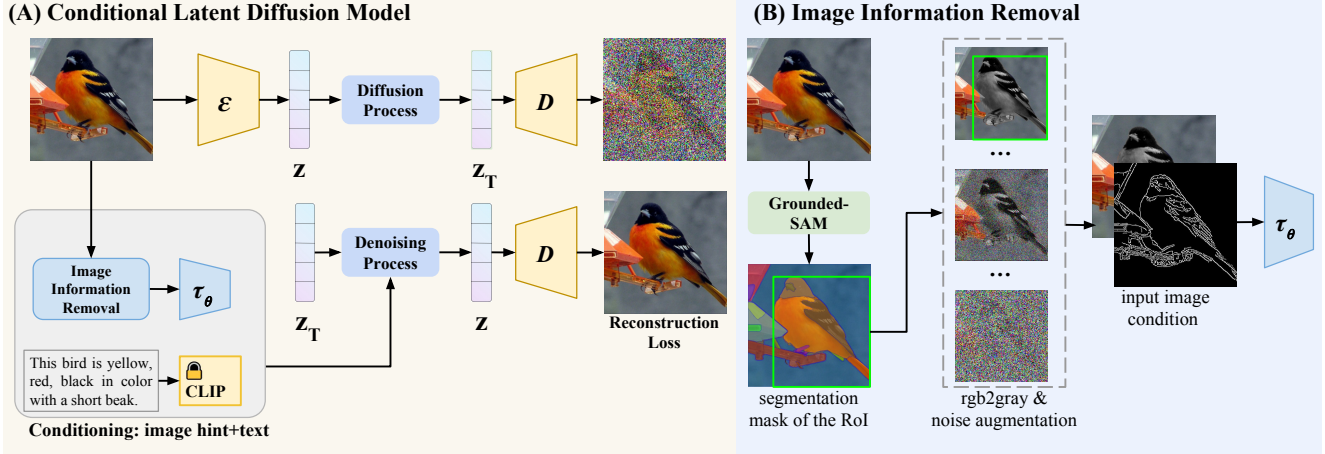


Figure 2. **IIR-Net Overview.** Our model mainly consists of two modules: (A) Conditional Latent Diffusion Model: We introduce the original image  $x$  as additional control to our model to preserve the text-irrelevant features of  $x$ . See Section 3.1 for detailed discussion; (B) Image Information Removal Module: We erase the image information mainly by two operations. First, we convert RGB values to the gray values in the ROI to exclude the color information. Second, we add Gaussian noise to the input image to partially erase the texture-related information. This module is applied to address the identical mapping issue. See Section 3.2 for detailed discussion.

condition  $x'_0$  of our model, denoted by

$$q(x'_K|x'_0) = \prod_{k=1}^K q(x'_k|x'_{k-1}); \quad (5)$$

$$q(x'_k|x'_{k-1}) = \mathcal{N}(x'_k; \sqrt{1 - \beta_k}x'_{k-1}; \beta_k \mathbf{I}), \quad (6)$$

where  $k$  denotes the time step applied to  $x'_k$ , which is different from the time step  $t$  applied to  $x_t$ . Note that  $x_t$  is obtained by adding noise to the original image  $x_0$  in diffusion models, whereas  $x'_k$  is obtained by adding noise to the image condition  $x'_0$  in diffusion models. During training we randomly sample  $x'_k$  from  $\{x'_0, \dots, x'_K\}$ .

While  $x'_k$  inherently preserves the structure information of  $x_0$ , we find that explicitly incorporating additional structural guidance, such as edges, helps the model better capture structural information. Thus, we concatenate  $x'_k$  with the predictions of a Canny Edge detector  $\mathbf{C}(x_0)$ . Thus, the output of our image information removal module is:

$$R(x_0) = [x'_k, \mathbf{C}(x_0)]. \quad (7)$$

Given the output of our image information removal module  $R(x_0)$ , the final objective of IIR-Net is defined as:

$$L_{\text{IIR-Net}} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_\theta(z_t, t, \tau_{\theta_1}(S), \tau_{\theta_2}(R(x_0)))\|_2^2 \right]. \quad (8)$$

## 4. Experiments

### 4.1. Datasets and Experiment Settings

**Datasets.** We evaluate the performance of our model on three standard datasets, CUB [35], Outdoor Scenes [11],

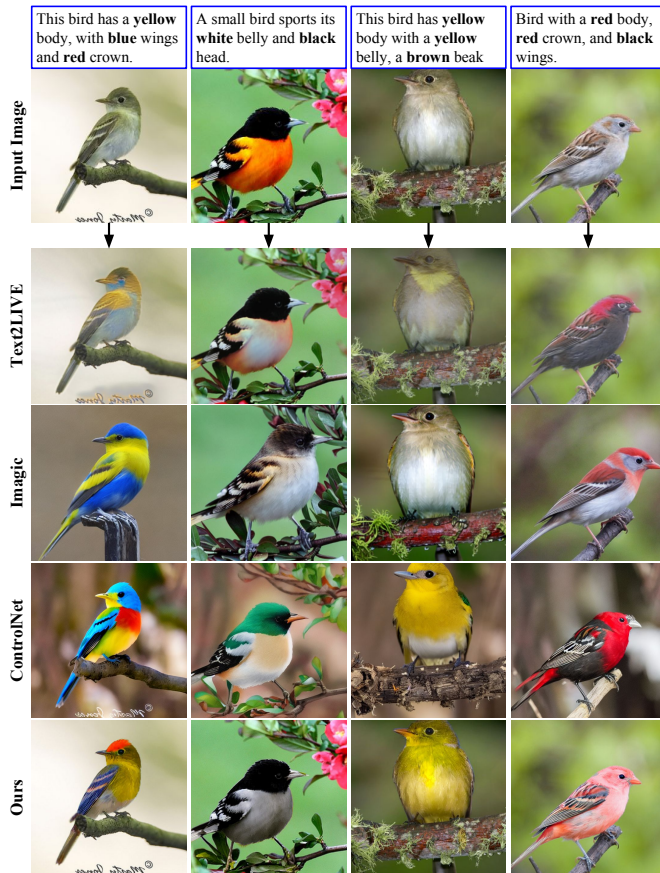
and COCO [17]. CUB [35] contains 200 bird species that we split into 8,855 training images and 2,933 test images. Outdoor Scenes [11] contains 8,571 images captured from 101 webcams, with each webcam collecting 60~120 images showcasing different attributes like weather, season, or time of day. COCO [17] contains 82,783 training images and 40,504 validation images. Following [9], we randomly select 150 test images from each dataset to evaluate the performance of each method.

**Metrics.** Following [9], we adopt the perceptual metric LPIPS [41] and CLIP score [23] as our quantitative metrics. LPIPS measures the image fidelity and CLIP evaluates the model's editability. Additionally, we perform quantitative experiments by user study and inference time to evaluate the effectiveness and efficiency of our model.

**Baselines.** We compare IIR-Net with three state-of-the-art approaches: Text2LIVE [1], Imagic [9], and ControlNet [40]. For Text2LIVE, we set the optimization steps to 600. For Imagic, both the text embedding optimization steps and model fine-tuning steps are set to 500. We sample the interpolation hyperparameter  $\eta$  from 0.1 to 1 with a 0.1 interval, and the guidance scale is set to 3. For ControlNet and IIR-Net, we generate images with a CFG-scale of 9.0, and DDIM steps of 20 by default.

**Implementation Details.** We initialized our model weights from Stable Diffusion 1.5 [28] and ControlNet [40]. During training, we applied a batch size of 8 and a maximum learning rate of  $1 \times 10^{-6}$ . We finetuned our models approximately 100 epochs on the CUB [35] dataset, and around 5 epochs on the Outdoor Scenes [11] and COCO [17] datasets. The training process was parallelized

### (A) CUB



### (B) Outdoor Scenes

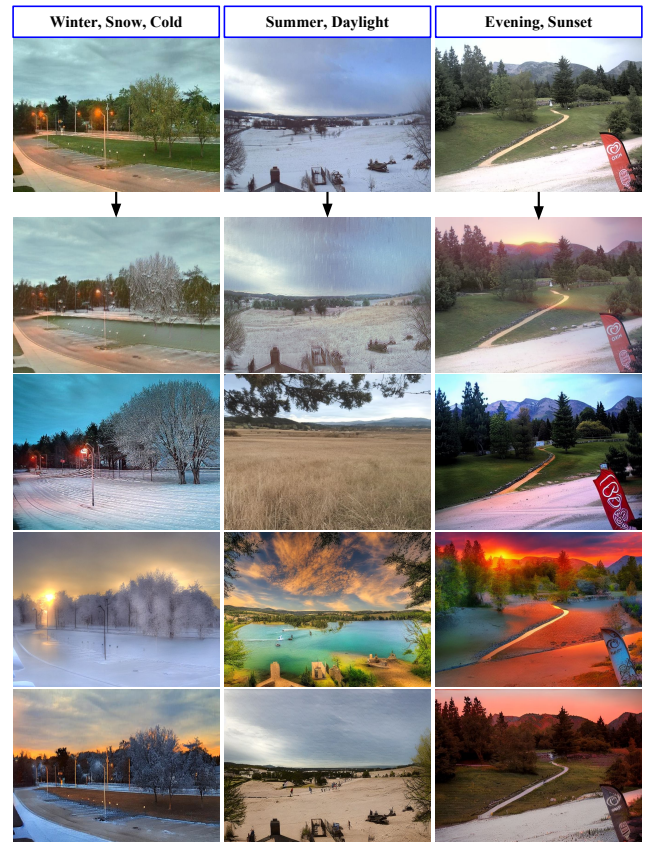


Figure 3. Qualitative comparison on CUB and Outdoor Scenes. From top to bottom: input image, Text2LIVE [1], Imagic [9], ControlNet [40], and ours. Generated images have 512 pixels on their shorter side. See Section 4.2 for discussion.

on 2 NVIDIA RTX-A6000s. To adapt the image conditions in our model, we configured the channel of the image encoder block to 4, with 3 channels for RGB images and 1 channel for the edge map. We finetuned the Stable Diffusion decoder for experiments on CUB, as these images primarily focus on various birds with a consistent style. We froze the Stable Diffusion Decoder for the Outdoor Scenes and COCO datasets, since these datasets comprising natural images with diverse objects and varying styles.

## 4.2. Qualitative Results

### Entire-image Editing on the CUB and Outdoor Scenes Datasets.

Figure 3 presents a qualitative comparison of the edited images generated by our model and the baselines. In Figure 3 (A), we present a comparison on the CUB [35] dataset. We observe that our model can accurately manipulate parts of the bird while preserving the text-irrelevant content of the original image. For example, in the first column of Figure 3 (A), while baselines such as ControlNet and Imagic can recognize “yellow” and “blue” from the text

prompt, they both fail to effectively apply them to the correct corresponding parts of the bird. Imagic generates a bird with a blue crown and yellow wings, while ControlNet generates a blue head and a red breast. In contrast, our model accurately edits the bird by parts according to the prompt and produce a bird with blue wings, yellow body, and red crown. In addition, we observe that the background of images generated by Imagic and ControlNet has been changed. This is due to the fact that Imagic and ControlNet do not directly use the original image as their input. *E.g.*, Imagic optimizes the text embeddings to get features that reflect the attributes of the original image, and ControlNet uses the Canny Edge map as input. Thus, it is challenging for these method to preserve the text-irrelevant content of the original image. In contrast, our model takes the original image as input and only erases the text-relevant content, thus preserving the text-irrelevant content effectively.

In Figure 3 (B), we present a comparison on the Outdoor Scenes [11] dataset. Consistent with our findings on the CUB dataset, we observe that baselines like Imagic and

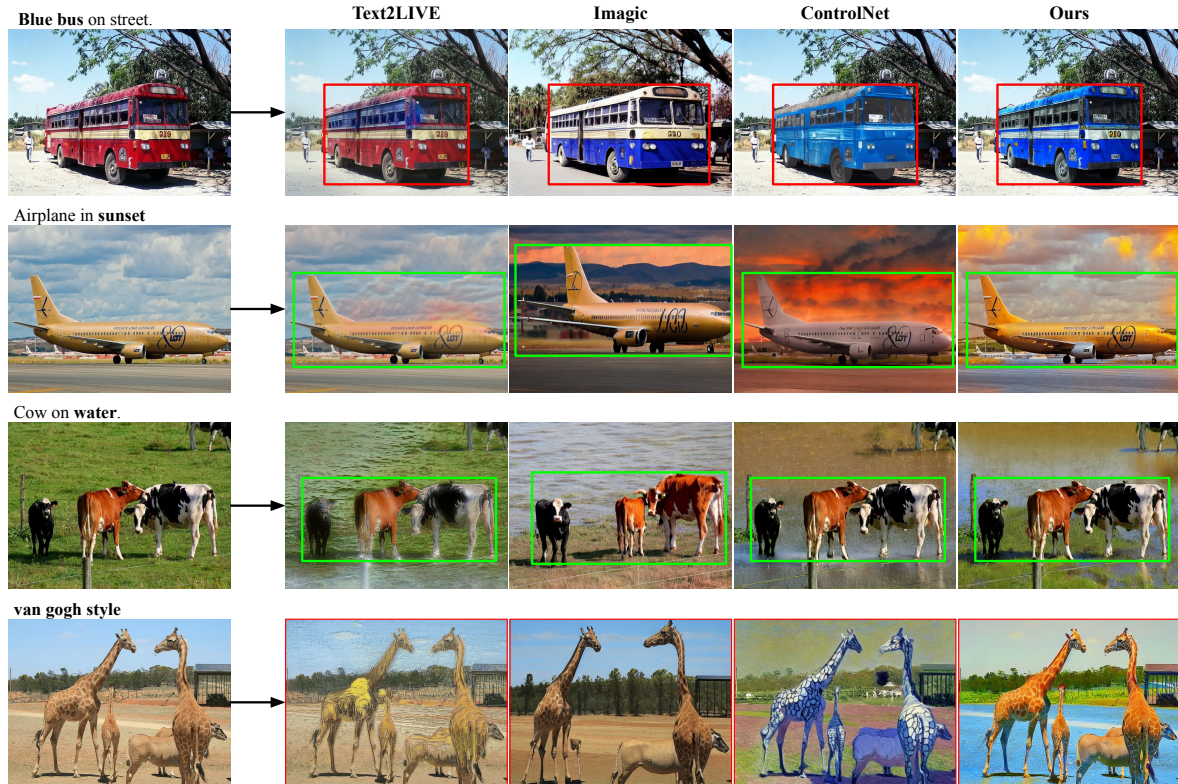


Figure 4. Qualitative comparison for various editing tasks on the COCO dataset. From top to bottom: color editing, scene attribute transfer, texture editing, and style transfer. Generated images have 512 pixels on their shorter side. Objects that are the target of modification are in red bounding boxes and whereas objects that should be preserved are in green bounding boxes. See Section 4.2 for discussion.

ControlNet tend to modify the text-irrelevant contents of the original image, such as the textures and background, while Text2LIVE only introduces limited visual effects to the original image and may fail to generate images aligned with the text descriptions. For example, in the second column of Figure 3 (B), images produced by Imagic and ControlNet are well aligned with text descriptions (“summer,” “daylight”), but they introduce unexpected objects such as trees or a lake to the image. In contrast, Text2LIVE preserves the original image well, but fails to align with text descriptions, as seen with the snow-covered field in summer. However, our method effectively modifies the desired content, such as changing “winter” to “summer,” while preserving the original content of the image.

**Region-based Image Editing on COCO.** Unlike object-centric datasets such as CUB and Outdoor Scenes, COCO images can contain complex scenes with many objects, yet only parts of the input image may require modification. Thus, we apply Grounding-DINO [18] and SAM [10] to localize the Region of Interest (RoI) that requires editing<sup>2</sup>.

Figure 4 presents a qualitative comparison of our method and prior work on various image editing tasks. We find that

<sup>2</sup>Since Text2LIVE and Imagic automatically localize the RoI, we apply Grounding-DINO and SAM to ControlNet and our method.

our method produces images that are well-aligned with the text descriptions while non-edited components better represent the original images. *E.g.*, in the color editing task, although Imagic and ControlNet generate a blue bus according to the text prompt, Imagic changes the original shape of the bus and ControlNet modifies the bus’ texture. In contrast, our method only modifies the color attribute while preserving irrelevant attributes. Furthermore, our model generates images that appear more natural and visually appealing. *E.g.*, in the scene attribute transfer task, the visual effect of “sunset” brought by our model is naturally aligned with the original image, whereas Text2LIVE introduces obvious artificial effects to the airplane.

Finally, we evaluate our model on tasks where the original ControlNet performs well, such as texture editing and style transfer. Our results show that adapting text-to-image generation models to image editing tasks does not notably compromise their capabilities. For example, in the style transfer examples on COCO images our method still retains the ability to transfer a photorealistic image to an artistic style. See the supplementary for additional examples.

**Ablation Study.** In Figure 5, we provide ablation study of IIR-Net. We find that without our unsupervised image content removal mechanism, the model always outputs the

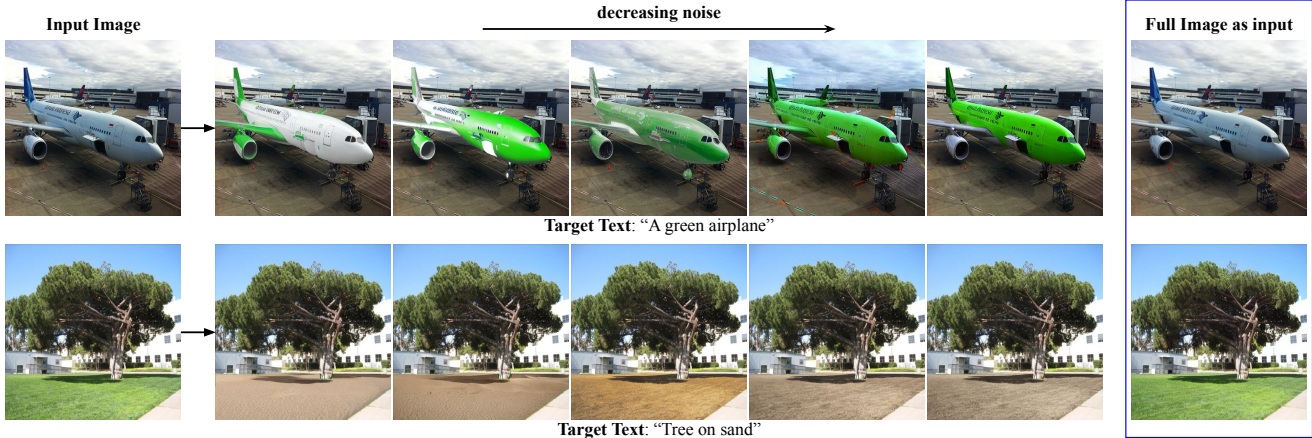


Figure 5. **Ablation Study.** We perform experiments to evaluate the effectiveness of our color removal and texture removal operations. Images generated without our image information removal module are outlined by the blue bounding box. See Section 4.2 for discussion.

input image as the predicted image, *i.e.*, the identical mapping issue [12]. *E.g.*, the images in the blue bounding box remain white airplane and green grass, showing a lack of alignment with the text descriptions. By incorporating the color removal mechanism (see images with low noise level), our model performs well on tasks such as color editing. For example, when changing the airplane’s color from white to green, our model preserves the most of the airplane’s attributes, only modifying the color. We observe that the color removal mechanism can find texture editing challenging. For example, as seen in the second row of the figure, the images generated with low noise level still exhibit the grass texture instead of the intended “sand” texture. Therefore, we incorporate noise augmentation to the input images to better handle such editing tasks. As shown in the second row, our model successfully modifies the grass texture to sand under high-level noise conditions. In practical applications, users can adjust the noise level according to different editing tasks to achieve optimal performance.

### 4.3. Quantitative Results

**Editability-fidelity Tradeoff.** Table 1 reports our quantitative results on CUB, Outdoor Scenes, and COCO. As observed in our qualitative experiments, our model achieves a better tradeoff between image fidelity and editability compared to other state-of-the-art methods. *E.g.*, our model achieves the best LPIPS scores (0.138 and 0.301) and comparable CLIP scores (29.57 and 24.30) on CUB and COCO. In Outdoor Scenes, our model achieves the highest CLIP score and the second best LPIPS score. Text2LIVE achieves better LPIPS score than our method on Outdoor Scenes. However, it may be due to the fact that Text2LIVE mainly augments the scenes with new visual effects, rather than directly modifying the attributes of the scenes. *E.g.*, Text2LIVE fails to change the grassland to a snowy landscape or convert lush trees to bare ones in the scenes.

**User Study.** We conducted a user study to quantitatively evaluate the performance of IIR-Net, as shown in Table 2. We randomly selected 30 images from COCO and applied each model to generate the modified images, resulting in a total of 120 generated images. Each image was annotated three times by users and we asked our annotators to judge whether the image is correctly manipulated based on the text guidance while preserving the text-irrelevant content of the original image. In the table, we report that IIR-Net significantly outperforms baselines. See the supplementary for additional details on our user study.

**Inference Time** Table 2 presents a comparison of the inference time and their standard error using the same Stable Diffusion v1.5 [28] backbone for Imagic, ControlNet, and our method. All methods are benchmarked on a NVIDIA RTX A6000 GPU. We find our method has significantly faster inference times compared to Imagic, boosting inference speed by two orders of magnitude when processing  $512 \times 512$  images. In addition, our method is approximately 50x faster than Text2LIVE. We note both ControlNet and our method have around 5s inference time, demonstrating that approach introduces negligible overhead to ControlNet.

## 5. Limitations & Broader Impacts.

**Limitations.** We identify three failure cases of our methods in this section: First, the attributes of the original image are likely to be modified in non-rigid image editing tasks. Second, it is challenging for our method to change the brightness of the input image drastically. Third, the target object may be localized and segmented inaccurately. We present examples of these three failure cases in Figure 6. As shown in the top row, though our method can achieve non-rigid image editing according to the input image and a modified structural guidance, we observe that the model fails to map some attributes to the correct parts. *E.g.*, the bird of the input image has a grey crown while the edited image gen-

	CUB		Outdoor Scenes		COCO	
	LPIPS ↓	CLIP ↑	LPIPS ↓	CLIP ↑	LPIPS ↓	CLIP ↑
Imagic [9]	0.406	27.03	0.551	22.85	0.567	21.53
Text2live [1]	0.162	<b>30.37</b>	<b>0.218</b>	22.64	0.495	<b>25.11</b>
ControlNet [40]	0.528	29.49	0.618	23.89	0.606	23.57
ours	<b>0.138</b>	29.57	0.479	<b>25.45</b>	<b>0.301</b>	24.30

Table 1. Quantitative experiments of image manipulation on CUB [35], Outdoor Scenes [11], and COCO [17] datasets. CLIP [23] is used to evaluate the image editing performance and LPIPS is applied to evaluate image fidelity. Generated images have been resized to  $224 \times 224$  resolution for CLIP score. We use the “ViT-B/32” version of CLIP. See Section 4.3 for discussion.

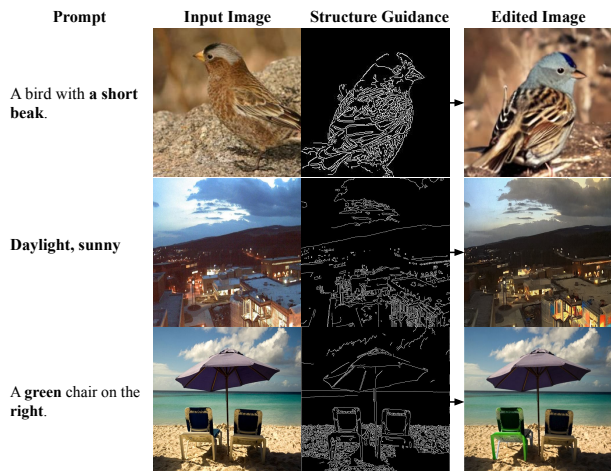


Figure 6. **Failure cases** include inconsistencies with the original image in non-rigid image editing task (top); challenges in notably modifying the brightness of the image (middle), and inaccurate localization or segmentation (bottom). See Section 5 for discussion.

Method	User Preference	Inference Time
Text2LIVE [1]	30.0%	281.6±1.72s
Imagic [9]	23.3%	483.4±1.31s
ControlNet [40]	33.3%	5.0±0.04s
IIR-Net (ours)	<b>68.3%</b>	5.0±0.03s

Table 2. We randomly select 30 images from COCO for user study and speed evaluation. Top row reports user judgments on the correctness of the image manipulation. Bottom row reports speed for our method v.s. the baselines. Our method has negligible overhead compared to ControlNet, and is significantly faster than Text2LIVE and Imagic. See Section 4.3 for discussion.

erate a bird whose head is gray. The color of wings is also slightly different from the input bird. In the middle row, we find that our model fails to change the brightness of the image in some cases. *E.g.*, the input image is a night view. Therefore, the brightness of the image is low in this image and the model tend to reconstruct an image with a low brightness even if the target text is “daylight,” “sunny.” In

the bottom row, we observe that our segmentation module fails to accurately localize the target object according to the prompt due to text ambiguity. While the prompt specifies the chair on the right-hand side, our model modifies the attributes of the chair on the left-hand side.

**Broader Impacts.** Our model is designed to perform image editing according to user-provided language descriptions. Thus, it enables modification of attributes such as colors, textures, or styles in the original images. As other image generation and editing approaches, our model may be used to synthesize images that contains misinformation. Therefore, it is important for practitioners to review and control how images are manipulated to avoid misinformation. Further research on detecting machine-generated images is needed to mitigate this potential issue.

## 6. Conclusion

In this paper, we propose IIR-Net, a text-to-image editing model that incorporates the original image by selectively erasing the image information. IIR-Net mainly consists of two stages: an conditional diffusion model that takes the original image as additional control, and an image information removal module to address the identical mapping issue. We demonstrate that IIR-Net outperforms the state-of-the-art in both qualitative and quantitative evaluations on CUB, Outdoor Scenes, and COCO datasets. For example, compared to Imagic, IIR-Net improves the LPIPS score from 0.57 to 0.30 and the CLIP score from 21.53 to 24.30 on COCO, with a speed improvement of two orders of magnitude. We also use qualitative examples to demonstrate the effectiveness of our model on various image editing tasks, validating that our model can modify the target attribute according to language descriptions while preserving the text-irrelevant content of the original image well.

**Acknowledgements** This material is based upon work supported, in part, by DARPA under agreement number HR00112020054. Any opinions, findings, and conclusions or recommendations are those of the author(s) and do not necessarily reflect the views of the supporting agencies.



## References

- [1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kashtan, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 1, 3, 4, 5, 8
- [2] Dina Bashkirova, Jose Lezama, Kihyuk Sohn, Kate Saenko, and Irfan Essa. Masksketch: Unpaired structure-guided masked image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3
- [3] Helisa Dharmo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5213–5222, 2020. 2
- [4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3
- [5] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022. 3
- [6] 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. *arXiv preprint arXiv:2208.01618*, 2022. 3
- [7] 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
- [8] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3
- [9] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 5, 8
- [10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 6
- [11] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33(4):1–11, 2014. 2, 4, 5, 8
- [12] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2, 7
- [13] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems*, 33:22020–22031, 2020. 2
- [14] Nannan Li and Bryan A Plummer. Supervised attribute information removal and reconstruction for image manipulation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 457–473. Springer, 2022. 2, 3
- [15] Nannan Li, Kevin J Shih, and Bryan A Plummer. Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures. *arXiv preprint arXiv:2210.01887*, 2022. 2
- [16] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 4, 8
- [18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 6
- [19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 2
- [20] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 3
- [21] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018. 2
- [22] 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. In *International Conference on Machine Learning (ICML)*, 2022. 2
- [23] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4, 8
- [24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

- [25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3
- [26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2, 3
- [27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 7
- [29] 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 preprint arXiv:2208.12242*, 2022. 1, 2, 3
- [30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2, 3
- [31] 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 preprint arXiv:2111.02114*, 2021. 2
- [32] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 3
- [33] Ming Tao, Bing-Kun Bao, Hao Tang, Fei Wu, Longhui Wei, and Qi Tian. De-net: Dynamic text-guided image editing adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 2
- [34] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 3
- [35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 4, 5, 8
- [36] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2
- [37] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2951–2960, 2021. 2
- [38] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13789–13798, 2021. 2
- [39] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2
- [40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 4, 5, 8
- [41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4
- [42] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022. 1, 2, 3
- [43] Zhongping Zhang, Huiwen He, Bryan A. Plummer, Zhenyu Liao, and Huayan Wang. Complex scene image editing by scene graph comprehension. In *British Machine Vision Conference (BMVC)*, 2023. 3