iEdit: Localised Text-guided Image Editing with Weak Supervision

Rumeysa Bodur[1*], Erhan Gundogdu[2], Binod Bhattarai[3], Tae-Kyun Kim[1,4], Michael Donoser[2], Loris Bazzani[2]
[1]Imperial College London, UK, [2]Amazon, [3]University of Aberdeen, UK, [4]KAIST, South Korea
r.bodur18@imperial.ac.uk, binod.bhattarai@abdn.ac.uk, kimtaekyun@kaist.ac.kr
{eggundog, donoserm, bazzanil}@amazon.de

Abstract

Diffusion models (DMs) can generate realistic images with text guidance using large-scale datasets. However, they demonstrate limited controllability on the generated images. We introduce iEdit, a novel method for text-guided image editing conditioned on a source image and textual prompt. As a fully-annotated dataset with target images does not exist, previous approaches perform subject-specific fine-tuning at test time or adopt contrastive learning without a target image, leading to issues on preserving source image fidelity. We propose to automatically construct a dataset derived from LAION-5B, containing pseudo-target images and descriptive edit prompts. The dataset allows us to incorporate a weakly-supervised loss function, generating the pseudo-target image from the source image's latent noise conditioned on the edit prompt. To encourage localised editing we propose a loss function that uses segmentation masks to guide the editing during training and optionally at inference. Trained with limited GPU resources on the constructed dataset, our model outperforms counterparts in image fidelity, CLIP alignment score, and qualitatively for both generated and real images.

1. Introduction

Significant progress has been made in developing large-scale text-to-image generative models [29, 32, 33, 36], enabling artists and designers to create realistic images without specialised expertise. Existing methods show limited level of controllability, as they are sensitive to the guiding prompt, *i.e.*, a small change in the input text yields a significantly different output image. Image editing with generative models increases controllability and help artists and designers make their work personal, creative and authentic. Recent methods for image editing [17, 28, 29, 35] incorporate various types of inputs (e.g., text, mask, stroke) to facilitate the generation of more specific content.

In this paper, we propose iEdit, a framework based on Latent Diffusion Models [33] (LDMs) for *text-guided image editing*: given a real or generated image and a user-provided textual prompt for editing, we generate a new image which includes targeted and localised modifications as displayed in Fig. 1. Such editing tasks involve adding or removing objects, changing the appearance of specific regions, or modifying the overall composition of the image. Text-guided image editing requires the ability to *preserve* the fidelity of the shape, style and semantics for some parts of the image, while *synthesising* realistic modifications that are consistent with the edit prompt.To this end, we introduce a method to automatically generate paired datasets specifically suited for image editing, customise LDMs [33] with the ability of performing editing by aligning the generated images with the target text, and provide shape and location awareness to the method by leveraging segmentation masks.

Figure 1. Our method can edit images with a textual prompt while preserving image fidelity in the regions not related to the edit.

A supervised approach to train image editing models requires a fully annotated dataset consisting of triples: source image, edit prompt and target image. These datasets are expensive to collect and include several challenges, such as, ambiguity due to vague annotation instructions or potential bias of too specific instructions. Hence, we follow the weakly-supervised approach and propose a method to automatically generate editing datasets by semantically pairing images from existing image-caption datasets (e.g., LAION-5B [37]) using multi-modal embeddings (e.g., CLIP [31]) to obtain pseudo-target images. Since captions poorly de-

*Work done during an internship with Amazon.

scribe images we propose to generate edit prompts by captioning the source image (e.g., BLIP [24]) and changing its adjectives and nouns. The full process is automatic and enables us to generate a potentially large number of triplets.

`iEdit` leverages the automatically-constructed dataset to fine-tune LDMs for editing, in contrast with subject or sample-specific training methods [2, 7, 14, 21, 28], which use a set of user-provided images and target labels/attributes to overfit the models to generate specific content. Hence, they lack generalisation capabilities and scalability, i.e., a specific model needs to be fine-tuned for each of the set of target images or attributes. Moreover, `iEdit` embeds the semantics of the edit prompt using a contrastive learning loss [21, 31] to align the generated image and edit prompt. To preserve the regions of the source image not mentioned in the edit prompt and apply changes only on relevant regions, we propose a loss function on masks automatically-extracted with an off-the-shelf segmentation model.

To summarise, the contributions of our paper are the following. We present a method to automatically construct paired datasets that are used during training in a weakly-supervised way. We present a novel image editing method that is trained to align source and pseudo-target images with edit prompts, and can edit both generated and real images. We introduce loss functions that use semantic masks to enable localised preservation and synthesis of semantics and regions for editing. Our method can be trained in a lightweight way by fine-tuning subparts of the network backbone. Qualitative and quantitative results demonstrate that our method outperforms state-of-the-art counterparts.

2. Related Work

Text-guided Image Generation. DMs [16, 40] have become the *de-facto* alternative to Generative Adversarial Networks [13] (GANs) for image generation. Stable training allows to increase the capacity of DMs, such as DALL-E 2 [32], GLIDE [29], Imagen [36], and to fairly compare [9] with the GAN-based counterparts [11, 47]. LDMs [33] address the computational limitation by working on a latent low-dimensional latent space. Similarly, we provide LDMs the ability of image editing, while keeping computational resources constrained (2 NVIDIA V100 GPUs), so that training is more attainable by the research community.

Text-guided Image Editing with GANs. GANs were adopted for image manipulation using text guidance. Some studies [1, 30, 46] effectively combine StyleGAN [19] and CLIP [31] latent embeddings. StyleGAN-NADA [12] proposes a text-driven method for out-of-domain generation. ManiGAN [22] proposes to train GAN models with multiple stages with text input. However, GANs are hard to train in a stable way to perform localised editing with large datasets and diversity of input types. To introduce local edits, Text2Live [3] proposes to automatically learn an edit layer that is then combined during image generation.

Text-guided Image Editing with DMs. DMs can be adapted for text-guided image editing to deal with the challenges of GANs. *Prompt-to-prompt* [14] aligns patches using cross-attention without fine-tuning but requires inversion [9] for real images and uses attention maps at lower resolution than masks used in our method. SDEdit [28] uses user guidance, like stroke painting, and can be conditioned to edit prompts [7]. It is sensitive to the strength parameter that can lead to over-preservation or forgetting the source image. These conditional DMs modify the image globally, undesirably changing regions that are not mentioned in the edit prompt. Recent work focuses on fine-tuning on a single image available [20] at test time, a small set of subject-specific images [35] or specialises on attribute specific modifications [35]. However, this leads to high computation at test time and limited scalability, since they need to fine-tune a model for each image, set or attribute. In contrast, our approach requires a single fine-tuning step on the dataset automatically generated for editing.

Other methods, Imagen Editor [44], Repaint [27] and Blended Diffusion [2], introduce the use of masks for image editing. However, in Blended Diffusion and Repaint, the masks are manually-provided and available only at test time while Imagen Editor uses an object detector to randomly mask objects whereas in our method the masks are automatically extracted with regard to the target modification. DiffEdit [7] predicts masks at test time from text-conditioned latent noise differences of the modification in textual descriptions. In contrast, our method uses automatically-extracted masks during training and optionally at test time to provide localised properties to the editing framework. Composer [17] decomposes images into a set of factors, including masks, and recomposes back to images. Our method is lighter (1.1B vs 3B parameters) and requires a smaller training set (200K vs 60M).

Target Data for Image Editing. While domain-specific translation works use paired datasets with conditional DMs [48] or Brownial Bridge DMs [23], there is no large-scale dataset for generic text-guided image editing. In order to automatically construct one, InstructPix2Pix [4] generates instructions from manually written caption-instruction pairs using GPT-3 [5] and uses a pre-trained LDM [14] to generate target images. In contrast, our dataset construction method is based on retrieval and, therefore, lightweight because it does not require DMs to generate synthetic target images. At the same time, we show that `iEdit` performs well with the synthetic dataset from InstructPix2Pix.

3. `iEdit`

We describe here our process for constructing a dataset with pseudo-target images for text-guided image editing. We briefly review LDMs, introduce the proposed image editing framework, and expand it to incorporate location awareness, guiding the editing process during training and inference.

3.1. Paired Dataset Construction

Most methods discussed in Sec. 2 are trained or fine-tuned without a target image, *i.e.*, a ground-truth image after the edit has been applied to the source image. To train image editing models in a weakly-supervised way, we propose a method to automatically obtain pseudo-target images. An optimal image editing dataset requires nearly identical image pairs, differing only in a specific attribute or object, and both source and target images should have captions that highlight the differences between the pairs. Given the absence of such an annotated dataset, and considering the cost and labor involved in manual creation, we automatically construct one using the publicly available LAION-5B image-caption dataset [37].

The LAION-5B source captions, being extracted from the web, are complex and noisy. Thus, we introduce a technique to convert these captions into a desired edit as depicted in Fig. 2. The first step is to generate a simplified version of each image caption using the state-of-the-art image captioning method BLIP [24]. In the second step, the generated captions are manipulated by replacing an adjective or noun with its antonym or a random co-hyponym, using the WordNet library [10], e.g., "a blue and orange *bus*"→ "a blue and orange *train*". The resulting text is our edit prompt. In the final step, we need to obtain a pseudo-target image, given the source image and the edit prompt. To this end, we extract CLIP [31] embeddings of the source image and the edit prompt. After qualitatively evaluating the results of retrieving targets with solely text or image embeddings, we opted for a weighted mean of both, empirically set at 0.6 for text and 0.4 for image. Finally, we use this weighted mean to retrieve the nearest-neighbour image of each image-prompt pair. With this set of operations, we constructed a dataset consisting of $\sim 200K$ image pairs and edit prompts. Compared with the original captions in LAION-5B, our edit prompts are simpler, shorter and contain only essential information (e.g., the original caption for the red dress in Figure 2 is "red spring and autumn female short-sleeve princess dress puff skirt one-piece dress costume dress 90 - 135"). As we prepared our dataset by using holistic image and caption representations, it includes the retrieved pseudo-target images which may not closely align with their source images. This is not ideal for our scenario where a considerable part of the source image should remain unchanged. To mitigate this shortcoming of preparing an in the-wild-dataset for image editing, we propose to incorporate masks as described in Sec. 3.4. We evaluate our dataset in Sec. 4 and provide examples in Fig. 2 and more in the Supplementary, including statistics of the dataset.

3.2. Background: Latent Diffusion Model

In the forward diffusion process of Denoising Diffusion Probabilistic Models (DDPM) [16], the training data is corrupted by adding Gaussian noise and a neural network is used to reverse this process. LDMs [33], use an autoencoder to operate in a lower-dimensional latent space to reduce the computational complexity and increase the inference speed of DMs. An encoder $\mathcal{E}$ compresses the input image x to a latent vector $z = \mathcal{E}(x)$. Both diffusion and denoising process happen in the latent space, and then a decoder $\mathcal{D}$ reconstructs the image from the latent vector, $x = \mathcal{D}(z)$. The denoising step can be *conditioned* with a domain-specific encoder τ_θ projecting the conditioning input to an intermediate representation that is exposed to the denoising U-Nets [34]. The loss for conditional LDM is:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||^2, \quad (1)$$

where ϵ_θ is a U-Net conditioned to the step t uniformly sampled from $\{1, 2, ..., T\}$. The network parameters θ are optimised to predict the noise $\epsilon_1 \sim \mathcal{N}(0, 1)$ that is used for corrupting the encoded version of the input image. At inference time, the trained model is sampled by iteratively denoising $z \sim \mathcal{N}(0, 1)$ using the deterministic DDIM [41].

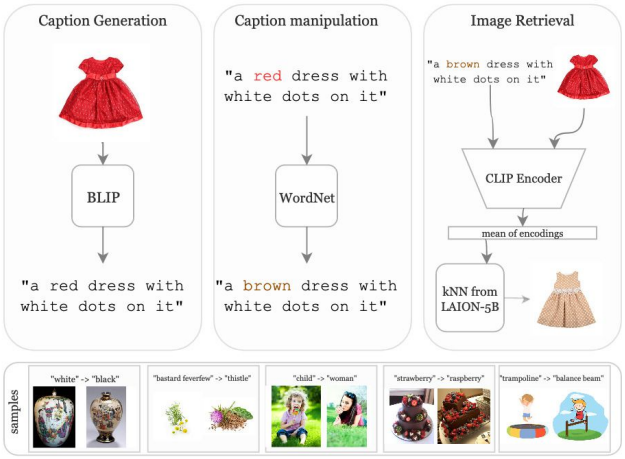

Figure 2. **Paired dataset construction.** To construct a paired dataset for training, we generate captions with BLIP. We manipulate these captions by replacing nouns or adjectives with antonyms or co-hyponyms. We assign the pseudo-target image by using CLIP embeddings of the edit prompt and the source image to retrieve nearest neighbours.

3.3. `iEdit` with Weak Supervision

We present here our method to fine-tune the LDM [33] for image editing using the dataset described in Sec. 3.1. Let x_1 and x_2 be the source and pseudo-target images, respectively, and y_2 the edit prompt of x_2, derived from x_1's caption y_1 as detailed in Sec. 3.1. We first obtain the noisy image z_t by adding noise ϵ_1 to $z_1 := \mathcal{E}(x_1)$, the source image encoded in the latent space:

$$z_t = \sqrt{\alpha_t} z_1 + \sqrt{1 - \alpha_t}\epsilon_1, \quad (2)$$

where $\epsilon_1 \sim \mathcal{N}(0,1)$, and α_t is the Gaussian transition sequence following the notation in [42]. We consider z_t as the noisy version of both the source and target image, since our aim is to generate the target image from z_t. We, then, calculate the ground truth noise for reconstructing $z_2 := \mathcal{E}(x_2)$, which is the target image encoded in the latent space:

$$\epsilon_2 = \frac{z_t - \sqrt{\alpha_t} z_2}{\sqrt{1-\alpha_t}}. \tag{3}$$

Hereby, we modify the objective in Eq. 1 to minimise the L2 loss between this ground truth noise, ϵ_2, and the noise predicted by the network, ϵ_θ, given the edit prompt:

$$\mathcal{L}_{paired} = \mathbb{E}_{\mathcal{E}(x),y_2,\epsilon_2,t}[||\epsilon_2 - \epsilon_\theta(z_t, t, \tau_\theta(y_2))||^2] \tag{4}$$

To further encourage the generated image to be aligned with the edit prompt, we introduce a global CLIP loss [30] between the edit prompt and the generated image $\hat{x}_1$:

$$\mathcal{L}_{global}(\hat{x}_1, y_2) = D_{CLIP}(\hat{x}_1, y_2) \tag{5}$$

where $\hat{x_1}$ is obtained with the decoder $\mathcal{D}$ as follows:

$$\hat{x_1} = \mathcal{D}(\hat{z}_1) \tag{6}$$

$$\hat{z}_1 = \frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t, \tau_\theta(y_2))}{\sqrt{\alpha_t}}. \tag{7}$$

With higher noise levels the reconstruction is more successful and therefore $\mathcal{L}_{paired}$ gives more reliable results. On the other hand, CLIP loss gives more reliable results on low noise levels as CLIP embeddings are ideal for noise-free inputs. Thus, we set inversely proportional weights for these losses based on the noise level t as $(1 - \frac{t}{T})\mathcal{L}_{global} + \frac{t}{T}\mathcal{L}_{paired}$, where T is the maximum number of noise steps.

3.4. `iEdit` with Location Awareness

As pointed out before in Sec. 3.1, the constructed dataset has a misalignment issue. In our preliminary experiments, we observed that it causes noisy results when the model is optimised using Eq.4. Hence we propose to extend `iEdit` to incorporate masks from images in order to enable localised image editing and better align source and pseudo-target images. Specifically, we introduce the use of masks during training to guide the learning process and optionally during inference to generate localised edits. To obtain the masks, we use CLIPSeg [26], a state-of-the-art method that generates image segmentations conditioned to text prompts at test time. We take the differences between the BLIP-generated source caption and the edit prompt along with the noun it describes in case it is an adjective. This difference prompt y_1^{diff} (or y_2^{diff}) and the corresponding image x_1 (or x_2) are fed into the CLIPSeg model to obtain masks for both source and target images, M_1 and M_2, respectively.

Training with Masks. In Figure 3, we visualise an overview of our method for incorporating masks during training. To enable localised edits, we modify the optimisation process to predict the target noise, ϵ_2, on the masked region and the source noise, ϵ_1, on the inverse mask region, rather than optimising the network to predict only the target noise. Specifically, Eq. 4 becomes:

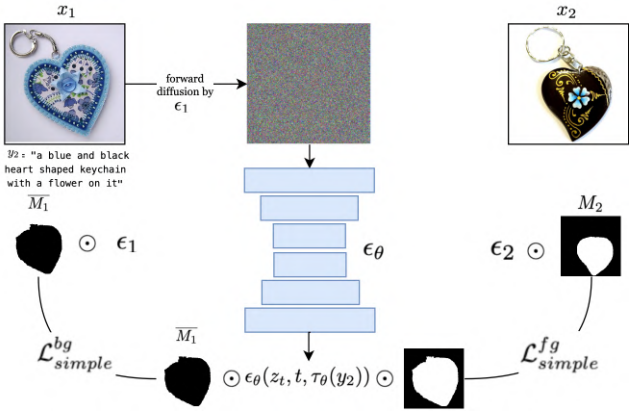

Figure 3. **Proposed image editing framework.** Our framework takes as input the image pairs from the constructed dataset, their corresponding masks and the edit prompt. We optimise ϵ_θ to predict the noises ϵ_1 and ϵ_2 for background and foreground. $\odot$ denotes element-wise multiplication of the image and mask.

$$\mathcal{L}_{mask} = \mathbb{E}_{\mathcal{E}(x),y_2,\epsilon_2,t}\mathcal{L}^{fg}_{mask} + \mathbb{E}_{\mathcal{E}(x),y_2,\epsilon_1,t}\mathcal{L}^{bg}_{mask} \tag{8}$$

where $\mathcal{L}^{fg}_{mask}$ and $\mathcal{L}^{bg}_{mask}$ represent the foreground and background loss terms as follows:

$$\begin{aligned} \mathcal{L}^{fg}_{mask} &= [||\epsilon_2 \odot M_2 - \epsilon_\theta(z_t, t, \tau_\theta(y_2)) \odot M_2||^2] \\ \mathcal{L}^{bg}_{mask} &= [||\epsilon_1 \odot \overline{M_1} - \epsilon_\theta(z_t, t, \tau_\theta(y_2)) \odot \overline{M_1}||^2] \end{aligned} \tag{9}$$

where $\overline{M}_i$ denotes the inverse of the mask, and $\odot$ represents element-wise multiplication. While localising the edit, the incorporation of masks in this loss will also mitigate the misalignment issue present in the constructed dataset, as only the overlapping areas will serve as supervision by the pseudo-target, while the rest of the image relies on the source image, as in the original LDM loss (Eq. 1).

In addition to $\mathcal{L}_{mask}$, we also employ two additional losses, namely a perceptual loss [18] and a localised CLIP loss, to better translate according to the edit prompt. The perceptual loss is used to ensure that the edited image has a similar visual appearance to the target image in their masked regions. Specifically, we use a pre-trained VGG network [39] $V(\cdot)$ to extract features from the edited image $\hat{x}_1$ and the target image x_2 and minimise the mean squared error between their feature representations:

$$\mathcal{L}_{perc} = \mathbb{E}_{\hat{x}_1,x_2}[||V(\hat{x}_1 \odot M_1) - V(x_2 \odot M_2)||^2]. \tag{10}$$

The localised CLIP loss aims to ensure that the masked area of the generated image is coherent with the difference of the edit prompt with the source caption, y_2^{diff}, in terms of semantic content. This loss is defined as:

$$\mathcal{L}_{loc}(\hat{x}_1, y_2^{diff}) = D_{CLIP}(\hat{x}_1 \odot M_1, y_2^{diff}). \quad (11)$$

In summary, our final loss for fine-tuning `iEdit` is:

$$(1 - \frac{t}{T})(\mathcal{L}_{loc} + \mathcal{L}_{global}) + \frac{t}{T}\mathcal{L}_{mask} + \lambda_{perc}\mathcal{L}_{perc}. \quad (12)$$

Inference with Masks. In SDEdit [28], during inference, random noise is added to the input image, and this corrupted image is denoised with the trained model to generate the edited version. We extend this approach by incorporating masks in order to generate localised edits while preserving the visual content of the source image in the inverse mask region. Given an input image, an edit prompt, and which term in the edit prompt describes the edit area, we apply Gaussian noise corruption to the input image for a certain number of iterations defined by a sampling ratio. We vary this hyper-parameter between $0.6 - 0.8$ throughout our experiments for both SDEdit and our method, where 0.0 and 1.0 correspond to using the input image or a random Gaussian noise as input. At each DDIM [42] sampling step, we replace the pixel values of the reconstructed image (*i.e.*, $\tilde{z}_t$) in the inverse mask region $\overline{M}$ with their corrupted version of the input image, z_t, at the given time step t, as

$$\tilde{z}_t = \overline{M} \odot z_t + M \odot \tilde{z}_t. \quad (13)$$

The inference with this additive masking described above helps to preserve the details of original latent input z_1, progressively at each noise level t. Finally, we follow the decoding step as in Eq. 6 to obtain the final image $\hat{x_1}$.

4. Experiments

In this section, we perform a thorough comparison of `iEdit` with state-of-the-art methods by quantitative and qualitative analysis on generated and real source images.

4.1. Experimental Setup

Datasets. We fine-tune `iEdit` with the paired dataset constructed from LAION-5B [37] as described in Sec. 3.1. For evaluation with generated images, we use LDM [33] to construct 70 image-prompt pairs, with 20 distinct generated images. For evaluation with real images, we build 52 image-prompt pairs consisting of 30 distinct images from COCO [25], ImageNet [8] and AFHQ [6]. Additional training details are provided in the supplemental material.

Comparison to Other Methods. We compare to the SDEdit [28] extension of SD since we use a similar diffusion process and to recent state-of-the-art methods, such as DALL-E 2[1] [32], DiffEdit [7] and InstructPix2Pix[2] [4]. For DALL-E 2, we manually provide ground-truth masks for each image during evaluation.

Evaluation Metrics. In text-based image editing, we have to evaluate with respect to: 1) how well the method synthesises the regions of the image explicitly mentioned in the edit prompt and 2) the quality of preservation of the rest of the image not mentioned in the prompt. We adapted the Structural Similarity Index (SSIM) [45] to capture this intuition as: the SSIM score on the edited area (SSIM-M) and on the rest of the image (SSIM-$\overline{M}$) by using manually created ground-truth masks. In addition, we use CLIPScore [15] to measure the alignment between the edit prompt and the edited image by computing the cosine similarity between their embeddings generated by the CLIP model. CLIPScore has been shown in [44] to be reliable due to its high agreement with human judgement. We compute the Fréchet Inception Distance (FID) [38] to evaluate the quality and fidelity of generated images.

4.2. Qualitative Evaluation

Editing Generated Images. Fig. 4 compares `iEdit` (our method) and `iEdit`-M (with predicted masks at inference) with SDEdit [28], DALL-E 2 [32], DiffEdit [7], InstructPix2Pix [4] on the generated images of the evaluation set (Sec. 4.1) along with heatmaps that visualise the difference between the input and generated images. We observe that SDEdit [28] often fails to perform the desired edit, e.g., (*"a heart-shaped candle"* in col. 2, *"latte art"* in col. 4). It also does not preserve fidelity, e.g., background in col. 1 and attributes of the plate in col. 4. This is due to SDEdit's trade-off between preservation of the input and editing according to the edit prompt. DALL-E 2 [32] shows better fidelity to the edit prompt, but it ignores the prior information regarding the object to edit from the source image. It produces results that completely replace the edited object with a visually different one (col. 1 and 2) while also failing in some cases (col. 1, 4 and 5). The user-provided masks in DALL-E 2 help localising the edit, but its in-painting nature ignores the structure of the edited region from the source images. This further causes the cases that do not blend well with the general structure of the image (col. 1). DiffEdit [7] relies on automatically detected masks at inference time which are sensitive to input/output prompts and model performance. Consequently, although it does a good job in preserving the background it fails in some cases to perform the manipulation, e.g. col. 1, 4 and 5. InstructPix2Pix[3] displays fidelity to both source image

[1]Web UI accessed in March 2023: https://openai.com/product/dall-e-2.

[2]Web UI accessed in March 2023: https://huggingface.co/spaces/timbrooks/instruct-pix2pix.

[3]Note that our method is trained with limited resources and a batch size of 1, while [4] uses a batch size of 1024.

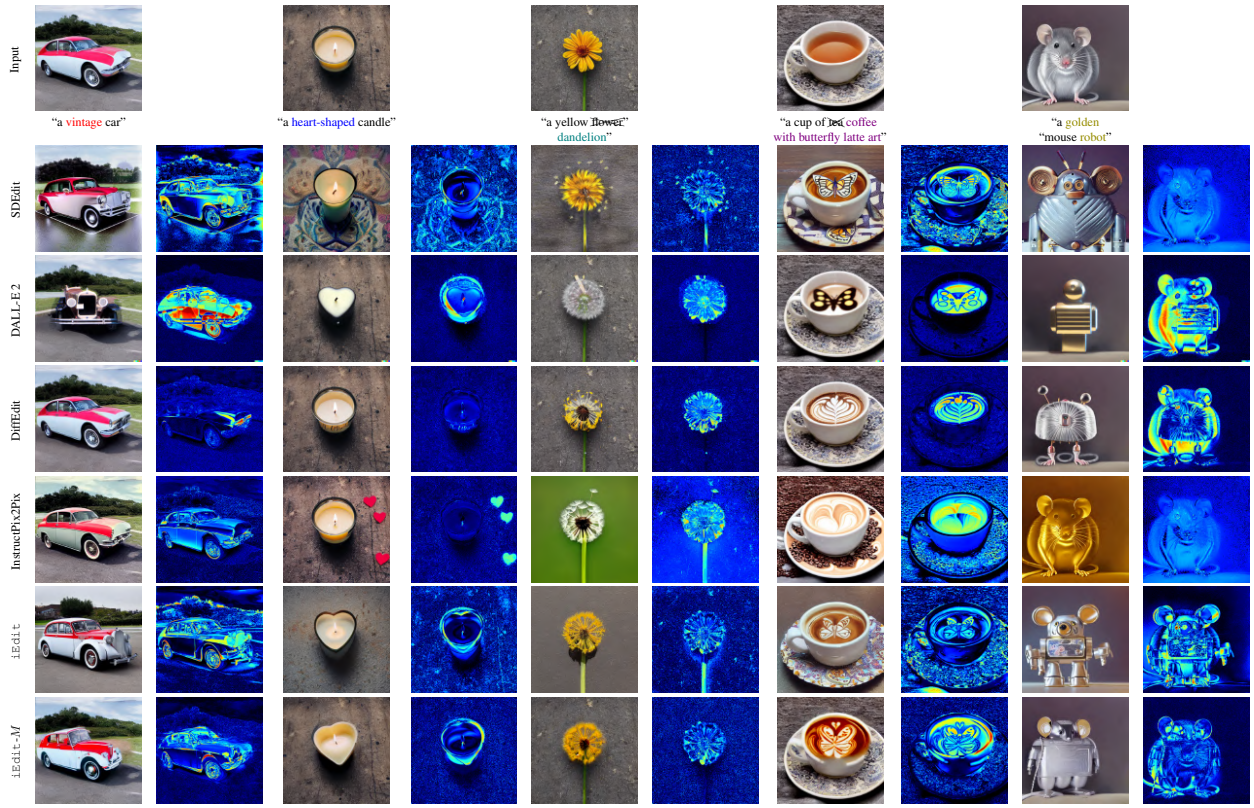

Figure 4. **Comparison to state-of-the-art on generated images.** Our method produces results with higher fidelity to the source image and the edit prompt compared to SDEdit [28], DALL-E 2 [32], DiffEdit [7] and InstructPix2Pix [4].

and edit prompt except for some failure cases (col. 2 and 5), however, it affects the whole layout in many cases, e.g., *"a vintage car"* in col. 1 produces a vintage-style image (including background), and the image in col. 4 results in coffee beans over the image. This is a limitation [43] of InstructPix2Pix inherited from Prompt-to-Prompt [14], used to generate the training set. It also fails when multiple changes are requested at once, (*"golden robot mouse"* in col. 5). Compared to these methods, `iEdit` performs more successful localised edits, while preserving the background, with small undesired changes. This is alleviated with the use of predicted masks in `iEdit`-M at inference.

Editing Real Images. One of the main advantages of our method is the ability to edit real images without inversions [14, 43]. Fig. 5 shows a qualitative comparison on the real images of the test set (Sec. 4). In accordance with the observations on editing generated images, SDEdit [28] achieves low fidelity to edit prompt (col. 3 and 4), or high deviation from the source image (background in col. 2 and 4, and *"bus"* in col. 8). DALL-E 2 [32] yields results that are inconsistent with the source or look unnatural (col. 1 and 3), and also fails in col. 6 and 8. As previously mentioned, DiffEdit [7] heavily relies on detected masks and provided prompts. Results indicate increased challenges with real images, leading to frequent failures in achieving satisfactory outcomes in col. 2, 3, 4, 5, and 8. The limitation of InstructPix2Pix [4] results do not comply with the style of the source image (col. 2), look unrealistic (col. 1) or affect the whole image (col. 3 and 4). `iEdit` shows high fidelity to the source image and the edit prompt on real images as well, where undesired background changes are averted with the introduction of masks in `iEdit`-M.

4.3. Quantitative Evaluation

The left side of Table 1 shows a comparison of state-of-the-art methods and `iEdit` for editing generated images in our evaluation set. In terms of CLIPScore, `iEdit` and `iEdit`-M outperform all compared methods, which means high alignment and fidelity between the generated image and the edit prompt. Note that CLIPScore is one of the main metrics for editing, since it was shown to have high agreement with human judgement [44]. We further measure FID, where our method is only marginally outperformed by DiffEdit. As SSIM-M measures the SSIM score of the edited area, its high values do not imply a better result, because the edited area could be too similar to that of the source image. In editing, we want the SSIM-M to be a trade-off between changing the source image accordingly to the edit prompt

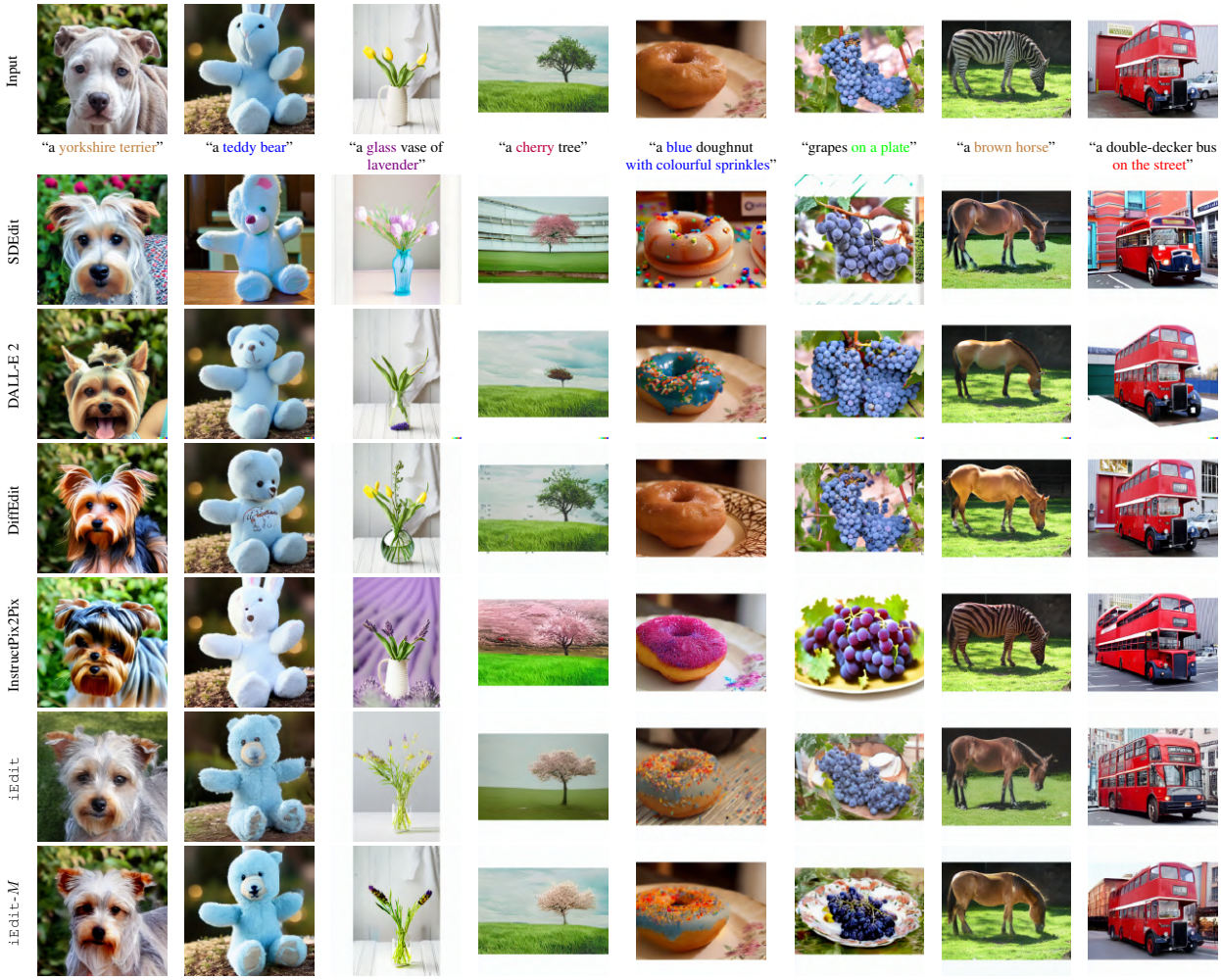

Figure 5. **Qualitative Results on Real Images.** iEdit outperforms compared methods showing high fidelity to the edit prompt and the input image.

Method	Generated Images				Real Images			
	CLIPScore (%) ↑	FID↓	SSIM-M(%)	SSIM-$\overline{M}$(%) ↑	CLIPScore (%) ↑	FID↓	SSIM-M(%)	SSIM-$\overline{M}$(%) ↑
SDEdit [28]	62.58	171	82.44	50.64	65.84	180	74.36	64.60
DALL-E 2 [32]	65.44	143	82.45	**94.76**	65.46	162	74.41	**93.97**
DiffEdit [7]	60.31	**95**	89.31	78.14	64.39	**100**	82.06	<u>91.88</u>
InstructPix2Pix [4]	65.12	108	88.62	76.43	66.91	145	80.59	79.92
iEdit (iP2P dataset)	63.99	106	84.73	76.66	65.62	132	81.25	79.13
iEdit-M (iP2P dataset)	63.04	<u>100</u>	84.87	77.33	65.93	<u>125</u>	80.82	80.18
iEdit (ours)	<u>65.76</u>	158	82.70	52.02	<u>67.02</u>	166	74.59	70.09
iEdit-M (ours)	**66.36**	114	83.08	<u>78.18</u>	**67.44**	147	74.98	80.44

Table 1. **Quantitative Results on Generated and Real Images.** Best numbers are marked in **bold**, second best <u>underlined</u>.

and not drastically. So, we attribute the high SSIM-M of InstructPix2Pix and DiffEdit to the fact that generated images are too similar to source images and might indicate that the editing was not fully successful. The lower values for iEdit(-M) imply more changes in the foreground mask, that combined with the higher CLIPScore and the qualitative results in Fig. 4 and 5, we can conclude that the editing is preserving information from the source image while considering the edit prompt correctly. In terms of preserving unrelated parts of images in the background, DALL-E 2 shows the best SSIM-$\overline{M}$ while iEdit-M has the second best result. DALL-E 2 has an unfair advantage in our experiments, since it uses manually-provided masks at inference which tightly defines the manipulation area while we use CLIPSeq masks generated by a prompt that consists of y^{diff} or desired location in case of adding an object.

Ablation Settings		Scores			
Losses	**Fine-tuning Dataset**	**CLIPScore (%)** ↑	**FID**↓	**SSIM-M(%)**	**SSIM-$\overline{M}$(%)** ↑
$\mathcal{L}_{global} + \mathcal{L}_{pairs}$	LAION-caption-200K	65.95	158	78.84	55.14
$\mathcal{L}_{global} + \mathcal{L}_{LDM}$+$\mathcal{L}_{loc}$	LAION-edit-200K	65.62	156	79.61	62.08
$\mathcal{L}_{global} + \mathcal{L}_{LDM}$+$\mathcal{L}_{loc}$+$\mathcal{L}_{perc}$	LAION-edit-200K	65.64	153	79.75	<u>62.20</u>
$\mathcal{L}_{global} + \mathcal{L}_{mask}$+$\mathcal{L}_{loc}$+$\mathcal{L}_{perc}$	LAION-edit-200K	<u>66.09</u>	<u>146</u>	79.31	59.54
$\mathcal{L}_{global} + \mathcal{L}_{mask}$+$\mathcal{L}_{loc}$+$\mathcal{L}_{perc}$ +Masked Inference	LAION-edit-200K	**66.97**	**128**	79.65	**77.99**

Table 2. Ablation study of `iEdit` on our loss functions and datasets.

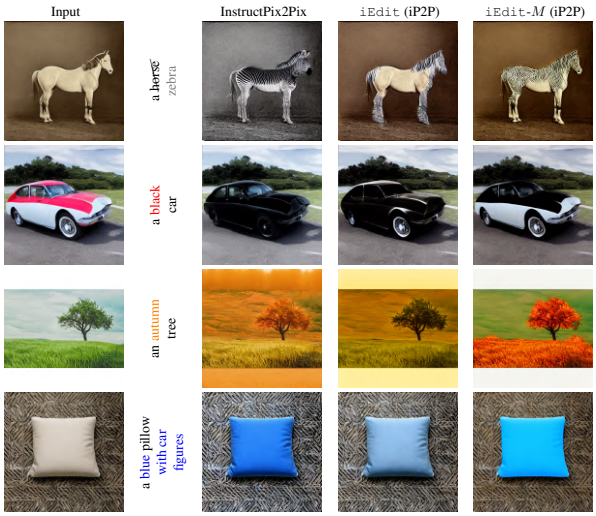

Figure 6. Comparison of InstructPix2Pix and iEdit trained on InstructPix2Pix dataset (iP2P).

On the right side of Table 1, we present a quantitative evaluation of our method and compared methods on editing the real images in our evaluation set. We observe a similar trend to the results obtained on generated images.As before, even without using masks at inference, our method produces results with higher fidelity to the edit prompt in terms of CLIPScore. Conclusions similar to the case of generated images can be drawn for FID, SSIM-M and SSIM-$\overline{M}$. To summarise, we conclude that our method is the best tradeoff between preserving the fidelity of the source image considering both foreground and background (SSIM and FID) and being more aligned with the edit prompt (CLIPScore).

In order to assess the utility of the proposed way to construct our dataset, we trained `iEdit` using the dataset constructed in InstructPix2Pix ("IP2P") with and without integrating masks during training. From Fig. 6, it's evident that certain limitations observed in InstructPix2Pix, such as the production of artificial-looking images, manipulations affecting the whole image (*"an autumn tree"*) and difficulties in concurrently altering multiple aspects (*"a blue pillow with car figure"*), are also observed in our approach. While the incorporation of masks does offer some mitigation, these challenges persist. This can be attributed to the dataset's construction via another generative method, Prompt-to-prompt [14], leading to the inheritance of its weaknesses. From the results of `iEdit` (IP2P) in Table 1we can infer that better aligned paired dataset helps preserving the background (low FID and high SSIM-$\overline{M}$), but falls short in translation especially when editing real images as supported by the low CLIPScore and SSIM-M scores. This highlights the ongoing significance of learning from real data which further motivates the usefulness of our dataset construction strategy and the applicability of our method on both synthetic and real datasets.

4.4. Ablation Study

Table 2 shows an ablation study that validates the effectiveness of each component of `iEdit` using the full evaluation set (generated and real images). First, we compare LAION-caption-200K (1^{st} row) and LAION-edit-200K (2^{nd} row). LAION-edit-200K is the dataset constructed as described in Sec. 3.1 and LAION-caption-200K is created in the same way but using the original captions from LAION-5B instead of the automatically-generated ones. We use the best experimental setup for LAION-caption-200K, which employs $\mathcal{L}_{global} + \mathcal{L}_{pairs}$. This is the best setup because the difference between these noisy source and target captions of LAION-5B are too high (even different languages sometimes) and often not grounded to images. This prevents the use of masks in $\mathcal{L}_{mask}$ and of $\mathcal{L}_{perc}$ and $\mathcal{L}_{loc}$. The 1^{st} and 2^{rd} rows of Table 2 show the benefit of automatically-generated prompts with higher CLIPScore and SSIM-$\overline{M}$. Moreover, we ablate our main losses and components forming the masked image editing framework, by training only with the localised CLIP loss $\mathcal{L}_{loc}$ (2^{nd} row), by adding $\mathcal{L}_{perc}$ (3^{rd} row) and by switching $\mathcal{L}_{LDM}$ in favour of $\mathcal{L}_{mask}$ (4^{th} row). The addition of $\mathcal{L}_{perc}$ marginally boosts the performance in all metrics, while $\mathcal{L}_{mask}$ leads to a large improvement in CLIPScore and FID. The final row of Table 2 that uses predicted masks at inference shows a large improvement in all metrics.

5. Conclusion

We presented `iEdit`, a novel method for text-guided image editing based on LDMs. In addition, we proposed a method to automatically create a dataset to train our model with weak supervision. We introduced loss functions that use masks to enable localised editing and preservation of fidelity. Our model is fine-tuned on the constructed dataset by using only 2 16GB GPUs. `iEdit` achieves favourable qualitative and quantitative results against state-of-the-art methods on editing of generated and real images.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*. Association for Computing Machinery, 2022. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 2

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision (ECCV)*, pages 707–723. Springer, 2022. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 2, 5, 6, 7, 1, 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1877–1901. Curran Associates, Inc., 2020. 2

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 5, 6, 7, 1, 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 2

[10] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 3

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022. 2

[12] 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)*, 41:1 – 13, 2021. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2014. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. 2, 6, 8

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[17] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions, 2023. 1, 2

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 2

[20] 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. *arXiv preprint arXiv:2210.09276*, 2022. 2

[21] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022. 2

[22] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H.S. Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[23] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models, 2023. 2

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. 2, 3

[25] 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 *European conference on computer vision*. Springer, 2014. 5

[26] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 4

[27] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. pages 11451–11461, 2022. 2

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 5, 6, 7, 3

[29] Alexander Quinn 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, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804. PMLR, 2022. 1, 2

[30] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2, 4

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[32] 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. 1, 2, 5, 6, 7, 3

[33] Robin Rombah, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5, 7

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[35] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 1, 2

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R. Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 5

[38] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 5

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[40] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. 2

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 4, 5

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 6, 1

[44] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting, 2022. 2, 5, 6

[45] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Ieee Transactions on Image Processing*, 13, 2004. 5

[46] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse image generation and manipulation. *CoRR*, abs/2012.03308, 2020. 2

[47] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *The 32nd British Machine Vision Conference (BMVC)*, 2021. 2

[48] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models, 2023. 2