This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Diffusion-Based Conditional Image Editing through Optimized Inference with Guidance

Hyunsoo Lee[1] Minsoo Kang[1] Bohyung Han[1,2]
[1]ECE & [2]IPAI, Seoul National University
{philip21, kminsoo, bhhan}@snu.ac.kr

Abstract

We present a simple but effective training-free approach for text-driven image-to-image translation based on a pretrained text-to-image diffusion model. Our goal is to generate an image that aligns with the target task while preserving the structure and background of a source image. To this end, we derive the representation guidance with a combination of two objectives: maximizing the similarity to the target prompt based on the CLIP score and minimizing the structural distance to the source latent variable. This guidance improves the fidelity of the generated target image to the given target prompt while maintaining the structure integrity of the source image. To incorporate the representation guidance component, we optimize the target latent variable of diffusion model's reverse process with the guidance. Experimental results demonstrate that our method achieves outstanding image-to-image translation performance on various tasks when combined with the pretrained Stable Diffusion model.

1. Introduction

Diffusion-based text-to-image generation models [12, 21, 22] have shown superior performance in generating high-quality images. These models have also been adapted for text-driven image editing tasks, aiming to translate a given source image into a target domain while preserving its overall structure and background. However, text-driven image-to-image translation remains a challenging task, as it requires selectively modifying specific parts of the image while maintaining the integrity of the background and structure. Moreover, achieving precise control over fine details further adds to the complexity of this task.

To address these tasks, many text-driven image editing methods [2, 12, 13, 30] rely on additional fine-tuning of pretrained diffusion models. While these methods yield promising results, they are impractical due to the substantial computational and memory overhead required for the fine-tuning process. In contrast, training-free algorithms [5, 8, 18, 19, 29] introduce unique inference techniques for diffusion models, achieving their objectives without fine-tuning. However, despite their efficiency, these approaches often struggle to preserve the structure of the source image and tend to produce blurry results.

We propose a simple yet effective training-free image-to-image translation method built upon pretrained diffusion models. Our approach incorporates representation guidance based on a triplet loss to enhance fidelity to the target task. Specifically, we modify the reverse process of the diffusion model by integrating the representation guidance, which comprises two key components: (a) a CLIP [20]-based objective that ensures the sampled target latent aligns semantically with the given target prompt, and (b) a structural constraint that enforces similarity between the source and target images' structural information, as extracted from the feature maps of the pretrained diffusion model. Experimental results demonstrate that the proposed reverse process effectively preserves the background and structure of image for both local and global editing tasks. The main contributions of our work are summarized below:

- We propose a novel representation guidance mechanism motivated by metric learning, leveraging features from the pretrained CLIP and Stable Diffusion for text-driven image-to-image translation.
- Our approach modifies the denoising process of the pretrained Stable Diffusion model without any additional training procedure.
- Experimental results on various image-to-image translation tasks using both real and synthetic images verify the outstanding performance of our approach compared to existing methods.

2. Related Work

2.1. Text-to-image diffusion models

Existing methods [21–23] based on diffusion models have demonstrated remarkable performance in text-to-image gen-

eration tasks. For instance, Stable Diffusion [22] leverages pretrained autoencoders [14] to project a given image onto a low-dimensional latent space and estimates its distribution within this manifold, rather than modeling the raw data distribution directly. Imagen [23] employs a large pretrained text encoder to generate text embeddings, which are then used to condition the diffusion model for image synthesis. Similarly, DALL·E 2 [21] predicts the CLIP image embedding from a text caption and synthesizes an image conditioned on both the estimated CLIP embedding and the textual input.

2.2. Diffusion-based image manipulation methods

Text-driven image manipulation aims to preserve the structure and background of the source image while selectively editing the image to align with the target prompt. Several works [5–8, 15, 16, 19, 29] employ publicly available pretrained text-to-image diffusion models, such as Stable Diffusion [22], to deal with the image editing tasks. For example, DiffEdit [5] adaptively interpolates between the source and target latents at each time step, guided by an estimated object mask. Prompt-to-Prompt [8] substitutes the self-attention and cross-attention maps of the source latents for those retrieved from the target latents. Plug-and-Play [29] injects the self-attention and intermediate feature maps obtained from the source latents into the target image generation process. Pix2Pix-Zero [19], on the other hand, optimizes the target latent to align with cross-attention maps extracted from the pretrained diffusion model, based on both the source and target latents. Null-text inversion [18] first inverts a source image using an optimization-based pivotal tuning procedure to achieve precise reconstruction and then applies the Prompt-to-Prompt technique to generate the target image based on the inverted image. MasaCtrl [4] modifies traditional self-attention mechanisms in diffusion models, allowing the model to utilize the local features extracted from the source image, hereby ensuring consistency across generated images. Conditional score guidance [15] derives a score function conditioned on both the source image and source prompt to guide target image generation. Prompt interpolation-based correction [16] refines the noise prediction for the target latent by progressively interpolating between the source and target prompts in a time-dependent manner. Unlike these approaches, InstructPix2pix [2] fine-tunes the pretrained Stable Diffusion using source images and generated pairs of text instructions and target images.

3. Diffusion-Based Image-to-Image Translation

This section describes a simple DDIM-based method tailored for the text-driven image-to-image translation task.

3.1. Inversion process of source images

Diffusion models [9, 25, 26] generate images through inversion and reverse processes, which correspond to the forward and backward processes, respectively. In the inversion process, the original data $\mathbf{x}_0$ is progressively perturbed with Gaussian noise, resulting in the sequence of intermediate latent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. For text-driven image-to-image translation, existing algorithms [5, 19] often employ the deterministic process of DDIM [26] using pretrained text-to-image diffusion models. The deterministic DDIM inversion process is defined as follows:

$$\begin{aligned}\mathbf{x}^{\text{src}}_{t+1} &= f^{\text{inv}}_t(\mathbf{x}^{\text{src}}_t) \\ &= \sqrt{\alpha_{t+1}}\left(\frac{\mathbf{x}^{\text{src}}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}^{\text{src}}_t, t, \mathbf{y}^{\text{src}})}{\sqrt{\alpha_t}}\right) \\ &\quad + \sqrt{1-\alpha_{t+1}}\epsilon_\theta(\mathbf{x}^{\text{src}}_t, t, \mathbf{y}^{\text{src}}),\end{aligned} \tag{1}$$

where $f^{\text{inv}}_t(\cdot)$ denotes the inversion process at time step t, $\epsilon_\theta(\cdot,\cdot,\cdot)$ is the noise prediction network, $\mathbf{x}^{\text{src}}_t$ is the noisy source image at time step t, and $\mathbf{y}^{\text{src}}$ is the CLIP [20] embedding of the source prompt p^{src}. Note that $\mathbf{x}^{\text{src}}_T$ is obtained by recursively applying Eq. (1) starting from the source image $\mathbf{x}^{\text{src}}_0$, and it is subsequently used in the reverse process to generate the target image.

3.2. Reverse process of target images

The target image $\mathbf{x}^{\text{tgt}}_0$ is generated from $\mathbf{x}^{\text{tgt}}_T$, which is set equal to $\mathbf{x}^{\text{src}}_T$, using the DDIM reverse process given by

$$\begin{aligned}\mathbf{x}^{\text{tgt}}_{t-1} &= f^{\text{rev}}_t(\mathbf{x}^{\text{tgt}}_t) \\ &= \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}^{\text{tgt}}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}^{\text{tgt}}_t, t, \mathbf{y}^{\text{tgt}})}{\sqrt{\alpha_t}}\right) \\ &\quad + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{x}^{\text{tgt}}_t, t, \mathbf{y}^{\text{tgt}}),\end{aligned} \tag{2}$$

where $f^{\text{rev}}_t(\cdot)$ is the reverse process function at time step t, and $\mathbf{x}^{\text{tgt}}_t$ and $\mathbf{y}^{\text{tgt}}$ denote the target image at time step t and the CLIP feature of the target prompt p^{tgt}, respectively. However, recursively applying Eq. (2) to synthesize the target image often fails to preserve the overall structure and background of the source image. Therefore, we modify the reverse process to generate the desired target image in a training-free manner.

4. Proposed Approach

This section elaborates on our sampling strategy, which optimizes the deterministic reverse process of DDIM with respect to the representation guidance term.

4.1. Overview

We propose the optimized objective $\mathcal{L}^{\text{dist}}_t$, referred to as representation guidance, to address the limitations of the recursive application of Eq. (2), which often fails to preserve the structure of the source image. The key idea is to regulate the reverse process, ensuring that the target image aligns

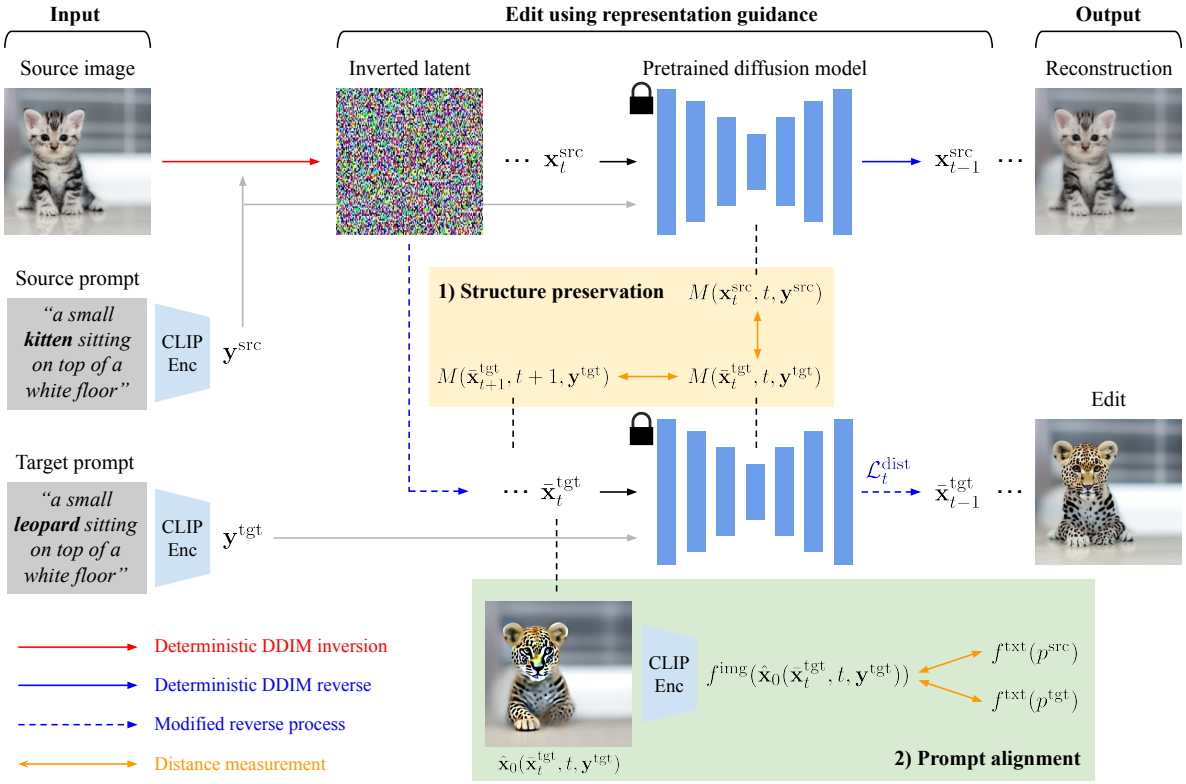


Figure 1. Overview of the proposed method about utilizing the representation guidance.

semantically with the target prompt while maintaining the overall structure of the source image through the guidance term. The modified reverse process is given by

$$\begin{aligned}\bar{\mathbf{x}}_{t-1}^{\text{tgt}} &= \bar{f}_t^{\text{rev}}(\bar{\mathbf{x}}_t^{\text{tgt}}) \\ &= \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\bar{\mathbf{x}}_t^{\text{tgt}} - \sqrt{1-\alpha_t}\gamma_t\epsilon_\theta(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) \\ &\quad - \nabla_{\bar{\mathbf{x}}_t^{\text{tgt}}}\mathcal{L}_t^{\text{dist}}, \end{aligned} \tag{3}$$

where $\bar{f}_t^{\text{rev}}(\cdot)$ is defined by our modified reverse process at time step t, and γ_t is equal to $\sqrt{\frac{\alpha_{t-1}}{\alpha_t}} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}$. We will discuss how $\mathcal{L}_t^{\text{dist}}$ is calculated in Section 4.3. Algorithm 1 summarizes the detailed procedures of the proposed method, referred to as Optimized Inference with Guidance (OIG).

4.2. Naïve representation guidance

During the modified reverse process, we simply derive a naïve distance objective, which is defined as

$$\begin{aligned}\mathcal{L}_t^{\text{naive-dist}} := &-\text{Sim}\Big(f^{\text{img}}(\hat{\mathbf{x}}_0(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})), f^{\text{txt}}(p^{\text{tgt}})\Big) \\ &+\beta_f \|M(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) - M(\mathbf{x}_t^{\text{src}}, t, \mathbf{y}^{\text{src}})\|_F, \end{aligned} \tag{4}$$

where $f^{\text{img}}(\cdot)$ and $f^{\text{txt}}(\cdot)$ are CLIP image and text encoders, respectively, $\text{Sim}(\cdot,\cdot)$ represents the cosine similarity between two vectors, $\|\cdot\|_F$ denotes the Frobenius norm,

Algorithm 1 Optimized Inference with Guidance (OIG)

Inputs: A source image $\mathbf{x}_0^{\text{src}}$, a source prompt $\mathbf{y}^{\text{src}}$, a target prompt $\mathbf{y}^{\text{tgt}}$
for $t \leftarrow 0, \cdots, T-1$ **do**
  Compute $\mathbf{x}_{t+1}^{\text{src}}$ using Eq. (1)
end for
$\bar{\mathbf{x}}_T^{\text{tgt}} \leftarrow \mathbf{x}_T^{\text{src}}$
for $t \leftarrow T, \cdots, 1$ **do**
  Compute $\hat{\mathbf{x}}_0(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})$ using Eq. (5).
  Compute $\mathcal{L}_t^{\text{dist}}$ using Eq. (6).
  Compute $\gamma_t \leftarrow \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}$.
  Compute $\bar{\mathbf{x}}_{t-1}^{\text{tgt}}$ using Eq. (3).
end for
$\mathbf{x}_0^{\text{tgt}} \leftarrow \bar{\mathbf{x}}_0^{\text{tgt}}$
Output: A target image $\mathbf{x}_0^{\text{tgt}}$

$M(\cdot,\cdot,\cdot)$ extracts an intermediate feature map from the noise prediction network $\epsilon_\theta(\cdot,\cdot,\cdot)$, and β_f is a hyperparameter. Also, $\hat{\mathbf{x}}_0(\cdot,\cdot,\cdot)$ is the estimated image sample based on the following Tweedie's formula [27]:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t, t, \mathbf{y}) = \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, \mathbf{y})}{\sqrt{\alpha_t}}. \tag{5}$$

Table 1. Quantitative comparisons of our method with DiffEdit [5], Plug-and-Play [29] Pix2Pix-Zero [19], Null-text inversion [18], and MasaCtrl [4] using the pretrained Stable Diffusion [22] and images sampled from the LAION-5B dataset [24]. InstructPix2pix [2] is included only for reference because it requires extra fine-tuning. For the the drawing → oil painting task, we do not report the BD score since the background is not clearly defined. Black and red bold-faced numbers represent the best and second-best performance in each column.

Method	cat → dog			dog → cat			dog → crochet dog			horse → zebra			zebra → horse			drawing → oil painting	
	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)
DiffEdit	0.297	0.025	0.191	0.291	0.033	0.118	**0.304**	0.031	0.107	**0.322**	0.030	0.090	0.287	0.033	0.167	0.278	0.047
Plug-and-Play	0.273	**0.023**	0.224	0.278	**0.019**	**0.106**	0.294	**0.020**	0.122	0.310	**0.022**	**0.067**	0.275	**0.025**	**0.122**	0.259	0.035
Pix2Pix-Zero	**0.300**	0.030	**0.189**	**0.295**	0.034	0.144	0.303	0.029	0.123	0.316	0.047	0.160	**0.289**	0.031	0.140	0.288	**0.023**
Null-text inv.	0.296	0.029	0.200	0.294	0.026	0.118	**0.304**	0.021	**0.093**	0.311	**0.025**	0.086	**0.299**	0.032	0.144	0.291	0.036
MasaCtrl	**0.299**	0.039	0.395	0.293	0.045	0.225	0.291	0.039	0.169	0.270	0.033	0.199	0.276	0.043	0.242	**0.293**	0.073
OIG (Ours)	0.298	**0.020**	**0.153**	**0.297**	**0.019**	**0.084**	**0.315**	**0.015**	**0.071**	**0.324**	0.026	**0.067**	**0.289**	**0.021**	**0.112**	**0.293**	**0.013**
InstructPix2pix	0.279	0.010	0.129	0.276	0.031	0.107	0.310	0.036	0.154	0.304	0.039	0.139	0.252	0.024	0.311	0.303	0.026

The second term of the right-hand side in Eq. (4) is inspired by the ability of intermediate feature maps extracted from the noise prediction network to capture the semantic information of the source image, as demonstrated in [29]. This term promotes the preservation of semantic features from the source image in the generated target image.

4.3. Representation guidance

Building on the metric learning framework [11], we refine the naïve objective in Eq. (4) by introducing a stricter constraint. Specifically, we guide the generation of the target image to align more closely with the target prompt p^{tgt} than with the source prompt p^{src}. Furthermore, we encourage the generated target latent $\bar{\mathbf{x}}_t^{\text{tgt}}$ to remain semantically closer to the source latent $\mathbf{x}_t^{\text{src}}$ than $\bar{\mathbf{x}}_{t+1}^{\text{tgt}}$, thereby effectively preserving critical information from the source image.

Since the target latent $\bar{\mathbf{x}}_{t+1}^{\text{tgt}}$ at the previous time step inherently contains less semantic information about the source image than the source latent $\mathbf{x}_t^{\text{src}}$, we propose a stricter constraint than simply aligning $\bar{\mathbf{x}}_t^{\text{tgt}}$ with $\mathbf{x}_t^{\text{src}}$. Specifically, we encourage $\bar{\mathbf{x}}_t^{\text{tgt}}$ to diverge from the target latent $\bar{\mathbf{x}}_{t+1}^{\text{tgt}}$ at the previous timestep. This design is motivated by the intuition that such a strategy better preserves the semantic content of the source image during the reverse process. To enforce this, we impose another constraint such that the distance between $M(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})$ and $M(\mathbf{x}_t^{\text{src}}, t, \mathbf{y}^{\text{src}})$ remains smaller than the distance between $M(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})$ and $M(\bar{\mathbf{x}}_{t+1}^{\text{tgt}}, t+1, \mathbf{y}^{\text{tgt}})$. Empirical results show that this strategy effectively helps preserve the structure and semantic information of the source image.

In summary, we define the representation guidance via the effective distance objectives to realize the aforementioned two constraints as follows:

$$
\begin{aligned}
&\mathcal{L}_t^{\text{dist}} := \\
&-\lambda_1 \min(0,\ \text{Sim}(f^{\text{img}}(\hat{\mathbf{x}}_0(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})), f^{\text{txt}}(p^{\text{tgt}})) \\
&\quad - \text{Sim}(f^{\text{img}}(\hat{\mathbf{x}}_0(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}})), f^{\text{txt}}(p^{\text{src}})) - \beta_p) \\
&+\lambda_2 \max(0,\ \beta_f \| M(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) - M(\mathbf{x}_t^{\text{src}}, t, \mathbf{y}^{\text{src}}) \|_F \\
&\quad - \| M(\bar{\mathbf{x}}_t^{\text{tgt}}, t, \mathbf{y}^{\text{tgt}}) - M(\bar{\mathbf{x}}_{t+1}^{\text{tgt}}, t+1, \mathbf{y}^{\text{tgt}}) \|_F), \qquad (6)
\end{aligned}
$$

where λ_1, λ_2, and β_p are hyperparameters.

5. Experiments

We compare the proposed algorithm, referred to as OIG, with existing methods [2,4,5,18,19,29], using the pretrained Stable Diffusion [22]. Additionally, we present an ablation study to analyze the effects of the proposed components.

5.1. Implementation details

We implement the proposed method using PyTorch, based on the publicly available code of Pix2Pix-Zero [19]. To speed up the translation of given source images, we reduce the number of denoising timesteps to 50 for all compared algorithms including the proposed method. Additionally, we replace the original captions with those generated by Bootstrapping Language-Image Pre-training (BLIP) [17], as the original captions often include languages other than English. The target prompts are constructed from the source prompts to reflect the tasks at hand. For instance, in the cat → dog task, we replace the token most closely related to "cat" in the source prompt with the "dog" token based on the CLIP text encoder. Note that we obtained results from the official codes of DiffEdit [5], Plug-and-Play [29], Pix2Pix-Zero [19], Null-text inversion [18] and MasaCtrl [4] using the same source and target prompts with the classifier-free guidance [10] and the pretrained Stable Diffusion v1-4 checkpoint for fair comparisons. For the InstructPix2pix [2] experiments, we use the official implementation and employ text instructions corresponding to the image-to-image translation tasks, as specified in the InstructPix2pix protocol, instead of using the source and target prompts.

5.2. Experimental settings

Dataset and tasks We select about 250 images for each task from the LAION-5B dataset [24], based on the highest CLIP similarity with the text description corresponding to the source task [1] to compare our method with state-of-the-art algorithms. For evaluation, we follow the standard experimental protocols of existing methods [18, 19] for image-to-image translation tasks. Specifically, we focus on

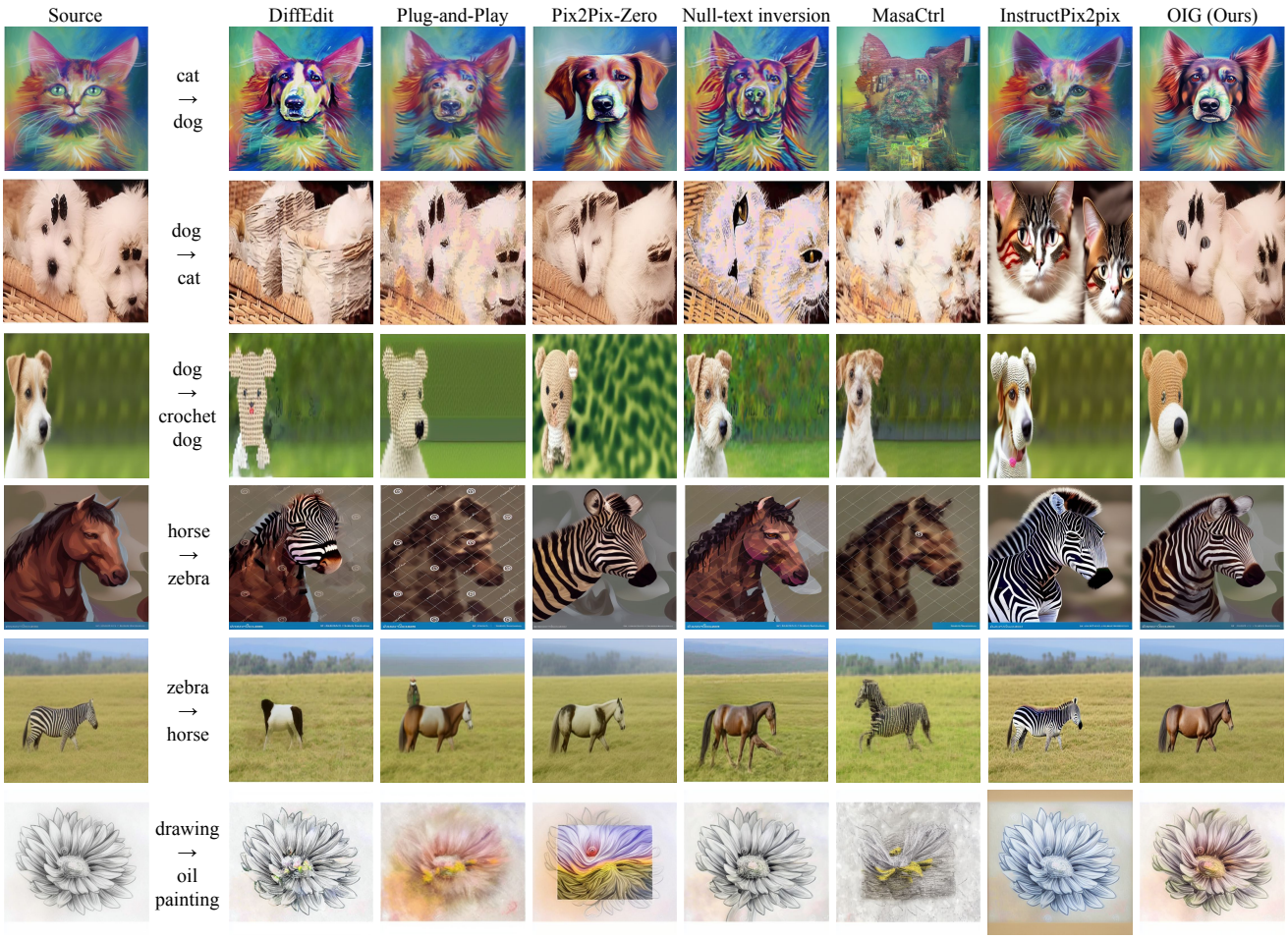


Figure 2. Qualitative comparisons between the proposed algorithm and state-of-the-art methods [2, 4, 5, 18, 19, 29] on the data sampled from the LAION-5B dataset [24] using the pretrained Stable Diffusion [22]. Note that we strictly keep the original aspect ratio of each image during all experiments, and we change the aspect ratio only for visualization.

object-centric tasks such as cat → dog, dog → cat, horse → zebra, and zebra → horse. Additionally, we include the dog → crochet dog task, where the goal is to transform the dog into one that resembles a yarn figure. We also test the proposed approach by transforming hand-drawn sketches into oil paintings in the drawing → oil painting task.

Evaluation metrics We evaluate 1) how well the synthesized target image aligns with the target prompt, and 2) how effectively the structure of the source image is preserved after translation. First, we measure the similarity between the target prompt and the generated target image using CLIP [20], which we call CLIP Similarity (CS). To assess the overall structural difference between the source and target images, we use the self-similarity map of ViT [28] extracted from both images. We then compute the squared Euclidean distance between their feature maps, which we call Structure Distance (SD). Additionally, to evaluate how well each algorithm preserves the background, we identify the background components of both the source and target images by removing the object parts using the pretrained segmentation model, Detic [31]. We measure the squared Euclidean distance between the background regions of the two images, which we denote as Background Distance (BD).

5.3. Quantitative results

We present quantitative results in Table 1 to compare the proposed method with state-of-the-art training-free approaches [4, 5, 18, 19, 29] on various tasks using the pretrained Stable Diffusion [22] and 250 real images sampled from the LAION-5B dataset for each task. As shown in the table, our method outperforms the compared algorithms on most metrics, demonstrating its effectiveness for text-driven image-to-image translation tasks. Additionally, we include the results of InstructPix2pix [2], which requires additional fine-tuning of GPT-3 [3] and Stable Diffusion, unlike the proposed training-free method. However, since the primary

Table 2. Ablation study results to analyze the effect of the proposed component using the pretrained Stable Diffusion [22] and real images sampled from the LAION-5B dataset [24] for various tasks. DDIM employs the reverse process defined in Eq. 2 while Naïve Distance replaces the triplet-based distance objective of the representation guidance in Eq. 6 with the naive distance term in Eq. 4 and Distance utilizes the representation guidance based on the triplet loss.

Method	cat → dog			dog → cat			dog → crochet dog			horse → zebra			zebra → horse			drawing → oil painting	
	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)	BD (↓)	CS (↑)	SD (↓)
DDIM	0.289	0.072	0.347	0.289	0.063	0.224	0.299	0.069	0.242	0.310	0.081	0.255	0.289	0.079	0.221	0.269	0.128
Naïve Distance	**0.298**	**0.020**	0.159	**0.297**	0.020	0.087	**0.316**	0.017	0.080	0.322	**0.023**	0.081	**0.290**	0.027	0.143	0.292	0.016
OIG (Ours)	**0.298**	**0.020**	**0.153**	**0.297**	**0.019**	**0.084**	0.315	**0.015**	**0.071**	**0.324**	0.026	**0.067**	0.289	**0.021**	**0.112**	**0.293**	**0.013**

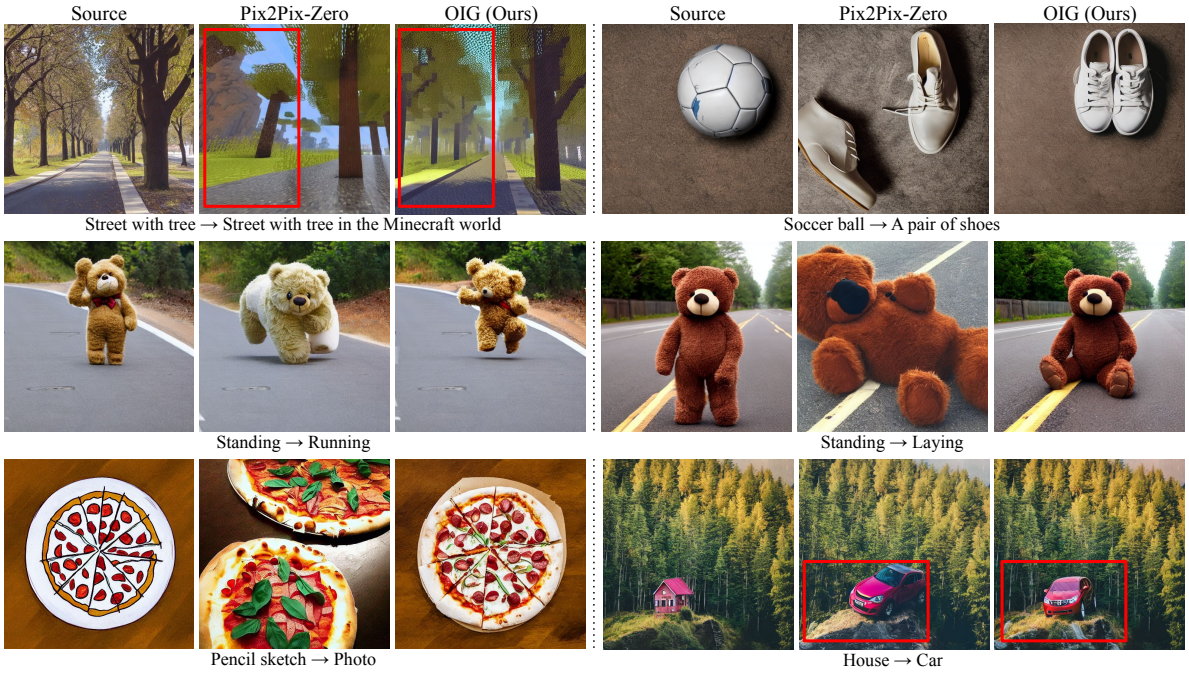


Figure 3. Qualitative comparisons between the proposed algorithm and Pix2Pix-Zero [19] on synthetic images given by Stable Diffusion [22].

objective of InstructPix2pix–mapping text instructions to image-to-image translations–differs from our goal, its results are reported only for reference.

5.4. Qualitative results

We visualize translated results in Figure 2 given by the state-of-the-art methods [2, 4, 5, 18, 19, 29] and the proposed approach, all using the pretrained Stable Diffusion and images sampled from the LAION-5B dataset. As presented in the figure, our method demonstrates superior text-driven image editing performance compared to other methods. Unlike our approach, competing algorithms often struggle to maintain the structural integrity of the source image in the translated outputs.

To demonstrate the superior performance of the proposed method, we qualitatively compare it with Pix2Pix-Zero using synthesized images generated by Stable Diffusion. As shown in Figure 3, our method significantly outperforms prior approaches, while Pix2Pix-Zero often fails to preserve structure and introduces noticeable artifacts. We also show additional editing examples of the proposed method on the synthesized images given by the pretrained Stable Diffusion in Figure 4, which verifies the effectiveness of our method. Notably, our method demonstrates outstanding performance even on multi-subject editing tasks. Additional qualitative comparisons and examples are provided in Appendix E.

Furthermore, we introduce coherence guidance, which is an additional component designed to improve the fine details of the target image. Motivated by CycleGAN [32], we leverage the concept of cycle-consistency between the source and target domains using a pretrained diffusion model. Empirical results demonstrate that the proposed approach effectively refines minor details in the target image generated through representation guidance, improving overall quality. Comprehensive explanations and qualitative results of the coherence guidance are provided in Appendix B. Note that the results shown in the main paper were obtained without incorporating the coherence guidance.

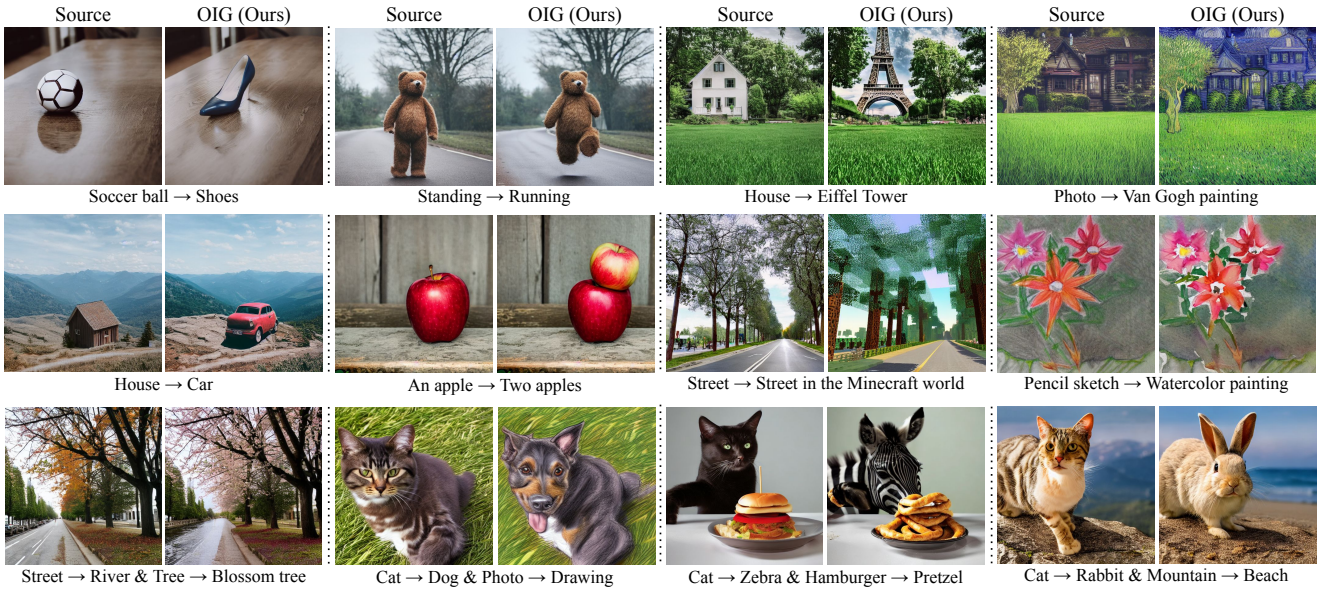


Figure 4. Editing examples of the proposed method on synthetic images given by the pretrained Stable Diffusion [22]. Given the source and target prompts, our method generates a target image while successfully preserving the overall structure of the source image and maintaining the background excluding the parts to be manipulated.

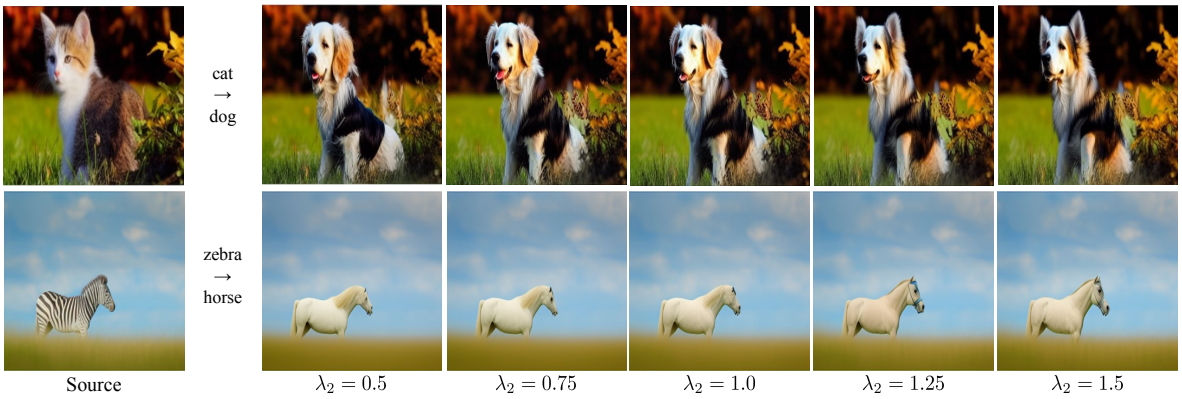


Figure 5. Qualitative reseults of the sensitivity analysis about the hyperparameter λ_2 using the pretrained Stable Diffusion [22] and real images sampled from the LAION-5B dataset [24].

5.5. Ablation study

To analyze the impact of each proposed component, we conduct an ablation study on various tasks using synthetic images generated by the pretrained Stable Diffusion and real images sampled from the LAION-5B dataset. As demonstrated in Table 2, Figure 6, and Appendix C, the triplet-based distance term in Eq. (6) produces target images with higher fidelity compared to the naïve distance objective described in Eq. (4).

Furthermore, we qualitatively analyze the impact of the hyperparameter λ_2 on the results of the proposed method. As illustrated in Figure 5, our approach exhibits robustness across a range of λ_2 values from 0.5 to 1.5. Empirical observations reveal that smaller λ_2 values enhance alignment between the target image and the target prompt but may compromise the structural consistency with the source image. In contrast, larger λ_2 values better preserve the structure of the source image in the generated results, albeit at the expense of slightly reduced fidelity to the target prompt.

6. Conclusion

We proposed a simple but effective method for text-driven image-to-image translation tasks based on pretrained text-to-image diffusion models. To enhance the fidelity of the target images to the given prompt, we introduced representation guidance based on a triplet-based distance objective for the reverse process. For the derivation of the distance objec-

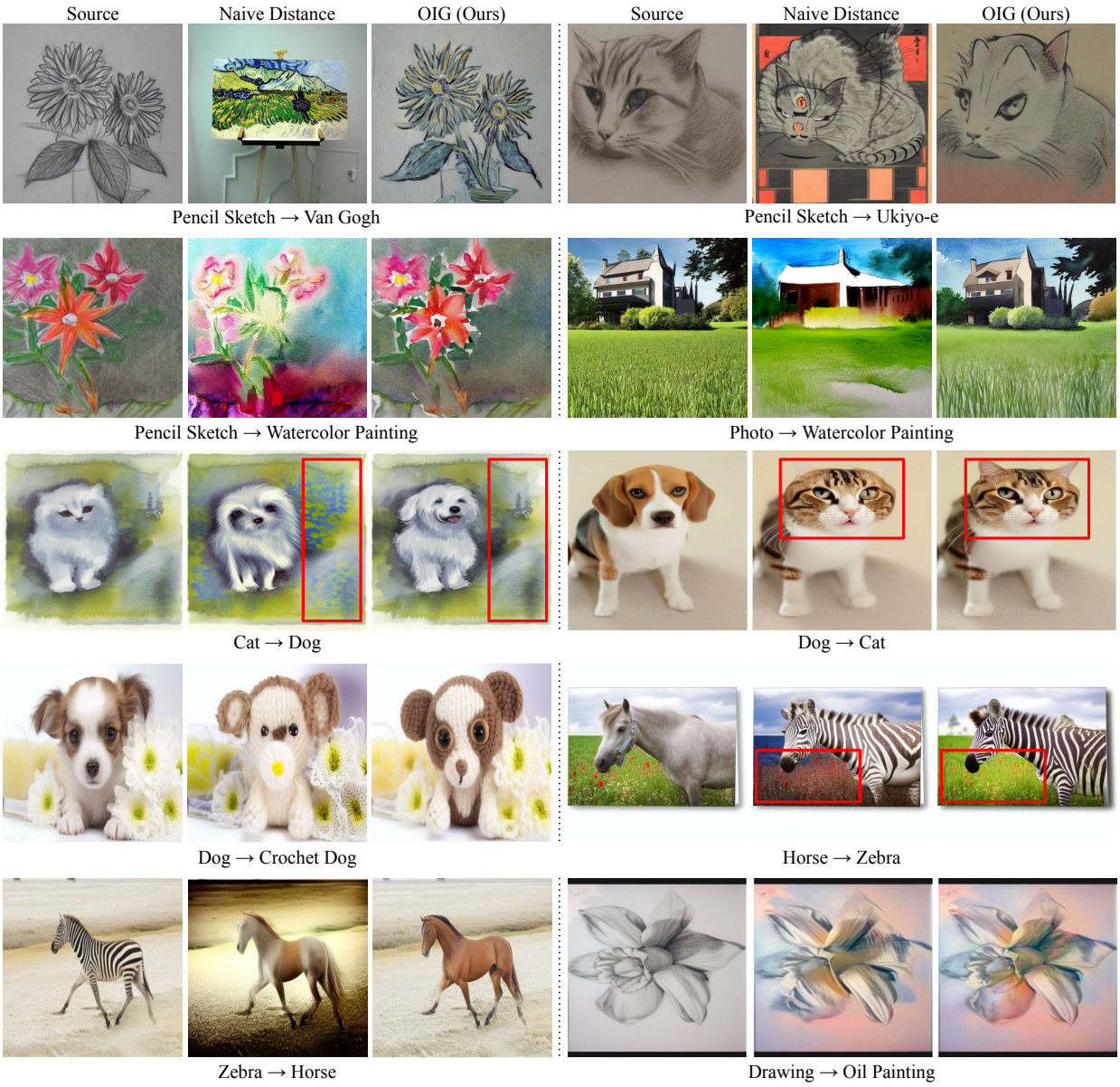


Figure 6. Qualitative results of our components on synthetic images (1st - 2nd row) and the data from LAION-5B dataset [24] (3rd - 5th row) using the pretrained Stable Diffusion [22]. Representation guidance significantly modifies the details of the target image, such as removing the artifacts, improving the backgrounds, and strictly fitting the structure into the source image.

tive, we encourage the target latents to align closely with the target prompts compared with the source prompts for fidelity while the target latents become closer to the source latents rather than the previous target latents for preserving the structure or background of the source image. To demonstrate the effectiveness of our approach, we conducted extensive experiments on various image manipulation tasks and compared the results with state-of-the-art image-to-image translation methods. Experimental results show that our framework achieves outstanding qualitative and quantitative performance in a variety of scenarios when combined with the pretrained Stable Diffusion model.

Acknowledgement

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) [No.RS2022-II220959 (No.2022-0-00959), No.RS-2021-II211343, No.RS-2021-II212068] and the National Research Foundation (NRF) [No.RS-2021-NR056445 (No.2021M3A9E408078222)] funded by the Korea government (MSIT).

References

[1] Romain Beaumont. CLIP Retrieval: Easily Compute CLIP Embeddings and Build a CLIP Retrieval System with Them. https://github.com/rom1504/clip-retrieval, 2022. 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to Follow Image Editing Instructions. In *CVPR*, 2023. 1, 2, 4, 5, 6

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020. 5

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 2, 4, 5, 6

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *ICLR*, 2023. 1, 2, 4, 5, 6

[6] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models. In *ICCV*, 2023. 2

[7] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 2

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *ICLR*, 2023. 1, 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 2

[10] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[11] Elad Hoffer and Nir Ailon. Deep Metric Learning using Triplet Network. In *SIMBAD*, 2015. 4

[12] 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 *CVPR*, 2023. 1

[13] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided Diffusion Models for Robust Image Manipulation. In *CVPR*, 2022. 1

[14] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 2

[15] Hyunsoo Lee, Minsoo Kang, and Bohyung Han. Conditional Score Guidance for Text-Driven Image-to-Image Translation. In *NeurIPS*, 2023. 2

[16] Junsung Lee, Minsoo Kang, and Bohyung Han. Diffusion-based image-to-image translation by noise correction via prompt interpolation. In *ECCV*, 2025. 2

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022. 4

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images Using Guided Diffusion Models. In *CVPR*, 2023. 1, 2, 4, 5, 6

[19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-Shot Image-to-Image Translation. In *SIGGRAPH*, 2023. 1, 2, 4, 5, 6

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 1, 2, 5

[21] 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

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8

[23] 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. In *NeurIPS*, 2022. 1, 2

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *NeurIPS Datasets and Benchmarks Track*, 2022. 4, 5, 6, 7, 8

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015. 2

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 2

[27] Charles M Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The annals of Statistics*, 1981. 3

[28] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT Features for Semantic Appearance Transfer. In *CVPR*, 2022. 5

[29] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, 2023. 1, 2, 4, 5, 6

[30] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-Driven Image Editing by Fine-tuning an Image Generation Model on a Single Image. In *SIGGRAPH*, 2023. 1

[31] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*, 2022. 5

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *CVPR*, 2017. 6