Making Text-to-Image Diffusion Models Zero-Shot Image-to-Image Editors by Inferring "Random Seeds"

Chen Henry Wu, Fernando De la Torre
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
{chenwu2,ftorre}@cs.cmu.edu

Abstract

Recent text-to-image diffusion models trained on large-scale data achieve remarkable performance on text-conditioned image synthesis (e.g., GLIDE, DALL·E-2, Imagen, Stable Diffusion). This paper introduces a simple method to use stochastic text-to-image diffusion models as zero-shot image editors. Our method, **CycleDiffusion**, is based on the finding that when all random variables (or "random seed") are fixed, two similar text prompts will produce similar images. The core of our idea is to infer the random variables that are likely to generate a source image conditioned on a source text. With the inferred random variables, the text-to-image diffusion model then generates a target image conditioned a target text. Our experiments show that CycleDiffusion outperforms SDEdit and the ODE-based DDIB method, and it can be further improved by Cross Attention Control.[1]

1 Introduction

It has been observed that given a fixed random seed, a text-to-image diffusion model will generate similar images when conditioned on two similar text prompts. In this paper, we show how to exploit this finding to make stochastic text-to-image diffusion models zero-shot image-to-image editors, without any assumptions on the model architecture.

We first formalize "random seeds" by defining the Gaussian latent codes $\boldsymbol{z}$ of *stochastic* diffusion models (as opposed to the ODE-based deterministic ones) as the concatenation of all Gaussian noises in the denoising process. We then propose DPM-Encoder to infer $\boldsymbol{z}$ from a real image $\boldsymbol{x}$. Finally, we propose CycleDiffusion, a method for zero-shot image editing. Given a pretrained text-to-image diffusion model, CycleDiffusion requires no finetuning to achieve this task. It first encodes the source image $\boldsymbol{x}$ as $\boldsymbol{z}$ using DPM-Encoder, conditioned on the source text $\boldsymbol{t}$; then it decodes $\boldsymbol{z}$ into the target image $\hat{\boldsymbol{x}}$ with the text-to-image diffusion model conditioned on the target text $\hat{\boldsymbol{t}}$.

Our experiments compared CycleDiffusion with SDEdit [6] and the ODE-based DDIB [16]. We simulated hyperparameter search and random trials *for each sample*, which is quite common for text-to-image diffusion models, and used the directional CLIP score as the criterion. Results show that CycleDiffusion outperforms SDEdit and DDIB. We also demonstrate that CycleDiffusion can be combined with the Cross Attention Control [2] to further preserve the image structure.

2 Method

2.1 Formalizing "Random Seeds" of Stochastic Diffusion Models

We first formalize the "random seeds" of stochastic diffusion probabilistic models (DPMs), such as DDPMs [3], non-deterministic DDIMs [14], and score-based SDEs [15]. These models generate

[1] https://huggingface.co/spaces/ChenWu98/Stable-CycleDiffusion

Score-based Method Workshop at Conference on Neural Information Processing Systems (NeurIPS 2022).

Algorithm 1: CycleDiffusion for zero-shot image editing

Input: source image $\boldsymbol{x} := \boldsymbol{x}_0$; source text $\boldsymbol{t}$; target text $\hat{\boldsymbol{t}}$; encoding step $T_{\text{es}} \leq T$
1. Sample noisy image $\hat{\boldsymbol{x}}_{T_{\text{es}}} = \boldsymbol{x}_{T_{\text{es}}} \sim q(\boldsymbol{x}_{T_{\text{es}}}|\boldsymbol{x}_0)$
for $t = T_{es}, \ldots, 1$ **do**
  2. $\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$
  3. $\boldsymbol{\epsilon}_t = \big(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_T(\boldsymbol{x}_t, t|\boldsymbol{t})\big)/\boldsymbol{\sigma}_t$
  4. $\hat{\boldsymbol{x}}_{t-1} = \boldsymbol{\mu}_T(\hat{\boldsymbol{x}}_t, t|\hat{\boldsymbol{t}}) + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}_t$
Output: $\hat{\boldsymbol{x}} := \hat{\boldsymbol{x}}_0$

images with a Markov chain structure. Given $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the image $\boldsymbol{x} := \boldsymbol{x}_0$ is generated through $\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_T(\boldsymbol{x}_t, t), \text{diag}(\boldsymbol{\sigma}_t^2))$. We can define the latent code $\boldsymbol{z}$ and the mapping $\boldsymbol{x} = G(\boldsymbol{z})$ as

$$\boldsymbol{z} := \big(\boldsymbol{x}_T \oplus \boldsymbol{\epsilon}_T \oplus \cdots \oplus \boldsymbol{\epsilon}_1\big) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{x}_{t-1} = \boldsymbol{\mu}_T(\boldsymbol{x}_t, t) + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}_t, \quad t = T, \ldots, 1, \tag{1}$$

where $\oplus$ is concatenation. Here, $\boldsymbol{z}$ has dimension $d = d_I \times (T+1)$, where d_I is the image dimension.

2.2 DPM-Encoder: an Encoder for Diffusion Models

To edit real images, we propose DPM-Encoder to infer $\boldsymbol{z}$ from the image $\boldsymbol{x}$. For each image $\boldsymbol{x} := \boldsymbol{x}_0$, a stochastic DPM have a predefined posterior $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ [3, 14]. Based on $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ and Eq. (1), we can directly derive $\boldsymbol{z} := \big(\boldsymbol{x}_T \oplus \boldsymbol{\epsilon}_T \oplus \cdots \oplus \boldsymbol{\epsilon}_2 \oplus \boldsymbol{\epsilon}_1\big) \sim \text{DPMEnc}(\boldsymbol{z}|\boldsymbol{x}, G)$ as

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T-1}, \boldsymbol{x}_T \sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0), \quad \boldsymbol{\epsilon}_t = \big(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_T(\boldsymbol{x}_t, t)\big)/\boldsymbol{\sigma}_t, \quad t = T, \ldots, 1. \tag{2}$$

2.3 CycleDiffusion: Image-to-Image Translation with DPM-Encoder

Since DPM-Encoder can infer $\boldsymbol{z}$ from a given real image $\boldsymbol{x}$, we use it to build our zero-shot image-to-image translation method CycleDiffusion. Let $G_{\boldsymbol{t}}$ be a text-to-image diffusion model conditioned on text $\boldsymbol{t}$. The task input is a source image $\boldsymbol{x}$, a source text $\boldsymbol{t}$ describing the source image $\boldsymbol{x}$, and a target text $\hat{\boldsymbol{t}}$ describing the target image $\hat{\boldsymbol{x}}$ to be generated. Like the GAN-based UNIT method [5], CycleDiffusion encodes $\boldsymbol{x}$ as $\boldsymbol{z}$ with DPM-Encoder and decodes it as $\hat{\boldsymbol{x}} = G_{\hat{\boldsymbol{t}}}(\boldsymbol{z})$. Formally, we have

$$\boldsymbol{z} \sim \text{DPMEnc}(\boldsymbol{z}|\boldsymbol{x}, G_{\boldsymbol{t}}), \quad \hat{\boldsymbol{x}} = G_{\hat{\boldsymbol{t}}}(\boldsymbol{z}). \tag{3}$$

Inspired by the realism-faithfulness tradeoff in SDEdit [6], we can truncate $\boldsymbol{z}$ towards a specified encoding step $T_{\text{es}} \leq T$. The full algorithm of CycleDiffusion with truncation is shown in Algorithm 1.

An analysis for image similarity with fixed $\boldsymbol{z}$. This part analyzes how the fixed $\boldsymbol{z}$ helps bound image distances. Suppose the text-to-image model has the following two properties:

1. Conditioned on the same text, similar noisy images lead to similar enough mean predictions. Formally, $\boldsymbol{\mu}_T(\boldsymbol{x}_t, t|\boldsymbol{t})$ is K_t-Lipschitz, i.e., $\|\boldsymbol{\mu}_T(\boldsymbol{x}_t, t|\boldsymbol{t}) - \boldsymbol{\mu}_T(\hat{\boldsymbol{x}}_t, t|\boldsymbol{t})\| \leq K_t\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|$.
2. Given the same image, the two texts lead to similar predictions. Formally, $\|\boldsymbol{\mu}_T(\hat{\boldsymbol{x}}_t, t|\boldsymbol{t}) - \boldsymbol{\mu}_T(\hat{\boldsymbol{x}}_t, t|\hat{\boldsymbol{t}})\| \leq S_t$. Intuitively, a smaller difference between $\boldsymbol{t}$ and $\hat{\boldsymbol{t}}$ gives us a smaller S_t.

Let B_t be the upper bound of $\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|_2$ at time step t when the same latent code $\boldsymbol{z}$ is used for sampling (i.e., $\boldsymbol{x}_0 = G_{\boldsymbol{t}}(\boldsymbol{z})$ and $\hat{\boldsymbol{x}}_0 = G_{\hat{\boldsymbol{t}}}(\boldsymbol{z})$). We have $B_T = 0$ because $\|\boldsymbol{x}_T - \hat{\boldsymbol{x}}_T\|_2 = 0$, and B_0 is the upper bound for the generated images $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2$. The upper bound B_t can be propagated through time, from T to 0. Specifically, by combining the above two properties, we have

$$B_{t-1} \leq (K_t + 1)B_t + S_t. \tag{4}$$

3 Experiments

This section provides experiments for zero-shot image-to-image translation. Following Section 2.3, we curated a set of 150 triplets $(\boldsymbol{x}, \boldsymbol{t}, \hat{\boldsymbol{t}})$ for this task, where $\boldsymbol{x}$ is the source image, $\boldsymbol{t}$ is the source text (e.g., "an aerial view of autumn scene." in Figure 1), and $\hat{\boldsymbol{t}}$ is the target text (e.g., "an aerial view of winter scene."). The target image to be generated is denoted as $\hat{\boldsymbol{x}}$.

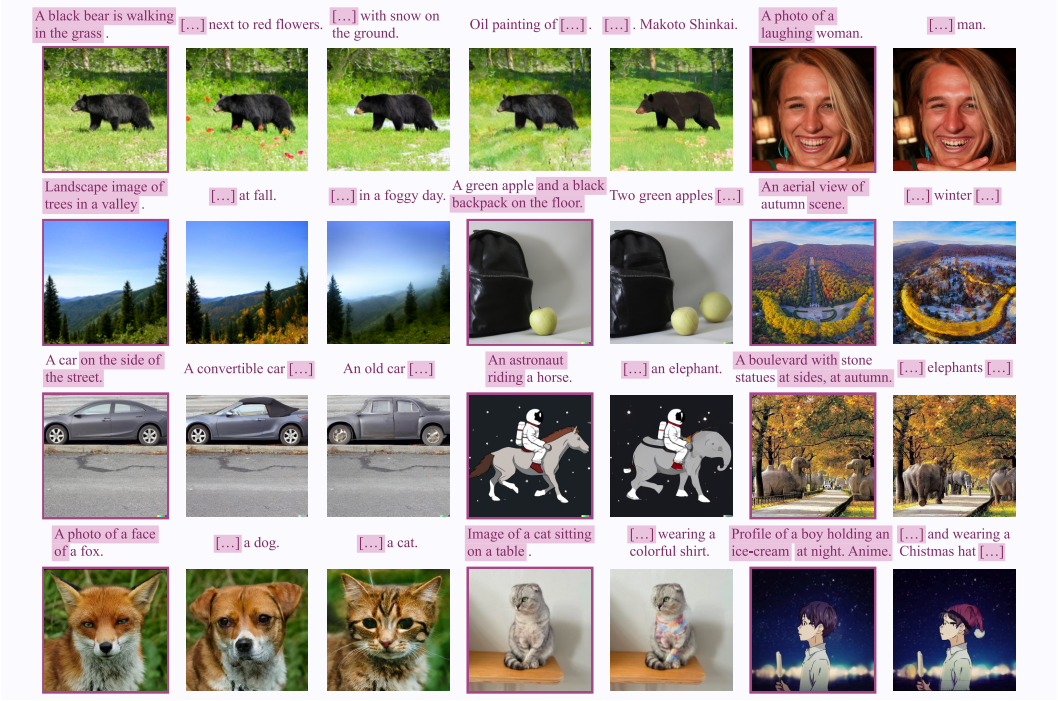

Figure 1: Text-to-image diffusion models can be zero-shot image-to-image editors. Source images $\boldsymbol{x}$ are displayed with a purple margin; the others are generated target images $\hat{\boldsymbol{x}}$. Overlapping text spans are marked in purple in source texts $\boldsymbol{t}$ and abbreviated as $[\ldots]$ in target texts $\hat{\boldsymbol{t}}$.

Metrics: To evaluate the faithfulness to source images, we reported PSNR and SSIM. To evaluate the authenticity to the target text, we used the CLIP score, i.e., the cosine similarity between CLIP embeddings: $\mathcal{S}_{\text{CLIP}}(\hat{\boldsymbol{x}}|\hat{\boldsymbol{t}}) = \cos\left\langle \text{CLIP}_{\text{img}}(\hat{\boldsymbol{x}}), \text{CLIP}_{\text{text}}(\hat{\boldsymbol{t}}) \right\rangle$. We also reported directional CLIP score [8], i.e., the cosine similarity between the differences of CLIP embeddings:

$$\mathcal{S}_{\text{D-CLIP}}(\hat{\boldsymbol{x}}|\boldsymbol{x}, \boldsymbol{t}, \hat{\boldsymbol{t}}) = \cos\left\langle \text{CLIP}_{\text{img}}(\hat{\boldsymbol{x}}) - \text{CLIP}_{\text{img}}(\boldsymbol{x}), \text{CLIP}_{\text{text}}(\hat{\boldsymbol{t}}) - \text{CLIP}_{\text{text}}(\boldsymbol{t}) \right\rangle. \tag{5}$$

Baselines: The baselines include SDEdit [6] and DDIB [16]. We used the same hyperparameters (and hyperparameter trials) for the baselines and CycleDiffusion whenever possible (see Appendix A).

Pre-trained text-to-image diffusion models: We experimented with two models: (1) `LDM-400M`, a 1.45B-parameter model trained on LAION-400M [13], (2) `SD-v1-4`, a 0.98B-parameter Stable Diffusion model trained on LAION-5B [12].

Table 1: Quantitative evaluation for zero-shot image editing. We did not use a fixed combination of hyperparameters, and neither did we plot the trade-off curve. The reason is that every input can have its best combination of hyperparameters and even random seed. Instead, **for each input**, we ran 15 trials for each combination of hyperparameters and report the one with the highest $\mathcal{S}_{\text{D-CLIP}}$. For a fair comparison, different methods share the same set of combinations of hyperparameters if possible, detailed in Appendix A.

	Method	$\mathcal{S}_{\text{CLIP}}\uparrow$	$\mathcal{S}_{\text{D-CLIP}}\uparrow$	PSNR↑	SSIM↑
`LDM-400M`	SDEdit [6]	0.332	0.264	13.68	0.390
	DDIB [16]	0.324	0.195	15.82	0.544
	CycleDiffusion w/ DDIM ($\eta = 0.1$; ours)	**0.333**	**0.275**	**18.72**	**0.625**
`SD-v1-4`	SDEdit [6]	**0.344**	0.258	15.93	0.512
	DDIB [16]	0.331	0.209	18.10	0.653
	CycleDiffusion w/ DDIM ($\eta = 0.1$; ours)	0.334	**0.272**	**21.92**	**0.731**

Results: Table 1 shows the results for zero-shot image-to-image translation. CycleDiffusion excels at being faithful to the source image (i.e., PSNR and SSIM); by contrast, SDEdit and DDIB have comparable authenticity to the target text (i.e., $\mathcal{S}_{\text{CLIP}}$), but their outputs are much less faithful. For all methods, we find that the pre-trained weights `SD-v1-1` and `SD-v1-4` have better faithfulness than `LDM-400M`. Figure 1 provides samples from CycleDiffusion, demonstrating that CycleDiffusion achieves meaningful edits that span (1) replacing objects, (2) adding objects, (3) changing styles, and (4) modifying attributes. See Figure 3 (Appendix B) for qualitative comparisons with the baselines.

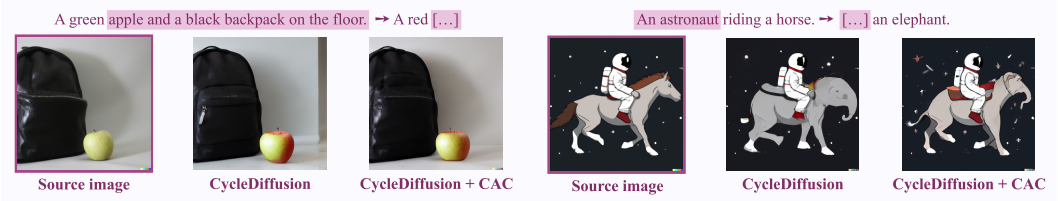

Figure 2: Cross Attention Control (CAC) [2] helps CycleDiffusion when the intended *structural* change is small. For instance, when the change is color (left), CAC helps CycleDiffusion preserve the background; when the change is horse → elephant (right), CAC makes the elephant look like a horse.

CycleDiffusion + Cross Attention Control: Besides fixing the random seed, [2] shows that fixing the cross attention map (this operation is called Cross Attention Control, or CAC) also improves the similarity between synthesized images. CAC is applicable to CycleDiffusion: in Algorithm 1, we can apply the attention map of $\boldsymbol{\mu}_T(\boldsymbol{x}_t, t|\boldsymbol{t})$ to $\boldsymbol{\mu}_T(\hat{\boldsymbol{x}}_t, t|\hat{\boldsymbol{t}})$. However, we cannot apply it to all samples because CAC puts requirements on $\boldsymbol{t}$ and $\hat{\boldsymbol{t}}$ (i.e., the target text is a subsequence of the source one, or the two texts differ in only one token [2]). Figure 2 shows that CAC helps CycleDiffusion when the intended *structural* change is small. For instance, when the intended change is color but not shape (left), CAC helps CycleDiffusion preserve the background; when the intended change is horse → elephant, CAC makes the generated elephant to look more like a horse in shape.

4 Related Work

Several recent methods for text-to-image synthesis are built upon diffusion probabilistic models (DPMs), such as GLIDE [7], DALL·E 2 [9], Imagen [11], Stable Diffusion [10]. It has been observed that when using the same random seed, text-to-image DPMs tend to generate similar images given two similar text prompts. It holds for both stochastic DPMs [3, 14] and deterministic DPMs [14, 15]. Based on the finding for deterministic DPMs, [16] proposed dual diffusion implicit bridge (DDIB) for unpaired image translation, and our CycleDiffusion is an extension of it to stochastic DPMs.

Besides random seeds, a concurrent work [2] found that fixing the cross-attention map in Transformer-based text-to-image DPMs further improves the similarity between images. To fix the cross-attention map for two text prompts, they proposed Cross Attention Control (CAC). To edit real images, they combined CAC with DDIB [16] or mask heuristics because "how to infer random seeds" for stochastic DPMs is non-trivial. In our experiments, we show that CAC can be applied to CycleDiffusion to improve the structural preservation of the image.

Two concurrent works, Imagic [4] and UniTune [17], showed that zero-shot image editing can also be achieved by finetuning the text-to-image DPM on the source image to be edited. Different from these two works, our CycleDiffusion is an optimization-free method that does not need finetuning the large text-to-image DPMs. Similar optimization-free methods include the Cross Attention Control [2] discussed above and DiffEdit [1], which automatically infers a mask before editing.

5 Conclusions

This paper proposes CycleDiffusion, a simple method for zero-shot image editing with pretrained text-to-image DPMs without optimization. Different from several works that use deterministic DPMs for this task, CycleDiffusion shows that stochastic DPMs can also be used for zero-shot image editing. Besides the promising results, we also analyze how the fixed $\boldsymbol{z}$ helps bound image distances. However, an analysis on needs further exploration.

References

[1] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, 2022.

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

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[4] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, 2022.

[5] Ming-Yu Liu, Thomas M. Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[6] 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. *ICLR*, 2022.

[7] 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. *ArXiv*, 2021.

[8] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and D. Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. *ICCV*, 2021.

[9] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *ArXiv*, 2022.

[10] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

[11] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

[12] Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, mehdi cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Richard Vencu, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS Datasets and Benchmarks*, 2022.

[13] 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*, 2021.

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021.

[15] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

[16] Xu Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *ArXiv*, 2022.

[17] Dani Valevski, Matan Kalman, Y. Matias, and Yaniv Leviathan. UniTune: Text-driven image editing by fine tuning an image generation model on a single image. *ArXiv*, 2022.

A Experimental Details

CycleDiffusion and the baselines for zero-shot image-to-image translation have some shared hyperparameters; each method also has its own unique hyperparameters. For both CycleDiffusion and the baselines, we can enumerate some of these parameters and select the best per sample based on a certain criterion. Moreover, for stochastic methods such as SDEdit and our CycleDiffusion, one may run several trials. In this section, we provide these details.

Per sample selection criterion: For each test sample, we allow each method to enumerate some combinations of hyperparameters (detailed below). To select the best combination for each sample, we used the directional CLIP score $\mathcal{S}_{\text{D-CLIP}}$ as the criterion (higher is better).

DDIB: Since DDIB only applies to deterministic DPMs, we used the deterministic DDIM sampler with 100 steps. We set the classifier-free guidance of the encoding step as 1; we enumerated the classifier-free guidance of the decoding step as $\{1, 1.5, 2, 3, 4, 5\}$.

SDEdit: For SDEdit, we used the DDIM sampler ($\eta = 0.1$) with 100 steps. We enumerated the classifier-free guidance of the decoding step as $\{1, 1.5, 2, 3, 4, 5\}$; we enumerated the SDEdit step as $\{15, 20, 25, 30, 40, 50\}$; we ran 15 trials for each hyperparameter combination.

CycleDiffusion: For our CycleDiffusion, we used the DDIM sampler ($\eta = 0.1$) with 100 steps. We set the classifier-free guidance of the encoding step as 1; we enumerated the classifier-free guidance of the decoding step as $\{1, 1.5, 2, 3, 4, 5\}$; we enumerated the early stopping step T_{es} as $\{15, 20, 25, 30, 40, 50\}$; we ran 15 trials for each hyperparameter combination.

B Additional Results for Zero-Shot Image-to-Image Translation

Figure 3 provides a qualitative comparison for zero-shot image-to-image translation. Compared with SDEdit and the ODE-based DDIB, CycleDiffusion improves the faithfulness to the source image.

Figure 3: Samples for zero-shot image-to-image translation. Notations follow Figure 1. Compared with DDIB and SDEdit, CycleDiffusion greatly improves the faithfulness to the source image.