MirrorDiffusion: Stabilizing Diffusion Process in Zero-shot Image Translation by Prompts Redescription and Beyond

Yupei Lin, Xiaoyu Xian, Yukai Shi[†], and Liang Lin, Fellow, IEEE



Fig. 1: Without any supervision, MirrorDiffusion realized three zero-shot image-to-image translations: w/o glasses \rightarrow w/ glasses, Male \rightarrow Female and Fox \rightarrow Dog.

Abstract-Recently, text-to-image diffusion models become a new paradigm in image processing fields, including content generation, image restoration and image-to-image translation. Given a target prompt, Denoising Diffusion Probabilistic Models (DDPM) are able to generate realistic vet eligible images. With this appealing property, the image translation task has the potential to be free from target image samples for supervision. By using a target text prompt for domain adaption, the diffusion model is able to implement zero-shot image-to-image translation advantageously. However, the sampling and inversion processes of DDPM are stochastic, and thus the inversion process often fail to reconstruct the input content. Specifically, the displacement effect will gradually accumulated during the diffusion and inversion processes, which led to the reconstructed results deviating from the source domain. To make reconstruction explicit, we propose a prompt redescription strategy to realize a mirror effect between the source and reconstructed image in the diffusion model (MirrorDiffusion). More specifically, a prompt redescription mechanism is investigated to align the text prompts with latent code at each time step of the Denoising Diffusion Implicit Models (DDIM) inversion to pursue a structure-preserving reconstruction. With the revised DDIM inversion, MirrorDiffusion is able to realize accurate zero-shot image translation by editing optimized text prompts and latent code. Extensive experiments demonstrate that MirrorDiffusion achieves superior performance over the stateof-the-art methods on zero-shot image translation benchmarks

by clear margins and practical model stability. Our project is available at https://mirrordiffusion.github.io/

1

Index Terms—Diffusion Process, Generative Model, Image-to-Image Translation, Zero-Shot.

I. INTRODUCTION

R ECENTLY, text-to-image diffusion model [1] becomes a new fashion in signal processing fields. With large-scale pre-training on text-to-image data pairs, Denoising Diffusion Probabilistic Models (DDPM) have a strong capacity to generate diverse image content [2]–[4]. Nevertheless, DDPM has achieved great success in image generation task, it still fail to achieve a desired performance on image-to-image translation, especially on zero-shot image-to-image translation.

Image translation [5], [6] aims to transform images from a source domain to a target domain, such as cat \rightarrow dog. This task inherently requires the target domain images for model adaption, which used to be accomplished by Generative Adversarial Nets (GAN) [7]–[12]. However, its difficult for traditional GANs to fully understand the target domain knowledge with limited number of samples, which often lead to the



Fig. 2: To show displacement effect, the reconstruction process of typical DDIM work [15] is visualized in Fig.2 (a), which can be formulated as: $z_0 \rightarrow z_T \rightarrow z'_0$. However, errors accumulate in typical diffusion methods, causing biases in latent codes $[z_0, z'_0]$ and deviations in $[I_{source}, I_{reco}]$. To align the latent codes, we propose a prompt redescription mechanism to realize a mirror effect between the source and reconstructed image in the diffusion model (MirrorDiffusion).

poor translation quality. DDPM resolves this question by using a large-scale pre-training with text-to-image data [2]–[4] and integrating multimodal information like large-scale language models [13], [14].

Given a target prompt, the diffusion model has the ability to generate realistic yet eligible content. This attractive property has considerable potential for realizing the image translation task with fewer or zero target samples.

To address zero-shot image translation, Pix2Pix-Zero [15] first adopts the diffusion pipeline. Specifically, Pix2Pix-Zero measures the distance between source and target domain sentences by applying a CLIP [16] model. By utilizing that domain gap between text embedding, Pix2Pix-Zero successfully achieves zero-shot image-to-image translation with a pre-trained Denoising Diffusion Implicit Models (DDIM) [1]. Literally, typical DDIM usually converts the image into latent code, and then performs a latent code re-sampling w.r.t target domain prompt to demonstrate zero-shot image translation. This pipeline is able to generate images of the target domain without target images. As shown in Fig. 2, the generated results often deviate from the structure of the source domain, which violates the structure consistency. To keep the structure of the generated results consistent with the original content, many efforts [17], [18] are devoted to investigating the reconstruction pipeline in DDIM. Prompt2Prompt [17] first reconstructs the image structure at the early sampling stage with the original prompt, and then applies a word swap mechanism to change the prompt for image detail generation. Instruct Pix2Pix [19] proposes a triple dataset, which contains caption, edit instruction and edited instruction for image editing. To keep the structural consistency, Instruct Pix2Pix utilizes a cross-attention mechanism to ensure the edited image becomes consistent with the original image. Null-text Inversion [18] optimizes the null-text embedding for classifier-free guidance to ensure that the reconstruction results consistent with the source image. This null-text optimization requires the high precision of provided text prompts, otherwise, the details of the generated results will be skewed. SDEdit [20] realizes fidelity and consistency in the image translation by using stochastic differential equation (SDE) for image encoding and decoding. Since the SDE is invertible, the inversion process from noise to image is naturally realized.

However, the forementioned methods gradually accumulate displacement during the diffusion and inversion processes, which makes the reconstructed results gradually deviate from the source image. As shown in Fig. 2, the reconstructed image appears a displacement effect, which further affects the accuracy of the translation result. To solve the displacement effect in image reconstruction, we propose a prompt redescription strategy to realize a mirror effect between the source and reconstructed image during the diffusion process (MirrorDiffusion). Specifically, we address the deviation problem of the reconstructed image by aligning the text prompts and latent codes at each time step of the reconstruction process. With the revised DDIM inversion, MirrorDiffusion obtains accurate target text embedding and latent code for zero-shot image translation. Our contributions can be summarized as:

- A prompt redescription mechanism is proposed to address the displacement problem of image reconstruction in DDIM Inversion. With the prompt redescription, we achieve a reliable yet effective image reconstruction.
- Based on the revised DDIM inversion, we align the latent code with the text prompt during the diffusion process to further ensure consistency in zero-shot image translation.
- Extensive experiments demonstrate that MirrorDiffusion achieves superior performance over the state-of-the-art diffusion models on zero-shot image translation benchmarks by clear margins and practical model stability.

II. METHODOLOGY

A. Displacement in Diffusion Inversion Process

In Denoising Diffusion Implicit Models (DDIM) and its derivatives [1], [15], the sampling and inversion processes [1], [21] are stochastic, and thus the inversion process often fail to reconstruct the input as shown in Fig. 2 (a). Specifically, given a source domain image I_{source} , we first apply the encoder $Dec(\cdot)$ of the diffusion model [1] to convert it into a latent code z_0 . And then send z_0 into DDIM for diffusion, the object function is as follow:

$$L_{DDIM} = \min_{\theta} E_{t \sim U(1,T)} \left\| N_{gaus} - \epsilon_{\theta} \left(z_t, t, c \right) \right\|_2^2, \quad (1)$$



Fig. 3: Visualization results. Compared with state-of-the-art diffusion approaches across four tasks, our method excels in generating highly realistic translation results with **excellent structure consistency**.

 $N_{gaus} \sim \mathcal{N}(0, 1)$ is Gaussian noise with standard normal distribution, t is time step in DDIM with a range of [1, T], c represents the embedding of a conditional text prompt, ϵ_{θ} is the noise prediction network in DDIM, z_t represents the latent code in high dimensional space after t times of diffusion process. The physical meaning of Equ. 1 is to minimize the difference between the noise predicted by ϵ_{θ} and the real distribution standard Gaussian noise.

After DDIM optimization, as show in Fig. 2 (a), we can input the latent code z_T into the DDIM for stepwise sampling and image reconstruction:

$$z'_{T-1} = Sample\left(\epsilon_{\theta}, z_T, c_T, T\right),\tag{2}$$

where the $Sample(\cdot)$ represents the stepwise sampling of DDIM, ϵ_{θ} is the noise prediction network in DDIM.

As shown in Fig. 2 (a), with T steps sampling, we can convert z_T into z'_0 , and obtain the reconstructed image by: $I_{reco} = Dec(z'_0)$. However, in the inversion process, ϵ_{θ} may deviate from the standard Gaussian noise, and this displacement will gradually accumulate on $[Z_T, Z'_{T-1}, ..., Z'_0]$, resulting in an collapse in I_{reco} .

To show this displacement effect and collapse in I_{reco} , we visualize reconstruction process of some typical DDIM works [1], [15]. As shown in Fig. 2 (a), the diffusion inversion process realizes the reconstruction by: $z_0 \rightarrow z_T \rightarrow z'_0$. However, the distribution between $[z_0, z'_0]$ appears a displacement effect, resulting in a deviation between $[I_{source}, I_{reco}]$.

B. Prompts Redescription

To make the reconstruction explicit, we propose a prompts redescription strategy. Specifically, in the reconstruction phrase as shown in Fig. 2 (b), we calculate the difference between the latent codes during inversion and reconstruction as:

$$\mathcal{L}_{rewrite} = \min_{\theta} E_{t \sim \text{Uniform } (1,T)} \left\| z_{t-1} - z'_{t-1} \right\|_2^2, \quad (3)$$

where the z'_{t-1} indicates the sampling noise of the current time step, which is obtained by: $z'_{t-1} = Sample(\epsilon_{\theta}, z_t, c_{rewrite}, t)$. Then, we use the $\mathcal{L}_{rewrite}$ to implement a redescription toward current text prompt c_t as:

$$c_{rewrite} = c_t - \lambda \nabla_c \mathcal{L}_{rewrite},\tag{4}$$

where ∇_c takes the partial derivative on c and obtain gradient for prompt redescription. According the rewritten text prompt $c_{rewrite}$, we re-sample z_t at each time step in inversion process as: $z'_{t-1} = Sample(\epsilon_{\theta}, z_t, c_{rewrite}, t)$. As shown in Fig. 2 (b), to our surprise, the prompt redescription mechanism ensures that $[z_0, z'_0]$, $[I_{source}, I_{reco}]$ are firmly aligned after T steps sampling.

C. Zero-shot Image Translation with MirrorDiffusion

With the prompts redescription mechanism, we can obtain the aligned combination $[c_{rewrite}, z'_0]$ toward ϵ_{θ} . As shown in Fig. 3, our model can implement zero-shot image translation by further editing $c_{rewrite}$ according to the target domain.

As shown in Fig. 4, we apply CLIP [16] to compute the domain gap Δc between the source domain and target domain. Specifically, the CLIP [16] is used to extract the highlevel features of source domain sentences and target domain sentences, respectively. And the mean difference, which is computed along those features, is represented as the domain gap Δc .



Fig. 4: The framework overview of MirrorDiffusion. With the prompt redescription mechanism, our model obtains the firmly aligned $[z_0, z'_0]$, $[I_{source}, I_{reco}]$ combinations. We apply CLIP [16] to compute the domain gap Δc between the source domain and target domain for image editing. Specifically, the CLIP [16] is used to extract the high-level features of source domain sentences and target domain sentences, respectively. And the mean difference, which is computed along those features, is represented as the domain gap Δc . Then, we apply the target text embedding $c_{rewrite} + \Delta c$ for zero-shot image translation with diffusion inversion process. With T-time inversion, MirrorDiffusion can obtain the corresponding latent code z'_0 , which corresponds to I_{trans} with $Dec(\cdot)$.

As shown in Fig. 4, we then use the updated target text embedding $c_{rewrite} + \Delta c$ for a latent code sampling as:

$$z'_{t-1} = Sample\left(\epsilon_{\theta}, z'_{t}, c_{rewrite} + \Delta c, t\right), \qquad (5)$$

where Δc represents the target domain direction, such as $Dog \rightarrow Cat$. And $c_{rewrite}$ is the source domain text embedding, which was firmly aligned by the prompt redescription mechanism. After *T*-time sampling, our model can obtain the corresponding latent code z'_0 . As shown in Fig. 4, we can easily obtain translated image with a renewed z'_0 as:

$$I_{trans} = Dec(z'_0), \tag{6}$$

where I_{trans} represents the translated image, $Dec(\cdot)$ is an image decoder [22].

III. EXPERIMENT

A. Datasets and Metric

To evaluate the gap between our model and baselines, we selected three sub-datasets from the LAION-5B dataset [23] containing cat, horse and sketch images, each containing 250 images. Subsequently, we formulated the following image transformation tasks, including, Translate Cat to Dog(C2D-F), Add Glasses to Cat(C2G-F), Translate Sketch to Oil Pastel(S2O-F), Translate Horse to Zebra(H2Z-F).

We conducted a comparative analysis of performance differences between our approach and several baselines, including: SDEdit [20], DDIM [1], InstructPix2pix [19], Pix2Pix-Zero [15] and NULL-Text inversion [18] methods, across these datasets.For quantitative quality evaluation, we used CLIP-ACC [24] ,Structure Dist [25] , Structure Similarity Index Measure [26] (SSIM) and Learned Perceptual Image Patch Similarity [27] (LPIPS), which are evaluated in terms of whether the translation is successful.

B. Implementation details

During the image editing process, we optimize only the variables in the text rewrite process and the pre-trained stable diffusion model remains in a frozen state. In this experiment, we established the weight parameter λ for $\mathcal{L}_{rewtire}$ as 1,

employing the Adam optimizer to update the value of $c_{rewrite}$, with a learning rate of 0.0001. In this paper, the number of iterations for the DDIM inversion, as well as for the DDIM editing sampling and reconstruction sampling are both set to 60. In the image editing process, we use classifier-free guidance [28] to predict the noise at each time step. All inputs and generated results are with a size of $512 \times 512 \times 3$.

C. Comparison

As shown in Fig. 3, we compare the visual appearance of our approach with the baseline models on the four tasks. It can be seen that SDEdit, DDIM and Instruct Pix2Pix struggle to maintain appearance consistency. Pix2Pix-Zero and Null-Text Inversion perform well in terms of general appearance, but fall short in terms of preserving details and translation accuracy. In contrast, our method performs well in both appearance preservation and translation correctness. Our quantitative evaluation results also achieved the best results. In Tab. I, the pink cells represent the best results and the orange cells represent the second-best results. It can be clearly seen that we achieved the best results in both translation quality and structure preservation. To demonstrate the effectiveness of our approach on diverse tasks, we supplement three additional tasks with different scenarios: w/o glasses \rightarrow w/ glasses, Male \rightarrow Female and Fox \rightarrow Dog. The image data for these tasks are sourced from CelebA-HQ [29] and AFHQ [30]. As shown in Fig. 1, our method effectively achieves high-quality image translation in these tasks.

D. Ablation Study

Effects of $\mathcal{L}_{rewrite}$ during reconstruction. As shown in Fig. 5, we show the attention maps during the reconstruction process of MirrorDiffusion. To verify the effectiveness of our method, we show the attention maps of 'w/ $\mathcal{L}_{rewrite}$ ' and 'w/o $\mathcal{L}_{rewrite}$ '. As shown in Fig. 5, without $\mathcal{L}_{rewrite}$, the attention map of the cat gradually deviates from itself, leading to a poor reconstruction result. With the proposed $\mathcal{L}_{rewrite}$, the cat's attention maps center on critical regions and complete a faithful reconstruction.



Fig. 5: Attention maps of 'w/o $\mathcal{L}_{rewrite}$ ' and 'w/ $\mathcal{L}_{rewrite}$ ' during reconstruction process.

TABLE I: Comparison of quantitative results. We evaluate the results in terms of the four evaluation metrics CLIP-ACC, Structure Dist, SSIM, and LPIPS, with the pink cells indicating the results of the best performance and the orange cells indicating the results of the second best performance

		C21	D-F			H2.	Z-F			C20	j-F			\$20)-F	
METHOD	Clip↑	Structure↓	SSIM↑	LPIPS↓	Clip↑	Structure↓	SSIM↑	LPIPS↓	Clip↑	Structure↓	SSIM↑	LPIPS↓	Clip↑	Structure↓	SSIM↑	LPIPS↓
SDEdit [20]	66.9	0.146	0.441	0.552	78.7	0.223	0.441	0.573	76.9	0.133	0.428	0.566	56.1	0.133	0.502	0.519
DDIM [1]	60.2	0.127	0.637	0.439	72.5	0.159	0.6441	0.491	68.8	0.114	0.589	0.435	53.5	0.122	0.633	0.441
Instruct Pix2Pix [19]	72.5	0.086	0.699	0.275	76.4	0.256	0.455	0.711	74.4	0.155	0.342	0.633	66/3	0.130	0.649	0.485
Pix2Pix-Zero [15]	75.2	0.071	0.718	0.272	78.3	0.106	0.671	0.385	80.3	0.047	0.653	0.246	70.7	0.059	0.741	0.240
NULL-Text-Inversion [18]	74.5	0.081	0.72	0.25	81.6	0.1369	0.625	0.36	83.2	0.054	0.743	0.225	69.7	0.066	0.743	0.248
Ours	77.0	0.04	0.782	0.15	82.1	0.0758	0.722	0.271	83.9	0.024	0.797	0.13	70.7	0.043	0.768	0.176



Fig. 6: Ablation study of $\mathcal{L}_{rewrite}$. It can be observed that $\mathcal{L}_{rewrite}$ plays a significant role in preserving the appearance. TABLE II: Ablation on $L_{rewrite}$. The results show that the structure can be well preserved with $L_{rewrite}$.

	C2G-F								
	Clip↑	structure↓	SSIM \uparrow	LPIPS \downarrow					
w/o $\mathcal{L}_{rewrite}$	82.6	0.055	0.729	0.255					
w/ $\mathcal{L}_{rewrite}$	83.9	0.024	0.797	0.132					

Effects of $\mathcal{L}_{rewrite}$ during editing. To show the effect of $\mathcal{L}_{rewrite}$ during editing, we present a comparison between the results obtained using 'w/ $\mathcal{L}_{rewrite}$ ' and 'w/o $\mathcal{L}_{rewrite}$ '. As depicted in Fig. 6, the differences between these two approaches are clearly visible. Without the use of $\mathcal{L}_{rewrite}$, the edited results are difficult to effectively preserve the original appearance. Conversely, when $\mathcal{L}_{rewrite}$ is employed, the results show improved preservation of the original appearance while accomplishing the desired image translation. **Quantitative results analysis.** Furthermore, we also present some quantitative metrics in the C2GF task. As shown in Tab. II, it can be observed that our image translation results demonstrate significant improvements across these four metrics, which highlights the effectiveness of the prompt redescription mechanism.



Fig. 7: Comparison of the generation results of 'w/ Simple Alignment' and 'w/ $\mathcal{L}_{rewrite}$ '.

Effects of prompt rewrite module. To verify the effectiveness of the alignment and rewrite modules, we have made an ablation study. Suppose we apply an independent alignment module instead of using the rewrite module, we show the generated results in Fig. 7. During the editing stage, there exists a gap between the prompt and expected prompt at each time step. A simple alignment strategy fails to compensate this gap among prompts, leading to a poor quality. Although an independent alignment module aligns the overall structure with the original image, it fails to preserve image details. As shown in Fig. 7, without rewrite module, the orientation of boy's face and dog's head exhibit an unreasonable rotation.



Fig. 8: The generated results of Pix2Pix-Zero 'w/o $\mathcal{L}_{rewrite}$ ' and 'w/ $\mathcal{L}_{rewrite}$ '.

Instead, our method aligns the prompt with expected prompt at each time step by rewrite module, achieving a successfully image translation that preserves the structural consistency.

Further investigations on the rewrite module. We apply the prompt rewrite module to DDIM [1] and Pix2Pix-Zero [15] for further investigation. As shown in Fig. 8, the structure of 'Pix2Pix-Zero w/ $\mathcal{L}_{rewrite}$ ' is consistent to the original image. Additionally, we show the results of DDIM with the proposed rewrite module. As shown in Fig. 9, 'DDIM w/ $\mathcal{L}_{rewrite}$ ' is more faithful to the structure of the original image.



Fig. 9: The generated results of DDIM 'w/o $\mathcal{L}_{rewrite}$ ' and 'w/ $\mathcal{L}_{rewrite}$ '.

IV. CONCLUSION

In this paper, we aims to address the deviation and displacement problems of current text-to-image diffusion models in image translation tasks. By introducing a new prompt redescription mechanism, our method surpasses state-of-theart diffusion-based image translation methods on both visual results and quantitative results.

REFERENCES

- J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.
- [2] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.
- [3] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subjectdriven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

- [4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint* arXiv:2204.06125, 2022.
- [5] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1161–1165, 2018.
- [6] H. Chen, L. Dong, H. Yang, X. He, and C. Zhu, "Unsupervised realworld image super-resolution via dual synthetic-to-realistic and realisticto-synthetic translations," *IEEE Signal Processing Letters*, vol. 29, pp. 1282–1286, 2022.
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [8] P. Xiang, L. Wang, F. Wu, J. Cheng, and M. Zhou, "Single-image deraining with feature-supervised generative adversarial network," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 650–654, 2019.
- [9] L.-H. Chen, C. G. Bampis, Z. Li, and A. C. Bovik, "Learning to distort images using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 27, pp. 2144–2148, 2020.
- [10] S. Hong and J. Ryu, "Unsupervised face domain transfer for lowresolution face recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 156–160, 2019.
- [11] H. Yin and J. Xiao, "Laplacian pyramid generative adversarial network for infrared and visible image fusion," *IEEE Signal Processing Letters*, vol. 29, pp. 1988–1992, 2022.
- [12] P. Zhou, L. Xie, B. Ni, and Q. Tian, "Searching towards class-aware generators for conditional generative adversarial networks," *IEEE Signal Processing Letters*, vol. 29, pp. 1669–1673, 2022.
- [13] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14974–14983.
- [14] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, "Improving zeroshot visual question answering via large language models with reasoning question prompts," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4389–4400.
- [15] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," arXiv preprint arXiv:2302.03027, 2023.
- [16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [17] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," arXiv preprint arXiv:2208.01626, 2022.
- [18] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Nulltext inversion for editing real images using guided diffusion models," *arXiv preprint arXiv:2211.09794*, 2022.
- [19] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.
- [20] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Image synthesis and editing with stochastic differential equations," arXiv preprint arXiv:2108.01073, 2021.
- [21] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794, 2021.
- [22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.
- [23] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.
- [24] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint* arXiv:2104.08718, 2021.
- [25] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, "Splicing vit features for semantic appearance transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10748–10757.
- [26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

- [28] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.
- [29] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.
- [30] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8188–8197.