On the Robustness of Diffusion Inversion in Image Manipulation

Anonymous authors
Paper under double-blind review

Abstract

Text-guided image editing is a rapidly growing field due to the development of large diffusion models. In this work, we present an effective approach to address the key step of real image editing, known as "inversion", which involves finding the initial noise vector that reconstructs the input image when conditioned on a text prompt. Existing works on conditional inversion is often unstable and inaccurate, leading to distorted image manipulation. To address these challenges, our method starts by analyzing the inconsistent assumptions and accumulative errors that contribute to the ill-posedness of mathematical inverse problems. We then introduce learnable latent variables as bias correction to approximate invertible and bijective inversion. We perform latent trajectory optimization with a prior to fully invert the image by optimizing the bias correction on the unconditional text prompt and initial noise vector. Our method is based on the publicly Stable Diffusion model and is extensively evaluated on a variety of images and prompt editing, demonstrating high accuracy, robustness, and quality compared to state-of-the-art baseline approaches.

1 Introduction

The foundation of text-guided image editing is *inversion*, which requires running the reverse of the generative process. That is, finding the initial noise vector that produces the input image when passed through the diffusion process. Inversion has been studied considerably for generative adversarial networks (GANs), i.e., GAN inversion (Cheng et al., 2022; Xia et al., 2022; Zhu et al., 2020), but has not yet been completely resolved for text-guided diffusion models. A naive solution is proposed by adding Gaussian noise to the input image and then performing a predefined number of diffusion steps using denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), but it results in significant distortions (Hertz et al., 2022). Instead of DDPMs, denoising diffusion implicit models (DDIMs) are seen as a more effective way for inversion (Song et al., 2020a). DDIMs invent a particular parameterization of the diffusion process, that creates a smooth, deterministic, and reversible mapping between images and their latent representations (Dhariwal & Nichol, 2021; Preechakul et al., 2022; Kwon et al., 2022; Wu et al., 2022). Therefore, one is able to perform the diffusion process in the reverse direction, that is deterministically noising an image to obtain the initial noise $\mathbf{x}_0 \rightarrow \mathbf{x}_T$ instead of $\mathbf{x}_T \rightarrow \mathbf{x}_0$.

Although DDIM inversion produces promising results, it is not sufficiently accurate in many cases, specifically when classifier-free guidance (Nichol et al., 2021; Ho & Salimans, 2022) is applied. This is mainly due to a distortion-editability trade-off (Tov et al., 2021; Hertz et al., 2022) where reducing the prompt influence improves reconstruction performance but limits the ability to perform significant manipulations through text. However, classifier-free guidance with a large scale is often necessary for non-trivial semantic edits to real images. Such a large guidance scale amplifies the accumulated errors caused by ordinary differential equation (ODE) discretization (Song et al., 2020a; Su et al., 2022) in DDIM such that the obtained noise vector might violate the assumption of Gaussian distribution (Ho et al., 2020; Song et al., 2020b;a). These drawbacks make the DDIM inversion difficult to be *bijective* and *invertible* as both properties are critical to achieving exact reconstruction while retaining semantic text-guide editing capabilities.

This paper proposes an effective inversion approach that mitigates the barriers associated with DDIM inversion given classifier-free guidance in real image editing. To achieve this feat, we mathematically formulate the DDIM inversion by solving an ill-posed inverse problem, which is originally non-bijective and non-invertible. We recognize the ill-posedness induced by classifier-free guidance and accumulated error during inversion and uniformly define them as *information loss*. Hence we introduce learnable latent variables to encode intrinsic information between input images $\mathbf{x}_0$ and noise vectors $\mathbf{x}_T$. To this end, an augmented formulation is built to convert ill-posed inversion to bijective inversion, as shown in Fig. 1. We then identify the learnable latent variables by analyzing the key components of DDIM inversion. Although all works focus on conditional text-based guidance, we observe the substantial effect induced by the unconditional part in classifier-free guidance. Hence, we correct the empty text embeddings by adding a learnable latent variable, which is referred to as "text bias correction". Similarly, the accumulated errors from noise latent vectors can be addressed by adding another learnable latent, which is referred to as "noise bias correction". We optimize both introduced latent variables to address the augmented inversion problem, i.e., to fully invert the input image and prompt.

2 Method

2.1 Augmented DDIM Inversion

In the mathematical perspective, conditional DDIM inversion can be seen as solving ill-posed inverse problems where the mapping between input image $\mathbf{x}_0$ and latent Gaussian variables $\mathbf{x}_T$ are non-bijective and the denoising process is non-invertible, as shown in Fig. 1. Such ill-posedness refers to a mismatch between input image $\mathbf{x}_0$ and reconstruction $\mathbf{x}_0^*$, which is typically caused by "information loss", which refers to the inconsistency, violation of assumptions, accumulative errors induced by inversion.

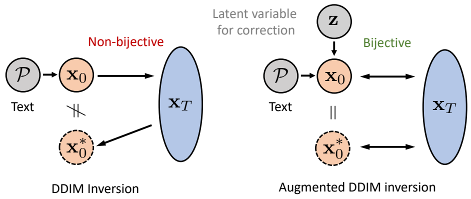

Figure 1: Augmented DDIM inversion (ADI).

Inspired by solving inverse problems via invertible neural networks (Ardizzone et al., 2018), we introduce new latent variables $\mathbf{z}$ to capture the information loss such that the conditional DDIM inversion is invertible. To achieve this goal, an *augmented* inverse problem is formulated based on such a bijective mapping:

$$\mathbf{x}_T = \mathcal{F}_\theta(\mathbf{x}_0, \mathbf{z}; \mathcal{P}), \quad \mathbf{x}_0^* = \mathcal{F}_\theta^{-1}(\mathbf{x}_T, \mathbf{z}, \mathcal{P}) \tag{1}$$

where $\mathcal{F}$ is a encoder function of $\mathbf{x}_0$ and z parametrized by a DDIM model with parameters θ, $\mathcal{P}$ is the textual prompt related to $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\varnothing$. Leveraging the benefits of latent variable z, augmented DDIM inversion becomes fully invertible and bijective. More specifically, our goal is to exactly reconstruct an input image $\mathbf{x}_0$ conditioning on text guidance $\mathcal{P}$ with additional latent variable $\mathbf{z}$ via augmented DDIM inversion $\mathbf{x}_T \rightarrow \mathbf{x}_0^*$, which yields to $\mathbf{x}_0^* = \mathbf{x}_0$.

More specifically, the information loss mainly results from classifier-free guidance and is partially caused by the accumulated error in each diffusion step. We exploit the key feature of the classifier-free guidance, where the result is highly affected by the unconditional prediction. Therefore, we propose to optimize the empty textual embedding by introducing a learnable bias correction, while keeping the conditional textual embedding unchanged. We also observe that a large guidance scale amplifies the accumulated error such that the obtained noise vector does not satisfy the Gaussian assumption. This issue decreases the ability to edit using the particular noise vector. To address this challenge, we define a correct noise vector by $\tilde{\mathbf{x}}_T = \mathbf{x}_T + \mathbf{z}_T$ where $\mathbf{z}_T$ is a learnable latent variable that is referred to as a noise bias correction. Combining the two bias correction together, $\mathbf{z} = (\mathbf{z}_\varnothing, \mathbf{z}_T)$, our augmented DDIM inversion is fully formulated and can be solved by latent trajectory optimization.

2.2 Latent Trajectory Optimization with a Prior

Augmented DDIM inversion provides a promising alternative for accurate reconstruction given conditional classifier-free guidance. Optimizing the bias correction terms $\mathbf{z} = (\mathbf{z}_\varnothing, \mathbf{z}_T)$ is the key aspect of this solution. Inspired by GAN inversion studies, we seek to perform a more efficient optimization consisting of the following three key components:

- **Trajectory optimization**. The reversed DDIM produces a T steps diffusion trajectory $\{\mathbf{x}_t\}_0^T$ between the image encoding $\mathbf{x}_0$ to a Gaussian noise vector $\mathbf{x}_T$. Instead of pursuing a direct matching of the reconstruction, i.e., $\mathbf{x}_0^* \approx \mathbf{x}_0$ at the last step via backward diffusion, we propose to optimize the entire trajectory such that the reconstructed trajectory $\{\mathbf{x}_t^*\}_0^T$ is close as possible to the original trajectory $\{\mathbf{x}_t\}_0^T$, i.e., $\mathbf{x}_t^* \approx \mathbf{x}_t, t = 0, ..., T$. This is defined by $\min \|\mathbf{x}_{t-1}^* - \mathbf{x}_{t-1}\|_2^2$ where $\mathbf{x}_{t-1}$ is the intermediate result of the optimization.

- **Optimization with a prior**. Recent work uses random noise vectors for each iteration of their optimization, aiming at mapping every noise vector to a single image. This method is inefficient and probably suboptimal. Instead, we see to perform a more local optimization by identifying a good starting point guided by "prior" knowledge — we observed that using DDIM inversion with guidance scale $\lambda = 1$ provides a rough approximation of the original image which is highly editable but far from accurate. Thus we refer to this initial DDIM inversion trajectory $\mathbf{x}_T^*$ with $\lambda = 1$ as our trajectory starting point and then perform our trajectory optimization around it with a guidance scale, $\lambda > 1$, enabling high editing capability.
- **Multiple starts strategy**. Differing from the trajectory starting points, we note that latent optimization at $\mathbf{z}$ with gradient descent is sensitive to the initial starting points. It is likely that converged solutions $\mathbf{z}^* = (\mathbf{z}_\varnothing^*, \mathbf{z}_T^*)$ may fall into one of many local minima rather than the global optimum solution. Therefore, we incorporate multiple initializations with several 'hops' within each initialization to more effectively sample through the local minima. In practice, we choose the one that minimizes the initial loss as starting point to perform optimization.

3 Experiments

Experiment setting Our framework is general and can be combined with different state-of-the-art large models, but here we use Stable Diffusion (Rombach et al., 2022) training on LAION-5B datasets (Schuhmann et al., 2022). We have used a subset of 200 images and captions pairs selected from the MS-COCO dataset (Chen et al., 2015) and TEdBench dataset (Kawar et al., 2022). The baseline methods include DDIB (Su et al., 2022) and CycleDiffusion (Wu & De la Torre, 2022) and we use the same hyperparameters to conduct a fair comparison given similar computing and memory costs.

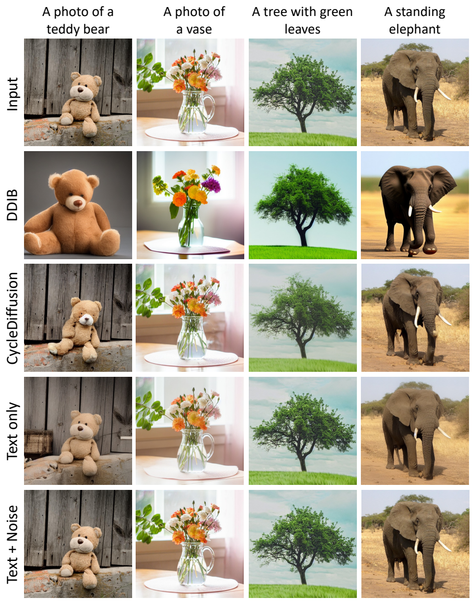

Figure 2: Comparison of diffusion inversion.

For our method in Stable Diffusion trained model, we use $T = 50$ steps for DDIM inversion and $m = 10$ iterations for each timestamp. The guidance scale for latent-TOP is $\mathcal{G} = 7.5$ which is the default value used by (Nichol et al., 2021). We generate 20 samples to perform multiple start initialization for text and noise bias correction. Note that we divide the entire optimization procedure into two steps: (1) obtain the prior trajectory via DDIM inversion and save the sequence and (2) execute the separate optimization by loading the saved trajectory information. Using this way, we achieve to finish the algorithm by using 1-2 minutes on V100 with 16G GPU memory.

Results Figure 2 shows the results for DDIM inversion. DDIB (Su et al., 2022) clearly fails to reconstruct the input images due to instability and inherent limitations of DDIM inversion. CycleDiffusion (Wu & De la Torre, 2022) seems better than DDIB (Su et al., 2022) and provides an approximate reconstruction of input images but is still struggling with exactly recovering some details, e.g., the posture of the teddy bear. The fourth and fifth rows in Figure 2 offer our solution for DDIM inversion where the accuracy outperforms the other two baseline methods.

3.1 Robustness to Image Perturbation

Out latent-TOP provides a promising inversion reconstruction with high accuracy but one may ask whether our method is robust to a small perturbation in either image pixel level or text caption prompt.

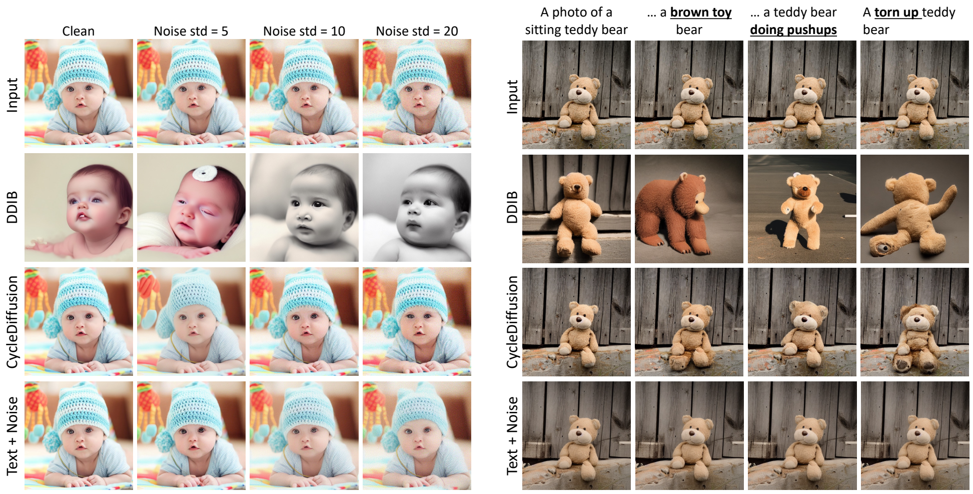

Figure 3: (Left) Noise perturbation on images. Given the fixed input source caption (”A photo of a baby”), we compare the inversion given various noise perturbations added to the input image. (Right) Prompt perturbation on text. Given the same input image, we compare the inversion under different source prompt guidance.

Here, we first investigate the robustness of our method on image perturbation, see Fig.3. In this case, we start from the clean input image and gradually add Gaussian noise with a variety of levels, including $\sigma = 5, 10, 20$ (pixel value ranging from 0 to 255). Given the same source prompt, we compare the reconstruction performance given noisy image inputs. Similar to the accuracy results, DDIB (Su et al., 2022) is clearly far away from the right inversion path where much information is lost, e.g., background, baby's hat, posture, and close, even though only a tiny perturbation is applied. When noise increases, DDIB (Su et al., 2022) fails to recognize the color style. CycleDiffusion (Wu & De la Torre, 2022) provides competitive results, specifically when the noise level is low. Although some details are not exactly covered, it greatly outperforms DDIB (Su et al., 2022). When the noise increases to 20, we expect our algorithm can recover the original input while capturing the noise style. A larger noise makes the inversion difficult but still preserves the global semantics and local details. Our method achieves superior performance but CycleDiffusion (Wu & De la Torre, 2022) starts to display distortion in some local details.

3.2 Robustness to Text Perturbation

Our method requires an input source caption as a prompt, so it is natural to analyze the sensitivity of chosen caption. Here we use the teddy bear as an example, keeping the input image unchanged but changing the original source caption in different ways. It's a non-trivial task to assign a reasonable and specific caption without introducing any bias or imprecise information when we see an image for the first time. The extreme case is sampling a random caption from the dataset for each image but it is undesired for text-based editing clearly. Figure 3 shows the inversion results using multiple captions. For example, we replace the “teddy bear” with “brown toy bear” or use “torn up” instead of “photo”. Our method shows strong robustness in defending these perturbations applied on textual prompts, i.e., source caption. This is due to the advantages of textual bias correction, which optimizes the empty text prompt to match different conditional textual prompts.

4 Conclusion

This paper provides a novel inversion method by formulating an augmented inversion problem with learnable bias corrections. We address this inversion problem through an efficient latent trajectory optimization with a prior. Leveraging the benefits of three core components in latent-TOP, we significantly improve optimization efficiency, convergence, and stability. By evaluating various input real images, source captions, and target prompt, our method achieves impressive performance on inversion accuracy, robustness to perturbation on text and noise, as well as real image editing.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8296–8305, 2020.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18511–18521, 2022.

Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations*, 2018.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11431–11440, 2022.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14347–14356. IEEE Computer Society, 2021.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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.

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*, pp. 2426–2435, 2022.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

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 preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pp. 319–345. Springer, 2020.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287–2296, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

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. *arXiv preprint arXiv:2210.08402*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022.

Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022.

Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. *arXiv preprint arXiv:2212.08698*, 2022.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022.

Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pp. 592–608. Springer, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A Appendix

A.1 Background

A.1.1 Diffusion Models

Diffusion Denoising Probabilistic Models (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) aim at modeling a distribution $p_\theta(\mathbf{x}_0)$ to approximate the data distribution $q(\mathbf{x}_0)$. The forward process performs a progressing procedure from $\mathbf{x}_0$ to $\mathbf{x}_T$ via a Markov chain, where we generate the latent variables $\mathbf{x}_1, ..., \mathbf{x}_T$ by gradually adding noise to the data via Gaussian transition. When T is large enough, the last noise vector $\mathbf{x}_T$ nearly follows an isotropic Gaussian distribution.

The forward process has a simple closed-form solution that expresses the latent variable $\mathbf{x}_t, t \in \{0, ..., T\}$ as a linear combination of noise and $\mathbf{x}_0$ (Ho et al., 2020):

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

where α_t is referred to as the noising schedule which defines the amount of noise present at each intermediate timestep, $0 = \alpha_T < \alpha_{T-1} < ..., < \alpha_1 < \alpha_0 = 1$. Each refinement step consists of an application of a neural network $f_\theta(\mathbf{x}, t)$ on the current sample $\mathbf{x}_t$, followed by a random Gaussian noise perturbation, obtaining $\mathbf{x}_{t-1}$. The network is trained for a simple denoising objective, aiming for $f_\theta(\mathbf{x}_t, t) = \epsilon_\theta^{(t)}(\mathbf{x}_t) \approx \epsilon_t$.

Sampling from distribution $q(\mathbf{x}_0)$ is defined by a reverse process, from isotropic Gaussian noise $\mathbf{x}_T$ to data, which is refined iteratively through $t \leq T$ passes through the network. There are various sampling strategies (Song et al., 2020a; Nichol & Dhariwal, 2021) that define the process of merging the noise prediction $\epsilon_\theta^{(t)}(\mathbf{x}_t)$ and current sample $\mathbf{x}_t$ to produce the previous sample $\mathbf{x}_{t-1}$. The final $\mathbf{x}_0$ sample is the resultant generated image.

A.1.2 DDIM Inversion

Unlike the commonly used DDPM, the generative sampling process in DDIMs is defined in a non-Markovian manner,

$$\mathbf{x}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\mathbf{x}_t + \left(\sqrt{\frac{1-\alpha_{t-1}}{\alpha_{t-1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\epsilon_\theta^{(t)}(\mathbf{x}_t) \tag{3}$$

which can be used for inversion, based on the assumption that the ordinary differential equation (ODE) process can be reversed in small steps:

$$\mathbf{x}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\mathbf{x}_t + \left(\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\epsilon_\theta^{(t)}(\mathbf{x}_t). \tag{4}$$

Thus, the diffusion process is performed in the reverse direction, deterministically noising an image to obtain the initial noise vector. In other words, DDIM inversion achieves $\mathbf{x}_0 \to \mathbf{x}_T$ instead of $\mathbf{x}_T \to \mathbf{x}_0$.

Empirically, the error of DDIM inversion is reasonably small since Eq. 4 can be treated as an Euler method over the following ODE, which is up to discretization errors of the ODE solvers:

$$\mathrm{d}\hat{\mathbf{x}}(t) = \epsilon_\theta^{(t)}\left(\frac{\hat{\mathbf{x}}(t)}{\sqrt{\sigma^2+1}}\right)\mathrm{d}\sigma(t) \tag{5}$$

where $\hat{\mathbf{x}} = \mathbf{x}/\sqrt{\alpha}$ and $\sigma = \sqrt{1-\alpha}/\sqrt{\alpha}$. However, in practice, a slight error is incorporated in every step, and eventually, the accumulated error might be non-negligible. In some cases, the obtained noise vector might be out of the Gaussian assumption (see the examples in Section 4).

A.1.3 Conditional DDIM Inversion

For unconditional diffusion models, DDIM inversion succeeds given negligible errors. However, the goal of text-guided diffusion models is to generate an output image given conditioning textual prompt $\mathcal{P}$. The critical challenge is how to amplify the effect induced by the conditional text. To this end, a classifier-free guidance prediction (Ho & Salimans, 2022; Nichol et al., 2021) is introduced

$$\tilde{\epsilon}_\theta^{(t)} = \epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\varnothing) + \mathcal{G} \cdot (\epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\mathcal{P}) - \epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\varnothing)) \tag{6}$$

where $\mathcal{C}_\mathcal{P} = \Psi(\mathcal{P})$ is the conditional textual embedding, $\mathcal{C}_\varnothing = \Psi(\text{“ ”})$ is the empty embedding, which is referred to as the unconditional textual guidance, and $\mathcal{G}$ is the guidance scale, which adjusts the balance between conditional prediction $\epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\mathcal{P})$ and unconditional prediction $\epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\varnothing)$.

Thus the unconditional DDIM in Eq. equation 3 can be naturally extended to the conditional formulation:

$$\mathbf{x}_{t-1} = \xi_t \cdot \mathbf{x}_t + \phi_t \cdot \tilde{\epsilon}_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\mathcal{P}, \mathcal{C}_\varnothing) \tag{7}$$

where $\xi_t = \sqrt{\alpha_{t-1}/\alpha_t}$ and $\phi_t = \sqrt{(1-\alpha_{t-1})/\alpha_{t-1}} - \sqrt{(1-\alpha_t)/\alpha_t}$. The above denoising process is approximately invertible as claimed in DDIM (Song et al., 2020a). Hence, conditional DDIM inversion can be performed by

$$\mathbf{x}_t = \frac{\mathbf{x}_{t-1}}{\xi_t} - \frac{\phi_t}{\xi_t} \cdot \tilde{\epsilon}_\theta^{(t)}(\mathbf{x}_t) \approx \frac{\mathbf{x}_{t-1}}{\xi_t} - \frac{\phi_t}{\xi_t} \cdot \tilde{\epsilon}_\theta^{(t)}(\mathbf{x}_{t-1}) \tag{8}$$

where $\mathbf{x}_t$ is approximately reversible from $\mathbf{x}_{t-1}$. The success of the approximation in Eq. equation 8 relies on the linearization assumption $\tilde{\epsilon}_\theta^{(t)}(\mathbf{x}_t) \approx \tilde{\epsilon}_\theta^{(t)}(\mathbf{x}_{t-1})$, which corresponds to reversing the first-order ODE solver (Su et al., 2022; Song et al., 2020b). Although higher-order ODE solvers can stabilize the generative and reverse processes, DDIM inversion with these ODE solvers still relies on the strength of the linearization assumption (Wallace et al., 2022). This assumption works for the unconditional DDIM models, but it fails in conditional DDIM inversion since the classifier-free guidance $\mathcal{G} \cdot (\epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\mathcal{P}) - \epsilon_\theta^{(t)}(\mathbf{x}_t, \mathcal{C}_\varnothing))$ is not consistent across time steps (Nichol et al., 2021). A trade-off solution is to decrease the guidance scale $\mathcal{G}$, e.g., from 7.5 to 1.0, for stabilization and better reconstruction accuracy, but it damages the strength of conditional editing in practice.

Moreover, as noted in the prompt-to-prompt work (Hertz et al., 2022), the DDIM inversion is unstable in many cases, specifically encoding from $\mathbf{x}_0$ to $\mathbf{x}_T$ and back often results in inexact reconstructions of the input images. This might be explained by our observation that image editing and manipulations through text using Stable Diffusion models (Rombach et al., 2022) require classifier-free guidance with a large guidance scale, and such a guidance scale amplifies the accumulated error. That is to say, an increase in corruption is correlated with the strength of the conditioning.

A.2 Related Work

Many works utilized GANs to perform a variety of image manipulations. Inversion has been well studied for GANs (Xia et al., 2022; Zhu et al., 2020), ranging from encoder-based methods (Richardson et al., 2021), latent-based optimization (Abdal et al., 2019; 2020), to fine-tuning of the models (Alaluf et al., 2022). However, these studies have limitations in editing a given real image while preserving the details of unedited parts.

Diffusion models show remarkable results for image manipulation tasks, e.g., image-to-image translation (Choi et al., 2021; Zhao et al., 2022). A desirable property of image translation is the cycle consistency (Zhu et al., 2017; Park et al., 2020). DDIB (Su et al., 2022) enforces exact cycle consistency by encoding an input image using DDIM inversion with a source class (or text) and decoding it back conditioned on the target class (or text) but often fails to retain the semantic structures precisely. CycleDiffusion (Wu & De la Torre, 2022) uses a deterministic DPM encoder to enable zero-shot image translation and editing by leveraging text-to-image diffusion models. However, the editing performance is sensitive to the choice of pre-defined prompt even though cycle consistency is satisfied. SDEit (Meng et al., 2021) adds intermediate noise to an image, then denoises it using a diffusion process conditioned on the desired edit, which is struggling with global edits.

Prompt-to-Prompt (Hertz et al., 2022) alters a text-to-image editing by manipulating cross-attention layers, providing more control over synthesized images, and can be extended to real image editing when DDIM inversion offers meaningful attention maps (Mokady et al., 2022). DiffEdit (Couairon et al., 2022) edits real/synthetic images using automatically generated masks for regions of the input image. DiffusionCLIP (Kim et al., 2022) edits image by using a combination of DDIM inversion, large model gradients, and model fine-tuning. Imagic demonstrates impressive editing results leveraging the Imagen model (Saharia et al., 2022) but also requires the restrictive fine-tuning of the model. SINE (Zhang et al., 2022) proposes model-based guidance with a patch-based fine-tuning scheme. Unlike these recent works, our method does not require fine-tuning the entire model, avoiding damaging the trained models for each image. More recently, EDICT (Wallace et al., 2022) proposes an exact DDIM inversion scheme via affine coupling layers but its computational cost is twice that of a baseline DDIM process due to double loops. Our method is more efficient as we only need to fine-tune the learnable latent via a local optimization guided by prior trajectory.