IterInv: Iterative Inversion for Pixel-Level T2I Models

Chuanming Tang[1,2,3], **Kai Wang**[3], **Joost van de Weijer**[3,4]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China
[3]Computer Vision Center, Barcelona, Spain
[4]Universitat Autonoma de Barcelona, Barcelona, Spain
tangchuanming19@mails.ucas.ac.cn, {kwang,joost}@cvc.uab.es

Abstract

Large-scale text-to-image diffusion models have been a ground-breaking development in generating convincing images following an input text prompt. The goal of image editing research is to give users control over the generated images by modifying the text prompt. Current image editing techniques predominantly hinge on DDIM inversion as a prevalent practice rooted in Latent Diffusion Models (LDM). However, the large pretrained T2I models working on the latent space suffer from losing details due to the first compression stage with an autoencoder mechanism. Instead, other mainstream T2I pipeline working on the pixel level, such as Imagen and *DeepFloyd-IF*, circumvents the above problem. They are commonly composed of multiple stages, typically starting with a text-to-image stage and followed by several super-resolution stages. In this pipeline, the DDIM inversion fails to find the initial noise and generate the original image given that the super-resolution diffusion models are not compatible with the DDIM technique. According to our experimental findings, iteratively concatenating the noisy image as the condition is the root of this problem. Based on this observation, we develop an iterative inversion (*IterInv*) technique for this category of T2I models and verify *IterInv* with the open-source *DeepFloyd-IF* model. By combining our method with a popular image editing method, we prove the application prospects of *IterInv*. The code will be released at `https://github.com/Tchuanm/IterInv.git`

1 Introduction

The Text-to-Image (T2I) field has witnessed significant advancements and demonstrated an unprecedented ability to generate realistic images [20, 26–30]. State-of-the-art T2I models undergo training on immensely large language-image datasets, demanding substantial computational resources. Nevertheless, notwithstanding their remarkable abilities, these models do not readily facilitate *real image editing*. *Text-guided image editing*, also referred to as prompt-based image editing, empowers users to effortlessly modify an image exclusively through text prompts [3, 6, 9, 10, 18, 32, 35]. Several of the existing methods use DDIM inversion [31] as a common practice to attain the initial latent code of the image and then apply their proposed editing techniques along the denoising phase [21, 23, 32]. Nevertheless, current text-guided editing methods are mainly working on latent-level T2I Latent Diffusion Models (LDM) [28]. And the first compression stage with an Autoencoder results in losing details of the original images as our experiments show. Instead, another mainstream of pixel-level T2I diffusion models [29, 30] avoids this problem by directly generating images over the pixel space. However, the inversion and text-guided image editing techniques are not well explored for this branch.

In this paper, we take advantage of the open-source *DeepFloyd-IF* model as the representative for pixel-level T2I models. We first observe that directly applying the DDIM inversion to the open-source T2I *DeepFloyd-IF* model will lead to failures in reconstructing the original images. We attribute

Neural Information Processing Systems (NeurIPS) 2023 Workshop on Diffusion Models.

this phenomenon to the concatenation conditioning with the noisy images in the super-resolution diffusion models. To solve this problem, we propose the iterative inversion (*IterInv*) technique, where we find the trace back in the diffusion process by iterative optimization to approximate the real image.

From our experiments over public datasets collected from previous text-guided editing papers [23, 32], we verify the extraordinary reconstruction ability of *IterInv* and successfully combine it with DiffEdit [6] to show its compatibility with editing methods. We prospect this proposal can help to innovate future research on text-guided image editing based on the pixel-level T2I models.

2 Related work

Large-scale T2I models. The pioneering text-to-image frameworks based on diffusion model can be roughly categorized considering where the diffusion prior is conducted, i.e., the pixel space or latent space. The first class of methods generate images directly from the high-dimensional pixel level, including GLIDE [22], Imagen [29] and *DeepFloyd-IF* [30]. Another stream of works proposes to first compress the image to a low-dimensional space, and then train the diffusion model on this latent space. Representative methods falling into this category include LDM [28] and DALL-E [26].

Inversion techniques for T2I models. DDIM inversion [31] shows significant potential in editing tasks by deterministically calculating and encoding context information into a latent space and then reconstructing the original image using this latent representation. However, DDIM is found lacking for text-guided diffusion models when classifier-free guidance (CFG) [13] is applied, which is necessary for meaningful editing. Leveraging optimization on Null-Text embedding, Null-Text Inversion (NTI) [21] further improved the image reconstruction quality when CFG is applied, and retained the rich text-guided editing capabilities of the Stable Diffusion model [28].

Text-guided image editing. Various methods [1, 5, 16–18] leverage CLIP [25] for image editing based on a pretrained *unconditional* diffusion model. However, they are limited to the generative prior which is only learned from visual data of a specific domain, whereas the CLIP text-image alignment information is from much broader data. This prior gap is mitigated by recent progress of T2I models [4, 7, 14, 26, 27, 29]. Nevertheless, these T2I models offer little control over the generated contents. This creates a great interest in developing methods [2, 10, 15, 19, 23, 24, 32, 34] to adopt such T2I models for controllable image editing. However, how to cooperate these methods with pixel-level T2I models is still an open topic, and we are the first to shed light on this direction.

3 Methodology

3.1 Preliminary

***DeepFloyd-IF*.** We develop our method for the publicly available *DeepFloyd-IF* method, but the method is general and can be extended to other diffusion models. Pixel-level T2I *DeepFloyd-IF* [30] is a cascaded diffusion model. It is composed of three cascaded pixel-level diffusion stages: a text-based low-resolution text-to-image generation stage S1, and two super-resolution diffusion modules as stage S2 and S3. The output resolutions of these stages are 64, 256, and 1024 respectively. In stage S1, given the conditional input prompt $\mathcal{P}$, a conditioning mechanism denoted as $\tau_\theta(\mathcal{P})$ is employed, which maps the condition $\mathcal{P}$ into a prompt vector.

DDIM Inversion. Inversion entails finding an initial noise z_T that reconstructs the code z_0 of the real image upon sampling. Existing image editing methods [2, 8] aim at precisely reconstructing a given image for editing, therefore employ the deterministic latent DDIM sampling [31]: $z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(z_t, t, \mathcal{C}) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(z_t, t, \mathcal{C})$. $\bar{\alpha}_{t+1}$ is noise scaling factor defined in DDIM [31] and $f_\theta(z_t, t, \mathcal{C})$ predicts the denoised latent code z_0 as $f_\theta(z_t, t, \mathcal{C}) = \left[z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, \mathcal{C})\right] / \sqrt{\bar{\alpha}_t}$. However, with the pixel-level inversion, the concatenation condition with a noisy image in stages S2 and S3 of the *DeepFloyd-IF* model complicates the inversion process, and the generated image based on the found initial noise image of DDIM inversion deviates from the original image.

3.2 *IterInv*: Iterative Inversion

DDIM inversion provides a deterministic process for conventional diffusion models [12, 31], similar to the first stage S1 of the *DeepFloyd-IF* model. However, directly applying DDIM to *DeepFloyd-IF* leads to failure reconstructions as shown in Fig. 2. In the second and third columns, we only apply DDIM inversion with real images in the super-resolution stages S2 and S3. It is evident that the reconstructions retain only the layout structure, while other details are lost. Then with DDIM inversion image as an output of stage S1, the reconstruction is even worse as in the fourth column. Based on this finding, we devise a novel iterative inversion framework (*IterInv*) that successfully injects the DDIM into the super-resolution stages S2 and S3. The detail of the cascaded network is shown in Fig. 1 and the algorithm diagram is shown in Algorithm 1. For each input image, we resize into three scales 64,256,1024 as inputs for different stages $x^{s_1}, x^{s_2}, x^{s_3}$.

Stage S1. In stage S1, with the low-resolution input image x^{s_1}, we employ Null-Text inversion [21] in conjunction with classifier-free guidance [13] to approximate the original image.

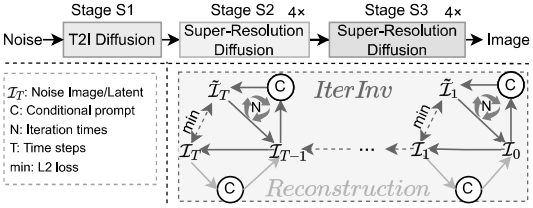

Figure 1: The network of *DeepFloyd-IF* pipeline and proposed *IterInv* inversion technology.

Stage S2/S3. In stages S2 and S3, the super-resolution diffusion models can operate at either pixel-level (i.e., *DeepFloyd-IF*-upscaler) or latent-level (i.e., SD-upscaler). In stage S2, we use $\mathcal{I}_0^{s_2}$ to formulate the real image x^{s_2} or latent code $z_0^{s_2}$. With the input as $\mathcal{I}_0^{s_2}$, we design the reverse iterative inversion process for each timestep. Specifically, in the process of $\mathcal{I}_{t-1}^{s_2} \rightarrow \mathcal{I}_t^{s_2}$, we design a N steps iterative optimization operation. We initialize $\tilde{\mathcal{I}}_t^{s_2} = \mathcal{I}_{t-1}^{s_2}$, and perform the following optimization for N times iterations:

$$\min ||\mathcal{I}_{t-1}^{s_2} - \tilde{\epsilon}_\theta(\tilde{\mathcal{I}}_t^{s_2}, \hat{\mathcal{I}}_0^{s_1}, t, \mathcal{C})||_2^2 \tag{1}$$

Here $\tilde{\epsilon}_\theta(\tilde{\mathcal{I}}_t^{s_2}, \hat{\mathcal{I}}_0^{s_1}, t, \mathcal{C})$ is the DDIM sampling on $\tilde{\mathcal{I}}_t^{s_2}$ and the noised reconstruction image $\hat{\mathcal{I}}_0^{s_1}$ from the previous stage. After the optimization, we update the input as:

$$\mathcal{I}_t^{s_2} = \tilde{\epsilon}_\theta(\tilde{\mathcal{I}}_t^{s_2}, \tilde{\mathcal{I}}_0^{s_1}, t, \mathcal{C}). \tag{2}$$

Afterwards, with the N iteration optimization of each timestep, we minimize the disparity between the predicted and excepted noise. Then, we regard the last iteration predict noise $\mathcal{I}_t^{s_2}$ as the trace of the inversion process of step t. The final noise $\mathcal{I}_T^{s_2}$ is the ideal trace for promoting the follow-up accurate image reconstruction and editing task. Stage S3 shares a similar mechanism as stage S2, and thus we employ the same *IterInv* strategy.

Image Edit. After successfully inverting real images with *DeepFloyd-IF*, we further enable the text editing technique on *DeepFloyd-IF*. To our best knowledge, our *IterInv* is the first to realize image editing with the pixel-level diffusion manner. With our deterministic iterative inversion, we introduce DiffEdit into *DeepFloyd-IF* to achieve text-guided image editing, as shown in the last column in Fig. 2.

Algorithm 1: Iterative Inversion (*IterInv*)

Input: A source prompt $\mathcal{P}$,
Three scale input images $x^{s_1}, x^{s_2}, x^{s_3}$
Initialize: $\varnothing_T = \tau_\theta(\text{“”}), \mathcal{C} = \tau_\theta(\mathcal{P})$;
$\omega_1 = 7.0, \omega_2 = 1.0, \omega_3 = 1.0$;
$\mathcal{I}_0^{s_1} = x^{s_1}, \mathcal{I}_0^{s_2} = x^{s_2}, \mathcal{I}_0^{s_3} = x^{s_3}$;
Stage S1:
$\mathcal{I}_T^{s_1}, \{\varnothing_t\}_{t=1}^T, \bar{\mathcal{I}}_0^{s_1} \leftarrow \textbf{NTI}(\mathcal{I}_0^{s_1}, \mathcal{C}, \varnothing_T)$
Stage S2:
$\hat{\mathcal{I}}_0^{s_1} \leftarrow \bar{\mathcal{I}}_0^{s_1} + noise$
for $t = 1, 2, \ldots, T$ **do**
  $\tilde{\mathcal{I}}_t^{s_2} \leftarrow \mathcal{I}_{t-1}^{s_2}$;
  for $j = 1, ..., N$ **do**
    $\tilde{\mathcal{I}}_{t-1}^{s_2} = \tilde{\epsilon}_\theta(\tilde{\mathcal{I}}_t^{s_2}, \hat{\mathcal{I}}_0^{s_1}, t, \mathcal{C})$;
    $L_t = ||\mathcal{I}_{t-1}^{s_2} - \tilde{\mathcal{I}}_{t-1}^{s_2}||_2$;
    $\tilde{\mathcal{I}}_t^{s_2} \leftarrow \tilde{\mathcal{I}}_t^{s_2} - \frac{\Delta L_t}{\Delta \tilde{\mathcal{I}}_t^{s_2}}$
  end
  $\mathcal{I}_t^{s_2} \leftarrow \tilde{\mathcal{I}}_t^{s_2}$;
end
$\bar{\mathcal{I}}_0^{s_2} \leftarrow (\mathcal{I}_T^{s_2}, \hat{\mathcal{I}}_0^{s_1}, \mathcal{C})$ ◁ *Reconstruction*
Stage S3: Similar as lines 8-18 computing $\mathcal{I}_T^{s_3}$
Return: $\mathcal{I}_T^{s_1}, \mathcal{I}_T^{s_2}, \mathcal{I}_T^{s_3}$

4 Experiments

Data and Implementation. We evaluate the proposed *IterInv* qualitatively and quantitatively. We experimented *IterInv* using *DeepFloyd-IF* with PyTorch and diffusers [33]. For quantitative

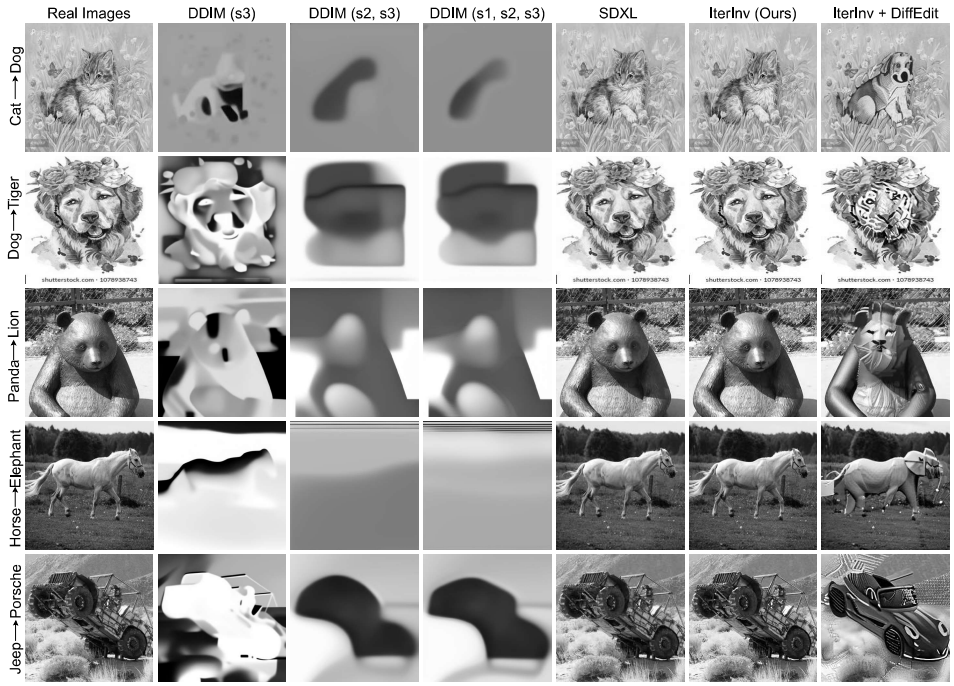

Figure 2: Visualization comparison of various inversion means and our editing results.

Base model	Method	CFG ω_1	MSE (↓)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	CLIP (↑)
SDXL	Autoencoder	-	0.009016	0.1269	0.9048	33.4993	**21.3172**
DeepFloyd-IF	DDIM (s3)	-	0.275662	0.7882	0.4213	6.8092	21.2448
	DDIM (s2,s3)	-	0.076924	0.6413	0.5865	11.8044	21.1808
	DDIM (s1,s2,s3)	1.0	0.079924	0.6393	0.5821	11.8366	21.2223
DeepFloyd-IF	Ours (*IterInv*)	1.0	0.000130	0.0356	0.9806	40.6102	21.2907
		3.0	0.000129	0.0353	0.9806	40.6459	21.3161
		5.0	0.000129	0.0353	0.9806	40.6484	21.3161
		7.0	**0.000129**	**0.0353**	**0.9806**	**40.6484**	21.3161

Table 1: Evaluation of reconstruction quality of each method on *ImageInversion* dataset.

comparison, we collect the *ImageInversion* dataset with 69 images: 50 images from PnP [32] and 19 images from Pix2pix-zero [23]. To ensure a fair comparison among different methods, all images are resized to 1024×1024. To comprehensively evaluate the quality of inversion methods, we employ five metrics including MSE, LPIPS [37], SSIM [36], PSNR, and CLIP score [11] to verify the effectiveness of *IterInv*. In *IterInv*, we set $T = 50$, $N = 20$ by default.

Reconstruction and Editing. The qualitative and quantitative measurements of reconstructions are shown in Fig. 2 and Tab. 1 respectively. Compared with DDIM inversion applied in various stages and the *first stage autoencoder* of the SDXL model, *IterInv* achieves more precise image construction and there is only the SDXL *compression autoencoder*[1] can reconstruct the target image closer to our final performance, as indicated by the five metric evaluations. Furthermore, we visualize the image editing results by applying DiffEdit to *IterInv* in the last column of Fig. 2. By these convincing image editing results, we are showing the application prospect of *IterInv* combined with text-guided image editing method in future research.

Ablation study. Current inversion methods on T2I model are easily influenced by the hyperparameter of classifier-free guidance (CFG [21]). We conduct the ablation study over various CFG values of stage S1 (ω_1), as shown in Tab. 1. With different values of ω_1, *IterInv* achieves nearly the same reconstruction performance, which shows the robustness of *IterInv* to the classifier-guidance scale.

[1]The first stage compression model makes the upper bound for any inversion techniques based on the SDXL.

5 Conclusions

In existing research, text-guided image editing mainly focuses on latent diffusion models (LDM), which might lead to missing details in the final results due to the low resolution of the latent space. Pixel-level T2I diffusion models can be applied to avoid this drawback, e.g. *DeepFloyd-IF* as utilized in this paper. However, the popular DDIM inversion technique is not able to guarantee the reconstruction of real images when applied to the pixel-level *DeepFloyd-IF* model. We attribute this problem to the condition that is concatenated with a noisy input. To address this issue, we propose iterative inversion (*IterInv*) as a novel inversion technique. The experimental results conducted on real images confirm that *IterInv* obtains superior inversion and editing performance.

Acknowledgments We acknowledge projects TED2021-132513B-I00 and PID2022-143257NB-I00, financed by MCIN/AEI/10.13039/501100011033 and FSE+ and the Generalitat de Catalunya CERCA Program. Chuanming acknowledges the Chinese Scholarship Council (CSC) No.202204910331.

References

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

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[3] 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. *Proceedings of the International Conference on Computer Vision*, 2023.

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the International Conference on Computer Vision*, pages 14347–14356. IEEE, 2021.

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[8] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023.

[9] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. *arXiv preprint arXiv:2304.07090*, 2023.

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*, 2023.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

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

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

[14] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *Proceedings of the International Conference on Computer Vision*, 2023.

[15] 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

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

[17] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*, 2023.

[18] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing, 2023.

[19] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[20] Midjourney.com. Midjourney. `https://www.midjourney.com`, 2022.

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[22] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022.

[23] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 2023.

[24] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *Proceedings of the International Conference on Computer Vision*, 2023.

[25] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] 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.

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[28] 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 (CVPR)*, pages 10684–10695, June 2022.

[29] 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.

[30] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if. `https://github.com/deep-floyd/IF`, 2023.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[32] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[33] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[34] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 2023.

[35] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path, 2023.

[36] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

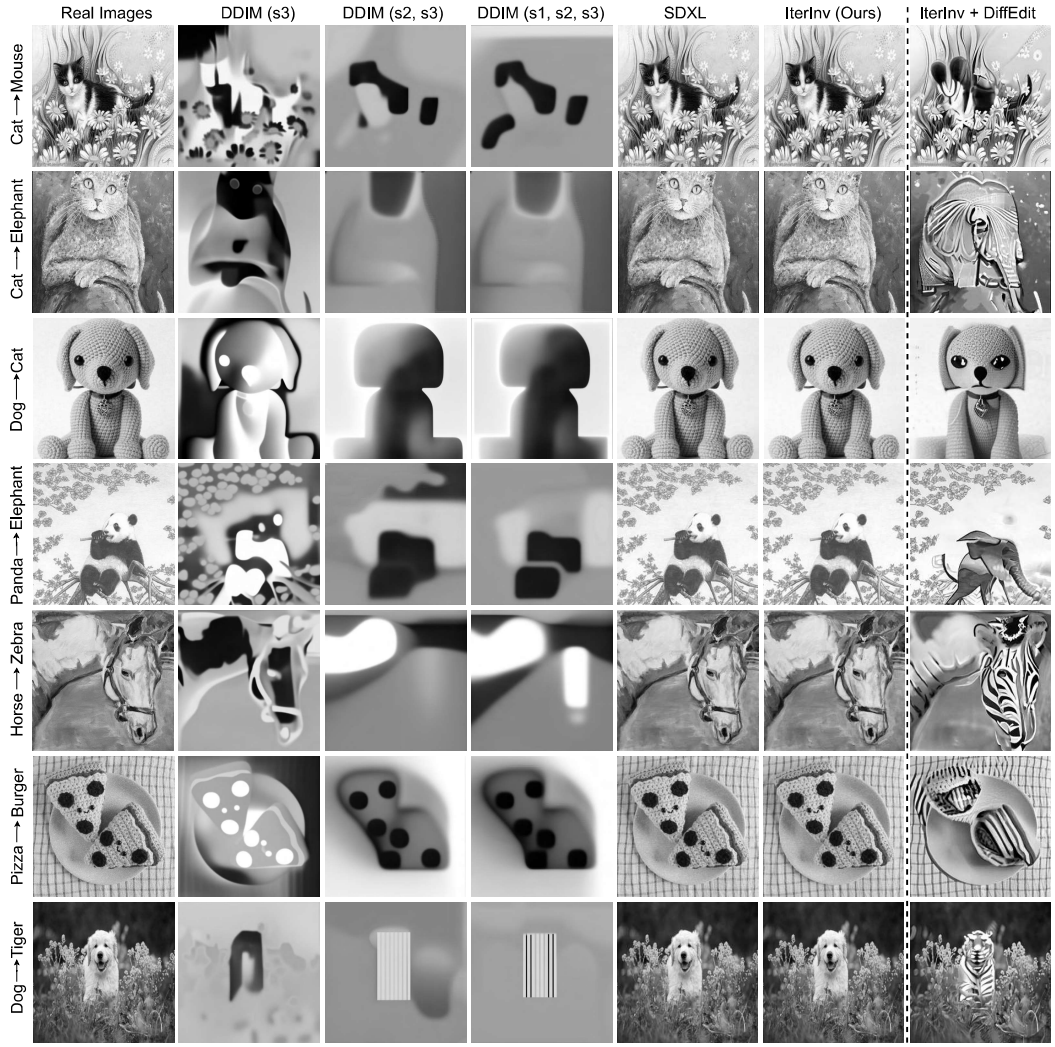

Figure 3: More visualization of reconstruction and editing examples of each method compared with *IterInv* on the *ImageInversion* dataset.

A Appendix A: More visualization

In Fig. 3, we present more inversion and editing examples as an extension for Fig. 2. In Tab. 1, MSE, LPIPS, SSIM, and PSNR all measure the quality of the original image with a reconstruction image, reflecting the inversion quality of different methods. However, the CLIP score is a method to measure an image's proximity to the provided prompt, which is not really related to the reconstruction quality compared with the original image. In Fig. 2 and Fig. 3, the reconstruction quality of SDXL Autoencoder and ours *IterInv* are all looks good. However, as compared in Tab.1, the MSE, LPIPS, SSIM, and PSNR of our *IterInv* all outperform SDXL with a significant margin, which shows the excellent reconstruction ability of our *IterInv* in pixel-level diffusion method.