Effective Real Image Editing with Accelerated Iterative Diffusion Inversion

Zhihong Pan   Riccardo Gherardi   Xiufeng Xie   Stephen Huang

Oppo Mobile Telecommunications Corp.

2479 E Bayshore Rd, Palo Alto, CA, USA

Abstract

Despite all recent progress, it is still challenging to edit and manipulate natural images with modern generative models. When using Generative Adversarial Network (GAN), one major hurdle is in the inversion process mapping a real image to its corresponding noise vector in the latent space, since it is necessary to be able to reconstruct an image to edit its contents. Likewise for Denoising Diffusion Implicit Models (DDIM), the linearization assumption in each inversion step makes the whole deterministic inversion process unreliable. Existing approaches that have tackled the problem of inversion stability often incur in significant trade-offs in computational efficiency. In this work we propose an Accelerated Iterative Diffusion Inversion method, dubbed AIDI*, that significantly improves reconstruction accuracy with minimal additional overhead in space and time complexity. By using a novel blended guidance technique, we show that effective results can be obtained on a large range of image editing tasks without large classifier-free guidance in inversion. Furthermore, when compared with other diffusion inversion based works, our proposed process is shown to be more robust for fast image editing in the 10 and 20 diffusion steps' regimes.*

1. Introduction

Diffusion models are a class of generative models that learn to generate high-quality images by iteratively applying a denoising process from a random noisy starting point. They are capable to achieve state-of-the-art (SOTA) image quality since the very early stage of denoising diffusion probabilistic models (DDPM) [21] and score-based generative modeling [42]. While the number of sampling steps required for high quality image generation was initially very large, several follow-up studies [34, 7, 29, 30] have reduced significantly the number of steps without degrading the image quality, making the widespread use of diffusion models possible. In particular, denoising diffusion implicit models (DDIM) [41] are widely used for their speed and flexibility in deterministic and stochastic generations. Further reduc-

ing the number of sampling steps for both image generation or editing is still nevertheless an open research problem.

Diffusion models were initially designed for image generation; for this reason, their usefulness for real image editing is limited without a proper inversion process, similar to the inversion challenge [46] faced by real image editing using a Generative Adversarial Network (GAN) [17]. GAN inversion is limited by the reduced dimensionality in latent space; diffusion inversion is comparably less restricted as the latent space has the same dimensionality as the original image. Naïve one-step inversion step, *i.e.* simply perturbing an image with random noise, was initially used for early image editing works such as SDEdit [31] and Blended Diffusion [6]. Later, an Euler method based inversion process that applies deterministic step-by-step noise injection was used for image-to-image translation in DDIB [43] and DiffusionCLIP [27]. However, as shown in the text-guided image editing tests in Prompt-to-Prompt (PTP) [19], such inversion is not reliable for real image editing because the inversion often leads to failed reconstruction even when no editing is performed. Follow-up works like null-text inversion (NTI) [32] and exact diffusion inversion (EDICT) [45] have focused on improving the reconstruction accuracy by introducing auxiliary variables like the learned null-text embedding in NTI, or processes like the coupled diffusion process in EDICT. The reconstruction accuracy improvements come however with regressions in computational complexity of the inversion and/or the editing processes.

In this paper, we are the first to look beyond the simple inversion process that is based on the Euler method and investigate a better numerical solver for improved inversion stability. Modeling the inversion process as an ordinary differential equation (ODE) problem, the implicit backward Euler method is well suited as its solution results in an exact reconstruction using Euler's method, assuming the same time steps are used. Given that, we propose an Accelerated Iterative Diffusion Iteration (AIDI) method that adopts a fixed-point iteration process for each inversion step. Combined with Anderson acceleration [5] method which helps convergence of this iteration, it is demonstrated through experiments of large test set that it results in significantly im-
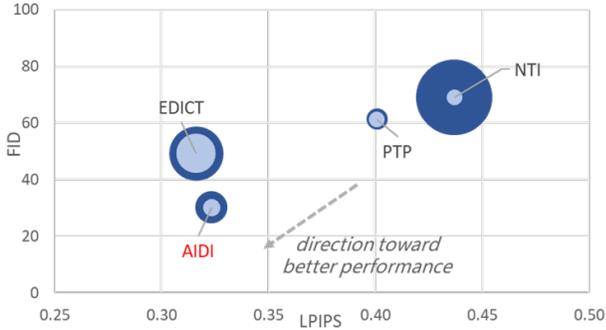
Figure 1. Quantitative assessment for various diffusion inversion based methods, using a challenging image editing task of swapping dog to cat in AFHQ test set. Image editing quality are assessed jointly by LPIPS and FID. Lower LPIPS score is preferred for perception similarity with original image and lower FID in reference to the AFHQ cat train set translates to better image editing quality. Circle areas and their outer ring represent average latency time for editing and inversion respectively.

proved reconstruction accuracy. Alternatively, an empirical acceleration method is invented for equivalent performance with less computational overhead.

While a large classifier-free guidance [22] scale is often needed for effective image editing, inversion with the same large guidance scale is not reliable even for our proposed AIDI. We have demonstrated nevertheless that it is possible to apply different guidance scales for inversion and editing respectively and still achieve effective editing using our proposed blended guidance strategy. Inspired by the cross-attention map replacement method proposed in PTP [19], we utilize the cross-attention map from the image reconstruction using the same guidance setting of inversion to blend different guidance scales in image editing. Higher guidance scales up to 7 are applied for pixels that require more editing and the low scale used in inversion, default as one, is used for irrelevant pixels.

As shown in Fig. 1, using a challenging image editing task that swaps dogs in high-resolution AFHQ [9] images to cats, our method is the best overall in terms of both editing quality, evaluated with the FID [20] score as related to the target cat domain in AFHQ, and perceptual similarity, evaluated using the LPIPS [47] metric in reference to the input image. Fig. 1 also shows the average latency time as proportional circular areas. Here the number of function evaluations (NFE) is not used because it is not an accurate metric of computational complexity for all methods. For example, in NTI there is substantial overhead in back propagation caused by learning the null-text embeddings in the inversion process. Here we fix the number of diffusion steps to assess average latency time instead, as it is correlated with image editing quality for all methods. Note that a small number of 20 inversion and editing steps is used for all methods here. As the inversion process is only needed once

per image while editing could be applied multiple times, the latency time is split to editing and inversion, represented as inner circle and outer ring respectively. For latency times, ours is only slower in inversion than the original PTP while as fast as PTP and NTI in editing.

In summary, based on pretrained text-to-image diffusion models, we propose a framework for text-based real image editing with the following key advantages:

- We are the first to our knowledge to apply fixed-point iteration and acceleration in diffusion inversion, showing significant improvements in reconstruction accuracy based on a large 5000 image COCO test set. The LPIPS value for a 20-step reconstruction is reduced to 0.063 compared to 0.148 for the baseline.

- For inversion without classifier-free guidance or low guidance scale, we propose a blended guidance method to apply larger guidance scales for effective editing in relevant areas, while maintaining fidelity elsewhere with low scales.

- Our proposed image editing method is still effective for inversion steps as low 10, where competing approaches exhibit significant artifacts.

2. Related Works

2.1. Text-to-Image Diffusion Models

The rapid progress of Diffusion-based generative models has advanced the state-of-the-art for many generative tasks, including text-to-image generations. A highly capable unconditional diffusion model was shown in [12], using sampling guidance to match the CLIP scores of the text input and generated image. More recently techniques such as GLIDE[33], DALL·E 2 [37] and Imagen[40], have used text embeddings from large language models to train conditional diffusion models, all of them capable of generating diverse and high quality images that match with the arbitrarily complex prompt texts. Both GLIDE and DALL·E 2 are conditional to CLIP textual embeddings, while DALL·E 2 trains instead a diffusion prior to first generate image embeddings from the input text CLIP embedding, before the image embedding is then fed into another diffusion model for image generation. To handle high-resolution image generation, both GLIDE and Imagen generate the text conditional image at low-resolution using cascaded diffusion models, which are conditional to both text and image to progressively increase resolution. In alternative to that, LDM [39] proposed to conduct the conditional text-to-image diffusion in a latent space of reduced dimensionality for faster training and sampling. Based on the LDM architecture, a large text-to-image model Stable Diffusion [13] was trained with a huge dataset and released for open research.

2.2. Real Image Editing in Generative Models

Since the successful implementation of disentangled representation by StyleGAN [24, 25], image editing in GAN has become very apt at separating different modes of variation such as pose, expression, color, and texture. Various methods [36, 16] have been published for text-guided image editing using the contrastive language-image model CLIP. While powerful for manipulating generated images, applying it to real image editing is not straightforward since the inversion of a real image to latent variables is trivial. Earlier GAN inversion methods focused on inversion without changing the GAN model weights, either via optimization [1, 2, 10, 23] or learning [8, 44, 38, 3]. Two recent works, HyperStyle [4] and HyperInverter [15], introduced hypernetworks [18] to modify the GAN weights for each image and help recovering lost details during reconstruction. It was shown that the modified weights have no adversarial effects on generated image quality when the inverted latent variables are modified for image editing. However, even with modified weights, GAN inversion is far from perfect due to the significantly reduced dimensionality of the latent space as compared to the original image pixel space.

Diffusion models are not subject to this limitation since there is no dimensionality changes between input and output for each inversion step. Earlier image editing methods using diffusion models, including SDEdit [31] and blended diffusion [6], didn't utilize the potential of accurate diffusion inversion as they inject random noises into input image to synthesize a noisy start. Another recent work Imagic [26] achieved mask-less real image editing without involving diffusion inversion, utilizing optimized textual embedding and diffusion model fine tuning instead.

DiffusionCLIP [27] was the first to adopt the more accurate step-by-step diffusion inversion process but it relied on diffusion model refinement to achieve text-guided image editing, without addressing the inversion accuracy challenge directly. DiffEdit [11] also avoided the inaccuracy concern by controlling an encoding ratio and applying a generated mask. Prompt-to-Prompt (PTP) [19] was the first to achieve comprehensive text guided image editing without diffusion model refinement, including local editing without a known mask. However it focused on generated image editing, citing that the simple step-by-step inversion is not reliable for real images, especially with larger classifier-free guidance scales. Null-text inversion (NTI) [32] proposed to change the constant null-text embedding to image-specific optimal ones in order to achieve accurate reconstruction and it then applies PTP techniques for real image editing. Later EDICT [45] proposed to use an auxiliary diffusion branch to achieve exact inversion and reconstruction of the input image, resulting in improved image editing quality. Most recently, pix2pix-zero [35] learns editing directions in the textual embedding space for corresponding image transla-

tion tasks but it adopts the original diffusion inversion process used in PTP without making efforts to avoid the known inversion divergence. To our knowledge, our paper is the first to address the inversion accuracy challenge without change in model configuration or system architecture so the proposed iterative inversion could be applied to and benefit other methods like pix2pix-zero that do not address the inversion accuracy issue yet.

3. Proposed Method

3.1. Diffusion Inversion Preliminaries

Our proposed generative compression method is built on pre-trained, unmodified text-to-image diffusion models. Without loss of generality, the publicly available Stable Diffusion model [13], which uses the latent diffusion model (LDM) [39] architecture, is adopted for all experiments. Given the LDM architecture, the diffusion process is conducted in the latent space z, which can be decoded to the image space x. Nevertheless our method is equally applicable to other diffusion models conducted in native image space instead of a latent space.

For a text-to-image diffusion model learned from a large set of paired image latent variable z and image caption p, the process is optimized using a simple standard mean-squared error (MSE) loss:

$$L^{\text{simple}} = E_{t,z,p,\epsilon}||\epsilon - \boldsymbol{\epsilon}_\theta(z_t, p, t)||^2 \tag{1}$$

where ϵ is a random Gaussian noise added to z to synthesize z_t, t is a randomly-set time-step chosen from $1, 2, \ldots, T$, and $\boldsymbol{\epsilon}_\theta$ denotes the diffusion model trained to estimate the injected noise using optimized parameters θ. As the goal of our proposed iterative inversion is accurate reconstruction, here we use a DDIM sampler where the sampling noise is $\sigma_t = 0$ for all t to achieve a deterministic sampling process. The iterative sampling step to generate an image latent z_0 from a random sample z_T is:

$$\begin{aligned} z_0^t &= (z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_t)/\sqrt{\bar{\alpha}_t} \\ z_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}z_0^t + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_t \end{aligned} \tag{2}$$

ϵ_t is typically calculated as follows:

$$\epsilon_t = \omega\boldsymbol{\epsilon}_\theta(z_t, p, t) + (1 - \omega)\boldsymbol{\epsilon}_\theta(z_t, \emptyset, t) \tag{3}$$

where ω is the classifier-free guidance scale and $\emptyset$ is the null-text reference.

Early image editing works like SDEdit use the same noise injection step used in model training for inversion at any t. That is:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \tag{4}$$

Since ϵ is a random noise independent of z_0, this is a stochastic inversion process that cannot reliably reconstruct
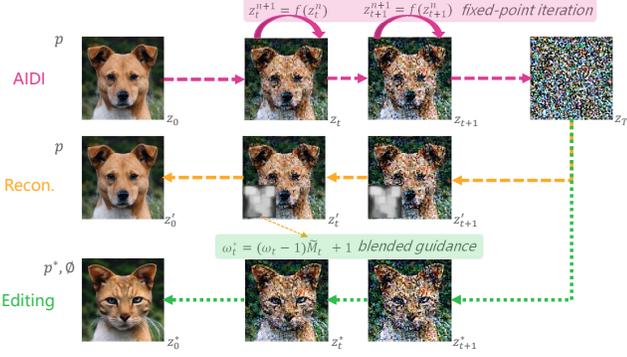
Figure 2. Flowchart of the proposed effective real image editing. From top to bottom, (i) the image is transformed to an inverted noise vector using AIDI; the inverted noise vector is used to either (ii) reconstruct the original image using the same prompt p or (iii) generate an edited image using prompts $p^*, \emptyset$ with classifier-free guidance, where the reconstruction process is also used to supply the mask for blended guidance and attention injection. Note that the visible noise is not Gaussian, as all images are decoded from latent space z to image space x for display.

z_0 from z_t when t is close to T. In later works, a simple DDIM inversion process is adopted base on a linear ODE solver:

$$z_0^t = (z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_{\tilde{t}})/\sqrt{\bar{\alpha}_t}$$
$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}}z_0^t + \sqrt{1 - \bar{\alpha}_{t+1}}\epsilon_{\tilde{t}}. \tag{5}$$

To allow for better reconstruction accuracy, here $\epsilon_{\tilde{t}}$ is instead approximated using $t+1$ as follows:

$$\epsilon_{\tilde{t}} = \omega\epsilon_\theta(z_t, p, t+1) + (1-\omega)\epsilon_\theta(z_t, \emptyset, t+1). \tag{6}$$

3.2. Accelerated Iterative Diffusion Inversion

For an inversion step where z_{t-1} is known, we aim to find the optimal z_t so that we can recover z_{t-1}. We can rewrite Equation 2 as:

$$z_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}z_{t-1} + \left[\sqrt{1 - \bar{\alpha}_t} - \sqrt{\frac{(1 - \bar{\alpha}_{t-1})\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\right]\epsilon_t. \tag{7}$$

Since ϵ_t depends on z_t (cf. Equation 3), it can be denoted as an implicit function $z_t = f(z_t)$. The ideal inversion step, finding z_t that results in z_{t-1} exactly, becomes a step to find a fixed-point solution for f. In numerical analysis, this is often solved via the iterative process:

$$z_t^{n+1} = f(z_t^n), n = 0, 1, \dots \tag{8}$$

The convergence of this iterative process can often be accelerated using established techniques such as Anderson acceleration. We also propose and employ an alternative empirical acceleration method which is simpler and faster than

---

Algorithm 1: Accelerated Iterative Diffusion Inversion

1 **Input:** A latent image z_0 and prompt p, acceleration method C, iteration parameters I, m.
2 **Function:** $f(z)$ is the implicit function defined in Equation 7, $g(z)$ is the residual function $f(z) - z$
3 **Output:** An inverted noise vector z_T.

---

4 **for** $t = 1, 2, \dots, T$ **do**
5 $\quad z_t^0, z_t^1 \leftarrow z_{t-1}, f(z_t^0)$;
6 $\quad$ **for** $i = 1, \dots, I$ **do**
7 $\quad\quad$ **if** C *is* AIDI_A **then**
8 $\quad\quad\quad m_i \leftarrow min(m, i)$;
9 $\quad\quad\quad G_i \leftarrow [g(z_t^{i-m_i}), \dots, g(z_t^i)]$;
10 $\quad\quad\quad \gamma_i \leftarrow \arg\min_{\gamma \in \Gamma_i} \|G_i \cdot \gamma\|_2$, where $\Gamma_i = \{(\gamma^0, \dots, \gamma^{m_i}) : \sum_{j=0}^{m_i} \gamma^j = 1\}$;
11 $\quad\quad$ **else if** C *is* AIDI_E **then**
12 $\quad\quad\quad m_i, \gamma_i \leftarrow 1, (0.5, 0.5)$;
13 $\quad\quad$ Set $z_t^{i+1} \leftarrow \sum_{j=0}^{m_i} \gamma_i^j f(z_t^{i-m_i+j})$;
14 $\quad$ **end**
15 $\quad z_t \leftarrow z_t^I$;
16 **end**
17 **Return** z_T

Anderson's. With appropriate acceleration, the iterative inversion process can become more efficient and stabler than a simple forward Euler method. We summarize our full proposed process in Alg. 1, and refer to it as *AIDI*, short for accelerated iterative diffusion inversion. Note that for the residual function $g(z)$ used in the AIDI_A variant with Anderson acceleration, it is defined as $g(z) = f(z) - z$. For the AIDI_E variant, it is a simplified version of Anderson acceleration with fixed setting for m and γ as 1 and $(0.5, 0.5)$, saving the additional optimization process to find γ.

3.3. Blended Guidance

While the proposed AIDI can significantly improve the inversion stability at different guidance scales, it is not sufficient on its own for reliably reconstructing and then eventually editing an image when the guidance scale is large. Adopting the PTP pipeline for real image editing, we first fix a small guidance scale like $\omega = 1$ for the inversion using source prompt p. For the following editing, we establish two parallel sampling processes, one conditional to prompt p with same small ω and the other conditional to the target prompt p^*. For the p^* process, we introduce a blended ω^* to apply larger guidance scales for pixels relevant to editing and lower ones for the rest to keep them unedited. To support this blended guidance, a soft editing mask $\tilde{M}_t$ is generated concurrently with the image editing process. First the cross-attention map M_t with an anchor to-
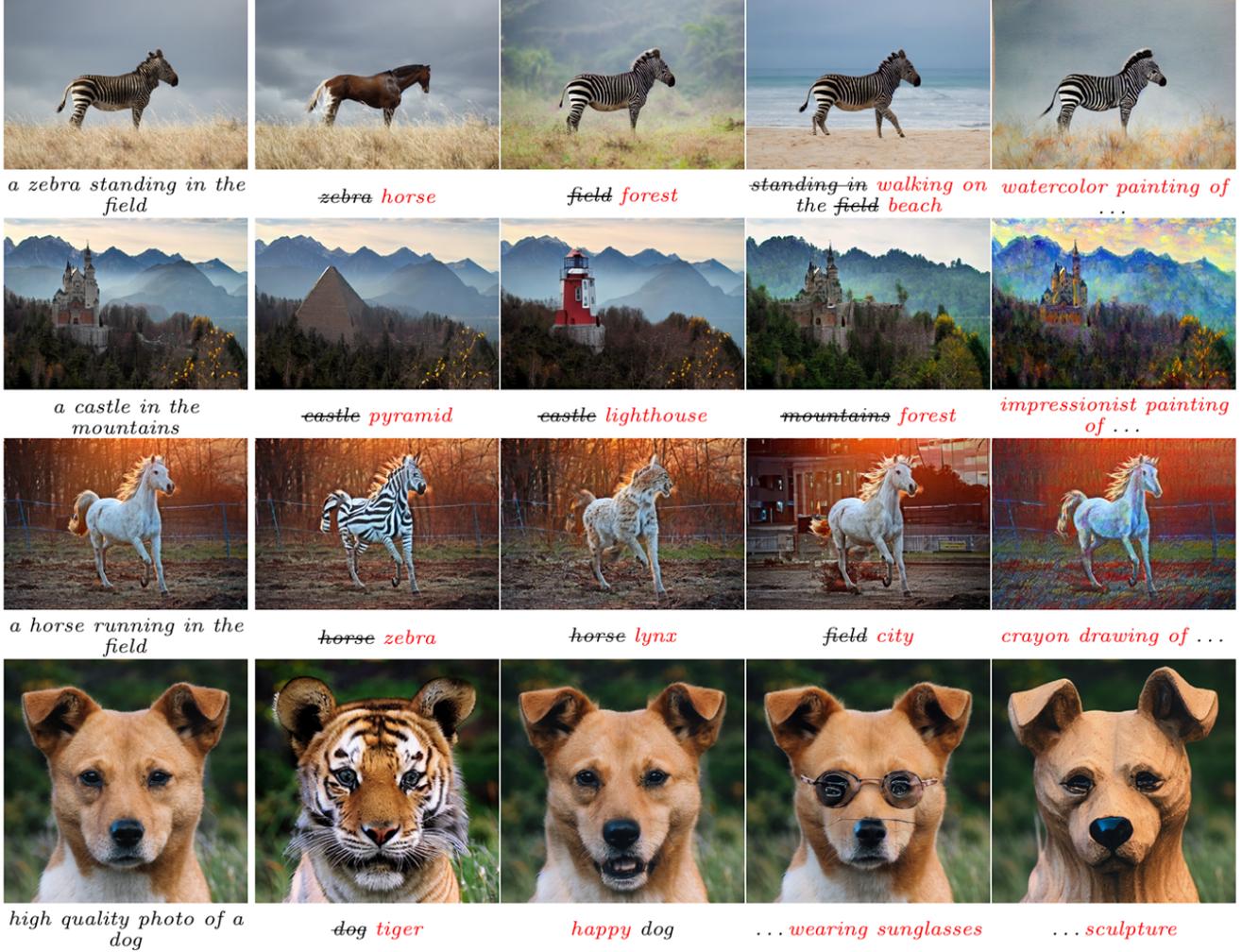
Figure 3. Visual examples of the effective editing abilities of our proposed real image editing based on AIDI, using only 20 editing steps.

ken is determined. Here the anchor token refers to the word most relevant to the intended editing area. This token could be positive, *i.e.* associating with areas to edit, or negative, which associates it with areas to keep unedited. In the case of *photo of a dog → photo of a cat*, *dog* is a positive token. In contrast, for *a dog → a dog on the beach*, *dog* is a negative one instead. The mask M_t is first normalized to $\overline{M}_t$, where all pixels smaller than a threshold δ are normalized to the range of $(-M, 0)$ and the others normalized to the range of $(0, M)$. Then a soft mask $\tilde{M}_t$ is defined as $Sigmoid(\overline{M}_t)$ for a positive token and $Sigmoid(-\overline{M}_t)$ for a negative one. Given that, the blended guidance scale ω^* is defined as

$$\omega_t^*(k) = (\omega_E - \omega)\tilde{M}_t(k) + \omega \qquad (9)$$

where ω_E is a large guidance scale intended for editing, and k refers to pixels of the mask image. Note that the soft mask approaches a binary one when M becomes very large. Combining our proposed AIDI and blended guidance, the overall process for our effective real image editing is illus-

trated in Fig. 2, assuming $\omega = 1$ in inversion for simplicity.

3.4. Stochastic Editing

For image editing methods based on deterministic DDIM inversion methods, the DDIM sampling process is commonly set as deterministic to better control editing effects. Similar to the proposed blended guidance, the same soft mask is adopted to control the affected area of stochastic sampling. The deterministic sampling process described in Equation 2 is made stochastic as:

$$q(z_{t-1}) \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}z_0^t + \sqrt{1 - \bar{\alpha}_{t-1} - \eta\sigma_t^2\tilde{M}_t}\epsilon_t, \eta\sigma_t^2\tilde{\mathbf{M}}_t) \quad (10)$$

where σ_t^2 represents the noise injected for stochastic sampling, and $\tilde{\mathbf{M}}_t$ is transformed from $\tilde{M}_t$ as a diagonal matrix. Here η and $\tilde{M}_t$ control the scale and range of stochastic editing respectively. While large η increases sampling diversity, it reduces the perceptual similarity with the original image which is often desired in cases such as replacing

| AE | PTP | AIDI_E | AIDI_A | NTI | EDICT |
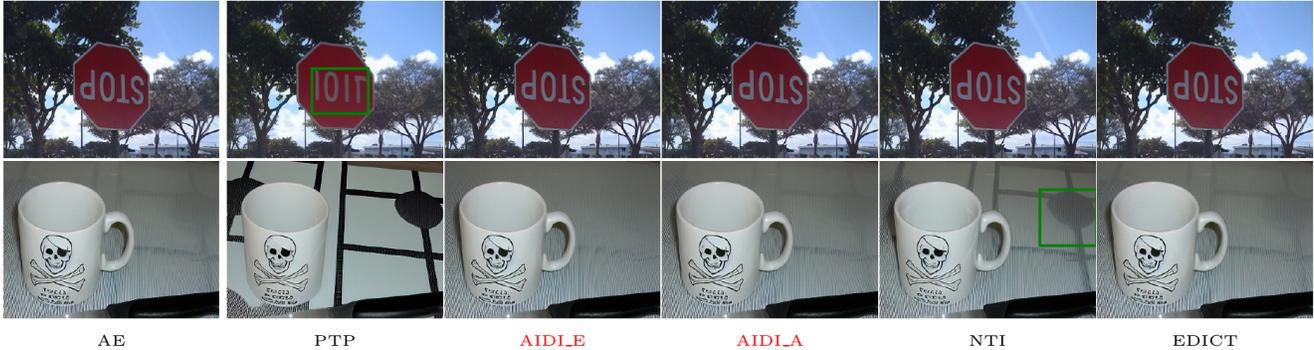
Figure 4. Visual examples of reconstruction accuracy test for various diffusion inversion methods. The AE images on the left are decoded from Stable Diffusion without inversion and used as the reference for perfect reconstruction. Selected local artifacts are highlighted with □ for quick reference.

an object with another in a different domain. Based on our experiments, a small η can result in satisfying results without loosing perceptual similarity when deterministic editing fails.

4. Experiments

All experiments conducted here are based on the released version `v1.4` of Stable Diffusion using NVIDIA A100 GPU card. For image reconstruction test, 5000 test images from COCO dataset [28] are used. For image editing, the test set of AFHQ [9] is used primarily for its high image quality and because it enables comparisons with GAN inversion-based editing (we denote it as HS-SCLIP, combining HyperStyle [4] and StyleCLIP [14]). Other than HS-SCLIP, we perform comprehensive comparisons in both reconstruction and editing quality with all related text-based real image editing techniques which use DDIM inversion, including PTP [19], NTI [32] and EDICT [45]. NTI's main innovation is the null-text inversion, so in image editing tests it uses PTP editing techniques after the inversion. For all visual examples, the default inversion and editing steps are 20 unless specified otherwise.

4.1. Image Editing

Visual examples from Fig. 3 showcase the diverse capabilities of our proposed real image editing process. Using a pair of text prompts as before and after-editing image captions, it can either perform local editing, like replacing an object or background, or global editing, like style transfer. As an example, for the *zebra* image, we are able to change the background to a *forest* while keeping the unedited area perceptually unchanged. For partial editing within one object, it can change the posture of the *zebra* from *standing* to *walking* in addition to editing the background, or change the expression of the *dog* to *happy*. It is also able to do a text-based style transfer, converting a photo to a *watercolor painting*. While similar range of editing capabilities

have been demonstrated before, we are the first to demonstrate effective editing using as few as 10 editing steps, a result enabled by the improved accuracy of AIDI and by the proposed blended guidance, which focuses editing effects to specific relevant areas. Note that without an exact binary mask, the local edits from these prompt-pair based processes would not be be able to keep certain areas intact. Consider for example the editing task of changing the background of *a castle in the mountains* to *forest*: the castle itself remains largely the same but some greenery is blended within its structure, increasing the overall coherence of the resulting image. This is arguably preferable to a hard blending with the background. The lack of a binary mask can sometimes give rise to hallucinated artifacts, such as the lines around the mouth area in the *dog wearing sunglasses* instance.

4.2. Reconstruction Accuracy

We conducted a reconstruction test using the COCO test set with diverse contents using their default captions as text prompts where the quantitative results are illustrated as a graphic table, Fig. 6. Compared to the simple inversion used in PTP, results from both our proposed AIDI variants are shown to significantly improve the accuracy across different classifier-free guidance scales and different diffusion steps. For both AIDI_E and AIDI_A, they are able to achieve near-exact inversion when there is no classifier-free guidance. While the accuracy drops when the guidance scale increases, this does not impact our proposed real image editing process since a low guidance scale like 1 is used in inversion. We confirm that EDICT is able to maintain exact inversion with as few as 10 iteration steps, although it was only tested in the 50 steps' regime at publication. As for NTI, it is only applicable to a guidance scale > 1. Despite its accuracy increasing along with the guidance scale, it is worse, even at an empirically optimal guidance scale of 3, than our proposed inversion and reconstruction with a guidance scale of 1. For visual examples in Fig 4, both PTP and
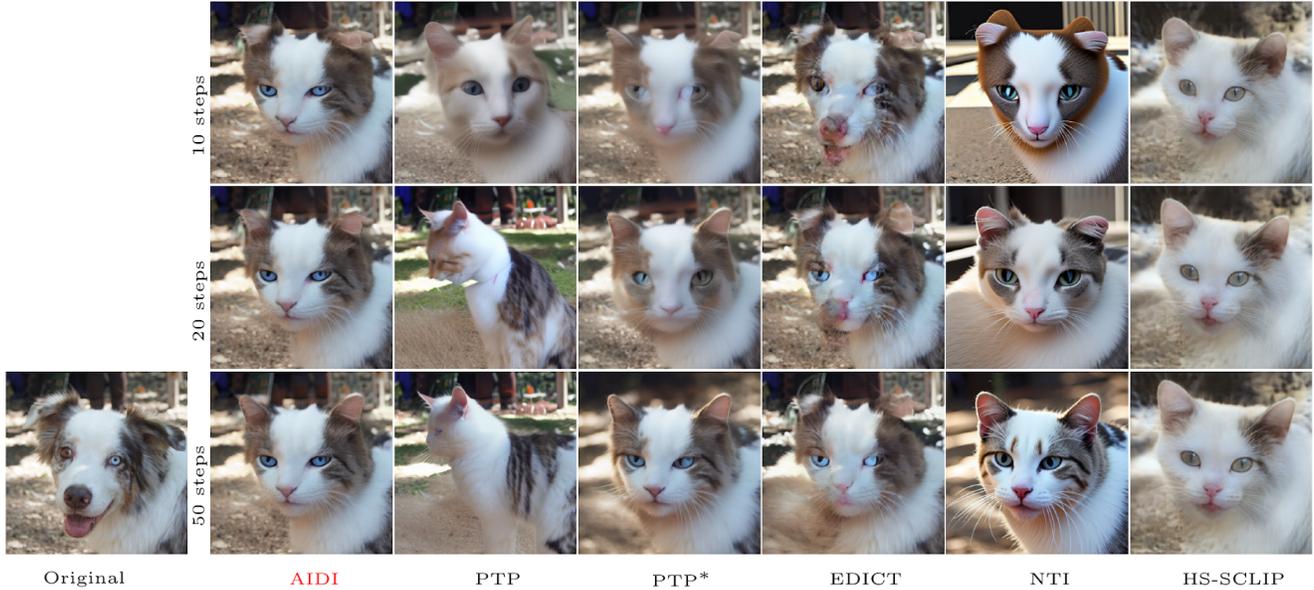
Figure 5. Visual examples for dog-to-cat editing test using AFHQ test set. Results from of different model are organized horizontally and results from different settings, 10/20/50 editing steps for all diffusion-based models or 3 hyperparameter settings for HS-SCLIP, are organized vertically.

NTI's results contain noticeable errors. Our AIDI_A result is as flawless as the EDICT one. Despite the AIDI_E result missing some minor detail, we find its differences from AIDI_A negligible from both a quantitative and qualitative perspective. Unless otherwise specified, all image editing tests are conducted using AIDI_E.

Steps	10	20	50	10	20	50	10	20	50	10	20	50	all
Scale ↓	PTP			AIDI_E			AIDI_A			NTI			EDICT
LPIPS 1	0.166	0.148	0.130	0.063	0.063	0.067	0.060	0.061	0.062				0.059
LPIPS 2	0.232	0.228	0.231	0.098	0.104	0.131	0.102	0.103	0.103	0.083	0.073	0.080	0.059
LPIPS 3	0.319	0.338	0.391	0.176	0.203	0.278	0.208	0.253	0.284	0.077	0.068	0.068	0.059
SSIM 1	0.728	0.745	0.763	0.821	0.821	0.817	0.823	0.822	0.822				0.824
SSIM 2	0.675	0.679	0.677	0.793	0.788	0.764	0.788	0.788	0.788	0.794	0.806	0.802	0.824
SSIM 3	0.604	0.586	0.539	0.735	0.705	0.635	0.699	0.658	0.630	0.800	0.811	0.814	0.824

Figure 6. Reconstruction accuracy test for different diffusion inversion methods using COCO test set. Results from combinations of different inversion steps and various classifier-free guidance scales are included. For the two assessed perceptual metrics, higher SSIM and lower LPIPS are preferred, both color-coded towards green color.

4.3. Quantitative Assessment

EDICT is the first one in diffusion inversion based image editing methods to report quantitative analysis for image editing. We have conducted similar quantitative assessments using the AFHQ dataset. Our rationale for choosing AFHQ is that the high quality close-up images do make visual artifacts more noticeable. Another advantage is that for a dog-to-cat image test, there is a real cat image set to serve as a photorealistic reference to evaluate editing results, an arguably better approach than using a CLIP score that relies on the indirect text reference. This also enables comparison with GAN inversion-based methods, since a pre-trained StyleGAN model is available for this dataset. Using the prompt pair *high quality photo of a dog* → *high quality photo of a cat*, the task is to replace the dog in the input images with a cat which not only looks realistic, but perceptually similar to the original dog in characteristics like color and pose, while maintaining the background unedited. To evaluate the consistency in color, pose and backgrounds, the LPIPS metric is evaluated between the edited and original dog image. To evaluate the editing effects, the FID score in reference to the AFHQ training set is used over CLIP score. Results for other metrics, including CLIP, are included as supplementary materials.

As shown in Fig. 7, our results are the best overall for any number of editing steps. Our result of 20 editing steps is better than the next best EDICT at 50 steps. In the extreme case of 10 steps, our method is still relatively effective while all other methods have significant regressions in quality. For PTP, in addition to the original setting where the same guidance scale is used for both inversion and editing, a PTP* version is also tested where the inversion is conducted without guidance, and it has significantly improved performance comparing to PTP. Similar improvement has also been reported in EDICT, denoted as the difference between UC PTP and C PTP. This dual-scale setting is similar to our blended guidance so the performance gap between PTP* and ours is mainly caused by the increased inversion
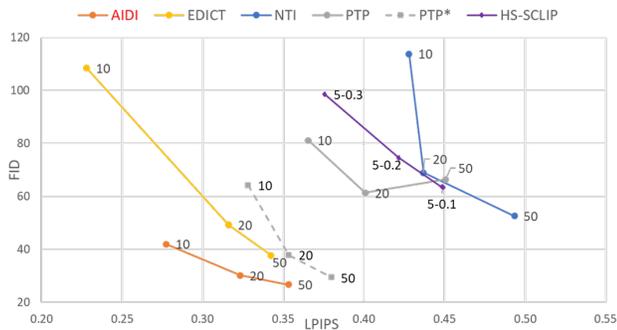
Figure 7. Quantitative assessments for dog-to-cat swap test using AFHQ test set. The data labels represent editing steps for all methods except for HS-SCLIP where they are hyperparameters.

accuracy from our AIDI. For HS-CLIP, which is based on HyperStyle inversion and StyleCLIP editing, three sets of hyperparameters are included to illustrate similar trade-off between high perceptual similarity comparing to the source image and high editing quality in reference to the target domain. To out knowledge, this is the first direct quantitative comparison between editing methods using diffusion inversion and GAN inversion respectively.

The results from NTI, PTP editing with null-text inversion, are worse than PTP baseline even with increased inversion stability. From the visual examples in Fig. 5, they seem to be able to generate a high quality cat image. However, these images are often too smooth and the lack of photorealistic details are causing degraded performance in both FID and LPIPS. One possible cause is that the learned null-text embedding deviates from the original one too much so it is out-of-distribution for the pre-trained diffusion model. Another consideration is that quantitative assessments like this require a fixed setting of hyperparameters so any method relying on fine-tuning of hyperparameters would not perform well on average.

The robustness of our methods for reduced number of inversion steps comparing to others is also noticeable in Fig. 5. For our method, although the 50-step result has better sharpness and more realistic ears, the 10-step version is of acceptable quality without any glaring artifacts, whereas all competing approaches are subject to significant degradations at 10-steps. In the case of EDICT, there are obvious artifacts, dog-like mouth area, for both 10-step and 20-step results. Note that for this experiment, we used similar grid-search strategy to find optimal settings for our method, PTP, PTP* and NTI, as we adopted attention map injection techniques from PTP. For EDICT, its own default optimal settings are used. For fair comparison, we didn't change hyperparameters for different editing steps, except for the fixed-point iteration steps in AIDI where 11, 6 and 5 are used for 10/20/50 editing steps respectively.

4.4. Stochastic Editing

When using deterministic DDIM inversion, image editing through the DDIM sampling process must also be set deterministic in order to keep the fidelity of unedited areas. As a result, there is no direct remedy for failed editing results. As shown in the middle image in Fig. 8, the original over-exposed background window was incorrectly edited as part of the cat face. Using our proposed stochastic editing, it is possible to improve editing results when such failure case happens. Different from other stochastic sampling practices which uses a larger η for increased sampling diversity, we used a small $\eta = 0.1$ for stochastic generation of the same editing task and resulted in a properly edited result as shown on the right. We conducted a quantitative analysis using the dog-to-cat test to measure this benefit. As shown in the chart in Fig. 8, $20 \times n$ represents average performance when n stochastic editings were repeated for each image to select the best one based on a combined rank of low FID and LPIPS. It is shown that the stochastic editing, when tested only once, is expected to have an equivalent FID score as the deterministic one but worse in LPIPS. With increased n, the selected result is expected to further improve in both FID and LPIPS.

Original | Failure case of deterministic editing | Corrected by stochastic editing
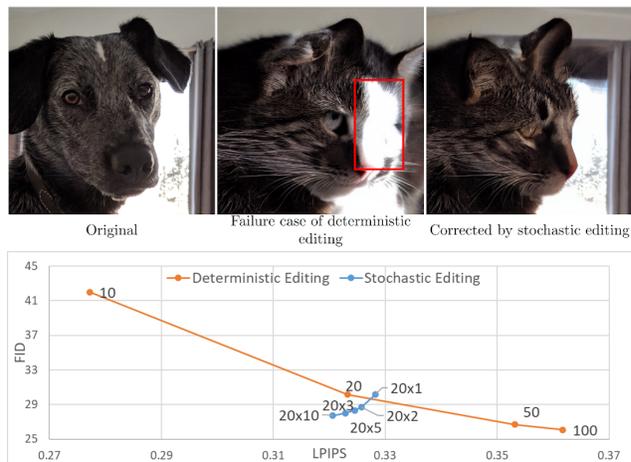
Figure 8. Top: visual examples of stochastic editing recovers from failure case of deterministic editing; Bottom: quantitative comparison between deterministic and stochastic editing.

5. Conclusions

In this work, we have proposed *AIDI*, an accelerated iterative diffusion inversion method capable of improving image reconstruction and editing accuracy while significantly improving the trade-offs between computational cost and quality. While our proposed real image editing is effective without auxiliary masks or sketches to specify editable area and content, it relies on the cross-attention maps between the image and prompt texts for general guidance, similar to

other text guided image editing work using diffusion models. In addition to logically set prompt pairs, the spatial resolution of these attention maps are very coarse. Detailed control of editable area remains a subject of future work.

It is also noted that while we don't use large classifier-free guidance scales for inversion with introduction of blended guidance, reducing the difference in guidance scales between inversion and editing is still useful to reduce potential artifacts caused by blended guidance. While our AIDI can improve inversion stability for large guidance scales as well, it is not sufficient for reliable inversion yet and this is also of future research interests.

Ethical considerations: we acknowledge that any image editing technique such as the one presented in this paper are unavoidably encumbered by ethical concerns, and will inherit any bias present in the training data from the underlying backbone. Deployments of these systems should employ appropriate safeguards in regards to allowed prompts and guidances to prevent malicious and illegal uses.

References

[1] 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*, pages 4432–4441, 2019. 3

[2] 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*, pages 8296–8305, 2020. 3

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. ReStyle: A residual-based StyleGAN encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 3

[4] 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*, pages 18511–18521, 2022. 3, 6

[5] Donald G Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4):547–560, 1965. 1

[6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 3

[7] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 1

[8] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in GANs. 2021. 3

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2, 6

[10] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 3

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[12] Katherine Crowson, Maxwell Ingham, Adam Letts, and Alex Spirin. Disco Diffusion. https://github.com/alembics/disco-diffusion, 2022. 2

[13] Katherine Crowson, Maxwell Ingham, Adam Letts, and Alex Spirin. Stable diffusion. https://github.com/CompVis/stable-diffusion, 2022. 2, 3

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 6

[15] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. HyperInverter: Improving StyleGAN inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. 3

[16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[18] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 3

[19] 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. 1, 2, 3, 6

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[23] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. GAN inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021. 3

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[26] 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. 3

[27] 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*, pages 2426–2435, 2022. 1, 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[29] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 1

[30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1

[31] 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. 1, 3

[32] 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. 1, 3, 6

[33] 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*, 2022. 2

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021. 1

[35] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 3

[36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[37] 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. 2

[38] 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*, pages 2287–2296, 2021. 3

[39] 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*, pages 10684–10695, 2022. 2, 3

[40] 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. 2

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[42] 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*, 2020. 1

[43] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2023. 1

[44] 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. 3

[45] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. 1, 3, 6

[46] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey.(2021). *arXiv preprint arXiv:2101.05278*, 2021. 1

[47] 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. 2