Eta Inversion: Designing an Optimal Eta Function for Diffusion-based Real Image Editing

Wonjun Kang*[1], Kevin Galim*[1], and Hyung Il Koo†[1,2]

[1] FuriosaAI, Seoul 06036, South Korea
{kangwj1995,kevin.galim,hikoo}@furiosa.ai
[2] Ajou University, Suwon 16499, South Korea
Code: https://github.com/furiosa-ai/eta-inversion

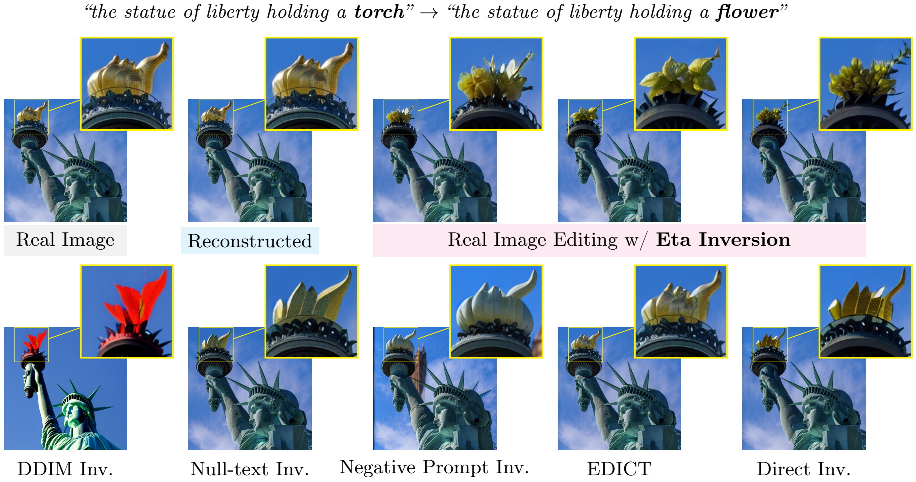

*"the statue of liberty holding a **torch**"* → *"the statue of liberty holding a **flower**"*
Real Image
Reconstructed
Real Image Editing w/ **Eta Inversion**
DDIM Inv.
Null-text Inv.
Negative Prompt Inv.
EDICT
Direct Inv.

Fig. 1: Eta Inversion for real image editing. We design an optimal time- and region-dependent η function for DDIM sampling [39] for superior results. In the example above, existing methods fail to change the torch into a flower or do not preserve the structure, while Eta Inversion creates various plausible results. Tested with PtP [13].

Abstract. Diffusion models have achieved remarkable success in the domain of text-guided image generation and, more recently, in text-guided image editing. A commonly adopted strategy for editing real images involves inverting the diffusion process to obtain a noisy representation of the original image, which is then denoised to achieve the desired edits. However, current methods for diffusion inversion often struggle to produce edits that are both faithful to the specified text prompt and

* Authors contribute equally. † Corresponding author.

closely resemble the source image. To overcome these limitations, we introduce a novel and adaptable diffusion inversion technique for real image editing, which is grounded in a theoretical analysis of the role of η in the DDIM sampling equation for enhanced editability. By designing a universal diffusion inversion method with a time- and region-dependent η function, we enable flexible control over the editing extent. Through a comprehensive series of quantitative and qualitative assessments, involving a comparison with a broad array of recent methods, we demonstrate the superiority of our approach. Our method not only sets a new benchmark in the field but also significantly outperforms existing strategies.

Keywords: Diffusion Models · Diffusion Inversion · Real Image Editing

1 Introduction

Text-guided image synthesis [5,6,18,26,33,34,38] is one of the essential tasks in computer vision due to its enormous potential for design and art industries. Recent breakthroughs in diffusion models [9,14,31,36,39] drastically increased text-to-image generation performance. Due to the success of diffusion-based image generation, text-guided image editing with diffusion models is also gaining interest in the research community [2,8,13,22,29,41]. However, editing a real image is challenging and existing methods still struggle to produce consistent high-quality results, yet insufficient to the industry's high demand and interest.

Given a source image, a source prompt describing that image, and a target prompt describing the desired output image, it is possible to invert the diffusion process for the source image and edit the inverse latent according to the target prompt. Similar to GAN inversion [35,44], diffusion inversion seeks to identify the latent noise corresponding to a particular image. Unlike GANs [11], which require a single generation step, diffusion models require many iterative steps, making inversion more challenging.

Despite recent advancements in diffusion inversion [24,39,42] and editing methods [2,13,41], proper quantitative evaluation is lacking, particularly studies on all combinations of these techniques. We address this gap by reformulating and integrating existing strategies within a single framework, categorizing existing methods into two distinct groups: perfect reconstruction methods and imperfect reconstruction methods. Using this framework, we conduct a thorough evaluation of all methods under consistent and fair conditions, employing a variety of metrics.

Unlike previous methods that use a fixed η value, such as 0 or 1, in the DDIM [39] sampling equation, our research explores whether a dynamic η function is superior. Consequently, we analyze the role of η in diffusion inversion and propose Eta Inversion, a perfect reconstruction method. Eta Inversion utilizes a time- and region-dependent η to introduce optimal noise during the backward process, achieving better editing diversity. To our knowledge, we are the first to investigate an optimal time-dependent η function to balance editing extent and source image similarity for improved performance. To prevent modifications to the background of the image, we make η region-dependent, applying $\eta > 0$ only to specific object regions based on their cross-attention map. Comprehensive experiments validate our findings, demonstrating state-of-the-art performance both quantitatively and qualitatively. Our contributions are:

- We formulate a generalized framework for diffusion inversion methods.
- We formally explore the role of η in diffusion inversion and real image editing.
- We design a time- and region-dependent η function to inject optimal real noise and achieve state-of-the-art performance in diffusion inversion.
- We provide an extensive benchmark for diffusion inversion by evaluating existing inversion methods using various image editing methods.

2 Related Work

2.1 Diffusion Models for Image Generation and Editing

Diffusion models offer more stable training and better diversity than GANs [11], making them a common choice for image generation. Denoising Diffusion Probabilistic Models (DDPM) [14] showcased the capabilities of diffusion models but require about 1000 inference steps for quality images. Denoising Diffusion Implicit Models (DDIM) [39] improve this by reducing inference steps to 50, removing stochastic elements from DDPM sampling. Although rooted in Variational Inference, diffusion models can also be viewed as score-based models using Stochastic Differential Equations (SDEs) [40]. Latent Diffusion Models [36] perform denoising in compressed latent space, greatly reducing inference cost and time. Stable Diffusion [36] has become a standard for text-to-image generation due to its public availability and impressive performance.

Text-guided image editing methods [1,2,8,13,29,41] aim to align an image with a target prompt while maintaining its original structure. We focus on methods that require no additional training or optimization for better flexibility. Prompt-to-Prompt (PtP) [13] edits images by injecting cross-attention maps from the source into the target prompt's denoising process. Similarly, Plug-and-Play (PnP) [41] not only injects cross-attention maps but also integrates spatial features. Furthermore, MasaCtrl [2] focuses on motion editing and employs self-attention maps instead of cross-attention maps.

2.2 Diffusion Inversion Methods

To perform real image editing, a noisy image or latent representation must first be obtained via diffusion inversion. DDIM Inversion [39] achieves low error reconstruction in an unconditional image generation setting, but classifier-free guidance [15] leads to significant differences from the input image.

To address this, Null-text Inversion (NTI) [24] optimizes the null-text embedding $\varnothing_t$ for each timestep, reducing the inversion gap but adding computational overhead. Negative Prompt Inversion (NPI) [23] replaces the null-text with the source text embedding, providing a fast, inference-only inversion pipeline.

ProxNPI [12] enhances NPI with regularization and reconstruction guidance, improving accuracy with minimal cost. EDICT [42] achieves exact inversion via an auxiliary diffusion path but doubles inference time. DDPM Inversion [16] and CycleDiffusion [43] use stored variance noise from the forward path for exact inversion, but the non-normal distribution of this noise affects editing performance. Direct Inversion [17] preserves similarity to the source image by replacing latents during denoising with those from the DDIM Inversion forward path, though this may limit the extent of editing. They also provided a dataset for editing evaluation.

Unlike previous methods that use a static η value, our contribution lies in enhancing editability by designing an optimal dynamic η function, an aspect not previously explored. Furthermore, we are the first to employ real noise injection for real image editing. This innovation allows us to optimally add real Gaussian noise during editing with minimal inference overhead, achieving balanced and precise image editing.

Table 1: Table of notation.

$⬚_t$:	⬚ at timestep t	🟨 :	noise prediction	$\boldsymbol{\epsilon}_{t,\theta}$:	noise prediction network	q_t :	marginal distribution of Eq. (2)
$⬚^{(s)}$:	⬚ of source	🟫 :	sampling	$\boldsymbol{s}_{t,\theta}$:	score estimation network	p_{t,η_t}:	marginal distribution of Eq. (6)
$⬚^{(t)}$:	⬚ of target	🟪 :	editing	α_t :	noise schedule	$\mathcal{M}_t$:	attention map
$⬚^*$:	inverted ⬚	🟩 :	forward path	$\bar{\alpha}_t$:	$\prod_{i=1}^{t} \alpha_i$	$\boldsymbol{w}$:	standard Wiener process (forward)
$⬚'$:	reconstructed ⬚	🟦 :	backward path	$\boldsymbol{\epsilon}_{\text{add}}$:	additional noise $\sim \mathcal{N}(0, I)$	$\bar{\boldsymbol{w}}$:	standard Wiener process (backward)

3 Preliminaries

3.1 Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) [14] are generative models consisting of a noising forward path and a denoising backward path. During the forward path, Gaussian noise $\boldsymbol{\epsilon}$ is gradually added to the sample data point. DDPM's backward path consists of a noise prediction step and a sampling step. Denoising Diffusion Implicit Models (DDIM) [39] are an extended version of DDPM which escape from the Markovian forward process. The general form of the sampling function of DDIM is given as below where $\boldsymbol{\epsilon}_t$ is the estimated noise at timestep t for latent $\boldsymbol{x}_t$, computed as $\boldsymbol{\epsilon}_t \leftarrow \boldsymbol{\epsilon}_{t,\theta}(\boldsymbol{x}_t)$:

$$\text{Sample}(\boldsymbol{x}_t, \boldsymbol{\epsilon}_t, \eta_t) = \sqrt{1/\alpha_t}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\boldsymbol{\epsilon}_t + \sigma_t \boldsymbol{\epsilon}_{\text{add}}. \quad (1)$$

σ_t is defined as $\sigma_t = \eta_t\sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$, and $\eta_t \geq 0$ is a controllable hyperparameter. DDPM is a special case of DDIM where $\eta_t = 1$ for all t, whereas DDIM sampling uses $\eta_t = 0$, making the sampling procedure deterministic. For conditional image generation such as text-to-image generation, the noise estimation network receives an additional conditional input $\boldsymbol{c}$ as

$\boldsymbol{\epsilon}_{t,\theta}(\boldsymbol{x}_t, \boldsymbol{c})$. However, it has been empirically shown that the above conditioning is insufficient for reflecting text conditions, so classifier-free guidance [15] is usually used to amplify the text condition as $\tilde{\boldsymbol{\epsilon}}_{t,\theta} = w \cdot \boldsymbol{\epsilon}_{t,\theta}(\boldsymbol{x}_t, \boldsymbol{c}) + (1-w) \cdot \boldsymbol{\epsilon}_{t,\theta}(\boldsymbol{x}_t, \varnothing)$, where w is the guidance scale parameter and $\varnothing$ is the empty prompt. We can summarize the text-to-image generation procedure as Noise Prediction $\boldsymbol{\epsilon}_t \leftarrow \tilde{\boldsymbol{\epsilon}}_{t,\theta}(\boldsymbol{x}_t, \boldsymbol{c}, \varnothing, w = 7.5)$ and Sampling $\boldsymbol{x}_{t-1} \leftarrow \mathrm{Sample}(\boldsymbol{x}_t, \boldsymbol{\epsilon}_t, \eta_t = 0)$ and simplify them as $\boldsymbol{x}_{t-1} \leftarrow \mathrm{DDIM}(\boldsymbol{x}_t, \boldsymbol{c}, \varnothing, w, \eta_t)$.

3.2 Score-based Models

Eq. (2) and Eq. (3) are the forward and backward SDE of score-based models corresponding to the forward and backward path of DDPM [40]. Eq. (4) is the probability flow ODE and corresponds to DDIM ($\eta = 0$) sampling [39, 40]. Eq. (5) is the extended version of the backward SDE which has the same marginal distribution q_t for any $\eta \geq 0$, and DDIM sampling (Eq. (1)) is a numerical method of Eq. (5) [46]. Similarly, we can train a score function $\boldsymbol{s}_{t,\theta}(\boldsymbol{x}) = -\boldsymbol{\epsilon}_{t,\theta}(\boldsymbol{x})/\sqrt{1-\alpha_t} \approx \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ and apply a numerical method to Eq. (6).

$$\mathrm{d}\boldsymbol{x} = f_t \boldsymbol{x}\,\mathrm{d}t + g_t\,\mathrm{d}\boldsymbol{w}, \quad \left(f_t = \frac{1}{2}\frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}, g_t = \sqrt{-\frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}}\right) \tag{2}$$

$$\mathrm{d}\boldsymbol{x} = [f_t \boldsymbol{x} - g_t^2 \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})]\,\mathrm{d}t + g_t\,\mathrm{d}\bar{\boldsymbol{w}} \tag{3}$$

$$\mathrm{d}\boldsymbol{x} = [f_t \boldsymbol{x} - 0.5 g_t^2 \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})]\,\mathrm{d}t \tag{4}$$

$$\mathrm{d}\boldsymbol{x} = [f_t \boldsymbol{x} - 0.5(1+\eta_t^2) g_t^2 \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})]\,\mathrm{d}t + \eta_t g_t\,\mathrm{d}\bar{\boldsymbol{w}} \tag{5}$$

$$\mathrm{d}\boldsymbol{x} = [f_t \boldsymbol{x} - 0.5(1+\eta_t^2) g_t^2 \boldsymbol{s}_{t,\theta}(\boldsymbol{x})]\,\mathrm{d}t + \eta_t g_t\,\mathrm{d}\bar{\boldsymbol{w}} \tag{6}$$

3.3 DDIM Inversion

DDIM Inversion [39] is an important technique for real image editing and can be derived from DDIM ($\eta = 0$) sampling (Eq. (1)) by approximating $\boldsymbol{\epsilon}_t \approx \boldsymbol{\epsilon}_{t+1}$:

$$\boldsymbol{x}_{t+1} = \sqrt{\bar{\alpha}_{t+1}/\bar{\alpha}_t}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}}\left(\sqrt{1/\bar{\alpha}_{t+1} - 1} - \sqrt{1/\bar{\alpha}_t - 1}\right)\boldsymbol{\epsilon}_t. \tag{7}$$

DDIM Inversion can be written as $\boldsymbol{x}_{t+1} \leftarrow \mathrm{DDIM}_{\mathrm{inv}}(\boldsymbol{x}_t, c, \varnothing, w)$. With $w = 1$, it encodes latent noise with negligible reconstruction error, but large w values (e.g., $w = 7.5$ in Stable Diffusion) result in significant error accumulation, leading to two issues:

Reconstruction The reconstructed image from the inverted noise differs from the source image and fails to maintain the source's features during image editing.

Editability The inverted noise deviates from a Gaussian distribution, causing poor editing results and unexpected behavior.

4 Generalized Framework

4.1 Existing Text-guided Image Editing Methods

We focus on training-free image editing using diffusion inversion for real image editing. We generate an edited image with a pre-trained text-to-image model $\boldsymbol{\epsilon}_\theta$ like Stable Diffusion and adjust certain input parameters for high-quality editing while maintaining the source image's structure. The process involves denoising with both the source and target prompts. By modifying $\boldsymbol{x}_t^{(t)}$ and $\boldsymbol{c}^{(t)}$ of the target process using information from the source branch, it is possible to steer the editing process for better results (notation in Tab. 1):

$$\boldsymbol{x}_{t-1}^{(t)} \leftarrow \mathrm{DDIM}(\boxed{\boldsymbol{x}_t^{(t)}, \boldsymbol{c}^{(t)}}, \varnothing, w; \boxed{\mathcal{M}_t^{(t)}}). \tag{8}$$

Source and target paths only differ in the modified input. In practice, existing methods like PtP [13], MasaCtrl [2] and PnP [41] inject U-Net's [37] attention maps of the source inference process into the target inference process.

4.2 Inversion and Real Image Editing

To perform real image editing, we need to acquire the inverted source noise $\boldsymbol{x}_T^{(s)}$ from the source image $\boldsymbol{x}_0^{(s)}$. Applying DDIM Inversion (forward path) can yield $\boldsymbol{x}_T^{(s)}$, but with considerable reconstruction errors as discussed in Sec. 3.3. Therefore, diffusion inversion methods aim to enhance the forward path for a precise and editable $\boldsymbol{x}_T^{(s)}$ and adjust the backward path to ensure accurate reconstruction and optimal editing. Details of existing inversion methods are provided in the supplementary materials.

Forward Path of Source (Inversion) The forward path can be expressed as $\boldsymbol{x}_{t+1}^{(s)^*} \leftarrow \mathrm{DDIM}_{\mathrm{inv}}(\boldsymbol{x}_t^{(s)^*}, \boldsymbol{c}^{(s)}, \boxed{\varnothing, w})$, with $\varnothing$ or w usually modified. The goal is to emulate the ideal forward path, which is unknown in practice. Many methods use $\mathrm{DDIM}_{\mathrm{inv}}(w = 1)$ to ensure $\boldsymbol{x}_T^{(s)^*}$ aligns well with a Gaussian distribution for better editability.

Backward Path of Source and Target (Reconstruction and Editing) The backward process aims to align with the ideal (unknown) or actual forward path. Existing methods focus on matching the actual forward path by controlling $\varnothing$ or w like NTI [24] and NPI [12,23]. When editing images, two backward paths are used: one for the source prompt and one for the target prompt, written as:

$$\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \mathrm{DDIM}(\boldsymbol{x}_t^{(s)'}, \boldsymbol{c}^{(s)}, \boxed{\varnothing, w}), \tag{9}$$

$$\boldsymbol{x}_{t-1}^{(t)'} \leftarrow \mathrm{DDIM}(\boxed{\boldsymbol{x}_t^{(t)'}, \boldsymbol{c}^{(t)}}, \boxed{\varnothing, w}; \boxed{\mathcal{M}_t^{(t)}}). \tag{10}$$

Inversion methods strive to reduce the gap between $\boldsymbol{x}_{t-1}^{(s)^*}$ and $\boldsymbol{x}_{t-1}^{(s)'}$, but perfect reconstruction remains challenging.

Perfect Reconstruction Methods To achieve perfect source reconstruction, intermediate latents from the forward path can be directly reused for image editing. By replacing the current latent in the backward source path with the corresponding latent from the forward path at each timestep by setting $\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \boxed{\boldsymbol{x}_{t-1}^{(s)^*}}$, we ensure that the backward source path precisely matches the forward path. This alignment guarantees perfect reconstruction, and is employed by CycleDiffusion [43], DDPM Inversion [16], and Direct Inversion [17].

5 Theoretical Analysis of the Role of $\boldsymbol{\eta_t}$

Diffusion inversion demands accurate source image reconstruction and editable target images (Sec. 3.3). Existing perfect reconstruction inversion methods (Sec. 4.2) satisfy the former but lack editability and often yield images too similar to the source. To enhance editability, we explore improving these methods without compromising diffusion model properties.

Motivation Using deterministic DDIM sampling, the source and target backward paths differ only in the estimated noise per timestep, leading to limited editing and target images resembling the source. We aim to enable the target path to diverge from the source path by introducing a stochastic term (additional noise) using non-zero η_t DDIM sampling. In particular, we investigate the optimal design of a function for η_t to achieve superior performance.

Proposition 1 (Proof in Supp.). *Let* $\delta_{\eta_t} = \|\boldsymbol{x}_{t-1}^{(s)'} - \mathrm{DDIM}(\boldsymbol{x}_t^{(t)'}, \boldsymbol{c}^{(t)}, \eta_t)\|_2$ *be the source-target branch distance at timestep* t*. If* δ_0 *is small, there exists an* $\eta_t > 0$ *that satisfies* $\mathbb{E}_{\boldsymbol{\epsilon}_{add}}[\delta_{\eta_t}] > \delta_0$*.*

Proposition 1 indicates that introducing a non-zero η_t can encourage the target path to escape from the source path without losing the property of diffusion models. We further study the role of η_t theoretically to address two major problems for real image editing: (i.) inaccurate inversion ($p_T^{(t)} \neq p_T^{(t)'}$) and (ii.) inaccurate editing ($\boldsymbol{s}_{t,\theta}^{(t)}(x) \neq \nabla_{\boldsymbol{x}} \log q_t^{(t)}(x)$). We use the continuous-time framework of score-based models and measure the sample quality of generation (editing) with KL Divergence D_{KL}.

5.1 Inaccurate Inversion ($\boldsymbol{p_T^{(t)} \neq p_T^{(t)'}}$)

As diffusion inversion methods fail to obtain the ideal inverted $p_T^{(t)}$, the image generation (editing) procedure starts from an inaccurately inverted $p_T^{(t)'}$.

Proposition 2 (Proof in Supp.). *Under mild conditions (see supp.), Eq. (11) is satisfied, wherein D_{Fisher} denotes the Fisher Divergence.*

$$D_{\text{KL}}(p_0^{(t)'} \parallel p_0^{(t)}) = D_{\text{KL}}(p_T^{(t)'} \parallel p_T^{(t)}) - \int_0^T \eta_t^2 g_t^2 D_{\text{Fisher}}(p_t^{(t)'} \parallel p_t^{(t)})\,\mathrm{d}t \tag{11}$$

Proposition 2 shares a similar concept to [21,27], which is generalized to Eq. (6). As $\int_0^T \eta_t^2 g_t^2 D_{\text{Fisher}}(p_t^{(t)'} \parallel p_t^{(t)})\,\mathrm{d}t \geq 0$, we can reduce $D_{\text{KL}}(p_0^{(t)'} \parallel p_0^{(t)})$ by applying η_t with $\int_0^T \eta_t \,\mathrm{d}t > 0$ (SDE) rather than setting $\eta_t = 0$ for all t (ODE). Proposition 2 indicates that introducing a non-zero η_t can improve the backward path of inaccurate diffusion inversion.

5.2 Inaccurate Editing ($\boldsymbol{s}_{t,\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}) \neq \nabla_{\boldsymbol{x}} \log \boldsymbol{q}_t^{(t)}(\boldsymbol{x})$)

If we would assume that the score estimation network is perfect, such that $\boldsymbol{s}_{t,\theta}^{(t)}(x) = \nabla_{\boldsymbol{x}} \log q_t^{(t)}(x)$, the choice of η_t would not change the marginal distribution as $p_{t,\eta_t}^{(t)} = q_t^{(t)}$ [3]. However, since we consider training-free image editing methods, and reuse the score estimation network from a pre-trained image generation model, a non-negligible score estimation error is introduced. As a result, η_t impacts the marginal distribution and good performance cannot be guaranteed by setting $\eta_t = 0$ [3]. Therefore, it is beneficial to optimize η_t for superior performance.

Proposition 3 (Proof in Supp.). *Assuming mild conditions (see supp.), if the score estimation function $\boldsymbol{s}_{t,\theta}^{(t)}(x)$ undergoes perturbations only near timestep T and near timestep 0, there exist a timestep T_a and a timestep T_b, along with a large constant $\eta_{\text{const}} > 0$, such that $D_{\text{KL}}(p_0^{(t)} \parallel q_0^{(t)})$ becomes reduced when employing η_t as Eq. (12), in comparison to $\eta_t = 0$ for all t or $\eta_t = \eta_{\text{const}}$ for all t.*

$$\eta_t = \begin{cases} \eta_{\text{const}} & \text{if } T \geq t \geq T_a \\ \eta_{\text{const}}(t - T_b)/(T_a - T_b) & \text{if } T_a > t \geq T_b \\ 0 & \text{if } T_b > t \geq 0 \end{cases} \tag{12}$$

Proposition 3 is inspired by several findings of [3]. Even though we need to make assumptions for the score estimation function for our theory, it reveals the insight that decreasing η during the backward process can better approximate the true target image distribution and lead to better editing results in practice.

6 Proposed Inversion Method

In this section, we discuss how to design an optimal η function based on our theoretical findings. Our full Eta Inversion algorithm is depicted in Algorithm 1. Fig. 2 provides an overview of our method.

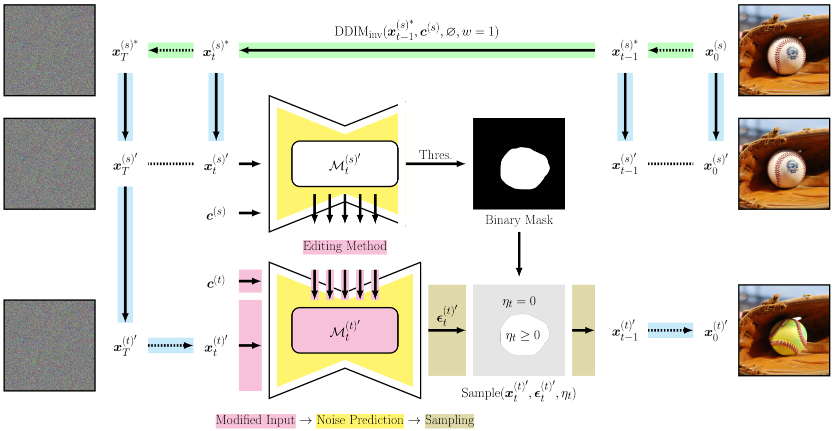

$\mathrm{DDIM}_{\mathrm{inv}}(\boldsymbol{x}_{t-1}^{(s)^*}, \boldsymbol{c}^{(s)}, \varnothing, w=1)$
$\boldsymbol{x}_T^{(s)^*}$
$\boldsymbol{x}_t^{(s)^*}$
$\boldsymbol{x}_{t-1}^{(s)^*}$
$\boldsymbol{x}_0^{(s)}$
$\boldsymbol{x}_T^{(s)'}$
$\boldsymbol{x}_t^{(s)'}$
$\mathcal{M}_t^{(s)'}$
Thres.
Binary Mask
$\boldsymbol{x}_{t-1}^{(s)'}$
$\boldsymbol{x}_0^{(s)'}$
$\boldsymbol{c}^{(s)}$
Editing Method
$\boldsymbol{c}^{(t)}$
$\boldsymbol{x}_T^{(t)'}$
$\boldsymbol{x}_t^{(t)'}$
$\mathcal{M}_t^{(t)'}$
$\boldsymbol{\epsilon}_t^{(t)'}$
$\eta_t = 0$
$\eta_t \geq 0$
$\mathrm{Sample}(\boldsymbol{x}_t^{(t)'}, \boldsymbol{\epsilon}_t^{(t)'}, \eta_t)$
$\boldsymbol{x}_{t-1}^{(t)'}$
$\boldsymbol{x}_0^{(t)'}$
Modified Input → Noise Prediction → Sampling

Fig. 2: Eta Inversion for real image editing. We design an optimal time- and region-dependent η function to inject real noise in the target path to improve editability.

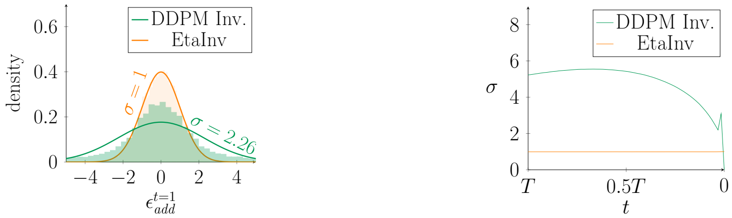
6.1 Exploring the Optimal η Function

Time-dependent η Image editing aims to modify high-level features (e.g., objects) while preserving low-level features (e.g., background details). High-level features are generated early (near timestep T), and low-level features are generated later (near timestep 0) [14, 39, 45]. Therefore, we employ a larger η_t value initially to edit high-level features and a smaller η_t value later to maintain finer details, aligning with Propositions 1 and 3 by progressively reducing η_t for smaller timesteps.

Region-dependent η (Masked η) To improve editing, we employ a region-dependent η inspired by existing editing methods [2, 13, 41] that use attention maps to propagate information from the source to the target path. Concurrent with our method, DiffEditor [25] also employs a region-dependent η but requires an input mask. Our method, on the other hand, uses cross-attention maps to selectively apply a non-zero η to targeted regions without requiring external input. By leveraging the cross-attention map for an object and applying noise ($\eta > 0$) only where the map exceeds a threshold (Fig. 2), we can edit the object while preserving the background. Adjusting the threshold changes the extent of the editing by modifying the region addressed.

6.2 Improving the Injected Noise ϵ_{add}

Although methods like CycleDiffusion [43], DDPM Inversion [16], and Direct Inversion [17] ensure perfect reconstruction by closing the gap between the forward and backward source path with $\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \boldsymbol{x}_{t-1}^{(s)^*}$, they can produce unexpected editing results if the distance $\|\boldsymbol{x}_{t-1}^{(s)^*} - \mathrm{DDIM}(\boldsymbol{x}_t^{(s)'}, \boldsymbol{c}^{(s)}, \eta_t)\|$ is too large. This



(a) Histograms of noise $\epsilon_{add}^{t=1}$ for one sample image from PIE-Bench [17] at $t = 1$. Only Eta Inversion shows a true Gaussian shape with a standard deviation (σ) of 1.

(b) The noise standard deviation $\sigma(\epsilon_{add}^{t})$, across timesteps, averaged over 100 images. Only Eta Inversion consistently maintains $\sigma = 1$ for all timesteps.

Fig. 3: Noise distribution of DDPM Inversion and Eta Inversion. Eta Inversion applies unit Gaussian noise, unlike DDPM Inversion, which applies noise such that $x'_t - x^*_t = 0$.

issue arises because such compensation violates the properties of diffusion models. Specifically, CycleDiffusion [43] and DDPM Inversion [16] calculate $\boldsymbol{\epsilon}_{\text{add}}$ to meet the condition $\boldsymbol{x}_{t-1}^{(s)^*} = \text{DDIM}(\boldsymbol{x}_t^{(s)'}, \boldsymbol{c}^{(s)}, \eta_t)$. However, this $\boldsymbol{\epsilon}_{\text{add}}$ deviates from a Gaussian distribution, which adversely impacts image generation and editing.

Our approach also employs the compensation strategy used in [16, 17, 43] to ensure perfect reconstruction but improves on it by sampling $\boldsymbol{\epsilon}_{\text{add}}$ directly from a Gaussian distribution (Fig. 3). To minimize the forward-backward gap and reduce the necessary compensation, we sample $\boldsymbol{\epsilon}_{\text{add}}$ multiple times and select the noise that minimizes this gap using $\arg\min \|\boldsymbol{x}_{t-1}^{(s)^*} - \text{DDIM}(\boldsymbol{x}_t^{(s)'}, \boldsymbol{c}^{(s)}, \eta_t; \boldsymbol{\epsilon}_{\text{add}})\|$ (Algorithm 1 Backward L. 5, 6).

Algorithm 1 Eta Inversion

```
Input: $\boldsymbol{x}_0^{(s)}$
Output: reconstructed $\boldsymbol{x}_0^{(s)'}$,
    edited $\boldsymbol{x}_0^{(t)'}$

Forward:
 1: initialize $\boldsymbol{x}_0^{(s)^*} \leftarrow \boldsymbol{x}_0^{(s)}$
 2: for $t = 0, 1, ..., T-1$ do
 3:     $\boldsymbol{x}_{t+1}^{(s)^*} \leftarrow \text{DDIM}_{\text{inv}}($
            $\boldsymbol{x}_t^{(s)^*}, \boldsymbol{c}^{(s)}, w = 1)$
 4: end for
 5: return $\boldsymbol{x}_T^{(s)^*}, \boldsymbol{x}_{T-1}^{(s)^*}, ..., \boldsymbol{x}_0^{(s)^*}$

Backward:
 1: initialize $\boldsymbol{x}_T^{(s)'}, \boldsymbol{x}_T^{(t)'} \leftarrow \boldsymbol{x}_T^{(s)^*}$
 2: define time- and region-dependent $\eta_t$
 3: for $t = T, T-1, ..., 1$ do
 4:     $\boldsymbol{x}_{t-1}^{(s)'}(\boldsymbol{\epsilon}_{\text{add}}) := \text{DDIM}($
            $\boldsymbol{x}_t^{(s)'}, \boldsymbol{c}^{(s)}, \eta_t, w = 7.5; \boldsymbol{\epsilon}_{\text{add}})$
 5:     $\{\boldsymbol{\epsilon}\} \leftarrow$ sample noise $n$ times $\sim \mathcal{N}(0, I)$
 6:     $\boldsymbol{\epsilon}_{\text{min}} \leftarrow \arg\min_{\boldsymbol{\epsilon}_{\text{add}} \in \{\boldsymbol{\epsilon}\}} ||\boldsymbol{x}_{t-1}^{(s)^*} - \boldsymbol{x}_{t-1}^{(s)'}(\boldsymbol{\epsilon}_{\text{add}})||$
 7:     $\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \boldsymbol{x}_{t-1}^{(s)^*}$
 8:     $\boldsymbol{x}_{t-1}^{(t)'} \leftarrow \text{DDIM}(\boldsymbol{x}_t^{(t)'}, \boldsymbol{c}^{(t)}, \eta_t, w = 7.5; \boldsymbol{\epsilon}_{\text{min}})$
 9: end for
10: return $\boldsymbol{x}_0^{(s)'}, \boldsymbol{x}_0^{(t)'}$ (satisfying $\boldsymbol{x}_0^{(s)'} = \boldsymbol{x}_0^{(s)}$)
```

Table 2: Evaluation results of inversion methods with various editing methods on PIE-Bench. Our method achieves the highest CLIP scores in most cases while maintaining relatively low structural similarity scores. **EtaInv (1)** and **EtaInv (2)** employ a region-dependent η, which further helps improve structural similarity compared to their versions without mask (w/o mask).

Metric ($\times 10^2$)	CLIP similarity ↑			CLIP accuracy ↑			DINO ↓			LPIPS ↓			BG-LPIPS ↓		
Method	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa
DDIM Inv. [39]	30.99	29.38	30.74	94.57	85.57	95.00	6.94	6.11	7.55	46.65	40.84	47.68	24.97	20.84	25.37
Null-text Inv. [24]	30.73	30.75	30.07	92.57	90.43	93.00	1.24	3.27	4.49	15.13	30.51	25.02	5.69	14.17	11.92
NPI [23]	30.49	30.73	29.54	92.71	91.29	87.29	2.03	2.67	4.51	19.28	26.18	26.03	8.24	11.57	12.41
ProxNPI [12]	30.31	30.54	29.49	92.43	90.71	88.14	1.92	2.29	3.92	17.69	21.76	22.99	7.76	9.57	10.99
EDICT [42]	29.28	24.69	29.68	92.71	63.43	93.29	0.41	4.26	0.79	6.65	30.22	8.59	3.10	14.96	4.20
DDPM Inv. [16]	29.43	30.26	29.57	92.71	94.86	93.00	0.42	1.04	0.75	6.87	12.50	8.65	3.27	5.84	4.12
Direct Inv. [17]	30.92	31.32	30.37	94.71	95.14	94.57	1.28	2.27	4.32	15.79	25.59	26.91	6.33	12.98	13.76
Eta Inversion (1)	31.01	31.33	30.39	95.00	94.86	93.14	1.34	2.34	3.66	16.58	27.33	23.12	6.57	14.05	11.57
Eta Inversion (1) w/o mask	31.00	31.34	30.37	95.29	95.00	92.71	1.37	2.37	3.69	16.85	27.68	23.40	6.74	14.33	11.79
Eta Inversion (2)	31.25	31.63	30.62	95.43	95.29	93.86	1.70	3.40	5.24	21.14	36.59	33.07	8.00	18.72	16.64
Eta Inversion (2) w/o mask	31.27	31.62	30.62	95.43	95.86	94.14	1.85	3.58	5.46	22.77	38.43	34.81	9.03	20.19	18.03

Table 3: Evaluation results on the change-style subset of PIE-Bench. **EtaInv (3)** is optimized for style transfer and uses a larger η to significantly outperform previous methods in terms of CLIP similarity. Since style transfer requires changing the whole image, **EtaInv (3)** does not use η masking.

Metric ($\times 10^2$)	CLIP similarity ↑			CLIP accuracy ↑			DINO ↓			LPIPS ↓		
Method	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa
DDIM Inv. [39]	31.00	30.21	30.67	83.75	73.75	86.25	6.47	6.09	6.90	46.76	42.63	47.42
Null-text Inv. [24]	32.06	32.79	29.97	88.75	91.25	86.25	1.60	3.98	4.07	19.60	37.26	25.81
NPI [23]	31.44	32.37	29.60	92.50	90.00	75.00	2.22	3.30	4.04	22.18	32.26	27.37
ProxNPI [12]	30.88	31.66	29.38	86.25	85.00	80.00	2.02	2.52	3.36	19.18	25.47	22.68
EDICT [42]	29.45	25.32	29.93	91.25	58.75	90.00	0.41	4.22	0.73	6.68	31.11	8.57
DDPM Inv. [16]	29.78	30.64	29.78	90.00	90.00	90.00	0.43	0.97	0.66	6.99	12.53	8.50
Direct Inv. [17]	31.71	32.51	30.37	91.25	93.75	85.00	1.64	2.47	3.79	19.87	27.22	26.56
Eta Inversion (3)	32.85	33.12	30.82	90.00	86.25	86.25	4.19	5.16	6.69	47.76	52.66	46.17

7 Experiments

7.1 Setup

We unify and re-implement existing diffusion inversion methods based on diffusers [30] and opt for Stable Diffusion v1.4 [36] with $T = 50$ steps, using default settings for all methods. For image editing, we apply PtP [13], PnP [41], and MasaCtrl [2] on the dataset PIE-Bench [17]. Evaluating image editing performance is challenging due to the lack of clear metrics. Prior works [17, 24, 41] focused on two factors: (i.) text-image alignment, indicating the output image's faithfulness to the target prompt; and (ii.) structural similarity, showing how well the output image preserves the source image's structure.

For text-image alignment we use: (i.) **CLIP similarity**: the dot product of normalized CLIP [32] embeddings of the target prompt and the output image; and (ii.) **CLIP accuracy**: ratio of output images where the text-caption similarity with the target prompt is higher than with the source prompt [29]. Text-caption similarity [7] is defined as the CLIP similarity between the target prompt and the BLIP-generated [19] caption of the output image. For structural similarity we

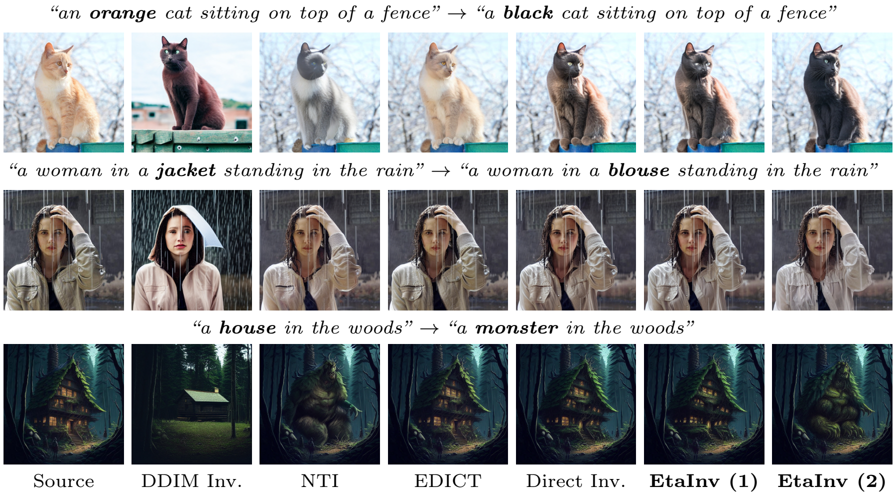

"an **orange** cat sitting on top of a fence" → "a **black** cat sitting on top of a fence"
"a woman in a **jacket** standing in the rain" → "a woman in a **blouse** standing in the rain"
"a **house** in the woods" → "a **monster** in the woods"
Source
DDIM Inv.
NTI
EDICT
Direct Inv.
EtaInv (1)
EtaInv (2)

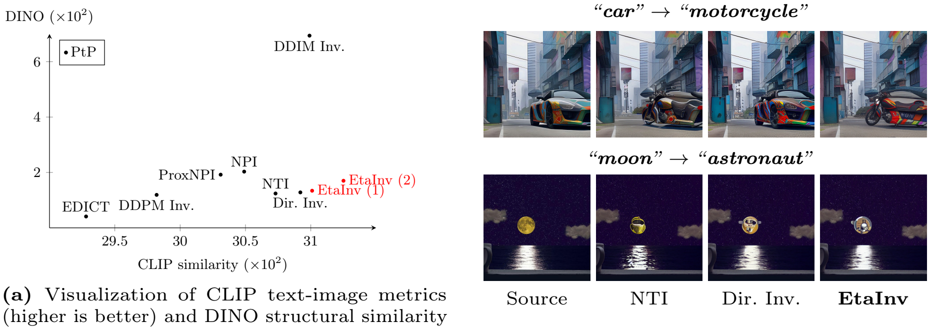
Fig. 4: Image editing qualitative results created with PtP [13] and various inversion methods. Our method, particularly **EtaInv (2)**, outperforms existing methods and edits the image to a greater degree. We preserve the structure of the source image while correctly editing the image to match the target prompt.

use: (i.) **DINOv1 ViT** [4]; (ii.) **LPIPS** [47]; and (iii.) **BG-LPIPS** [17], which computes LPIPS only on the background part (mask is provided by PIE-Bench).

We present our results on the complete PIE-Bench dataset, as well as on the change-style subset of PIE-Bench, which focuses exclusively on style transfer. In general, we found that a decreasing linear η schedule improves results, and that a larger η results in more editing, which aligns with our findings. Additionally, a larger noise sample count n achieves better structural similarity scores and more stable editing overall. We propose three distinct linear η functions, each optimized for a specific objective: structural similarity (EtaInv (1)), target prompt alignment (EtaInv (2)), and style transfer (EtaInv (3)). The η functions used, additional qualitative and quantitative results, and comprehensive hyperparameter grid search results are included in the supplementary materials.

7.2 PIE-Bench Results

Tab. 2 presents our results on PIE-Bench with EtaInv (1) and (2). For PtP, our method balances text-image alignment and structural similarity, achieving the highest CLIP text-image score and a low structural similarity score. PnP also shows our method as the best in CLIP similarity and accuracy. While our structural metrics are inferior, a too low score may indicate insufficient editing (like EDICT's PtP result in Fig. 4). Lastly, with MasaCtrl, we achieve the second-best CLIP similarity but worse structural similarity compared to other techniques. Fig. 5a visualizes the trade-off between text-image and structural similarity for PtP (see supplementary for PnP and MasaCtrl).



(a) Visualization of CLIP text-image metrics (higher is better) and DINO structural similarity metrics (lower is better) on PIE-Bench for PtP.

(b) Failure cases of our proposed method.

Fig. 5: CLIP-DINO trade-off plot and failure cases.

Fig. 4 showcases qualitative results for the top-performing methods. Our proposed Eta Inversion demonstrates superior editing performance. Notably, EtaInv (2), which employs a higher η, promotes more editing. Furthermore, utilizing a region-dependent η enhances structural similarity metrics by preserving more background (Fig. 6) while introducing a slight decrease in CLIP metrics.

Style Transfer Results Style transfer requires changing the whole image to a higher degree than other tasks (e.g., object replacing). Thus, we disable η masking and increase η to introduce more noise to further enlarge the gap between the source and target branch for a better editing effect. Tab. 3 shows that EtaInv (3) significantly improves CLIP similarity over previous methods, which we attribute to the injected real noise. Although DINO and LPIPS scores suggest underperformance, these metrics are less useful for style transfer, which requires complete image editing. Fig. 7 further demonstrates that EtaInv (3) achieves more impactful and faithful style transfer.

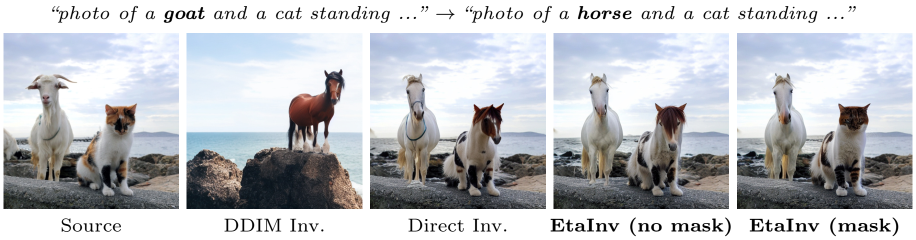

"photo of a goat and a cat standing ..." → "photo of a horse and a cat standing ..."
Source
DDIM Inv.
Direct Inv.
EtaInv (no mask)
EtaInv (mask)

Fig. 6: Effectiveness of a region-dependent (masked) η function. Only **EtaInv (mask)** preserves the cat in the original image.

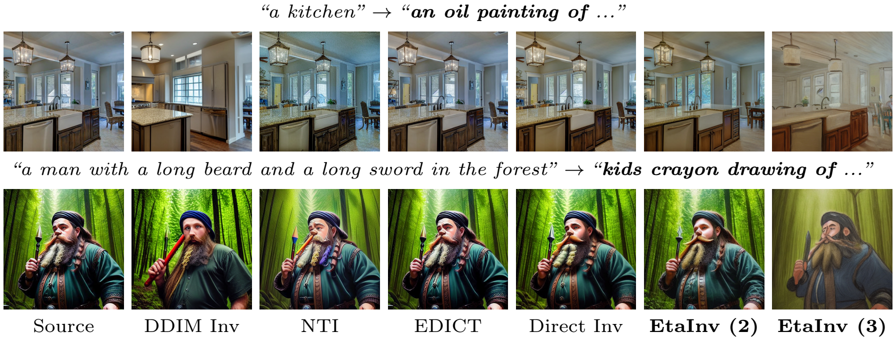

"a kitchen" → *"**an oil painting of** ..."*
"a man with a long beard and a long sword in the forest" → *"**kids crayon drawing of** ..."*
Source
DDIM Inv
NTI
EDICT
Direct Inv
EtaInv (2)
EtaInv (3)

Fig. 7: Style transfer results created with PtP [13] and various inversion methods. **Eta Inversion (3)** with a larger η function improves style transfer.

8 Limitations

Some image edits yield unrealistic outcomes or insufficient changes, despite preserving the original structure (Fig. 5b). Adjusting the seed and η function can improve results, but no universal setting works for every edit. Future efforts will focus on automating the optimal η selection. Furthermore, existing metrics for evaluating image editing are limited, as none measure both structural similarity with the source image and faithfulness to the target prompt. We propose exploring Multimodal Large Language Models [10, 20, 28] for more effective image editing assessment in future research.

9 Conclusion

In this paper, we propose a unified framework for diffusion inversion and introduce Eta Inversion, a novel approach for real image editing. Our method incorporates real noise into the editing process by utilizing an optimally designed η function within DDIM sampling for faithful image editing. Through detailed comparison and analysis of the role of η, we demonstrate state-of-the-art performance in real image editing across various metrics, offering both compelling qualitative outcomes and precise editing control.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
2. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023)

3. Cao, Y., Chen, J., Luo, Y., ZHOU, X.: Exploring the optimal choice for generative processes in diffusion models: Ordinary vs stochastic differential equations. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
5. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K.P., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: International Conference on Machine Learning. pp. 4055–4075. PMLR (2023)
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
8. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: The Eleventh International Conference on Learning Representations (2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making LLaMA SEE and draw with SEED tokenizer. In: The Twelfth International Conference on Learning Representations (2024)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
12. Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., et al.: Proxedit: Improving tuning-free real image editing with proximal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4291–4301 (2024)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
16. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12469–12478 (2024)
17. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Direct inversion: Boosting diffusion-based editing with 3 lines of code. arXiv preprint arXiv:2310.01506 (2023)
18. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
19. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)

21. Lu, C., Zheng, K., Bao, F., Chen, J., Li, C., Zhu, J.: Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In: International Conference on Machine Learning. pp. 14429–14460. PMLR (2022)
22. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
23. Miyake, D., Iohara, A., Saito, Y., Tanaka, T.: Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. arXiv preprint arXiv:2305.16807 (2023)
24. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
25. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8488–8497 (2024)
26. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
27. Nie, S., Guo, H.A., Lu, C., Zhou, Y., Zheng, C., Li, C.: The blessing of randomness: SDE beats ODE in general diffusion-based image editing. In: The Twelfth International Conference on Learning Representations (2024)
28. OpenAI: Gpt-4 technical report (2023)
29. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
30. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
34. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
35. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: 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)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)

37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
40. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
41. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
42. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22532–22541 (2023)
43. Wu, C.H., De la Torre, F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 (2022)
44. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3121–3138 (2022)
45. Yue, Z., Wang, J., Sun, Q., Ji, L., Chang, E.I.C., Zhang, H.: Exploring diffusion time-steps for unsupervised representation learning. In: The Twelfth International Conference on Learning Representations (2024)
46. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: The Eleventh International Conference on Learning Representations (2023)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)