Supplementary Materials

A Proofs of Propositions

A.1 Proof of Proposition 1

Proposition 1 *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* $\mathbb{E}_{\boldsymbol{\epsilon}_{add}}[\delta_{\eta_t}] > \delta_0$*.*

Proof. Given a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, it is known that the random variable $|X|$ follows a *Folded Normal Distribution* with

$$\mathbb{E}[|X|] = \sigma\sqrt{\frac{2}{\pi}}e^{-\mu^2/2\sigma^2} + \mu\,\mathrm{erf}(\frac{\mu}{\sqrt{2\sigma^2}}), \tag{1}$$

$$\arg\min_{\mu} \mathbb{E}[|X|] = 0, \tag{2}$$

where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,\mathrm{d}t$. Let $\boldsymbol{x} \in \mathbb{R}^d$ and

$$\boldsymbol{\mu}_t \coloneqq \frac{(\boldsymbol{x}_t^{(t)'} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^{(t)'})}{\sqrt{\alpha_t}} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\boldsymbol{\epsilon}_t^{(t)'} - \boldsymbol{x}_{t-1}^{(s)'}. \tag{3}$$

As Eq. (3) requires $1-\bar{\alpha}_{t-1}-\sigma_t^2 \geq 0$, using the definition of $\sigma_t(\eta_t)$ we write

$$\sqrt{1-\bar{\alpha}_{t-1}} \geq \eta_t\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}, \tag{4}$$

resulting in the following condition for η_t:

$$\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}} \geq \eta_t \geq 0. \tag{5}$$

Assuming δ_0 is sufficiently small with $\sqrt{1-\bar{\alpha}_{t-1}}\sqrt{\frac{2d}{\pi}} > \delta_0$, we show that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_{add}}[\delta_{\eta_t}] &= \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_{add}}[\|\boldsymbol{\mu}_t + \sigma_t \boldsymbol{\epsilon}_{\text{add}}\|_2] && (6)\\
&\geq \frac{1}{\sqrt{d}} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}_{add}}[\|\boldsymbol{\mu}_t + \sigma_t \boldsymbol{\epsilon}_{\text{add}}\|_1] && \text{(Cauchy–Schwarz inequality)} \quad (7)\\
&= \frac{1}{\sqrt{d}} \mathbb{E}[\Sigma_{i=1}^{d} |\mu_{t,i} + \sigma_t \epsilon_{\text{add},i}|] && \text{(Sum of } i^{\text{th}} \text{ dimension's)} \quad (8)\\
&= \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \mathbb{E}[|\mu_{t,i} + \sigma_t \epsilon_{\text{add},i}|] && (9)\\
&= \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \mathbb{E}[|X_i|] && (X_i \sim \mathcal{N}(\mu_{t,i}, \sigma_t^2)) \quad (10)\\
&\geq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \mathbb{E}[|X_i|]_{\mu_{t,i}=0} && (Eq.\ (2)) \quad (11)\\
&= \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \sigma_t \sqrt{\frac{2}{\pi}} && (Eq.\ (1)) \quad (12)\\
&= \sigma_t \sqrt{\frac{2d}{\pi}} && (13)\\
&= \eta_t \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} \sqrt{\frac{2d}{\pi}}. && (14)
\end{aligned}$$

Thus, our proposition $\mathbb{E}_{\epsilon_{add}}[\delta_{\eta_t}] > \delta_0$ holds if we choose an η_t, which satisfies

$$\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}} \geq \eta_t > \frac{\delta_0}{\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} \sqrt{\frac{2d}{\pi}}}. \quad (15)$$

□

A.2 Proof of Proposition 2

Assumption 1 *We rewrite the following assumptions from prior works [17, 20, 28] using our notation for completeness.*

1. $q_0(\boldsymbol{x}) \in \mathcal{C}^3$ *and* $\mathbb{E}_{q_0(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^2] < \infty$.
2. $\forall t \in [0,T]$ *:* $\boldsymbol{f}_t(\cdot) \in \mathcal{C}^2$. *And* $\exists C > 0$, $\forall \boldsymbol{x} \in \mathbb{R}^d$, $t \in [0,T]$ *:* $\|\boldsymbol{f}_t(\boldsymbol{x})\|_2 \leq C(1 + \|\boldsymbol{x}\|_2)$.
3. $\exists C > 0$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ *:* $\|\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}_t(\boldsymbol{y})\|_2 \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_2$.
4. $g \in \mathcal{C}$ *and* $\forall t \in [0,T]$, $|g(t)| > 0$.
5. *Open bounded set* $\forall O$, $\int_0^T \int_O \|q_t(\boldsymbol{x})\|_2^2 + d \cdot g(t)^2 \|\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t < \infty$.
6. $\exists C > 0$, $\forall \boldsymbol{x} \in \mathbb{R}^d$, $t \in [0,T]$ *:* $\|\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})\|_2^2 \leq C(1 + \|\boldsymbol{x}\|_2)$.

7. $\exists C > 0,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\ :\ \|\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}) - \nabla_{\boldsymbol{y}} \log q_t(\boldsymbol{y})\|_2 \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_2$.
8. $\exists C > 0,\ \forall \boldsymbol{x} \in \mathbb{R}^d,\ t \in [0, T]\ :\ \|\boldsymbol{s}_{t,\theta}(\boldsymbol{x})\|_2 \leq C(1 + \|\boldsymbol{x}\|_2)$.
9. $\exists C > 0,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\ :\ \|\boldsymbol{s}_{t,\theta}(\boldsymbol{x}) - \boldsymbol{s}_{t,\theta}(\boldsymbol{y})\|_2 \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_2$.
10. *Novikov's condition:* $\mathbb{E}[\exp(\frac{1}{2}\int_0^T \|\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}) - \boldsymbol{s}_{t,\theta}(\boldsymbol{x})\|_2^2)] < \infty$.
11. $\forall t \in [0, T],\ \exists k > 0\ :\ q_t(\boldsymbol{x}) = \mathcal{O}(e^{-\|\boldsymbol{x}\|_2^k}),\ p_{t,\eta_t}(\boldsymbol{x}) = \mathcal{O}(e^{-\|\boldsymbol{x}\|_2^k})$ *as* $\|\boldsymbol{x}\|_2 \to \infty$.

Proposition 2 *Under Assumption 1, Eq. (16) holds, wherein* D_{Fisher} *denotes the Fisher Divergence.*

$$D_{\text{KL}}(p_0^{(t)'} \parallel p_0^{(t)}) = D_{\text{KL}}(p_T^{(t)'} \parallel p_T^{(t)}) - \int_0^T \eta_t^2 g_t^2 D_{\text{Fisher}}(p_t^{(t)'} \parallel p_t^{(t)})\,\mathrm{d}t \tag{16}$$

Proof. Given the general form of the SDE (Eq. (17)), Eq. (18) and Eq. (19) are the equations for the probability flow ODE, and Eq. (20) is the link between the Fokker–Planck equation [28] and the probability flow.

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{f}_t(\boldsymbol{x})\,\mathrm{d}t + \boldsymbol{G}_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{w} \tag{17}$$

$$\mathrm{d}\boldsymbol{x} = \tilde{\boldsymbol{f}}_t(\boldsymbol{x})\,\mathrm{d}t \tag{18}$$

$$\tilde{\boldsymbol{f}}_t(\boldsymbol{x}) \leftarrow \boldsymbol{f}_t(\boldsymbol{x}) - \frac{1}{2}\nabla \cdot [\boldsymbol{G}_t(\boldsymbol{x})\boldsymbol{G}_t(\boldsymbol{x})^\intercal] - \frac{1}{2}\boldsymbol{G}_t(\boldsymbol{x})\boldsymbol{G}_t(\boldsymbol{x})^\intercal \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \tag{19}$$

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{f}}_t(\boldsymbol{x}) p_t(\boldsymbol{x})] \tag{20}$$

Recall the forward path (Eq. (21)) and the extended backward path (Eq. (22)) of diffusion models (score-based models). In practice, we use Eq. (23) as the backward path to sample data with the score estimation network $\boldsymbol{s}_{t,\theta}(\boldsymbol{x})$ instead of the unknown ground truth $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$.

$$\mathrm{d}\boldsymbol{x} = f_t \boldsymbol{x}\,\mathrm{d}t + g_t\,\mathrm{d}\boldsymbol{w} \quad \left(f_t = \frac{1}{2}\frac{\mathrm{d}\log \alpha_t}{\mathrm{d}t}, g_t = \sqrt{-\frac{\mathrm{d}\log \alpha_t}{\mathrm{d}t}}\right) \tag{21}$$

$$\mathrm{d}\boldsymbol{x} = [f_t \boldsymbol{x} - \frac{1+\eta_t^2}{2} g_t^2 \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})]\,\mathrm{d}t + \eta_t g_t\,\mathrm{d}\bar{\boldsymbol{w}} \tag{22}$$

$$\mathrm{d}\boldsymbol{x} = \underbrace{[f_t \boldsymbol{x} - \frac{1+\eta_t^2}{2} g_t^2 \boldsymbol{s}_{t,\theta}(\boldsymbol{x})]}_{\boldsymbol{A}_t(\boldsymbol{x})}\,\mathrm{d}t + \eta_t g_t\,\mathrm{d}\bar{\boldsymbol{w}} \tag{23}$$

Using $\boldsymbol{f}_t(\boldsymbol{x}) \leftarrow \boldsymbol{A}_t(\boldsymbol{x})$ and $\boldsymbol{G}_t(\boldsymbol{x}) \leftarrow \eta_t g_t$, we rewrite the probability flow ODE (Eq. (18), Eq. (19)) as

$$\mathrm{d}\boldsymbol{x} = \tilde{\boldsymbol{A}}_t(\boldsymbol{x})\,\mathrm{d}t, \tag{24}$$

$$\tilde{\boldsymbol{A}}_t(\boldsymbol{x}) = \boldsymbol{A}_t(\boldsymbol{x}) + \frac{1}{2}\eta_t^2 g_t^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}), \tag{25}$$

and the Fokker-Plank equation (Eq. (20)) as

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{A}}_t(\boldsymbol{x}) p_t(\boldsymbol{x})]. \tag{26}$$

For real image editing (the backward path of the target), we are interested in the unknown true inverted marginal distribution $p_T^{(t)}$ and the inaccurately inverted marginal distribution $p_T^{(t)'}$, which is obtained by diffusion inversion. Since both distributions use the same backward path (Eq. (23)), we can formulate the probability flow ODE for both as

$$\boldsymbol{A}_t^{(t)}(\boldsymbol{x}) = f_t\boldsymbol{x} - \frac{1+\eta_t^2}{2} g_t^2 \boldsymbol{s}_{t,\theta}^{(t)}(\boldsymbol{x}), \tag{27}$$

$$\tilde{\boldsymbol{A}}_t^{(t)}(\boldsymbol{x}) = \boldsymbol{A}_t^{(t)}(\boldsymbol{x}) + \frac{1}{2}\eta_t^2 g_t^2 \nabla_{\boldsymbol{x}} \log p_t^{(t)}(\boldsymbol{x}), \tag{28}$$

$$\tilde{\boldsymbol{A}}_t^{(t)'}(\boldsymbol{x}) = \boldsymbol{A}_t^{(t)}(\boldsymbol{x}) + \frac{1}{2}\eta_t^2 g_t^2 \nabla_{\boldsymbol{x}} \log p_t^{(t)'}(\boldsymbol{x}), \tag{29}$$

and the Fokker-Plank equation for both as

$$\frac{\partial}{\partial t} p_t^{(t)}(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{A}}_t^{(t)}(\boldsymbol{x}) p_t^{(t)}(\boldsymbol{x})], \tag{30}$$

$$\frac{\partial}{\partial t} p_t^{(t)'}(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{A}}_t^{(t)'}(\boldsymbol{x}) p_t^{(t)'}(\boldsymbol{x})]. \tag{31}$$

Finally, we show

$$\frac{\partial D_{\mathrm{KL}}(p_t^{(t)'} \parallel p_t^{(t)})}{\partial t} \tag{32}$$

$$= \frac{\partial}{\partial t} \int p_t^{(t)'}(\boldsymbol{x}) \log \frac{p_t^{(t)'}(\boldsymbol{x})}{p_t^{(t)}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \tag{33}$$

$$= \int \frac{\partial}{\partial t} p_t^{(t)'}(\boldsymbol{x}) \log \frac{p_t^{(t)'}(\boldsymbol{x})}{p_t^{(t)}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} - \int \frac{p_t^{(t)'}(\boldsymbol{x})}{p_t^{(t)}(\boldsymbol{x})} \frac{\partial}{\partial t} p_t^{(t)}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \tag{34}$$

$$\begin{aligned} = &- \int \nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{A}}_t^{(t)'}(\boldsymbol{x}) p_t^{(t)'}(\boldsymbol{x})] \log \frac{p_t^{(t)'}(x)}{p_t^{(t)}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \\ &+ \int \frac{p_t^{(t)'}(\boldsymbol{x})}{p_t^{(t)}(\boldsymbol{x})} \nabla_{\boldsymbol{x}} \cdot [\tilde{\boldsymbol{A}}_t^{(t)}(\boldsymbol{x}) p_t^{(t)}(\boldsymbol{x})] \,\mathrm{d}\boldsymbol{x} \end{aligned} \tag{35}$$

$$\begin{aligned} = &\int [\tilde{\boldsymbol{A}}_t^{(t)'}(\boldsymbol{x}) p_t^{(t)'}(\boldsymbol{x})]^{\intercal} \nabla_{\boldsymbol{x}} \log \frac{p_t^{(t)'}(x)}{p_t^{(t)}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \\ &- \int [\tilde{\boldsymbol{A}}_t^{(t)}(\boldsymbol{x}) p_t^{(t)}(\boldsymbol{x})]^{\intercal} \nabla_{\boldsymbol{x}} \frac{p_t^{(t)'}(\boldsymbol{x})}{p_t^{(t)}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \qquad \text{(Assumption 1)} \end{aligned} \tag{36}$$

$$= \int p_t^{(t)'}(\boldsymbol{x}) [\tilde{\boldsymbol{A}}_t^{(t)'}(\boldsymbol{x})^{\intercal} - \tilde{\boldsymbol{A}}_t^{(t)}(\boldsymbol{x})^{\intercal}] [\nabla_{\boldsymbol{x}} \log p_t^{(t)'}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_t^{(t)}(\boldsymbol{x})] \,\mathrm{d}\boldsymbol{x} \tag{37}$$

$$= \frac{1}{2}\eta_t^2 g_t^2 \int p_t^{(t)'}(\boldsymbol{x}) \|\nabla_{\boldsymbol{x}} \log p_t^{(t)'}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_t^{(t)}(\boldsymbol{x})\|_2^2 \,\mathrm{d}\boldsymbol{x} \tag{38}$$

$$= \eta_t^2 g_t^2 D_{\mathrm{Fisher}}(p_t^{(t)'} \parallel p_t^{(t)}). \qquad \text{(See [17, 20])} \tag{39}$$

Thus, Eq. (16) holds by integrating Eq. (39).

□

A.3 Proof of Proposition 3

We omit the superscript $\square^{(t)}$ in this section for simplicity. We express the scale of the score estimation error as ϵ, which we assume is sufficiently small:

$$\boldsymbol{s}_{t,\theta}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}) + \epsilon\, \mathrm{Error}(\boldsymbol{x}), \tag{40}$$

with $\mathrm{Error}(x) = \mathcal{O}(1)$. It is known that $D_{\mathrm{KL}}(p_0 \parallel q_0)$ has order of ϵ^2 as below [5,6]:

$$D_{\mathrm{KL}}(p_0 \parallel q_0) = \epsilon^2 L(\eta_t) + \mathcal{O}(\epsilon^3). \tag{41}$$

Assumption 2 *We rewrite the following assumptions from prior work [2] using our notation for completeness.*

1. *Without loss of generality (time re-scaling),* $f_t = -\frac{1}{2}$, $g_t = 1$.
2. $\exists c_U \in \mathbb{R}$, $\forall \boldsymbol{x} \in \mathbb{R}^d$: $-\log p_0(\boldsymbol{x}) - \frac{|\boldsymbol{x}|^2}{2} \geq c_U$.
3. $\forall t \in [0, T]$, $-\log p_t(\boldsymbol{x})$ *is strongly convex.*
4. $\forall t \in [0, T]$, $m_t \boldsymbol{I} \preceq \nabla^2(-\log p_t(\boldsymbol{x})) \preceq M_t \boldsymbol{I}$ *where* $m_t \geq 1$ *for* $t \in (0, T]$ *and* $m_0 > 1$.

The proof for Proposition 3 is based on two propositions from prior work [2], which we reformulate to match our notation (Lemma 1 and Lemma 2). Under Assumption 2 (1), η_t in our notation corresponds to h_t from [2]. For the rest of this section, δ represents the Dirac delta function.

Lemma 1 (Proposition 3.4 of [2]). *Suppose the score estimation function* $\boldsymbol{s}_{t,\theta}(x)$ *only undergoes perturbation at some fixed arbitrary timestep* $t_a \in (0, T]$ *with* $\mathrm{Error}(\boldsymbol{x}) = \delta_{t-t_a} E(\boldsymbol{x})$. *Let* $\eta_t = \eta$ *(η_t is constant for all t). Under Assumption 2, and if η is large enough, there exists an upper bound* $L_{\mathrm{ub}}(\eta) \geq L(\eta)$, *which is an exponentially decreasing function converging to 0. Thus, there exists an η with* $L(\eta) < \min(\epsilon, L(0))$.

Lemma 2 (Proposition 3.5 of [2]). *Suppose the score estimation function* $\boldsymbol{s}_{t,\theta}(x)$ *only undergoes perturbation at some fixed timestep* $t_b \ll 1$ *near timestep 0 with* $\mathrm{Error}(\boldsymbol{x}) = \delta_{t-t_b} E(\boldsymbol{x})$. *Let* $\eta_t = \eta$ *(η_t is constant for all t). Under Assumption 2, and if η is large enough, we have* $L(0) \ll L(\eta)$.

Proposition 3 *Under Assumption 2, if the score estimation function* $\boldsymbol{s}_{t,\theta}(x)$ *undergoes perturbations only near the timestep T and near the timestep 0, there exists a timestep T_a and a timestep T_b, along with a large constant* $\eta_{\mathrm{const}} > 0$, *such that* $D_{\mathrm{KL}}(p_{0,\eta_t} \parallel q_0)$ *becomes reduced when employing η_t as Eq. (42), in comparison to $\eta_t = 0$ for all t or $\eta_t = \eta_{\mathrm{const}}$ for all t.*

$$\eta_t = \begin{cases} \eta_{\text{const}} & \text{if } T \geq t \geq T_a \\ \eta_{\text{const}}(t - T_b)/(T_a - T_b) & \text{if } T_a > t \geq T_b \\ 0 & \text{if } T_b > t \geq 0 \end{cases} \tag{42}$$

Proof. We define the following cases for three different possible η_t functions (Fig. 1):

- Case 1: Eq. (42),
- Case 2: $\eta_t = \eta_{\text{const}}$ for all t,
- Case 3: $\eta_t = 0$ for all t.

We write $\text{Error}(\boldsymbol{x})$ as below assuming perturbations only at t_a (near timestep T) and at t_b (near timestep 0):

$$\text{Error}(\boldsymbol{x}) = (\delta_{t-t_a} + \delta_{t-t_b})E(\boldsymbol{x}). \tag{43}$$

Let T_a be an arbitrary timestep with $t_a > T_a > t_b$. Assume we perform a shorter diffusion backward pass from T to T_a by setting the final diffusion step to T_a instead of 0 and measure its sample quality with $D_{\text{KL}}(p_{T_a} \parallel q_{T_a})$. Since we now only operate in the time interval $[T_a, T]$, we can ignore the perturbation at t_b and rewrite the error function as $\text{Error}(\boldsymbol{x}) = \delta_{t-t_a}E(\boldsymbol{x})$. By applying Lemma 1 to our new diffusion pass in $[T_a, T]$, without loss of generality, there exists a constant $\eta_{\text{const,a}}$ so that

$$L(\eta_{\text{const,a}}) < \min(\epsilon, L(T_a)) \leq L(T_a), \tag{44}$$

$$D_{\text{KL}}(p_{T_a,\eta_{\text{const,a}}} \parallel q_{T_a}) = \epsilon^2 L(\eta_t) + \mathcal{O}(\epsilon^3) = \mathcal{O}(\epsilon^3) \approx 0. \tag{45}$$

Similarly, let T_b be an arbitrary timestep with $T_a > T_b > t_b$ and assume we perform a shorter diffusion backward pass starting from T_b and ending at 0. From Lemma 2, there exists a constant $\eta_{\text{const,b}}$ that satisfies $L(0) \ll L(\eta_{\text{const,b}})$.

Let $\eta_{\text{const}} = \max(\eta_{\text{const,a}}, \eta_{\text{const,b}})$. We compare case 1 to case 2 and case 3 and show that case 1 has the best sampling quality among those three.

i) Comparison to case 2 ($\eta_t = \eta_{\text{const}}$ for all t).

1. $T \geq t \geq T_a$: By Lemma 1 and Eq. (45), we have $D_{\text{KL}}(p_{T_a,\eta_t} \parallel q_{T_a}) \approx 0$ for case 1 and $D_{\text{KL}}(p_{T_a,\eta_{\text{const}}} \parallel q_{T_a}) \approx 0$ for case 2.
2. $T_a \geq t \geq T_b$: Since we have $D_{\text{KL}}(p_{T_a} \parallel q_{T_a}) \approx 0$ for both cases and the score function is accurate, the η_t function does not affect p_{T_b,η_t}, thus we have $D_{\text{KL}}(p_{T_b} \parallel q_{T_b}) \approx 0$ for both case 1 and case 2.
3. $T_b \geq t \geq 0$: Since $\eta_t = 0$ for $t \leq T_b$ for case 1, case 1's $D_{\text{KL}}(p_{0,\eta_t} \parallel q_0)$ is smaller than case 2's $D_{\text{KL}}(p_{0,\eta_{\text{const}}} \parallel q_0)$ by Lemma 2.

ii) Comparison to case 3 ($\eta_t = 0$ for all t).

1. $T \geq t \geq T_a$: By Lemma 1, Eq. (44) and Eq. (45), case 1 satisfies $D_{\text{KL}}(p_{T_a,\eta_t} \parallel q_{T_a}) \approx 0$ while we have $D_{\text{KL}}(p_{T_a,0} \parallel q_{T_a}) > D_{\text{KL}}(p_{T_a,\eta_t} \parallel q_{T_a})$ for case 3.
2. $T_a \geq t \geq T_b$: Since the score function is accurate and $D_{\text{KL}}(p_{T_a,\eta_t} \parallel q_{T_a}) \approx 0$ for case 1, $D_{\text{KL}}(p_{T_b,\eta_t} \parallel q_{T_b}) \approx 0$ holds while we have $D_{\text{KL}}(p_{T_a,0} \parallel q_{T_a}) > D_{\text{KL}}(p_{T_a,\eta_t} \parallel q_{T_a})$ and therefore $D_{\text{KL}}(p_{T_b,0} \parallel q_{T_b}) > D_{\text{KL}}(p_{T_b,\eta_t} \parallel q_{T_b})$ for case 3.

3. $T_b \geq t \geq 0$: For case 1 we have $D_{\mathrm{KL}}(p_{T_b,\eta_t} \parallel q_{T_b}) \approx 0$ and for case 3 we have $D_{\mathrm{KL}}(p_{T_b,0} \parallel q_{T_b}) > D_{\mathrm{KL}}(p_{T_b,\eta_t} \parallel q_{T_b})$. Since both case 1 and case 3 follow the same ODE ($\eta_t = 0$ for $t \leq T_b$), case 1's $D_{\mathrm{KL}}(p_{0,\eta_t} \parallel q_0)$ is smaller than case 3's $D_{\mathrm{KL}}(p_{0,0} \parallel q_0)$.

Following **i)** and **ii)**, case 1 has the best sample quality, since its $D_{\mathrm{KL}}(p_{0,\eta_t} \parallel q_0)$ is smaller than $D_{\mathrm{KL}}(p_{0,\eta_{\mathrm{const}}} \parallel q_0)$ of case 2 and $D_{\mathrm{KL}}(p_{0,0} \parallel q_0)$ of case 3.

□

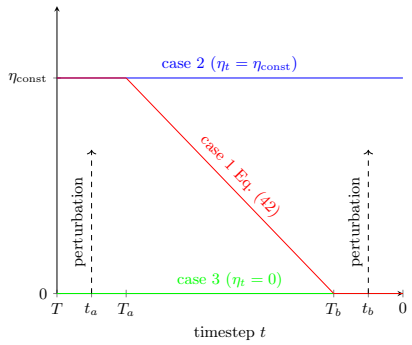


Fig. 1: Proposition 3. We assume that the score estimation model is accurate and only has two perturbations at t_a and t_b (t_b close to 0). We show that the η function of case 1 provides better sample quality than case 2 and case 3.

B Experimental Details

We provide our hyperparameters for diffusion inversion and real image editing in Tab. 1a and Tab. 1b. In general, all hyperparameters follow the official code implementation of the respective method. Additionally, Tab. 1c shows which backbone we used for each metric. Fig. 2 visualizes the η function for our three proposed Eta Inversion configurations. For EtaInv (1) and (2), we set the cross-attention map threshold to 0.2 and use a sampling count of $n = 10$. For EtaInv (3), we do not use region-dependent η and use a sampling count of $n = 1$.

C Searching the Optimal Eta Function

In this section, we present several hyperparameter study results on how we searched the optimal η function for EtaInv (2). We initialize all hyperparameters to EtaInv (2) by default. Tests are performed using PyTorch [24] on an NVIDIA V100 32GB GPU in 32-bit precision. Fig. 3 shows an overview of our experiments.

Table 1: Experimental details.

(a) Inversion hyperparameters. In general, parameter values follow the official implementation.

Inversion Method	Parameter	Value
DDPM Inv. [12]	`skip`	18
EDICT [30]	`init_image_strength`	1.0
	`leapfrog_steps`	`True`
	`mix_weight`	0.93
Null-text Inv. [19]	`early_stop_epsilon`	1e-05
	`num_inner_steps`	10
ProxNPI [10]	`dilate_mask`	1
	`prox`	10
	`quantile`	0.7
	`recon_lr`	1
	`recon_t`	400

(b) Editing hyperparameters. In general, parameter values follow the official implementation.

Editing Method	Parameter	Value
PtP [11]	`cross_replace_steps`	0.4
	`self_replace_steps`	0.6
	`equilizer_params_values`	2.0
PnP [29]	`pnp_f_t`	0.8
	`pnp_attn_t`	0.5
MasaCtrl [1]	`step`	4
	`layer`	10

(c) Backbone models for metric computation.

Metric	Backbone
CLIP similarity [26]	ViT-B16 [9]
DINO structural similarity [3]	ViT-B8 [9]
Perceptual Similarity (LPIPS) [32]	AlexNet [14]

C.1 Sign of Slope $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t}$

We first formulate η as a linear function for simplicity and explore how the slope $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t}$ of the graph affects the editing performance. Fig. 3a displays the tested η functions and Tab. 2a shows the metric values for each function. The results demonstrate that $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t} < 0$ shows better text-image alignment performance and a better editing effect than $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t} \geq 0$, which aligns with our theoretical findings. Therefore, we use a decreasing slope with $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t} < 0$ for further experiments.

C.2 Optimal $t, \eta-$intercepts

Next, we analyze how different linear η functions affect performance. We define several intercepts on the time axis $(0.3, 0.4, 0.5, 0.6, 0.7)$ and on the η axis (0.6, 0.7, 0.8, 0.9, 1.0) and linearly interpolate between two intercepts as displayed in Fig. 3b. Tab. 2b shows that a larger η and a smaller t (corresponding to applying noise even at later timesteps) improve text-image alignment while sacrificing structural similarity. EtaInv (2) with (η-intercept $= 0.7, t$-intercept $= 0.6$) provides a good balance for both.

C.3 Non-zero Concavity $\frac{\mathrm{d}^2\eta_{(T-t)}}{\mathrm{d}t^2}$

Furthermore, we perform several grid search experiments with non-linear η functions by introducing an exponent p (1/3, 1/2, 1, 2, 3) resulting in a concave

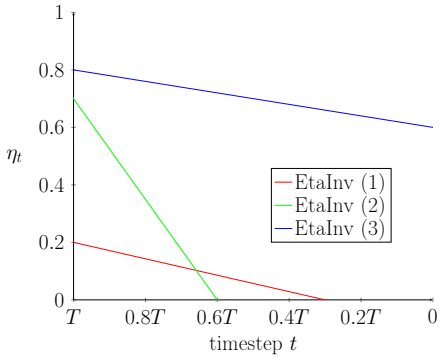


Fig. 2: η function for our three different EtaInv settings. **EtaInv (1)** uses smaller η favoring structural similarity, **EtaInv (2)** uses larger η favoring prompt alignment, and **EtaInv (3)** uses very large η even at later steps, optimized for style transfer.

(if $p < 1$) or convex (if $p > 1$) η function. First, we fix the $t, \eta-$intercept to EtaInv (2) and compute metrics for different exponents in Tab. 2c. We can observe that making EtaInv (2) concave improves image alignment since more total noise is injected and that a convex EtaInv (2) achieves better structural similarity since less noise is injected. Second, we provide an extensive grid search over various intercepts and exponents in Tab. 3 where each power shows a similar trade-off for alignment and similarity when altering the η and t intercept. Based on these experiments we find that there is no immediate benefit of introducing a non-linear η function and decide to fix it to linear for the remaining tests.

C.4 Sampling Count n

We test several different noise sampling counts $(1, 10, 10^2, 10^3, 10^4)$ in Tab. 4a and observe that a larger sampling count improves structural similarity while reducing text-image CLIP similarity. We argue that a larger sample count reduces the randomness in Eta Inversion by finding a noise that better approximates the true source-target branch distance.

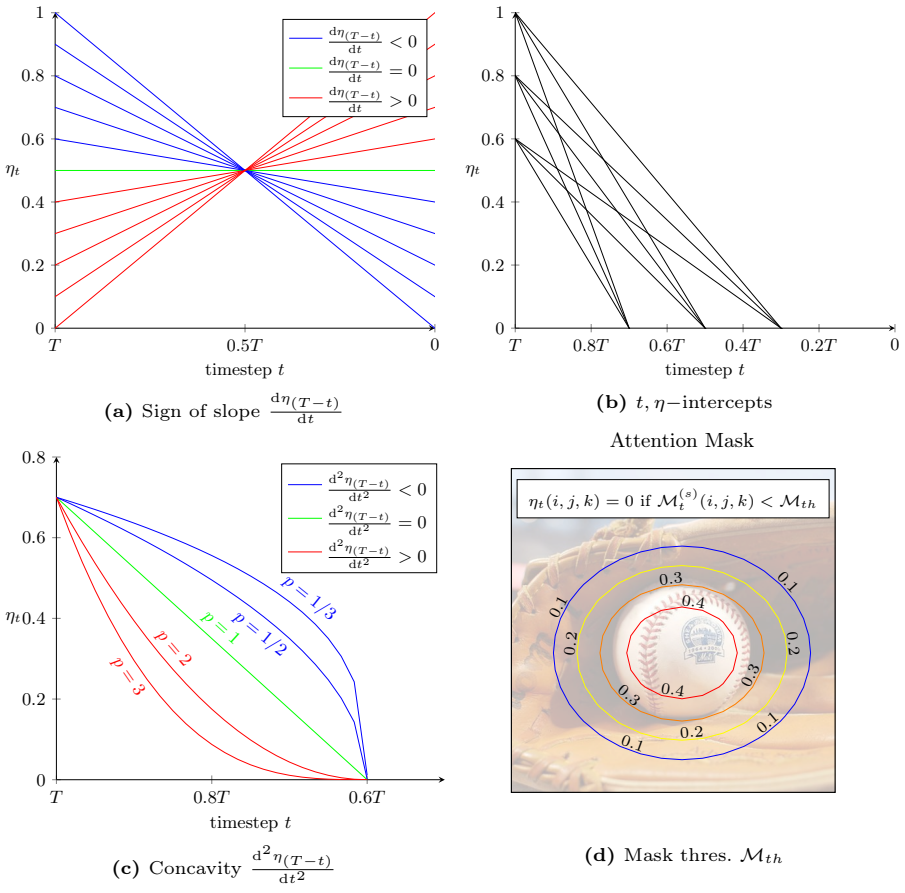
C.5 Cross-attention Map Source

There are three different sources for cross-attention maps: (i.) from the forward (inversion) path; (ii.) from the the backward path of the source latent; and (iii.) from the backward path of the target latent. Therefore, we provide results for each cross-attention map source in Tab. 4b while additionally including two more tests: GT, which uses the ground-truth foreground-background segmentation map provided by the dataset instead of cross-attention; and Source+Target which combines the backward attention maps from the source and the target branch with a max operation. We found that averaged attention masks from the forward path (i.) are most accurate and stable, since Eta Inversion injects no noise in

the forward path, leading to a balanced trade-off of text-image alignment and structural similarity.

C.6 Mask Threshold $\mathcal{M}_{th}$

Finally, Tab. 4c shows text-image alignment and structural similarity metrics for different attention map thresholds. Additionally, Smooth does not threshold the attention map but instead multiplies it to η, reducing η at low attention values, which did not achieve good results. For the threshold experiments, a larger attention threshold reduces the region where $\eta > 0$ and noise is injected (see Fig. 3d), consequently showing worse text-image alignment and better structural similarity. We find that the threshold $\mathcal{M}_{th} = 0.2$ achieves the best results.



(a) Sign of slope $\frac{d\eta_{(T-t)}}{dt}$

(b) $t, \eta-$intercepts

(c) Concavity $\frac{d^2\eta_{(T-t)}}{dt^2}$

(d) Mask thres. $\mathcal{M}_{th}$

Fig. 3: Exploring the optimal η function.

Table 2: Extensive parameter study for slope, intercept, and concavity of the η function evaluated on PIE-Bench with PtP. Hyperparameters are set to EtaInv (2) by default.

(a) Slope $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t}$ results. A negative/decreasing slope leads to better text-image alignment. An increasing slope may lead to better similarity but, in practice, fails to edit the image sufficiently since no noise is injected in the early diffusion steps, which is needed to edit high-level features.

	Metric ($\times 10^2$)				
	Text-Image Alignment (CLIP)		Structural Similarity		
Slope $\frac{\mathrm{d}\eta_{(T-t)}}{\mathrm{d}t}$	text-img ↑	text-cap. ↑	DINOv1 ↓	LPIPS ↓	BG-LPIPS ↓
−1.0	31.49	95.71	2.65	29.73	10.76
−0.8	31.45	96.14	2.49	28.50	10.37
−0.6	31.38	95.14	2.36	27.35	10.00
−0.4	31.33	94.71	2.22	26.22	9.65
−0.2	31.33	94.14	2.10	25.08	9.30
0.0	31.28	95.00	2.01	24.02	8.97
0.2	31.28	94.57	1.92	23.15	8.69
0.4	31.24	94.14	1.86	22.45	8.46
0.6	31.23	95.00	1.82	21.90	8.28
0.8	31.20	95.00	1.79	21.55	8.17
1.0	31.16	94.57	1.78	21.41	8.13

(b) $t, \eta-$intercept results. A larger η improves alignment while sacrificing similarity. A larger t reduces the total injected noise and improves similarity while worsening alignment. We chose $\eta = 0.7, t = 0.6$ for Eta Inversion (2).

	Metric ($\times 10^2$)									
	Text-Image Alignment (CLIP) ↑					Structural Similarity (DINOv1) ↓				
η \ t	0.3	0.4	0.5	**0.6**	0.7	0.3	0.4	0.5	**0.6**	0.7
0.6	31.22	31.19	31.19	31.14	31.12	1.76	1.71	1.67	1.60	1.53
0.7	31.30	31.30	31.26	**31.25**	31.18	1.93	1.88	1.82	**1.70**	1.61
0.8	31.32	31.35	31.34	31.26	31.24	2.11	2.03	1.96	1.83	1.71
0.9	31.42	31.41	31.40	31.33	31.29	2.29	2.20	2.14	1.97	1.84
1.0	31.45	31.43	31.43	31.44	31.33	2.47	2.39	2.33	2.14	1.95

(c) Concavity results. An exponent $p > 1$ leads to a convex graph, which reduces η and thus the total noise injected. Consequently, text-image alignment worsens while similarity improves. A linear η function ($p = 1$) is sufficient for a good balance of text-image alignment and structural similarity.

	Metric ($\times 10^2$)				
	Text-Image Alignment (CLIP)		Structural Similarity		
Exponent p	text-img ↑	text-cap. ↑	DINOv1 ↓	LPIPS ↓	BG-LPIPS ↓
1/3	31.30	95.43	1.98	23.86	8.89
1/2	31.29	94.29	1.89	22.99	8.61
1	31.25	95.43	1.70	21.14	8.00
2	31.14	95.00	1.55	19.34	7.41
3	31.08	95.29	1.48	18.38	7.11

Table 3: $t, \eta-$intercept results for various concavity/exponents. Every exponent shows a similar trade-off for alignment and similarity, e.g., when increasing η and decreasing t. We conclude that a linear function ($p = 1$) is sufficient for good editing results. Remaining hyperparameters are set to match Eta Inversion (2).

		Metric ($\times 10^2$)									
		Text-Image Alignment (CLIP) ↑					Structural Similarity (DINOv1) ↓				
Exponent p	$\eta \backslash t$	0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7
1/3	0.6	31.25	31.26	31.28	31.28	31.21	1.98	1.94	1.89	1.80	1.70
	0.7	31.37	31.33	31.35	31.30	31.26	2.19	2.13	2.07	1.98	1.86
	0.8	31.39	31.42	31.40	31.37	31.31	2.40	2.34	2.28	2.17	2.01
	0.9	31.43	31.46	31.45	31.40	31.39	2.60	2.54	2.46	2.36	2.19
	1.0	31.43	31.49	31.52	31.52	31.46	2.81	2.74	2.68	2.55	2.38
1/2	0.6	31.25	31.28	31.27	31.22	31.18	1.92	1.88	1.81	1.73	1.63
	0.7	31.31	31.30	31.31	31.29	31.26	2.10	2.05	1.98	1.89	1.77
	0.8	31.40	31.40	31.36	31.36	31.31	2.31	2.25	2.17	2.05	1.90
	0.9	31.45	31.46	31.40	31.39	31.38	2.50	2.44	2.35	2.25	2.06
	1.0	31.46	31.51	31.49	31.44	31.43	2.71	2.66	2.55	2.43	2.24
1	0.6	31.22	31.19	31.19	31.14	31.12	1.76	1.71	1.67	1.60	1.53
	0.7	31.30	31.30	31.26	31.25	31.18	1.93	1.88	1.82	1.70	1.61
	0.8	31.32	31.35	31.34	31.26	31.24	2.11	2.03	1.96	1.83	1.71
	0.9	31.42	31.41	31.40	31.33	31.29	2.29	2.20	2.14	1.97	1.84
	1.0	31.45	31.43	31.43	31.44	31.33	2.47	2.39	2.33	2.14	1.95
2	0.6	31.15	31.13	31.11	31.09	31.05	1.60	1.57	1.53	1.48	1.44
	0.7	31.23	31.22	31.16	31.14	31.10	1.71	1.65	1.60	1.55	1.49
	0.8	31.28	31.28	31.24	31.17	31.10	1.84	1.78	1.69	1.62	1.54
	0.9	31.34	31.30	31.28	31.24	31.14	1.99	1.90	1.82	1.72	1.60
	1.0	31.41	31.39	31.34	31.28	31.18	2.14	2.05	1.93	1.83	1.68
3	0.6	31.11	31.10	31.08	31.05	31.04	1.53	1.50	1.46	1.43	1.39
	0.7	31.16	31.15	31.11	31.08	31.05	1.60	1.56	1.52	1.48	1.43
	0.8	31.24	31.21	31.16	31.09	31.05	1.69	1.64	1.58	1.52	1.47
	0.9	31.28	31.26	31.20	31.14	31.07	1.81	1.74	1.67	1.58	1.51
	1.0	31.33	31.29	31.25	31.18	31.10	1.92	1.85	1.77	1.66	1.56

Table 4: Extensive parameter study for noise sample count, attention source, and attention threshold evaluated on PIE-Bench with PtP. Hyperparameters are set to EtaInv (2) by default.

(a) Noise sample count n results. Large sample counts generally lead to better similarity, while lower sample counts achieve better prompt alignment.

	Metric ($\times 10^2$)				
	Text-Image Alignment (CLIP)		Structural Similarity		
Sample count n	text-img ↑	text-cap. ↑	DINOv1 ↓	LPIPS ↓	BG-LPIPS ↓
1	31.19	95.57	1.72	21.26	8.08
10	31.25	95.43	1.70	21.14	8.00
10^2	31.22	95.00	1.73	21.27	8.07
10^3	31.12	94.71	1.70	21.11	8.04
10^4	31.13	95.71	1.70	21.06	7.99

(b) Cross-attention map source results. **GT** uses ground-truth foreground-background maps from PIE-Bench. **Forward** are cross-attention maps collected during the forward path. **Forward (mean)** averages all forward attention maps to one map. **Backward Source** and **Backward Target** are the attention maps from the backward source and target path respectively. **Backward Source+Target** combines both activations to one map via the max-operator. We found that forward (mean) provides the best balance for text alignment and similarity.

	Metric ($\times 10^2$)				
	Text-Image Alignment (CLIP)		Structural Similarity		
Attention source	text-img ↑	text-cap. ↑	DINOv1 ↓	LPIPS ↓	BG-LPIPS ↓
No mask	31.27	95.43	1.85	22.77	9.03
GT	31.22	94.71	1.67	20.67	7.00
Forward	31.24	95.43	1.75	21.59	8.27
Forward (Mean)	31.25	95.43	1.70	21.14	8.00
Backward Source	31.26	95.14	1.81	22.30	8.74
Backward Target	31.22	94.71	1.80	22.27	8.73
Backward Source+Target	31.25	94.86	1.83	22.48	8.84

(c) Cross-attention threshold results. A higher threshold reduces the region where noise is being injected, resulting in less editing and better similarity while negatively affecting alignment. **Smooth** multiplies η with the attention map activations instead of thresholding it, resulting in smaller η being applied to less activated regions. $\mathcal{M}_{\mathbf{th}} = \mathbf{0.2}$ provides a good tradeoff between alignment and similarity metrics.

	Metric ($\times 10^2$)				
	Text-Image Alignment (CLIP)		Structural Similarity		
Attention threshold $\mathcal{M}_{th}$	text-img ↑	text-cap. ↑	DINOv1 ↓	LPIPS ↓	BG-LPIPS ↓
No mask	31.27	95.43	1.85	22.77	9.03
0.1	31.25	95.71	1.81	22.21	8.59
0.2	31.25	95.43	1.70	21.14	8.00
0.3	31.15	94.71	1.64	20.25	7.62
0.4	31.08	95.29	1.56	19.35	7.24
0.5	31.05	95.71	1.51	18.82	7.02
Smooth	31.05	94.99	1.51	18.77	7.02

D Existing Diffusion Inversion Methods

We summarize existing diffusion inversion methods in Tab. 5. The table divides each forward and backward path into (η, w) and strategy. Former indicates the used η parameter (DDIM or DDPM) with the guidance scale parameter w, while strategy indicates additional effort by the inversion method to reduce the gap to the (ideal) forward path.

Table 5: Summary of existing diffusion inversion methods.

	forward path		backward path	
Inversion Method	(η, w)	strategy	(η, w)	strategy
DDIM Inv. [27]	$(0, 1)$	-	$(0, 7.5)$	-
NTI [19]	$(0, 1)$	-	$(0, 7.5)$	optimized $\varnothing_t$
NPI [18]	$(0, 1)$	-	$(0, 7.5)$	$\varnothing \leftarrow c^{(s)}$
ProxNPI [10]	$(0, 1)$	-	$(0, 7.5)$	$\varnothing \leftarrow c^{(s)}$, modified $\tilde{\epsilon}_\theta$
Direct Inv. [13]	$(0, 1)$	-	$(0, 7.5)$	$\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \boldsymbol{x}_{t-1}^{(s)^*}$
EDICT [30]	$(0, 3)$	Coupled Transformations	$(0, 3)$	Coupled Transformations
DDPM Inv. [12]	-	$q(\boldsymbol{x}_t \vert \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I)$	$(1, 3.5)$	$\boldsymbol{x}_{t-1}^{(s)'} \leftarrow \boldsymbol{x}_{t-1}^{(s)^*}$, modified ϵ_{add}

E Text-guided Image Editing Methods

In this section, we introduce commonly used training-free text-guided image editing methods. We also state the inversion method used by the original paper for real image editing.

E.1 Prompt-to-Prompt (PtP) [11]

Editing To forward information from the source to the target backward path, PtP replaces the cross-attention maps of the target with the maps from the source. Since PtP needs to know which attention maps to exchange, each word in the source prompt must be matched to a word in the target prompt. Thus, PtP introduces restrictions for specifying an appropriate source and target prompt.

Inversion The proposed Prompt-to-Prompt in [11] applies DDIM Inversion ($w = 1$) and DDIM sampling ($w = 7.5$) without any additional strategy, which negatively affects structural similarity.

E.2 MasaCtrl [1]

Editing MasaCtrl focuses on non-rigid real image editing (motion editing). Unlike PtP, MasaCtrl replaces self-attention maps instead of cross-attention maps. For real image editing, MasaCtrl uses an empty prompt for the source prompt. Otherwise, when performing synthetic editing and if a source prompt is specified, MasaCtrl optionally uses masked guidance by obtaining a mask from the cross-attention maps for the word to edit.

Inversion MasaCtrl applies DDIM Inversion ($w = 1$) and replaces the source prompt $c^{(s)}$ with an empty prompt $\varnothing$. Consequently, MasaCtrl does not require a source prompt for inversion.

E.3 Plug-and-Play (PnP) [29]

Editing Plug-and-Play also focuses on modifying self-attention maps similar to MasaCtrl. In addition, PnP performs spatial feature injection in U-Net's decoder layers from the source to the target branch. They thereby report better structural preservation than PtP.

Inversion Same as MasaCtrl, Plug-and-Play applies DDIM Inversion ($w = 1$) and replaces the source prompt $c^{(s)}$ with an empty prompt $\varnothing$.

F Image Editing Metrics

It is unclear how to evaluate image editing performance. Instead of having a single metric measuring both text-image alignment with the target prompt and structural similarity with the source image, previous methods focus on evaluating these two concepts separately. For text-image alignment, CLIP [26] is commonly used, while for structural similarity, a strong feature extractor like DINO [3] is utilized along with more classical evaluation metrics such as MS-SSIM [31]. Below, we give a detailed explanation of all our metrics.

F.1 Text-Image Alignment

Text$^{(t)}$ - Image$^{(t)}$ CLIP Similarity (text-img) This metric first computes the target prompt text embeddings and the output image embeddings using CLIP, normalizes them, and finally computes the dot product of the two embeddings. The larger the dot product value, the better the target prompt and the output image are aligned.

Text$^{(t)}$ - Image Caption$^{(t)}$ [4] CLIP Similarity (text-cap) Similar to text-image CLIP similarity, but instead of using the output image embeddings, a caption of the output image is obtained via BLIP [16]. We then embed the generated caption by CLIP and compute the dot product of the target prompt text embeddings and the generated caption embeddings. Since there is a gap between CLIP's image space and CLIP's text space, this has the advantage that both embeddings lie in CLIP's text space.

Directional CLIP Similarity [25] (directional) Unlike the above two metrics, this metric additionally incorporates the source prompt and the source image. The idea of this metric is that, in CLIP space, the direction from the source prompt to the target prompt should match the direction from the source image to

the output image. Therefore, directional CLIP similarity computes the respective embeddings of both prompts and images and then retrieves the text direction and image direction by taking their respective difference. Lastly, the dot product of those two directions serves as the metric value.

CLIP Accuracy (acc) This metric was originally introduced in [23] and computes the ratio of output images where the text-image CLIP similarity is higher with the target prompt than with the source prompt. We noticed in our experiments that most inversion and editing methods reach a perfect score of 1. Thus, we decided to use text-caption similarity via BLIP instead of text-image similarity.

F.2 Structural Similarity

DINO [3] Self-similarity DINO self-similarity measures the similarity between the source image and the target image by obtaining embeddings via DINO and computing their MSE loss. DINO effectively extracts structural features of images compared to other feature extraction models. We provide metrics for both DINOv1 [3] and DINOv2 [22] but focus on the more commonly used DINOv1.

LPIPS [32] LPIPS is another metric similar to DINO self-similarity operating on AlexNet [14] and focusing on matching human perception.

BG-LPIPS [13, 23] BG-LPIPS computes the LPIPS only on the background part of the source and output image, which should not be edited. This metric works well for edits like replacing or editing single objects rather than performing style transfer, where there is no clear background. The background mask is provided by the user or dataset.

MS-SSIM [31] Multi-scale structural similarity (MS-SSIM) is an improved version of SSIM and computes similarity over various image scales by consecutive downsampling.

F.3 Text-Image Alignment and Structural Similarity

VIEScore [15] VIEScore is specifically introduced for evaluating image generation and editing performance using GPT-4V(ision) [21]. For image editing, VIEScore assesses both text-image alignment and structural similarity. It employs a carefully designed prompt that rates the editing performance with three separate scores, ranging from 0 to 10, where higher is better: overall score, alignment score, and similarity score. The overall score evaluates the image editing in general, whereas the alignment score and similarity score focus on text-image alignment and structural similarity, respectively.

G Additional Quantitative and Qualitative Results

G.1 Additional Quantitative Results

Tab. 6 provides results for PIE-Bench editing with additional metrics. Fig. 4 shows trade-off plots for PtP, PnP and MasaCtrl. PIE-Bench also divides all 700 edits into ten categories, such as random edits or changing objects. We thus evaluate and present our metrics for each category individually in Tab. 7. Additionally, we provide supplementary metrics for the GPT-4 [21] based VIEScore [15] on subsets of PIE-Bench in Tab. 8, and pairwise compare our method with Direct Inversion through human evaluation in Tab. 9. Furthermore, Tab. 10 provides metrics for the ImageNetR-TI2I [29] dataset. In all additional experiments, our method achieves state-of-the-art results in most cases.

In Tab. 11, we evaluate the reconstruction accuracy and inference time of our method. We observe that our method achieves perfect reconstruction, matching VAE reconstruction, which serves as the upper bound. Additionally, our method introduces negligible overhead and offers an inference speed similar to standard DDIM Inversion, when compared to both standard inversion and all three tested editing methods.

G.2 Additional Qualitative Results

In addition to the above quantitative results, we provide more qualitative comparisons of our method with various other inversion and editing approaches. Fig. 5, Fig. 6, and Fig. 7 show qualitative results on PIE-Bench [13] images using PtP [8], PnP [29], and MasaCtrl [1] respectively. Additionally, Fig. 8 shows examples where a larger η as in EtaInv (3) leads to better editing results (e.g., style transfer). Finally, Fig. 9 shows the impact of η on image editing where higher η values result in better target prompt alignment while negatively influencing structural similarity.

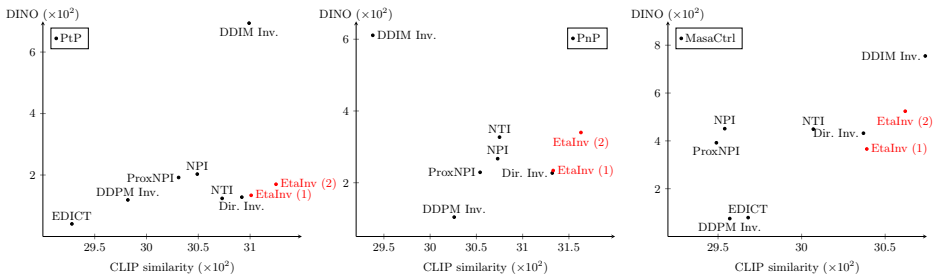


Fig. 4: Visualization of CLIP text-image metrics (higher is better) and DINO metrics (lower is better) on PIE-Bench for PtP (left), PnP (middle), and MasaCtrl (right).

Table 6: PIE-Bench [13] evaluation with additional metrics.

		Metric ($\times 10^2$)								
		Text-Image Alignment (CLIP)				Structural Similarity				
Editing	Inversion	text-img ↑	text-cap ↑	directional ↑	acc ↑	DINOv1 ↓	DINOv2 ↓	LPIPS ↓	BG-LPIPS ↓	MS-SSIM ↑
PtP	DDIM Inv. [27]	30.99	75.73	1.49	94.57	6.94	0.98	46.65	24.97	61.93
	Null-text Inv. [19]	30.73	74.41	3.35	92.57	1.24	0.32	15.13	5.69	88.76
	NPI [18]	30.49	74.28	3.80	92.71	2.03	0.43	19.28	8.24	85.93
	ProxNPI [10]	30.31	74.00	3.48	92.43	1.92	0.41	17.69	7.76	86.83
	EDICT [30]	29.28	72.69	0.83	92.71	0.41	0.16	6.65	3.10	93.71
	DDPM Inv. [12]	29.43	73.12	0.69	92.71	0.42	0.18	6.87	3.27	93.38
	Direct Inv. [13]	30.92	75.76	2.34	94.71	1.28	0.32	15.79	6.33	88.34
	EtaInv (1)	31.01	76.08	2.39	95.00	1.34	0.34	16.58	6.57	87.60
	EtaInv (2)	31.25	76.20	2.67	95.43	1.70	0.40	21.14	8.00	83.19
	EtaInv (3)	31.52	76.27	3.21	95.00	3.05	0.59	33.57	13.31	68.86
PnP	DDIM Inv. [27]	29.38	69.38	4.17	85.57	6.11	0.94	40.84	20.84	69.30
	Null-text Inv. [19]	30.75	73.77	4.62	90.43	3.27	0.61	30.51	14.17	78.37
	NPI [18]	30.73	73.83	4.84	91.29	2.67	0.53	26.18	11.57	81.81
	ProxNPI [10]	30.54	74.02	4.12	90.71	2.29	0.46	21.76	9.57	84.43
	EDICT [30]	24.69	60.09	2.58	63.43	4.26	0.73	30.22	14.96	76.46
	DDPM Inv. [12]	30.26	73.72	2.02	94.86	1.04	0.29	12.50	5.84	89.35
	Direct Inv. [13]	31.32	76.13	3.10	95.14	2.27	0.48	25.59	12.98	82.34
	EtaInv (1)	31.33	76.47	2.66	94.86	2.34	0.50	27.33	14.05	80.85
	EtaInv (2)	31.63	76.36	3.22	95.29	3.40	0.64	36.59	18.72	70.83
	EtaInv (3)	31.92	77.10	3.98	94.57	5.16	0.88	50.39	26.61	50.53
Masa	DDIM Inv. [27]	30.74	75.22	1.37	95.00	7.55	1.02	47.68	25.37	60.37
	Null-text Inv. [19]	30.07	72.79	2.64	93.00	4.49	0.68	25.02	11.92	77.35
	NPI [18]	29.54	71.17	2.95	87.29	4.51	0.70	26.03	12.41	77.93
	ProxNPI [10]	29.49	71.50	2.59	88.14	3.92	0.63	22.99	10.99	80.51
	EDICT [30]	29.68	73.46	0.81	93.29	0.79	0.23	8.59	4.20	92.32
	DDPM Inv. [12]	29.57	73.17	0.70	93.00	0.75	0.22	8.65	4.12	91.51
	Direct Inv. [13]	30.37	74.50	1.39	94.57	4.32	0.66	26.91	13.76	76.95
	EtaInv (1)	30.39	74.00	1.62	93.14	3.66	0.59	23.12	11.57	79.61
	EtaInv (2)	30.62	74.32	1.96	93.86	5.24	0.80	33.07	16.64	68.18
	EtaInv (3)	30.72	74.24	2.22	94.00	7.21	1.06	44.97	23.65	47.64

Table 7: Per category evaluation for PIE-Bench [13]: Categories are defined as: (0) random; (1) change object; (2) add object; (3) delete object; (4) change attribute content; (5) change attribute pose; (6) change attribute color; (7) change attribute material; (8) change background; and (9) change style.

		Metric ($\times 10^2$)																			
		Text-Image alignment (CLIP) ↑										Structural similarity (DINOv1) ↓									
Editing	Inversion	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
PtP	DDIM Inv.	30.73	30.79	31.28	29.93	31.29	32.21	31.18	32.48	30.80	31.00	6.79	6.71	6.70	7.23	6.62	6.21	7.50	8.02	7.56	6.47
	NTI	30.68	30.42	30.24	29.52	30.33	31.63	31.31	31.60	30.55	32.06	1.41	1.45	0.93	1.15	0.90	0.76	1.17	1.11	1.31	1.60
	NPI	30.61	30.29	30.12	29.27	30.49	31.37	30.89	30.97	30.20	31.44	1.98	2.65	1.91	1.97	1.37	1.52	2.00	1.84	2.15	2.22
	ProxNPI	30.44	30.21	29.97	29.29	30.39	31.32	30.75	30.87	29.97	30.88	1.85	2.55	1.85	1.90	1.29	1.48	1.91	1.81	1.98	2.02
	EDICT	28.50	28.31	29.52	29.05	29.77	31.77	29.58	30.45	29.21	29.45	0.47	0.46	0.35	0.40	0.35	0.38	0.42	0.38	0.41	0.41
	DDPM Inv.	28.55	28.49	29.74	29.08	29.80	32.09	29.80	30.61	29.28	29.78	0.45	0.43	0.37	0.41	0.36	0.41	0.44	0.40	0.42	0.43
	Direct Inv.	30.75	30.43	30.86	29.98	30.50	31.98	31.27	32.14	30.79	31.71	1.37	1.24	1.12	1.23	1.01	1.15	1.18	1.13	1.32	1.64
	EtaInv (1)	30.89	30.50	30.98	30.03	30.66	31.96	31.40	32.23	30.74	31.89	1.44	1.27	1.21	1.34	1.09	1.22	1.21	1.17	1.33	1.74
	EtaInv (2)	31.07	31.02	31.20	30.11	30.90	31.92	31.63	32.27	31.15	32.20	1.70	1.55	1.65	1.80	1.32	1.52	1.45	1.45	1.70	2.33
	EtaInv (3)	31.26	31.09	31.41	30.34	31.00	32.17	32.01	32.24	31.67	32.85	2.92	2.74	2.85	3.42	2.26	2.68	2.53	2.28	3.47	4.19
PnP	DDIM Inv.	29.71	29.15	29.28	28.12	29.35	29.70	30.01	30.92	28.32	30.21	6.34	6.22	5.58	6.21	4.93	5.71	5.92	6.08	6.97	6.09
	NTI	30.55	30.51	30.07	29.47	30.52	30.60	30.87	31.60	30.99	32.79	3.23	3.71	2.97	3.30	2.24	2.43	3.28	2.92	3.55	3.98
	NPI	30.67	30.19	30.22	29.75	30.45	31.24	31.05	31.48	30.55	32.37	2.61	3.46	2.24	2.82	1.78	1.71	2.54	2.44	2.75	3.30
	ProxNPI	30.54	30.09	30.09	29.70	30.30	31.32	30.89	31.19	30.36	31.66	2.17	3.09	2.02	2.52	1.57	1.51	2.28	2.07	2.34	2.52
	NSI	30.81	30.76	31.37	30.54	30.90	32.23	31.66	32.27	31.35	32.34	2.04	2.33	1.85	2.05	2.10	1.88	2.22	2.20	2.35	2.18
	EDICT	24.53	24.03	24.46	23.86	25.53	25.79	23.37	26.37	24.94	25.32	4.17	4.59	4.28	4.40	3.33	4.18	4.47	3.73	4.60	4.22
	DDPM Inv.	29.72	29.79	30.26	29.66	30.23	32.09	30.56	31.64	30.14	30.64	1.10	1.17	0.87	1.14	0.97	0.92	1.05	1.01	1.07	0.97
	Direct Inv.	30.81	30.76	31.37	30.54	30.89	32.23	31.66	32.27	31.35	32.34	2.04	2.33	1.85	2.05	2.10	1.88	2.22	2.20	2.35	2.18
	EtaInv (1)	30.95	30.76	31.20	30.61	30.74	32.16	31.85	32.27	31.56	32.33	2.26	2.56	2.04	2.32	2.32	2.08	2.37	2.35	2.56	2.46
	EtaInv (2)	31.44	31.38	31.39	30.76	30.72	32.33	32.30	32.24	31.54	32.85	3.15	3.60	3.08	3.55	3.28	3.09	3.45	3.20	3.73	3.77
	EtaInv (3)	31.57	31.62	31.69	31.00	30.99	32.28	32.52	33.12	32.17	33.12	4.86	5.69	4.90	5.81	4.75	4.90	4.97	4.86	5.32	5.16
Masa	DDIM Inv.	29.85	30.81	31.18	30.96	30.93	32.31	30.06	32.22	30.38	30.67	7.64	7.65	6.76	8.88	6.58	6.36	7.94	8.16	7.96	6.90
	NTI	29.71	30.29	30.13	29.81	29.89	31.78	29.84	31.14	29.58	29.97	5.09	5.15	3.98	5.18	2.89	2.93	5.20	4.26	4.39	4.07
	NPI	29.19	29.88	29.55	29.34	29.63	31.09	28.55	30.74	29.07	29.60	4.88	5.77	4.28	5.24	2.99	3.51	4.41	4.20	4.03	4.04
	ProxNPI	29.08	29.85	29.50	29.49	29.68	31.13	28.55	30.70	28.93	29.38	4.15	5.23	3.71	4.83	2.41	3.21	3.84	3.68	3.35	3.36
	NSI	29.23	29.75	30.32	30.01	29.97	31.52	29.86	30.94	29.60	29.96	3.94	4.37	3.30	5.00	3.03	3.22	3.56	4.74	4.03	3.55
	EDICT	28.75	28.45	30.25	29.63	29.88	32.11	29.91	30.81	29.74	29.93	0.80	0.83	0.65	1.08	0.70	0.77	0.68	0.83	0.71	0.73
	DDPM Inv.	28.72	28.57	30.09	29.51	29.93	32.02	29.58	30.53	29.48	29.78	0.77	0.94	0.61	1.02	0.62	0.64	0.63	0.69	0.71	0.66
	Direct Inv.	29.63	30.30	30.65	30.38	30.17	32.04	30.00	31.74	29.87	30.22	3.40	3.99	2.82	4.40	2.71	2.94	2.87	4.17	3.26	2.84
	EtaInv (1)	29.81	30.38	30.51	30.51	30.19	32.09	30.24	31.84	29.83	30.32	3.64	4.25	2.88	4.59	2.95	3.29	3.22	4.44	3.78	3.23
	EtaInv (2)	30.04	30.74	30.81	30.42	30.71	31.94	30.30	31.84	30.15	30.89	5.15	5.54	4.64	6.42	4.29	4.75	5.24	5.39	5.48	4.90
	EtaInv (3)	30.10	30.81	30.83	30.67	30.74	32.10	30.15	31.88	30.54	30.82	7.30	7.61	6.58	8.40	6.06	6.44	7.14	7.56	7.41	6.69

Table 8: VIEScore [15] for the PIE-Bench random (0) and change style (9) subsets. VIEScore ranges from 0 to 10 (higher is better) and provides three scores: overall, text-image alignment (align.), and structural similarity (sim.). For PtP and MasaCtrl, our method achieves state-of-the-art overall and alignment scores.

		PIE-Bench Random			PIE-Bench Style		
Editing	Inversion	overall ↑	align. ↑	sim. ↑	overall ↑	align. ↑	sim. ↑
PtP	DDIM Inv. [27]	4.31	4.78	5.36	1.72	1.79	3.10
	Null-text Inv. [19]	5.43	5.83	8.07	4.03	4.16	8.01
	NPI [18]	5.87	6.35	7.83	3.85	4.09	7.18
	ProxNPI [10]	5.24	5.65	7.63	3.18	3.33	7.56
	EDICT [30]	1.48	1.56	8.06	0.43	0.43	9.19
	DDPM Inv. [12]	1.40	1.48	8.37	0.38	0.38	8.63
	Direct Inv. [13]	5.27	5.66	8.08	3.18	3.24	7.09
	Eta Inversion (1)	4.98	5.34	8.13	3.16	3.23	7.09
	Eta Inversion (2)	5.60	6.07	8.01	3.48	3.65	6.70
	Eta Inversion (3)	6.21	6.73	7.65	4.24	4.54	6.06
PnP	DDIM Inv. [27]	4.26	4.99	5.13	4.25	4.60	5.43
	Null-text Inv. [19]	6.02	6.63	7.23	5.99	6.30	7.80
	NPI [18]	6.37	6.91	7.68	5.94	6.34	7.96
	ProxNPI [10]	5.79	6.29	7.83	4.70	4.88	8.10
	EDICT [30]	2.60	2.87	4.17	2.35	2.69	3.90
	DDPM Inv. [12]	4.12	4.39	8.49	2.03	2.06	7.48
	Direct Inv. [13]	5.04	5.43	7.46	4.74	4.93	7.94
	Eta Inversion (1)	5.06	5.44	7.18	3.53	3.63	7.33
	Eta Inversion (2)	5.56	5.99	6.91	4.44	4.81	6.24
	Eta Inversion (3)	5.88	6.49	6.63	5.34	5.78	6.55
Masa	DDIM Inv. [27]	2.55	2.91	3.93	1.29	1.39	2.63
	Null-text Inv. [19]	3.86	4.13	7.64	1.84	1.99	6.81
	NPI [18]	3.81	4.19	6.97	2.74	2.91	6.61
	ProxNPI [10]	3.65	4.01	7.37	1.68	1.74	6.38
	EDICT [30]	1.56	1.63	8.64	0.57	0.58	8.34
	DDPM Inv. [12]	1.31	1.39	8.91	0.43	0.43	8.71
	Direct Inv. [13]	2.61	2.78	7.48	1.24	1.29	5.78
	Eta Inversion (1)	3.19	3.33	8.04	1.18	1.21	7.30
	Eta Inversion (2)	4.36	4.53	7.48	2.15	2.18	6.03
	Eta Inversion (3)	3.83	4.11	6.66	2.48	2.55	5.55

Table 9: Human evaluation on PIE-Bench. We conducted 740 comparisons on PIE-Bench's random (0) and change style (9) subsets, where participants were asked to choose if Direct Inversion's output is better, if Eta Inversion's output is better, or if it is a tie. In non-tie cases, Eta Inversion was preferred approximately 2 to 3.5 times more than Direct Inversion.

Editing method: PtP	Tie	Direct Inversion	**Eta Inversion**
PIE-Bench (random)	69.76%	10.70%	19.53%
PIE-Bench (change style)	52.50%	10.63%	36.86%

Table 10: ImageNet-R-TI2I [29] evaluation with various metrics. For all three editing methods, our approach achieves the best CLIP text-image and text-caption metrics.

		Metric ($\times 10^2$)							
		Text-Image Alignment (CLIP)				Structural Similarity			
Editing	Inversion	text-img ↑	text-cap ↑	directional ↑	acc ↑	DINOv1 ↓	DINOv2 ↓	LPIPS ↓	MS-SSIM ↑
PtP	DDIM Inv. [27]	30.06	69.87	2.49	98.89	8.35	1.07	50.48	55.17
	Null-text Inv. [19]	30.32	70.58	5.40	94.44	2.18	0.53	26.10	81.51
	NPI [18]	30.34	68.43	6.15	94.44	3.25	0.63	28.71	79.33
	ProxNPI [10]	29.99	69.09	5.41	92.22	3.08	0.59	25.61	80.94
	EDICT [30]	29.39	67.69	2.62	94.44	0.94	0.32	12.40	89.85
	DDPM Inv. [12]	28.98	67.14	0.69	97.78	0.46	0.22	7.87	92.22
	Direct Inv. [13]	30.34	70.23	3.27	97.78	1.80	0.45	21.98	82.99
	EtaInv (1)	30.54	70.89	3.40	97.78	1.89	0.46	22.68	82.17
	EtaInv (2)	30.94	71.17	4.24	94.44	3.70	0.67	38.66	64.84
	EtaInv (3)	31.44	73.02	5.25	98.89	5.05	0.81	45.65	56.46
PnP	DDIM Inv. [27]	28.83	67.13	7.91	87.78	7.45	1.11	46.03	62.50
	Null-text Inv. [19]	30.35	70.75	7.49	90.00	4.75	0.79	41.87	68.19
	NPI [18]	30.82	70.77	8.11	92.22	4.23	0.74	37.62	72.69
	ProxNPI [10]	30.20	69.28	6.28	96.67	3.68	0.64	31.20	76.82
	EDICT [30]	23.44	58.15	5.65	57.78	6.54	0.95	43.21	62.90
	DDPM Inv. [12]	29.91	69.18	3.20	93.33	1.43	0.40	16.63	85.14
	Direct Inv. [13]	30.80	71.40	5.36	98.89	3.02	0.60	30.71	77.23
	EtaInv (1)	30.97	71.64	5.57	97.78	3.30	0.62	32.50	75.15
	EtaInv (2)	31.27	73.31	6.86	97.78	5.24	0.86	47.84	54.70
	EtaInv (3)	31.61	73.85	7.33	97.78	5.86	0.96	53.60	44.77
Masa	DDIM Inv. [27]	29.67	68.81	1.23	98.89	9.80	1.18	53.96	51.68
	Null-text Inv. [19]	29.61	65.74	2.97	90.00	6.79	0.93	34.57	67.39
	NPI [18]	28.07	63.98	3.91	84.44	7.39	1.02	36.60	67.09
	ProxNPI [10]	27.42	63.35	2.70	87.78	6.42	0.90	32.28	70.95
	EDICT [30]	29.38	68.50	3.33	95.56	1.96	0.44	16.93	85.68
	DDPM Inv. [12]	29.28	67.91	0.93	100.00	0.93	0.31	11.15	89.09
	Direct Inv. [13]	29.40	67.05	0.53	97.78	5.70	0.80	30.59	71.67
	EtaInv (1)	29.49	67.67	0.62	97.78	5.99	0.83	31.98	69.71
	EtaInv (2)	29.71	67.88	0.92	98.89	7.82	1.05	42.91	52.07
	EtaInv (3)	29.92	68.86	1.19	97.78	8.88	1.19	48.30	42.22

Table 11: Reconstruction benchmark of inversion methods on the COCO training set [7] (left). Inference time of inversion methods (right). Our method matches VAE reconstruction metrics, thus demonstrating perfect reconstruction. Additionally, our inference time is only slightly higher than DDIM Inversion for inversion (Inv.) and the three editing methods PtP, PnP, and MasaCtrl. Inference time is measured on an NVIDIA V100.

	Reconstruction error			Inference time (s)			
Method	PSNR ↑	LPIPS ↓	SSIM ↑	Inv.	PtP	PnP	Masa
DDIM Inv. [27]	14.3	0.500	0.469	22.1	23.5	19.3	25.6
Null-text Inv. [19]	26.3	0.072	0.745	200.9	202.7	198.4	204.6
NPI [18]	24.0	0.147	0.697	22.1	23.6	19.3	25.6
ProxNPI [10]	26.6	0.067	0.751	25.7	27.3	22.8	29.3
EDICT [30]	26.6	0.067	0.751	44.0	53.9	45.1	55.8
DDPM Inv. [12]	26.6	0.067	0.751	32.1	41.0	35.5	43.0
Direct Inv. [13]	26.6	0.067	0.751	22.1	23.5	19.3	25.6
Eta Inversion	26.6	0.067	0.751	22.8	24.3	19.9	26.4
Eta Inversion w/o mask	26.6	0.067	0.751	22.3	23.9	19.4	25.9
VAE Reconstruction	26.6	0.067	0.751	-	-	-	-

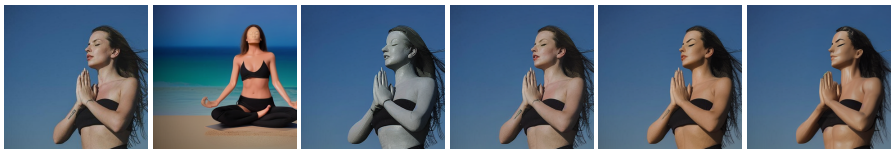
*"a **kitten** walking through the grass" → "a **duck** walking through the grass"*

*"painting of a **shepherd** dog sitting in a laundry room next to a washing machine" → "painting of a **poodle** dog sitting in a laundry room next to a washing machine"*

*"a detailed oil painting of a **calm** beautiful woman with stars in her hair" → "a detailed oil painting of a **laughing** beautiful woman with stars in her hair"*

*"a **cat** sitting next to a mirror" → "a **tiger** sitting next to a mirror"*

*"a woman in a black bikini top and yoga pants is meditating" → "a **wax statue of** woman in a black bikini top and yoga pants is meditating"*



*"a slanted mountain bicycle on the road in front of a building" → "a slanted **rusty** mountain bicycle on the road in front of a building"*

Source | DDIM Inv [27] | NTI [19] | EDICT [30] | Dir. Inv. [13] | **EtaInv (2)**

Fig. 5: Additional qualitative results for PtP editing.

"a man wearing a tie" → *"a man wearing a* ***black and yellow stripes*** *tie "*

"a woman in front of a glowing yellow light" → *"a woman* ***riding a lion*** *in front of a glowing yellow light"*

"a basket of books and a ***cup****"* → *"a basket of books and a* ***candle****"*

"two ***red and white*** *toy gnomes are sitting on a snow covered surface"* → *"two* ***blue and green*** *toy gnomes are sitting on a snow covered surface"*

"a view of the mountains covered in ***snow****"* → *"a view of the mountains covered in* ***leaves****"*

"a boat is docked on a lake in the ***heavy fog****"* → *"a boat is docked on a lake in the* ***sunny day****"*

Source | DDIM Inv [27] | NTI [19] | EDICT [30] | Dir. Inv. [13] | **EtaInv (2)**

Fig. 6: Additional qualitative results for PnP editing.

*"the christmas illustration of a santa's **laughing** face"* → *"the christmas illustration of a santa's **angry** face"*

*"an illustration of a **cat** sitting on top of a rock"* → *"an illustration of a **bear** sitting on top of a rock"*

"a cat standing on fence" → *"a cat **wearing hat** standing on fence"*

*"a glass of red **drink** on the beach"* → *"a glass of red **wine** on the beach"*

*"a painting of a **cabin** in the snow with mountains in the background"* → *"a painting of a **car** in the snow with mountains in the background"*

*"a lion in a suit sitting at a table **with a laptop**"* → *"a lion in a suit sitting at a table"*

Source | DDIM Inv [27] | NTI [19] | EDICT [30] | Dir. Inv. [13] | **EtaInv (2)**

Fig. 7: Additional qualitative results for MasaCtrl editing.

*"a **collie dog** is sitting on a **bed**" → "a **garfield cat** is sitting on a **sofa**"*

*"a living room with a couch and a table" → "**a watercolor of** a living room with a couch and a table"*

*"a **black and white** drawing of a woman with long hair" → "a **colorful and detailed** drawing of a woman with long hair"*

*"a woman with **black** hair and red lipstick holding a flower" → "a woman with **silver** hair and red lipstick holding a flower"*

*"a **dry** tree in the wild" → "a **blooming** tree in the wild"*

*"a little girl wearing sunglasses and a gray **shirt** leaning against a wall" → "a little girl wearing sunglasses and a gray **dress** leaning against a wall"*

Source DDIM Inv [27] NTI [19] EDICT [30] Dir. Inv. [13] **EtaInv (3)**

Fig. 8: Additional qualitative results for EtaInv (3) PtP editing.

Fig. 9: Impact of η on image editing. We use a different linear η function for each generation by linearly interpolating η on the interval $[\eta(T), \eta(0)]$. Increasing η leads to better target prompt alignment while sacrificing background similarity. We disabled masking for demonstration purposes.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. Chen, H., Lee, H., Lu, J.: Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In: International Conference on Machine Learning. pp. 4735–4763. PMLR (2023)
6. Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., Zhang, A.: Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In: The Eleventh International Conference on Learning Representations (2023)

7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
8. Dong, W., Xue, S., Duan, X., Han, S.: Prompt tuning inversion for text-driven image editing using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7430–7440 (2023)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
10. 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)
11. 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)
12. 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)
13. 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)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. p. 1097–1105. NIPS'12, Curran Associates Inc., Red Hook, NY, USA (2012)
15. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation (2024), https://arxiv.org/abs/2312.14867
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. OpenAI: Gpt-4 technical report (2023)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)

23. 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)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
25. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2085–2094 (2021)
26. 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)
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
28. 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)
29. 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)
30. 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)
31. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
32. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)