ProxEdit: Improving Tuning-Free Real Image Editing with Proximal Guidance

Ligong Han¹ Song Wen¹ Qi Chen² Zhixing Zhang¹ Kunpeng Song¹ Mengwei Ren³ Ruijiang Gao⁴ Anastasis Stathopoulos⁴ Xiaoxiao He¹ Yuxiao Chen¹ Di Liu¹ Qilong Zhangli¹ Jindong Jiang¹ Zhaoyang Xia¹ Akash Srivastava⁵ Dimitris Metaxas¹ ¹Rutgers University ²Laval University ³New York University ⁴UT Austin ⁵MIT-IBM AI Lab

Abstract

DDIM inversion has revealed the remarkable potential of real image editing within diffusion-based methods. However, the accuracy of DDIM reconstruction degrades as larger classifier-free guidance (CFG) scales being used for enhanced editing. Null-text inversion (NTI) optimizes null embeddings to align the reconstruction and inversion trajectories with larger CFG scales, enabling real image editing with cross-attention control. Negative-prompt inversion (NPI) further offers a training-free closed-form solution of NTI. However, it may introduce artifacts and is still constrained by DDIM reconstruction quality. To overcome these limitations, we propose proximal guidance and incorporate it to NPI with cross-attention control. We enhance NPI with a regularization term and inversion guidance, which reduces artifacts while capitalizing on its training-free nature. Additionally, we extend the concepts to incorporate mutual self-attention control, enabling geometry and layout alterations in the editing process. Our method provides an efficient and straightforward approach, effectively addressing real image editing tasks with minimal computational overhead.

1. Introduction

Diffusion-based methods have emerged as popular approaches for real image editing, with many of these methods utilizing DDIM inversion (a deterministic inversion method proposed in Denoising Diffusion Implicit Models [52]). DDIM inversion is known to yield accurate reconstructions when using null embeddings or source prompts with a classifier-free guidance [25] (CFG) scale of 1. However, in order to achieve better editing capabilities, it is often necessary to use a CFG scale significantly larger than 1. Unfortunately, this scaling can lead to inaccurate reconstructions of the source image, which hinders the editing quality. This phenomenon is also observed in prompt-to-prompt [23] editing scenarios. To address this limitation, Null-text inversion [34] (NTI) was introduced as a solution. NTI employs pivotal inversion by optimizing the null embedding(s), ensuring that the reconstruction trajectory aligns with the inversion trajectory even under a larger CFG scale. While NTI has a lightweight parameter set, it requires per-image optimization, which can be time-consuming. To eliminate the need for optimization in NTI, Negative-prompt inversion [33] (NPI) offers a closed-form solution. By assuming equal predicted noises between
Building upon the remarkable results of NPI, we enhance it by incorporating a regularization term to improve the reconstruction of the source image. Moreover, we recognize that NPI is still constrained by the reconstruction quality of DDIM inversion, unable to correct errors introduced during the reconstruction process. To overcome this, we introduce a inversion guidance technique that performs one-step gradient descent on the current latent, aligning it with the inversion latents. The resulting algorithm offers a straightforward approach with negligible computational overhead.

Furthermore, as NTI and NPI are primarily designed for Cross-Attention Control [23], which focuses on texture and appearance changes, we extend our method to integrate proximal guidance into the Mutual Self-Attention Control framework [6]. This integration allows for geometry and layout alterations in real image editing tasks. In summary, our proposed method combines the benefits of NPI, inversion guidance, and a regularization term to provide an effective and efficient optimization-free solution for real image editing. We demonstrate its applications in NPI with Cross-Attention Control and Mutual Self-Attention Control, showcasing its versatility and potential usecases.

2. Related Work

Image generation with text guidance has been well-explored in image synthesis field [1, 2, 15–22, 38, 40, 44, 45, 53, 57, 63–65, 67, 73]. Recent development of text-to-image (T2I) diffusion models [7, 14, 24, 37, 50, 52, 54–56] introduced new solutions to this task. In particular, T2I diffusion models trained with large-scale image-caption pairs have shown impressive generation ability [36, 43, 46, 48]. The development of large-scale T2I models provides a giant and flexible design space for image manipulation methods leveraging the pre-trained model. Recent work proposes novel controlling mechanisms tailored for these T2I models [3, 8, 11, 27, 30, 61, 66, 68, 72].

Diffusion-based image editing. Many recent works fine-tune the pre-trained T2I models with a few personalized images to keep the context information [9, 29, 31, 39, 49, 51, 70]. Wide design choices have been explored in this direction. Textual-Inversion [10, 13, 60]-based methods propose fine-tuning the text embedding. Dreambooth [47] fine-tunes the whole model. [29] fine-tunes the cross-attention layers in the UNet of Stable-Diffusion model. These methods require hundreds of iterations at the fine-tuning stage to capture the identity information. For better efficiency, more techniques [17, 26, 32, 35] are developed by reducing the number of parameters optimized at fine-tuning stage. While fine-tuning the pre-trained T2I model shows extraordinary results, the test-time efficiency of these methods remains a great challenge. SEGA [5] discovers that target concept can be encoded using latent dimensions falling into the upper and lower tail of the distribution.

Inversion-based image editing. DDIM inversion [52] exhibits great potential in editing tasks by deterministically calculating and encoding the context information in a latent and reconstructing the original image with it. Applying editing prompt upon the inverted latent code greatly improved the test-time efficiency. Leveraging optimization on null-text embedding, Null-text Inversion [34] further improved the identity preservation of the edit. However, all these methods rely on optimization at test-time for accurate reconstruction, which typically requires several minutes. Negative-prompt inversion (NPI) [33] further reduces the computation cost for the inversion step while generates similarly competitive reconstruction results as Null-text inversion.
3. Method

3.1. Background

DDIM inversion. DDIM is a widely used deterministic sampling (if chosen to be) of DDPM. While DDPM follows a stochastic differential equation (SDE) process, DDIM corresponds to its ordinary differential equation (ODE) counterpart. The reverse DDIM process can be written as

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} z_t + \frac{1}{\sqrt{\alpha_t}} \left( \frac{1}{\alpha_{t-1}} - 1 - \frac{1}{\alpha_t} \right) \epsilon_\theta(z_t, t, C),$$  \hspace{1cm} (1)

where $C$ is the given conditioning. To invert the given image, the latent variables can be estimated by reversing the above discrete ODE sampling process. By rearranging Eq. (1), we have

$$z_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}} z_{t-1} + \frac{1}{\sqrt{\alpha_t}} \left( \frac{1}{\alpha_{t-1}} - 1 - \frac{1}{\alpha_t} \right) \epsilon_\theta(z_t, t, C).$$  \hspace{1cm} (2)

Note that $z_t$ appears at both sides. A common technique is to approximate $\epsilon_\theta(z_t, t, C)$ with $\epsilon_\theta(z_{t-1}, t-1, C)$, such that the inversion process can be solved by forward Euler method. Then, denote the sequence of latent variables from $z_0$ via DDIM inversion as $\{z_t^\star\}_{t=1}^T$, we have

$$z_t^\star = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}} z_{t-1}^\star + \frac{1}{\sqrt{\alpha_t}} \left( \frac{1}{\alpha_{t-1}} - 1 - \frac{1}{\alpha_t} \right) \epsilon_\theta(z_{t-1}^\star, t-1, C).$$  \hspace{1cm} (3)

Null-text inversion. Using the classifier-free guidance (CFG [25]), the noise is estimated by

$$\tilde{\epsilon}_\theta(z_t, t, C, \emptyset) = w \epsilon_\theta(z_t, t, C) + (1-w) \epsilon_\theta(z_t, t, \emptyset)$$  \hspace{1cm} (4)

If $w > 1$, the accumulated error on DDIM inversion will affect reconstruction accuracy. To address the problem, null-text inversion [34] (NTI) optimizes a set of per-timestep null-text embeddings $\{\theta_i\}$ to track the DDIM inversion trajectory even under a large $w$. It first computes $\{z_t^\star\}_{t=1}^T$ using DDIM inversion with $w = 1$. Then, after initializing $\tilde{z}_T = z_T^\star$, null-text inversion solves $\theta_t$ by performing the following optimizations for $t = T, \ldots, 1$:

$$\min_{\theta_t} ||z_{t-1}(\tilde{z}_t, \theta_t, C) - z_{t-1}^\star||_2^2.$$  \hspace{1cm} (5)

Negative-prompt inversion (NPI [33]) overcomes the limitation of per-image optimization in null-text inversion by providing a closed-form solution, $\theta_t = C$, with minimal approximation. NPI validates this solution through induction: if $\theta_t = C$ and $z_t = z_t^\star$ hold for timestep $t$, they also hold for timestep $t-1$, by assuming $\epsilon_\theta(z_t^\star, t, C) \approx \epsilon_\theta(z_{t-1}^\star, t-1, C)$. We can verify that with $\theta_t = C$, NPI reconstruction with $w > 1$ recovers the DDIM reconstruction,

$$\tilde{\epsilon}_\theta(z_t, t, C, \emptyset) = w \epsilon_\theta(z_t, t, C) + (1-w) \epsilon_\theta(z_t, t, \emptyset) \epsilon_\theta(z_t, t, C).$$  \hspace{1cm} (6)

In fact, as demonstrated in Fig. 2 and the subsequent remark, NPI provides an exact solution without any approximation when tracking the DDIM reconstruction trajectory instead of the inversion trajectory:

Remark 3.1. Negative-prompt inversion is the exact closed-form solution if we solve null-text inversion optimizations to track the DDIM reconstruction trajectory $\{\tilde{z}_t\}$.
Figure 5. **Qualitative comparisons of inversion methods.** The figure showcases qualitative comparisons among Null-text inversion (NTI) [34], Negative-prompt inversion (NPI) [33], and our proposed method (ProxNPI). Each row demonstrates the reconstruction results (columns 2-4) and editing results (columns 5-7) for the respective methods. Inversion guidance is employed to address minor errors in DDIM reconstruction. Errors or artifacts are marked using red circles or boxes. The comparisons highlight instances where NPI fails to retain specific image details (a), both NTI and NPI introduce undesired changes (b), the inversion guidance aids in recovering missing details (c), our method exhibits better background preservation (d), and NTI/NPI exhibit reconstruction errors (e).

### 3.2. Proximal Negative-Prompt Inversion

Negative-prompt inversion provides an elegant closed-form solution for computing null-text inverted null-embeddings, $∅_t = C$. This solution intuitively aligns with the DDIM reconstruction process. Fig. 3 illustrates a single inference step using classifier-free guidance (CFG) with a scale parameter $w = 2$. Initially, all methods employ DDIM inversion [52] with the source prompt (and $w = 1$). In Fig. 3(a), we depict a baseline approach where direct sampling is performed using the target prompt. Fig. 3(b) demonstrates the inference step of negative-prompt inversion, where CFG amplifies the editing direction of $(\tilde{ε}_{tar} - \tilde{ε}_{src})$. Intuitively, when the target prompt is close to the source prompt, the inference trajectory for editing should closely resemble the DDIM reconstruction trajectory. In fact, when the target condition $C' = C$, negative-prompt inversion exactly recovers DDIM reconstruction. However, we observe that negative-prompt inversion occasionally over-amplifies the editing direction $(\tilde{ε}_{tar} - \tilde{ε}_{src})$. To address this, we propose the addition of an extra loss term...
that encourages the CFG noise $\hat{\epsilon}$ to align with $\hat{\epsilon}_{\text{src}}$.

To accomplish this, we draw inspiration from the proximal gradient method [12, 58] and introduce a regularization term to constrain $(\hat{\epsilon}_{\text{tar}} - \hat{\epsilon}_{\text{src}})$. This regularization is achieved through the use of a proximal function,

$$\text{prox}_{\lambda, L_p}(x) = \text{argmin}_z \frac{1}{2} \| z - x \|^2 + \lambda \| z \|_p,$$  
(7)

which encourages desired properties in the editing process. When $p = 1$ (corresponding to $L1$ regularization), the solver takes the form of a soft-thresholding function,

$$[\text{prox}_{\lambda, L_1}(x)]_i = [S_\lambda(x)]_i = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda \\ 0 & \text{if } -\lambda \leq x_i \leq \lambda \\ x_i + \lambda & \text{if } x_i < -\lambda \end{cases}$$  
(8)

with $[\cdot]_i$ denoting the $i$-th component. If $p = 0$ (representing $L0$ regularization), the solver takes the form of a hard-thresholding function,

$$[\text{prox}_{\lambda, L_0}(x)]_i = \begin{cases} x_i & \text{if } |x_i| > \sqrt{2\lambda} \\ 0 & \text{otherwise} \end{cases}$$  
(9)

Since the value range of $(\hat{\epsilon}_{\text{tar}} - \hat{\epsilon}_{\text{src}})$ does not follow a standard Gaussian distribution, we employ a dynamic threshold rather than a fixed one by selecting a quantile of the absolute values $|\hat{\epsilon}_{\text{tar}} - \hat{\epsilon}_{\text{src}}|$. Fig. 3(c) provides a visualization of a 2-D case when soft-thresholding is employed. If all values are clamped to zero, our method reduces to DDIM reconstruction. Conversely, when all values are retained after thresholding, our method simplifies to negative-prompt inversion.

**Inversion guidance.** As discussed previously, NPI is still upper-bounded by the quality of DDIM reconstruction. Even if $\hat{\epsilon}$ converges to $\epsilon_{\text{src}}$, it cannot correct errors in cases where DDIM reconstruction is imperfect. On the other hand, NTI tracks the DDIM inversion trajectory and thus
Our proposed proximal guidance offers improvements in cases where MasaCtrl exhibits instability or introduces undesired changes that users wish to retain. MasaCtrl may introduce slight color shifts in the main subject(s) or background, as depicted in (a), (b), (c), and (f). However, with proximal guidance, the background and intended details are better preserved. For instance, in (a), the two steel bowls at the bottom and the person holding a phone at the right edge; in (b), the fence on the upper right and the dog’s mouth; in (c), the reins on the horse’s head; and in (d), the cup and vase in the background.

does not have such limitation. This motivates us to introduce a inversion guidance by performing a single step of gradient descent on the current latent \(\tilde{z}_{t-1}\), aiming to align it with the inversion latent \(z^*_{t-1}\). This gradient descent step is applied only to the “unedited” region identified by the mask \(M = |\tilde{e}_{\text{tar}} - \tilde{e}_{\text{src}}| \leq \lambda\), where we reuse the notation \(\lambda\) to represent the threshold value. By choosing a step size \(\eta\), the update can be expressed as \(\tilde{z}_{t-1} \leftarrow \tilde{z}_{t-1} - \eta M \odot (\tilde{z}_{t-1} - z^*_{t-1})\), where \(\eta = 1\) corresponds to a complete replacement. The complete algorithm is outlined in Algorithm 1. The algorithm can be thought of as an ADMM [4] type of method that solves NTI on the DDIM inversion trajectory:

\[
\min_{\theta_{e}} \| z_{t-1}(\tilde{z}_t, \theta_{e}, C) - \tilde{z}_{t-1}\|_2^2 \quad \text{s.t.} \quad \tilde{z}_{t-1} = z^*_{t-1} \tag{10}
\]

where the objective is solved by NPI (see Remark 3.1) and the constraint is enforced by inversion guidance.

### 3.3. Proximal Mutual Self-Attention Control

Both NTI and NPI are designed to be used with Cross-Attention Control (or Prompt-to-Prompt [23]) for real image editing. While Cross-Attention Control primarily focuses on changing the texture or appearance of a subject, recent methods have explored self-attention controlling mechanisms for manipulating geometry or layout [6, 41, 59, 71]. MasaCtrl [6] proposes a Mutual Self-Attention Control mechanism that queries image content from the source input image. In this section, we aim to integrate proximal guidance into the MasaCtrl framework.

Although NTI/NPI and MasaCtrl operate through different mechanisms, they share the same goal of preserving specific content from the source image. Initially, we observe that directly incorporating NPI with MasaCtrl by substituting the null embedding with the source prompt embedding can lead to artifacts. This occurs because, without any cross-attention or self-attention control, this is equivalent to setting the source prompt as negative prompt. As illustrated in an example in Fig. 4, using the source prompt as the negative prompt (“NPI w/o MasaCtrl”) generates a jumping person unrelated to the source image. When combined with MasaCtrl (“NPI w/ MasaCtrl”), the model is compelled to query cat features to render the same jumping person. Therefore, we propose using NPI solely in the reconstruction branch while retaining the null embedding in...
Figure 8. Qualitative Comparison of Prompt-to-Prompt [23] and MasaCtrl [6]. Column 2 presents the editing results obtained using NPI + Cross-Attention Control (“Prompt-to-Prompt”), while column 3 displays the editing results of MasaCtrl with proximal guidance (“ProxMasaCtrl”). In (b), Prompt-to-Prompt introduces a new texture on the cake that is absent in the input image, while (Prox-)MasaCtrl “recycles” the original texture.

Figure 9. Ablation study of thresholds. The figure shows visual results obtained by varying the threshold $\lambda$ from the 60% to 95% quantiles of the absolute noise difference. The first row represents the impact of hard-thresholding ($L_0$), while the second and third rows show the effects of soft-thresholding ($L_1$). Soft-thresholding induces less noticeable changes in the edited images compared to hard-thresholding, aligning with our expectations. Alternatively, increasing the CFG scale, such as with $w = 15$, can enhance the prominence of the target attribute, although it may introduce an intensified contrast ratio and shifted color tone.

Figure 10. Ablation study of inversion guidance. The figure shows visual results obtained by varying the stepsize of performing inversion guidance $\eta$ from the 0.1 to 1. The threshold is set to the 70% quantile and hard-thresholding is used.

4. Experiment

4.1. Cross-Attention Control

For quantitative evaluations, we measure the PSNR (peak signal-to-noise-ratio) and LPIPS [69] of the reconstructions of Negative-prompt inversion (NPI) [33] and our method (ProxNPI). As shown in Tab. 1, the proposed proximal guidance and inversion guidance improve both metrics by a large margin. For editing task, we report the CLIP [42] score (text alignment) and LPIPS [69] score (image alignment) in Tab. 2. It is observed that while ProxNPI marginally reduces CLIP similarity, it significantly improves the LPIPS score. Further details are provided in the supplementary material.

We further provide qualitative comparisons among Null-text inversion (NTI) [34], NPI, and ProxNPI in Fig. 5. Each row in the figure showcases the reconstruction results (columns 2-4) and editing results (columns 5-7) for each method. It is worth noting that NPI reconstruction is equivalent to DDIM reconstruction [52], as discussed previously. For (c-e) in Fig. 5, we utilize reconstruction guidance since DDIM reconstruction still introduces minor errors. These errors or artifacts are highlighted using red circles or boxes.

Additional visual results. We present more visual editing results in Fig. 6, along with the corresponding prompts provided underneath the images. Among the eight examples, inversion guidance is employed for examples (c), (e), (f),
4.2. Mutual Self-Attention Control

We conducted qualitative comparisons between MasaCtrl [6] and our proposed ProxMasaCtrl, as shown in Fig. 7. Our ProxMasaCtrl incorporates proximal guidance to address the issues of instability and undesired changes occasionally observed with MasaCtrl. As shown, MasaCtrl can introduce slight color shifts in the main subject(s) and background, as demonstrated in examples (a), (b), (c), and (f). However, with proximal guidance, we achieve better preservation of the background and intended details. For instance, in example (a), the two steel bowls at the bottom and the person holding a phone at the right edge are preserved. Similarly, in example (b), the fence on the upper right and the dog’s mouth are retained with improved fidelity. Additionally, in example (c), the reins on the horse’s head are better maintained, and in example (d), the cup and vase in the background are better preserved.

Comparing with Prompt-to-Prompt. As mentioned earlier, Prompt-to-Prompt [23] is designed to alter the texture of a subject, whereas Mutual Self-Attention Control [6] targets geometry and layout modifications. Fig. 8 shows a visual comparison between these two attention controlling mechanisms for geometry editing. This comparison aims to illustrate their respective behaviors rather than establish the superiority of one over the other. Interestingly, in Fig. 8 (b), Prompt-to-Prompt introduces a new texture on the cake, resembling its cross-section that deviates from the source image, while MasaCtrl preserves the original appearance. Additionally, Prompt-to-Prompt confines the rendered cake to a square shape within the cross-attention map of the original round cake, whereas MasaCtrl allows rendering outside the boundaries of the original cake. Note that the base of the cake is also changed to square.

4.3. Extensions

We further explore a few extensions of our method to the (a) DDPM inversion [28] framework, (b) personalized editing, as well as (c) simultaneous editing of texture and geometry. For (b), we employ an amortized encoder, ELITE [62], to enable training-free personalization. Due to space limit, results and details are presented in supplementary.

4.4. Ablations

Thresholding. In Fig. 9 we present visual results obtained by setting the threshold $\lambda$ to the 60%, 70%, ..., 95% quantiles of the absolute values of the noise difference. The first row illustrates the case of hard-thresholding (labeled as $L0$), while the second and third rows display the case of soft-thresholding (labeled as $L1$). As anticipated, we observe that soft-thresholding tends to introduce fewer changes to the edited images compared to hard-thresholding. Alternatively, we can increase the CFG scale, such as using $w = 15$, to enhance the prominence of the target attribute. However, this approach may lead to an amplified contrast ratio and shifted color tone. We empirically find using hard-thresholding with quantile 0.7 usually gives good results.

Inversion guidance. In Fig. 10 we present visual results obtained by varying the stepsize $\eta$ for the inversion guidance, applied when $t < T_{inv}$. As observed, when the guidance strength $\eta$ and $T_{inv}$ are small, the reconstruction of chopsticks is incomplete, and the pattern on the bowl is missing. We empirically find that setting $T_{inv} \geq 600$ and $\eta \geq 0.2$ generally yields satisfactory results.

5. Discussion and Conclusion

In this paper, we introduced proximal guidance, a versatile technique for enhancing diffusion-based tuning-free real image editing. We applied this technique to two concurrent frameworks: Negative-prompt inversion (NPI) and Mutual Self-Attention Control. The resulting algorithms, ProxNPI and ProxMasaCtrl, addressed limitations and achieved high-quality editing while maintaining computational efficiency. However, there are still considerations, as the performance of proximal guidance can be sensitive to hyperparameters. Exploring heuristics or automated methods for parameter selection could enhance the usability and generalizability of the proposed method. Our work demonstrates the potential of proximal guidance and opens avenues for further research in tuning-free real image editing.

Acknowledgements. This research has been partially funded by research grants to D. Metaxas through NSF: 2310966, 2235405, 2212301, 2003874, 1951890 and NIH 2R01HL127661.
References


