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

Ligong Han[1] Song Wen[1] Qi Chen[2] Zhixing Zhang[1] Kunpeng Song[1] Mengwei Ren[3]
Ruijiang Gao[4] Anastasis Stathopoulos[4] Xiaoxiao He[1] Yuxiao Chen[1] Di Liu[1]
Qilong Zhangli[1] Jindong Jiang[1] Zhaoyang Xia[1] Akash Srivastava[5] Dimitris Metaxas[1]
[1]Rutgers University [2]Laval University [3]New York University [4]UT Austin [5]MIT-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.

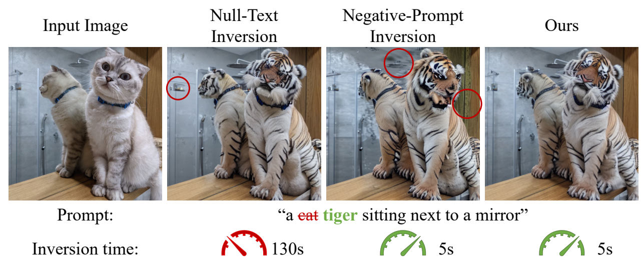

Figure 1. **Proximal Negative-Prompt Inversion.** A comparison of editing quality between Null-text inversion (NTI), Negative-prompt inversion (NPI), and our proposed method (ProxNPI). The bottom row represents the time required for inversion. Our approach incorporates the fast inversion capability of NPI without the need for test-time optimization, thereby incurring only minimal additional cost during inference.

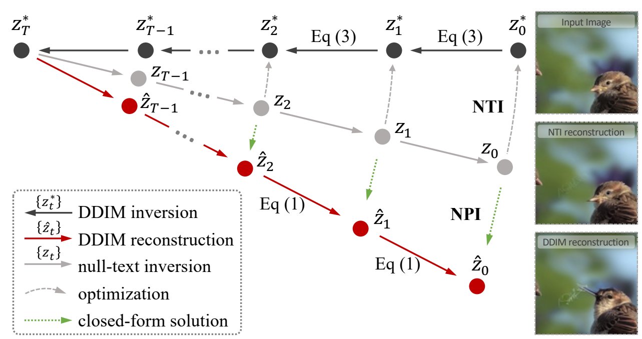

Figure 2. Negative-Prompt Inversion ("NPI") is the exact *closed-form* solution if we solve Null-text inversion ("NTI") on the DDIM reconstruction sequence $\{\hat{z}_t\}$.

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

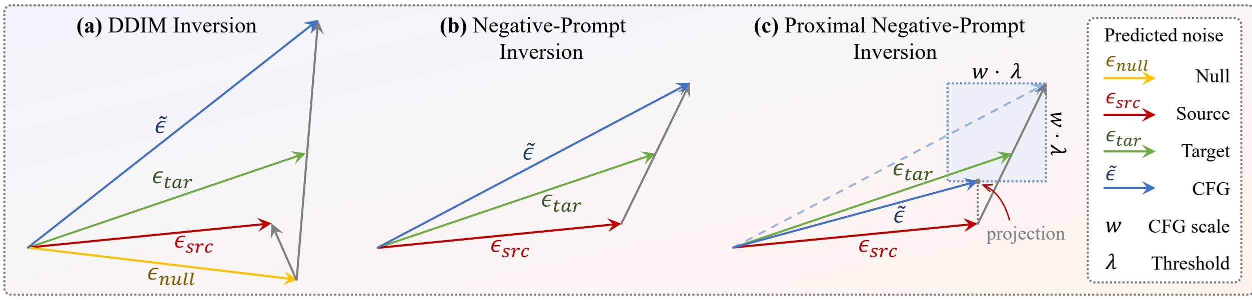

Figure 3. Illustration of a single inference step using classifier-free guidance (CFG) with a scale $w = 2$. All methods initially utilize DDIM inversion [52] with the source prompt (and $w = 1$). During the inference process: (a) direct sampling is performed using the target prompt; (b) the null embedding is replaced with the source prompt embedding; (c) a proximal gradient step is applied to the scaled noise difference $(\epsilon_{tar} - \epsilon_{src})$ following step (b). Here, we are visualizing soft-thresholding with a threshold λ, which corresponds to L1 regularization on $\tilde{\epsilon}$. If all values are clamped to zero, resulting in ProxNPI reducing to DDIM reconstruction. Conversely, when all values are retained after thresholding, ProxNPI reduces to NPI.

consecutive timesteps of the diffusion model, NPI elegantly demonstrated that the solver of NTI is equivalent to the source prompt embedding. However, NPI may occasionally introduce artifacts due to its underlying assumptions.

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 works propose 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.

Algorithm 1 Proximal Negative-Prompt Inversion

Input: Given source original sample z_0, source condition C, target condition C', denoising model ϵ_θ, proximal function $\text{prox}_\lambda(\cdot)$.

$\bar{z}_T = \text{DDIMInvert}(z_0, C, w = 1)$
$\tilde{z}_T = \bar{z}_T$
for $t = T$ to 1 **do**
  $\tilde{\epsilon}_{src} = \epsilon_\theta(\tilde{z}_t, t, C)$
  $\tilde{\epsilon}_{tar} = \epsilon_\theta(\tilde{z}_t, t, C')$
  $\tilde{\epsilon} = \tilde{\epsilon}_{src} + w \cdot \text{prox}_\lambda(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$
  $M = |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$
  $\tilde{z}_0 = \frac{1}{\sqrt{\alpha_t}}\tilde{z}_t - \sqrt{\frac{1}{\alpha_t} - 1}\tilde{\epsilon}$
  $\tilde{z}_{t-1} = \sqrt{\alpha_{t-1}}\tilde{z}_0 + \sqrt{1 - \alpha_{t-1}}\tilde{\epsilon}$
  if inversion guidance **and** $t < T_{inv}$ **then**
    $\tilde{z}_{t-1} = \tilde{z}_{t-1} - \eta M \odot (\tilde{z}_{t-1} - z^*_{t-1})$
  end if
end for
return $\tilde{z}_0$

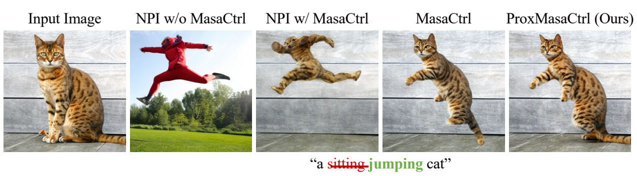

Figure 4. **Applying Negative Prompt Inversion (NPI) to Mutual Self-Attention Control (MasaCtrl [6]).** Directly applying NPI to MasaCtrl by substituting the null embedding with the source prompt embedding leads to the presence of strange artifacts (labeled as "NPI w/ MasaCtrl"). In our approach, we solely replace the null embedding with the source prompt in the DDIM reconstruction branch.

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

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

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

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

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 $\{z^*_t\}^T_{t=1}$, we have

$$z^*_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}} z^*_{t-1} + \sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right)\epsilon_\theta(z^*_{t-1}, t-1, C). \tag{3}$$

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

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

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 $\{\emptyset_t\}$ to track the DDIM inversion trajectory even under a large w. It first computes $\{z^*_t\}^T_{t=1}$ using DDIM inversion with $w = 1$. Then, after initializing $\bar{z}_T = z^*_T$, null-text inversion solves $\emptyset_t$ by performing the following optimizations for $t = T, \ldots, 1$:

$$\min_{\emptyset_t} ||z_{t-1}(\bar{z}_t, \emptyset_t, C) - z^*_{t-1}||^2_2. \tag{5}$$

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

$$\begin{aligned}\tilde{\epsilon}_\theta(z_t, t, C, C) &= w\epsilon_\theta(z_t, t, C) + (1 - w)\epsilon_\theta(z_t, t, C) \\ &= \epsilon_\theta(z_t, t, C).\end{aligned} \tag{6}$$

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* $\{\hat{z}_t\}$.

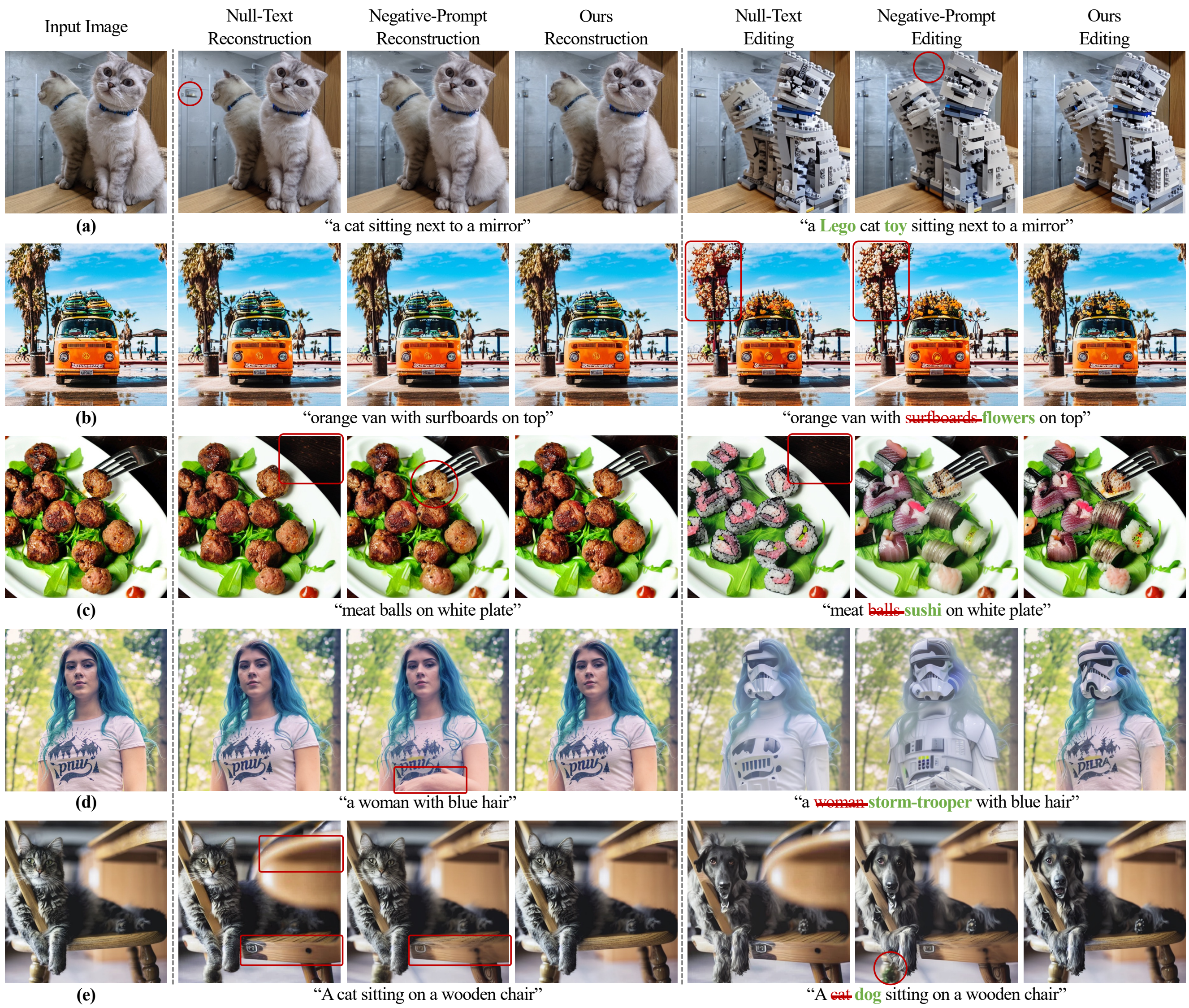

Figure 5. **Qualitative comparisons of inversion methods.** The figure showcases qualitative comparisons among Null-text inversion (NTI) [34], Negative-prompt inversion (NPI) [33], and our proposed method (ProxNPI). Each row demonstrates the reconstruction results (columns 2-4) and editing results (columns 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, $\emptyset_t = C$. 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 $(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$. 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 $(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$. To address this, we propose the addition of an extra loss term

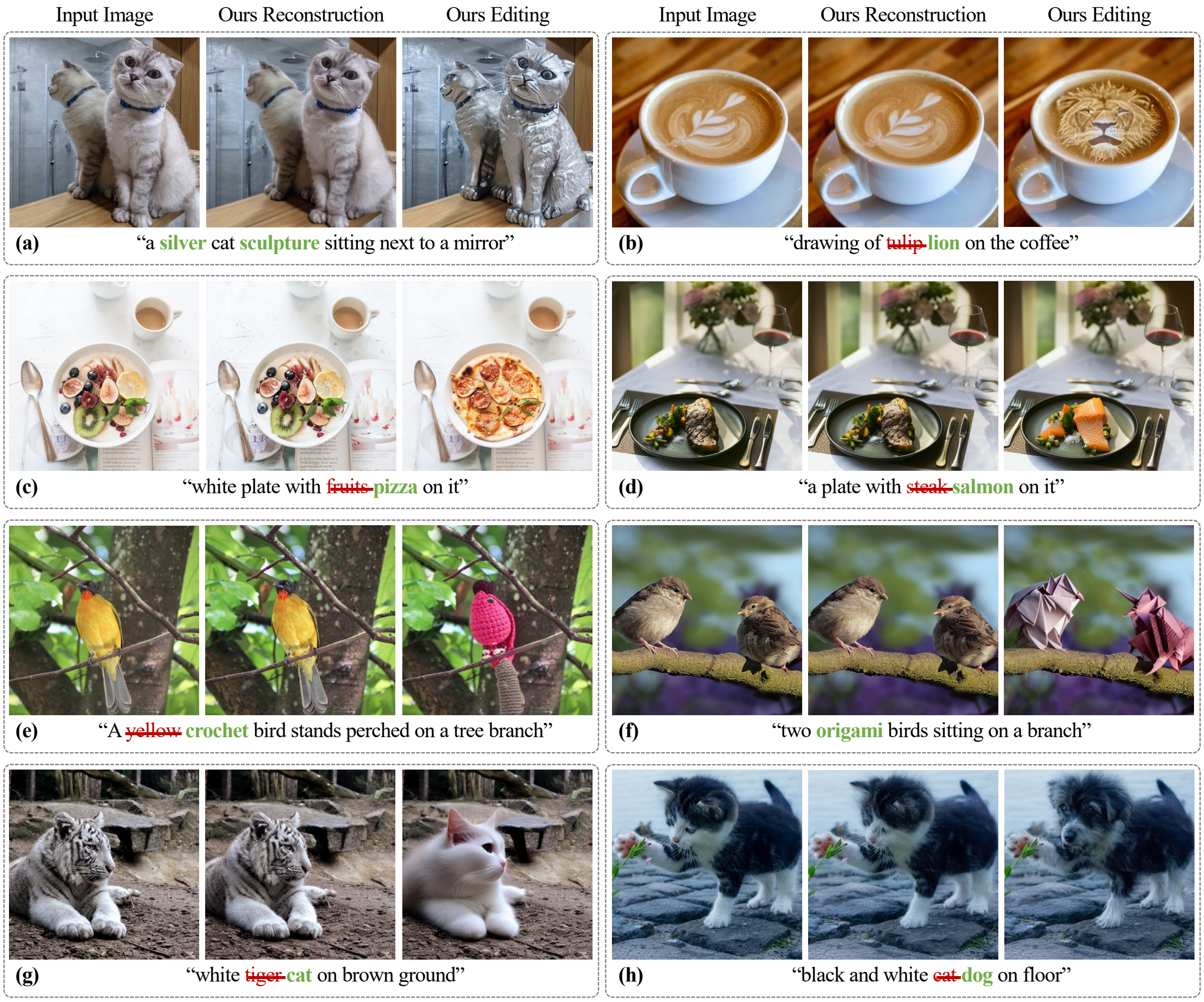

Figure 6. **Additional visual editing results.** More visual editing results for our method are presented, along with their corresponding prompts. Inversion guidance is applied for examples (c), (e), (f), and (g) due to imperfect DDIM reconstructions in these cases.

that encourages the CFG noise $\tilde{\epsilon}$ to align with $\tilde{\epsilon}_{src}$.

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

$$\text{prox}_{\lambda, L_p}(x) = \underset{z}{\text{argmin}} \frac{1}{2}\|z - x\|_2^2 + \lambda\|z\|_p. \tag{7}$$

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,

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

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

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

Since the value range of $(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$ 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 $|\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}|$. 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 $\tilde{\epsilon}$ converges to ϵ_{src}, it cannot correct errors in cases where DDIM reconstruction is imperfect. On the other hand, NTI tracks the DDIM inversion trajectory and thus

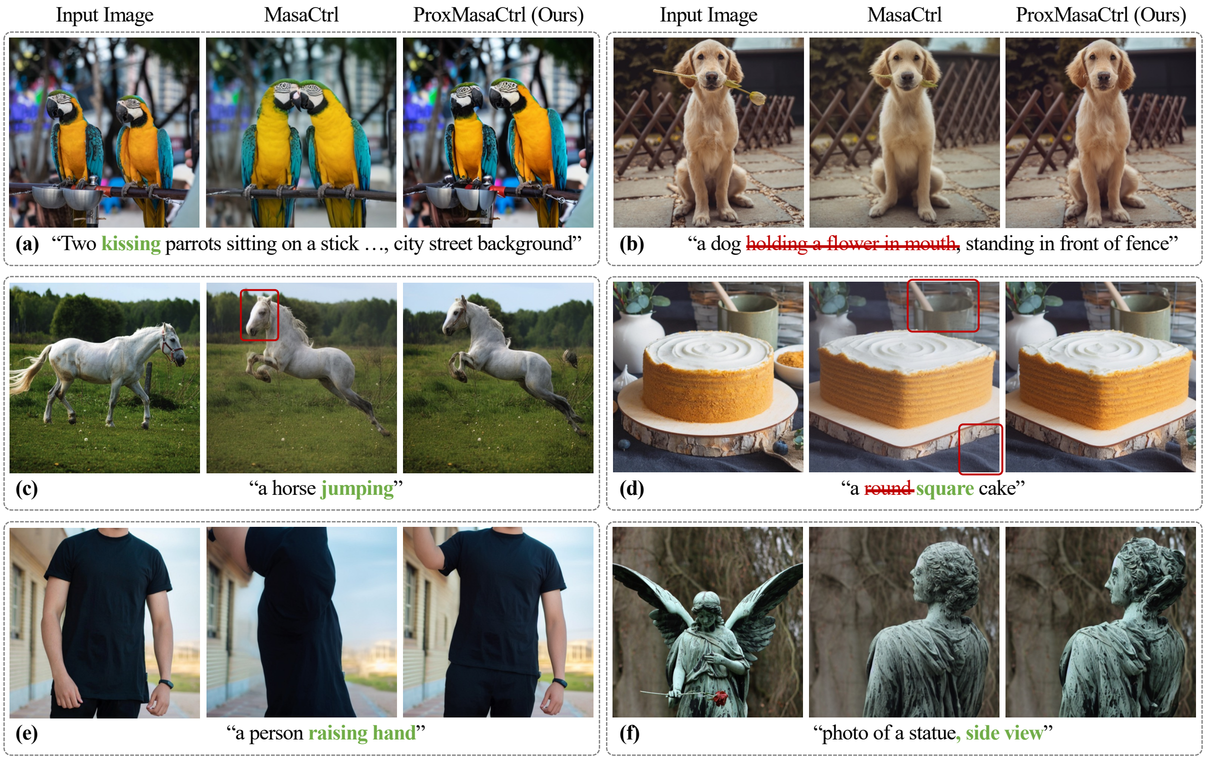

Figure 7. **Enhancing MasaCtrl [6] with Proximal Guidance.** Our proposed proximal guidance offers improvements in cases where MasaCtrl exhibits instability or introduces undesired changes that users wish to retain. MasaCtrl may introduce slight color shifts in the main subject(s) or background, as depicted 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 $M = |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$, where we reuse the notation λ to represent the *threshold* value. By choosing a step size η, 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:

$$\min_{\emptyset_t} ||z_{t-1}(\bar{z}_t, \emptyset_t, C) - \hat{z}_{t-1}||_2^2 \quad \text{s.t.} \quad \hat{z}_{t-1} = z^*_{t-1} \quad (10)$$

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.

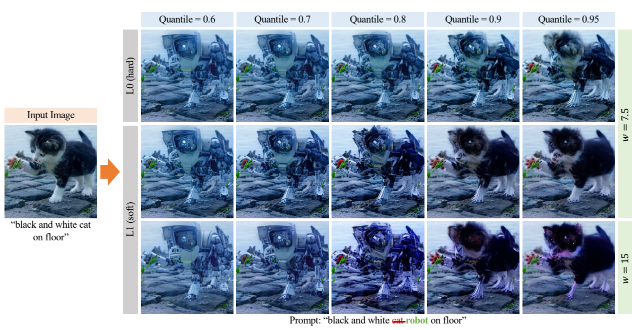

Figure 9. **Ablation study of thresholds.** The figure shows visual results obtained by varying the threshold λ from the 60% to 95% quantiles of the absolute noise difference. The first row represents the impact of hard-thresholding ($L0$), while the second and third rows show the effects of soft-thresholding ($L1$). 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.

the synthesis branch:

$$\begin{aligned} \hat{\epsilon} &= \hat{\epsilon}_{src} + 1 \cdot (\hat{\epsilon}_{src} - \hat{\epsilon}_{src}), && \text{[reconstruction]} \\ \tilde{\epsilon} &= \tilde{\epsilon}_{null} + w \cdot \operatorname{prox}_{\lambda}(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{null}). && \text{[synthesis]} \end{aligned} \tag{11}$$

Here, we introduce proximal guidance to the term $(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{null})$. During model forward passes, the MasaCtrl mechanism forces both $\tilde{\epsilon}_{null}$ and $\tilde{\epsilon}_{tar}$ to query features from $\hat{\epsilon}_{src}$ (in the original MasaCtrl, the unconditional part $\tilde{\epsilon}_{null}$ queries from $\hat{\epsilon}_{null}$ in the reconstruction branch). By setting λ to the 100% quantile, $\tilde{\epsilon}$ converges to $\tilde{\epsilon}_{null} \approx \hat{\epsilon}_{src}$, degrading to a DDIM reconstruction. Hence, similar to Prox-NPI, the introduced proximal guidance here also controls the proximity of the synthesized image to the source image.

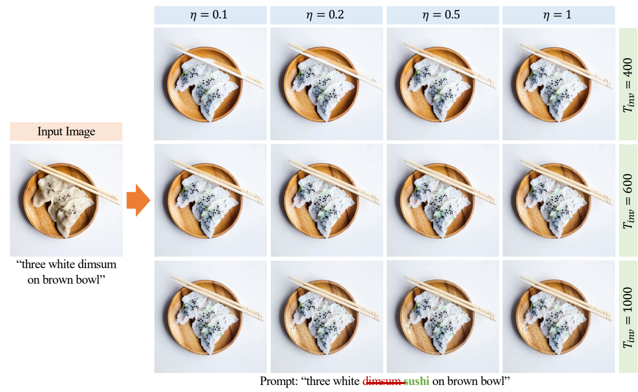

Figure 10. **Ablation study of inversion guidance.** The figure shows visual results obtained by varying the stepsize of performing inversion guidance η from the 0.1 to 1. The threshold is set to the 70% quantile and hard-thresholding is used.

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 Prox-NPI 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), and (g) due to imperfections in DDIM reconstructions for these specific cases.

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 λ 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

	Original		VAE	
Method	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
NPI	25.214 ± 4.308	0.134 ± 0.083	28.802 ± 5.640	0.095 ± 0.092
ProxNPI	$\mathbf{28.297 \pm 4.139}$	$\mathbf{0.057 \pm 0.029}$	$\mathbf{70.984 \pm 2.021}$	$\mathbf{0.000 \pm 0.000}$

Table 1. **Reconstruction.** For "original", metrics are measured between reconstructed and the original image; for "VAE", metrics are measured between reconstructed image and the VAE reconstruction.

	Cat → X		Dog → Y	
Method	CLIP ↑	LPIPS ↓	CLIP ↑	LPIPS ↓
NPI	$\mathbf{27.371 \pm 2.181}$	0.268 ± 0.077	$\mathbf{28.229 \pm 3.195}$	0.288 ± 0.106
ProxNPI	27.097 ± 2.507	$\mathbf{0.205 \pm 0.045}$	27.813 ± 3.414	$\mathbf{0.217 \pm 0.063}$

Table 2. **Editing.** CLIP [42] score measures similarities between edited images and target text prompts. LPIPS [69] score measures the structural similarity between edited and original images. $\mathtt{X} =$ {"dog", "hamster", "fox", "badger", "lion", "bear", "pig"}, and $\mathtt{Y} =$ {"cat", "hamster", "fox", "badger", "lion", "bear", "pig"}.

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 η for the inversion guidance, applied when $t < T_{inv}$. As observed, when the guidance strength η 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

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 6

[5] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*, 2023. 2

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 2, 3, 6, 7, 8

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[8] Hila Chefer, Oran Lang, Mor Geva, Volodymyr Polosukhin, Assaf Shocher, Michal Irani, Inbar Mosseri, and Lior Wolf. The hidden language of diffusion models. *arXiv preprint arXiv:2306.00966*, 2023. 2

[9] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation, 2023. 2

[10] Giannis Daras and Alexandros G Dimakis. Multiresolution textual inversion. *arXiv preprint arXiv:2211.17115*, 2022. 2

[11] Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. *arXiv preprint arXiv:2305.08252*, 2023. 2

[12] Jeffrey A. Fessler. Lecture notes on proximal methods. https://web.eecs.umich.edu/~fessler/course/598/l/n-05-prox.pdf, 2021. 5

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[15] Jun Han, Martin Renqiang Min, Ligong Han, Li Erran Li, and Xuan Zhang. Disentangled recurrent wasserstein autoencoder. In *International Conference on Learning Representations*, 2020. 2

[16] Ligong Han, Ruijiang Gao, Mun Kim, Xin Tao, Bo Liu, and Dimitris Metaxas. Robust conditional gan from uncertainty-aware pairwise comparisons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10909–10916, 2020. 2

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[18] Ligong Han, Martin Renqiang Min, Anastasis Stathopoulos, Yu Tian, Ruijiang Gao, Asim Kadav, and Dimitris N Metaxas. Dual projection generative adversarial networks for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14438–14447, 2021. 2

[19] Ligong Han, Sri Harsha Musunuri, Martin Renqiang Min, Ruijiang Gao, Yu Tian, and Dimitris Metaxas. Ae-stylegan: Improved training of style-based auto-encoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3134–3143, 2022. 2

[20] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 2

[21] Ligong Han, Anastasis Stathopoulos, Tao Xue, and Dimitris Metaxas. Unbiased auxiliary classifier gans with mine. *arXiv preprint arXiv:2006.07567*, 2020. 2

[22] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 2

[23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 6, 7, 8

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 3

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[27] Yihao Huang, Qing Guo, and Felix Juefei-Xu. Zero-day backdoor attack against text-to-image diffusion models via personalization. *arXiv preprint arXiv:2305.10701*, 2023. 2

[28] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023. 8

[29] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[30] Pengzhi Li, QInxuan Huang, Yikang Ding, and Zhiheng Li. Layerdiffusion: Layered controlled image editing with diffusion models. *arXiv preprint arXiv:2305.18676*, 2023. 2

[31] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 2

[32] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 2

[33] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 1, 2, 3, 4, 7

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1, 2, 3, 4, 7

[35] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[38] Byong Mok Oh, Max Chen, Julie Dorsey, and Frédo Durand. Image-based modeling and photo editing. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 433–442, 2001. 2

[39] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023. 2

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 2

[41] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7, 8

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 2

[45] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Image deblurring with domain generalizable diffusion models. *arXiv preprint arXiv:2212.01789*, 2022. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[49] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[51] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 7

[53] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2

[54] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[57] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2

[58] Ryan Tibshirani. Lecture notes on proximal gradient descent. https://www.stat.cmu.edu/~ryantibs/convexopt/lectures/prox-grad.pdf, 2019. 5

[59] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 6

[60] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[61] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023. 2

[62] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 8

[63] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[64] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 2

[65] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[66] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 2

[67] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[68] Jiaxin Zhang, Kamalika Das, and Sricharan Kumar. On the robustness of diffusion inversion in image manipulation. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 2

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 8

[70] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *arXiv preprint arXiv:2305.16225*, 2023. 2

[71] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *arXiv preprint arXiv:2305.18729*, 2023. 6

[72] Zhongping Zhang, Jian Zheng, Jacob Zhiyuan Fang, and Bryan A Plummer. Text-to-image editing by image information removal. *arXiv preprint arXiv:2305.17489*, 2023. 2

[73] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 2