ProxEdit: Improving Tuning-Free Real Image Editing with Proximal Guidance (Supplementary Materials)

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

A. Proof of Remark 3.1

Remark A.1. *Negative-prompt inversion is the* exact *closed-form solution if we solve null-text inversion optimizations to track the DDIM reconstruction trajectory* $\{\hat{z}_t\}$*, with* $\bar{z}_T$ *initialized as* $\hat{z}_T = z_T^*$*,*

$$C = \underset{\emptyset_t}{\operatorname{argmin}} ||z_{t-1}(\bar{z}_t, \emptyset_t, C) - \hat{z}_{t-1}||_2^2. \tag{12}$$

Proof. Following negative-prompt inversion [4], we prove this by induction. Suppose at timestep t, $\emptyset_t = C$ and $\bar{z}_t = \hat{z}_t$ hold, then we derive $\bar{z}_{t-1}$ for timestep $t-1$. By definition (Eq. (1) with classifier-free guidance),

$$\bar{z}_{t-1} = z_{t-1}(\bar{z}_t, t, C, \emptyset_t) = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}}\bar{z}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right)\tilde{\epsilon}_\theta(\bar{z}_t, t, C, \emptyset_t). \tag{13}$$

Since $\bar{z}_t = \hat{z}_t$ and by Eq. (2),

$$\bar{z}_t = \hat{z}_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}}\hat{z}_{t-1} + \sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right)\epsilon_\theta(\hat{z}_t, t, C). \tag{14}$$

Substituting the above into Eq. (13), we have

$$\bar{z}_{t-1} = \hat{z}_{t-1} + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \left(\tilde{\epsilon}_\theta(\hat{z}_t, t, C, \emptyset_t) - \epsilon_\theta(\hat{z}_t, t, C)\right). \tag{15}$$

Since

$$\begin{aligned}&\tilde{\epsilon}_\theta(\hat{z}_t, t, C, \emptyset_t) - \epsilon_\theta(\hat{z}_t, t, C)\\ =&(w-1)\left(\epsilon_\theta(\hat{z}_t, t, C) - \epsilon_\theta(\hat{z}_t, t, \emptyset_t)\right),\end{aligned}$$

we have $\bar{z}_{t-1} = \hat{z}_{t-1}$ if $\emptyset_t = C$. □

Figure 11. **Proximal guidance with DDPM inversion [3].**

B. Extension to DDPM Inversion

The concept of proximal guidance is also applicable to the DDPM inversion framework [3]. By performing an inversion of the input image with null text and setting λ to 100%, the exact DDPM reconstruction can be restored as follows:

$$\begin{aligned}&\hat{\epsilon} = \hat{\epsilon}_{null}, && \text{[inversion]}\\ &\tilde{\epsilon} = \tilde{\epsilon}_{null} + w \cdot \text{prox}_\lambda(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{null}). && \text{[synthesis]}\end{aligned} \tag{16}$$

This approach introduces an additional control parameter, allowing the edited image to more closely resemble the input image. An illustrative example is provided in Fig. 11.

C. Extension to Personalized Editing

We explore an extension of proximal guidance in personalized image editing. Here the target concept is specified by another reference image. This is realized by integrating an amortized encoder, as demonstrated using ELITE [5] in our experiments. Visual examples of this extension are presented in Fig. 12.

D. Simultaneous Texture and Geometry Editing

We extend our approach by sequentially applying ProxNPI (Cross-Attention Control) and ProxMasaCtrl (Mutual Self-Attention Control) to enable simultaneous editing of

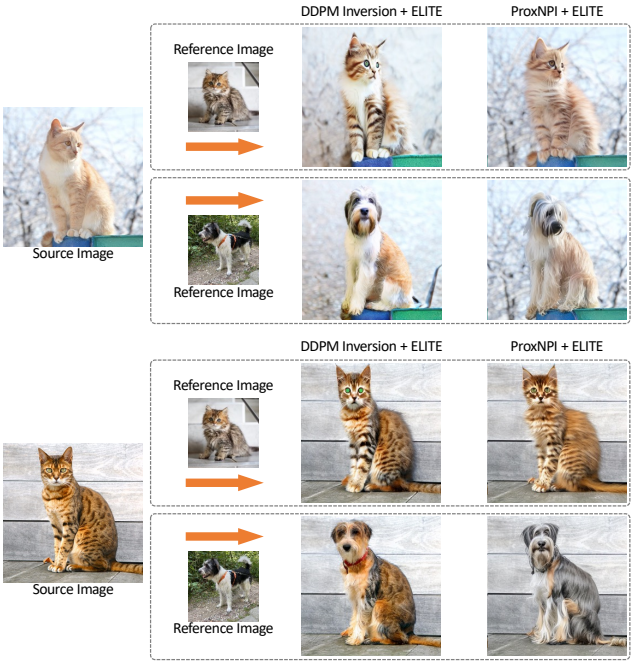

Figure 12. **Personalized image editing** with the ELITE [5] encoder in DDPM inversion [3] and prompt-to-prompt [2] frameworks.

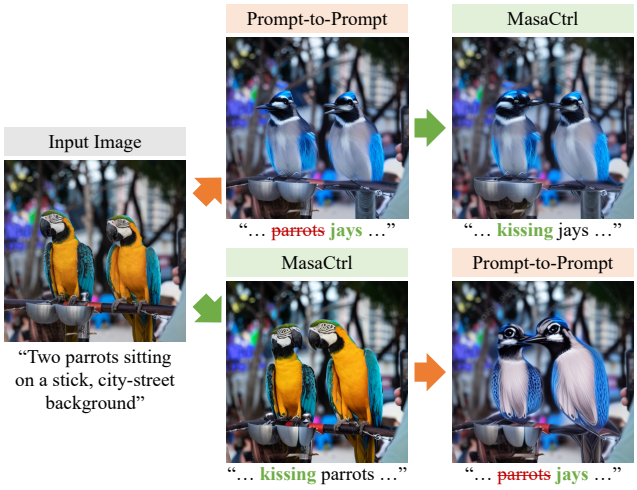

Figure 13. **Editing texture and pose.** Sequentially applying Prompt-to-Prompt and MasaCtrl (first row) or in the reverse order (second row) to edit both texture and pose.

both texture and geometry, as shown in Fig. 13. While this represents a preliminary exploration, and the integration of these two controlling mechanisms in a more efficient and effective manner is an avenue for future research.

E. Ablation of MasaCtrl feature injection strategies

In Fig. 14, we conduct ablations of different mutual self-attention feature injection strategies in MasaCtrl [1]. The synthesis branch utilizes the "null embedding" denoted as $C = \text{interp}(\alpha, C_{src}, C_{null})$, where the default setting uses the original null embedding with $\alpha = 1$. We explore the visual effects by varying α and querying different feature sets, including "source" (the default strategy), "joint" (querying from both branches), and "none" (no feature injection). While $\alpha = 1$ with "source" generally produces good results for both unconditional and conditional noises, we observe that using "joint" or "none" for the unconditional noise occasionally improves the outcomes.

F. Reconstruction Guidance

Algorithm 2 Proximal Negative-Prompt Inversion with reconstruction guidance

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

We have introduced the concept of "reconstruction guidance" as an additional solution to address the issue of imperfect DDIM reconstruction. Another straight-forward solution is *reconstruction guidance*. To do so, we perform one step of gradient descent on the current predicted original sample $\tilde{z}_0$ to align it with the source sample z_0. Similarly, this gradient descent step is applied to the "unedited" region identified by the mask $M = |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$. The update can be expressed as $\tilde{z}_0 \leftarrow \tilde{z}_0 - \eta M \odot (\tilde{z}_0 - z_0)$. The algorithm with reconstruction guidance is outlined in Algorithm 2. In Fig. 15, we present visual results obtained by varying the stepsize η. The guidance is applied when $t < T_{rec}$. As observed, when the guidance strength is small (with a small η), the reconstruction of chopsticks is incom-

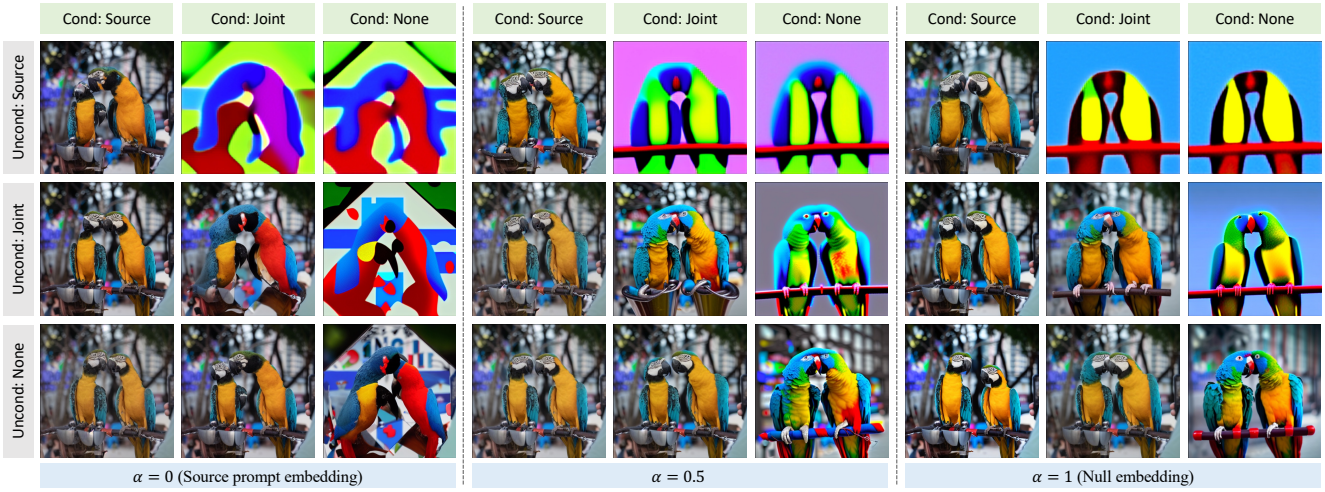

Figure 14. **MasaCtrl feature injection ablation.** We ablate different feature injection strategies in MasaCtrl by varying the α parameter (in the synthesis branch) and querying different feature sets ("source", "joint", "none"). For all α we use C_{src} in reconstruction branch. In the above example, we find that using "joint" or "none" for the unconditional noise improves results.

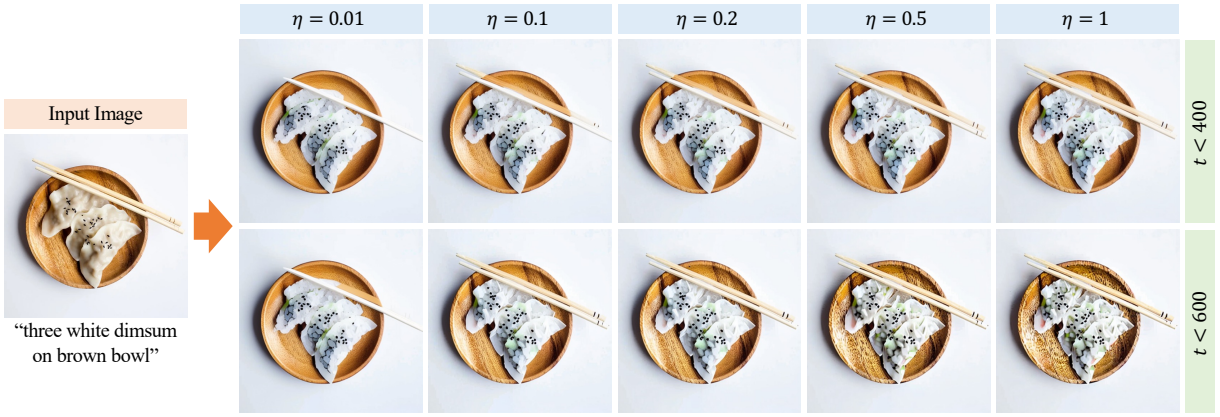

Figure 15. **Ablation study of reconstruction guidance.** The figure shows visual results obtained by varying the stepsize of performing reconstruction guidance η from the 0.01 to 1. The first row represents performing guidance when $t < 400$, while the second shows the effects of $t < 600$. The threshold is set to the 70% quantile and hard-thresholding is used.

plete. Increasing T_{rec} results in accurate reconstruction of the chopsticks, however, a large η may introduce artifacts such as an over-amplified contrast ratio. Based on empirical findings, we generally set $T_{rec} = 400$ and $\eta = 0.1$, although inversion guidance is still preferred over reconstruction guidance.

References

[1] 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

[2] 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. 2

[3] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023. 1, 2

[4] 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

[5] 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. 1, 2