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[1], Xiaoxiao He[1], Yuxiao Chen[1], Di Liu[1], Qilong Zhangli[1], Jindong Jiang[1], Zhaoyang Xia[1], Akash Srivastava[5], Dimitris Metaxas[1]

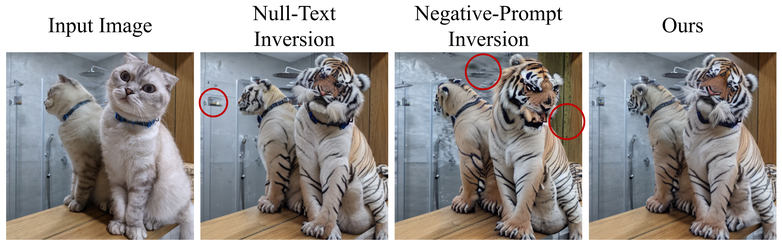
[1]Rutgers University [2]Laval University [3]New York University [4]UT Austin [5]MIT-IBM AI Lab

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

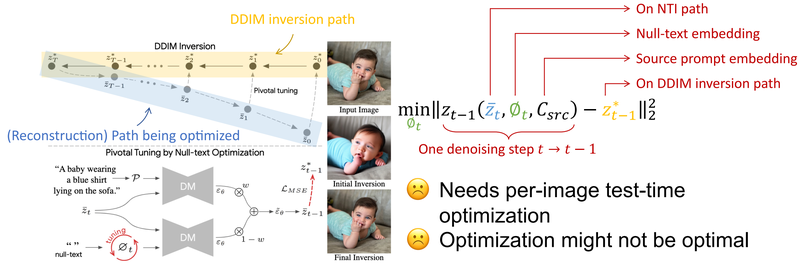
JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Introduction & Motivation

TL;DR: ≈ Tuning-free **Null-Text Inversion**



Recap of Null-Text Inversion:



☹ Needs per-image test-time optimization

☹ Optimization might not be optimal

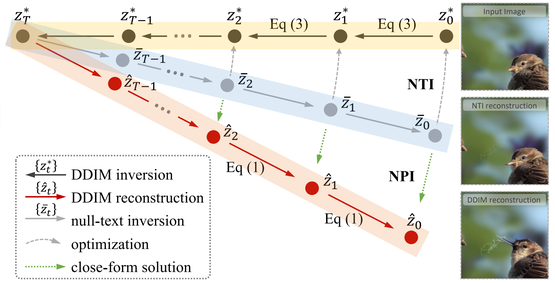
Our reformulation (*splitting variable*):

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

*Negative-Prompt Inversion (NPI)

Closed-form solution: $\emptyset_t = C_{src}$

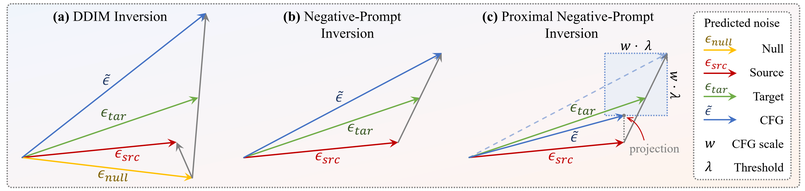
One-step gradient descent no grad, no UNet forward, very efficient



Code

Method

Illustration of a single inference step using classifier-free guidance (CFG) with a scale $w = 2$



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 $\tilde{z}_0$

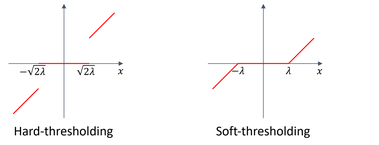
$\min_{\emptyset_t} \|z_{t-1}(\bar{z}_t, \emptyset_t, C_{src}) - \hat{z}_{t-1}\|_2^2$ *s.t.* $\hat{z}_{t-1} = z^*_{t-1}$

Inversion guidance to enforce constraint

Proximal Guidance

- Negative-Prompt Inversion:
  $\tilde{\epsilon} = \epsilon_{src} + w \cdot (\epsilon_{tar} - \epsilon_{src})$
- Ours:
  $\tilde{\epsilon} = \epsilon_{src} + w \cdot \textbf{prox}_\lambda(\epsilon_{tar} - \epsilon_{src})$
- Proximal operator:
  $\textbf{prox}_{\lambda, L_p}(x) = \underset{z}{\text{argmin}} \frac{1}{2}\|z - x\|_2^2 + \lambda\|z\|_p$

$\tilde{\epsilon} = \epsilon_{src} + 0$



	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	**28.297 ± 4.139**	**0.057 ± 0.029**	**70.984 ± 2.021**	**0.000 ± 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	**27.371 ± 2.181**	0.268 ± 0.077	**28.229 ± 3.195**	0.288 ± 0.106
ProxNPI	27.097 ± 2.507	**0.205 ± 0.045**	27.813 ± 3.414	**0.217 ± 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. X = {"dog", "hamster", "fox", "badger", "lion", "bear", "pig"}, and Y = {"cat", "hamster", "fox", "badger", "lion", "bear", "pig"}.

Experiments