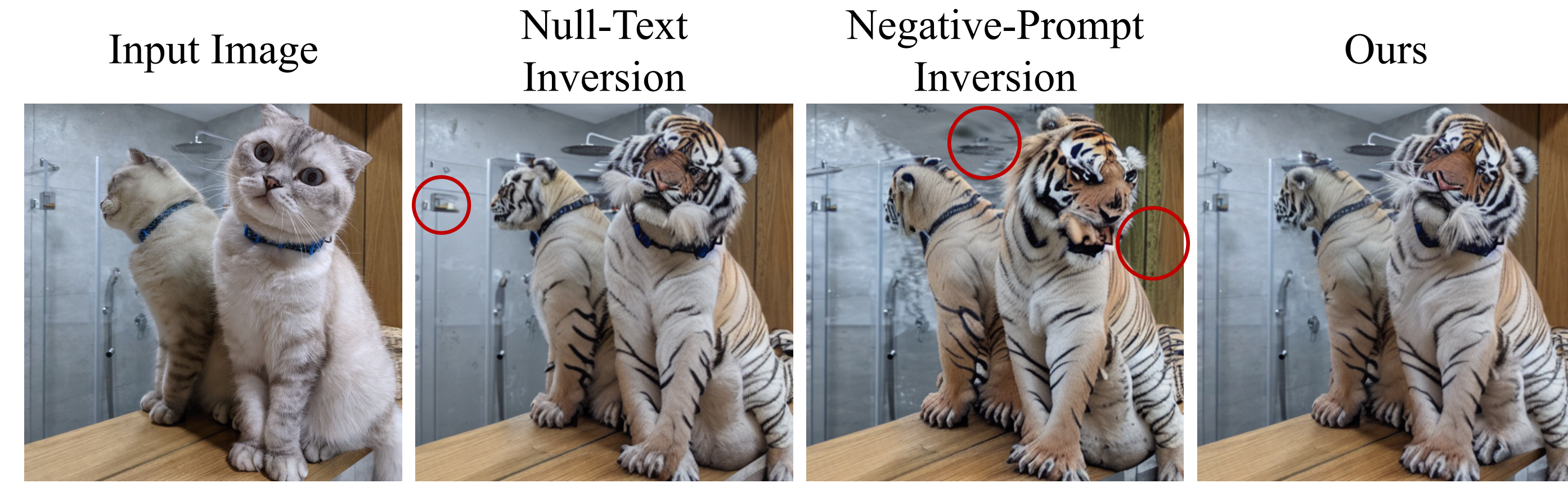


Introduction & Motivation

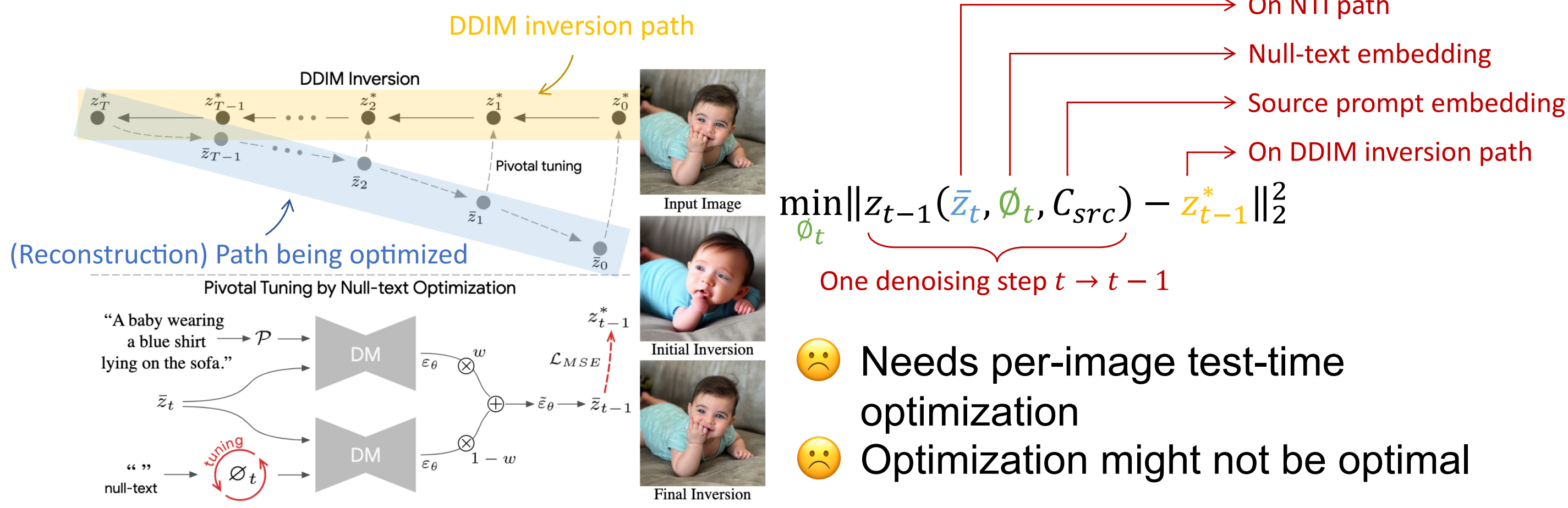
TL;DR: $\approx$ Tuning-free **Null-Text Inversion**



Prompt: "a ~~eat~~ **tiger** sitting next to a mirror"

Inversion time: 130s 5s 5s

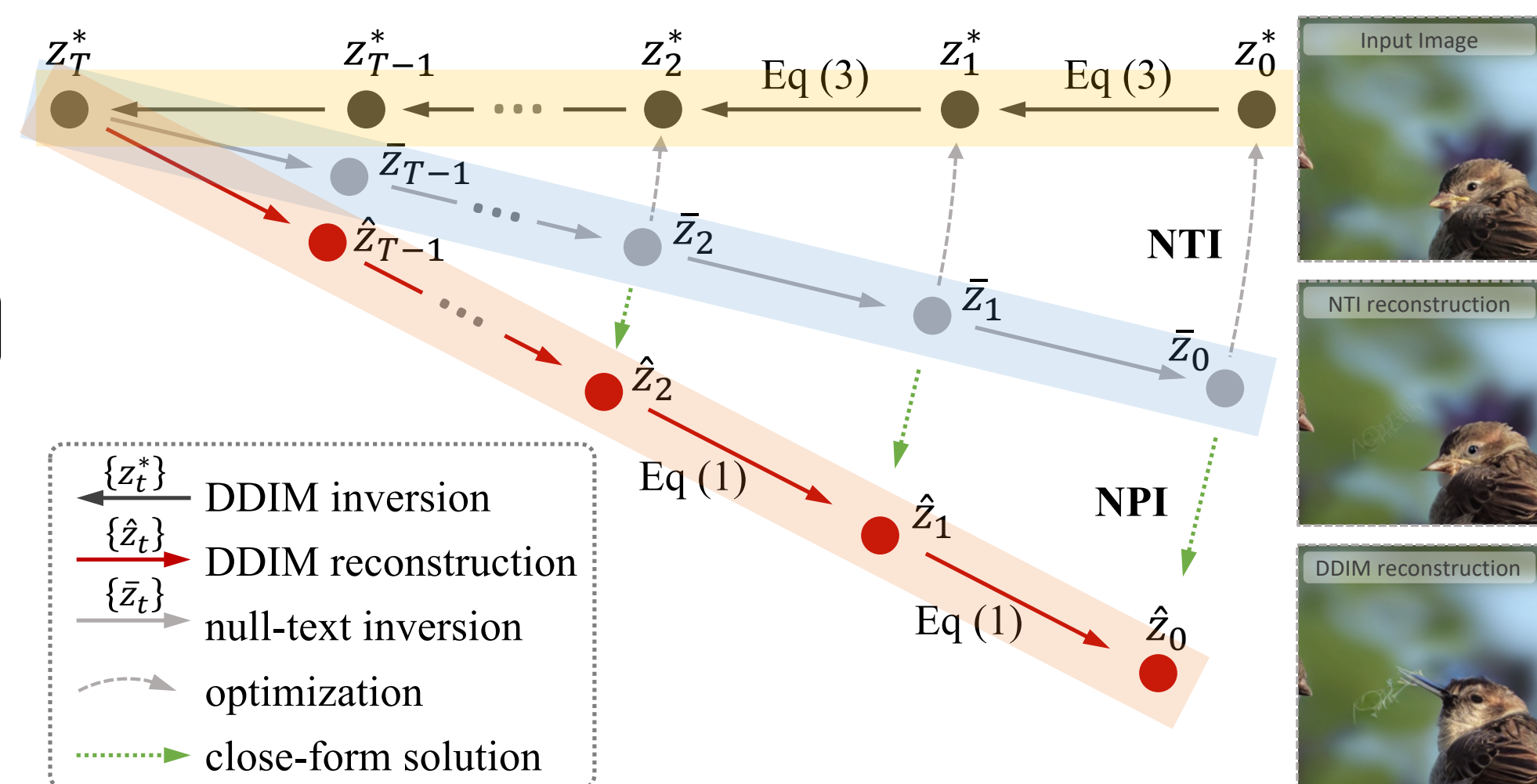
Recap of Null-Text Inversion:



Our reformulation (*splitting variable*):

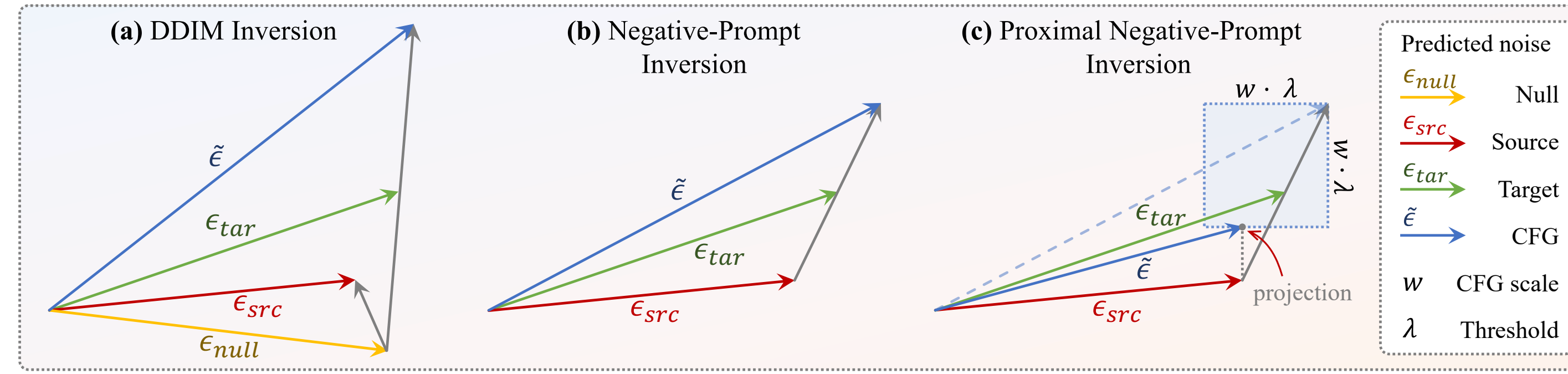
$$\min_{\hat{\phi}_t} \|z_{t-1}(\bar{z}_t, \hat{\phi}_t, C_{src}) - \hat{z}_{t-1}\|_2^2 \quad s.t. \quad \hat{z}_{t-1} = z_{t-1}^*$$

*Negative-Prompt Inversion (NPI) Closed-form solution: $\hat{\phi}_t = C_{src}$ One-step gradient descent no grad, no UNet forward, very efficient



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 $\text{prox}_\lambda(\cdot)$.

1: $\bar{z}_T = \text{DDIMInvert}(z_0, C, w = 1)$

2: $\tilde{z}_T = \bar{z}_T$

3: **for** $t = T$ to 1 **do**

4: $\tilde{\epsilon}_{src} = \epsilon_\theta(\tilde{z}_t, t, C)$

5: $\tilde{\epsilon}_{tar} = \epsilon_\theta(\tilde{z}_t, t, C')$

6: $\tilde{\epsilon} = \tilde{\epsilon}_{src} + w \cdot \text{prox}_\lambda(\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src})$

7: $M = |\tilde{\epsilon}_{tar} - \tilde{\epsilon}_{src}| \leq \lambda$

8: $\tilde{z}_0 = \frac{1}{\sqrt{\alpha_t}} \tilde{z}_t - \sqrt{\frac{1}{\alpha_t} - 1} \tilde{\epsilon}$

9: $\tilde{z}_{t-1} = \sqrt{\alpha_{t-1}} \tilde{z}_0 + \sqrt{1 - \alpha_{t-1}} \tilde{\epsilon}$

10: **if** inversion guidance and $t < T_{inv}$ **then**

11: $\tilde{z}_{t-1} = \tilde{z}_{t-1} - \eta M \odot (\tilde{z}_{t-1} - z_{t-1}^*)$

12: **end if**

13: **end for**

14: **return** $\tilde{z}_0$

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

Inversion guidance to enforce constraint

Proximal Guidance

• Negative-Prompt Inversion:

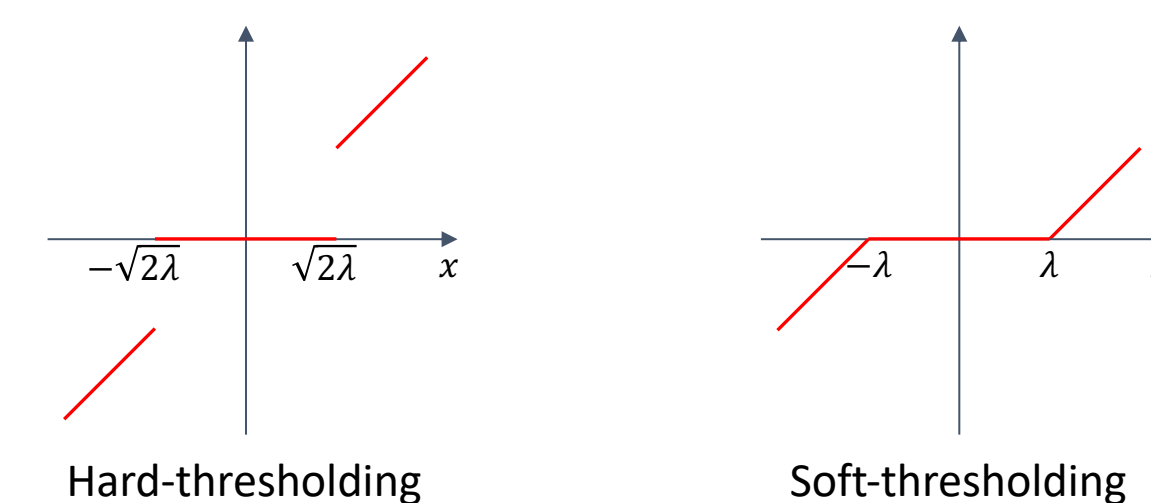
$$\tilde{\epsilon} = \epsilon_{src} + w \cdot (\epsilon_{tar} - \epsilon_{src})$$

• Ours:

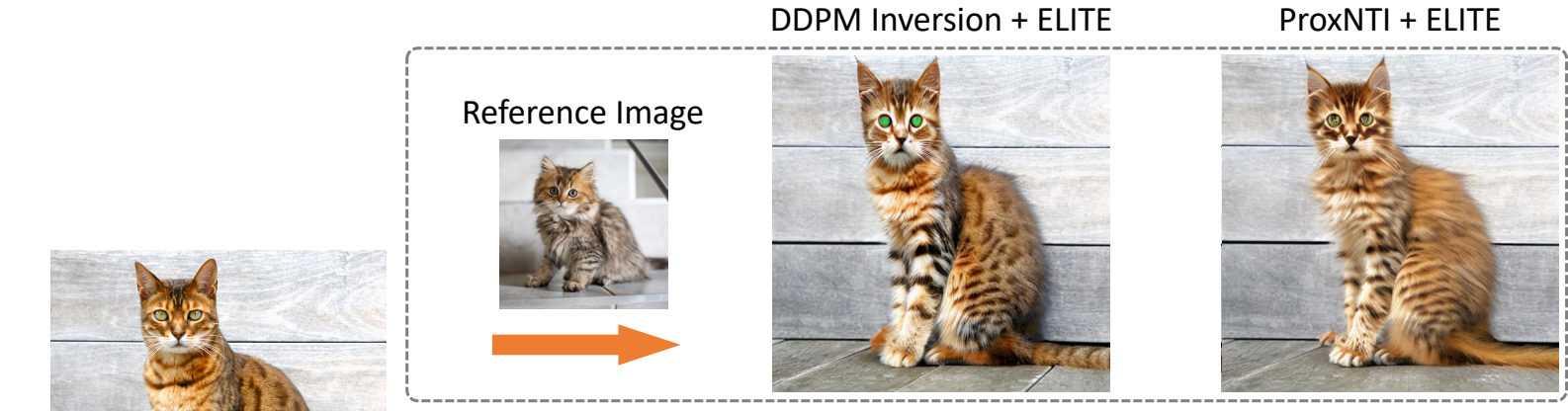
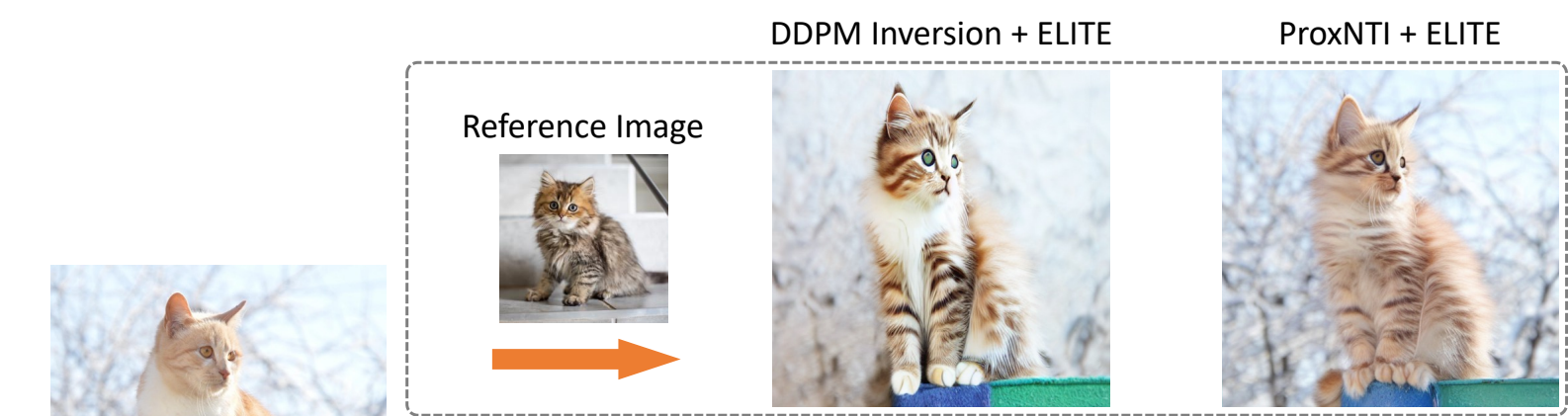
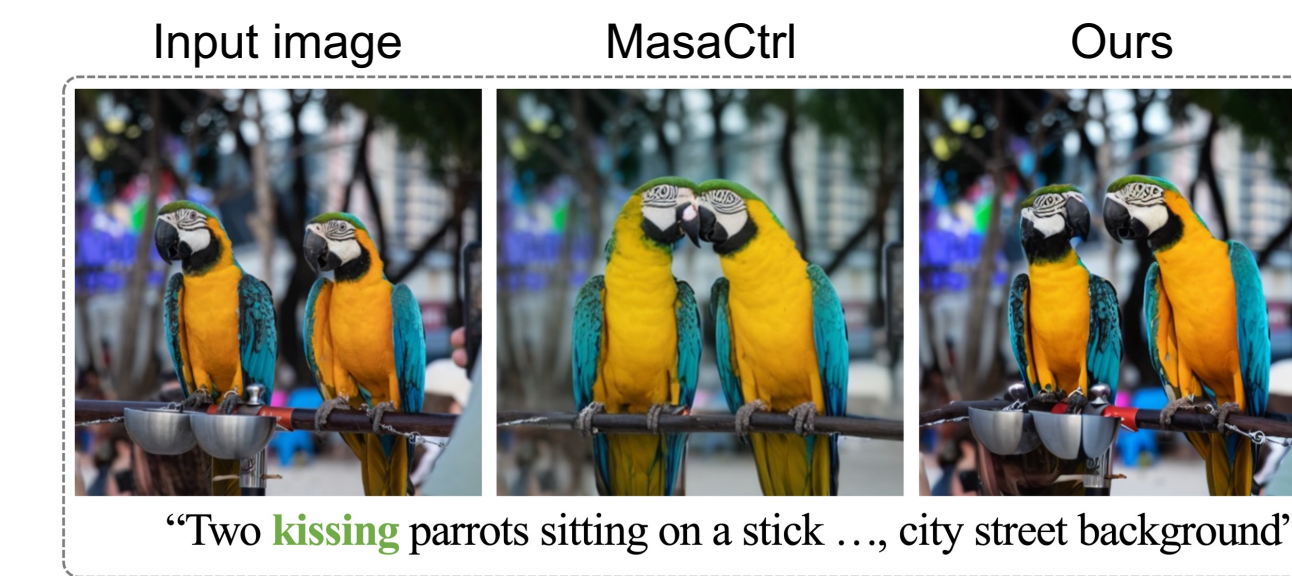
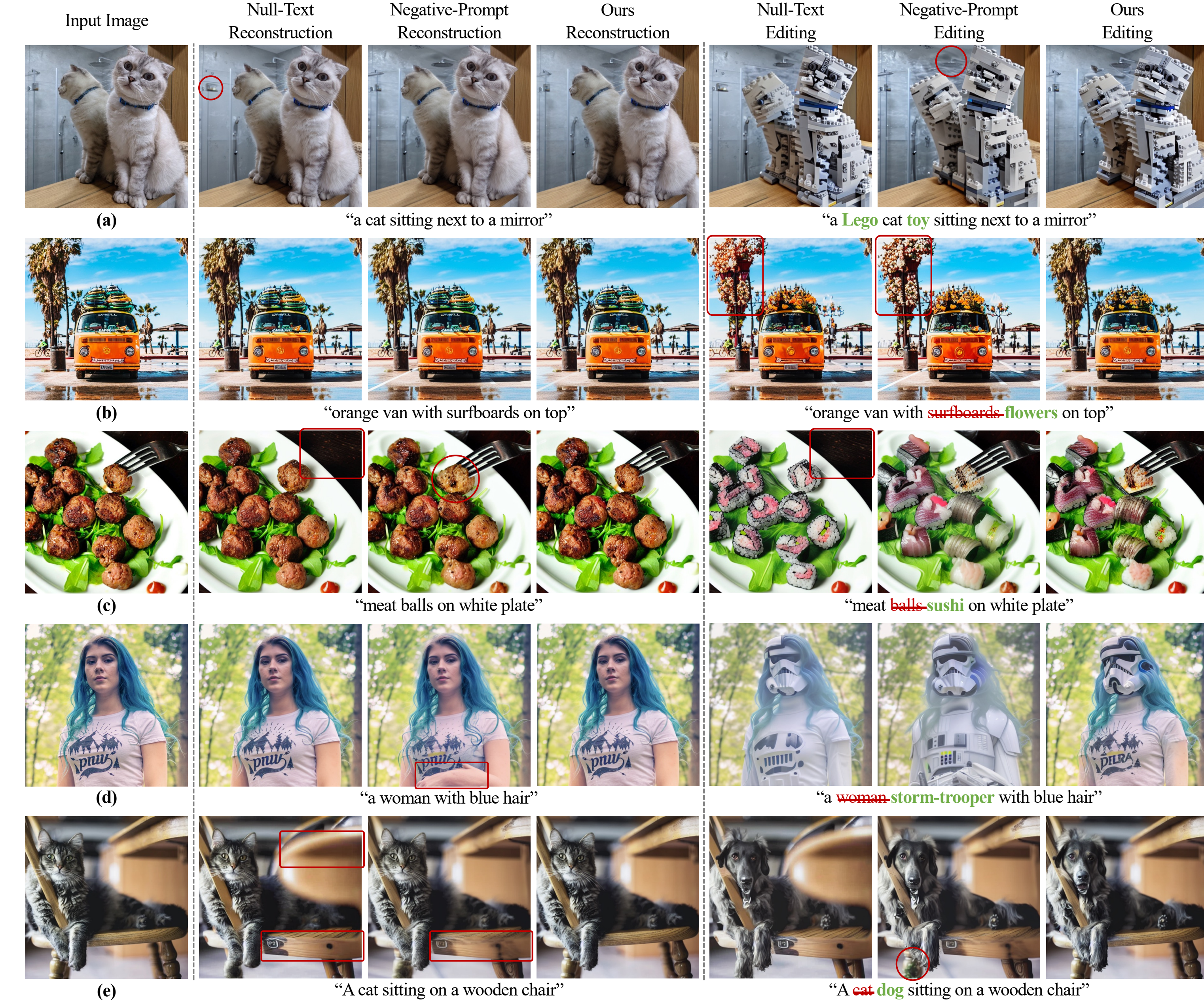
$$\tilde{\epsilon} = \epsilon_{src} + w \cdot \text{prox}_\lambda(\epsilon_{tar} - \epsilon_{src})$$

• Proximal operator:

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



Experiments



	Original		VAE	
Method	PSNR $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
NPI	25.214 $\pm$ 4.308	0.134 $\pm$ 0.083	28.802 $\pm$ 5.640	0.095 $\pm$ 0.092
ProxNPI	28.297 $\pm$ 4.139	0.057 $\pm$ 0.029	70.984 $\pm$ 2.021	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 $\rightarrow$ X		Dog $\rightarrow$ Y	
Method	CLIP $\uparrow$	LPIPS $\downarrow$	CLIP $\uparrow$	LPIPS $\downarrow$
NPI	27.371 $\pm$ 2.181	0.268 $\pm$ 0.077	28.229 $\pm$ 3.195	0.288 $\pm$ 0.106
ProxNPI	27.097 $\pm$ 2.507	0.205 $\pm$ 0.045	27.813 $\pm$ 3.414	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. $X = \{\text{“dog”, “hamster”, “fox”, “badger”, “lion”, “bear”, “pig”}\}$, and $Y = \{\text{“cat”, “hamster”, “fox”, “badger”, “lion”, “bear”, “pig”}\}$.

