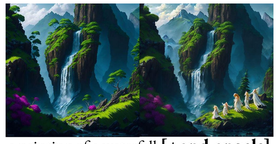
Inversion-Free Image Editing with Natural Language

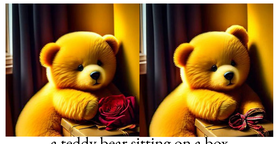
Sihan Xu*[1], Yidong Huang*[1], Jiayi Pan[2], Ziqiao Ma[1], Joyce Chai[1]
[1]University of Michigan; [2]University of California, Berkeley *equal contributions


bread meat on a wooden table with tomatoes and a napkin


a red **star heart** is sitting in the snow


a woman in a coat **[+and dress]** is dancing


a painting of a waterfall **[+and angels]** in the mountains


a teddy bear sitting on a box ~~**with a rose**~~


a lone tree is reflected in the water at night ~~**with a bright moon**~~

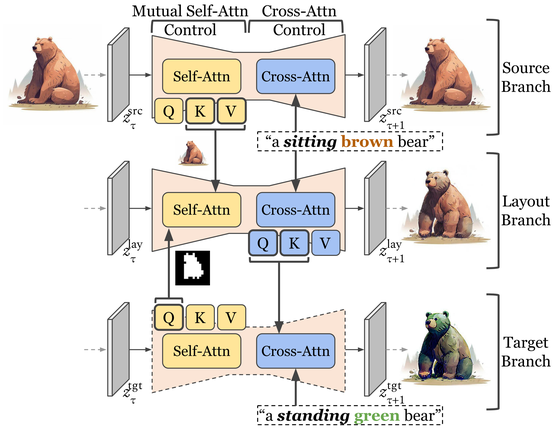
Motivation

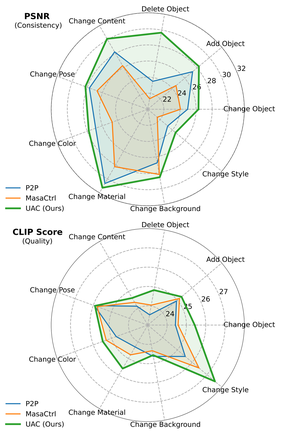
Inversion-based image editing methods, requiring an inversion branch as a series of anchors, are limited for real-time and real-world language-guided image editing applications.

- The inversion process is time-consuming;
- Balance between consistency and faithfulness is hard, even with optimization/calibration;
- Not compatible with *consistency sampling* using *consistency models*, which is more efficient.

Unified Attention Control

UAC unifies cross-attention control [2] and mutual self-attention control [3] with an additional latent *layout branch*, which serves as an intermediate to host the desired composition and structural information in the target image.





References

[1] Denoising Diffusion Implicit Models. *Jiaming Song et al.* In ICLR 2020.
[2] Prompt-to-Prompt Image Editing with Cross Attention Control. *Amir Hertz et al.* In ICLR 2022.
[3] MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. *Mingdeng Cao et al.* In ICCV 2023.

Links

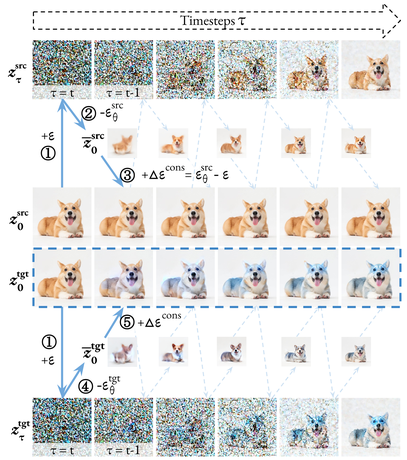
DDCM and Virtual Inversion

Sampling from a diffusion model is an iterative process that progressively denoises the data. Following Eq (12) in DDIM [1], the denoising step at t is formulated as:

$$z_{t-1} = \underbrace{\sqrt{\alpha_{t-1}}\left(\frac{z_t - \sqrt{1-\alpha_t}\varepsilon_\theta(z_t,t)}{\sqrt{\alpha_t}}\right)}_{\text{(predicted } z_0\text{)}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\varepsilon_\theta(z_t,t)}_{\text{(direction to } z_t\text{)}} + \underbrace{\sigma_t\varepsilon_t}_{\text{(random noise)}} \quad \text{where } \varepsilon_t \sim \mathcal{N}(0,I)$$

When σ_t is chosen as $\sqrt{1-\alpha_{t-1}}$ across all time t, the second term vanishes. The forward process naturally aligns with the form of Multistep (Latent) Consistency Sampling, consider $f(z_t,t;z_0) = (z_t - \sqrt{1-\alpha_t}\varepsilon'(z_t,t;z_0))/\sqrt{\alpha_t}$, where the initial z_0 is available in image editing:

$$\hat{z}_{\tau_i} = \sqrt{\alpha_{\tau_i}} z_0^{(\tau_{i+1})} + \sigma_{\tau_i}\varepsilon, \qquad z_0^{(\tau_i)} = f_\theta(\hat{z}_{\tau_i}, \tau_i, c)$$



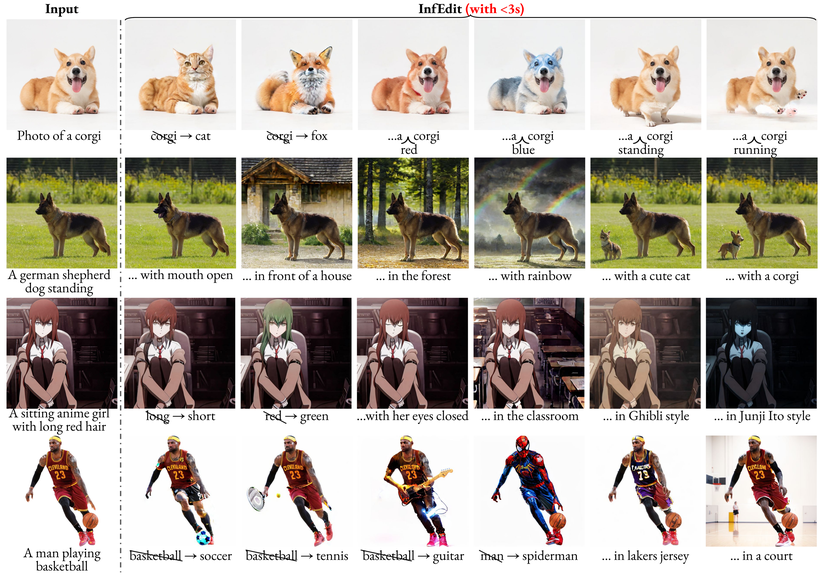
Algorithm 2 DDCM for inversion-free image editing

Input:
Conditional Diffusion/Consistency Model $\varepsilon_\theta(\cdot,\cdot,\cdot)$
Sequence of timesteps $\tau_1 > \tau_2 > \cdots > \tau_{N-1}$
Reference initial input z_0^{src}
Source/target prompts as conditions $c^{\text{src}}, c^{\text{tgt}}$

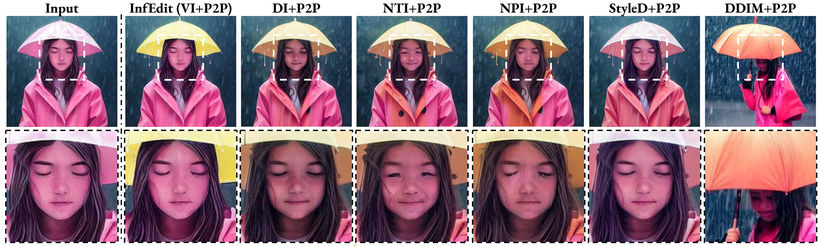
Sample a random terminal noise $z_{\tau_1}^{\text{src}} = z_{\tau_1}^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\varepsilon_{\tau_1}^{\text{cons}} = (z_{\tau_1}^{\text{src}} - \sqrt{\alpha_{\tau_1}} z_0^{\text{src}})/\sqrt{1-\alpha_{\tau_1}}$
$\varepsilon_{\tau_1}^{\text{src}}, \varepsilon_{\tau_1}^{\text{tgt}} = \varepsilon_\theta(z_{\tau_1}^{\text{src}}, \tau_1, c^{\text{src}}), \varepsilon_\theta(z_{\tau_1}^{\text{tgt}}, \tau_1, c^{\text{tgt}})$
$z_0^{\text{tgt}} = f_\theta(z_{\tau_1}^{\text{tgt}}, \tau_1, \varepsilon_{\tau_1}^{\text{tgt}} - \varepsilon_{\tau_1}^{\text{src}} + \varepsilon_{\tau_1}^{\text{cons}})$
for $n = 2$ to $N-1$ **do**
  Sample noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  ① $z_{\tau_n}^{\text{src}} = \sqrt{\alpha_{\tau_n}} z_0^{\text{src}} + \sqrt{1-\alpha_{\tau_n}}\varepsilon$
  ① $z_{\tau_n}^{\text{tgt}} = \sqrt{\alpha_{\tau_n}} z_0^{\text{tgt}} + \sqrt{1-\alpha_{\tau_n}}\varepsilon$
  ② $\varepsilon_{\tau_n}^{\text{src}} = \varepsilon_\theta(z_{\tau_n}^{\text{src}}, \tau_n, c^{\text{src}})$
  ③ $\varepsilon_{\tau_n}^{\text{cons}} = (z_{\tau_n}^{\text{src}} - \sqrt{\alpha_{\tau_n}} z_0^{\text{src}})/\sqrt{1-\alpha_{\tau_n}}$
  ④* $\varepsilon_{\tau_n}^{\text{tgt}} = \varepsilon_\theta(z_{\tau_n}^{\text{tgt}}, \tau_n, c^{\text{tgt}})$
  ⑤ $z_0^{\text{tgt}} = f_\theta(z_{\tau_n}^{\text{tgt}}, \tau_n, \varepsilon_{\tau_n}^{\text{tgt}} - \varepsilon_{\tau_n}^{\text{src}} + \varepsilon_{\tau_n}^{\text{cons}})$
end for
Output: z_0^{tgt}
*Vanilla target noise prediction, no attention control.

panda bear ~~**sitting**~~ **standing** on the ground holding a heart

Experiments and InfEdit Demos



Ablation Study



A girl with a pink yellow umbrella in the rain.

Method		Structure	Background Preservation				CLIP Similarity		Efficiency (sec / #)		
Inverse	Edit	Distance$_{10^3}$ ↓	PSNR ↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole ↑	Edited ↑	Inverse Time↓	Forward Time↓	Steps↓
NT	P2P	13.44	27.03	60.67	35.86	84.11	24.75	21.86	132.39 ± 7.69	12.90 ± 0.01	50
VI	P2P	14.22	**27.52**	**47.98**	**34.17**	**85.05**	**24.89**	**22.03**	N/A	**4.50** ± 0.01	**32**
VI*	P2P	15.61	26.64	55.85	41.15	84.66	24.57	21.69	N/A	2.60 ± 0.00	15
VI*	UAC	13.78	**28.51**	**47.58**	**32.09**	**85.66**	**25.03**	**22.22**	N/A	**2.22** ± 0.02	**12**

* Using the Latent Consistency Model (LCM) as the base model. Otherwise, Stable Diffusion (SD) v1.4 is adopted.