EditCrafter: Tuning-free High-Resolution Image Editing via Pretrained Diffusion Model

Kunho Kim[1] Sumin Seo[2] Yongjun Cho[3] Hyungjin Chung[4]

1 NCAI 2 medipixel 3 maum.ai 4 EverEx

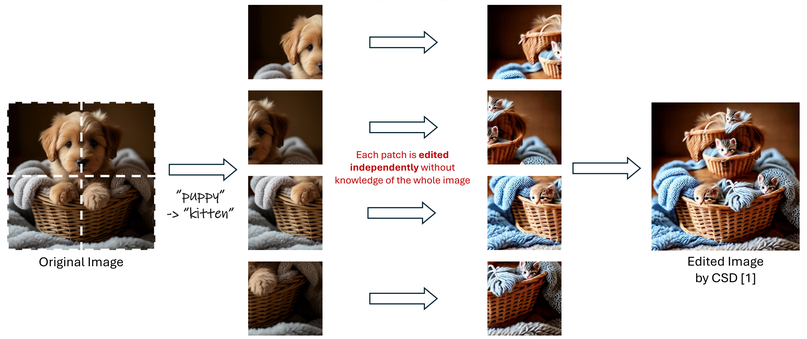
CVPR JUNE 3-7, 2026 DENVER COLORADO

Project Page

Motivation

Problem: Editing resolution is still the bottleneck

- Diffusion based editing methods inherit the **512×512 / 1024×1024 training resolution limit**.
- Real applications need arbitrary aspect ratios and 2K ~ 4K outputs with fine detail preservation.
- **Naïve patch-wise editing** breaks global structure, causing seams and **object repetitions**.
- Beyond image generation, there has been an emergence of high-resolution image generation. Nevertheless, the field of high-resolution image editing remains relatively underexplored.

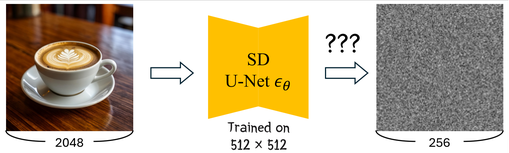


Challenge

1. How can we invert a high-resolution image into a latent using a fixed-size pretrained diffusion model?

Pretrained text-to-image diffusion models trained on 512×512 or 1024×1024 images.

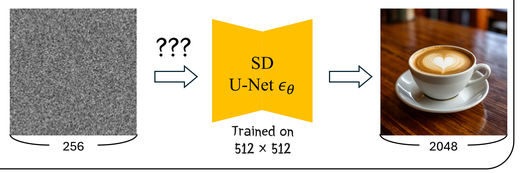
Direct DDIM inversion of high-resolution images yields poor latent.



2. If we succeed in creating an edit-friendly high-resolution latent, how can we guide the reverse process from this latent using a fixed-size pretrained diffusion model?

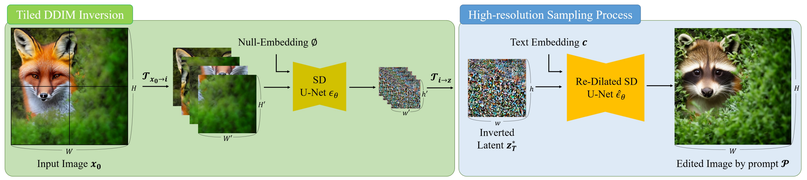
The fixed-size U-Net cannot process high-resolution latents directly.

Patch-wise reverse diffusion guided by a single prompt causes object repetition → CFG propagates the object to every patch even where it is absent.



EditCrafter Pipeline

Make a fixed resolution generator act as a high-resolution editor



Tiled DDIM Inversion

① Split high-resolution image x_0 into non-overlapping tiles S matching the T2I diffusion model's training resolution.
② Invert each tile independently with guidance scale $\omega = 0$ (no text conditioning)
→ avoids prompt–patch mismatch that arises from partial object content.
③ Concatenate the per-tile inverted latents to form edit-friendly high-resolution latent z_T^*.

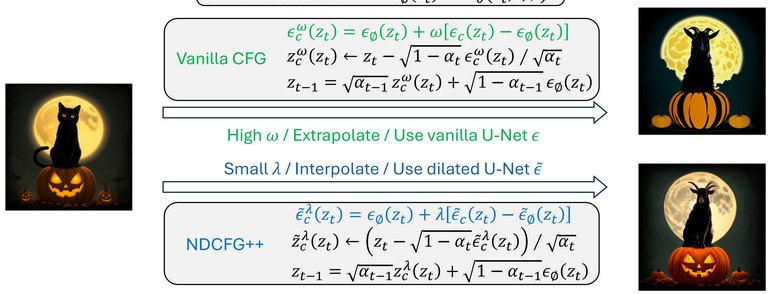
→ Tiled text-guidance-free inversion preserves identity of the original input image.

High-Resolution Reverse Diffusion with NDCFG++

① Adopt re-dilated convolutions throughout U-Net to expand the receptive field for higher-resolution latents while reusing pretrained parameters.
② Manifold-Constrained Noise-Damped CFG (NDCFG++):
- Small guidance scale $\lambda \in [0, 1]$ — interpolation (stays on manifold) vs. NDCFG's large $\omega \geq 1$ — extrapolation (may deviate).
- Renoising with vanilla noise estimator ϵ_θ → smoother editing trajectory.
- Applied for initial timesteps ($t \leq \tau$); standard CFG++ for remaining steps.

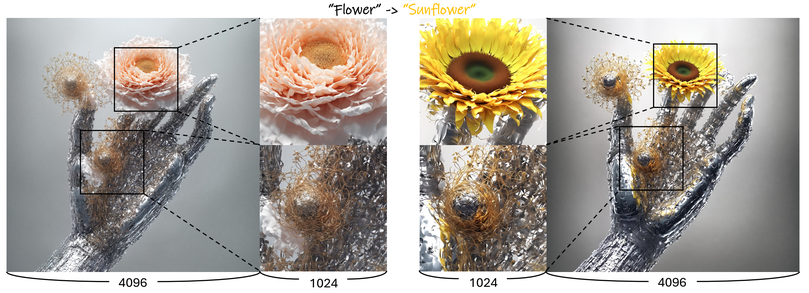
→ NDCFG++ interpolates between unconditional and conditional predictions, keeping estimates closer to the data manifold and producing faithful edits that preserve background textures and object identity in high-resolution editing.

$\epsilon := Vanilla\ U\text{-}Net \qquad \epsilon_c(z_t) := \epsilon_\theta(z_t, t, c)$
$\tilde{\epsilon} := Dilated\ U\text{-}Net \qquad \epsilon_\emptyset(z_t) := \epsilon_\theta(z_t, t, \emptyset)$



Results

Qualitative Results: SDXL 16× (4096×4096)



Quantitative Results

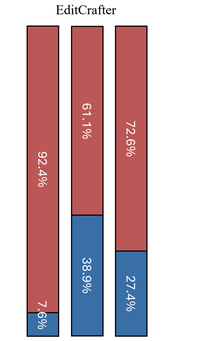
Model	Res	Method	ImageReward ↑	HPSv2 ↑	CLIPScore ↑
SD 2.1	4× 1:1	CSD	0.5538	0.2883	32.8353
		Ours	**1.4831**	**0.2935**	**34.8039**
	8× 1:2	CSD	0.7165	0.2782	32.2794
		Ours	**1.4238**	**0.2824**	**34.5303**
	16× 1:1	CSD	0.6304	0.2934	32.7795
		Ours	**1.6689**	**0.3017**	**35.3194**
SDXL 1.0	4× 1:1	CSD	0.6304	0.2934	32.7795
		Ours	**1.6242**	**0.2991**	**34.8067**
	8× 1:2	CSD	0.2939	0.2767	32.0854
		Ours	**1.4133**	**0.2842**	**33.9795**
	16× 1:1	CSD	0.3699	0.2877	32.8440
		Ours	**1.4919**	**0.2949**	**34.4959**

Ablation Study

Method	ImageReward ↑	HPSv2 ↑	CLIP Score ↑
Tiled Inv. + ScaleCrafter [2]	1.2595	0.2962	34.9431
Ours w/o NDCFG++	1.6273	0.2911	35.0254
Ours	**1.6689**	**0.3017**	**35.3194**

Comparison to SOTA Editing Method with Super-Resolution Upsampler

User Study Preference



[1] Kim et al., Collaborative Score Distillation for Consistent Visual Synthesis, NeurIPS 2023.
[2] He et al., ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models, ICLR 2024.
[3] Wang et al., Exploiting Diffusion Prior for Real-World Image Super-Resolution, IJCV 2024.
[4] Xu et al., Inversion-Free Image Editing with Natural Language. In CVPR, 2024.
[5] Han et al., ProxEdit: Improving Tuning-Free Real Image Editing with Proximal Guidance. In WACV, 2024.