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

Kunho Kim[1] Sumin Seo[2] Yongjun Cho[3] Hyungjin Chung[4]
[1]NC AI [2]Medipixel, Inc. [3]MAUM.AI [4]EverEx

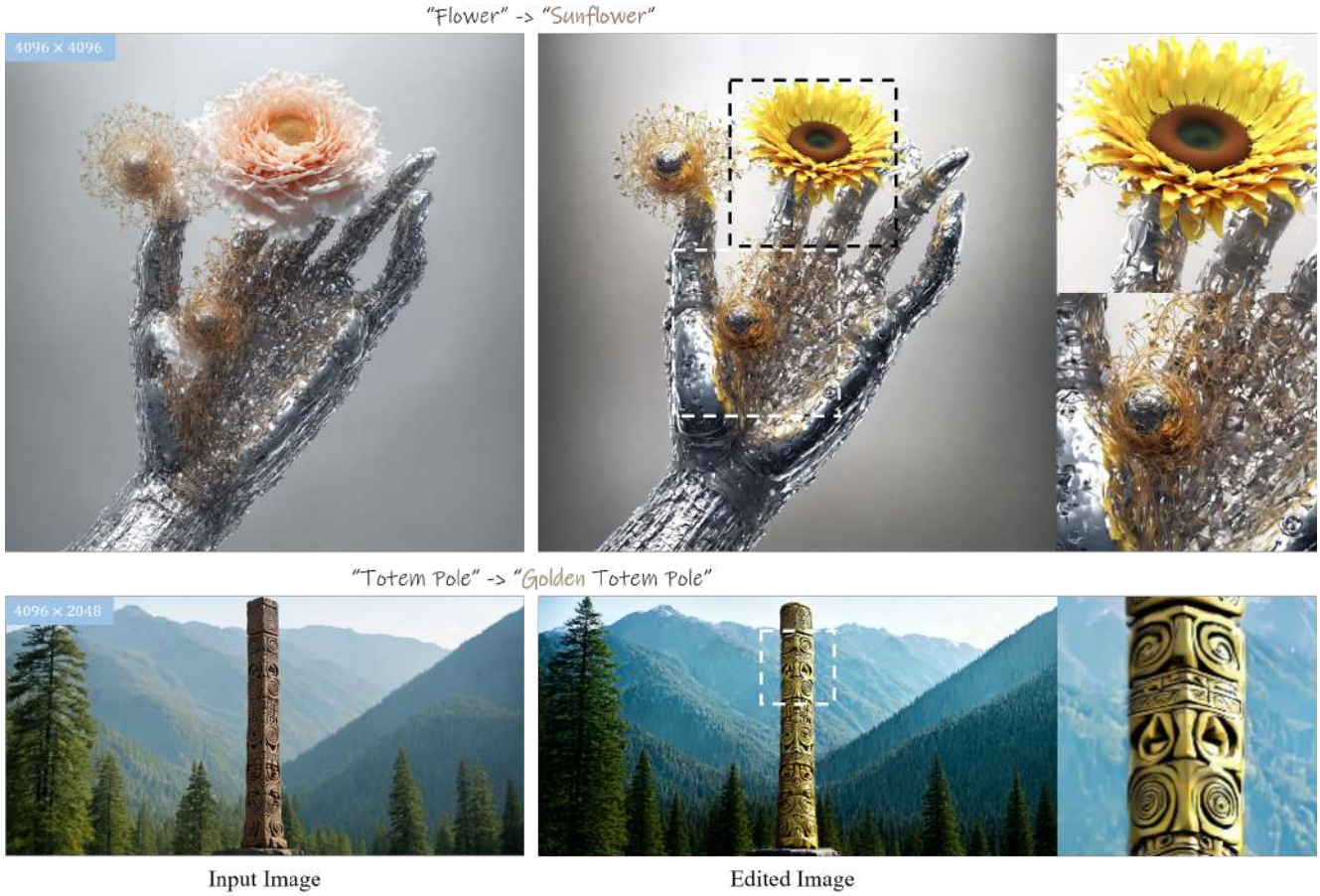


Figure 1. Our proposed framework, EDITCRAFTER, facilitates text-guided image editing at resolutions up to 4K while meticulously preserving the high-resolution details of the input images using only a single editing prompt.

Abstract

We propose EDITCRAFTER*, a high-resolution image editing method that operates without tuning, leveraging pretrained text-to-image (T2I) diffusion models to process images at resolutions significantly exceeding those used during training. Leveraging the generative priors of large-scale T2I diffusion models enables the development of a wide array of novel generation and editing applications. Although numerous image editing methods have been proposed based on diffusion models and exhibit high-quality editing results, they are difficult to apply to images with arbitrary aspect ratios or higher resolutions since they only work at the training resolutions (512×512 or 1024×1024). Naively applying patch-wise editing fails with unrealistic object structures and repetition. To address these challenges, we introduce* EDITCRAFTER*, a simple yet effective editing pipeline.* EDITCRAFTER *operates by first performing tiled inversion, which preserves the original identity of the input high-resolution image. We further pro-*

pose a noise-damped manifold-constrained classifier-free guidance (NDCFG++) that is tailored for high resolution image editing from the inverted latent. Our experiments show that the our EDITCRAFTER *can achieve impressive editing results across various resolutions without fine-tuning and optimization. Our project page is at* https://editcrafter.github.io

1. Introduction

Recent advancements in image synthesis, particularly with text-to-image (T2I) diffusion models trained on large-scale data, have garnered substantial attention from both academia and industry. In particular, T2I generation models such as Stable Diffusion [45], SDXL [38], Imagen [47], DALLE·2 [43], SD3.5 [12], and FLUX [29] have gained widespread popularity since its accessibility and notable image quality. Despite their success, these models are constrained to resolutions of 512×512 or 1024×1024 since they are trained on those resolutions.

Beyond image synthesis, there has been an emergence of intuitive and powerful text-based editing methods [3, 5, 10, 13, 15, 17, 19, 25, 30, 33, 34, 37, 46, 50] for semantically modifying images within pretrained T2I diffusion models, thereby enhancing users' control over the generated content. High-resolution image editing is particularly crucial across various domains, including digital content creation and industrial design, where maintaining fine-grained details and structural coherence is essential. However, similar to image synthesis, text-guided editing predominantly relies on pretrained diffusion models and therefore inherits limitations related to their training resolutions, which are insufficient for real-world high-resolution applications. These limitations pose challenges in applying existing editing methods to such high-resolution scenarios.

To overcome the resolution limitations of pretrained T2I diffusion models, one could directly train on high-resolution images. However, increasing the model size and training data scale demands substantial computational resources and extended development time. Consequently, previous studies [1, 31, 41] have explored the potential for generating arbitrary-sized images and panoramas, with their latent spaces designed to interact through a patch-wise joint diffusion process due to the inherent limitations of the fixed-resolution of T2I diffusion models. Nevertheless, the field of high-resolution image editing remains relatively under-explored. Similar to patch-wise high-resolution generation method [1, 31, 41], one approach [27] utilizes a joint diffusion process with kernel using the pretrained diffusion model to achieve consistency across a set of patch images. However, this approach often struggles to the object repetition issue as shown in Fig. 4. This issue of object repetition arises from guiding each patch with the same text prompt using classifier-free guidance [21], even when the object is absent in the patch. One naive approach is to independently apply text-guided image editing to each patch separately and then merge the patches to reconstruct the original high-resolution image. However, dividing a high-resolution image into local patches may result in each patch containing only a partial representation of the object rather than the full object, and the content of the patch may not align with the text prompt.

Building upon these observations, we introduce a tuning-free high-resolution image editing strategy, EDITCRAFTER. We carefully analyze the reconstruction capabilities of the inversion method in its patch-wise implementation to determine how to preserve the original information of the input high-resolution image. By employing appropriate tiled inversion techniques, we obtain high-resolution latent representations that are conducive to editing.

A subsequent challenge is guiding these high-resolution latents with a single text prompt to achieve high-quality edits. To address this, we adopt the high-resolution image generation method [18], which replaces the standard convolution layers with dilated convolution layers throughout the entire U-Net, utilizing the pretrained parameters as the editing framework. Furthermore, to effectively utilize the generator as an editor, we propose a manifold-constrained noise-damped classifier-free guidance (NDCFG++) approach. This sampling process enhances the ability to guide the high-resolution latents accurately, ensuring high-quality image edits aligned with the provided text prompts.

To investigate the effectiveness of our proposed approach, we conduct extensive experiments on high-resolution image editing, evaluating our method both quantitatively and qualitatively. Ablation studies demonstrate the superiority of our method over a direct approach including super-resolution upsampler with editing method using down-sized input images. We curated image-text pairs using a high-resolution generation model [44] to quantitatively evaluate our method, showing that ours outperforms baseline approaches in terms of human preference and image-text alignment, and further indicating that our design choice yields superior results. Qualitative results illustrate that our method effectively modifies the target object while avoiding the object repetition, achieving strong alignment with the text prompt. This improvement in inversion strategy and text guidance adherence highlights the contribution of our approach, which better follows the specified instructions compared to a previous method [27], without visible seams between patches or unwanted object repetition.

2. Related Work

2.1. Text-to-Image Diffusion Models

Diffusion models [22, 45] have demonstrated remarkable success in generating high-quality images, providing a foundation for later advancements in various image generation tasks including text-guided image generation [35, 43, 47] and image-to-image translation [16, 26, 36, 56]. Recent work has extended T2I diffusion models to various domains, including video generation [4, 23] and 3D generation [28, 32, 39, 48], where the focus lies in ensuring temporal and spatial consistency across generated frames and structures. Furthermore, ControlNet [56] has introduced end-to-end architecture for integrating various conditioning inputs—such as sketches, depth maps, or poses—within the diffusion process, thereby enhancing the training efficiency and adaptability of large-scale T2I models to follow complex editing conditions. Our method enables higher-resolution image editing which utilizes the generation capability of pretrained large-scale T2I diffusion models [45].

2.2. High-resolution Image Generation

High-resolution image generation has progressed through efforts to address key challenges, including computational inefficiency, high training costs, and structural artifacts. Training models directly on high-resolution images [6, 7, 44, 53] is indeed a viable approach. However, this method inherently requires a substantial increase in both model complexity and the volume of training data.

High-resolution image generation also encompasses diffusion models for panoramic view synthesis, which face significant challenges in achieving seamless patch integration and maintaining structural coherence across expansive fields of view [1, 31, 41]. Latent merging methods [1, 31] have introduced a foundational approach to panoramic image generation by employing an averaging technique to smooth transitions between patches, resulting in coherent extended views. However, despite achieving high-fidelity panoramas, they still encounters seam artifacts where patches meet, impacting the overall continuity of the image. To address lack of preserving semantic coherence across intricate scenes and refining structural detail where needed, merging-and-splitting diffusion method [41] builds on an attention-based mechanism that adapts to complex scene structures. These approaches are inherently limited by the appearance of seams between patches. To address this, we introduce a high-resolution image generation module that integrates patch-wise inverted latents, avoiding the need to generate each patch separately.

Another approach for high-resolution image generation employs kernel dilation [18, 24, 40] in conjunction with fixed-size pretrained diffusion models. ScaleCrafter [18] introduces a training-free upscaling approach leveraging kernel dilation, which eliminates the need for model retraining. Concurrently, FouriScale [24] has applied a frequency-domain perspective to diffusion models with kernel dilation, successfully reducing repetitive patterns and structural distortions, while maintaining visual integrity across scales. Building on collective advancements in high-resolution image generation [18], we design a high-resolution editing module adapted from these approaches.

2.3. Image Editing with Diffusion Models

Recent studies show that controlling the attention mechanism [3, 5, 17, 19] in the pretrained diffusion models provide fine-grained control over attention maps to guide the generation process to achieve desired edited details while maintaining the overall consistency of the image. In parallel, utilizing the inversion process in editing [17, 25, 33, 34, 37, 49], which maps a source image to the latent space of the model, enables both generated images from the generation models and real images editing with conditional inputs such as text prompts and reference images. Yet, the inversion process is excessively lengthy and underperforms when the data distribution differs from the training data, highlighting the need for investigating text-guidance-free inversion method [33]. However, inversion-based image editing methods are limited by the capabilities of pretrained diffusion models [38, 45], which significantly restrict high-resolution image editing.

To the best of our knowledge, CSD [27] is the first to achieve high-resolution image editing, utilizing a score distillation approach for synchronous patch-wise generation. CSD proposes score functions based on the Stein Variational Gradient Descent (SVGD) method, which are applied across patches to enable more coherent generative priors for high-resolution images, including panoramic views. However, boundary artifacts between patches remain a challenge due to the quality of patch-wise sampling and are limited by the dependency on pretrained models [2]. Our study addresses this limitation by incorporating patch-wise editing along with high-resolution generation models, allowing seamless, artifact-free editing across image patches, suggested with extensive high-quality image editing results on qualitative analysis.

3. Method

Our goal to edit the high-resolution image $x_0 \in \mathbb{R}^{H\times W\times 3}$ using the text prompt $\mathcal{P}$ with a text-guided diffusion model [38, 45] which are trained by the fixed-size low-resolution image $x_0' \in \mathbb{R}^{H'\times W'\times 3}$. To perform such editing operations, it is essential to first invert the image x_0 into the latent domain. The primary challenge lies in accurately inverting a high-resolution image x_0 with the fixed-size pretrained T2I diffusion model, while simultaneously preserving the model's intuitive text-based editing capabilities.

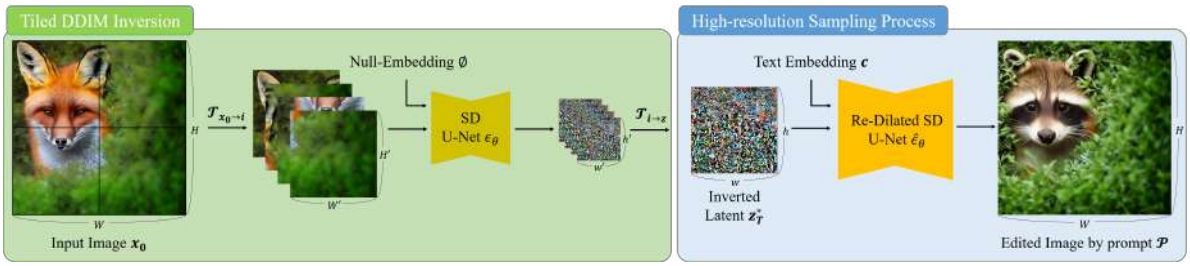


Figure 2. The overview of EDITCRAFTER pipeline. Since direct inversion of high-resolution images using the pretrained Stable Diffusion (SD) model is not feasible, we first perform tiled DDIM inversion to generate a high-resolution latent representation. Utilizing this latent, the reverse diffusion process is carried out with a re-dilated noise estimator. To enhance the quality of text-guided editing, we propose manifold-constrained noise-damped classifier-free guidance (NDCFG++). In this figure, editing prompt $\mathcal{P}$ is *"A raccoon peeking out from behind a bush"*.

Diffusion models are typically trained on resolutions of 512×512 or 1024×1024, which limits their capability to directly invert high-resolution images x_0. To address this limitation, we propose a straightforward yet effective method called tiled DDIM inversion. This approach generates high-resolution latent that is amenable to editing while preserving the original properties of the input high-resolution image x_0.

After inverting to obtain an edit-friendly high-resolution latent, an additional challenge emerges: the fixed-size diffusion model cannot handle high-resolution latents. To solve this problem, we adopt the high-resolution image generator [18], which employs re-dilation techniques to adjust the network's receptive field for higher-resolution images. However, our observations indicate that directly utilizing the high-resolution image generator does not yield adequate editing capabilities. Therefore, we modify the guidance mechanism of the image generator, enabling it to function effectively as an image editing tool.

In Sec. 3.1, we outline the foundational concepts of latent diffusion models. Sec. 3.2 introduces our tiled DDIM inversion technique for generating high-resolution latents. Finally, we describe how to guide the high-resolution latent using fixed-size pretrained diffusion models for editing purposes in Sec. 3.3.

3.1. Preliminaries

The latent diffusion model (LDM) uses the encoder $\mathcal{E}$ to encode image $x_0' \in \mathbb{R}^{H'\times W'\times 3}$ into a latent representation $z_0 = \mathcal{E}(x_0') \in \mathbb{R}^{h'\times w'\times c}$ and the decoder $\mathcal{D}$ to reconstruct the image from the latent $\tilde{x}_0' = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x_0'))$. Similar to the denoising diffusion probabilistic models (DDPM) [22], LDM executes a forward diffusion process across time steps t ranging from 0 to T in the latent space rather than the image space. The forward process that add the noise to the original sample z_0 to generate the noisy tractable sample z_t as follows:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}}z_0, (1-\bar{\alpha}_t)I), \tag{1}$$

where α_t is a predefined variance scheduler and $\bar{\alpha} = \prod_{i=1}^{t}\alpha_i$. After the sufficient time steps T, $q(z_t|z_0)$ converges to a unit Gaussian $\mathcal{N}(0, I)$. The reverse process gradually remove the noise and predict a clean sample z_{t-1} from the previous sample z_t:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \beta_t I), \tag{2}$$

where μ_θ and β_t represent the mean and the time-dependent constant variance, respectively. The mean $\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$ can be calculated by the noise estimator predicting the noise ϵ given z_t.

The goal is training the noise estimator (U-Net) to predict the noise ϵ from the noisy latents z_t. In text-guided diffusion models, we can control the reverse process through input a text embedding c of a text prompt P, which is obtained using a text encoder such as CLIP [42].

Since Stable Diffusion (SD) is trained with text conditioning, we can modulate the influence of the conditional text prompt $\mathcal{P}$ during the reverse diffusion process by employing classifier-free guidance (CFG) [21]. Utilizing the null-text embedding $\varnothing$ extracted from a null-text "" and the guidance scale parameter ω, the noise prediction incorporating classifier-free guidance is computed as follows:

$$\begin{aligned}\epsilon_\theta(z_t, t, c, \varnothing) &= \epsilon_\theta(z_t, t, \varnothing) \\ &\quad + \omega\cdot(\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \varnothing)),\end{aligned} \tag{3}$$

where the guidance scale $w \geq 0 \in \mathbb{R}$ regulates the relative influence of the conditional prediction $\epsilon_\theta(z_t, t, c)$ compared to the unconditional prediction $\epsilon_\theta(z_t, t, \varnothing)$.

For simplicity in notation throughout the subsequent sections, we define the conditional prediction $\epsilon_c(z_t) := \epsilon_\theta(z_t, t, c)$ and the unconditional prediction $\epsilon_\varnothing(z_t) := \epsilon_\theta(z_t, t, \varnothing)$. Consequently, Eq. 3 can be expressed as follows:

$$\epsilon_c^\omega(z_t) = \epsilon_\varnothing(z_t) + \omega[(\epsilon_c(z_t) - \epsilon_\varnothing(z_t)]. \tag{4}$$

The reverse process in SD is inherently stochastic, adhering to the training framework established by DDPM [22]. To accelerate the sampling speed, the deterministic reverse process of the DDIM sampling method [49] is commonly employed. The reverse DDIM process that generates the latent of previous time step z_{t-1} from the current time step latent variable z_t can be written as:

$$z_{t-1} = \sqrt{\alpha_{t-1}} z_c^{\omega}(z_t) + \sqrt{1-\alpha_{t-1}}\epsilon_c(z_t), \tag{5}$$

where $z_c^{\omega}(z_t) = (z_t - \sqrt{1-\alpha}\epsilon_c^{\omega}(z_t)) / \sqrt{\alpha_t}$. This reverse process operates in the direction $z_t \rightarrow z_0$.

As detailed in previous work [11, 49], the reverse DDIM process described in Eq. 5 is approximately invertible. By rearranging Eq. 5, we can derive the DDIM inversion process as follows:

$$z_{t+1} = \sqrt{\alpha_{t+1}} z_c^{\omega}(z_t) + \sqrt{1-\alpha_{t+1}}\epsilon_c(z_t), \tag{6}$$

which operates in the direction $z_0 \rightarrow z_t$. Consistent with prior work [9, 17, 33, 37], we employ this DDIM inversion for the editing process.

3.2. Tiled DDIM Inversion

Since the noise estimator (U-Net) in Stable Diffusion is trained on low-resolution images, directly inverting an encoded high-resolution image $z_0 = \mathcal{E}(x_0)$ into a high-resolution latent z_t is not feasible. To overcome this limitation, we split the high-resolution image into the low-resolution tiles and invert each tile separately. Let the tile size $S \in \mathbb{R}^{H' \times W' \times 3}$ match the training resolution of the noise estimator. We define the cropping function $\mathcal{T}_{x \rightarrow i}(S) : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H' \times W' \times 3}$ to extract the i-th region $x^{(i)}$ from a high-resolution image x. Its inverse function, $\mathcal{T}_{i \rightarrow x}(S) : \mathbb{R}^{H' \times W' \times 3} \rightarrow \mathbb{R}^{H \times W \times 3}$ reintegrates the tiled image into the i-th region of the high-resolution image. Utilizing the function $\mathcal{T}_{x \rightarrow i}(S)$, high-resolution image is partitioned into non-overlapping tiles with size S thereby facilitating DDIM inversion using Stable Diffusion. During the inversion of each tile, we minimize the influence of the text condition by setting the CFG guidance scale to $\omega = 0$ in Eq. 4, thereby ensuring that $\epsilon_c^{\omega}(z_t) = \epsilon_{\varnothing}(z_t)$. Consequently, the Eq. 6 becomes:

$$z_{t+1} = \sqrt{\alpha_{t+1}} z_c^{\omega}(z_t) + \sqrt{1-\alpha_{t+1}}\epsilon_{\varnothing}(z_t) \tag{7}$$

After inverting each latent $z_T^{(i)*}$, we concatenate all latents to form the high-resolution inverted latent z_T^* using the function $\mathcal{T}_{i \rightarrow z}(S / 8)$.

This tiled DDIM inversion approach enables the DDIM inversion regardless of input image size as shown in Fig. 2. Our key observation is that concatenating each inverted latent and creating the high-resolution inverted latent can provide a good initial point for the subsequent reverse process during editing. The pseudocode of tiled DDIM inversion is detailed in Alg. 1.

Algorithm 1 Tiled DDIM Inversion

Require: Real image $x_0 \in \mathbb{R}^{H \times W \times 3}$, Patch size $S \in \mathbb{R}^{H' \times W' \times 3}$

$\{x^{(0)}, x^{(1)}, \cdots, x^{(n)}\} = \mathcal{T}_{x_0 \rightarrow i}(S)$ ▷ Tiling
for $i = 0$ **to** n **do**
  $z_0^{(i)} = \mathcal{E}(x^{(i)})$ ▷ Encode image
  for $j = 0$ **to** $T-1$ **do**
    $\epsilon_c^{\omega}(z_t^{(i)}) = \epsilon_{\varnothing}(z_t^{(i)})$
    $z_c^{\omega}(z_t^{(i)}) \leftarrow (z_t^{(i)} - \sqrt{1-\bar{\alpha}_t}\epsilon_{\varnothing}(z_t^{(i)})) / \sqrt{\bar{\alpha}_t}$
    $z_{t+1}^{(i)} = \sqrt{\bar{\alpha}_{t+1}} z_c^{\omega}(z_t^{(i)}) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_c^{\omega}(z_t^{(i)})$
  end for
  $z_T^{(i)*} = \mathcal{T}_{i \rightarrow z}(S / 8)$ ▷ Mapping
end for
return z_T^*

3.3. High-resolution Sampling Process with Kernel Dilation

Similar challenges arise during the inversion process, as fixed-size diffusion models are incapable of handling high-resolution latents in the sampling (reverse) process. An existing method [27] addresses this limitation by employing a patch-wise joint reverse process, where overlapping regions between patches interact. However, we observe that this approach leads to an unintended issue with object repetition, as it typically relies on a single editing prompt $\mathcal{P}$. This causes SD to propagate the object information specified by the text condition across all patches. To mitigate this problem, we adopt kernel re-dilation method from ScaleCrafter [18], which enables seamless high-resolution image generation guided by text. A kernel re-dilation technique adjust the network's receptive field for higher-resolution images, ensuring consistency with the receptive field used in the original lower-resolution generation. With the dillated kernel, we start the reverse process from a high-resolution latent $z_T \sim \mathcal{N}(0, I_d) \in \mathbb{R}^{h \times w \times c}$ to generate a high-resolution image $x_0 \in \mathbb{R}^{H \times W \times 3}$.

Increasing the convolutional receptive field within dilated blocks impairs the model's denoising capabilities. To accurately reconstruct fine structural details while preserving the original denoising performance, noise-damped classifier-free guidance (NDCFG) [18] incorporates a vanilla noise estimator with strong denoising capabilities ϵ_θ, and a kernel re-dilated noise estimator that generates fine content structures $\tilde{\epsilon}_\theta$ as follows:

$$\tilde{\epsilon}_c^{\omega}(z_t) = \epsilon_{\varnothing}(z_t) + \omega[\tilde{\epsilon}_c(z_t) - \tilde{\epsilon}_{\varnothing}(z_t)] \tag{8}$$

$$\tilde{z}_c^{\omega}(z_t) \leftarrow (z_t - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\omega}(z_t)) / \sqrt{\alpha_t} \tag{9}$$

$$z_{t-1} = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\omega}(z_t) + \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_c^{\omega}(z_t). \tag{10}$$

When we try to edit with CFG $\omega = 7.5$ as typically used in SD, we observe that it cannot preserve the original input

image's information since it was originally devised for the generation purpose. Yet, Chung et al. [8] has demonstrated that an excessive guidance scale in CFG can degrade the generation quality of T2I diffusion models, whereas their method achieves improved fidelity by reformulating text guidance.

To integrate the idea that controls detailed information at smaller text guidance scale while preserving object-level information during sampling, we propose manifold-constrained noise-damped classifier-free guidance (ND-CFG++) as follows:

$$\tilde{\epsilon}_c^{\lambda}(z_t) = \epsilon_{\varnothing}(z_t) + \lambda[\tilde{\epsilon}_c(z_t) - \tilde{\epsilon}_{\varnothing}(z_t)] \tag{11}$$

$$\tilde{z}_c^{\lambda}(z_t) \leftarrow (z_t - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\lambda}(z_t)) \,/\, \sqrt{\alpha}_t \tag{12}$$

$$z_{t-1} = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\lambda}(z_t) + \sqrt{1-\alpha_{t-1}}\epsilon_{\varnothing}(z_t), \tag{13}$$

where $\lambda \in [0, 1]$ is a small guidance scale. This NDCFG++ is applied during the initial time steps when $t \leq \tau$, starting from our inverted latent z_T^*. When $t \geq \tau$, we adhere to the standard CFG++ reverse steps to preserve consistency.

Since NDCFG extrapolates beyond and unconditional noise prediction provided by vanilla noise estimator ϵ_θ and the difference between conditional and unconditional noise predictions from dilated noise estimator $\tilde{\epsilon}_\theta$ with the large guidance scale $\omega \geq 1$ (Eq. 8), the resulting estimates may potentially deviate from the data manifold. While ND-CFG++ interpolates between that unconditional prediction and the difference conditional and unconditional predictions with the small guidance scale $\lambda \in [0, 1]$, the resulting estimates are less likely deviate from the data manifold. Additionally, a key distinction between NDCFG and ours NDCFG++ lies in the renoising process (Eq. 10 & 13). Using the unconditional noise $\epsilon_{\varnothing}(z_t)$ predicted by vanilla noise estimator ϵ_θ instead of $\tilde{\epsilon}_c^{\omega}(z_t)$ provides smoother trajectory of editing as shown in Fig. 3. The whole algorithm of our sampling process is outlined in Alg. 2.

Algorithm 2 Reverse Diffusion with Ours

Require: Inverted latent z_T^*, $\lambda \in [0, 1]$, $\tau \leq T \in \mathbb{R}$

for $t = T$ **to** 1 **do**
  if $t \leq \tau$ **then** ▷ NDCFG++
    $\tilde{\epsilon}_c^{\lambda}(z_t^*) = \epsilon_{\varnothing}(z_t^*) + \lambda[\tilde{\epsilon}_c(z_t^*) - \tilde{\epsilon}_{\varnothing}(z_t^*)]$
    $\tilde{z}_c^{\lambda}(z_t^*) \leftarrow (z_t^* - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\lambda}(z_t^*)) \,/\, \sqrt{\alpha}_t$
    $z_{t-1}^* = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\lambda}(z_t^*) + \sqrt{1-\alpha_{t-1}}\epsilon_{\varnothing}(z_t^*)$
  else ▷ Vanilla CFG++
    $\tilde{\epsilon}_c^{\lambda}(z_t^*) = \tilde{\epsilon}_{\varnothing}(z_t^*) + \lambda[\tilde{\epsilon}_c(z_t^*) - \tilde{\epsilon}_{\varnothing}(z_t^*)]$
    $\tilde{z}_c^{\lambda}(z_t^*) \leftarrow (z_t^* - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\lambda}(z_t^*)) \,/\, \sqrt{\alpha_t}$
    $z_{t-1}^* = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\lambda}(z_t^*) + \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_{\varnothing}(z_t^*)$
  end if
end for
$x_0 = \mathcal{D}(z_0^*)$ ▷ Decode latent
return x_0

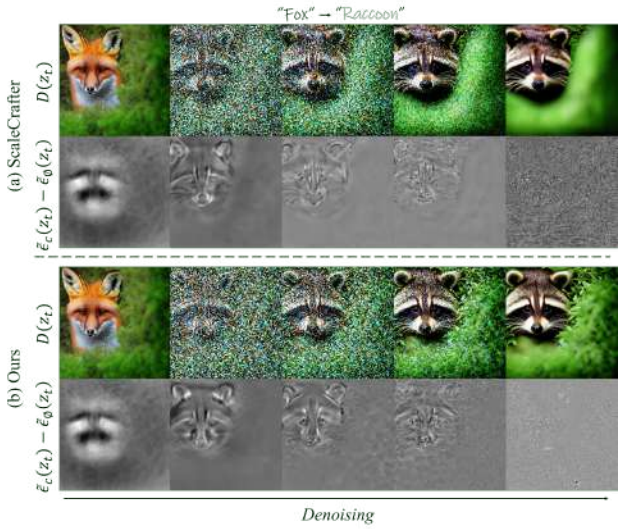


Figure 3. The first and third rows visualize the decoded latents over successive denoising steps. The second and fourth rows show the guidance residual—i.e., the difference between the dilated conditional and unconditional predictions $\epsilon_c(z_t) - \epsilon_\varnothing(z_t)$. As denoising progresses, our method (NDCFG++) preserves more semantically faithful signal and suppresses background noise, compared with directly applying ScaleCrafter (NDCFG) after tiled inversion.

4. Experiments

4.1. Experiment Setup

We conduct experiments using two recent pretrained versions of Stable Diffusion, specifically SD 2.1 [45] and SDXL 1.0 [38], to perform editing at resolutions exceeding their respective training resolutions 512×512, and 1024 × 1024. We scale the number of pixels by factors of 4, 8, and 16, resulting in editing resolutions of 1024×1024, 2048×1024, and 2048×2048 for SD 2.1, and 2048×2048, 4096×2048, and 4096×4096 for SDXL. All experiments are conducted on a single RTX 4090, demonstrating that our method does not require extensive VRAM (ranging from 3.8GB at 1024×1024 to 18.2GB at 4096×4096). Unless otherwise specified, all of our experiments are performed using a small guidance scale parameter $\lambda = 0.5$. The ablation study on the guidance scale is provided in the **supplementary material**.

4.1.1. Benchmark

To evaluate the high-resolution editing performance, we collect an image editing dataset from high-resolution generation model [44], known for its high visual fidelity and close resemblance to the natural images. The dataset includes 30 images with manually selected text prompts for each resolution including high-resolution square images and wide panoramic images, yielding a total of 150 prompt-image pairs. For creating editing prompts, we applied a word-swapping technique to the original prompts used for image

Figure 4. Qualitative comparisons. (1) Original image, (2) Ours, and (3) CSD in 4×, 8× and 16× settings. Best viewed on screen with zoom. The high-quality versions are provided in the **supplementary material**.

generation, replacing nouns that describe the main object or phrases depicting the background.

4.1.2. Baselines

To the best of our knowledge, the only existing baseline specifically designed for high-resolution image editing is CSD [27]. CSD leverages a joint diffusion approach to generate fixed-size patches using InstructPix2Pix [2], a fine-tuned variant of Stable Diffusion. Throughout the experiments, we use the default parameter settings for CSD.

4.1.3. Evaluation Metrics

We utilize HPSv2 [52] and ImageReward [54] to assess text-to-image alignment based on human preferences. Both models are trained on datasets consisting of human preference selections for images corresponding to given text prompts, using 645k and 137k text-image pairs, respectively. We also evaluate baseline methods using CLIPScore [20] to assess editing quality. CLIPScore measures the similarity between the edited image embedding and the text embedding extracted from the editing prompt $\mathcal{P}$.

In addition to the metrics mentioned above, user evaluation is a crucial aspect of image editing tasks. Therefore, we conducted a user study using Amazon MTurk to evaluate the effectiveness of our method. A detailed example of the user study is provided in the **supplementary material**.

4.2. Text-Guided High-Resolution Image Editing

4.2.1. Qualitative Evaluation

We present qualitative results to compare our method with the baseline in Fig. 4. CSD [27] frequently exhibits repetitive objects due to its patch-wise generation scheme. This limitation is illustrated by instances such as pandas appearing on the head of a tiger in edits at 2048×1024 resolution and koalas spanning the entire body of a chameleon in edits

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**

Table 1. Quantitative comparisons.

at 2048×2048 resolution. Furthermore, as the editing resolution increases, CSD tends to adhere to the original image rather than the editing prompt and produce white grids indicating patch boundaries. In contrast, our EDITCRAFTER demonstrates the capability to modify the shape of the target object or adjust background layout as shown in the edited results for object-centric and panoramic view images, respectively. It consistently generates highly faithful editing images that are well-aligned with the editing prompts while preserving the intricate details of the original images. We present more visual editing results in the **supplementary material**.

We also compare our method to a naive baseline for high-resolution image editing, which involves downsampling the image to apply current state-of-the-art editing methods, followed by upsampling to the desired higher resolution using super-resolution techniques. To compare ours with state-of-the-art fixed-size editing method, we applied 16× upsampling using StableSR [51] on images edited by InfEdit [55] which takes the input images resized by 512×512 from 2048×2048.

Fig. 5 showcases a qualitative comparison highlighting

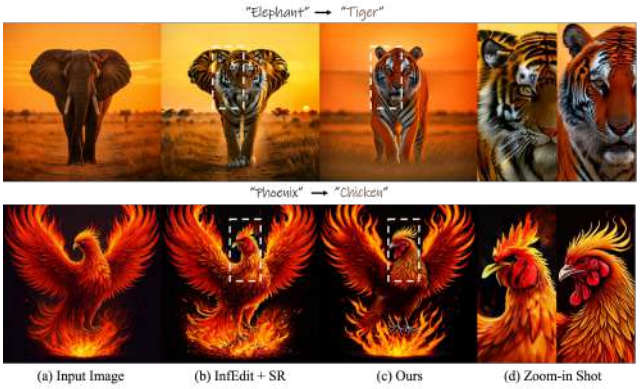


Figure 5. Comparison of 16× super-resolution upsampler [51] with InfEdit [55] with our approach based on the 16× SD 2.1.

Figure 6. Ablation study qualitative results on the 16× SD 2.1 .

Method	ImageReward ↑	HPSv2 ↑	CLIP Score ↑
Tiled Inv. + ScaleCrafter [18]	1.2595	0.2962	34.9431
Ours w/o NDCFG++	1.6273	0.2911	35.0254
Ours	**1.6689**	**0.3017**	**35.3194**

Table 2. Ablation study quantitative results on the 16× SD 2.1.

the superiority of our method, particularly in accurately editing the semantic target while preserving high-resolution detail. In both the "Tiger" and "Chicken" examples, our method produces coherent and realistic outputs with structural integrity and fine-grained texture details, most notably in the eye regions and furs. In contrast, InfEdit + StableSR produces distorted facial features, indicating that performing edits at low resolution followed by upsampling leads to a loss of fine-grained details. Comprehensive comparisons with this approach are provided in the **supplementary material**.

4.2.2. Quantitative Results

Tab. 1 represents a quantitative comparison between CSD [27] and EDITCRAFTER. As shown, EDITCRAFTER demonstrates superior performance over CSD on all resolution. The evaluations using ImageReward [54], HPSv2 [52], and CLIPScore [20] demonstrate that our editing results are highly aligned with the editing prompts and human preferences. On the other hand, CSD fails to modify the target object as specified by the text prompt, resulting in limited editing responsiveness to object-level changes as shown in the Fig. 4.

For the user study, we collected a total of 25 responses, including 5 vigilance tasks, from 112 participants. The results indicate that human evaluators preferred our EDITCRAFTER method in **72.61%** of cases compared to CSD. These findings suggest that EDITCRAFTER more effectively applies the requested edits, demonstrating stronger alignment with user expectations.

4.2.3. Ablation Study

To evaluate contribution of each component in our method, we conducted an ablation study on NDCFG++ within the 16× SD 2.1 setup. As demonstrated in the second row of Tab. 2, removing NDCFG++ ($\tau = 0$ in Alg. 2) results in performance degradation, with our method achieving the highest performance on text-to-image alignment. Without NDCFG++, performance slightly drops in human preference scores, and CLIP text-image matching score. As illustrated in Fig. 6, the contribution of NDCFG++ to qualitative enhancement is evident in its capability in accurately positioning the target object's head in the location of the original object within the source image, ensuring alignment with the object's identity.

On the one hand, the ablated version of our method, which excludes NDCFG++ and adopts the high CFG generation settings from ScaleCrafter [18], demonstrates diminished performance as shown in the first row of Tab. 2. This decline is evidenced by a substantial reduction in ImageReward scores, as well as decreases in other text-based image quality metrics. This generation setting overlooks the critical role of text guidance in image editing, adversely affecting both background integrity and the semantic consistency of the object. Our method, in contrast, effectively preserves the original background and object textures, including pattern of water droplets or color pattern, resulting in improved fidelity and alignment with the desired text-to-image prompts as shown in the Fig. 6 (d).

5. Conclusion

We present EDITCRAFTER, a tuning-free and optimization-free editing pipeline using pretrained diffusion models. By utilizing an straightforward but effective approach with tiled-inversion and NDCFG++, our method demonstrates consistently superior performance across high-resolution image editing methods. Notably, we conduct the first extensive quantitative and qualitative evaluation in high-resolution image editing, showcasing the effectiveness of our method's design. We anticipate that our proposed framework, EDITCRAFTER, can be effectively integrated into real-world applications, thereby enhancing the performance in practical settings.

Acknowledgments

Kunho Kim was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI).

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *ICML*, 2023. 2, 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 3, 7

[3] 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. In *ICCV*, 2023. 2, 3

[4] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2Video: Video Editing using Image Diffusion. In *ICCV*, 2023. 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.*, 2023. 2, 3

[6] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. In *ECCV*, 2024. 3

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *ICLR*, 2024. 3

[8] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. In *ICLR*, 2025. 6

[9] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023. 5

[10] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. TurboEdit: Text-Based Image Editing Using Few-Step Diffusion Models. In *ACM SIGGRAPH Asia 2024 Conference Proceedings*, 2024. 2

[11] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 5

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *ICML*, 2024. 2

[13] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing One-step Text-to-Image Models through Reward-based Noise Optimization. In *NeurIPS*, 2024. 2

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM TOG*, 2022. 12

[15] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. ReNoise: Real Image Inversion Through Iterative Noising. In *ECCV*, 2024. 2

[16] Jong Chul Ye Gihyun Kwon. CLIPstyler: Image Style Transfer with a Single Text Condition. In *CVPR*, 2022. 3

[17] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, and Dimitris Metaxas. ProxEdit: Improving Tuning-Free Real Image Editing with Proximal Guidance. In *WACV*, 2024. 2, 3, 5, 13

[18] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. In *ICLR*, 2024. 2, 3, 4, 5, 8, 12

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *ICLR*, 2023. 2, 3

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021. 7, 8, 12, 13

[21] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022. 2, 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 3, 4, 5

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J.

Fleet. Video Diffusion Models. In *NeurIPS*, 2022. 3

[24] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis. In *ECCV*, 2024. 3

[25] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. In *CVPR*, 2024. 2, 3

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2017. 3

[27] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative Score Distillation for Consistent Visual Synthesis. In *NeurIPS*, 2023. 2, 3, 5, 7, 8, 13, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29

[28] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. SALAD: Part-Level Latent Diffusion for 3D Shape Generation and Manipulation. In *ICCV*, 2023. 3

[29] Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. 2

[30] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, 2025. 2

[31] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions. In *NeurIPS*, 2023. 2, 3

[32] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 3

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-Text Inversion for Editing Real Images Using Guided Diffusion Models. In *CVPR*, 2023. 2, 3, 5, 13

[34] Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. SwiftEdit: Lightning Fast Text-Guided Image Editing via One-Step Diffusion. In *ICCV*, 2025. 2, 3

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 2022. 3

[36] Dani Lischinski Omri Avrahami, Ohad Fried. Blended Latent Diffusion. *ACM TOG*, 2023. 3

[37] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot Image-to-Image Translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3, 5

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 6, 12

[39] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*, 2023. 3

[40] Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. FreeScale: Unleashing the Resolution of Diffusion Models via Tuning-Free Scale Fusion. In *ICCV*, 2025. 3

[41] Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas. In *ECCV*, 2024. 2, 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 4

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[44] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. UltraPixel: Advancing Ultra-High-Resolution Image Synthesis to New Peaks. In *NeurIPS*, 2024. 2, 3, 6

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 2, 3, 6, 12

[46] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In *ICLR*, 2025. 2

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara

Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*, 2022. 2, 3

[48] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3D Neural Field Generation using Triplane Diffusion. In *CVPR*, 2023. 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 3, 5

[50] Nikita Starodubcev, Mikhail Khoroshikh, Artem Babenko, and Dmitry Baranchuk. Invertible Consistency Distillation for Text-Guided Image Editing in Around 7 Steps. In *NeurIPS*, 2024. 2

[51] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 7, 8, 13

[52] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 7, 8

[53] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers. In *ICLR*, 2025. 3

[54] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *NeurIPS*, 2023. 7, 8, 13

[55] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-Free Image Editing with Natural Language. In *CVPR*, 2023. 7, 8, 13

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 3

A. Implementation Details

We provide additional implementation details of Alg. 2. To highlight the distinguishing factors between ScaleCrafter [18] and our proposed method, we present both reverse processes. The DDIM sampling steps are configured to $T = 50$. For $\times 4$ editing, we set $\tau = 10$ and for both $\times 8$ and $\times 16$ editing, $\tau = 37$. These settings are applied to both SD 2.1 [45] and SDXL 1.0 [38]. We follow the re-dilated convolution configurations for each resolution as implemented in ScaleCrafter. For both CLIPScore [20] and CLIP Image Similarity [14], we employ the "ViT-B/32" model as the foundational architecture.

Algorithm 3 Reverse Diffusion with ScaleCrafter

Require: $z_T \sim \mathcal{N}(0, \mathbf{I}_d)$, $0 \leq \omega \in \mathbb{R}$, $\tau \leq T \in \mathbb{R}$

for $i = T$ **to** 1 **do**
  if $i \leq \tau$ **then**
    $\tilde{\epsilon}_c^{\omega}(z_t) = \epsilon_{\varnothing}(z_t) + \omega[\tilde{\epsilon}_c(z_t) - \tilde{\epsilon}_{\varnothing}(z_t)]$
  else
    $\tilde{\epsilon}_c^{\omega}(z_t) = \tilde{\epsilon}_{\varnothing}(z_t) + \omega[\tilde{\epsilon}_c(z_t) - \tilde{\epsilon}_{\varnothing}(z_t)]$
  end if
  $\tilde{z}_c^{\omega}(z_t) \leftarrow (z_t - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\omega}(z_t)) \,/\, \sqrt{\alpha}_t$
  $z_{t-1} = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\omega}(z_t) + \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_c^{\omega}(z_t)$
end for
$x_0 = \mathcal{D}(z_0)$ ▷ Decode latent
return x_0

Algorithm 4 Reverse Diffusion with Ours

Require: Inverted latent z_T^*, $\lambda \in [0, 1]$, $\tau \leq T \in \mathbb{R}$

for $i = T$ **to** 1 **do**
  if $i \leq \tau$ **then** ▷ NDCFG++
    $\tilde{\epsilon}_c^{\lambda}(z_t^*) = \epsilon_{\varnothing}(z_t^*) + \lambda[\tilde{\epsilon}_c(z_t^*) - \tilde{\epsilon}_{\varnothing}(z_t^*)]$
    $\tilde{z}_c^{\lambda}(z_t^*) \leftarrow (z_t^* - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\lambda}(z_t^*)) \,/\, \sqrt{\alpha}_t$
    $z_{t-1}^* = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\lambda}(z_t^*) + \sqrt{1-\alpha_{t-1}}\epsilon_{\varnothing}(z_t^*)$
  else ▷ Vanilla CFG++
    $\tilde{\epsilon}_c^{\lambda}(z_t^*) = \tilde{\epsilon}_{\varnothing}(z_t^*) + \lambda[\tilde{\epsilon}_c(z_t^*) - \tilde{\epsilon}_{\varnothing}(z_t^*)]$
    $\tilde{z}_c^{\lambda}(z_t^*) \leftarrow (z_t^* - \sqrt{1-\alpha_t}\tilde{\epsilon}_c^{\lambda}(z_t^*)) \,/\, \sqrt{\alpha_t}$
    $z_{t-1}^* = \sqrt{\alpha_{t-1}}\tilde{z}_c^{\lambda}(z_t^*) + \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_{\varnothing}(z_t^*)$
  end if
end for
$x_0 = \mathcal{D}(z_0^*)$ ▷ Decode latent
return x_0

B. Effect of Classfier-Guidance Scale

We investigate the effect of small guidance scale $\lambda \in [0, 1]$ in our sampling process. We examine the impact of varying the small guidance scale parameter, λ, within the range [0, 1] on our sampling process. As depicted in Fig. A7, the reconstruction produced with $\lambda = 0$ does not exactly replicate the original image; however, it serves as a promising initial foundation for subsequent editing tasks. Notably, as λ increases, the edited images progressively conform more closely to the specified edit prompt "wolf". This tendency is also reflected when measure the metric. Fig. A8 illustrates that increasing the guidance scale λ leads to higher values in edited image-text alignment metrics, while simultaneously reducing the preservation of the original image as measured by CLIP Image Similarity [14].

This behavior indicates that higher guidance scales enhance the alignment between the generated modifications, thereby facilitating more precise and controlled image editing. Based on our observations, we set the guidance scale parameter $\lambda = 0.5$ to achieve an optimal balance between adhering to the editing prompt and preserving the original identity for all experiments. However, users may adjust this setting to better suit real-world editing applications.

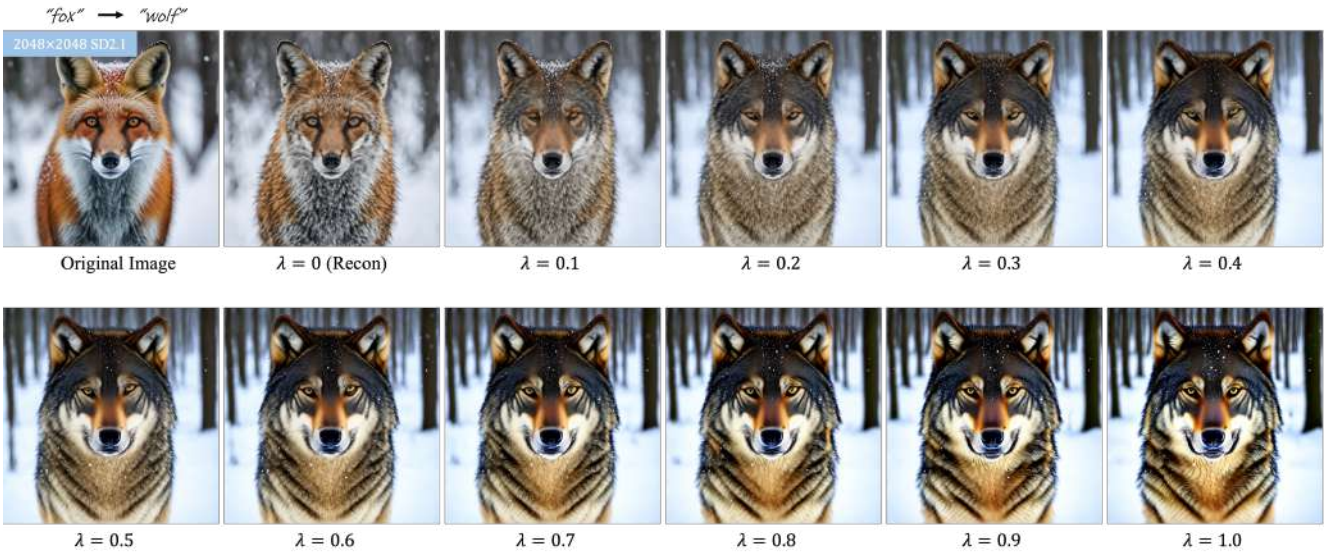


Figure A7. The effect of CFG scale.

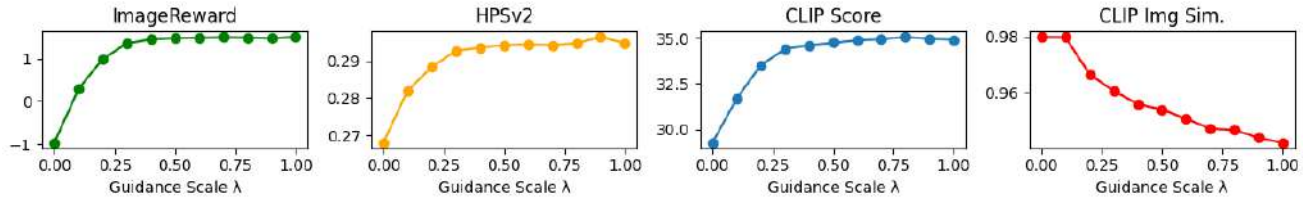


Figure A8. The effect of CFG scale λ in 4 $\times$ SD 2.1.

C. User Study

In Sec. 4 of the main paper, we reported the preference statistics collected from 112 user study participants who passed the vigilance tests from Amazon MTurk. We provide additional details of the user study in the following. We instructed participants to select the most anticipated outcome when the displayed source image is edited by the text prompt with the question used in [33]: `Which one better applies the requested edit to the input image on top, while preserving most of the details from the input image?` The example of user study screen is shown in Fig. A9.

D. Quantitative Evaluation of Low-Resolution Editing Combined with Super-Resolution

To the best of our knowledge, apart from CSD [27], no existing work directly addresses high-resolution image editing. However, for a comprehensive evaluation, we present quantitative comparisons in Tab. A4 on our dataset against ProxEdit [17]+StableSR [51] and InfEdit [55]+StableSR [51], which are currently state-of-the-art image editing methods. Our method, EDITCRAFTER, achieves the highest scores in both the ImageReward [54] and CLIPScore [20] metrics. Furthermore, although InfEdit + StableSR attains high HPSv2 scores, it is unable to capture intricate details because resizing disrupts high-level information and subsequent super-resolution fails to recover these details, as demonstrated in Fig. A10.

Furthermore, we conducted two user studies to compare our method against InfEdit + StableSR and ProxEdit + StableSR, respectively, using Amazon MTurk, following the same setup described in Sec. C. We collected a total of 25 responses, including 5 vigilance tasks, from 124 participants for the comparison between our method and InfEdit + StableSR, and from 117 participants for the comparison with ProxEdit + StableSR. The results demonstrate that human evaluators preferred our EDITCRAFTER method in **61.12%**, and **92.38%** of cases when compared to InfEdit + StableSR, and ProxEdit + StableSR, respectively. These results demonstrate that EDITCRAFTER more effectively applies the intended edits, achieving better alignment with user expectations compared to low-resolution editing combined with super-resolution.

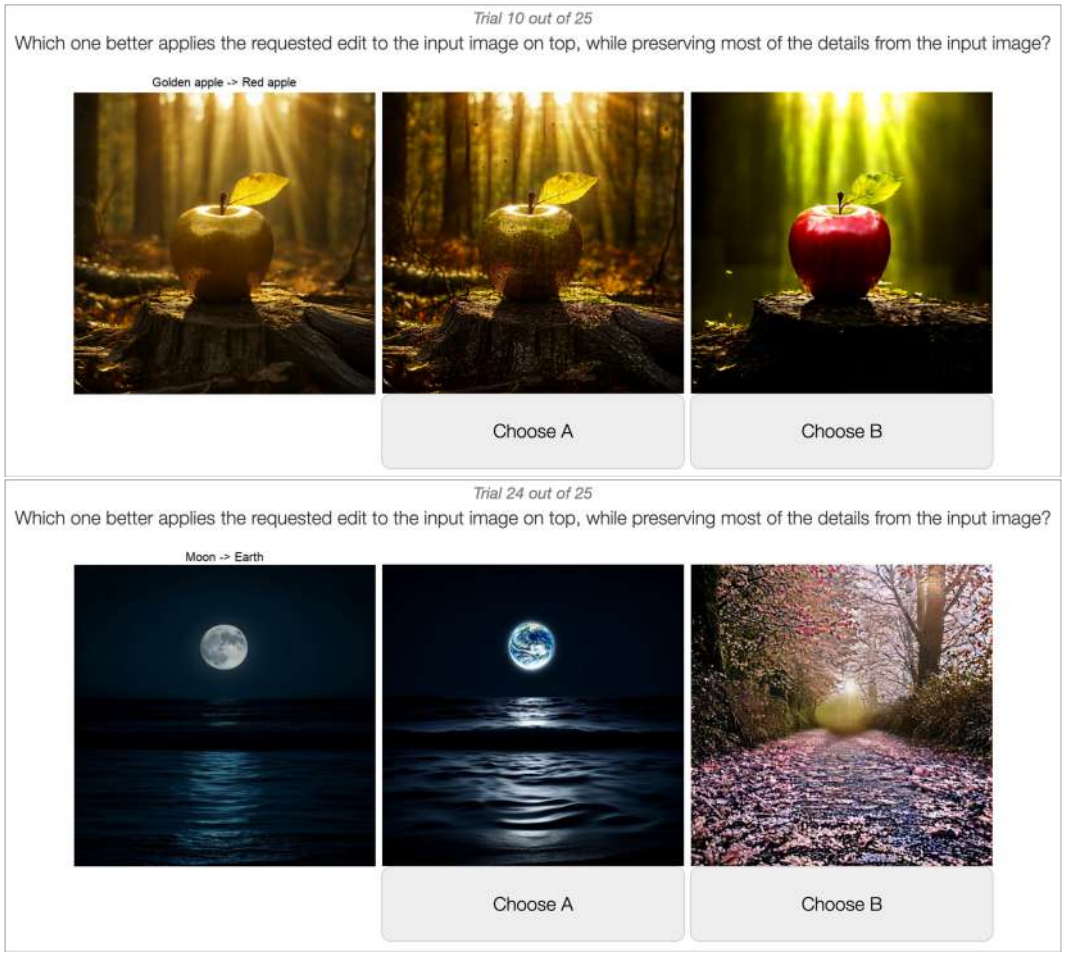


Figure A9. Screen captures of user study. The top example illustrates a main question from the user study, while the bottom example represents a vigilance question.

CSD	Ours	InfEdit + StableSR	Ours	ProxEdit + StableSR	Ours
27.39%	**72.61%**	38.88%	**61.12%**	6.62%	**92.38%**

Table A3. User study results. Participants were instructed to select the most preferred editing outcome based on its fidelity to both the original input image and the given textual edit instruction.

Res	Method	ImageReward ↑	HPSv2 ↑	CLIPScore ↑
4× 1:1	CSD	0.5538	0.2883	32.8353
	InfEdit+SR	1.2212	**0.2982**	33.6893
	ProxEdit+SR	-0.5561	0.2833	30.3980
	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
	InfEdit+SR	1.6670	**0.3021**	35.1438
	ProxEdit+SR	0.5440	0.2873	32.6323
	Ours	**1.6689**	0.3017	**35.3194**

Table A4. Quantitative comparisons on SD2.1.

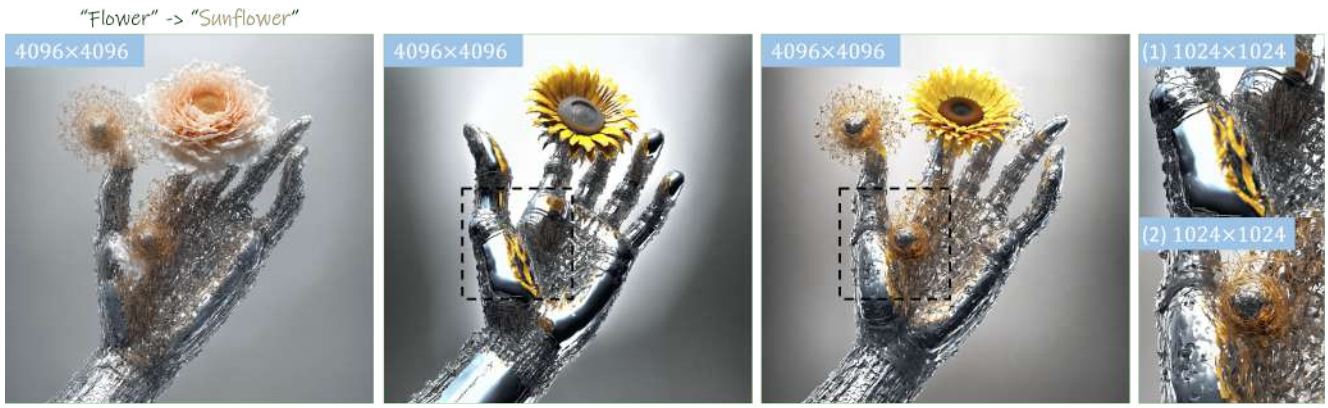
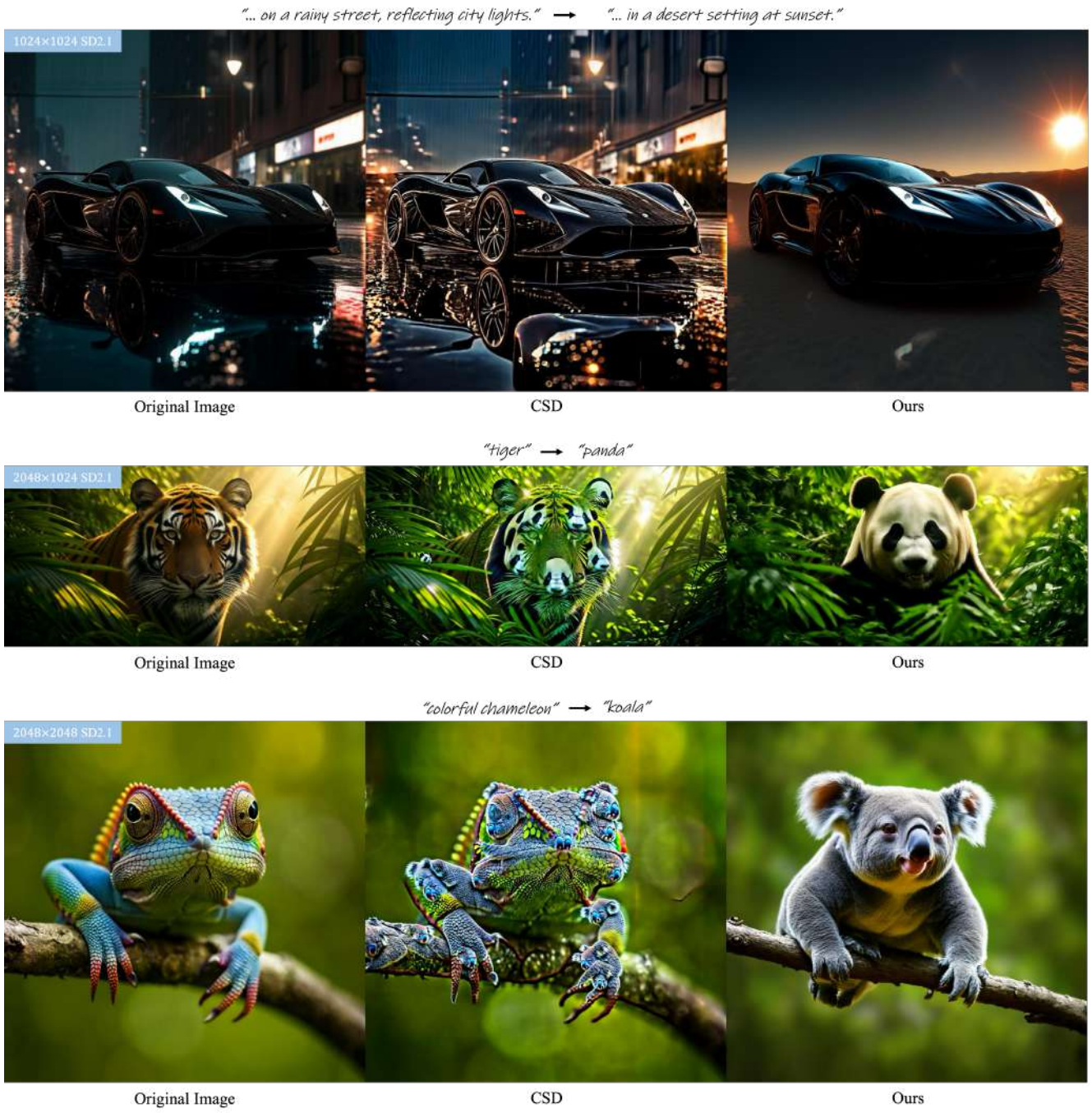


Figure A10. InfEdit+StableSR vs. EditCrafter for the teaser image.

E. High Quality Version of Fig. 4



"dandelion seeds" → "balloon"

Original Image | CSD | Ours

"cherry blossom" → "maple"

Original Image

CSD [27]

EDIT
-CRAFTER

"forest" → "burning forest"

Original Image

CSD [27]

EDIT
-CRAFTER

F. More Qualitative Comparisons

Original Image	CSD [27]	EDITCRAFTER (Ours)

SD2.1 $\times$**4** *"moon"* $\rightarrow$ *"earth"*

SD2.1 $\times$**4** *"blanket"* $\rightarrow$ *"grass"*

SD2.1 $\times$**4** *"cactus"* $\rightarrow$ *"aloe"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SD2.1 $\times$**4** *"lemon"* $\rightarrow$ *"cucumber"*

SD2.1 $\times$**4** *"tulips"* $\rightarrow$ *"roses"*

SD2.1 $\times$**4** *"vilage"* $\rightarrow$ *"castle"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SD2.1 $\times$**8** *"vilage"* $\rightarrow$ *"castle"*

SD2.1 $\times$**8** *"fox"* $\rightarrow$ *"lion"*

SD2.1 $\times$**8** *"owl"* $\rightarrow$ *"hawk"*

SD2.1 $\times$**8** *"parm tree"* $\rightarrow$ *"umbrella"*

SD2.1 $\times$**8** *"shark"* $\rightarrow$ *"dolphin"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SD2.1 $\times$**16** *"berrys"* $\rightarrow$ *"roses"*

SD2.1 $\times$**16** *"cat"* $\rightarrow$ *"goat"*

SD2.1 $\times$**16** *"soccer ball"* $\rightarrow$ *"crystal ball"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SDXL $\times$**4** *"asphalt"* $\rightarrow$ *"dessert"*

SDXL $\times$**4** *"gems"* $\rightarrow$ *"bones"*

SDXL $\times$**4** *"phoenix"* $\rightarrow$ *"chicken"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SDXL $\times$**8** *"cloud"* $\rightarrow$ *"mushroom"*

SDXL $\times$**8** *"lion"* $\rightarrow$ *"tiger"*

SDXL $\times$**8** *"shell"* $\rightarrow$ *"crab"*

SDXL $\times$**8** *"snow globe"* $\rightarrow$ *"jungle globe"*

SDXL $\times$**8** *"whale"* $\rightarrow$ *"turtle"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SDXL $\times$**16** *"apple"* $\rightarrow$ *"pink peach"*

SDXL $\times$**16** *"bee"* $\rightarrow$ *"hummingbird"*

SDXL $\times$**16** *"bird"* $\rightarrow$ *"owl"*

Original Image | CSD [27] | EDITCRAFTER (Ours)

SDXL $\times$**16** *"mountain"* $\rightarrow$ *"sand dune"*

SDXL $\times$**16** *"stone"* $\rightarrow$ *"Stonehenge"*

SDXL $\times$**16** *"waterfall"* $\rightarrow$ *"lava flow"*