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

A. Implementation Details

We provide additional implementation details of Alg. 2 in the main paper. To highlight the distinguishing factors between ScaleCrafter [3] 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 [8] and SDXL 1.0 [7]. We follow the re-dilated convolution configurations for each resolution as implemented in ScaleCrafter. For both CLIPScore [4] and CLIP Image Similarity [1], we employ the "ViT-B/32" model as the foundational architecture.

Algorithm 1 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 2 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. 1, 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. 2 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 [1].

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.

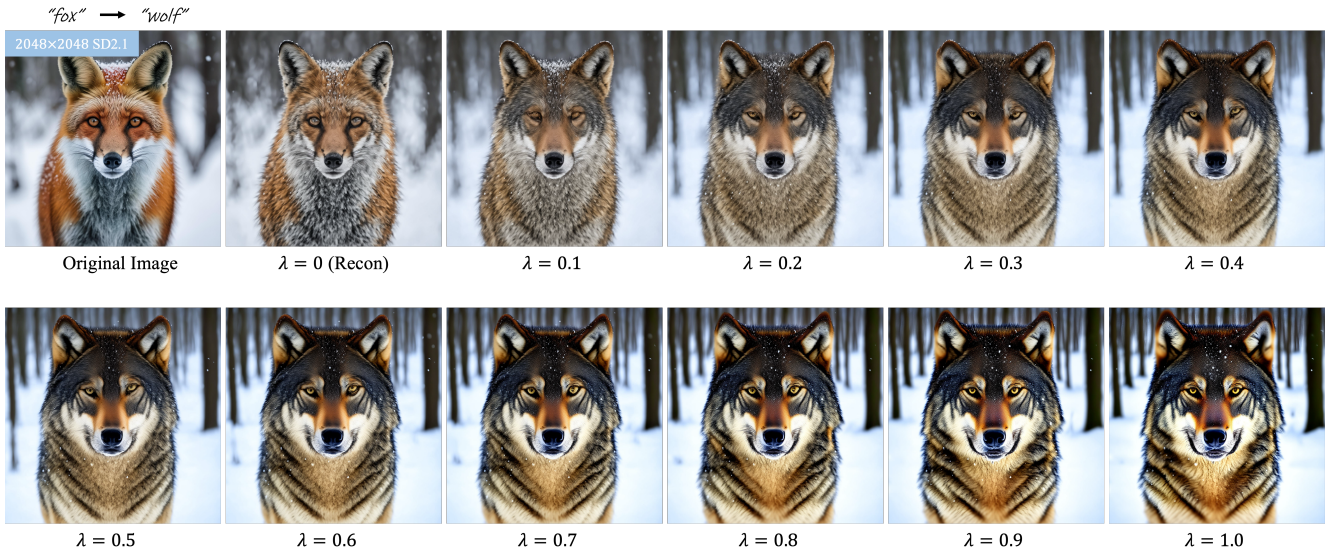


Figure 1. The effect of CFG scale.

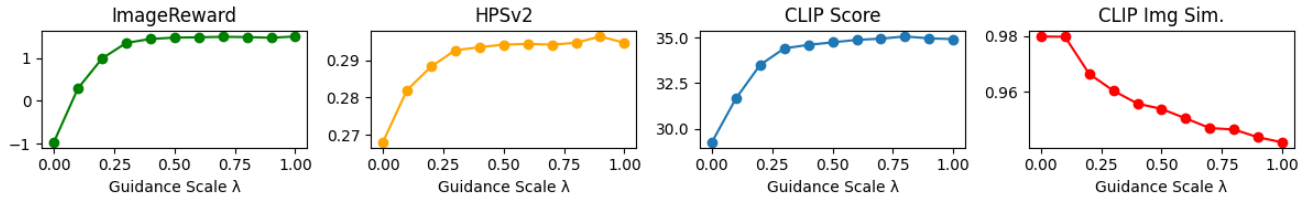


Figure 2. The effect of CFG scale λ in 4 $\times$ SD 2.1.

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 [6]: `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. 3.

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

To the best of our knowledge, apart from CSD [5], no existing work directly addresses high-resolution image editing. However, for a comprehensive evaluation, we present quantitative comparisons in Tab. 2 on our dataset against ProxEdit [2]+StableSR [9] and InfEdit [11]+StableSR [9], which are currently state-of-the-art image editing methods. Our method, EDITCRAFTER, achieves the highest scores in both the ImageReward [10] and CLIPScore [4] 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. 4.

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.

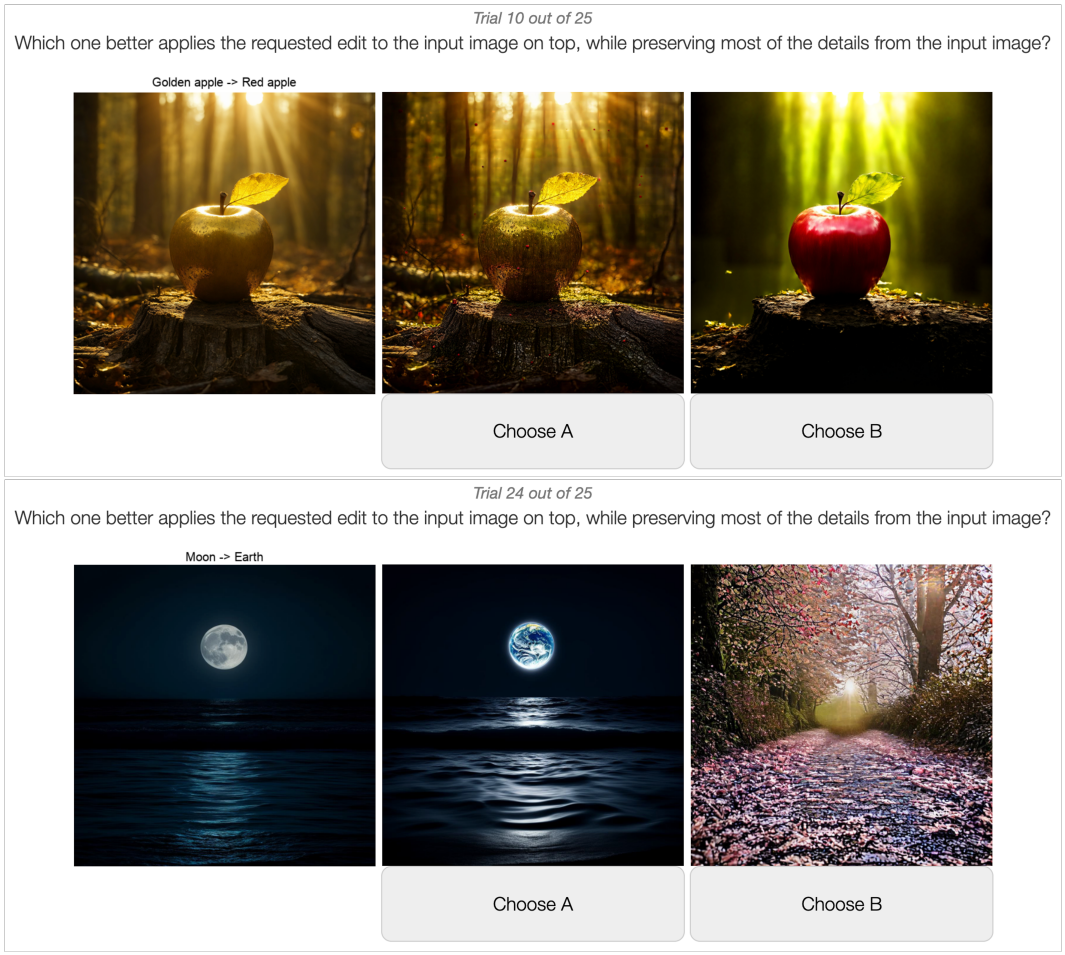


Figure 3. Screen captures of user study. The top example illustrates a main question from the user study, while the bottom example represents a vigilance question.

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

Table 1. 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 2. Quantitative comparisons on SD2.1.

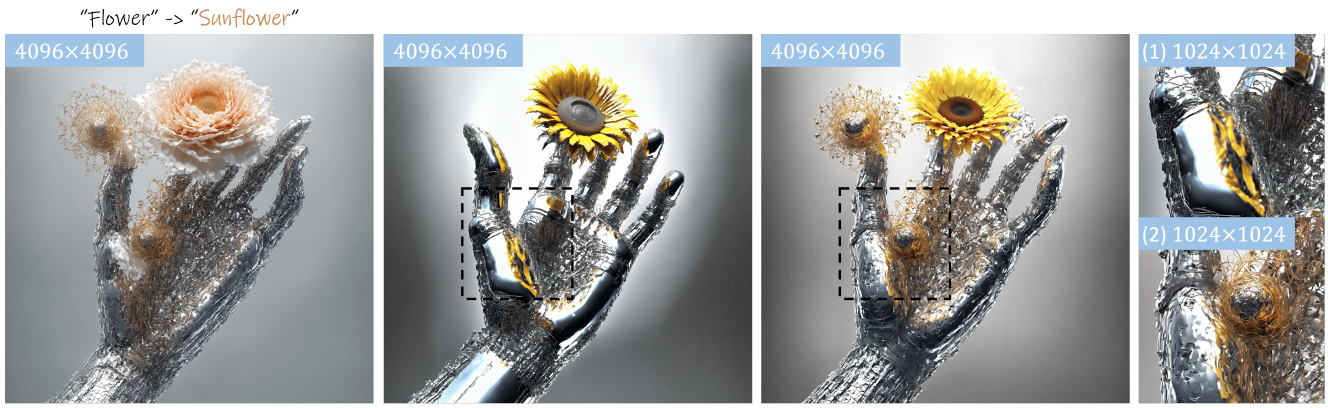
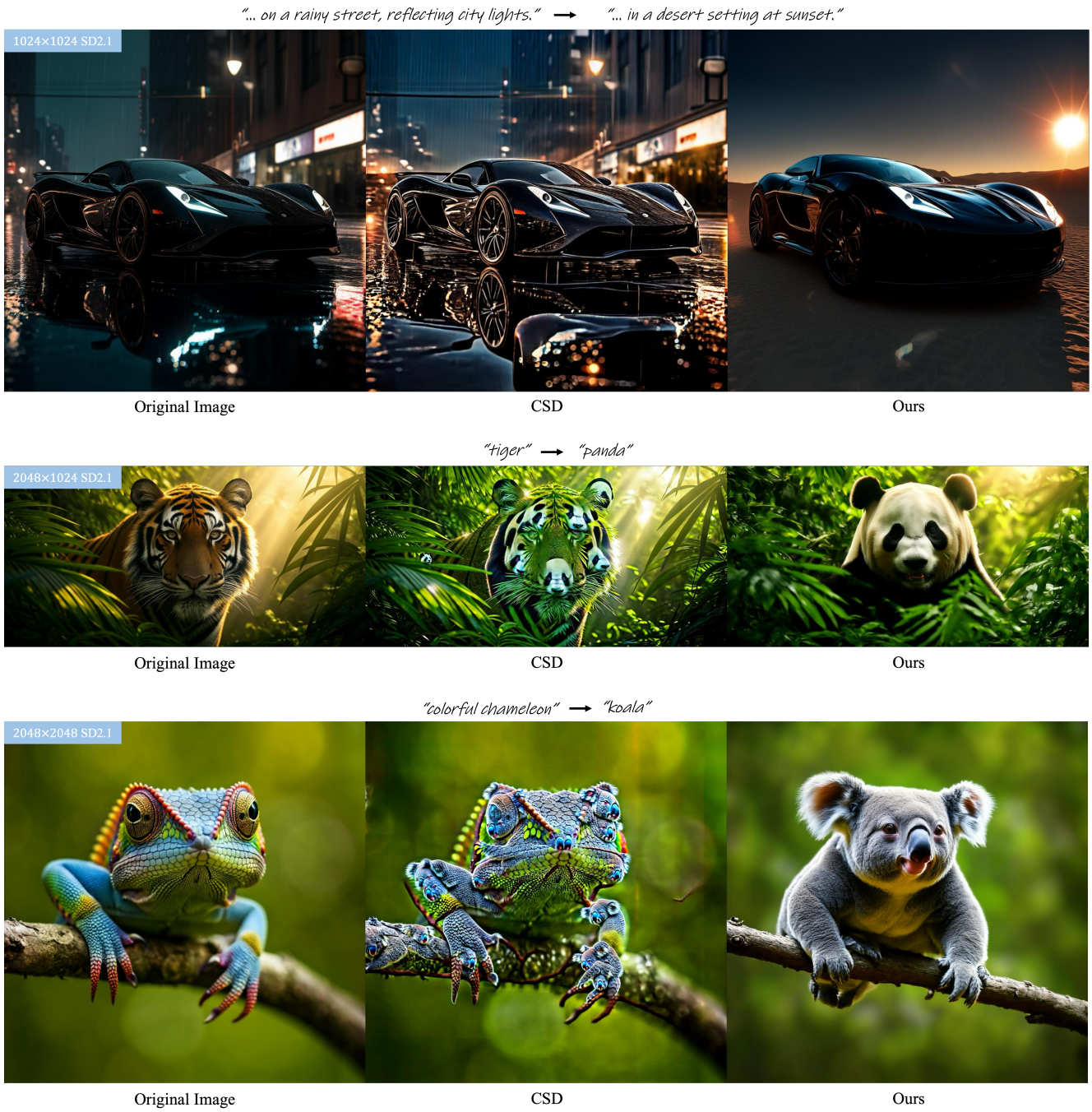


Figure 4. InfEdit+StableSR vs. EditCrafter for the teaser image.

E. High Quality Version of Fig. 3



Original Image | CSD | Ours

Original Image

CSD [5]

EDIT
-CRAFTER

"forest" → "burning forest"

Original Image

CSD [5]

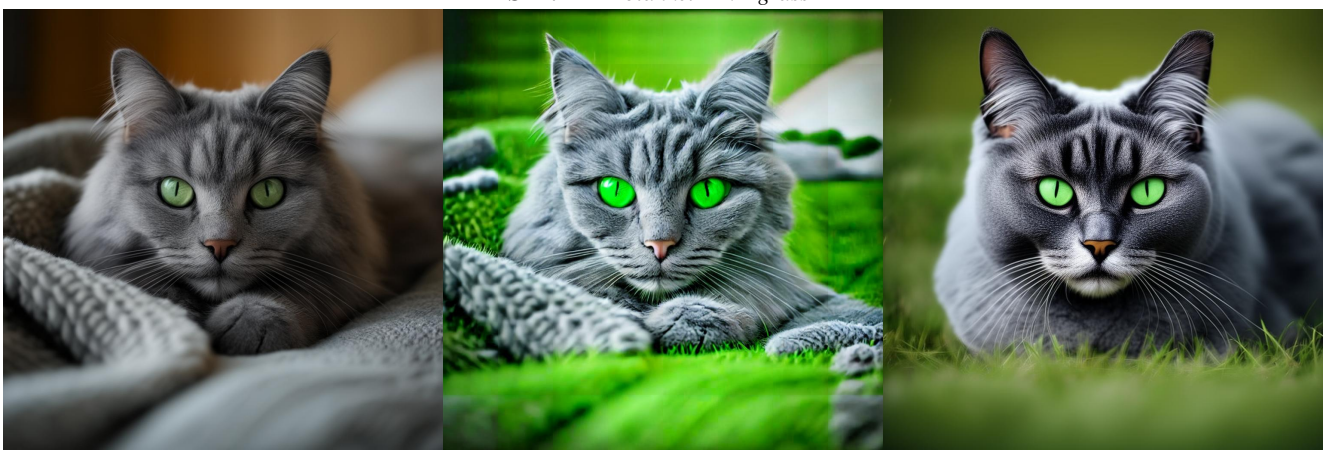
EDIT
-CRAFTER

F. More Qualitative Comparisons

Original Image	CSD [5]	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 [5] | 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 [5]	EDITCRAFTER (Ours)

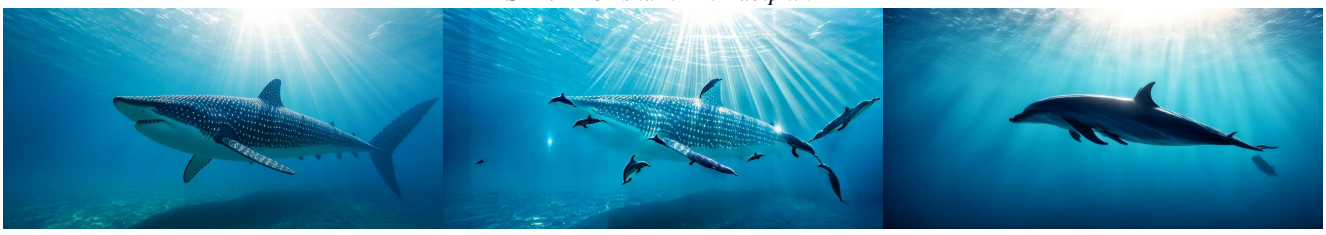
SD2.1 $\times$**8** *"vilage"* $\rightarrow$ *"castle"*

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 [5] | 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 [5] | EDITCRAFTER (Ours)

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

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

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

Original Image | CSD [5] | EDITCRAFTER (Ours)

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

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

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

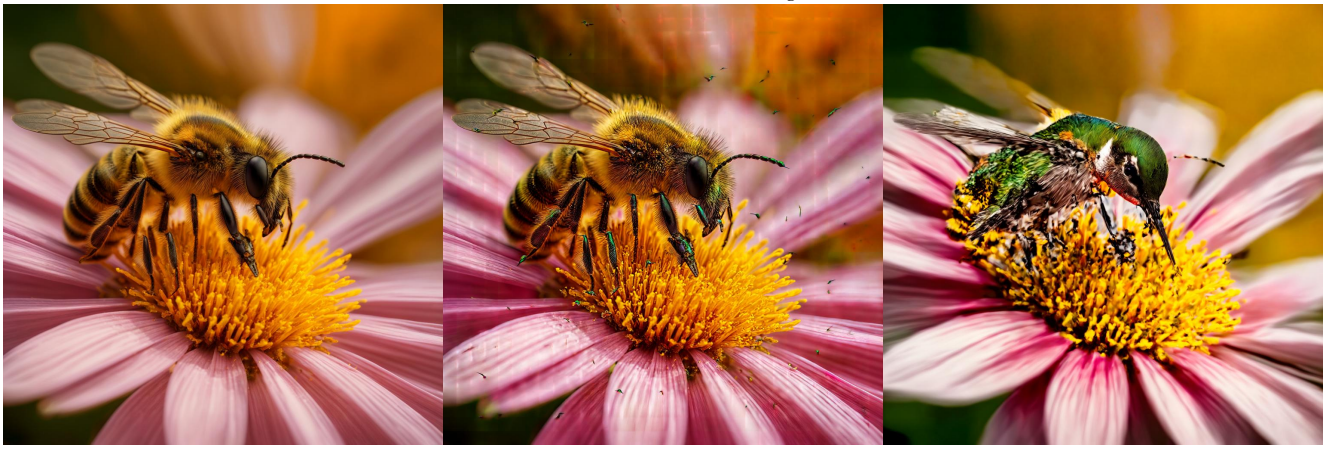
SDXL $\times$**8** *"snow globe"* $\rightarrow$ *"jungle globe"*

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

Original Image | CSD [5] | EDITCRAFTER (Ours)

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

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



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

Original Image | CSD [5] | EDITCRAFTER (Ours)

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

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

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

References

[1] 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. 1

[2] 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] 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. 1

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021. 1, 2

[5] 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, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18

[6] 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

[7] 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. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1

[9] 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. 2

[10] 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. 2

[11] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-Free Image Editing with Natural Language. In *CVPR*, 2023. 2