HierEdit : Region-Aware Hierarchical Diffusion for Efficient High-Resolution Editing

Supplementary Material

1. Comparison with Upscaler Baselines.

Here we present a qualitative (Figure 1) example where the lady's facial expression and posture totally changed. Both upscalers show artifacts (eyelash discontinuity and blurriness), while ours does not. We will add complete quantitative comparisons in the final version.

2. Example on Full Editing

Figure 2 further illustrates our ability to edit entire images, even though our focus is on local editing.

3. More Ablation Study

Ablation on Denoising Step Reduction. We further examine how aggressively denoising steps may be skipped while preserving output quality. Figure 3 presents results obtained by skipping $\{0, 10, 14, 18, 22, 24\}$ out of 28 total steps. Skipping 18 steps offers the best balance between visual quality and efficiency.

4. Implementation Details

4.1. Experimental Setup.

We adopt FLUX.1-dev [1] as our base model. The network is fine-tuned using LoRA with a rank of 16, trained for 10,000 steps with a batch size of 6 on six NVIDIA RTX A6000 Ada GPUs (48 GB each). All evaluations are performed on a NVIDIA RTX 6000 Ada Pro GPU (96 GB) server. For training, we employ the IPA300K dataset [8], which provides paired source and target images along with textual prompts and bounding-box annotations on 1024×1024 resolution. For the off the shelf low-resolution model selection, we choose either FLUX-Kontext.dev or OminiControl2. For kernel adaptation, we permute the token sequence to a "window-first" way, so that the tokens within one local window will stick to each other in the sequence and we also adapt our code to Flash Sparse Attention [5], to skip the masked blocks for speeding up

4.2. Instructional Editing Benchmarks

We provide additional details regarding the quanitative comparisons in Table 1 in the main paper. We evaluate on I2EBench [3], ImgEdit [6], CompBench [2], and EmuEdit [4]. For the composite benchmarks (those measuring multiple subtasks), we average the scores of the local editing tasks. The details for each benchmark are listed below.

I2EBench. The I2EBench benchmark, which encompasses 16 diverse image editing tasks spanning both low-level restoration and high-level semantic modifications. We evaluate on the the first category includes 9 low-level tasks: Deblurring, HazeRemoval, Lowlight, NoiseRemoval, RainRemoval, ShadowRemoval, SnowRemoval, WatermarkRemoval, and RegionAccuracy. We evaluate using the Structural Similarity Index (SSIM), comparing edited images against ground truth references.

CompBench. We evaluate on the local subset of the CompBench benchmark, which includes add, remove, and replace tasks. We report the average scores across those three tasks. For each task, we measure text-image alignment through the CLIP Score and structural similarity in background regions using SSIM.

EmuEdit. For the EmuEdit benchmark, we utilize the test split of the facebook/emu_edit_test_set dataset from HuggingFace, and report two metrics. The CLIP_{dir} assesses directional alignment of the edit. The DINO score measures feature-level preservation.

ImgEdit. We evaluate our method on the ImgEdit benchmark using its basic suite, which consists of single-turn editing tasks. The primary evaluation relies on a GPT-4 judge. The judge model receives three inputs (the original image, the edited image, and the textual instruction) and produces three integer scores ranging from 1 to 5 for different aspects of the edit. These scores are averaged. We report these averages across a range of local editing tasks: add, adjust, compose, extract, remove, and replace.

5. More Examples

We present additional qualitative examples in Figures 5. The original images are upsampled using the BSRGAN super-resolution model [7]. For each example, the low-resolution edited result is displayed inside the black box located at the lower-left corner. We also include 4K (Figure 6) and 2K (Figure 7) editing results on synthetic data.

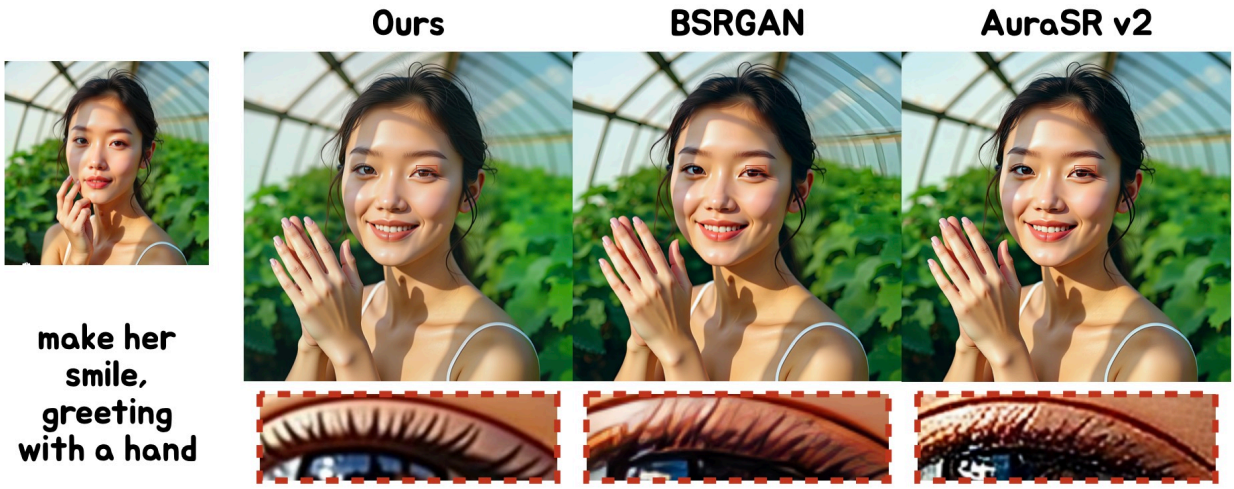

Figure 1. Comparison with low-resolution editing followed by upscaling.

Figure 2. Our model also supports full image editing by removing the mask although we focus on local editing.

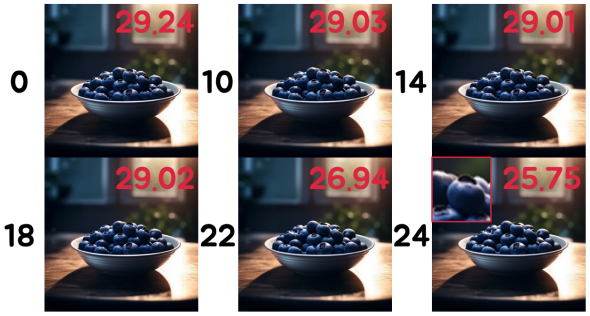

Figure 3. Ablation on the number of timesteps skipped. Results shows 18 steps yields an optimal solution balancing the generation quality and speed. Numbers in red are PSNR values.

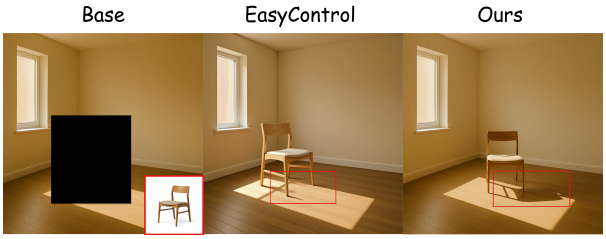

Figure 4. The inappropriate bounding box will lead to issues in correct shadowing or other artifacts, therefore we need to refine the bounding box.

References

[1] Black Forest Labs. Flux. Black Forest Labs GitHub repository, 2024. Model release. 1

[2] Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, Lei Bai, Wanli Ouyang, Lin Chen, Fei Zhao, Zihan Wang, Yuan Xie, and Shaohui Lin. Compbench: Benchmarking complex instruction-guided image editing. *arXiv preprint arXiv:2505.12200*, 2025. 1

[3] Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 41494–41516. Curran Associates, Inc., 2024. 1

[4] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 1

[5] Jingze Shi, Yifan Wu, Bingheng Wu, Yiran Peng, Liangdong Wang, Guang Liu, and Yuyu Luo. Trainable dynamic mask sparse attention. *arXiv preprint arXiv:2508.02124*, 2025. 1

[6] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2025. 1

[7] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1

Figure 5. More examples on 4K in-the-wild data.

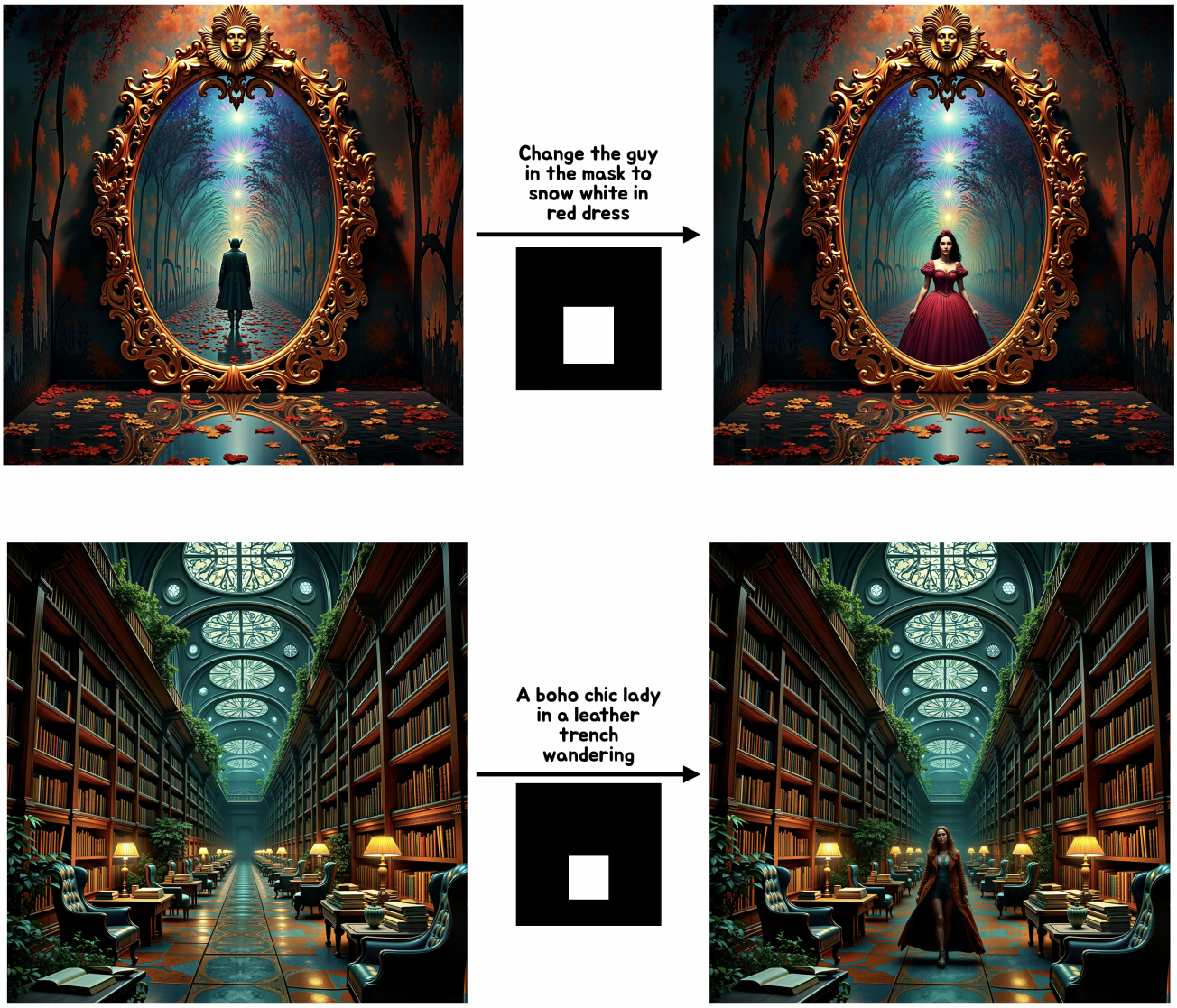

Figure 6. More examples on 4K synthetic data.

[8] Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1

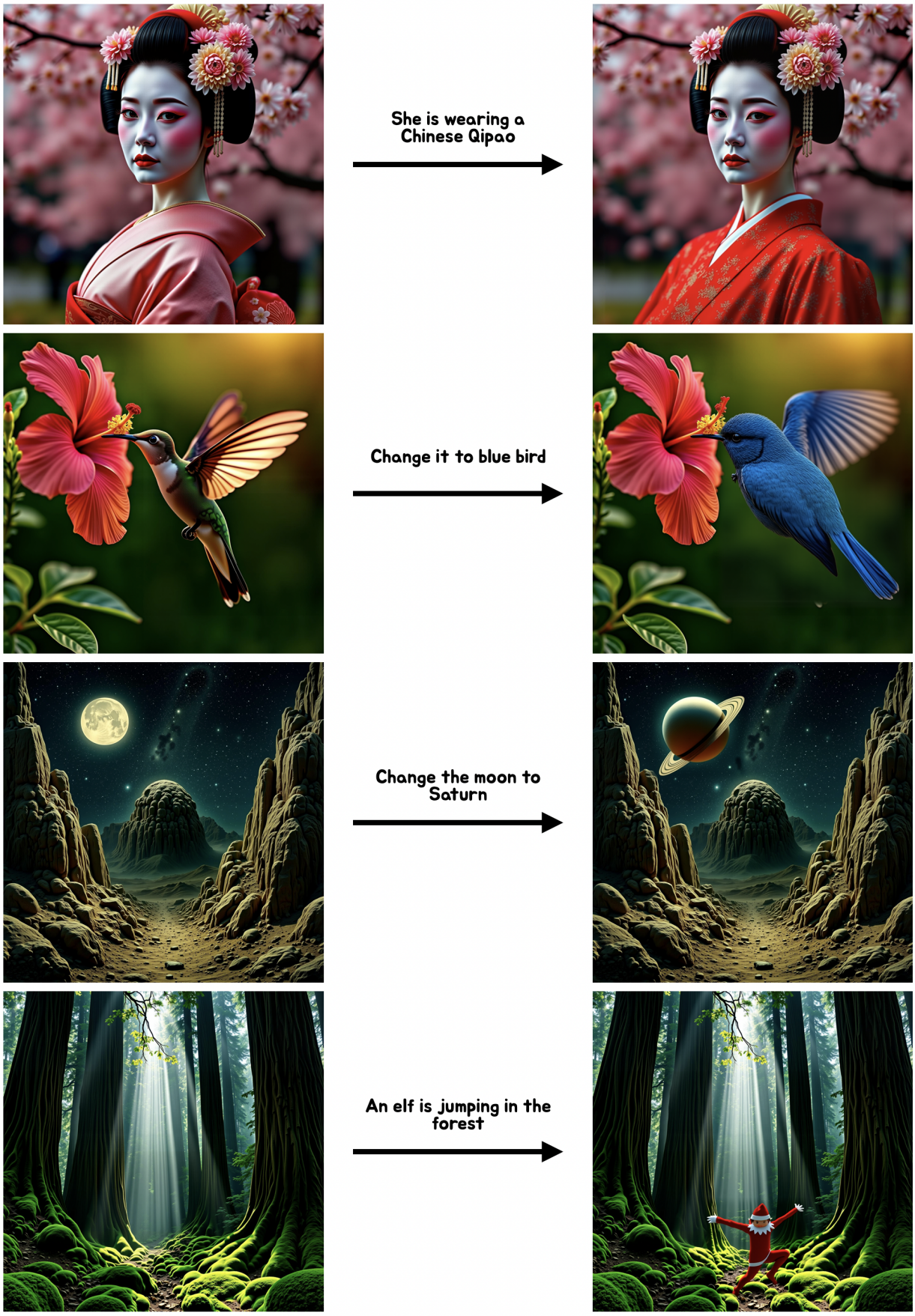

Figure 7. More examples on 2K synthetic data.