FuLLaMa: Training-free Diffusion-based Object Removal with Context Preservation Supplementary Material

İlke Demir
Cauth AI
ilke@cauth.ai

Umur Aybars Çiftçi
Binghamton University
uciftci@binghamton.edu

Figure 1. Comparison on COCO. Our method is compared against 6 SOTA object removal approaches on general object removal (milkshake), large area removal (the whole table), and multiple instance removal (people). FuLLaMa creates artifact-free, realistic, and consistent images without additional objects, preserving the context of the remaining area.

A. Additional Comparison

Fig. 1 samples our removal results on COCO, compared to six SOTA methods, visually confirming the results in Tab. 3 of the main paper. Rows 1, 2, and 3 are generated with verbal descriptions "milkshake", "everything on the table", and "people"; whereas the rest use existing masks. FuLLaMa can *clean* realistically, observing the reflections on the table (row 2), the completed bar (row 4), and the cupcake wrapper (row 5). In contrast, painters inpaint inconsistently (ClipAway contrast), may add NSFW content (PowerPaint), or fail to remove at all (SmartEraser cupcake and glass). Erasers perform better when deleted objects are smaller. When the area gets larger, repetitive artifacts (AttentiveEraser hands) and blurring (LaMa cupcake and ZITS++ vase) start to occur. Similar to the quantitative balance reported in Tab. 3 and the image quality superiority reported in Tab. 1, our visual results validate how FuLLaMa bridges the gap between erasers and painters.

B. Implementation Details

For the optimization, the $\{l, u\} \in \mathbb{R}^n$ is set as $\{\omega, \phi\} \in [0, 1]$, as normalized from $0.4 \leq \omega \leq 0.9$ and $1.5 \leq \phi \leq 35.0$ to provide flexibility to the DM. The initial guesses are set as $\omega_0 = 0.5, \phi_0 = 0.5$ in the normalized scale. The trust region is defined by updates to algorithm parameters η_1, η_2 which are initialized by the trust region radii $\rho_{init} = 0.5$ and $\rho_{end} = 0.01$. For faster convergence, we suggest $\rho_{init} = 0.2$ and $\rho_{end} = 0.1$ which flexes expanding the trust region. We set $k_{max} = 30$ for all results in the paper, although earlier convergence is observed. We implement our pipeline using pyiqa [4], pybobyqa [3], and diffusers libraries [19] for NR-IQA, optimization, and diffusion backbones.

C. LVLM Evaluation Details

We picked 22×7 random samples from COCO results generated by FuLLaMa, LaMa [17], AttentiveEraser [16], SmartEraser [8], PowerPaint [24], ClipAway [6], and ZITS++ [2]. Using the prompt below on 7+1 (original) image sets, we let Claude Sonnet report which result is the winner. We use the same prompt for all sets. We report the preference percentages in Tab. 4 of the main paper, where FuLLaMa result is selected as the best for 72.72% of the queries. If the answer is reported as a tie in rare cases, we include both approaches as winners, hence total of Tab. 4 overflows 100%.

```
You are a world renowned photographer,
invited to judge in a generative AI
contest.  The task is removing an object or
some area from the original photo (always
the first photo).  Investigate each 7
image carefully, capture light reflections,
physical realism, context preservation,
image quality, and artistic expression.
Which image would you prefer from both
technical and artistic standpoints and why?
```

The question "what should have been there?" has ambiguous answers which creates the need to evaluate with several metrics, capturing different requirements. To understand how LVLM evaluate these images, we interpret its preference by sampling its reasonings below. Note that it outputs 4-10 reasons per query from technical and artistic perspectives, so we select the interesting ones that inspire us to think beyond our traditional removal metrics.

- Creates better visual flow and reduces clutter
- Maintains perfect spatial relationships and natural lighting
- Tonal gradations are natural and the result feels like an intentional artistic choice
- The overall mood and atmosphere feel completely undisturbed
- Maintains the geometric integrity of the original container
- More natural-looking foreground-to-background transition
- Creates a perfect minimalist geometric composition
- Most natural-looking car arrangement and spacing
- Better use of the plate's surface area, creating more pleasing proportional relationships
- The brick pattern feels completely uninterrupted and authentic
- The scene maintains its authentic city street character while feeling complete rather than empty

D. NSFWR Computation

As mentioned in Sec. 4.5, we compute the ratio of NSFW outputs to all images as an evaluation metric, using stable-diffusion-safety-checker [5] as the detector. As all ratios (including the original COCO images) were higher than expected, we investigated the images and found out that the safety checker is not very accurate. We compile one false positive per method in Fig. 2, which are everyday images or incorrect generations. Food items or elongated artifacts seem to trigger the label. Consequently, we manually checked all positives and re-computed NSFWR manually, as reported in Tab. 3 col. 8 of the main paper. As discussed in Sec. 5.3, FuLLaMa inherently avoids painting NSFW content as it would significantly effect L_e.

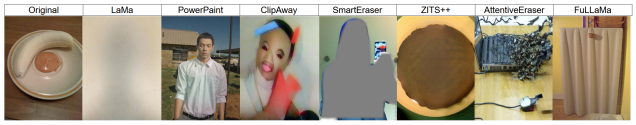

Figure 2. **NSFW False Positives.** Samples from each method, mislabeled by the safety checker.

E. Diffusion Model Flexibility

As mentioned in Sec. 6.2, we test 15 diffusion backbones for D on LaMa samples as listed in Tab. 2. DD stands for

LVLM	SmolVL[12]	MoonDream[18]	DeepSeekVL2[21]	Qwen2-VL-2B-Instruct -GPTQ-Int8[20]	Qwen2-VL-7B-Instruct[20]	Llama3-llava-next-8b[10]
FID↓	148.73	148.11	148.48	143.15	149.78	148.94
LPIPS↓	0.159	0.155	0.159	0.156	0.156	0.155
CLIP↓	0.189	0.185	0.186	0.183	0.184	0.190
CMMD↓	0.539	0.508	0.587	0.489	0.569	0.552
IQA↓	38.69	38.49	38.66	38.41	38.78	38.51
Time↓	429.62	366.95	376.55	448.65	411.18	397.08

Table 1. **LVLM Ablation.** Distribution, quality, and time impacts of six LVLMs are compared with no significant winner.

Backbone	ReMOVE↑	CMMD↓	FID↓
Kolors	0.7673	0.7896	176.94
RealVisXLV5	0.7604	0.6125	170.21
Kardinsky2	0.7781	0.6052	173.63
Dreamshaper8	0.7771	0.5131	160.45
SD1.5	0.7721	0.5414	154.88
SDXL	0.7591	0.6322	173.59
SDXL (DD)	0.7652	0.5909	168.97
SD2	0.7718	0.5367	156.05
SD2.1	0.7742	0.6216	168.57
SD3.5 (CN)	0.7770	0.6779	168.66
SD3.5 (DD)	0.7740	0.6294	167.79
SD3.5	0.7636	0.6568	163.91
Flux1Dev (DD)	0.7619	0.5773	158.09
Flux1Dev	0.7617	0.5404	154.80
FluxFill	**0.7905**	**0.5114**	**154.47**

Table 2. **Diffusion Backbone Performance.** FuLLaMa can use any diffusion backbone, FluxFill is chosen empirically.

DifferentialDiffusion version and CN stands for ControlNet version. FluxFill (without the optimization) quantifiably beats others in distribution and removal metrics, thus we employ it as our DM backbone.

F. Image Quality Metric Comparison

As no-reference image quality metrics are blooming with the advances in learning-based methods [11], each metric is better suited for different domains, tasks, and deformations. Hence, we select Q empirically. First, we generate an initial I_0 removal (Fig. 3 right) and a DM-based perfect removal I (Fig. 3 left). Then, we compute NR-IQA scores with 42 methods in Tab. 3 with the compute time. As in Sec. 6.3, we aim to produce $Q(I) > Q(I_0)$ (for increasing scores, or vice versa), work fast, and have significant $Q(I) - Q(I_0)$.

We pick the top 3 metrics partially satisfying these conditions, namely PaQ-2-PiQ [22], WaDIQaM [15], and ARNIQA [1]. Then, we compute the full dataset metrics plugging each into the full pipeline as in Sec. 6.3, which makes PaQ-2-PiQ the winner. It also makes perfect sense that PaQ-2-PiQ works best for our formulation, because it is built on learned human preferences. From 40000 real-world distorted pictures and 120000 patches, they collect 4M human judgments of picture quality, which is then used to train deep region-based architectures for global picture quality predictions. We also demonstrate a result with PaQ-2-PiQ vs. ARNIQA in Fig. 5 of the main paper, cols 7 vs. 3, where Q=ARNIQA produces a visibly blurred result.

Original I **Initial** I_0

Figure 3. **NR-IQA Test Sample**. Images 1 and 2 corresponding to the Q pre-selection in Tab. 3.

G. LVLM Ablation

As V is another crucial part of FuLLaMa, we compare six LVLMs in terms of their impact to our distribution, context, visual quality, and processing time. Between SmolVLM [12], MoonDream [18], DeepSeekVL2 [21], Qwen2-VL-2B and 7B-Instruct [20], and LLama3-llava-next-8B [10], no significant difference is observed in Tab. 1. We proceed with LLaMa, coherent with the system name.

H. Mask Robustness

Both painters and erasers need context to inpaint as coherent as possible; hence non-object-defining masks, small objects, large areas, random strokes, multiple discontinuous blobs, or tight vs. dilated masks all affect their performance. We demonstrate FuLLaMa's robustness on different masks in Fig. 4. In addition to its visual and contextual success in large area (the whole table) and small object (glass cups) removals, it is also capable of erasing non-scene conforming frames (outpaint) and inpaint random strokes. Furthermore, we investigate performance with automatic (tight but inaccurate) and manual (dilated but accurate) masks. On bottom left, automatic mask behave like an eraser and manual mask behave like a painter for removing the grandma. Opposite tendency emerges when the mask covers small baked goods.

Method	Better	Image 1 Score	Image 2 Score	Winner	Time (s)	Difference
Traditional No-Reference Metrics						
brisque	↓	42.36	41.79	Image 2	0.43	0.57
brisque_matlab	↓	41.94	41.07	Image 2	0.06	0.87
niqe	↓	4.00	4.06	Image 1	0.31	0.06
niqe_matlab	↓	3.49	3.44	Image 2	0.06	0.05
ilniqe	↓	29.25	34.25	Image 1	0.57	5.00
pi	↓	4.08	4.41	Image 1	2.31	0.33
piqe	↓	41.92	40.06	Image 2	0.04	1.86
Learning-Based Metrics						
nrqm	↑	5.56	6.05	Image 2	2.27	0.49
maniqa	↑	0.296	0.269	Image 1	1.97	0.027
maniqa-kadid	↑	0.338	0.355	Image 2	1.76	0.017
maniqa-pipal	↑	0.492	0.464	Image 1	1.74	0.028
musiq	↑	63.22	59.44	Image 1	0.22	3.78
musiq-spaq	↑	65.43	63.49	Image 1	0.21	1.94
musiq-paq2piq	↑	70.99	68.85	Image 1	0.21	2.14
musiq-ava	↑	4.35	4.33	Image 1	0.21	0.02
nima	↑	4.89	4.68	Image 1	0.63	0.21
nima-vgg16-ava	↑	4.79	4.36	Image 1	1.11	0.43
dbcnn	↑	0.404	0.380	Image 1	0.87	0.024
cnniqa	↑	0.504	0.475	Image 1	0.70	0.029
hyperiqa	↑	0.396	0.341	Image 1	0.47	0.055
wadiqam_nr	↑	-0.623	-0.631	**Image 1**	**0.09**	**0.008**
paq2piq	↑	68.56	64.41	**Image 1**	**0.20**	**4.15**
CLIP-Based and Vision-Language Metrics						
topiq_nr	↑	0.443	0.383	Image 1	0.62	0.060
topiq_nr-flive	↑	0.729	0.706	Image 1	0.54	0.023
topiq_nr-spaq	↑	0.503	0.468	Image 1	0.47	0.035
clipiqa	↑	0.557	0.363	Image 1	1.53	0.194
clipiqa+	↑	0.589	0.484	Image 1	1.48	0.105
clipiqa+_vitL14_512	↑	0.470	0.344	Image 1	6.00	0.126
clipiqa+_rn50_512	↑	0.341	0.252	Image 1	1.46	0.089
liqe	↑	2.81	2.11	Image 1	2.06	0.70
liqe_mix	↑	3.45	3.75	Image 2	2.08	0.30
arniqa	↑	0.646	0.562	Image 1	0.37	0.084
arniqa-kadid	↑	0.639	0.632	Image 1	0.32	0.007
arniqa-csiq	↑	0.842	0.785	**Image 1**	**0.32**	**0.057**
tres	↑	44.92	42.71	Image 1	1.50	2.21
tres-flive	↑	85.27	83.35	Image 1	1.51	1.92
unique	↑	0.446	0.396	Image 1	0.26	0.050
laion_aes	↑	4.93	4.91	Image 1	5.43	0.02
qalign	↑	3.95	3.58	Image 1	6.99	0.37
qalign_8bit	↑	3.82	3.44	Image 1	7.12	0.38
qualiclip	↑	0.747	0.607	Image 1	1.44	0.140
qualiclip+	↑	0.593	0.560	Image 1	1.54	0.033

Table 3. **NR-IQA Test.** Scores of I and I_0 in Fig. 3, their difference, and processing time, are comprehensively compared to define a smaller candidate set for Q between 42 no reference image quality metrics.

I. Input Modalities

Fig. 5 top row sequentially demonstrates when M is created with the point-and-segment input of FuLLaMa using SAM2 [13]. One click (the green star) causes automatic segmentation of the gate as in the mid column for removal in the last column. Bottom row demonstrates a language-based input using Sa2VA [23] where "`Segment the most prominent thing in the scene`" (as vague as it sounds) leads to segmenting and removing Riley.

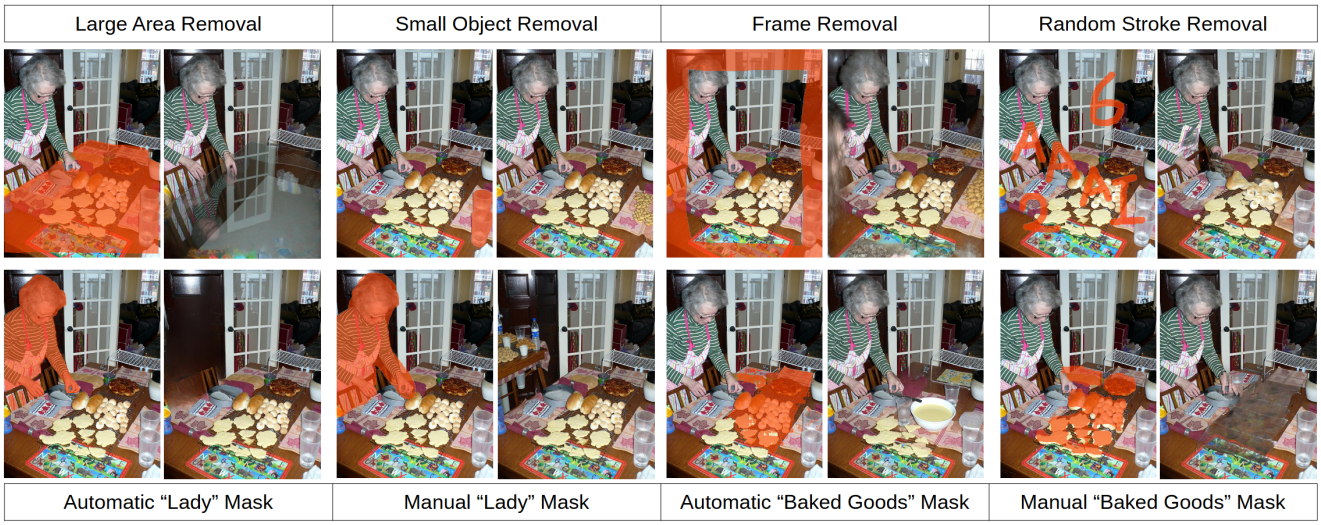

Figure 4. **Mask Dependency.** FuLLaMa can remove large areas, small objects, outer frames, and random strokes consistently (top row). Mask tightness and accuracy can change the painter-like or eraser-like behavior (bottom row).

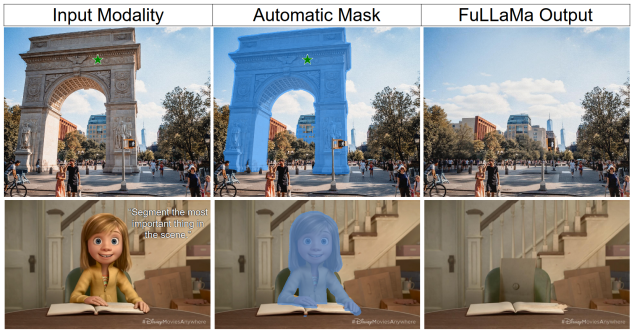

Figure 5. **Input Modalities.** Point-and-segment (top row) and language-based input options of FuLLaMa.

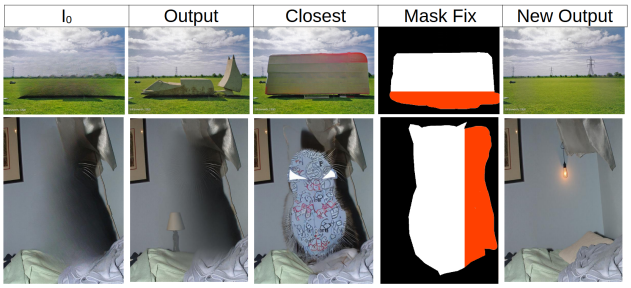

Figure 6. **Shadow Limitation.** FuLLaMa may conform to shadows, which can be fixed by including it in the mask.

J. Shadow Limitation & Fix

If an object to be removed has a shadow not included in the mask, initialization may be incorrect (Fig. 6 col. 1) causing the navigation to conform to the shadow instead of the environment, making it behave like a painter (adding a statue creating the exact shadow in Fig. 6 col. 2 top) or an eraser (blurring from the cat shadow to the wall). We show the closest SOTA eraser and painter results on the same limitation case with ZITS++ (col. 3 top) and SmartEraser (col. 3 bottom). We propose a quick fix as adding the shadow to the mask (red areas), which creates seamless, coherent, and high quality results (last col.).

K. Additional Results

Finally we demonstrate more examples and comparisons in the following pages. Fig. 7 samples scenes from places and animations with hard cases of removing multiple instances, transparent areas, and complex/stylized/blurred backgrounds. Fig. 8 demonstrates commercial and social media use cases modifying ads, text, offensive content, and of course dogs. Fig. 9 depicts the non-cropped version of Fig. 3 of the main paper for reference. Note that the perfection of FuLLaMA is hidden in the details, so we chose to place zoomed-in version in the main paper.

L. Comparison to Newest Image Editors

Thanks to reviewer feedback, we compare FuLLaMa to newest LVLM-based image editing models [7, 9, 14] in Fig. 10. FuLLaMa carefully balances what and how much to edit, as opposed to under-editing (e.g., Nano Banana for planes or Flux Kontext for the pizza and fire hydrants) or over-creation (e.g., Nano Banana imagining the table of the pizza or Seedream making up a full grass scene and a new museum), while preserving the context much better (e.g., the bench drawings and the glass table). Same simple prompts are used for all, such as “remove all planes”.

Figure 7. **Animation Samples.** Comparison of renderings for SOTA models and FuLLaMa in animation captures. Input modality is listed for each column. Zoom-in for artifacts.

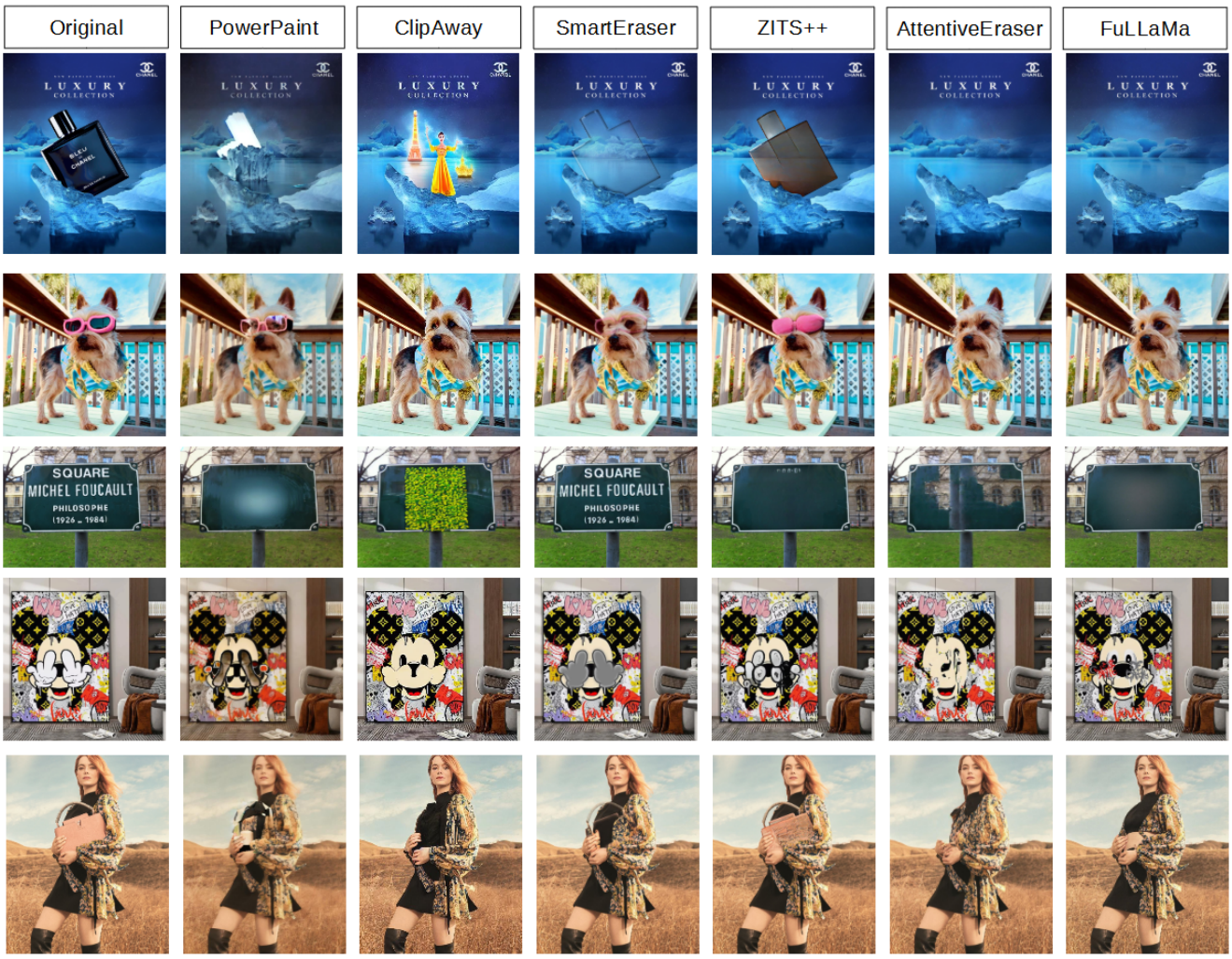

Figure 8. **Branding Samples.** Comparison of renderings for SOTA models and FuLLaMa for advertisement images.

Figure 9. **Qualitative Comparison** on LaMa samples. FuLLaMa creates picture-perfect removals while preserving patterns, context, reflections, and perspective; as opposed to erasers creating blurry areas and painters infilling irrelevant structures. Zoom-ins are provided in Fig. 3 of the main paper.

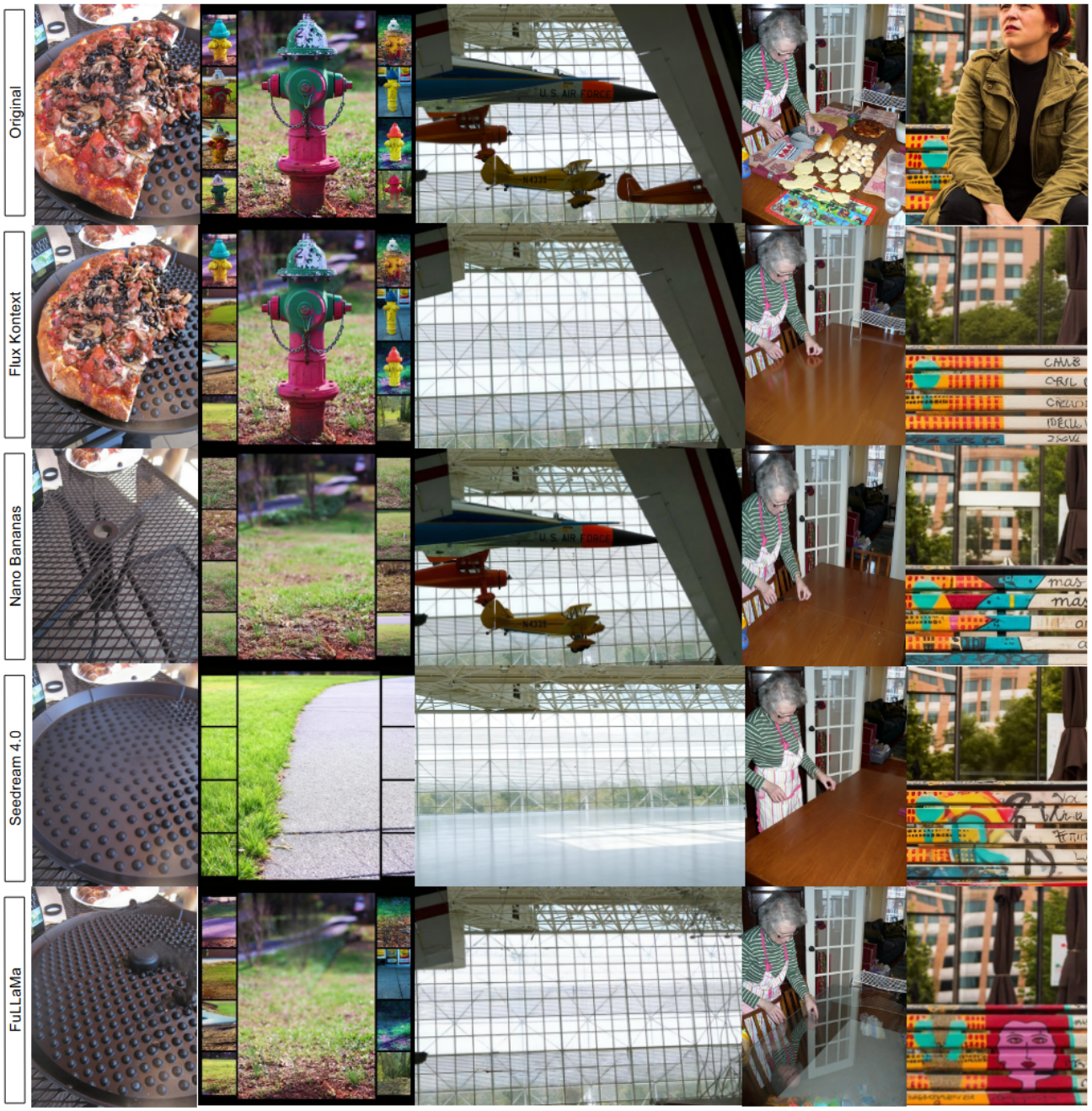

Figure 10. **Qualitative Comparison** to latest image editing models. FuLLaMa carefully balances what and how much to edit, as opposed to under-editing or over-generation by others, while preserving the context much better.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 189–198, 2024. 3

[2] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Zits++: Image inpainting by improving the incremental transformer on structural priors. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12667–12684, 2023. 2

[3] Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Trans. Math. Softw.*, 45(3), 2019. 2

[4] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 2

[5] CompVis. Stable diffusion safety checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker, 2022. 2

[6] Yiğit Ekin, Ahmet Burak Yildirim, Erdem Eren Çağlar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models. *Advances in Neural Information Processing Systems*, 37:17572–17601, 2024. 2

[7] Google DeepMind and Contributors. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[8] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24452–24462, 2025. 2

[9] 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. *arXiv preprint arXiv:2506.15742*, 2025. 5

[10] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[11] Qingyu Mao, Shuai Liu, Qilei Li, Gwanggil Jeon, Hyunbum Kim, and David Camacho. No-reference image quality assessment: Past, present, and future. *Expert Systems*, 42(3): e13842, 2025. 3

[12] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 3

[13] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[14] Team Seedream, :, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 5

[15] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021. 3

[16] Wenhao Sun, Xue-Mei Dong, Benlei Cui, and Jingqun Tang. Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20734–20742, 2025. 2

[17] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[18] Vikhyat Korrapati. moondream2 (revision 92d3d73), 2024. 3

[19] Patrick von Platen, Suraj Patil, Kandasamy Rajagopal, Olivier Dehaene, Lewis Tunstall, Kashif Rasul, Scott Rome, Theo Blodgett, Ayush Thakur, Gabriel Brodie, Merve Noyan, Anastasiia Kornilova, Shafiullah Qureshi, Joachim Schmidt, Marc Brummer, Sylvain Gugger, and Hugging Face. Diffusers: State-of-the-art diffusion models for image and audio generation in pytorch and flax. https://github.com/huggingface/diffusers, 2022. 2

[20] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[21] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu,

Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 3

[22] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585, 2020. 3

[23] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 4

[24] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024. 2