Erase Diffusion: Empowering Object Removal Through Calibrating Diffusion Pathways

Supplementary Material

1. Derivation of Chain-Rectifying Algorithms

The Chain-Rectifying Algorithms presented in this study significantly enhance the visual coherence and elimination of artifacts in the erase inpainting task. A critical aspect of this enhancement is the domain transform from the noise ϵ predicted by standard diffusion techniques to predicted noise ϵ_θ generated by our method. In the subsequent sections, we will thoroughly elucidate this difference and its implications for the inpainting process.

Based on Equation 6 and Equation 7 in the main text of the paper, we can obtain

$$\begin{aligned}\boldsymbol{x}_{t-1}^{mix} = \sqrt{\bar{\alpha}_{t-1}}\left(\bar{\alpha}_{t-1}\boldsymbol{x}_0^{ori} + (1-\bar{\alpha}_{t-1})\boldsymbol{x}_0^{obj}\right) \\ + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon.\end{aligned} \tag{1}$$

Additionally, by transforming Equation 7, we can derive

$$\tilde{\boldsymbol{x}}_t^{mix} = \frac{\boldsymbol{x}_t^{mix} - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}. \tag{2}$$

Substituting Equation 2 into Equation 7 in the main text and replacing λ_t with $1-\bar{\alpha}_t$ according to the experimental setup, we can obtain

$$\frac{\boldsymbol{x}_t^{mix} - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}} = \bar{\alpha}_t\boldsymbol{x}_0^{ori} + (1-\bar{\alpha}_t)\boldsymbol{x}_0^{obj}. \tag{3}$$

By combining Equation 1, Equation 2, and Equation 3, we can further infer that

$$\begin{aligned}\boldsymbol{x}_{t-1}^{mix} = \frac{1}{\alpha_t\sqrt{\alpha_t}}\boldsymbol{x}_t^{mix} + \frac{\sqrt{\bar{\alpha}_{t-1}}(\alpha_t-1)}{\alpha_t}\boldsymbol{x}_0^{obj} \\ + (\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\alpha_t\sqrt{\alpha_t}})\epsilon.\end{aligned} \tag{4}$$

According to the DDIM inversion [10], we can also obtain

$$\boldsymbol{x}_{t-1}^{mix} = \sqrt{\bar{\alpha}_{t-1}}\frac{\boldsymbol{x}_t^{mix} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta, \tag{5}$$

where ϵ_θ is the prediction of our erase diffusion model.

Finally, by combining Equation 4 and Equation 5, we can obtain

$$\mathrm{G}\epsilon_\theta = \mathrm{A}\boldsymbol{x}_t^{mix} + \mathrm{B}\boldsymbol{x}_0^{obj} + \mathrm{C}\epsilon, \tag{6}$$

where

$$\mathrm{G} = \sqrt{\alpha_t - \bar{\alpha}_t} - \sqrt{1-\bar{\alpha}_t},$$

$$\mathrm{A} = \frac{1}{\alpha_t} - 1,$$

$$\mathrm{B} = \frac{\sqrt{\bar{\alpha}_t}(\alpha_t - 1)}{\alpha_t},$$

$$\mathrm{C} = \sqrt{\alpha_t - \bar{\alpha}_t} - \frac{\sqrt{1-\bar{\alpha}_t}}{\alpha_t}.$$

From the equation 6, we can find that our network adjusts the standard predictions ϵ by leveraging the object information $\boldsymbol{x}_0^{obj}$ and the latent state $\boldsymbol{x}_t^{mix}$.

2. Comparison with Standard Diffusion

To perform a thorough comparative analysis of the performance between the standard diffusion training method and the erase diffusion training method in the context of the erase inpainting task, we first trained the SD2-Inpaint model (hereafter referred to as StandDiff) using the conventional training approach on the OpenImages V5 dataset [3]. The masking strategies employed include rectangular, elliptical, and irregular masks, as well as their random combinations. Additionally, in subsequent experiments, we investigated the impact of constraining the masked regions to the background areas of the original images (designated as BGDiff). This approach is intended to enhance the SD2-Inpaint model's tendency to recover background information during the denoising process, thereby visually improving the effectiveness of eliminating the target object. Figure 1 presents the relevant masking strategies, while the other training methodologies remain consistent with those utilized in EraDiff to ensure a fair comparison.

Table 1 presents the relevant experimental results. The data clearly indicate that the diffusion models trained using these methodologies achieve relatively low and comparable LPIPS scores, suggesting that these models effectively maintain visual coherence in the erased images, which represents a significant advantage of diffusion-based approaches. However, when employing the standard training method, the diffusion model exhibits suboptimal performance in object elimination. Conversely, restricting the masked regions during training to the background areas of the original images leads to a noticeable improvement in the model's object elimination capabilities, although it still falls short compared to EraDiff. This phenomenon arises from the observation that both StandDiff and BGDiff emphasize the restoration of identical masked regions at each time step during the training process. In contrast, EraDiff introduces a subtle shift in these regions between the time

Method	LPIPS	Local FID	GPT score
StandDiff	0.285	6.981	35.65%
BGDiff	0.259	6.588	51.06%
EraDiff w/o **SRA**	0.198	4.950	78.54%
EraDiff (ours)	**0.192**	**3.799**	**83.43%**

Table 1. Quantitative assessment of different training methods for diffusion models on the OpenImages V5 test set. Optimal results are highlighted in bold, with runner-up performance underlined.

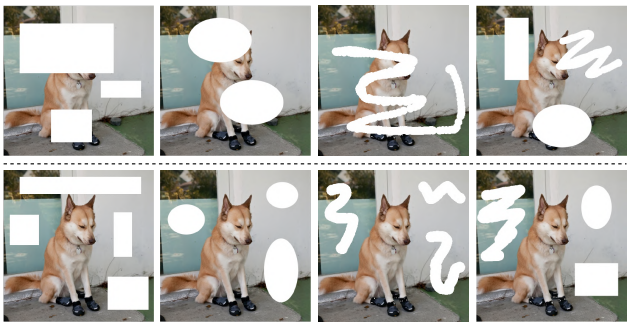

Figure 1. Masking strategies employed in the training processes of StandDiff (first row) and BGDiff (second row).

steps t and $t-1$. This variation can be interpreted as a minor perturbation within the masked areas, which enhances the model's capacity for continuous reasoning as opposed to merely reproducing prior outputs. Consequently, during the denoising process, when artifacts manifest in the erasure-target region, EraDiff dynamically adjusts to progressively eliminate these artifacts. Conversely, models trained using previous methodologies tend to rely heavily on and reinforce their descriptions, resulting in high tolerance for error pixels within the model. It is noteworthy that when t is significantly large, the associated noise will also be substantial, leading to a higher probability of generating artifacts or other non-target content during denoising. Therefore, these high-tolerance models exhibit limited capability in eliminating unwanted objects. The experimental results presented in Figure 3 of the main text further substantiate this observation.

3. Supplementary Experiments

To enhance the credibility of the comparison results between baseline models and EraDiff, we introduce a new testing dataset, FSS-1000 [4]. The FSS-1000 dataset encompasses 1,000 distinct categories with a total of 10,000 samples. It features a rich variety of images of animals and everyday objects, in addition to items such as merchandise and logos, which are relatively underrepresented in other existing segmentation datasets. Regarding evaluation metrics, we utilize LPIPS, FID, and Local FID to measure visual coherence in erased images, which is consistent with the metrics used in the OpenImages V5 dataset [3].In additional experiments, we employ GPT-4o to evaluate the effectiveness of the top three performing models in visual coherence for eliminating erasure targets. All models are compared equitably, without any fine-tuning on the FSS-1000 dataset. Tables 2 and 3 present the relevant experimental results. Additionally, a visual comparison of these methods is illustrated on the FSS-1000 dataset in Figure 2.

Method	FID↓	LPIPS↓	Local FID↓
SD2-Inpaint	6.982	0.248	1.201
SD2-Inpaint*	**6.874**	0.231	1.053
PowerPaint	12.885	0.395	1.759
Inst-Inpaint	7.320	0.336	2.284
LaMa	8.093	0.142	1.185
EraDiff (ours)	7.751	**0.127**	**0.869**

Table 2. Quantitative evaluation of baseline models and EraDiff on the FSS-1000 dataset. The optimal results are indicated in bold, and the sub-optimal results are indicated with underlines.

Method	Superior	Comparable	Inferior
SD2-Inpaint	2.42%	19.36%	78.22%
SD2-Inpaint*	3.78%	25.97%	70.25%
LaMa	12.93%	33.16%	53.91%

Table 3. Quantitative results of FSS-1000 dataset among SD2-Inpaint, SD2-Inpaint*, LaMa, and EraDiff. This table delineates a comparative analysis of the elimination performance results obtained by these methodologies relative to ours, highlighting whether their outcomes are superior, comparable, or inferior to those achieved by our approach.

The results presented in Table 2 illustrate the performance of the models SD2-Inpaint, SD2-Inpaint*, LaMa, and EraDiff, all of which demonstrate commendable outcomes as assessed through the LPIPS and FID metrics. Notably, EraDiff attains the best scores in both LPIPS and Local FID, further validating its capability to ensure visual coherence in erased images while maintaining high visual fidelity in the erased regions. Table 3 offers a comparative analysis of the performance of SD2-Inpaint, SD2-Inpaint*, and LaMa relative to our proposed model, EraDiff, in eliminating specified objects. The results unequivocally demonstrate that these alternative models exhibit significantly inferior performance relative to EraDiff. One of the primary reasons for this performance gap is the presence of undesirable artifacts in the erased regions of images processed by these alternative methods. Such limitations are visually substantiated by the evidence presented in Figure 2. These findings further substantiate that EraDiff, through the calibration of sampling pathways, effectively accomplishes tar-

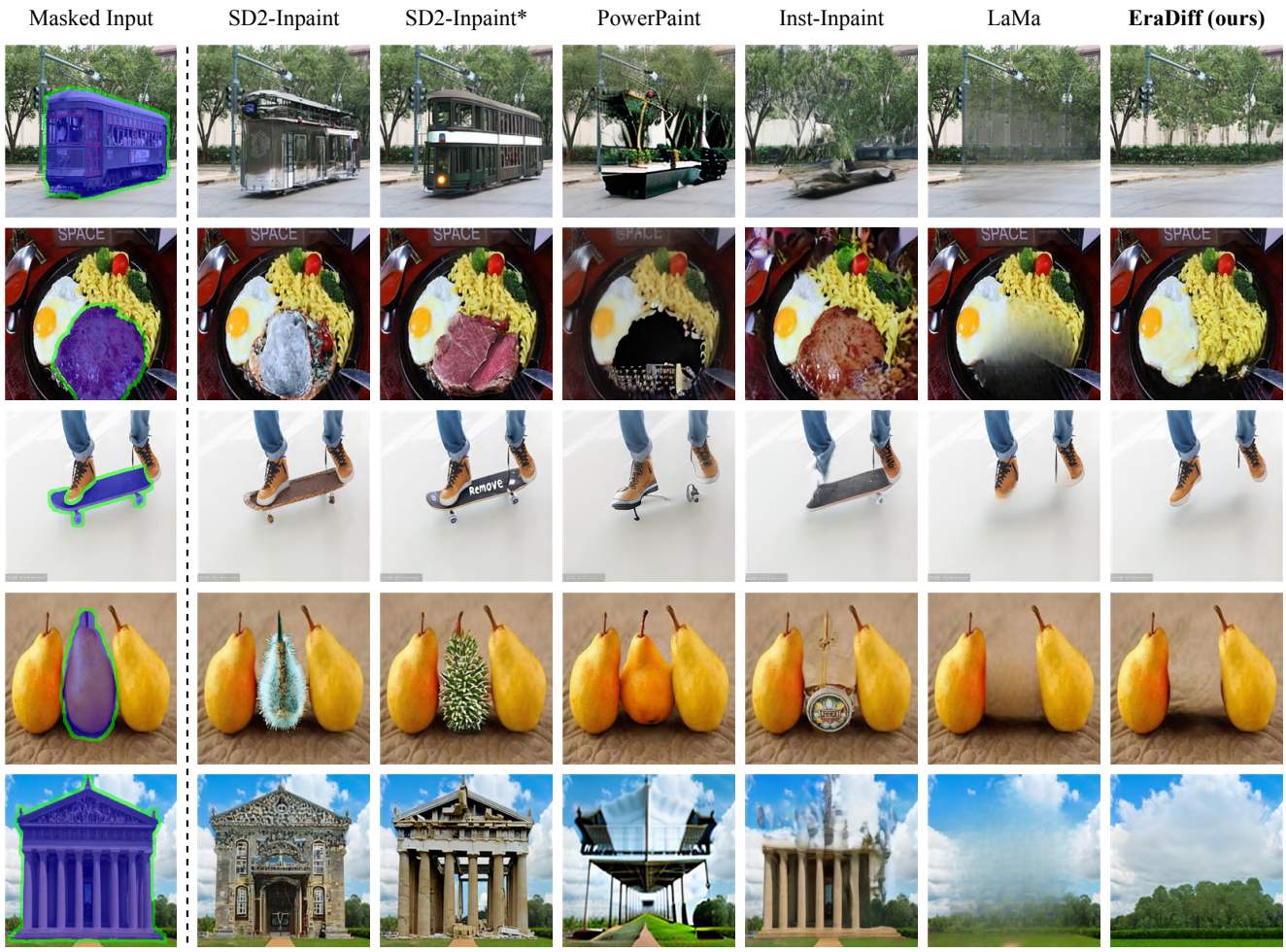

Figure 2. Qualitative results of FSS-1000 dataset compared among SD2-Inpaint [6], SD2-Inpaint with prompt guidance [6], PowerPaint [15], Inst-Inpaint [13], LaMa [11], and our approach.

get removal during the task of erase inpainting and excels in achieving both high visual quality and consistency in the generated images.

4. Generalization in Varied Scenarios

To rigorously evaluate the generalization capability of EraDiff, we conducted comprehensive experiments utilizing two distinct datasets, each exhibiting unique distribution characteristics. The first dataset comprises marketable product images sourced from online e-commerce platforms, while the second dataset includes cartoon character images drawn from the publicly available ATD-12k dataset [9]. We meticulously analyzed the performance of EraDiff in comparison to baseline models, with the resulting visualizations provided in Figures 5 and 6, respectively. These comparisons highlight the efficacy of our proposed model across diverse image distributions.

5. Experimental Details

Data synthesis. As shown in Figure 3, the data synthesis process required during the training phase is as follows: First, we employed matting techniques [1, 12] to extract the foreground $\boldsymbol{x}_0^{obj}$ from the $\boldsymbol{x}_0^{ori}$ and obtained a corresponding mask necessary for distinguishing between the foreground and background. Next, we scaled the $\boldsymbol{x}_0^{obj}$ by a random ratio ranging from 50% to 120%, followed by a random rotation from 0 to 360 degrees. Finally, we utilized the earlier obtained mask to seamlessly mix-up the modified $\boldsymbol{x}_0^{obj}$ and $\boldsymbol{x}_0^{ori}$ in the background.

Training. The training process utilized a batch size of 32 across a cluster of 16 A100 GPUs, for a total of 5 epochs. The noise scheduler was set to DDIM [10]. In the experiment, we fine-tuned the U-Net [7] of SD2-Inpaint, while the parameters of the other modules were frozen.

Inference. During the inference phase, we employed the

Figure 3. Data synthesis process for model training in this study.

Figure 4. Failure cases of our approach.

DPMSolverMultistepScheduler [5]. Importantly, we deliberately omitted the use of classifier-free guidance (CFG) and auxiliary prompts to simplify the reverse process. All images in this paper were generated at a resolution of 512×512 pixels, utilizing 20 denoising steps with a denoising strength of 0.95.

6. More Quantitative Results

To comprehensively evaluate EraDiff's performance in object elimination, we created a new test set of 10,000 samples using the original data synthesis method from training. This set is based on the OpenImages V5 dataset, with synthesized images requiring objects to be removed and original images serving as the ground truth for comparison.

In this experiment, we introduced novel evaluation metrics to thoroughly assess the quality of both object elimination and image coherence. Specifically, we employed two aesthetic evaluation metrics, AES [8] and NIMA [2], as well as PIDS and UIDS [14], which measure the distinction between the erased images and the target images.

Method	PIDS(%)	UIDS(%)	AES	NIMA
SD2-Inpaint	13.67	23.62	4.721	5.29
SD2-Inpaint*	14.35	22.19	4.793	5.536
PowerPaint	4.36	11.28	4.814	5.205
Inst-Inpaint	0.0	0.500	4.459	5.267
LaMa	18.04	25.92	4.616	5.605
EraDiff	**25.25**	**33.72**	**5.082**	**5.988**

Table 4. Quantitative evaluation of baseline models and EraDiff.

Method	Params	RT(s)
SD2-Inpaint	1.29B	2.61
SD2-Inpaint*	1.29B	2.63
PowerPaint	2.08B	22.86
Inst-Inpaint	0.51B	1.80
LaMa	0.05B	0.43
EraDiff	1.29B	1.74

Table 5. Comparison of model parameters and inference time.

The experimental results in Table 4 demonstrate that EraDiff outperforms the baseline models. This indicates that EraDiff not only effectively removes objects but also maintains the visual consistency of the resulting images.

7. Complexity Evaluation

In this section, we evaluated the computational complexity of the EraDiff model with that of baseline models. The results of this analysis are presented in Table 5. We found that EraDiff exhibits a moderate level of both parameter count and inference time. Specifically, because EraDiff shares the same architectural structure as SD2-Inpaint, their parameter counts are identical. However, EraDiff benefits from reduced inference time due to the absence of CFG utilization, enhancing its real-time performance efficiency.

8. Limitation and Failure Cases

The EraDiff method encounters certain challenges in some situations, as illustrated in Figure 4. Specifically, for document-type data erasure, it tends to produce text-like artifacts in the target erasure regions due to surrounding texts. Additionally, the method may underperform in tasks involving completion. For instance, removing a coat from an individual and reconstructing the arm beneath may yield suboptimal results. Furthermore, when dealing with large-scale background erasure (background replacement), the absence of reference information often leads to less desirable outcomes. We aim to address and overcome these limitations in future research endeavors.

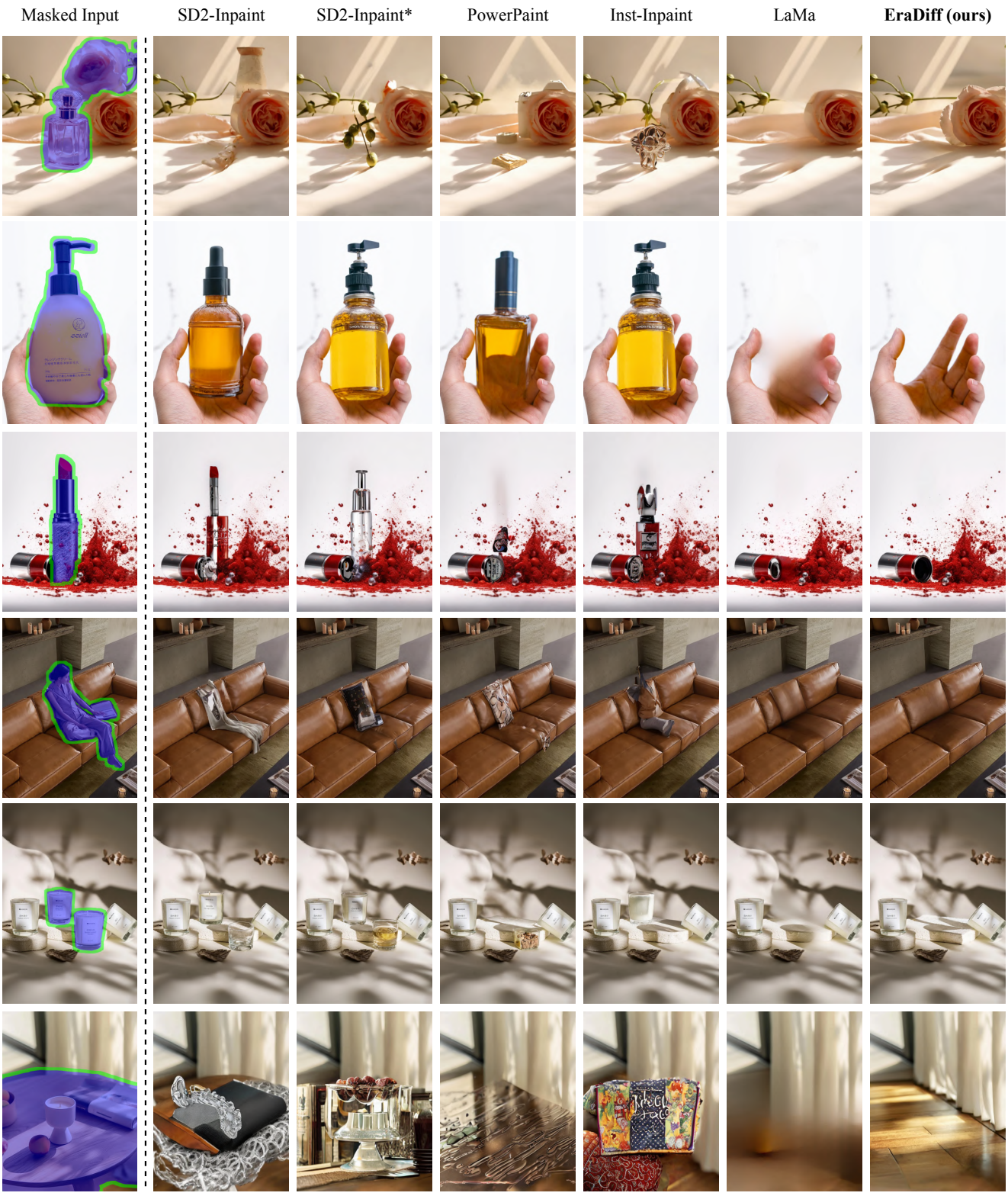

Figure 5. Comparison of visualizations for baseline models and the proposed EraDiff in scenarios of marketable products.

Figure 6. Comparison of visualizations for baseline models and the proposed EraDiff in scenarios of cartoon images.

References

[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3

[2] Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *IEEE Trans. Image Process.*, 27 (8):3998–4011, 2018. 4

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. 1, 2

[4] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, pages 2866–2875. Computer Vision Foundation / IEEE, 2020. 2

[5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 4

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 3

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[8] Christoph Schuhmann. Aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2023. 4

[9] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, 2021. 3

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 1, 3

[11] 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 *WACV*, pages 3172–3182, 2022. 3

[12] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 3

[13] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *CoRR*, abs/2304.03246, 2023. 3

[14] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2021. 4

[15] 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. *CoRR*, abs/2312.03594, 2023. 3