ORIDa: Object-centric Real-world Image Composition Dataset

Supplementary Material

A. Additional Data Samples

A.1. Dataset Samples

Figure A and Figure B display examples of Factual-Counterfactual (F-CF) sets and Factual-Only (F-Only) images. Figure A includes multiple objects, where F-CF sets (left) contain multiple object positions and their corresponding background-only images, while F-Only images (right) show object variations in different scenes. Figure B, on the other hand, focuses on a single object (object #2), illustrating its placement across various contexts, emphasizing the dataset's ability to capture real-world diversity for individual objects.

A.2. Object Categories

Figure C illustrates the diversity of ORIDa through objects categorized by attributes such as the number of colors, semantic class, transparency, reflectivity, and roughness. Objects are grouped by color complexity (1–2 to 7–8 colors) and semantic classes, including daily/office supplies, human-related items, animal-related objects, figures, and miscellaneous categories. Transparency, reflectivity, and roughness levels further represent diverse material properties and surface textures, capturing variations in light interaction and texture. These categories reflect the thoughtful curation of ORIDa, enabling support for a broad range of image editing tasks.

A.3. ISP-augmented data

Figure A also demonstrates images processed through five distinct ISP settings using Adobe Lightroom Classic. These settings include: (1) as-shot (default settings), (2) higher temperature, (3) lower temperature, (4) higher vibrance, and (5) lower vibrance. These augmentations expand the dataset's variety, enabling better generalization for training models under varying lighting and color conditions.

B. Experimental Details

B.1. Train Schedule

We fine-tune our models starting from the pre-trained SD-Inpaint model [4, 8] using the Adam optimizer [6] with a learning rate of 5e-5 and a cosine scheduler [7]. The batch size is set to 64 for both object removal and insertion tasks.

For object removal, the model is fine-tuned for 5,000 steps, resulting in 320,000 training samples, significantly fewer than ObjectDrop's 6.4 million samples, generated over 50,000 steps with a batch size of 128. For object insertion, we train the model for 500,000 steps, resulting in 32 million training samples. This is still fewer than ObjectDrop's 56.3 million samples, generated through a two-stage process: 100,000 iterations with a batch size of 512 on synthetic data, followed by 40,000 iterations with a batch size of 128 on real-captured data [11]. It is also fewer than Paint-by-Example's 76 million samples, produced over 40 epochs using a synthetic dataset of 1.9 million images [12]. Training the object insertion model takes about 150 hours using 4 NVIDIA A100-PCIE (40GB) GPUs.

B.2. Model Inputs

Our framework is built upon the pre-trained SD-Inpaint model [4, 8], where the U-Net processes a 9-channel input: four channels for the input latent, four for the condition latent, and one for the object mask.

Object Removal. For training, the input latent is a perturbed version of the original image's latent representation The condition latent is created by masking the input latent using the object mask. During inference, the model inputs for object removal remain identical to the training setup.

Object Insertion. For training, the input latent is the perturbed latent of the ground truth (real-captured) object-included image. The condition latent is the latent of a Copy & Paste image, which is generated by masking and pasting the source object into the target image. During inference, as the ground truth object-included image is unavailable, we use the Copy & Paste image as the input latent. All other settings remain consistent with the training configuration.

B.3. Diffusion Model

We primarily follow the pipeline of SD-Inpaint [4], making minimal modifications only during the inference stage. To better preserve the source object's identity in the object insertion task, we employ *skip residual* connections inspired by DemoFusion [3]. This method combines the noised latent reference $z_t' \sim q(z_t|z_0)$, derived from the original input image's latent representation z_0, with the current denoised latent $z_t \sim p_\theta(z_t|z_{t+1})$. The contributions of z_t' and z_t are dynamically weighted using a cosine scheduling mechanism, which adjusts the balance between the two throughout the denoising process. For more details on this weighting approach, refer to the original paper [3].

By leveraging information from z_0, the model retains the source object's identity while seamlessly blending it into the target scene. Apart from this inference-stage adjustment using skip residuals, the underlying SD-Inpaint model remains unchanged.

C. Additional Experiments

C.1. Object Removal

Figure E presents further object removal results using multiple models, including SD-Inpaint [4, 8], LaMa [10], MGIE [5], and SD-Ours$_r$. The examples highlight how different methods handle challenges such as background reconstruction, shadow removal, and artifact elimination.

C.2. Object Insertion

Additional qualitative results for object insertion are shown in Figure F with ORIDa test set and Figure G with in-the-wild data from internet and an external dataset, MureCom [1]. The results demonstrate the effectiveness of different models – Copy & Paste, Paint-by-Example [12], AnyDoor [2], ObjectStitch [9], and SD-Ours$_i$ – in maintaining object identity, harmonizing colors, and generating shadows and reflections for seamless integration.

C.3. Ablation Study on Data Scale

We performed an ablation study to evaluate the impact of dataset scale (25%, 50%, 100%) on object insertion performance, focusing on shadow generation and source object preservation (Figure H). Models trained on 25% of the dataset struggled with context-aware shadow generation and exhibited artifacts in object appearance. Increasing to 50% improved performance, however, some inconsistencies still remained. Training on the full dataset (100%) yielded the best results, with accurate shadows and faithful preservation of object identity and appearance. This demonstrates the importance of dataset scale for achieving high-quality, context-aware object insertion.

D. Limitations and Future Works

Excluded object types. Our dataset excludes certain categories of objects to maintain consistency and feasibility during data collection. First, human subjects are excluded due to complexities related to appearance variability and ethical considerations. Additionally, we have excluded objects characterized by significant temporal variability (e.g., living organisms, perishable food items, and deformable or flexible materials), as well as large-scale objects impractical for repeated captures. Addressing these exclusions in future dataset versions could significantly expand the range of applicable research and practical scenarios.

Dataset scale-up. We introduced a dataset for object-centric image composition at an unprecedented scale, containing 200 unique objects across 30,000 images. Despite this significant advancement, it remains insufficient to fully represent the vast diversity and complexity of real-world visual scenarios. We envision that both our dataset and the methodologies developed to capture it will serve as foundational resources for future datasets. An important direction is to simplify and streamline our data collection process, enabling scalable crowd-sourced dataset acquisition.

Limited 3D information and pose variability. Our dataset does not include explicit 3D information, such as multi-view captures or ground truth depth maps. Additionally, the dataset intentionally restricts variation in object poses to maintain consistency and highlight object placement across scenes, resulting in limited diversity in pose dynamics. Future datasets might address this limitation by incorporating more comprehensive pose variations and additional 3D-related annotations.

Limited novelty in model development. To emphasize the value of our high-quality real-world dataset, which can directly serve as training samples for advanced diffusion models, we primarily employed vanilla models [4, 8] with minimal modifications. Consequently, our research did not explore potential performance improvements achievable through advanced architectural innovations or customized model enhancements. Future work may benefit from integrating novel model architectures specifically tailored for object-centric image composition tasks.

E. Broader Impact

ORIDa advances realistic image compositing, supporting augmented/virtual reality and AI-driven content production. However, its realism raises concerns about misuse, such as deepfakes or deceptive media. We encourage responsible use and adherence to ethical guidelines to balance innovation with societal safeguards.

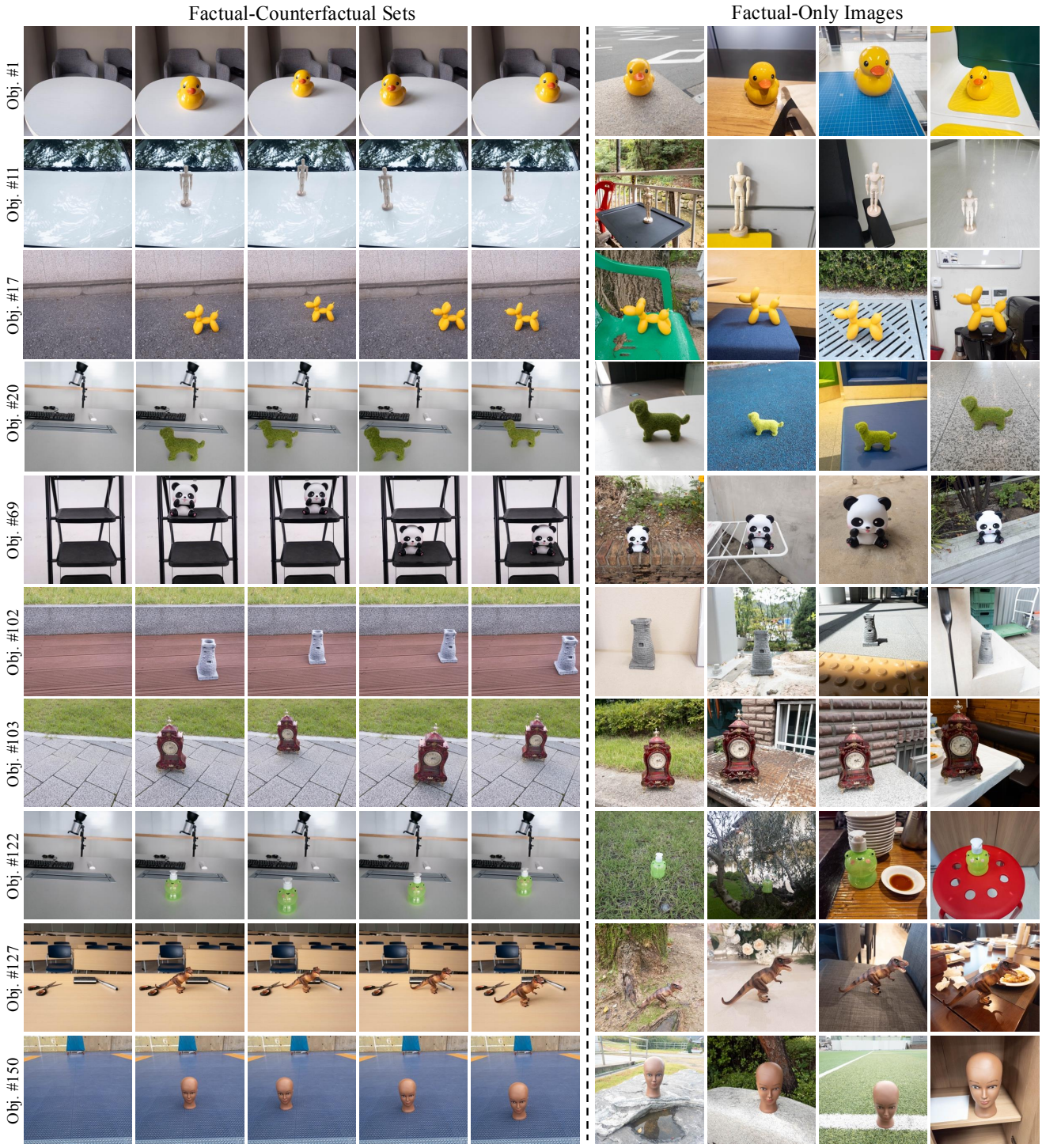

Figure A. Additional examples of Factual-Counterfactual (F-CF) sets and Factual-Only (F-Only) images.

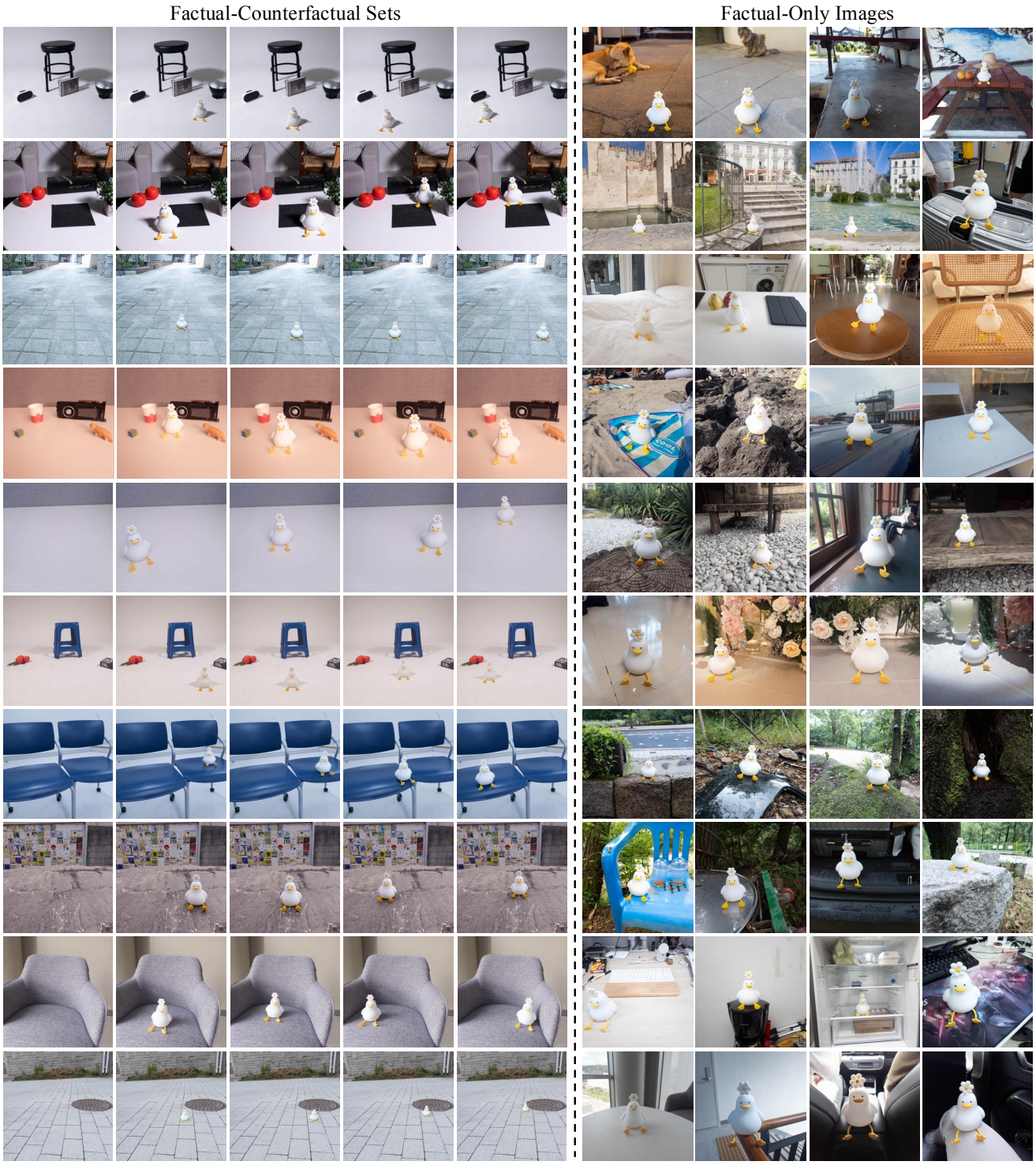

Figure B. Additional examples of Factual-Counterfactual (F-CF) sets and Factual-Only (F-Only) images for object #2.

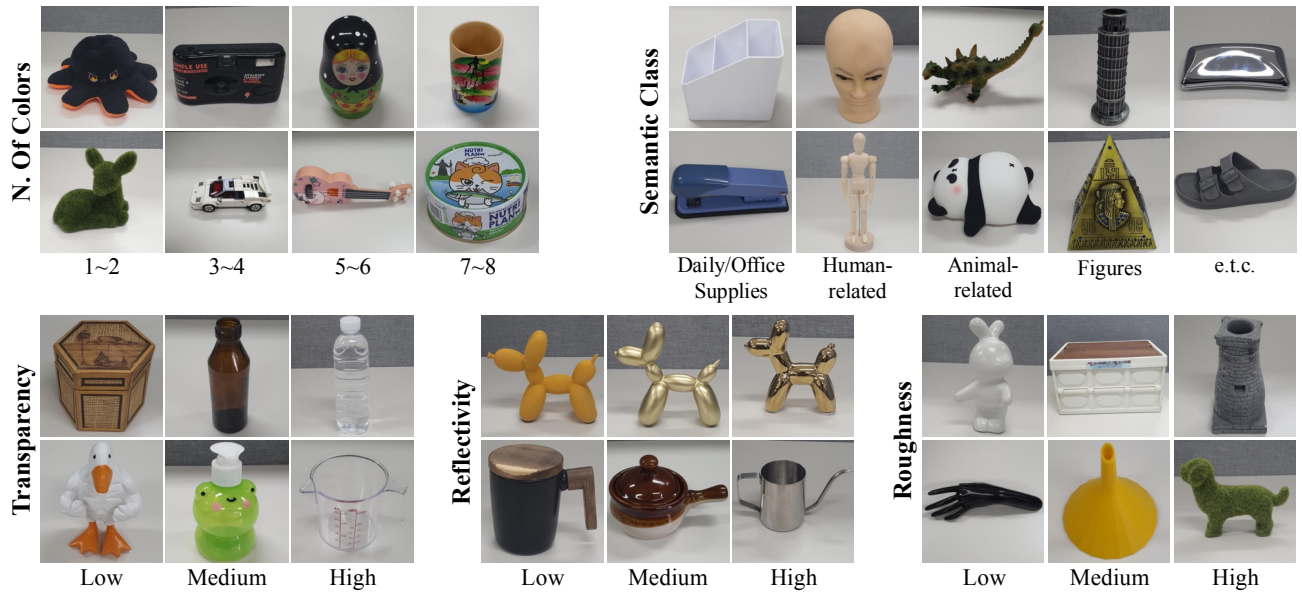

Figure C. Examples of objects categorized by number of colors, semantic class, transparency, reflectivity, and surface roughness.

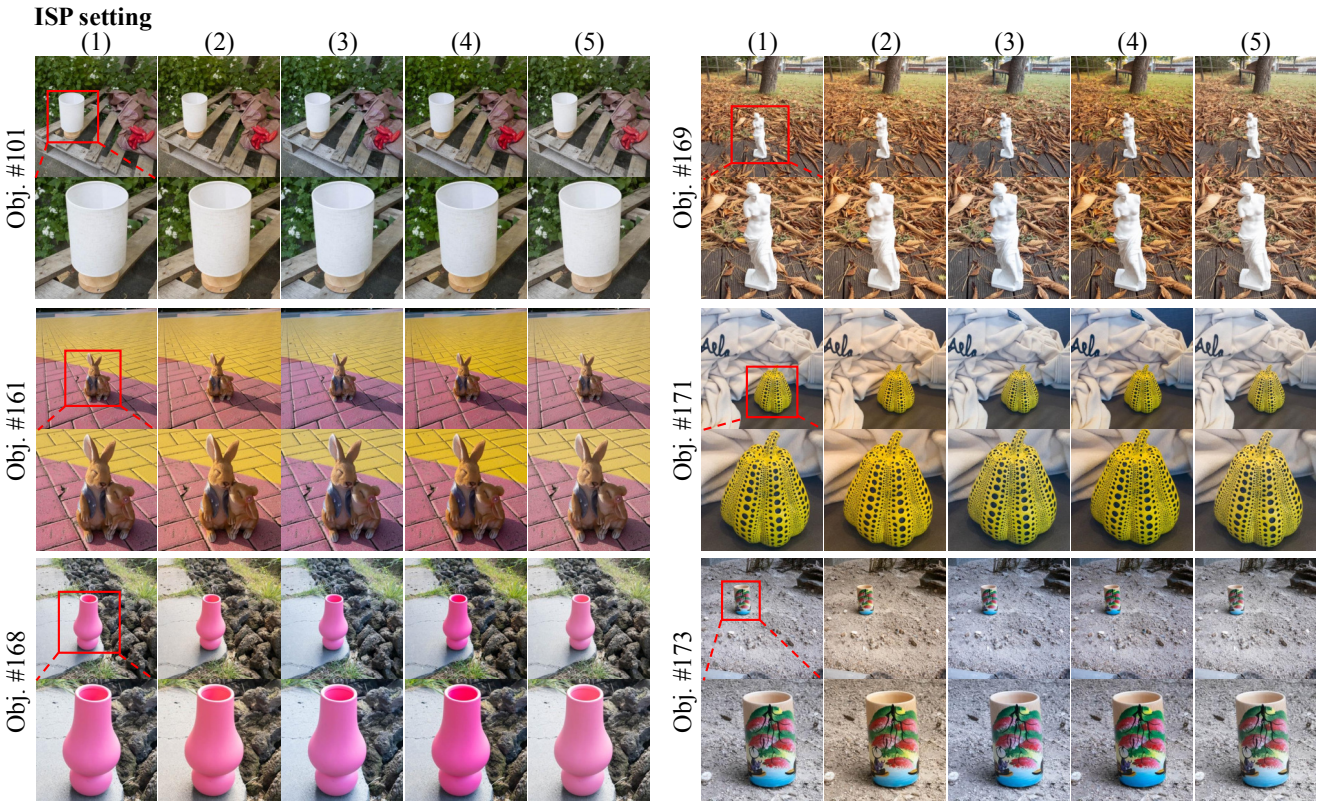

Figure D. Example images data for five different ISP settings: (1) as-shot, (2) higher temperature, (3) lower temperature, (4) higher vibrance, and (5) lower vibrance.

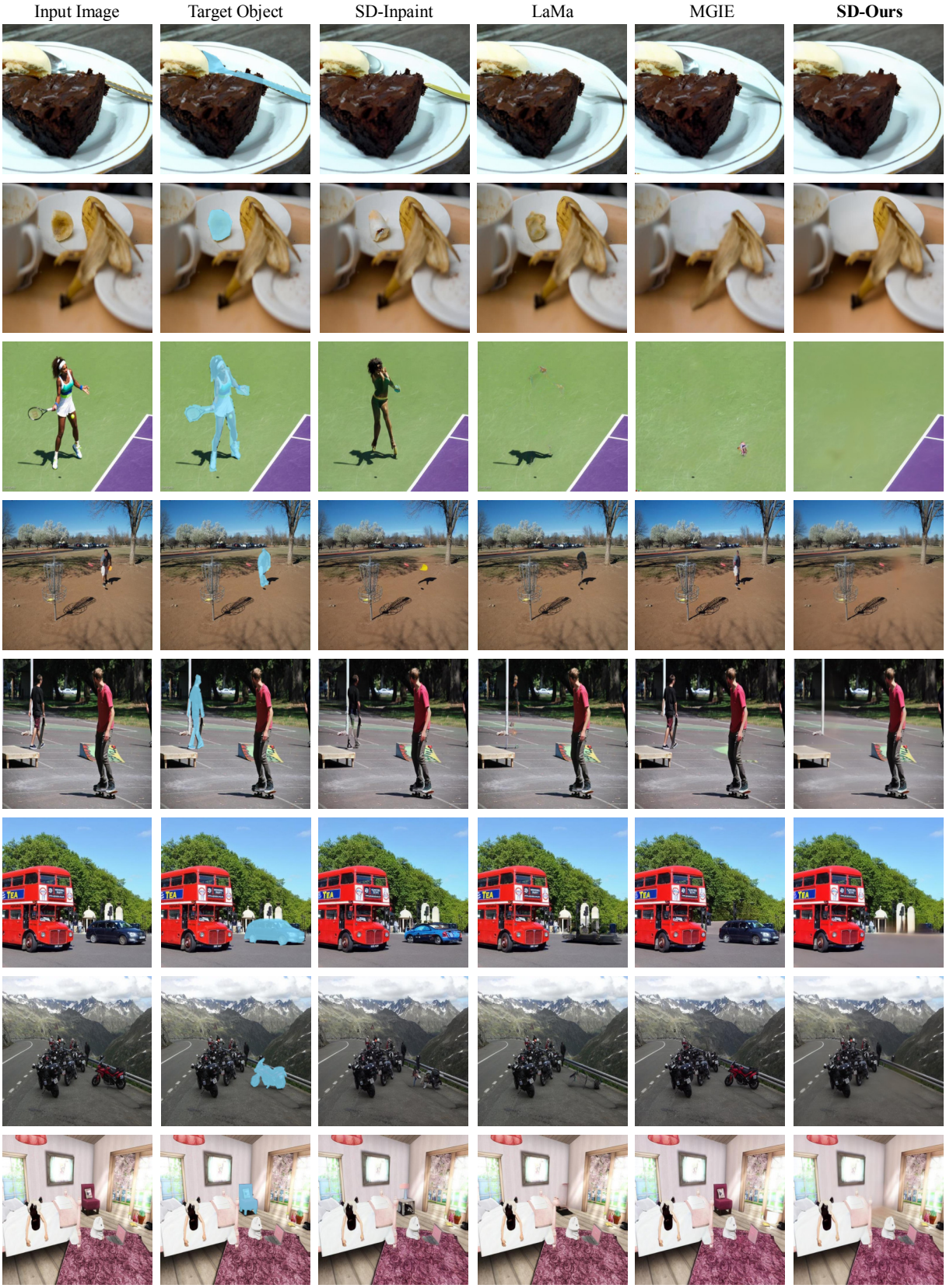

Figure E. Additional results of object removal with various models.

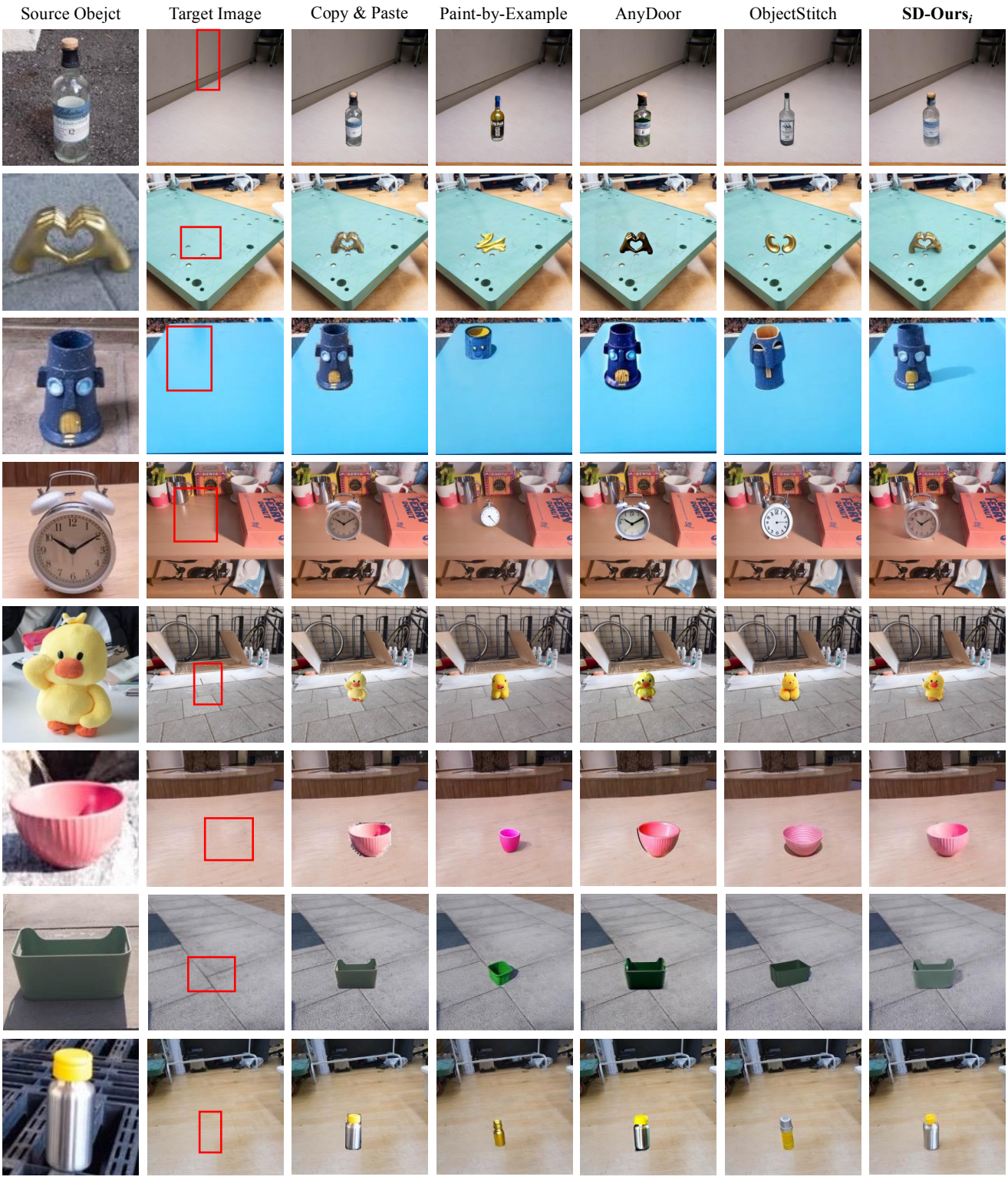

Figure F. Additional results of object removal generated by various models.

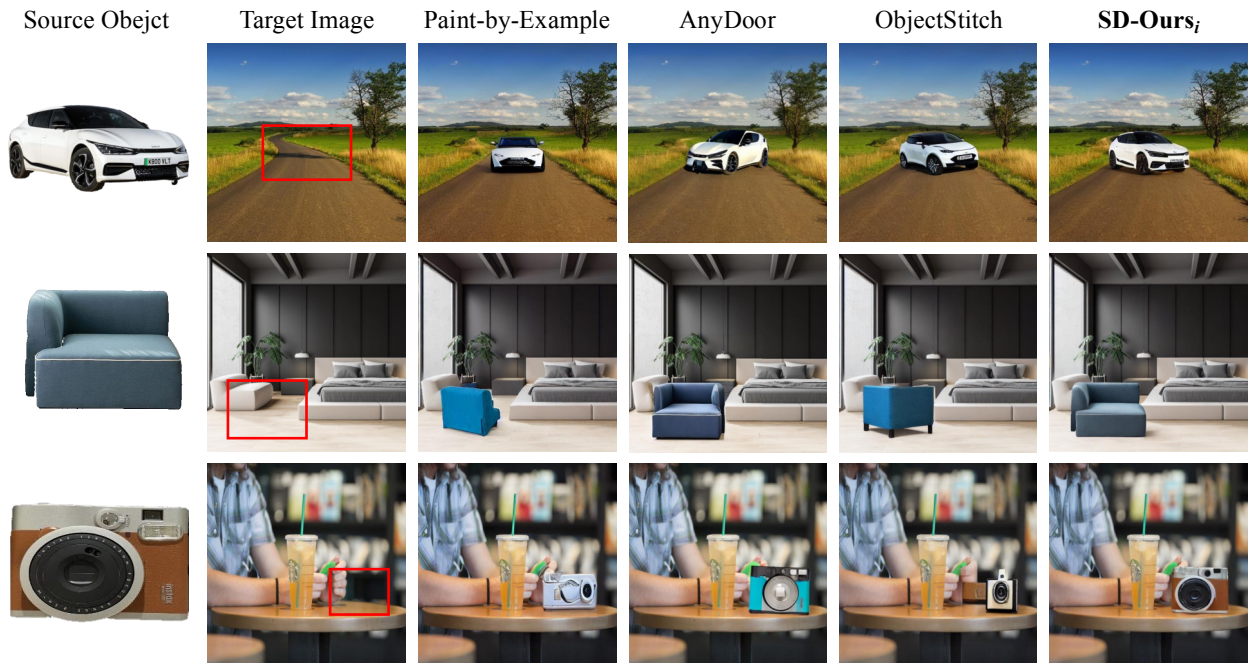

Figure G. Additional results for object insertion with objects and scenes from internet and an external dataset.

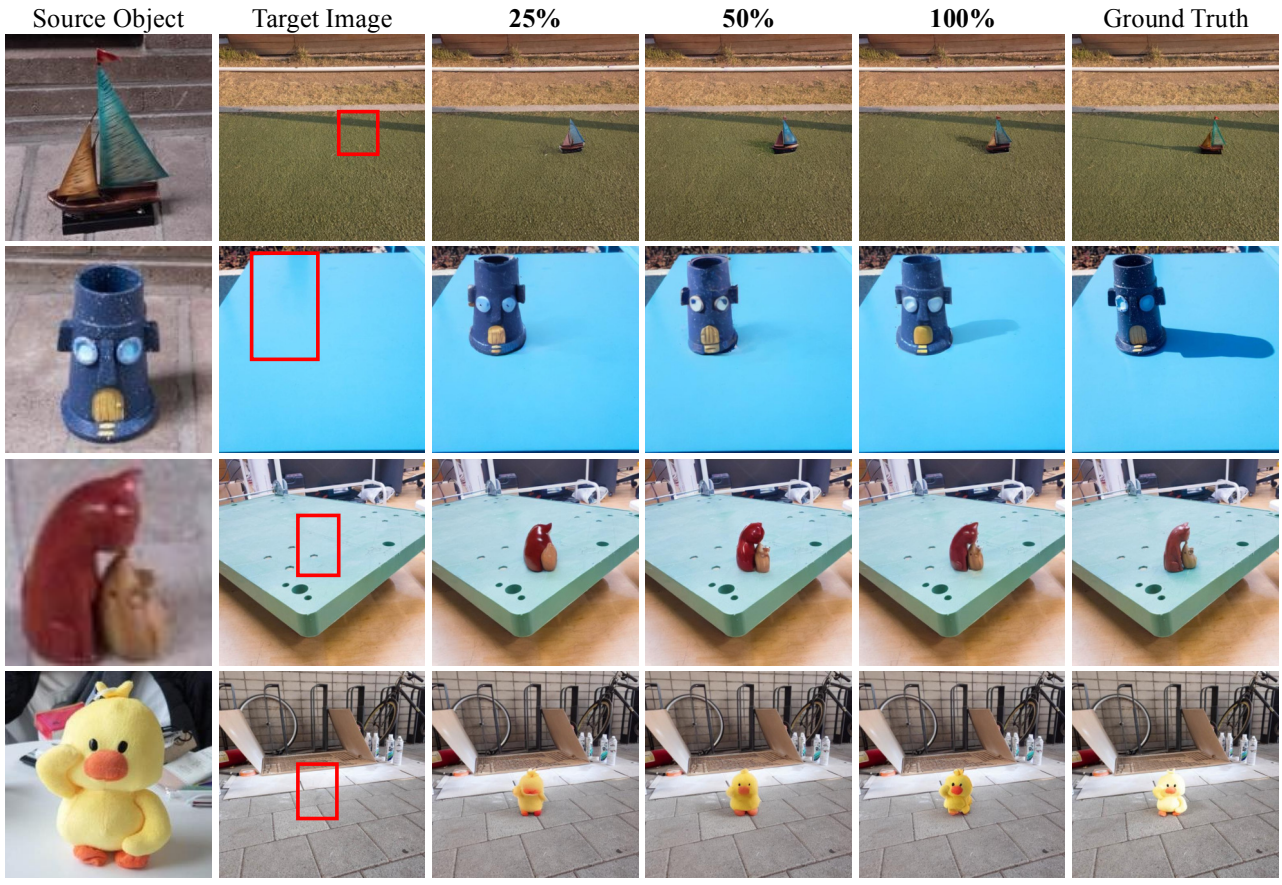

Figure H. Ablation study results on the effect of training dataset size for the object insertion task.

References

[1] Jiaxuan Chen, Bo Zhang, Qingdong He, Jinlong Peng, and Li Niu. Mureobjectstitch: Multi-reference image composition. *arXiv preprint arXiv:2411.07462*, 2024. 2

[2] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 2

[3] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6159–6168, 2024. 1

[4] Hugging Face. Inpainting with diffusers. https://huggingface.co/docs/diffusers/using-diffusers/inpaint, 2024. Accessed: 2024-11-13. 1, 2

[5] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 2

[6] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[9] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2

[10] 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

[11] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv preprint arXiv:2403.18818*, 2024. 1

[12] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1, 2