Generative Image Layer Decomposition with Visual Effects

Supplementary Material

Table S1. Additional ablation study of LAYERDECOMP on held-out test set based on random re-composition.

Model	FID ↓	CLIP-FID ↓
V_0:RGB-only	-	-
V_1:V_0+RGBA FG (obj.)	45.758	3.756
V_2:V_0+RGBA FG (obj.+v.e.)	45.123	3.739
Ours:$V_2 + \mathcal{L}_{\text{consist}}$	44.260	3.173

Table S2. Comparison of LAYERDECOMP with instruction-driven object removal methods on Emu-Edit Remove Set [6].

Model	FID ↓	CLIP-FID ↓
Emu-Edit [6]	47.555	6.711
OmniGen [8]	48.116	6.283
Ours	38.998	5.622

1. Additional Results for the Ablation Study

Test Set Details. The test set is a held-out subset of our camera-captured data consisting of 635 image pairs (composite image and background image). To construct this dataset, we manually collected real-world examples comprising photos of scenes captured before and after the removal of an object, while ensuring all other elements in the scene remain unchanged. We also manually labeled the binary object mask for the removal target. As illustrated in Fig. S2, the dataset encompasses both indoor and outdoor scenarios, effectively reflecting real-world phenomena such as shadows and reflections. This test set allows us to evaluate not only the quality of the decomposed background naturally but also the quality of the foreground. By re-compositing the background and foreground layer output, we can effectively assess the fidelity and visual coherence of the foreground, including the visual effect components.

Qualitative Comparison. To more intuitively demonstrate the effectiveness of our design in LAYERDECOMP, in addition to the quantitative analysis represented in Table 1 of the main paper, we provide more visual results of the four model variants in Fig. S3. For each variant, we present the decomposed background and foreground layers, along with the re-composited image obtained by alpha blending the two layers. For the RGB-only model (V_0), which lacks an RGBA foreground, we show only the decomposed background for reference. From the visual comparison among V_1, V_2, and "Ours", it is evident that our method, which leverages consistency loss to explicitly model visual effects in the foreground layer, produces: (i) background layers with cleaner removal and less artifacts, (ii) foreground layers with more accurate extraction of transparent visual effects, resulting in re-composited results that are more plausible and realistic.

Quantitative Comparison. To more comprehensively evaluate the quality of the decomposed foreground, we randomly move/resize the foreground prediction and then re-composite it onto the decomposed background to evaluate the fidelity of the resulting image. Specifically, there are three parameters to randomly adjust: ΔX, ΔY, and ΔS. $\Delta X \in [-0.3, +0.3]$ and $\Delta Y \in [-0.3, +0.3]$ specify the horizontal and vertical location changes as proportions of the input dimensions, while $\Delta S \in [0.5, 1.5]$ specifies a scaling ratio w.r.t. the original size. For each image, we randomly select three parameters and apply the same adjustment to all model variants' foreground prediction. The FID and CLIP-FID of the randomly re-composite images are reported in Table S1. Comparing with other model variants, leveraging consistency loss to explicitly model visual effects in the foreground layer indeed improves re-composition quality.

2. Additional Results for the Mask-Based Object Removal Experiment

Benchmarks Details. Here, we provide more details for the mask-based object removal benchmarks used to calculate the metrics presented in Table 2 of the main paper.

- RORD [5]: We randomly select 1,029 images from the original test set to reduce data redundancy caused by sampling from the same video. The dataset provides both manually labeled loose masks and tight masks for the real-world object removal task. The average area of the loose mask is 3.70 times that of the tight mask in each image. As shown in Fig. S4, RORD includes diverse indoor and outdoor scenes, featuring removal of various target objects with soft shadows or reflections in real-world settings.
- MULAN [7]: We randomly select 1,000 images from MULAN-COCO for our evaluation. For each image, the dataset provides multiple object layers in RGBA format, and we select the object in the top most layer as the removal target. To reduce hallucination problems in traditional inpainting methods caused by tight object mask, we further dilate the object mask by 10 pixels. As shown in Fig. S5, MULAN data also includes diverse indoor and outdoor scenes, featuring object removal in more cluttered settings.

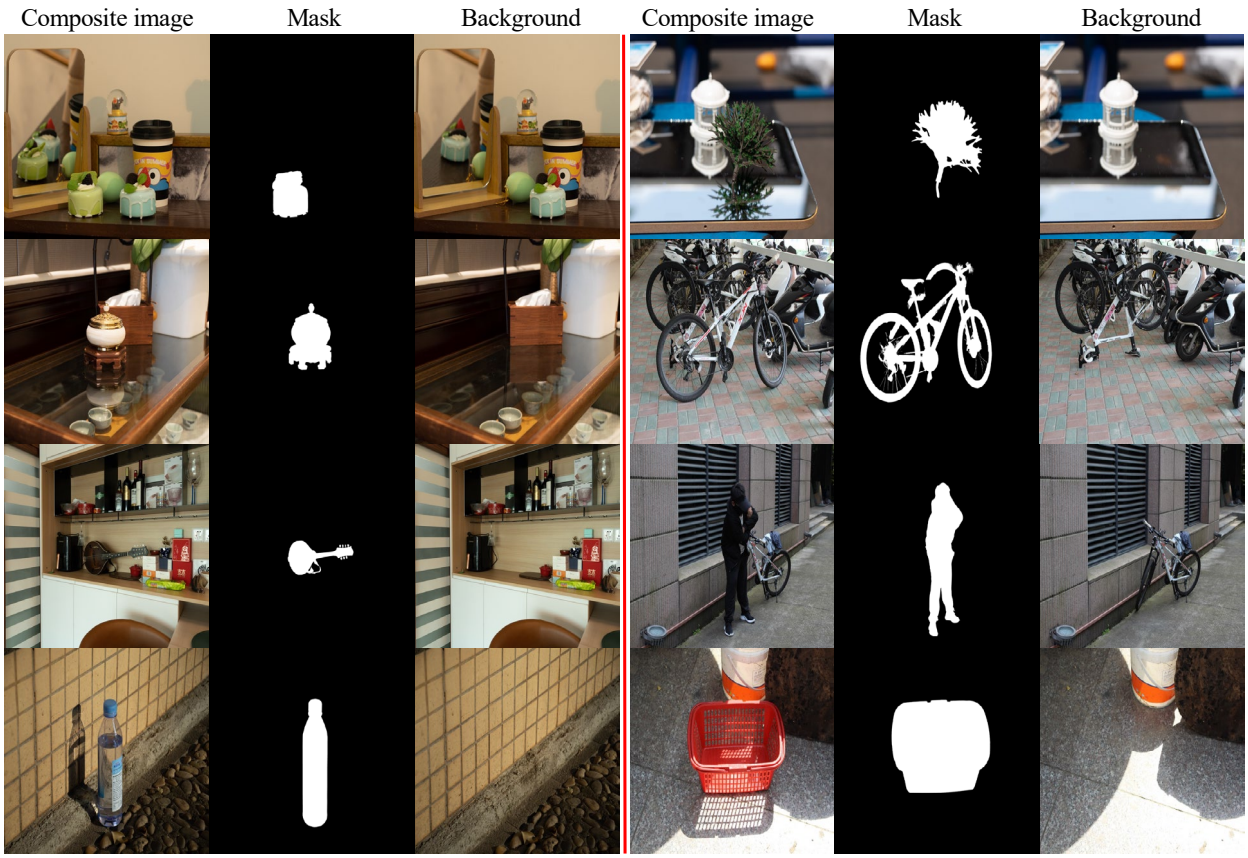

Figure S2. Examples of the real-world camera-captured image pairs used in our model training and ablation study. Images are captured by a camera in real-world scenes. The masks are manually annotated, indicating the target objects to remove.

- DESOBAv2 [3]: There are 750 images in the test set including binary object masks and paired shadow masks. We use the binary object masks as tight mask to input to LAYERDECOMP and merge the object mask and the corresponding shadow mask to create loose mask to input to other inpainting methods. Similarly, to reduce hallucination problems in traditional inpainting methods, the loose masks are further dilated by 10 pixels. The average area of the loose mask is 2.35 times that of the tight mask. As shown in Fig. S6, DESOBAv2 mostly features outdoor scenes with hard object shadows cast on surfaces with different materials and textures, adding more challenges to decompositing the visual effects.

More Visual Results. More visual comparison with ControlNet Inpainting [9], SD-XL Inpainting [4], and PowerPaint [10] on the three public benchmarks is provided in Fig. S4, Fig. S5, and Fig. S6. It can be observed that, with the assistance of the loose mask, the three baselines are able to remove most parts of the target object. However, they struggle to eliminate it entirely and face challenges in removing the shadows associated with the target object. Additionally, achieving photorealistic background completion in human plausible style remains a significant challenge. In contrast, our model, using only the tight mask, performs consistently better across a wide range of data sources.

3. Evaluating Real-World Generalization in Visual Effect Removal

To quantitatively evaluate the effectiveness of our model in removing visual effects, we randomly sampled 500 images from the MuLAN dataset. After filtering out images without prominent shadows or reflections, we obtained a final test set of 44 images. A user study was conducted with four participants, who evaluated whether our model successfully generated clean background layers and faithful foreground elements. The model achieved an overall success rate of 78.98%, demonstrating strong generalization capabilities for visual effect removal on real-world data.

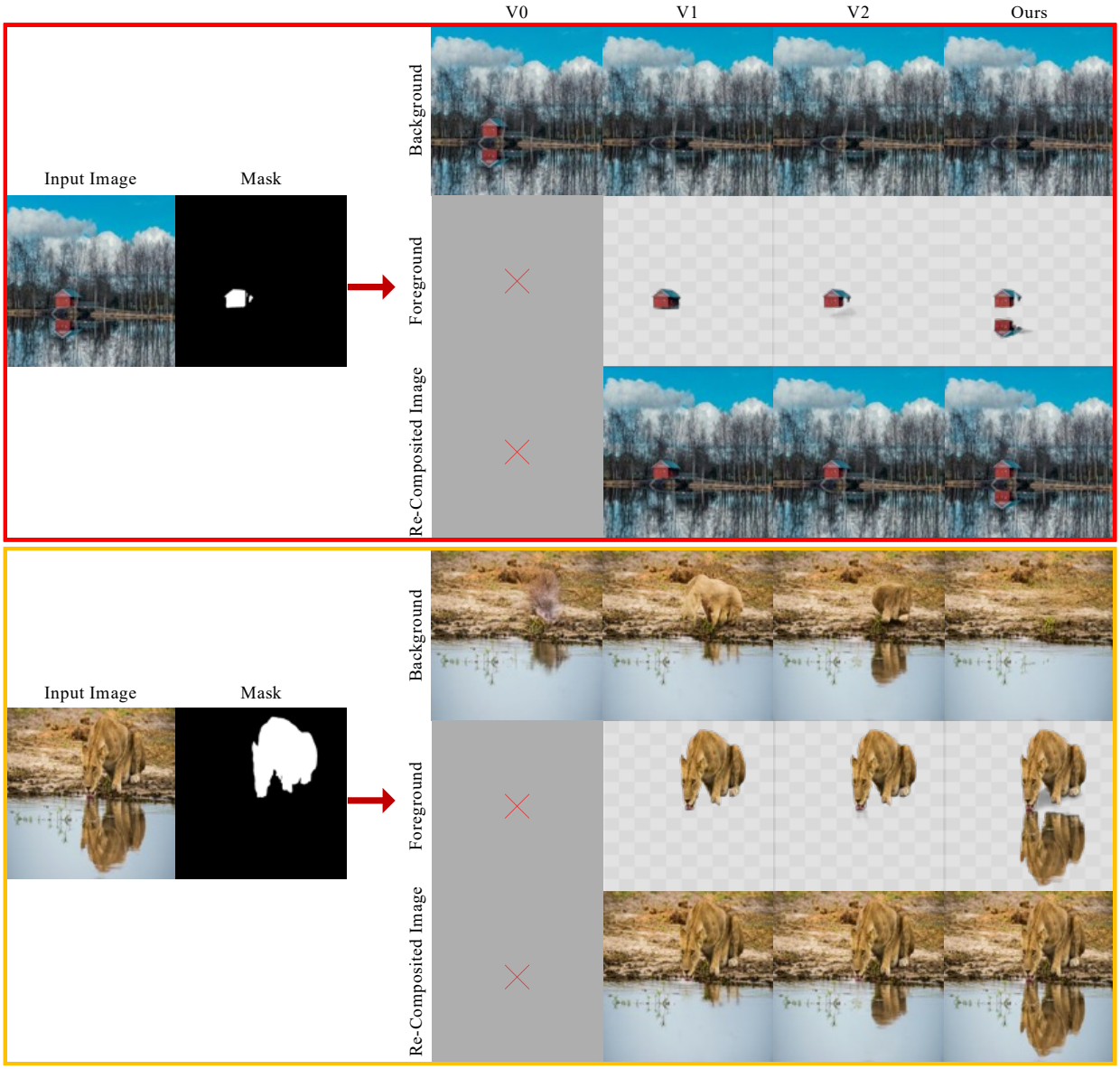

Figure S3. Visualization of results generated by model variants presented in Table 1 of the main paper. Our model can generate higher-quality foreground and background layers and produce more plausible and realistic re-composite results by effectively modeling the visual effects.

4. Additional Results for the Instruction-Driven Object Removal Experiment

Qualitative Comparison. Fig. S7 presents additional comparison results with instruction-driven methods on the object removal task on Emu-Edit Remove Set [1, 6]. Beyond showcasing the superior object removal performance of our model, these results further highlight its enhanced background integrity and completion capabilities.

Quantitative Comparison. We also perform a quantitative comparison with instruction-driven methods, specifically Emu-Edit [6] and OmniGen [8]. Using the released generation results from Emu-Edit Remove Set [1], we evaluate the performance based on FID and CLIP-FID metrics. For a fairer comparison, we use text-based masks as input to our model. As shown in Table S2, our model outperforms existing approaches by a large margin.

5. More Image Layer Decomposition Results from LAYERDECOMP

As shown in Fig. S8, we provide comprehensive visualization results from various data sources, including web images, public datasets, and the held-out test set. These results demonstrate that our model is robust across diverse scenarios.

References

[1] Facebook AI. Emu edit test set generations. https://huggingface.co/datasets/facebook/emu_edit_test_set_generations, 2024. Accessed: 2024-11-20. 3

[2] 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. In *International Conference on Learning Representations (ICLR)*, 2024. 8

[3] Qingyang Liu, Jianting Wang, and Li Niu. Desobav2: Towards large-scale real-world dataset for shadow generation. *arXiv preprint arXiv:2308.09972*, 2023. 2

[4] 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. 2, 5, 6, 7

[5] Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Won Jung, and Sung-Jea Ko. Rord: A real-world object removal dataset. In *BMVC*, page 542, 2022. 1

[6] 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, 3, 8

[7] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22413–22422, 2024. 1

[8] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 1, 3, 8

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 5, 6, 7

[10] 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. *arXiv preprint arXiv:2312.03594*, 2023. 2, 5, 6, 7

Figure S4. **Object removal - comparison with mask-based methods on the RORD dataset.** Our model, using tight input masks, generates more visually plausible results with fewer artifacts compared to ControlNet Inpainting [9], SD-XL Inpainting [4], and PowerPaint [10], which all require loose mask input.

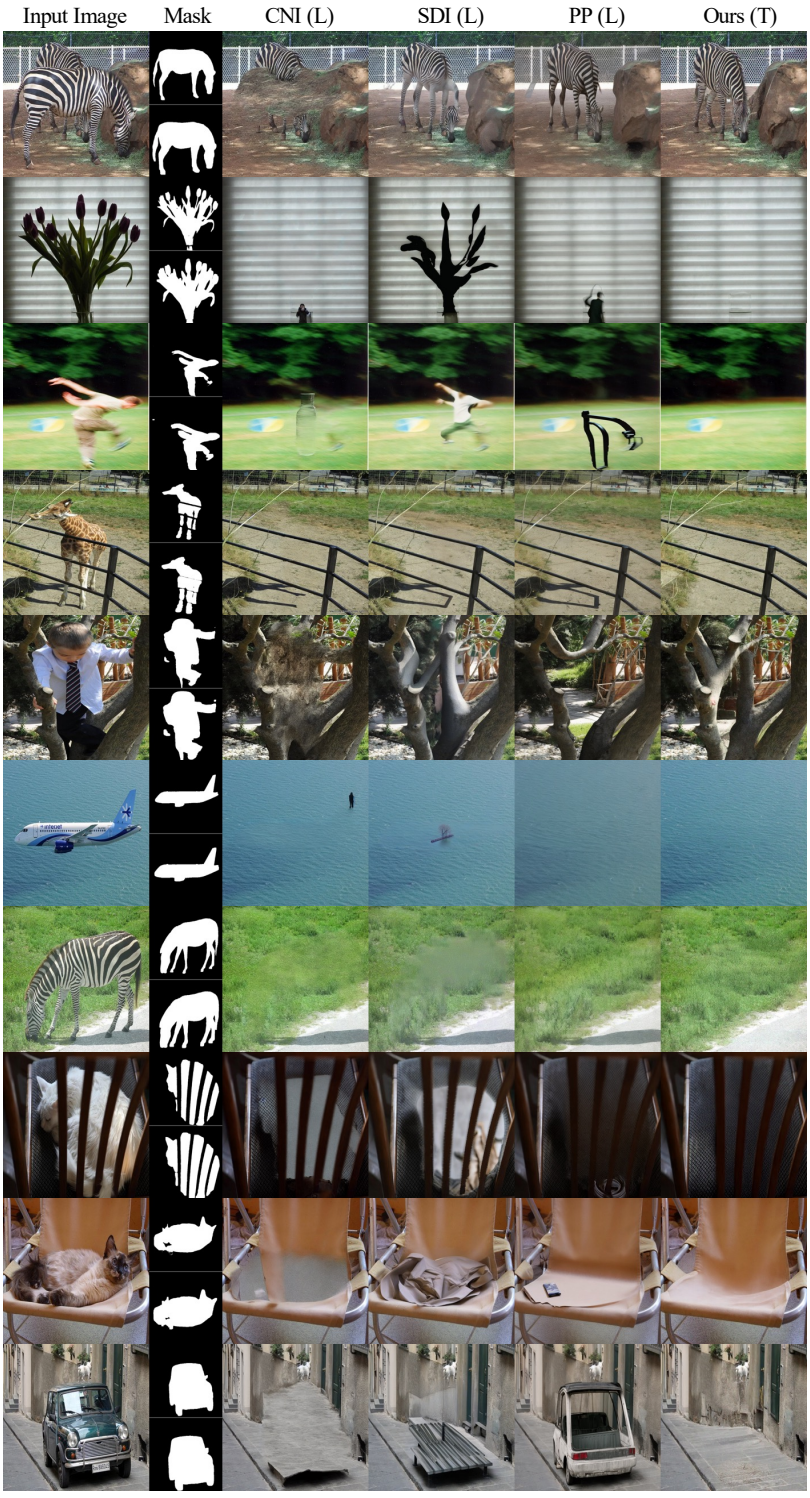

Figure S5. **Object removal - comparison with mask-based methods on the MULAN dataset.** Our model, using tight input masks, generates more visually plausible results with fewer artifacts compared to ControlNet Inpainting [9], SD-XL Inpainting [4], and PowerPaint [10], which all require loose mask input.

Figure S6. **Object removal - comparison with mask-based methods on the DESOBAv2 dataset.** Our model, using tight input masks, generates more visually plausible results with fewer artifacts compared to ControlNet Inpainting [9], SD-XL Inpainting [4], and PowerPaint [10], which all require loose mask input.

Figure S7. **Object removal - more comparison with instruction-driven methods on Emu-Edit [6] removal set.** Combining with a text-based grounding method, our model can effectively remove target objects and preserve background integrity, while existing instruction-based editing methods, such as Emu-Edit [6], MGIE [2], and OmniGen [8], may struggle to fully remove the target or maintain background consistency.

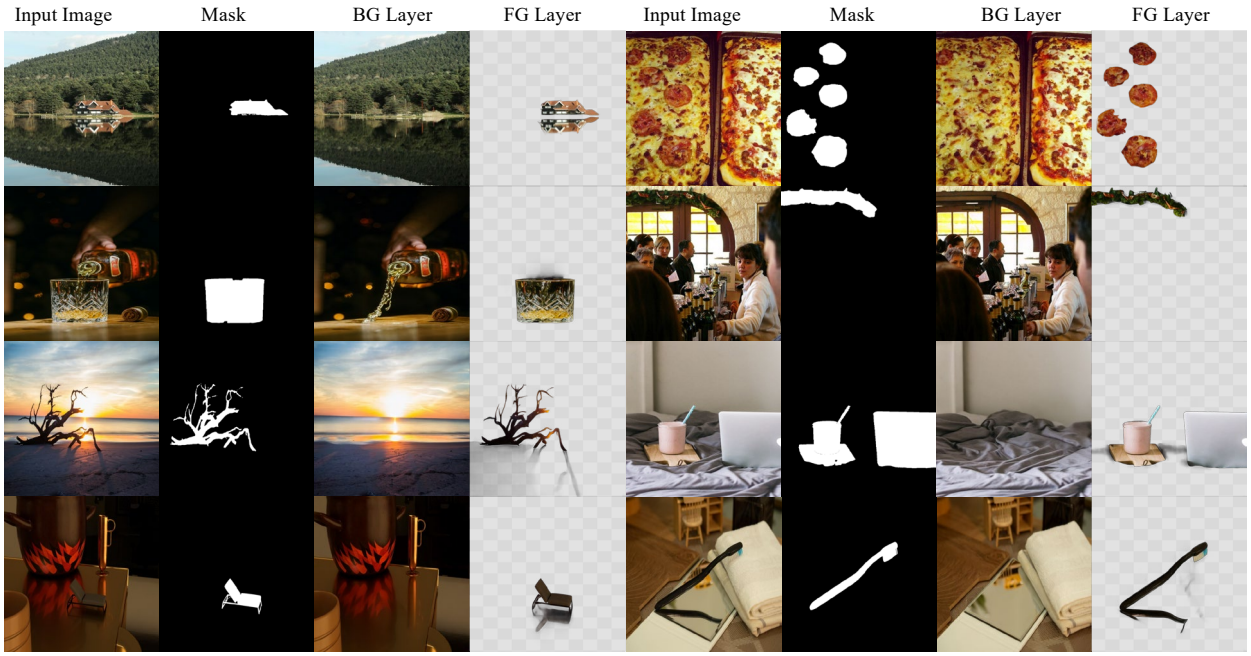

Figure S8. Additional image decomposition results of our model on public benchmarks and web images. These results demonstrate the robustness of our model across diverse data sources.