From Inpainting to Layer Decomposition: Repurposing Generative Inpainting Models for Image Layer Decomposition
Supplementary Material

Contents

A. Similarity Between Outpainting and Foreground Generation

As shown in Fig. 3, foreground generation closely resembles outpainting: in many cases, the target object is partially occluded, requiring the model to reconstruct missing regions beyond simple copy-and-paste. This demands the model to outpaint the foreground using the incomplete mask as a structural guide.

B. Technical Contributions and Differences from the Prior Approach

Our key contribution compared with LayerDecomp is a new perspective on layer decomposition and a data- and parameter-efficient way to achieve it. LayerDecomp requires full finetuning on a large-scale high-quality dataset, making it impractical for general users with limited resources. In contrast, our method introduces lightweight LoRA-based finetuning strategy with multi-modal context inputs. This allows us to repurpose existing pre-trained inpainting models for layer decomposition without large-scale training and democratize the capability to a broader range of users.

C. Application Limitations Related to Lighting and Shadows

It is true that the current model struggles with complex lighting conditions. However, we attribute this limitation to the lack of such samples in our training data, rather than to an intrinsic flaw of the method. Our main contribution is a new perspective and adaptation strategy that repurposes an inpainting model for layer decomposition in a data- and parameter-efficient manner. In fact, even without explicit training on these effects, our model already shows some ability to handle shadows and reflections, as illustrated in Figure 1 of the main paper and Figure 2 of the supplementary material.

D. Example prompts for generating foreground materials with LayerDiffuse

animal	common object	complex object	machine	human
dog	apple	coral	microwave	tall man
cat	banana	brain	refrigerator	slim man
cow	robot	pinecone	freezer	young woman
...	...	...	...	...
profession	**household**	**clothes**	**sci-fi objects**	**traffic objects**
businessman	sofa	shirt	laser gun	car
policeman	armchair	pants	plasma rifle	truch
fireman	loveseat	shorts	ion blaster	motorcycle
...	...	...	...	...

Table 1. Prompt categories and examples used to generate foreground object layers with LayerDiffuse.

In our data curation process, we use LayerDiffuse to generate synthetic foreground layers. Prompts are categorized into major themes, with ChatGPT-4o generating examples for each. We collected around 300 prompts per category. Sample prompts are shown in Table 1.

E. Additional Data Curation Details

We collect 15K real and 15K synthetic foreground objects with corresponding masks, along with 100K background images. To construct each sample, we sequentially overlay 1–3 foregrounds onto a background. When overlaps occur, we update the underlying foreground mask by subtracting the overlapping region. This process yields 100K training tuples of (composite image, target foreground, mask, target background).

F. More Qualitative Examples

In Figure 1, we present examples in the user study where our foreground layers are compared against two matting methods: Matting-Anything and DiffMatte. Our method superior in preserving the detail shapes and recovering occlusions.

In Figure 2, we present more examples of object removal as the same setting in the main paper, with comparisons against SD-XL Inpainting, PowerPaint and Flux.1-Fill-dev.

G. Effectiveness of Pre-trained VAE Encoding for Multi-Modal Inputs

In Figure 4, we visually confirm that the pre-trained VAE of the FLUX model can effectively encode and decode edge maps, segmentation maps, and depth maps, modalities used in our multi-modal context input, without noticeable detail loss. This validates our design choice of using the pre-trained VAE to tokenize these modalities efficiently.

H. Failure Cases

We show examples of failure cases in Figure 5. Our method has challenge in complex cases involving cluttered objects, large objects or hand-object interaction. We believe the performance can be further enhanced with better training data covering these challenging cases.

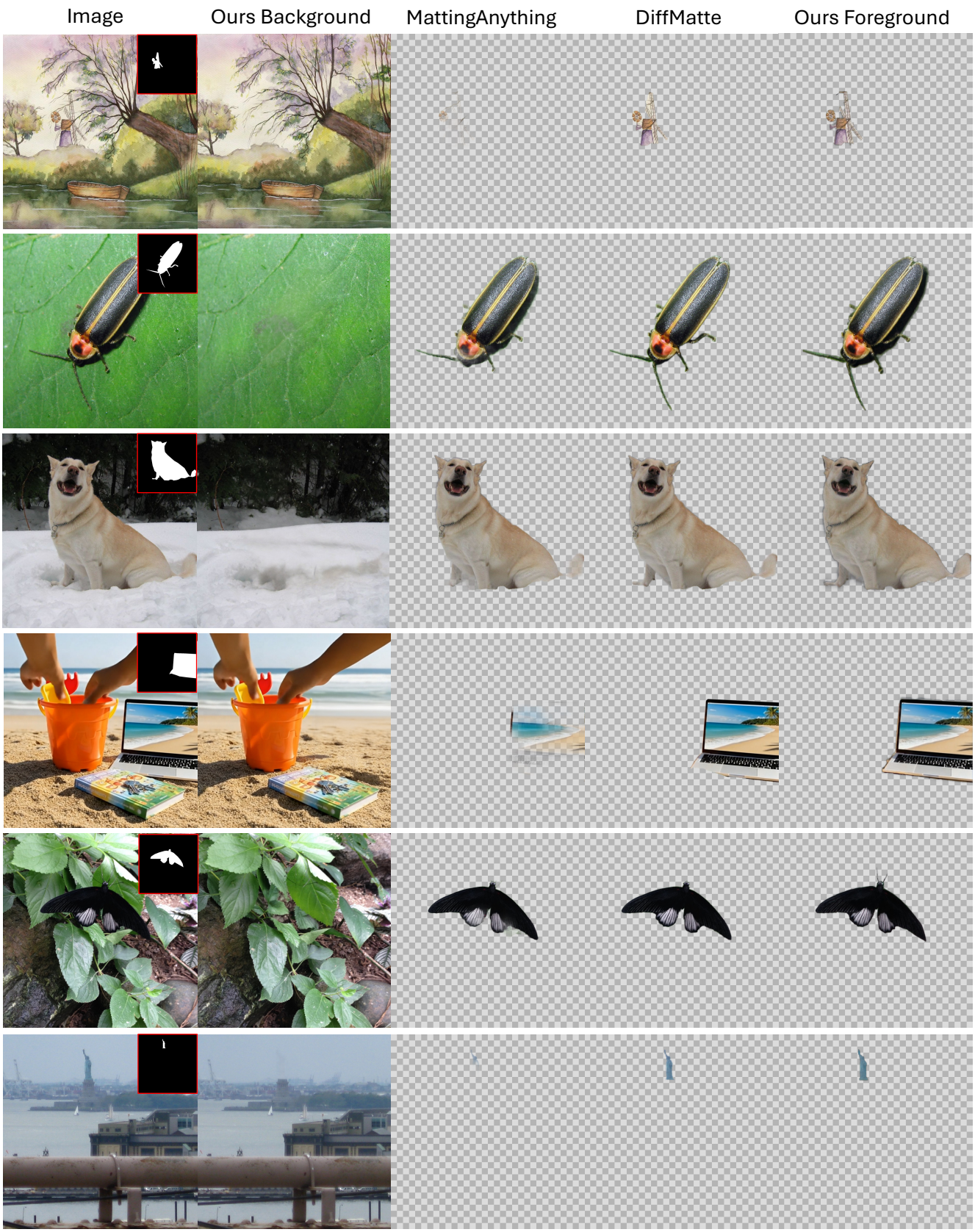

Figure 1. We present additional comparisons of foreground extraction using two matting methods, Matting-Anything and DiffMatte, both of which produce RGBA foreground layers.

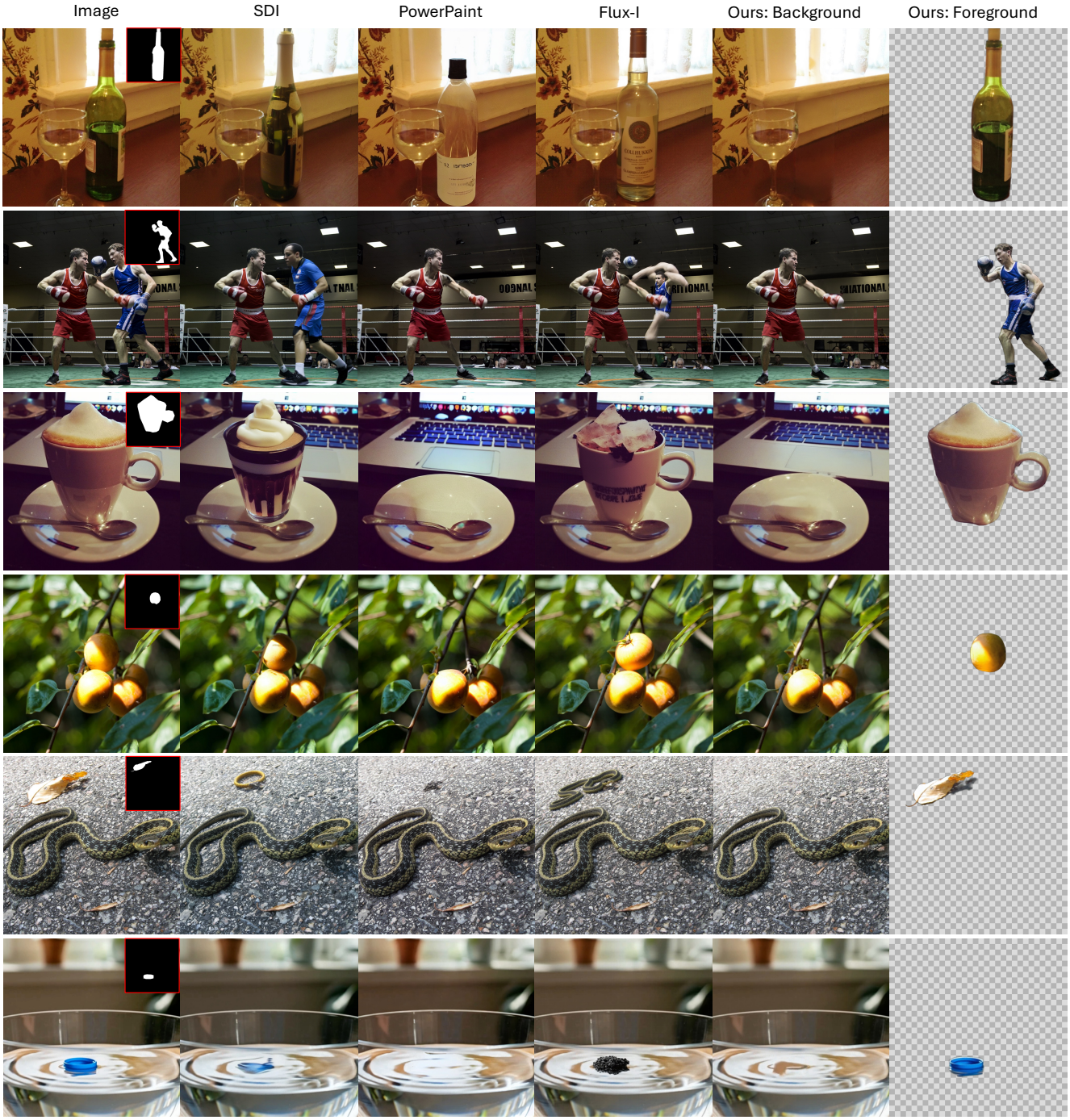

Figure 2. We present additional comparisons of object removal as the same setting in the main paper.

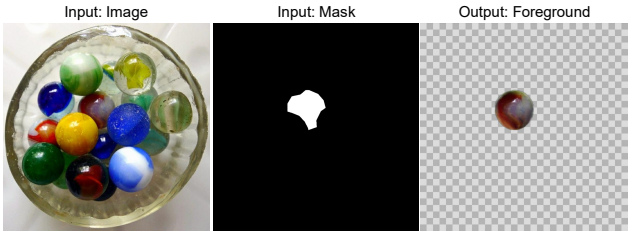

Figure 3. Similarity Between Outpainting and Foreground Generation

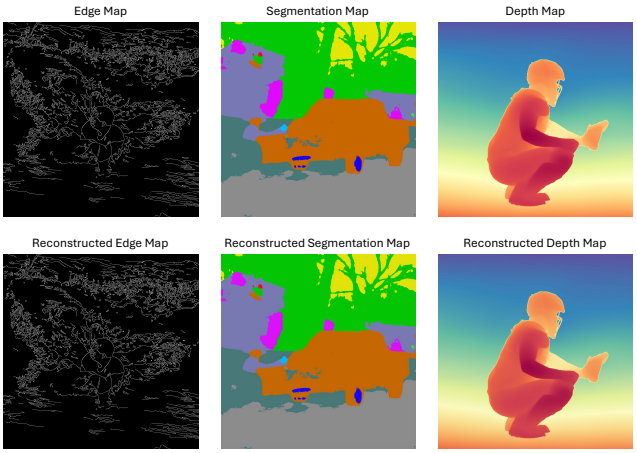

Figure 4. Visual results showing the pre-trained FLUX VAE's ability to reconstruct various modalities, edge map, segmentation map, and depth map, used in our multi-modal context.

Figure 5. We show examples of failure cases. The model tends to fail on complex images that involve cluttered objects, large objects with occlusion, and hand-object interaction.