Appendices

A. Overview

The following sections are covered in this supplementary material to support our main paper:

- Quantitative results of our method and the baselines;
- Real-world dataset collection;
- Comparison with prior image-blending works;
- Our model's robustness against low-quality images;
- Additional qualitative examples;
- Implementation details.

B. Quantitative Results

As mentioned in the paper, to demonstrate the realism and faithfulness of our model based on human perception, we conduct a user study with a real-world dataset and collect quantitative results.

To further support our conclusion, we prepared a synthetic test dataset based on Pixabay Dataset which is generated in a similar way to how we prepared the training dataset for our framework. The main difference is that we apply larger spatial perturbations on the input object. More specifically, the object is randomly rotated within the range $[-\theta, \theta]$ where $\theta = 40°$. We apply this change to better evaluate the model's ability to correct large geometric inconsistencies.

We test our model and two baseline methods (BLIP [4] and SDEdit [5]) on 1500 images randomly chosen from the synthetic test dataset. The baselines are trained on the same pretrained diffusion model, synthetic dataset, and using the same data augmentation method as ours. For SDEdit we use a noise strength of 0.85, enabling it to apply larger spatial transformations, which better adapts to the synthetic test dataset. In Tab. 1, we use FID [2] as a measurement of fidelity, and LPIPS [12] to measure the feature distance between the generation and ground truth. We also employ a modified CLIP score [1, 7] to measure semantic similarity between the given and generated object. The column *Crop* indicates whether we compare the performance with a cropped square patch that covers the generated area. Focusing on the cropped region, we can better evaluate the generation quality; using full image, we can assess the matching performance between the generated area and the background. As shown in this table, our model achieves in all cases the best performance in fidelity and preservation. Despite using a combination of metrics well suited to our task, we observe there are still limitations in these evaluation methods. For example, they cannot measure the correctness of the geometric transformation applied to the object. We leave the design of a better metric for generative object compositing as future work.

C. Real-World Dataset Collection

Fig. 1 depicts the real dataset labeling process. The bounding box annotations obtained by this tool are used when generating the result images for our user study. We show that this annotation process directly corresponds to the real-world use case where the user edits a pair of images (an object and a background image) for object compositing. Using this interface, the user can first choose an object image and a background image (displayed in the left panel and the middle panel). Afterward, the user drags the object to a target location on the background image. Then, the location is determined and the object can also be scaled at will. During the whole process, the right panel will display the copy-and-paste image as a preview. Finally, we extract the bounding box of the object to record the location and scale.

D. Comparison with Prior Image-Blending Works

In addition to BLIP and SDEdit, we further compare with several image blending methods: Deep Image Blending [11] (DIB), GP-GAN [10] and Poisson Blending [6]. To obtain better composition results, we also use SGRNet [3] to generate shadows for the blended objects.

Following the user study in the main paper, we conduct another user study on these prior works using our real dataset, where we display side-by-side composition results to users and ask them to choose the one that looks more realistic. Tab. 2 shows the percentage of users choosing our results when comparing to the baselines. This user study on realism demonstrates that our model outperforms the image blending-based baselines in the task of object composition. We also provide a qualitative example in Fig. 2, where only our model can synthesize a novel view, so the object is able to match the geometry of the background.

E. Robustness against Low-quality Images

We mention in the paper that the traditional compositing pipeline [8, 9] cannot address the problem of geometry harmonization and view synthesis, which are advantages of our generation-based method. Another advantage of our model over the traditional pipeline is that it is robust against low-quality input object images. In the real-world scenario of object compositing, it is a common case that the quality of the input object is not perfect. We categorize low-quality input object images into four scenarios:

- the input object image is blurry due to lens blur;

Method	Crop	FID ↓	LPIPS ↓	CLIP text score ↑	CLIP image score ↑
BLIP	✗	18.3673	0.0923	29.6719	95.5625
SDEdit	✗	17.4963	0.0870	29.6563	96.1250
Ours	✗	**15.8191**	**0.0835**	**29.8594**	**97.0000**
BLIP	✓	28.0690	0.2463	29.0313	91.1250
SDEdit	✓	27.0630	0.2312	29.0625	91.8750
Ours	✓	**24.4719**	**0.2223**	**29.4844**	**93.7500**

Table 1. Quantitative comparison with baselines. We measure the performance of our model against two baselines (BLIP and SDEdit) through FID, LPIPS, and modified CLIP scores. The results further demonstrate the effectiveness of our model in addition to the user study results in the paper. More visual comparisons with baselines on the real dataset are shown in Figs. 5 and 6.

Figure 1. The user interface of the real-world data labeling tool. This figure demonstrates our data collection process closely simulates the real-world use case in object compositing. The real dataset (including object-background pairs and bounding box annotations) collected using this interface is used for our user study. It consists of three panels: 1) the user can select the object from the left panel; 2) the middle panel shows the background image; 3) the right panel previews copy-and-paste results in real time. Users can drag the object to any location in the background as well as alter the scale of the object.

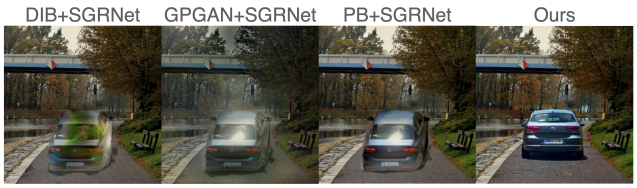

Figure 2. Qualitative comparisons of our model and image blending-based baselines. Shadow is generated by [3] to each baseline. Our model generates the most realistic compositing result since it can predict a new view of the object and do geometry correction.

Method	DIB+SGRNet	GPGAN+SGRNet	PB+SGRNet
Ours	82.93%	84.74%	76.91%

Table 2. A user study on *realism*. Similar to the user study in the main paper, we further compare to three image-blending baselines: 1) Deep Image Blending [11] (DIB); 2) GP-GAN [10] and 3) Poisson Blending [6] (PB); also applying SGRNet [3] for shadow generation. This table shows the percentage of users choosing our results in side-by-side comparisons. The high preference rates demonstrate the advantage of our model over the baselines.

- some parts of the input object are invisible such as when the object is partially occluded;
- the segmentation model fails to extract an accurate segmentation mask of the object, thus the object image includes some background image content; and
- the object is too small and thus has low resolution.

In our self-supervised training scheme, the synthetic training data we collected covers all the above situations so that the content adaptor will not be constrained by the

Figure 3. Robustness. We show our model's robustness against low-quality input objects in the real world. The figure includes three examples. In each example, the top row shows the input object under different conditions: 1) blur, 2) partial occlusion, 3) inaccurate segmentation, and 4) low resolution. The bottom row shows the compositing results corresponding to the input above. Compared to the original input object, our model produces similar high-quality generation results under all conditions.

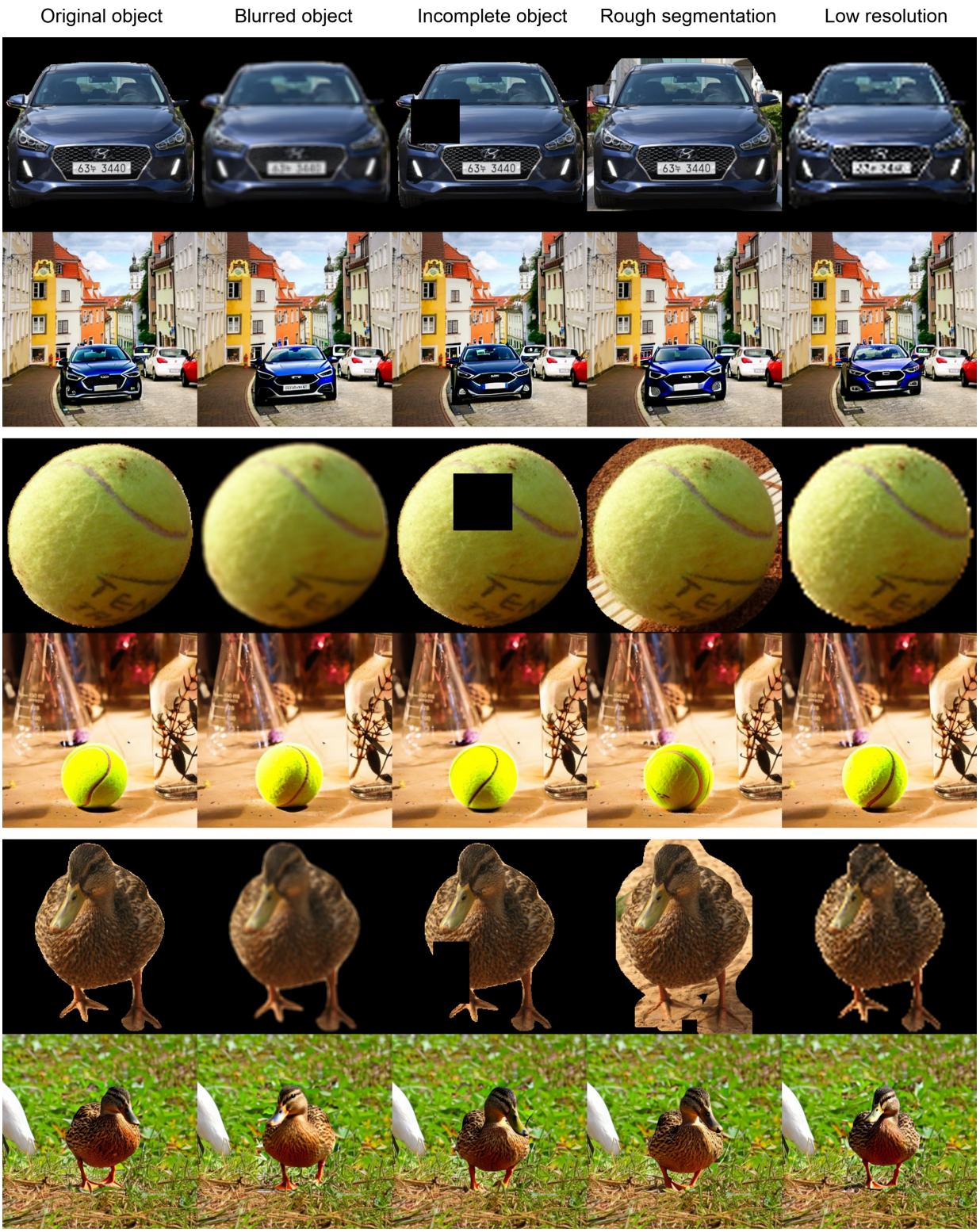

Figure 4. Robustness. We show our model's robustness against low-quality input objects in the real world. The figure includes three examples. In each example, the top row shows the input object under different conditions: 1) blur, 2) partial occlusion, 3) inaccurate segmentation, and 4) low resolution. The bottom row shows the compositing results corresponding to the input above. Compared to the original input object, our model produces similar high-quality generation results under all conditions.

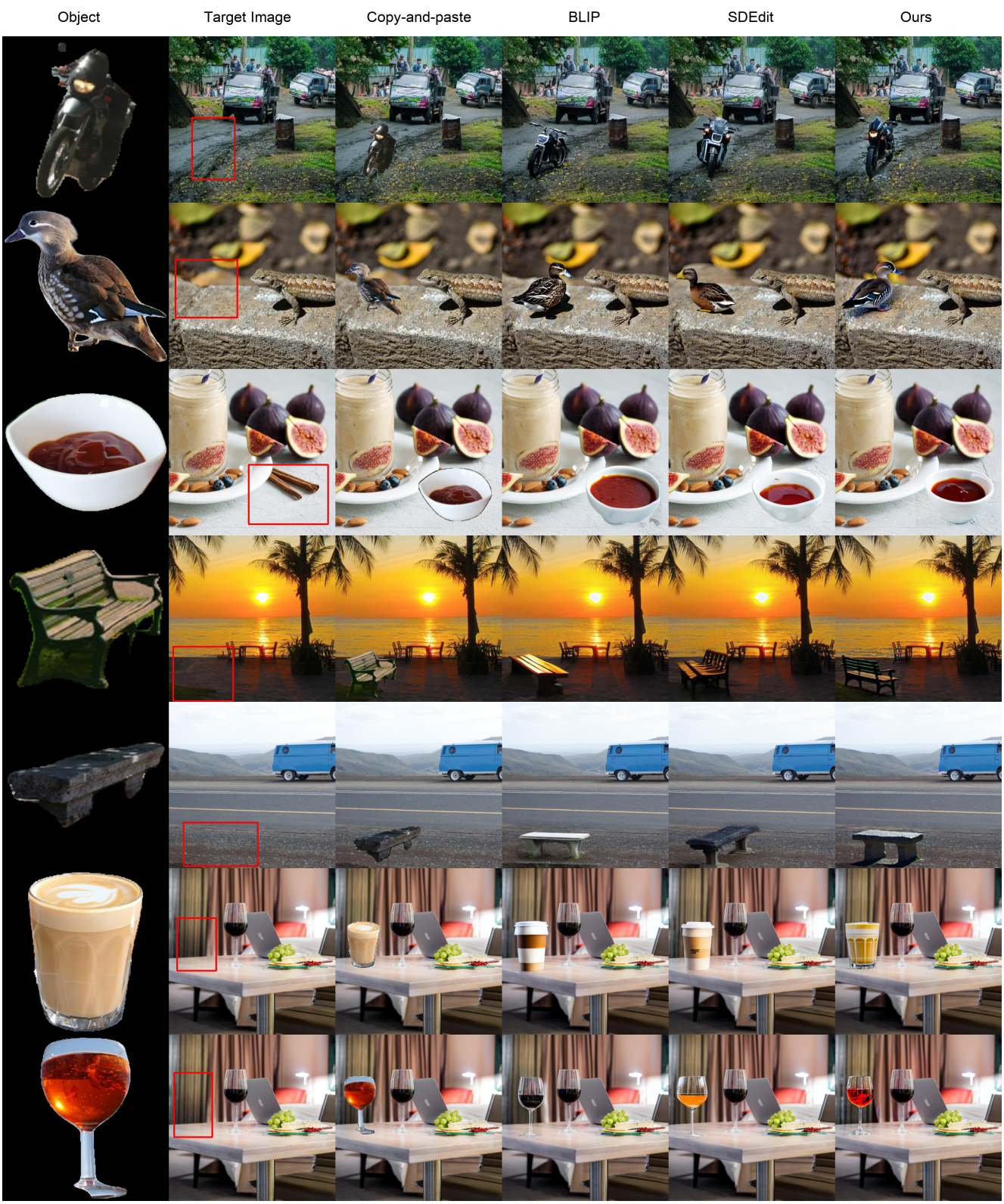

Figure 5. Qualitative comparison with baseline methods on the real-world test dataset. Our model better *preserves a similar appearance* to the reference object (the first column) while generating *realistic* content that is more consistent with the background.

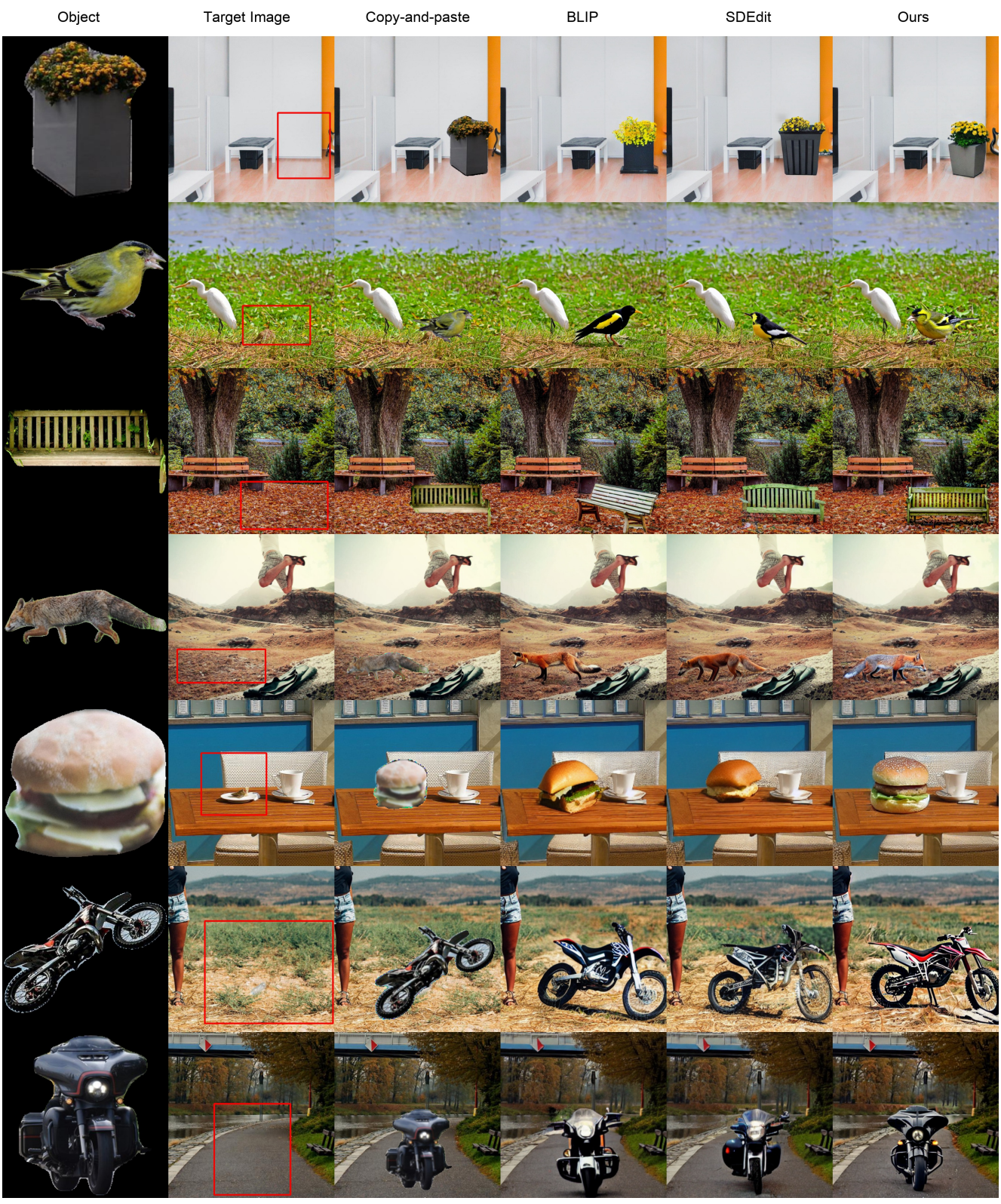

Figure 6. Qualitative comparison with baseline methods on the real-world test dataset. Our model better *preserves a similar appearance* to the reference object (the first column) while generating *realistic* content that is more consistent with the background.

flaws of the low-level features. In Figs. 3 and 4 we show examples where low-quality input objects are provided to simulate the aforementioned four scenarios. It is illustrated in the figures that despite the flaws in the input objects, our model is robust to such extreme cases and is still able to generate realistic content.

F. Additional Qualitative Examples

In addition to the qualitative results in the paper, we show more visual comparisons with baselines in Figs. 5 and 6.

G. Implementation Details

Content Adaptor. In the pretraining of the content adaptor, we use a pretrained ViT-L/14 image encoder from [7]. This image encoder has been trimmed and only 6 of its 12 attention blocks are kept. We apply this change to better preserve the details of the input object. In the adaptor, we use one attention layer with 8 heads. Its embedding dimension is 768.
Diffusion steps. We use $t = 100$ time steps when generating images for the user study and for all qualitative results; $t = 50$ is used when testing our model and baselines on the synthetic test data mentioned in Sec. B.
Baselines from the main paper. In the second baseline from the main paper, we integrate both BLIP and SDEdit to a stable diffusion model. When only using SDEdit, the generated object will be limited to follow its original shape and pose, so it is harder to match to the background geometry. Combining BLIP and SDEdit improves the performance (*e.g.* view synthesis) since both high-level semantic data and low-level texture features are included. To implement this, we use the semantic text embedding obtained by BLIP (from the input image) as context in attention blocks, and insert spatial guidance (the copy-and-paste image) during the denoising stage.

References

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 914–922, 2022. 1, 2

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[5] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[6] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 1, 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 7

[8] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, pages 240–256. Springer, 2022. 1

[9] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 1

[10] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019. 1, 2

[11] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 231–240, 2020. 1, 2

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1