IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation

Supplementary Material

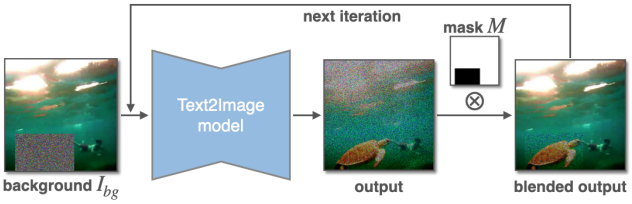


Figure 9. Illustration of the background-blending process.

1. Overview

The following sections will be discussed to further support our paper:

- The background-blending process;
- Mask types (used for shape-guided generation);
- Ablation study on two alternative architectures;
- Additional results of shape-guided generation;
- Additional qualitative comparison results.
- Additional comparisons with AnyDoor [6];
- Failure cases;

2. Background Blending

This process is illustrated in Fig. 9. At each denoising step, the background area of the denoised latent is masked and blended with unmasked area from the clean background (intuitively, the model is only denoising the foreground).

3. Mask Types

As discussed in Sec. 3.2, to enable more user control, we define four levels of coarse masks, including the bounding box mask. Fig. 10 shows all the mask types. As the coarse level increases (from mask 1 to mask 4), the model has more freedom to generate the object.

4. Ablation Study on Alternative Architectures

When making efforts for better identity preservation, we also explore two alternative architectures (Fig. 11) that are more intuitive to inject object features (due to the page limitation, they are removed from the main paper): 1) concatenation and 2) ControlNet [63]. To provide extra features in this two pipelines, a naive idea is to use the same segmented object I_{obj} as the additional input. However, both the structures of concatenation and ControlNet will result in a spatial correspondence between the output and the additional input (*i.e.*, the generated object tends to have the same size and position as the input), and using I_{obj} which is much larger than the mask M destroys such correspondence. For this reason, we use I^*_{obj}, the *inserted object* image as the additional hint to provide extra features, where the cropped and resized object I_{obj} is fitted in the mask area of the background image I_{bg}. To replace the text encoder branch, we use a combination of a CLIP encoder (ViT-L/14) and an adapter as the image encoder, fine-tuned together with the UNet backbone following the sequential collaborative training strategy discussed in Sec. 3.4. Furthermore, the two pipelines are trained on the same datasets (Pixabay and the video datasets) as our proposed model in the second stage.

4.1. Concatenation

The first architecture is illustrated in Fig. 11a. An additional feature injection branch is added for the purpose of better identity preservation: I^*_{obj} is concatenated with the background image I_{bg}. After this modification, the UNet encoder has 8 channels, where the extra 4 channels are initialized as 0.0 at the start of the training.

4.2. ControlNet

The second architecture is illustrated in Fig. 11b. ControlNet is another structure to enhance spatial conditioning control, such as depth maps, Canny edges, sketches and human poses. In this pipeline, the extra inputs are fed into a trainable copy of the original UNet encoder to learn the condition. In our task of generative object compositing, we use the concatenation of the inserted object I^*_{obj} and a mask $1 - M$ indicating the area to generate the object.

4.3. Quantitative Comparison

To quantize the effects of these two architectures, an evaluation is conducted on the DreamBooth dataset, just as in Sec. 4.3. Tab. 5 shows the results, where "Baseline" is setting 3 in the ablation study of the main paper (Sec. 4.7). Our model outperforms the rest pipelines in all three metrics that measure identity preservation, demonstrating the effectiveness of IMPRINT in memorizing object details.

To further assess the compositing effects, we perform another user study with the same configuration as in the main paper (Sec. 4.5), comparing the realism and fidelity of our results against the concatenation pipeline and ControlNet pipeline. Tab. 6 displays the user preferences for different frameworks in the two questions. The results validates the superiority of our model in both ID-preserving and compositing.

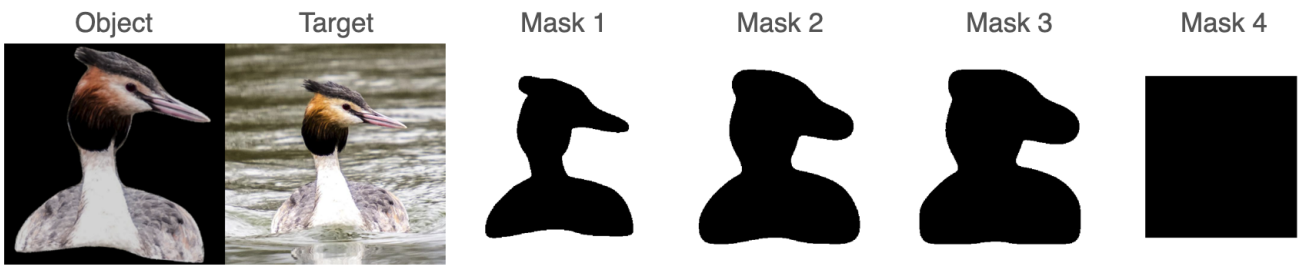
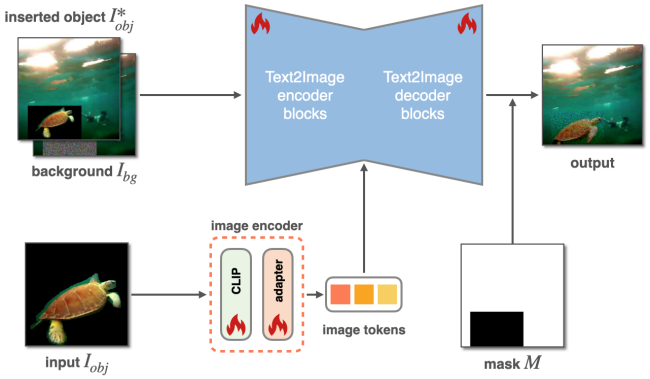
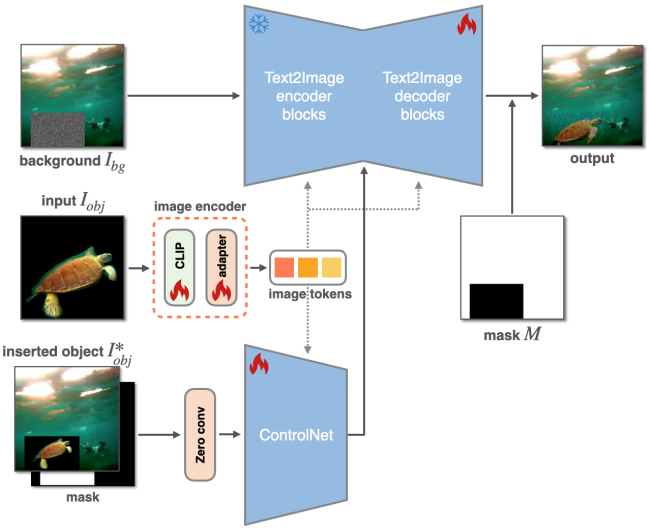


Figure 10. The four types of mask used in the second compositing stage. The generation is constrained in the masked area so the user-provided mask is able to modify the pose, view and shape of the subject.



(a) The concatenation-based pipeline. Aside from the embedding branch, an additional input (the inserted object I^*_{obj}) is concatenated with I_{bg}. Note that the UNet backbone encoder has 8 input channels, where the extra 4 channels are initialized as 0.0.



(b) The ControlNet-based pipeline. In the new ControlNet branch, the concatenation of I^*_{obj} and a mask is given as the additional input.

Figure 11. The pipelines of the two alternative architectures for feature injection: Concatenation and ControlNet.

4.4. Qualitative Comparison

Fig. 12 provides a qualitative comparison between our model and the other two pipelines. Although the nature of structural correspondence in these two pipelines enhances ID preservation, it also constrains their ability to make spatial adjustments. Thus, in the figure their compositing effects are worse than our model (in the first three examples, our outputs have larger pose changes). Moreover, owing to the pretraining stage, our model achieves better performance in keeping details.

Method	CLIP-score↑	DINO-score↑	DreamSim ↓
Baseline	76.6250	39.7837	0.3073
Concat	76.8125	40.3884	0.2945
ControlNet	76.8750	40.1471	0.2984
Ours	**77.0625**	**43.4463**	**0.2898**

Table 5. Quantitative comparison on the DreamBooth test set. *Baseline* refers to setting 3 in the ablation study section of the main paper. Detail preservation is measured and displayed in this table, comparing our proposed model with three different architectures.

	Ours	Concat	Ours	ControlNet
Realism	**50.68**	49.32	**53.38**	46.62
Fidelity	**55.41**	44.59	**54.73**	45.27

Table 6. User study results (in percentage). In the two questions that evaluates reality and similarity, the workers are presented with side-by-side results from different models and are asked to make comparison.

5. Additional Results of Shape-Guided Generation

5.1. Ablation Study

Shape-guidance is an important feature supported by our model that enables more user control. This feature is not independent of our efforts in identity preservation. Instead, the overall performance (realism and fidelity) of shape-guided generation is improved by our pretraining stage, as demonstrated by Tab. 7.

This ablation study is conducted on the video datasets

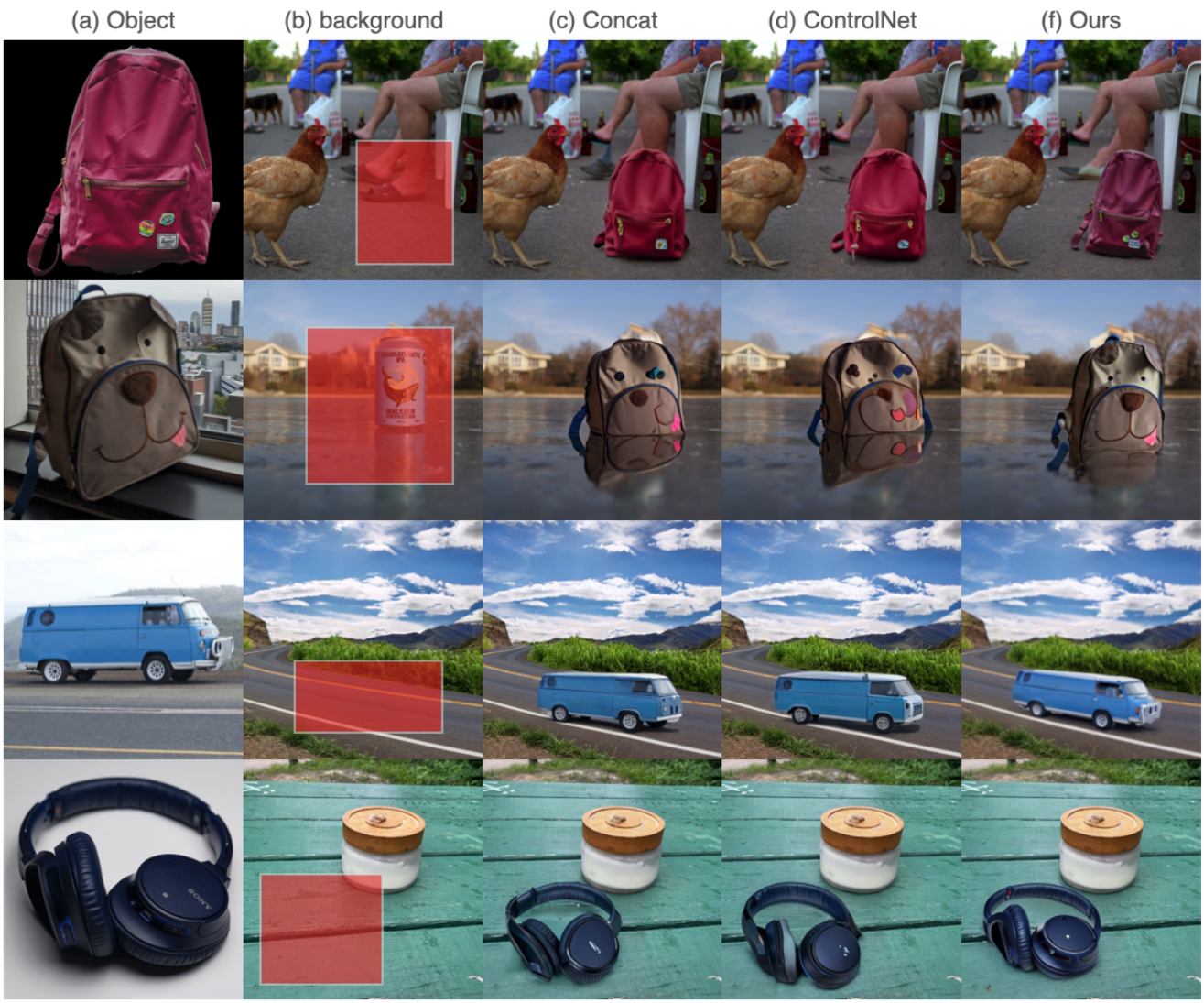


Figure 12. Qualitative comparisons with concatenation-based pipeline and ControlNet-based pipeline. Our model shows stronger ability in geometric adjustments (especially in the first three examples) as well as better performance in identity preservation.

(the test sets). We follow the same data generation pipeline in Sec. 3.3: the target image and the input object are taken from frames I_{n1}, I_{n2} respectively, with $n1 \neq n2$. The guidance mask M is a coarse mask of the object segmentation in the target frame $n1$. We compare our proposed model with another model that is only trained on the second compositing stage. The quantitative results show the improvement of the pretraining stage.

6. Additional Qualitative Results

To further show the advantages of our model against the baseline methods (Paint-by-Example or PbE [58], ObjectStitch or OS [53] and TF-ICON [35]), we include more qualitative results in Fig. 13 and Fig. 14.

Method	FID ↓	CLIP-score↑	DINO-score↑	DreamSim ↓
No PRE	70.0528	91.5625	83.8687	0.1723
PRE	**59.6255**	**91.9375**	**84.7372**	**0.1589**

Table 7. Ablation study on the pretraining stage in shape-guided generation. *PRE* means the pretraining. When the pretraining is finished, the model shows stronger capabilities in ID-preserving and realism, highlighting the fact that our pretraining boosts the performance of shape-guided generation.

7. Additional Comparisons with AnyDoor

We provide additional comparisons below using the official implementation of AnyDoor. We observe that IMPRINT

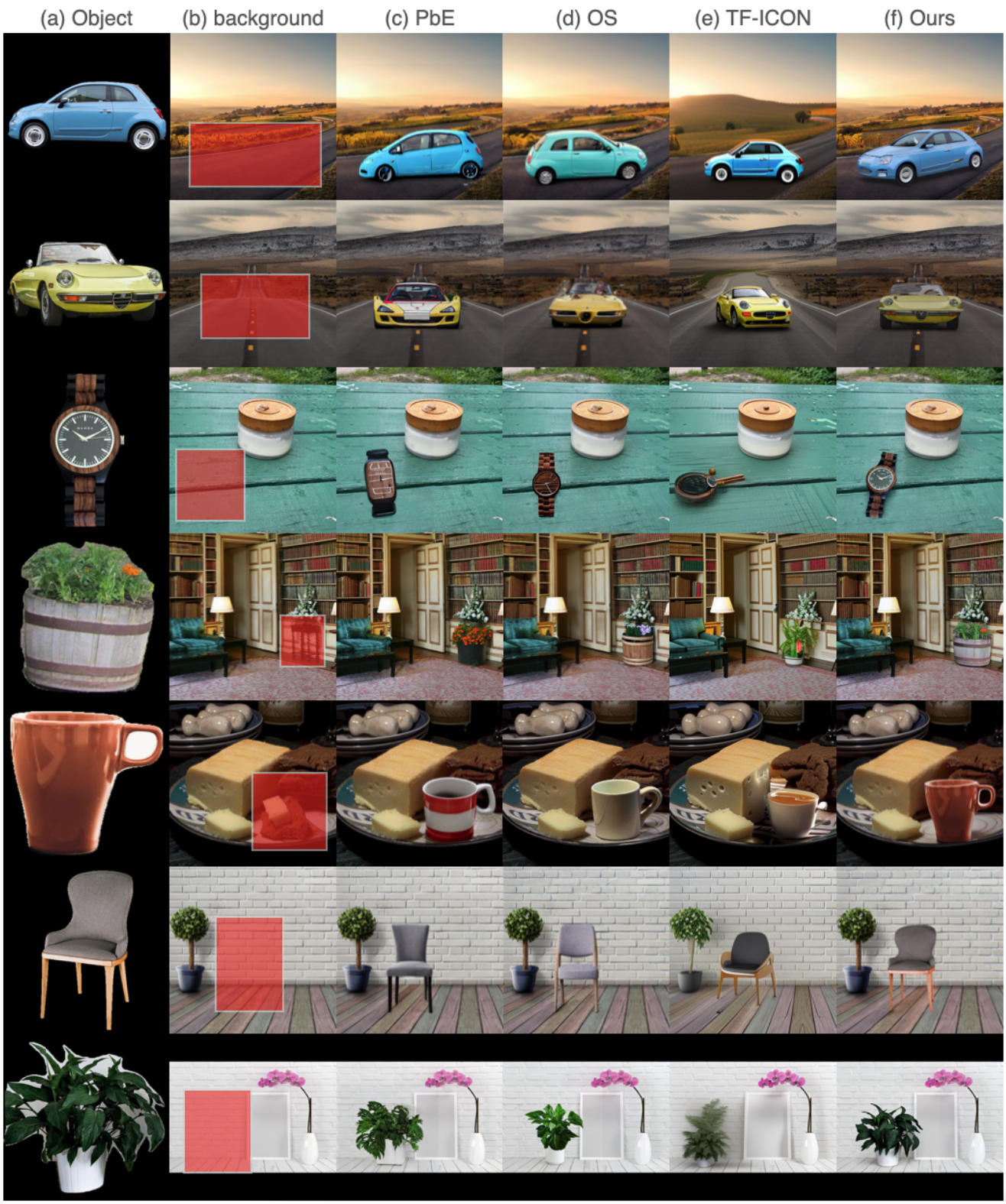


Figure 13. More qualitative comparisons. We compare our proposed model with Paint-by-Example (PbE), ObjectStitch (OS) and TF-ICON. IMPRINT better preserves object identity and the generated object is more consistent with the background.

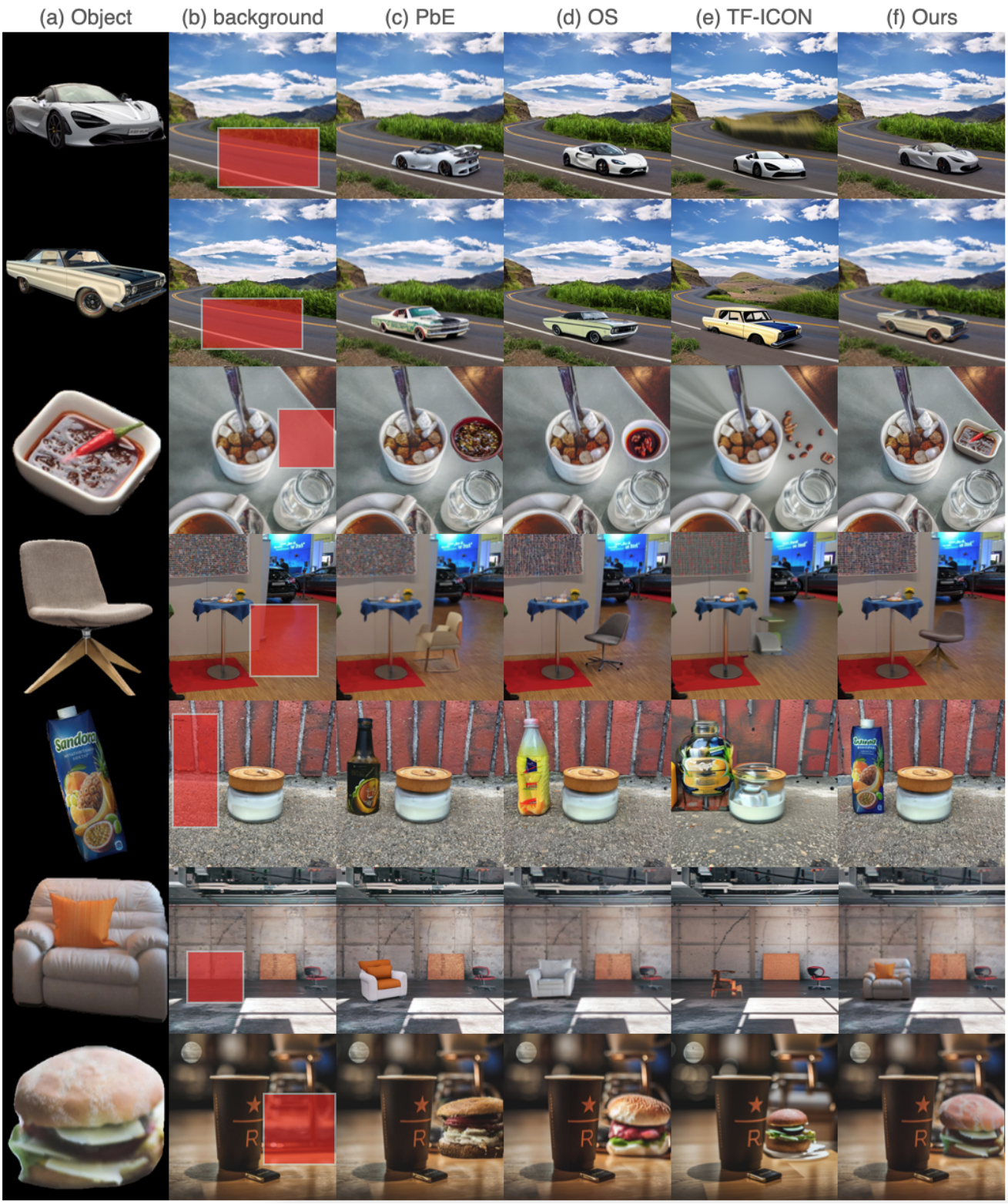


Figure 14. More qualitative comparisons. We compare our proposed model with Paint-by-Example (PbE), ObjectStitch (OS) and TF-ICON. IMPRINT better preserves object identity and the generated object is more consistent with the background.

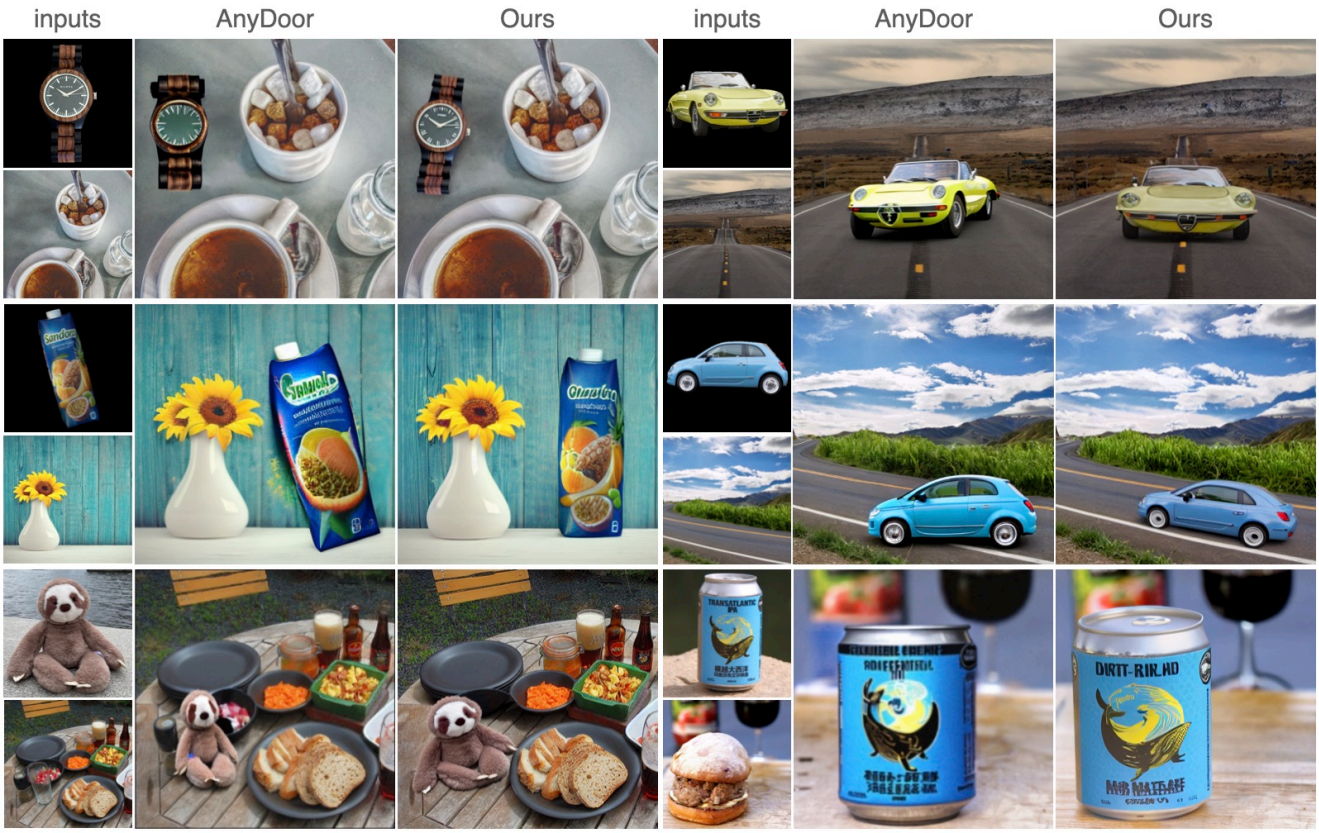


Figure 15. Additional qualitative comparisons with AnyDoor.

Method	CLIP↑	DINO↑
AnyDoor	83.563	83.598
Ours	**85.813**	**86.589**

Method	Realism	Fidelity
AnyDoor	40.71	35.18
Ours	**59.29**	**64.82**

Table 8. Left: Quantitative comparison on the DreamBooth test set. Right: User study results (in percentage).

significantly outperforms AnyDoor in the following experiments:

- We calculate CLIP score and DINO score on the DreamBooth test set to measure the identity preservation (as shown in the left of Tab. 8). Note that to get more accurate results, we masked the background of all generated images when performing the evaluation on the DreamBooth set.
- We conduct a new user study under the same setting as the user study in the main paper (shown in the right of Tab. 8). The users have higher preference rate in our results in both realism and detail preservation.
- In the additional visual comparisons in Fig. 15, our model demonstrates greater adaptability in adjusting the object's pose to match the background, while preserving the details.

8. Failure Cases

Fig. 16 shows the limitations of IMPRINT, as discussed in Sec. 5. In the first example, Though the vehicle is well aligned with the background, its structure is deformed and partially lose its identity due to the large spatial transformation. In the second example, the small logos and texts on the item cannot be fully maintained and exhibits small artifacts, mainly caused by the decoder in Stable Diffusion [43].

Figure 16. Limitations. 1) The first example shows identity loss when making large geometric corrections. The structure of the vehicle changes after generation. 2) The second example shows the degradation of small logos and texts after decoding from the latent space.