IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation

Yizhi Song[1], Zhifei Zhang[2], Zhe Lin[2], Scott Cohen[2], Brian Price[2],
Jianming Zhang[2], Soo Ye Kim[2], He Zhang[2], Wei Xiong[2], Daniel Aliaga[1]
Purdue University[1], Adobe Research[2]

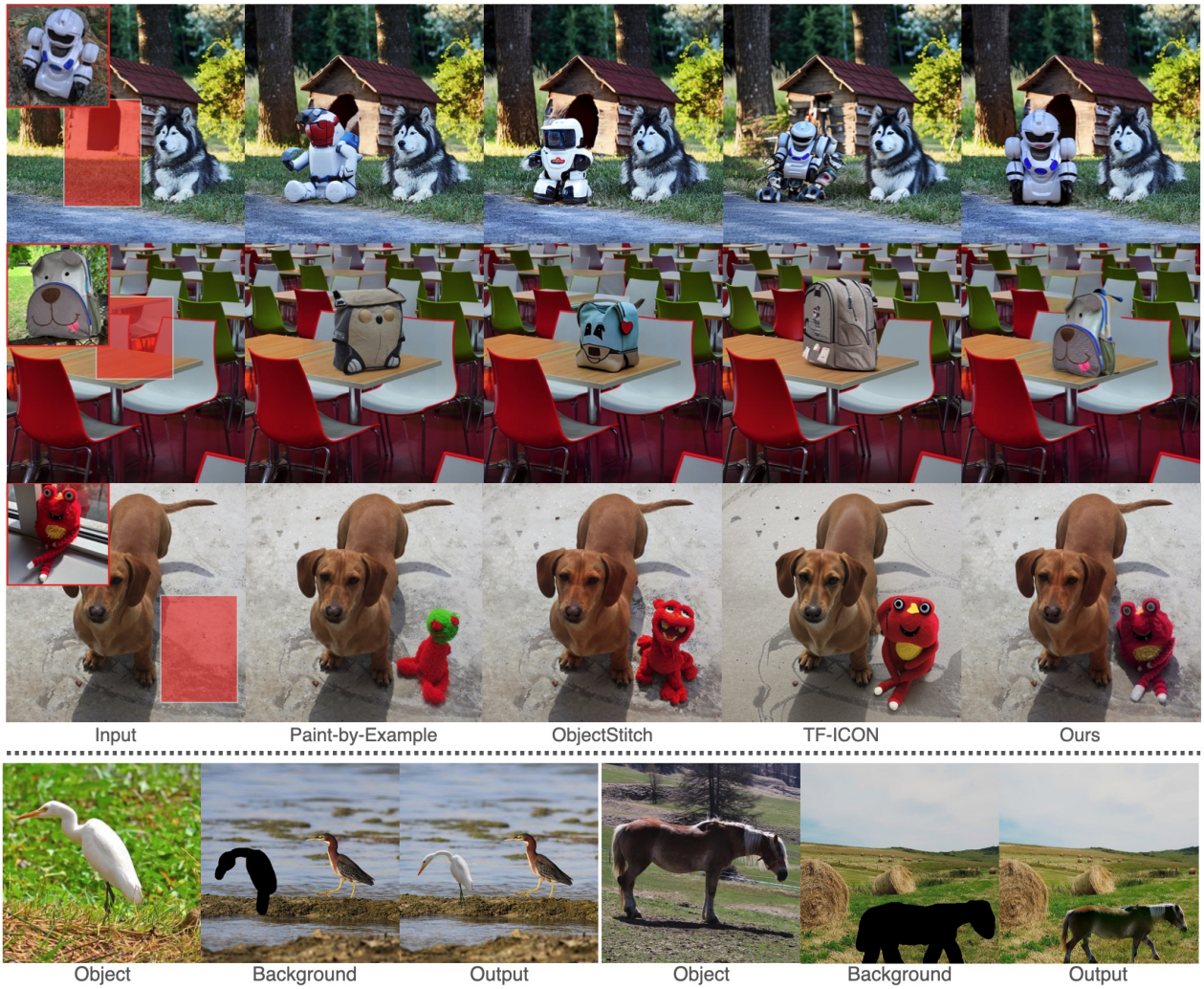


Figure 1. Top: Comparison with three prior works, *i.e.*, Paint-by-Example [58], ObjectStitch [53], and TF-ICON [35]. Our method **IMPRINT** outperforms others in terms of identity preservation and color/geometry harmonization. Bottom: Given a coarse mask, **IMPRINT** can change the pose of the object to follow the shape of the mask.

Abstract

*Generative object compositing emerges as a promising new avenue for compositional image editing. However, the requirement of object identity preservation poses a significant challenge, limiting practical usage of most existing methods. In response, this paper introduces **IMPRINT**, a novel diffusion-based generative model trained with a two-stage learning framework that decouples learning of identity preservation from that of compositing. The first stage is*

targeted for context-agnostic, identity-preserving pretraining of the object encoder, enabling the encoder to learn an embedding that is both view-invariant and conducive to enhanced detail preservation. The subsequent stage leverages this representation to learn seamless harmonization of the object composited to the background. In addition, IMPRINT incorporates a shape-guidance mechanism offering user-directed control over the compositing process. Extensive experiments demonstrate that IMPRINT significantly outperforms existing methods and various baselines on identity preservation and composition quality. Project page: https://song630.github.io/IMPRINT-Project-Page/

1. Introduction

Image compositing, the art of merging a reference object with a background to create a cohesive and realistic image, has witnessed transformative advancements with the advent of diffusion models (DM) [17, 40, 43, 46]. These models have catalyzed the emergence of generative object compositing, a novel task that hinges on two critical aspects: identity (ID) preservation and background harmonization. The goal is to ensure that the object in the composite image retains its identity while adapting its color and geometry for seamless integration with the background. Existing methods [35, 53, 58] demonstrate impressive capabilities in generative compositing; however, they often fail in ID-preservation or context consistency.

Recent works [53, 58], typically struggle with balancing ID preservation and background harmony. While these methods have made strides in spatial adjustments, they predominantly capture categorical rather than detailed information. TF-ICON [35] and two concurrent works [6, 60] have advanced subject fidelity but at the expense of limiting pose and view variations for background integration, thus curtailing their applicability in real-world settings.

To address the trade-off between identity preservation with pose adjustment for background alignment, we introduce IMPRINT, a novel two-stage compositing framework that excels in ID preservation. Diverging from previous works, IMPRINT decouples the compositing process into ID preservation and background alignment stages. The first stage involves a novel context-agnostic ID-preserving training, wherein an image encoder is trained to learn view-invariant features, crucial for detail engraving. The second stage focuses on harmonizing the object with the background, utilizing the robust ID-preserving representation from the first stage. This bifurcation allows for unprecedented fidelity in object detail while facilitating adaptable color and geometry harmonization.

Our contributions can be summarized as follows:

- We introduce a novel context-agnostic ID-preserving training, demonstrating superior appearance preservation through comprehensive experiments.
- Our two-stage framework distinctively separates the tasks of ID preservation and background alignment, enabling realistic compositing effects.
- We incorporate mask control into our model, enhancing shape guidance and generation flexibility.
- We conduct an extensive study on appearance retention, offering insights into various factors influencing identity preservation, *e.g.*, image encoders, multi-view datasets, training strategies, etc.

2. Related Work

2.1. Image Compositing

Image compositing, a pivotal task in image editing applications, aims to insert a foreground object into a background image seamlessly, striving for realism and high fidelity.

Traditionally, image harmonization [11, 21, 24, 57] and image blending [38, 54, 61, 62] focus on color and lighting consistency between the object and the background. However, these approaches fall short in addressing geometric adjustments. The emergence of GANs [13, 22, 65] has inspired numerous studies [3, 5, 30] that employ GANs to address issues of geometric inconsistency, yet are often domain-specific (*e.g.*, indoor scene) and limited in handling complex transformations (*e.g.*, out-of-plane rotation). Shadow synthesis methods [18, 47–49] mainly focus on realistic lighting effects.

With the advent of diffusion models [1, 2, 12, 17, 41–43, 51, 52], recent research has shifted towards unified frameworks encompassing all aspects of image compositing. Methods like [53, 58] employ CLIP-based adapters for leveraging pretrained models, but they struggle in preserving the object's identity due to their focus on high-level semantic representations. While TF-ICON [35] improves fidelity by incorporating noise modeling and composite self-attention injection, it faces limitations in object pose adaptability.

Recent research is increasingly centering on appearance preservation in generative object compositing. Two concurrent works, AnyDoor [6] and ControlCom [60], have made strides in this area. AnyDoor combines DINOv2 [37] and high-frequency filter, and ControlCom introduces a local enhancement module. However, these models have limited spatial correction capabilities. In contrast, our model designs a novel approach that substantially enhances visual consistency of the object while maintaining geometry and color harmonization, representing a significant advancement in the field.

2.2. Subject-Driven Image Generation

Subject-driven image generation, the task of creating a subject within a novel context, often involves customizing

subject attributes based on text prompts. Based on diffusion models, [9, 23] have led to techniques like using placeholder words for object representation, enabling high-fidelity customizations. Subsequent works [25, 34, 44, 45] extend this by fine-tuning pretrained text-to-image models for new concept learning. These advancements have facilitated diverse applications, such as subject swapping [10], open-world generation [28], instruction-based editing [19] using Large Language Models [31, 32, 64] and non-rigid image editing [4]. However, these methods usually require inference-time fine-tuning or multiple subject images, limiting their practicality. In contrast, our framework offers a fast-forward and background-preserving approach that is versatile for a broad spectrum of real-world data.

3. Approach

The proposed object compositing framework, IMPRINT, is summarized in Fig. 2. Formally, given input images of object $I_{obj} \in \mathbb{R}^{H\times W\times 3}$, background $I_{bg} \in \mathbb{R}^{H\times W\times 3}$, and mask $M \in \mathbb{R}^{H\times W}$ that indicates the location and scale for object compositing to the background, we aim to learn a compositing model $\mathcal{C}$ to achieve a composite image $I_{out} = \mathcal{C}(I_{obj}, I_{bg}, M) \in \mathbb{R}^{H\times W\times 3}$. The ideal outcome is an I_{out} that appears visually coherent and natural, *i.e.*, $\mathcal{C}$ should ensure that the composited object retains the identity of I_{obj}, aligns to the geometry of I_{bg}, and blends seamlessly into the background.

In this section, we expand upon our approach. To leverage pretrained text-to-image diffusion models, we design a novel image encoder to replace the text-encoding branch, thus retaining much richer information from the reference object (see Sec. 3.1). Distinct from existing works, our pipeline bifurcates the task into two specialized sub-tasks to concurrently ensure object fidelity and allow for geometric variations. The first stage defines a context-agnostic ID-preserving task, where the image encoder is trained to learn a unified representation of generic objects (Sec. 3.1). The second stage mainly trains the generator for an image compositing task (Sec. 3.2). In addition, we delve into various aspects contributing to the detail retention capability of our framework: Sec. 3.3 discusses the process of paired data collection, and Sec. 3.4 details our training strategy.

3.1. Context-Agnostic ID-preserving Stage

Distinct from prior methods, we introduce a supervised object view reconstruction task as the first stage of the training that help identity preservation. The motivation behind this task is based on the following key observations:

- Existing efforts [6, 35, 60], which successfully improve detail preservation, are limited in geometry harmonization and tend to demonstrate copy-and-paste behavior.
- There is a fundamental trade-off between identity preservation and image compositing: the object is expected to be altered, in terms of color, lighting, and geometry, to better align with the background, while simultaneously, the object's original pose, color tone, and illumination effects are memorized by the model and define its appearance.
- Multi-view data plays a significant role in keeping identity, yet acquiring such datasets is costly. Most large-scale multi-view datasets ([7, 59]) lack sufficient contextual information for compositing; they either lack a background entirely or have a background area that is too limited.

Based on the above insights, we give a formal definition of the task (as depicted in Fig. 2a): given an object of two views I_{v1}, I_{v2} and their associated masks M_{v1}, M_{v2}, the background is removed and the segmented object pairs are denoted as $\hat{I}_{v1} = I_{v1} \bigotimes M_{v1}$, $\hat{I}_{v2} = I_{v2} \bigotimes M_{v2}$. We build a view synthesis model $\mathcal{S} = \{\mathcal{E}_u, \mathcal{G}_\theta\}$ conditioned on $\hat{I}_{v1}$ to generate the target view $\hat{I}_{v2}$, where $\mathcal{E}_u$ is the image encoder and $\mathcal{G}_\theta$ is the UNet backbone parameterized by θ.

Image Encoder $\mathcal{E}_u$ consists of a pretrained DINOv2 [37] and a content adapter following [53]. DINOv2 is a SOTA ViT model outperforming its predecessors [20, 39, 50] which extracts highly expressive visual features for reference-based generation. The content adapter allows the utilization of pretrained T2I models by bridging the domain gap between image and text embedding spaces.

Image Decoder $\mathcal{G}_\theta$ takes the conditional denoising autoencoder $\mathcal{G}_\theta$ from Stable Diffusion [43] and fine-tune its decoder during training. The objective function is defined as (based on [43]):

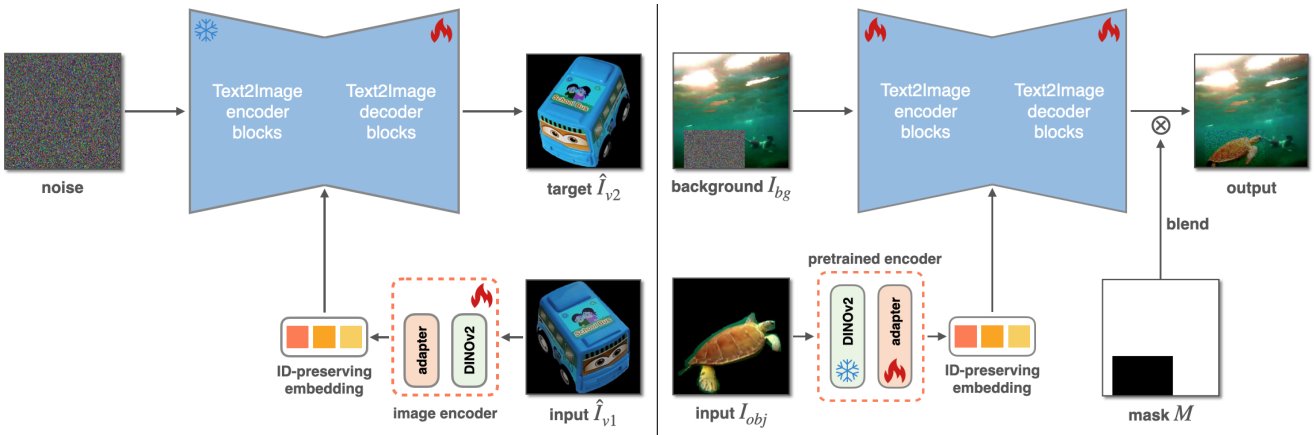
$$\mathcal{L}_{\text{id}} = \mathbb{E}_{\hat{I}_{v1},\hat{I}_{v2},t,\epsilon}\left[\left\|\epsilon - \mathcal{G}_\theta\left(\hat{I}_{v2}, t, \mathcal{E}_u\left(\hat{I}_{v1}\right)\right)\right\|_2^2\right], \quad (1)$$

where $\mathcal{L}_{\text{id}}$ is the ID-preserving loss and $\epsilon \sim \mathcal{N}(0,1)$. The image encoder $\mathcal{E}_u$ and the decoder blocks of $\mathcal{G}_\theta$ are optimized in this process. Intuitively, the encoder trained for this task will always extract representations that are view-invariant while keeping identity-related details that are shared across different views. The qualitative results of this stage are shown in Sec. 4.7. Unlike previous view-synthesis works [33], our context-agnostic ID-preserving stage does not require any 3D information (*e.g.*, camera parameters) as conditions, and we mainly focus on ID-preservation instead of geometrical consistency to background (which will be handled in the second stage). Therefore, only the image encoder will be taken to the next stage.

3.2. Compositing Stage

Fig. 2b illustrates the pipeline of the second stage which is trained for the compositing task, comprising the finetuned image encoder $\mathcal{E}_u$ and a generator $\mathcal{G}_\phi$ (parameterized by ϕ) conditioned on the ID-preserving representations.

A simple approach is to ignore the view synthesis stage, training the encoder and generator jointly in a single-stage



(a) Stage of context-agnostic ID-preserving: we design a novel image encoder (with pre-trained DINOv2 as backbone) trained on multi-view object pairs to learn view-invariant ID-preserving representation.

(b) Stage of object compositing: taking the learned image encoder from the first stage and freezing its backbone, the whole model is trained for compositing the object to the masked region (see Fig. 9 for the blending process).

Figure 2. The two-stage training pipeline of the proposed IMPRINT.

framework. Unfortunately, we found quality degradation from two aspects in this naive endeavor (see Sec. 4.7):

- When DINOv2 is trained in this stage, the model exhibits more frequent copy-paste-like behavior that composites the object in a very similar view as its original view.
- When object-centric multi-view datasets, *e.g.*, MVImgNet [59], are enabled in the training set, the model tends to produce more artifacts and exhibit poorer blending results due to the absence of background information in such datasets.

To overcome the issues above, we freeze the backbone of the image encoder (*i.e.*, DINOv2) in the second stage and carefully collect a training set (see Sec. 3.3 for details).

In this stage, we also leverage a pretrained T2I model as the backbone of the generator, which uses the background I_{bg}, a coarse mask M as inputs, and is conditioned on a ID-preserving object tokens $\hat{E}_u = \mathcal{E}_u(I_{obj})$, where I_{obj} indicates a masked object image. The generation is guided by injecting object tokens into the cross attention layers of $\mathcal{G}_\phi$. The coarse mask also allows the synthesis of shadows, and interactions of the object and the nearby objects.

As $\hat{E}_u$ already encompasses structured view-invariant details of the object, color and geometric adjustments are no longer limited by identity preservation efforts. This freedom allows for greater variation in compositing.

We define the objective function of this stage as:

$$\mathcal{L}\text{comp} = \mathbb{E}_{I_{obj}, I^*_{bg}, M, t, \epsilon}\left[M \left\| \epsilon - \mathcal{G}_\phi\left(I^*_{bg}, t, \hat{E}_u\right)\right\|_2^2\right] \quad (2)$$

where $\mathcal{L}_{comp}$ is the compositing loss, I^*_{bg} is the target image. $\mathcal{G}_\phi$ and the adapter are optimized.

The Background-blending Process To ensure that the transition area between the object and the background is smooth, we adopt a background-blending strategy. Refer to the Appendix for details.

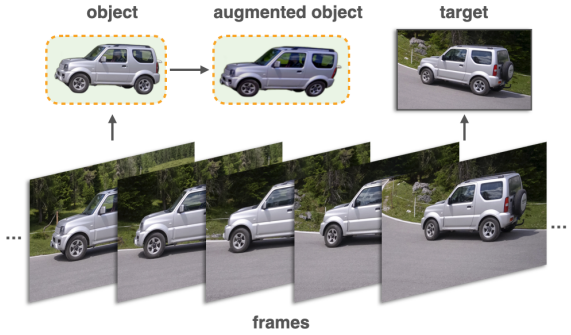


Figure 3. Illustration of the data augmentation pipeline.

Shape-guided Controllable Compositing could enable more practical guidance of the pose and view of the generated object by drawing a rough mask. However, most prior works [6, 35, 53] have no such control. In our proposed model, following [55], masks are defined at four levels of precision (see the Appendix), where the most coarse mask is a bounding box. Incorporating multiple levels of masks replicates real-world scenarios, where users often prefer more precise masks. Results are shown in Fig. 1.any

3.3. Paired Data Generation

The dataset quality is another key to better identity preservation and pose variation. As proved by [6], multi-view datasets can significantly improve the generation fidelity. In practice, we use a combination of image datasets (Pixabay), panoptic video segmentation datasets (YoutubeVOS [56], VIPSeg [36] and PPR10K [29]) and object-centric datasets (MVImgNet [59] and Objaverse [7]). They are incorporated in different training stages and associated with various processing procedures in our self-supervised training.

The image datasets we collected have high resolution and rich background information, so they are only utilized in the second stage for better compositing. Inspired by

[53, 58], to simulate the lighting and geometry changes in object compositing, we design an augmentation pipeline $\hat{I}_{obj} = \mathcal{P}(\mathcal{T}(I_{obj}))$, where $\mathcal{T}$ are the affine transformations, and $\mathcal{P}$ is color and light perturbation, supported by the look-up table in [21]. The perturbed object $\hat{I}_{obj}$ is used as the input and the natural image I^*_{bg} containing the original object is used as the target.

Video segmentation datasets usually suffer from low resolution and motion blur, which harm the generation quality. Nevertheless, they provide object pairs which naturally differ in lighting, geometry, view and even provide non-rigid pose variations. As a result, they are also used in the second stage. Illustrated by Fig. 3, each training pair comes from one video with instance-level segmentation labels. Two distinct frames are randomly sampled; one serves as the target image, while the object is extracted from the other frame as the augmented input.

Object-centric datasets offer a significantly larger scale than video segmentation datasets and provide more intricate object details. However, they are only used in the first stage due to the limited background information available in these datasets. During training, each pair I_{v1}, I_{v2} are also randomly sampled from the same video with $|v1 - v2| \leq n$, where n is the temporal sampling window. Empirically, we observe a loss in the generation quality as n increases, and $n = 7$ strikes a balance between fidelity and quality.

3.4. Training Strategies

All previous (or concurrent) training-free methods [6, 53, 58, 60] use a *frozen* transformer-based image encoder, either using DINOv2 or CLIP. However, freezing the encoder will limit their capability in extracting the object details: i) CLIP only encodes the semantic features of the object; ii) DINOv2 is trained on a dataset that is constructed based on image retrieval, allowing objects that are not entirely identical to be treated as the same instance. To overcome this challenge, we fine-tune the encoder specifically for compositing, ensuring the extraction of instance-level features.

Due to the extensive scale of the aforementioned encoders, they are prone to overfitting. The implementation of appropriate training strategies can effectively stabilize the training process and improve identity preservation. To this end, we design a novel training scheme: Sequential Collaborative Training.

More specifically, the object compositing stage is further divided into two phases: 1) in the first n epochs, we assign the adapter a larger learning rate of 4×10^{-5}, and assign the UNet a smaller learning rate of 4×10^{-6}; 2) in the next n epochs, we swap the learning rate of these two components (and the training finishes). This strategy focuses on training one component at each phase, with the other component simultaneously trained at a lower rate to adapt to the changed domain; the generator is trained in the end to ensure the synthesis quality.

Datasets	Pixabay	VIPSeg	YoutubeVOS	PPR10K
Training	116,820	51,743	42,868	6,020
Validation	6,490	5,487	3,690	102

Table 1. Statistics of the datasets used in the second stage.

4. Experiments

4.1. Training Details

The first stage is trained on 1,409,545 pairs and validated on 11,175 pairs from MVImgNet, which takes 5 epochs to finish. The learning rate associated with DINOv2 (ViT-g/14 with registers) is 4×10^{-6}, and the batch size is 256. The image embedding is dropped at a rate of 0.05.

The second stage is fine-tuned on a mixture of image datasets and video datasets, including a training set of 217,451 pairs and a validation set of 15,769 pairs (listed in Tab. 1), where we apply [26] to obtain the segmentation masks as labels. It is trained for 15 epochs with a batch size of 256. The embedding is dropped at a rate of 0.1.

In both stages, the images are resized to 512×512. During inference, the DDIM sampler generates the composite image after 50 denoising steps using a CFG [16] scale of 3.0. The model is trained on 8 NVIDIA A100 GPUs. The model is built on Stable Diffusion v1.4 ([43]).

4.2. Evaluation Benchmark

Datasets are collected from Pixabay and DreamBooth [44] for testing. More specifically, Pixabay testing set has 1,000 high-resolution images and has no overlap with the training set. A foreground object is selected from each image and perturbed through the data augmentation pipeline as in Sec. 3.3. The DreamBooth testing set consists of 25 unique objects with various views. Combined with 59 background images that are manually chosen, 113 pairs are generated for this test set. This dataset is challenging since most objects are of complex texture or structure. We also conduct a user study on this dataset.

Metrics measuring fidelity and realism are adopted to evaluate the effectiveness of different models in terms of identity preservation and background harmonization. We utilize CLIP-score [14], DINO-score, and DreamSim [8] as the measurements of generation fidelity. To obtain more precise comparison results, we always crop the output images so that the generated object is located in the center of the image. FID [15] is employed to measure the realism which indicates the compositing quality.

4.3. Quantitative Evaluation

To demonstrate the effectiveness of our model, we test our model and three baseline methods (Paint-by-Example [58],

Method	FID ↓	CLIP-score↑	DINO-score↑	DreamSim ↓
PbE	-	71.5000	31.3765	0.4954
OS	-	73.6250	32.9739	0.4297
T-I	-	75.1250	39.2863	0.3661
Ours	-	**77.0625**	**43.4463**	**0.2898**
PbE	23.2663	93.6250	85.2260	0.1907
OS	22.4934	94.9375	90.3853	0.1422
T-I	63.9730	88.3125	73.2155	0.3219
Ours	**16.4487**	**96.1875**	**94.705**	**0.0831**

Table 2. Quantitative comparison with prior works. IMPRINT and the baselines are tested on two datasets for realism and ID-preserving measurement: DreamBooth (top) and the Pixabay test set (bottom). The results on both datasets demonstrate the advance of our model in both ID-preserving and realistic harmonization with the background.

	Ours	OS	Ours	PbE	Ours	T-I
Realism	**50.68**	49.32	**62.84**	37.16	**53.38**	46.62
Fidelity	**80.41**	19.59	**86.49**	13.51	**73.65**	26.35

Table 3. User study results. We design two questions to measure the realism and fidelity of the generation. In both questions, the user is presented side-by-side comparisons of our generated image and another image randomly chosen from one of the baselines. The results in the table show user preference percentage. Our model not only achieves better realism, but also outperforms the baselines in ID-preserving by a large margin.

ObjectStitch [53], and TF-ICON [35]) on the two aforementioned test sets. The same inputs (a mask and a reference object) are used in all models. For fair comparison, we further fine-tune Paint-by-Example (PbE) on our second-stage training set.

When testing on TF-ICON, we employ the parameter set in "same domain" mode, as suggested by the official implementation. It also requires a text prompt as an additional input, so we apply BLIP2 [27], a state-of-the-art vision-language model to generate captions for the images. Moreover, the captions for the DreamBooth test set are manually refined to improve the performance. As shown in Tab. 2, IMPRINT achieves the best performance in both realism and fidelity. See the Appendix for quantitative comparisons with AnyDoor.

4.4. Qualitative Evaluation

Qualitative comparisons are shown in Fig. 4, comparing our model against prior methods. Although PbE and ObjectStitch show natural compositing effects, they often fail to capture the finer details of the objects. When the object has complex texture or structure, their generated object becomes less recognizable and even suffers from artifacts. In contrast, TF-ICON shows better consistency between the input and output, especially in keeping surface textures and captions. However, the background adaptation ability is also strictly restricted. As can be observed, TF-ICON has less variation in color and geometry changes, which results in a degradation in compositing effects. We further compare to AnyDoor with both qualitative and quantitative assessments (see the Appendix). The results show that IMPRINT achieves better ID-preservation and shows the flexibility in adapting to the background in terms of color and geometry.

We also show the synthesis results of the first stage in Fig. 5. Using the ID-preserving representation, our model is able to generate high-fidelity objects with large view variations. This process requires no extra condition such as camera parameters.

4.5. User Study

We also conduct a user study using Amazon Mechanical Turk, comparing our method against the three baselines on the challenging DreamBooth dataset. The user study consists of side-by-side comparisons of our result and a randomly chosen result from the baselines. We design two questions: 1) Which image is more realistic? (the input objects are hidden from the users) 2) Which image is more similar to the reference object? Each question has 111 comparisons. We received more than 880 votes from over 130 users. The results are shown in Tab. 3. In terms of realism, our model outperforms PbE and TF-ICON, while comparable with ObjectStitch. We also evaluate the visual similarity. The preference rate in the table demonstrates that our method has a significant advantage over the baselines.

4.6. Additional Visual Results of Shape-control

Shape-guided generation introduces a lot more flexibility for image editing, as the user now gains control over the shape, view and pose of the objects, and the transformation can be either rigid or non-rigid. Fig. 6 illustrates the diverse usage of image editing given a mask as guidance.

4.7. Ablation Study

When pursuing better identity preservation and background harmonization in the field of generative object compositing, we gain valuable experience in a wide range of techniques that contribute to this task. In Tab. 4, we provide a complete analysis and insights of all the factors, as well as demonstrate the effectiveness of our proposed method. The same metrics are utilized as explained in Sec. 4.2.

Training strategies. In setting 2, we also optimize CLIP encoder. The results of settings 1 and 2 show that the optimized CLIP can capture better object identity. However, this improvement comes at the cost of variation. Setting 5 and 6 also demonstrate improved identity and less variation. For this reason, the encoder backbone is frozen in our second stage.

Dataset. Dataset is another component that significantly affects the performance. After adding the video datasets,

Figure 4. Qualitative comparison on the DreamBooth test set. Paint-by-Example and ObjectStitch lose most object details and only maintain categorical information. TF-ICON tends to copy the pose of the input subject. The comparison highlights the advantage of IMPRINT in keeping identity and making geometric changes.

No.	(PRE)	Encoder	Adapter	(PRE) Tune encoder	(PRE) Tune UNet	Tune encoder	Video data	MVImgNet	FID ↓	DINO score ↑
1	✗	CLIP	✓	-	-	✗	✗	✗	22.493	90.385
2	✗	CLIP	✓	-	-	✓	✗	✗	19.538	92.695
3	✗	CLIP	✓	-	-	✓	✓	✗	17.847	94.216
4	✗	DINOv2	✗	-	-	✗	✓	✗	20.748	92.512
5	✗	DINOv2	✓	-	-	✗	✓	✗	20.131	92.846
6	✗	DINOv2	✓	-	-	✓	✓	✗	17.477	94.164
7	✗	DINOv2	✓	-	-	✓	✓	✓	17.947	94.023
8	✓	DINOv2	✓	✓	✗	✗	✓	✗	17.847	93.908
9	✓	DINOv2	✓	✗	✓	✗	✓	✗	19.286	93.273
10	✓	DINOv2	✓	✓	✓	✗	✓	✗	**16.449**	**94.705**

Table 4. Ablation study on our methodologies and other common components. *PRE* means whether the setting has our pretraining stage; *MVImgNet* and *video data* mean whether they are used in the compositing stage.

the model develops a stronger capability in engraving the details (setting 2 and 3). Nevertheless, if there are too many training pairs from object-centric datasets (MVImgNet), the generation quality will degrade (setting 6 and 7) since the background information is insufficient.

Architecture. Inspired by [53], we also use an adapter to connect the encoder with the generator. Setting 4 and 5 indicates that using the adapter will boost the overall performance in both realism and fidelity. We also observed the model converges faster when using the adapter.

Pretraining. In our framework, the first stage pretraining is a key component in improving ID-preservation and harmonization effects. To demonstrate the effectiveness, we test the original DINOv2 and our finetuned DINOv2 on a Objaverse test set. In this evaluation, the encoders generate embeddings for diverse views of 20 objects from various categories. The embeddings are then clustered and visualized in t-SNE figures (Fig. 8). This figure shows that the finetuned encoder produces better clustering results, demonstrating that our ID-preserving representation effectively encodes

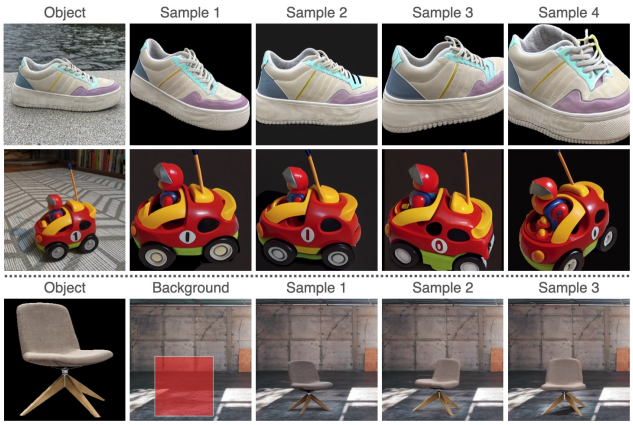


Figure 5. Top: Results of context-agnostic ID-preserving pretraining (after the first stage); IMPRINT generates view pose changes while memorizing the details of the object. Bottom: Diverse poses of the object after the second stage.

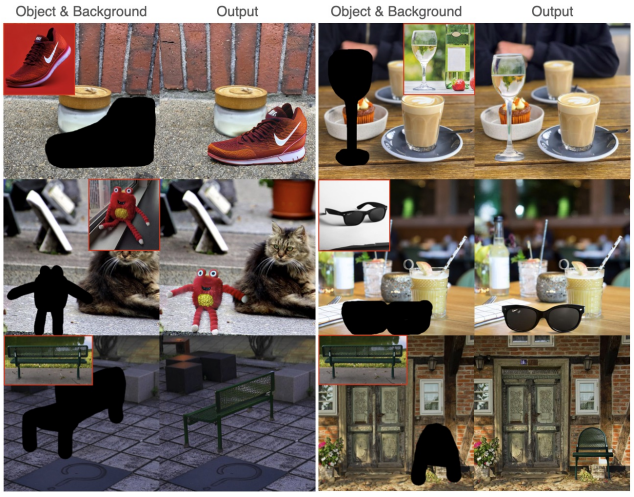


Figure 6. More shape-control results. IMPRINT introduces more user control by using a user-provided mask as input. Inspired by [55], we define four types of mask (including bounding box). In addition to compositing, our model also performs edits on the input object. Depending on the shape of the coarse mask, IMPRINT can operate different types of editing, including changing the view of an object, and applying non-rigid transformation on the object.

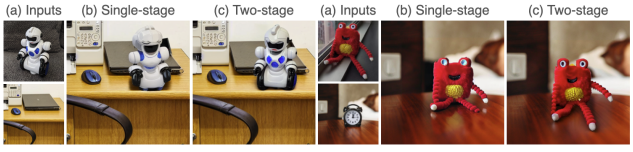
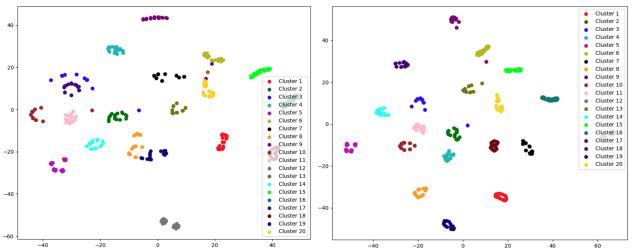


Figure 7. Ablation study on our two-stage training scheme. In (b) MVImgNet is added to the training set and simply trains the whole network in one stage. Compared with two-stage training, single-stage has a notable degradation in quality and loses more details.

the key details of the objects. We further ablate on the first stage training using setting 7 (where there is only the compositing stage) and 10 (two-stage). Without the first stage, there is a notable drop in the compositing quality (Fig. 7). Additionally, we assess the effect of freezing some components (*i.e.*, UNet or DINOv2) during the pretraining. Compared with Setting 10, setting 8 and 9 exhibit a drop in both harmonization and ID-preservation, validating the effectiveness of our training scheme.



(a) Clustering results using the original DINOv2.

(b) Clustering results using our finetuned DINOv2.

Figure 8. Ablation study on our first stage training. We use DINOv2 (before and after the first stage) to predict embeddings of different views of 20 Objaverse objects. The embeddings are then clustered using the same algorithm and visualized using t-SNE figures. The improved clustering results demonstrate that the embeddings produced by finetuned DINOv2 have higher quality.

5. Conclusion, Limitation and Future Work

In this paper, we propose IMPRINT, a novel two-stage framework that achieves state-of-the-art performance in identity preservation and background harmonization for generative object compositing. We design a new pretraining scheme where the model learns a view-invariant identity-preserving representation that efficiently captures the details of the object. By decoupling the task into an identity-preserving stage and a harmonization stage, IMPRINT can generate large color and geometry variations to better align with the background. Through visual and numerical comparison results, we show that IMPRINT significantly outperforms the previous methods in this task. Furthermore, we add shape guidance as an additional user control. Although IMPRINT effectively addresses both identity preservation and background alignment, it has several limitations. When the required view change is too large, there could be a notable drop in identity preservation, which can be improved by exploring and incorporating a 3D model or NERF representation into our model. Another limitation is that the model may degrade consistency of small texts or logos. Potential ideas to improve this are to employ more accurate latent auto-encoder to avoid loss of information in the latent space and learn object encoders at higher resolution to encode small local details more accurately.

Acknowledgements

This research is partially supported by NSF grants 1835739 and 210671.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 2

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 3

[5] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2019. 2

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2, 3, 4, 5, 1

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3, 4

[8] Stephanie Fu*, Netanel Tamir*, Shobhita Sundaram*, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv:2306.09344*, 2023. 5

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 3

[10] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. Photoswap: Personalized subject swapping in images. *arXiv preprint arXiv:2305.18286*, 2023. 3

[11] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5917–5926, 2023. 2

[12] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part I*, page 361–378, Berlin, Heidelberg, 2023. Springer-Verlag. 2

[13] Liu He, Jie Shan, and Daniel Aliaga. Generative building feature estimation from satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2

[14] 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. 5

[15] 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. 5

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] 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*, pages 914–922, 2022. 2

[19] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *arXiv preprint arXiv:2312.06739*, 2023. 3

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. Openclip, july 2021. 2(4):5, 2021. 3

[21] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 2, 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[24] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*, pages 690–706. Springer, 2022. 2

[25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[26] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 5

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[28] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[29] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[30] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[32] Jiayi Liu, Tinghan Yang, and Jennifer Neville. Cliqueparcel: An approach for batching llm prompts that jointly optimizes efficiency and faithfulness. *arXiv preprint arXiv:2402.14833*, 2024. 3

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3

[34] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 3

[35] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 1, 2, 3, 4, 6

[36] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 4

[37] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3

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

[39] 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. 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[41] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10721–10733, 2023. 2

[42] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement, 2023.

[43] 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, 3, 5, 6

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3, 5

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 3

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[47] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4390, 2021. 2

[48] 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.

[49] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. Pixht-lab: Pixel height based light effect generation for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16643–16653, 2023. 2

[50] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van

Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022. 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[52] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[53] 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. 1, 2, 3, 4, 5, 6, 7

[54] 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. 2

[55] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model, 2022. 4, 8

[56] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 4

[57] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. *arXiv preprint arXiv:2207.04788*, 2022. 2

[58] 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, 5, 3

[59] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9150–9161, 2023. 3, 4

[60] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 2, 3, 5

[61] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 365–374, 2021. 2

[62] 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. 2

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[64] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation

Supplementary Material

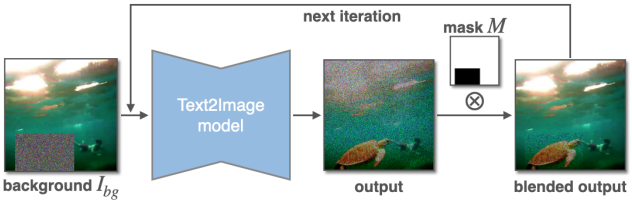


Figure 9. Illustration of the background-blending process.

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.

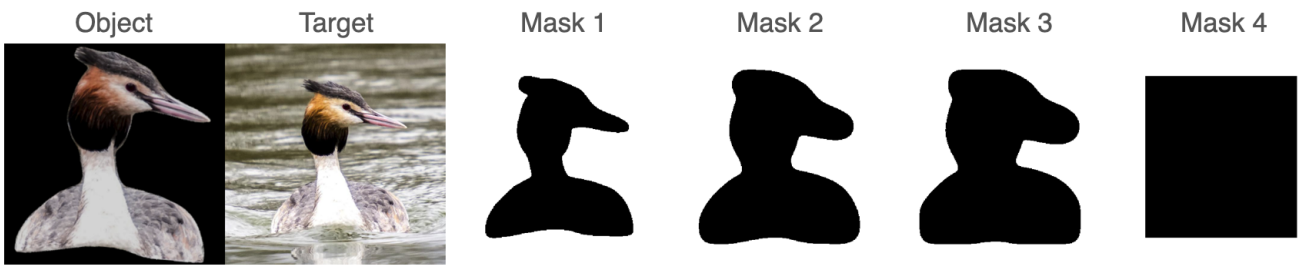
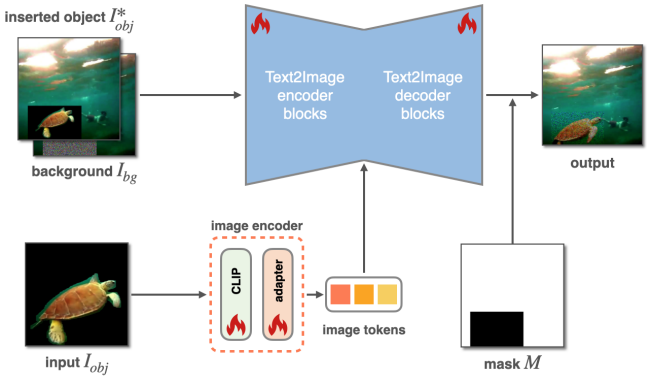
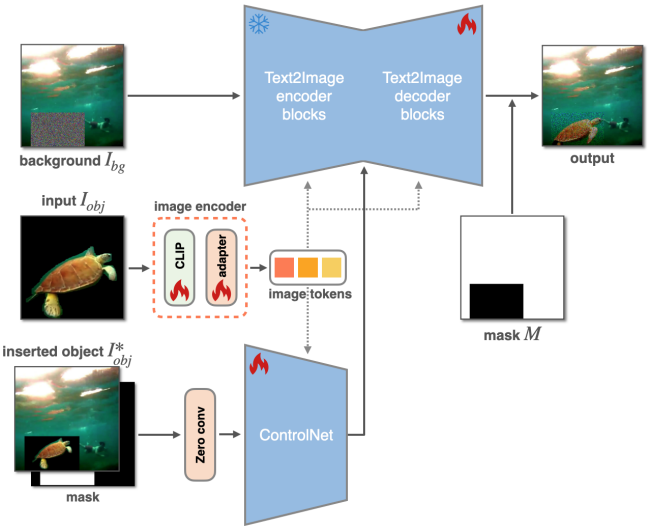


Figure 10. The four types of mask used in the second compositing stage. The generation is constrained in the masked area so the user-provided mask is able to modify the pose, view and shape of the subject.



(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

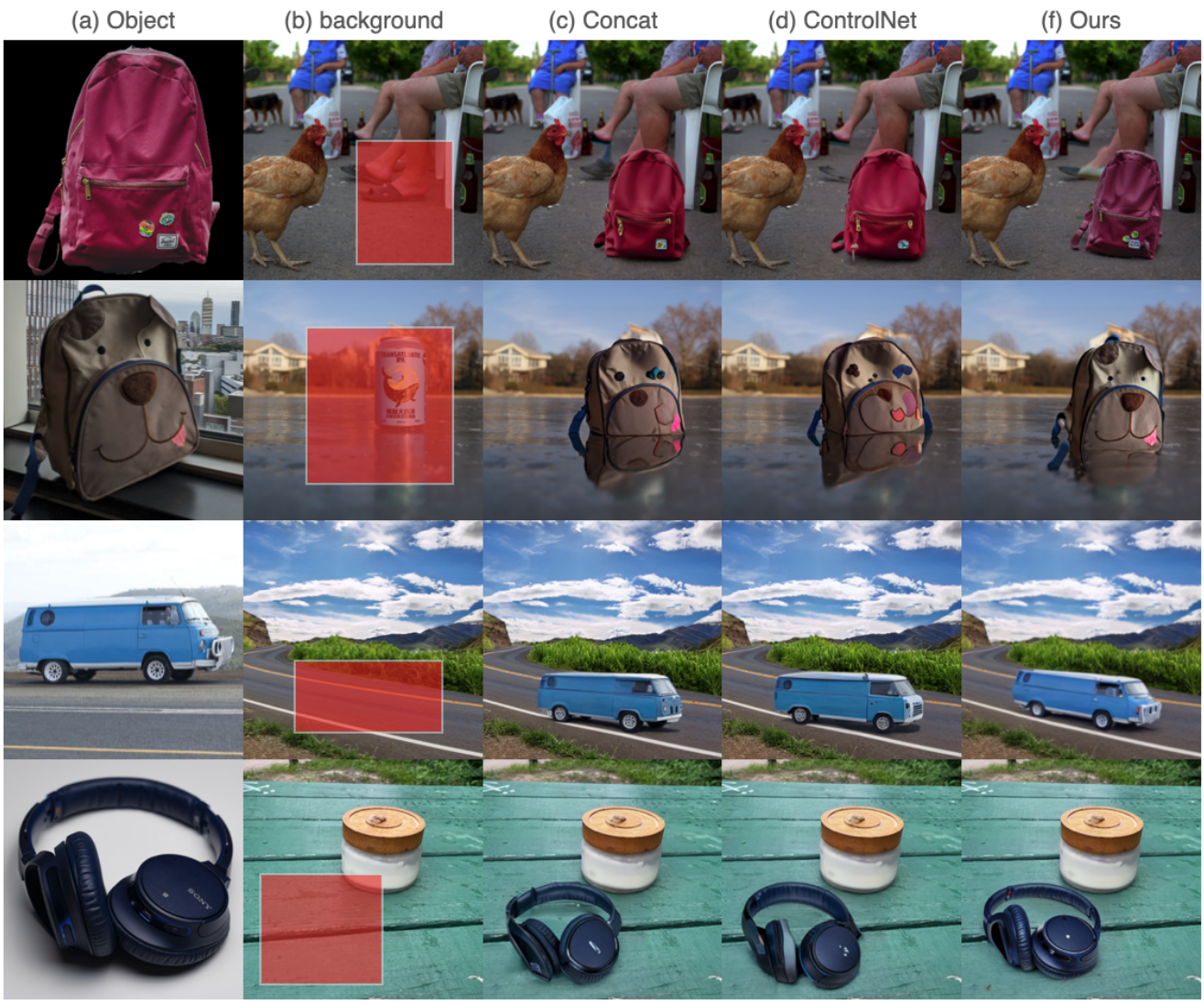


Figure 12. Qualitative comparisons with concatenation-based pipeline and ControlNet-based pipeline. Our model shows stronger ability in geometric adjustments (especially in the first three examples) as well as better performance in identity preservation.

(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

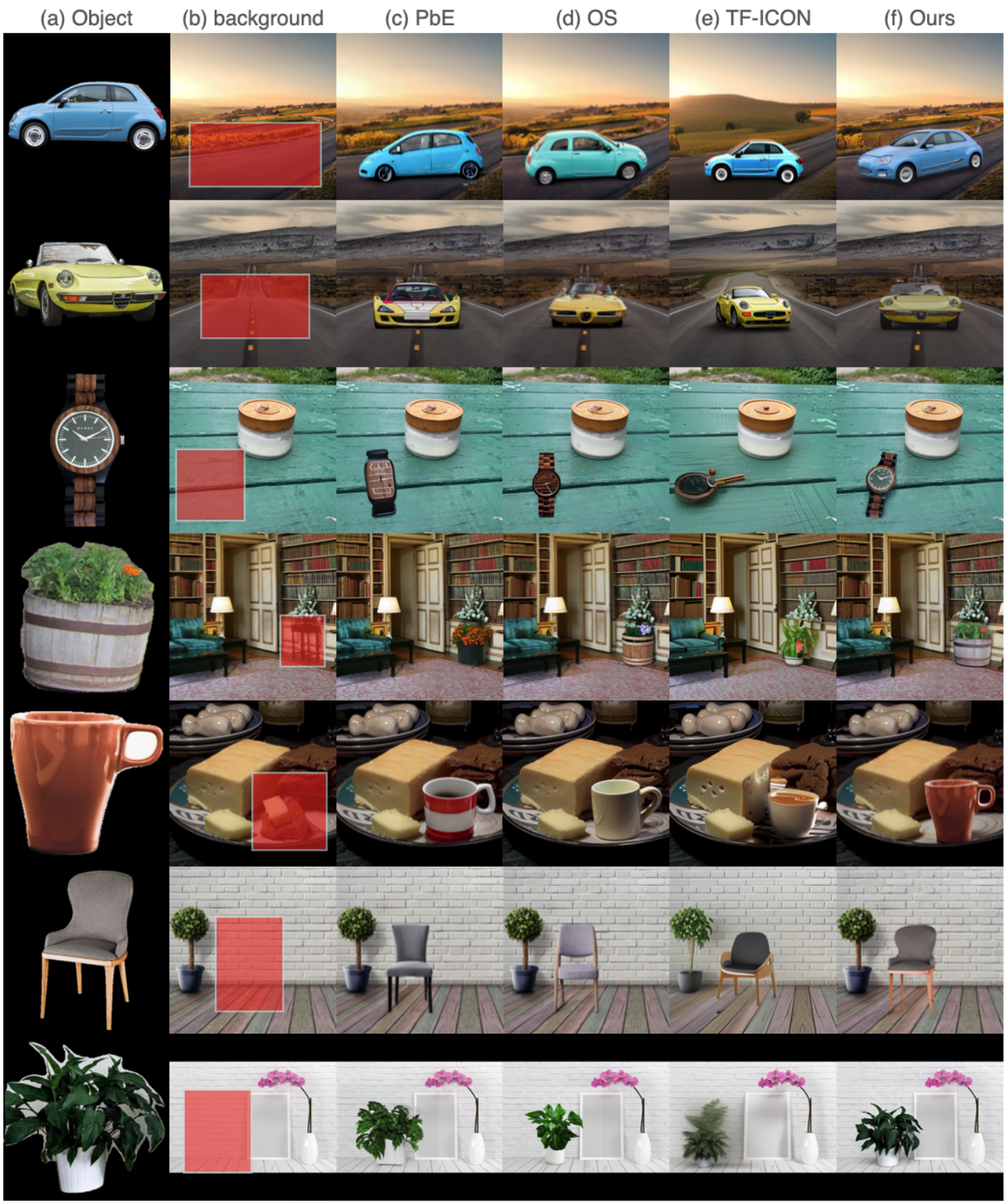


Figure 13. More qualitative comparisons. We compare our proposed model with Paint-by-Example (PbE), ObjectStitch (OS) and TF-ICON. IMPRINT better preserves object identity and the generated object is more consistent with the background.

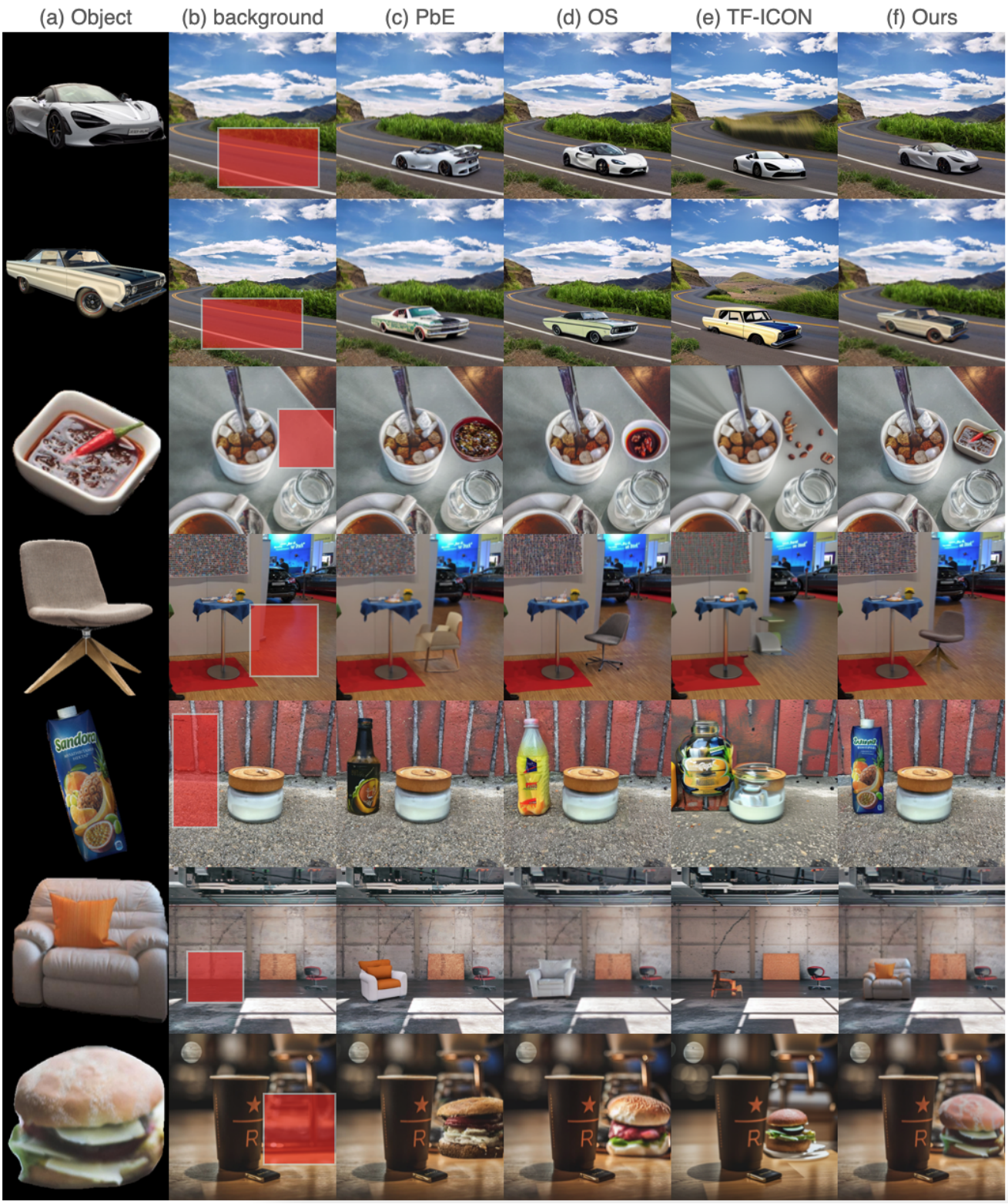


Figure 14. More qualitative comparisons. We compare our proposed model with Paint-by-Example (PbE), ObjectStitch (OS) and TF-ICON. IMPRINT better preserves object identity and the generated object is more consistent with the background.

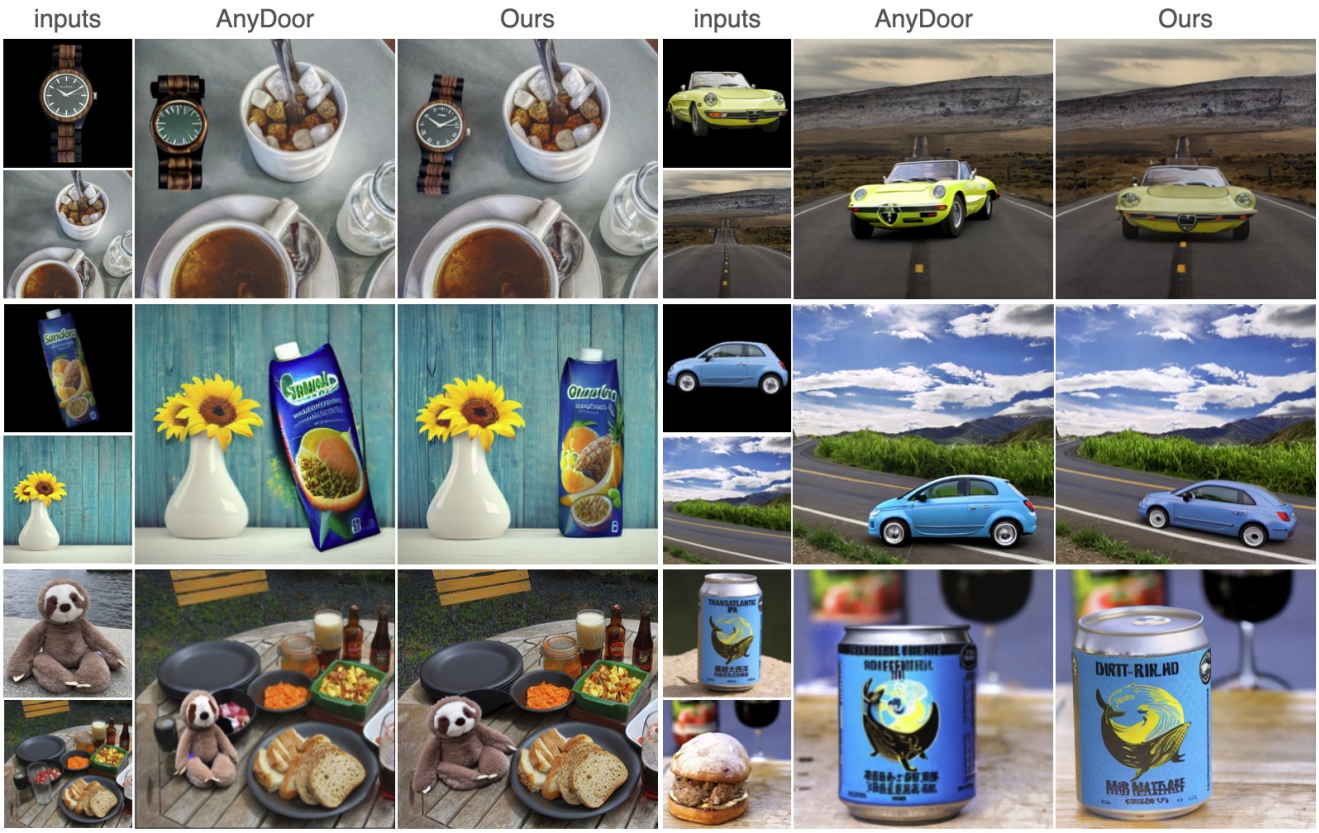


Figure 15. Additional qualitative comparisons with AnyDoor.

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.