InsertDiffusion: Identity-Preserving Visualization of Objects through a Training-Free Diffusion Architecture

Phillip Mueller[1,2], Jannik Wiese[3], Ioan-Daniel Craciun[4], and Lars Mikelsons[2]

[1] BMW Group, Knorrstrasse 147, 80788 Munich, Germany
phillip.mueller@bmw.de
[2] University of Augsburg, Am Technologiezentrum 8, 86159 Augsburg, Germany,
[3] Ludwig-Maximilians University Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany,
[4] Technical University of Munich, Boltzmannstraße 15, 85748 Garching, Germany.

Fig. 1: Realistic object representations in existing and generated backgrounds wit **InsertDiffusion** without the necessity for training or finetuning any parts of the architecture.

Abstract. Recent advancements in image synthesis are fueled by the advent of large-scale diffusion models. Yet, integrating realistic object visualizations seamlessly into new or existing backgrounds without extensive training remains a challenge. The purpose of this work is to develop a customizable approach that simplifies object insertion while maintaining identity and structural integrity, making high-quality visual compositions more accessible for engineering, design, and marketing applications. We therefore introduce InsertDiffusion, a novel training-free diffusion architecture that efficiently embeds objects into images while preserving their structural and identity characteristics. Our approach utilizes off-the-shelf generative models and eliminates the need for fine-tuning, making it ideal for rapid and adaptable visualizations in product design and marketing. We demonstrate superior performance over existing methods in terms of image realism and alignment with input conditions. By decomposing the generation task into independent steps, InsertDiffusion offers a scalable solution that extends the capabilities of diffusion models for practical applications, achieving high-quality visualizations that maintain the authenticity of the original objects.

1 Introduction

Image generation is undergoing remarkable advancements with the rise of diffusion models, achieving unprecedented levels of realism and naturalness in synthetic images [13, 39, 9, 34]. The evolution of latent diffusion models, especially Stable Diffusion (SD) [31] and its variants such as Stable Diffusion XL (SDXL) [28], continues to further improve generalization, quality, and allows for a variety of conditioning mechanisms such as text and reference images. A crucial component in these advancements is the development of CLIP [29], which provides a foundation for referencing text to visual concepts. Due to its generative capabilities and adaptability, SD sparked a wave of subsequent modifications and extensions to further increase the levels of image quality, customization and user-control.

Example conditions for controlled generation include sketches and spatial maps [52], shape-based guidance [24], as well as semantic segmentations and keypoints [22]. Image editing has also seen significant progress. Besides text-driven image editing [45, 1, 42], point-based dragging approaches [37, 21] and inpainting methods [17] extend the capabilities for image editing.

In this paper, we investigate the task of realistic object visualization, which involves inserting a given object into an existing or newly generated background and merging both representations to create a perceptually appealing scene while maintaining the object's key structure and characteristics. Some examples are visualized in Fig. 1. This task is particularly relevant for applications in product and engineering design, as well as customer-oriented marketing. Proposed applications include rendering geometric or CAD-like objects as realistic images and enabling customization in advertising and personalization (e.g.: visualizing a new car or bike in a customer's driveway). We specifically aim to visualize technical representations of products like bicycles as well as design-representations of consumer-products like cars.

Our approach[5] aims to decouple the highly customized models and workflows required for generating technically accurate images from the visualization and scenic representation. By leveraging publicly available, large-scale diffusion models, we can visualize the results in realistic scenes without the need for training or fine-tuning. This allows smaller, domain-specific models to focus on generating specific types of images while utilizing the extensive capabilities of models trained on millions of images for realistic rendering.

Previous studies have explored this task from various angles. However, none have addressed it from the perspective of fully leveraging existing capabilities in foundation models as straightforward as possible while utilizing publicly-available implementations only and avoiding training or finetuning altogether. Methods such as TF-Icon [16], AnyDoor [6], and PrimeComposer [46] aim to inject objects into given backgrounds. TF-Icon and PrimeComposer, which are both training-free, modify the injected object to align with the background style, altering its characteristics noticeably. AnyDoor learns detail- and ID-extractors

[5] Code is found under: https://anonymous.4open.science/r/InsertDiffusion-C377

to achieve mitigate this issue but is, therefore, not training-free. CollageDiffusion [35] merges multiple images into one collage but is also not training-free. For realistic image insertion into newly generated backgrounds, the state-of-the-art methods ReplaceAnything [3] and ObjectDrop [47] both require training or finetuning.

Despite the remarkable results of the existing methods, we find that they do not fully leverage the inherent capabilities of SD. Numerous studies have demonstrated these extensive capabilities in other domains [27, 40, 18]. To ensure consistency, adaptability, and ease of use, we propose a significantly simpler method that utilizes off-the-shelf generative models available through the Diffusers-library on HuggingFace [43] for all tasks. Our architecture is designed to be adaptable to the fast-evolving field of diffusion models, allowing for the replacement of any component in the architecture as new, improved versions become available.

In essence, we create a mask from the object and pass it, along with the object, to SD using the inpainting function. The inverse of the object mask defines the area in the background-image that the model can modify, while the object itself remains unchanged. After generating an intermediate image composition, we apply an image-to-image transformation that noises and then denoises the composed image again to optimize high-frequency structures.

The remainder of this work is structured into the following sections. Section 2 provides an overview on previous works which we build upon as well as existing methods for object insertion. Section 3 discusses the architecture of InsertDiffusion, its components and the specific design choices. In Section 4, we conduct qualitative and quantitative experiments and compare the performance of InsertDiffusion to alternative methods. We subsequently discuss the limitations and directions for future research in Section 5 and the potential societal impact of our work in Section 6 and draw a conclusion about our contribution in Section 7.

2 Related Work

2.1 Image-to-Image Transformation

Image-to-image transformations include a variety of tasks like local image editing, colorization, inpainting, uncropping, upscaling, and style changes. Tumanyan [42] manipulate the spatial features and self-attention layers of a pretrained SD model during the generation process. They inject features from the initial image into the text-guided image generation. Palette [33] pursues a different approach in proposing a unified framework for image-to-image translations using conditional diffusion models. The input image is partially noised and then iteratively denoised. The denoising process starts at an intermediate, noisy representation of the input image and is conditioned on text or other modalities. For latent diffusion models like SD, the input image is encoded and noise is added to the latent representation. The Diffusers library offers image-to-image implementations based on SD which, therefore, also operates with latent images [43]. Despite

their versatility, existing methods for image-to-image style transformation lack the possibility to explicitly maintain the identity of objects within the image and do not allow for the harmonic composition of objects and novel backgrounds.

2.2 Inpainting

Inpainting is a common method for local image editing. It relies on a mask to determine which regions in an image can be modified by the diffusion model. In each generation step, the initial image is noised according to the current timestep. Its unmasked regions are merged with the masked regions modified by the diffusion model and forwarded into the next denoising timestep [17]. Inpainting functionalities based on SD are provided in Diffusers [31, 43].

2.3 Object Insertion

Object insertion is essentially an image blending task. A notable early work that is not based on machine learning is Poisson image editing [26] which aims to seamlessly merge an object from a source image into a destination image by minimizing the differences in gradient across the boundary. This is done by solving Poisson's equation to ensure smooth transitions. However, for complex technical and CAD-like objects, this technique produces unrealistic results, as shown in Figure 8 in the appendix. Being a method for harmonizing both images, the method also fails to allow for adjustments of the background image to fit the object accurately. The same is true for more advanced harmonization approaches based on machine learning [41, 8, 54].

Most diffusion-based works for object insertion employ finetuning or training of an additional adapter. AnyDoor [6] is designed for zero-shot "object teleportation" into a given scene at specifiable locations by utilizing identity features from the target image and detail features of the target-scene composition. The identity features are extracted using a finetuned visual encoder (DINO-V2 [23]). The detailed features are represented using high-frequency maps. These are then mapped into the diffusion U-Net by a finetuned, ControlNet-style encoder [52].

AnyScene [5] develop a foreground injection module that guides a pretrained diffusion model to generate cohesive scenes in harmony with the provided object in the foreground. To enhance robust generation, they implement a layout control strategy that prevents distortions of foreground elements. By training the foreground injection module, the method is not training-free. ObjectDrop [47] leverages a dataset of "counterfactual" pairs of images that show the scene before and after object removal. This model is used to synthetically create a larger dataset of counterfactual image pairs and subsequently finetune SD for object insertion. PrimeComposer [46] steers the attention weights at different noise levels to preserve the object appearance while composing it with the background in a natural way. Additionally, they employ Classifier-Free Guidance [12] to enhance the quality of the composed images. Paint-by-Example [50] leverages SD [31] and Classifier-Free Guidance [12] and employ self-supervised training to disentangle and re-organize the background image and the object.

TF-Icon [16] is a training-free method that inverts the real images into latent codes using an exceptional prompt that contains no information. The latent codes are then used as a starting point for the text-guided image generation process. Composite self-attention maps are injected to infuse contextual information from the background image into the inserted object.

InpaintAnything [51] combines the inpainting mechanism used in Stable Diffusion [31] and RePaint [17] with SegmentAnything (SAM) [14] for object-adaptive segmentation. Although being training-free, the approach fails to seamlessly integrate the object into the new background, as neither the background nor the object are semantically modified to fit together.

Shopify-Background-Replacement (SBR) [38], first extracts the object from the original background using a depth estimation model. Then the depth image and the text prompt are passed to SDXL-Turbo [36] augmented by a ControlNet [52] which handles the depth map. After inferring a new background the foreground is pasted on top it to generate the final image.

For the task of background replacement Chen et al. [3] present ReplaceAnything. It is based on their previous work VirtualModel [4], which visualizes consumer products in new backgrounds as if they were held by a human. The VirtualModel consists of a Content-guided branch to ensure the consistency of the product, and an Interaction-guided branch to guide the model in creating realistic product-human interactions and is specifically trained for this task. ReplaceAnything is currently only available as a HuggingFace demo as no paper or code have been released.

To summarize the limitations, multiple existing methods for object insertion are not training-free [3, 5, 6, 46, 47, 50], which poses limitations for customization and practical applications where little or no training-data is available. The training-free methods [16, 38, 51] have limitations in the quality of the image composition and in maintaining the identity of the inserted object. With InsertDiffusion, we aim to address these limitations by proposing a modular and training-free framework for object insertion into novel backgrounds. Our method is presented in Section 3.

3 Method

The architecture of InsertDiffusion is shown in Fig. 2. The main idea is to modify image characteristics, such as shadows, lighting and texture in both object and background to obtain a realistic composition. This is done without training an additional adapter or finetuning the diffusion model. Given an object to insert, one can either use an existing background image or generate a new background with SDXL [28]. Once the object is isolated from its original background, we create a mask to "reserve" the desired location in the new background and pass the mask and the original object to the background refinement. This is done to adapt the background s.t. it seamlessly incorporates the object. In a second step, we refine this composition by adding some noise and subsequently denoising again.

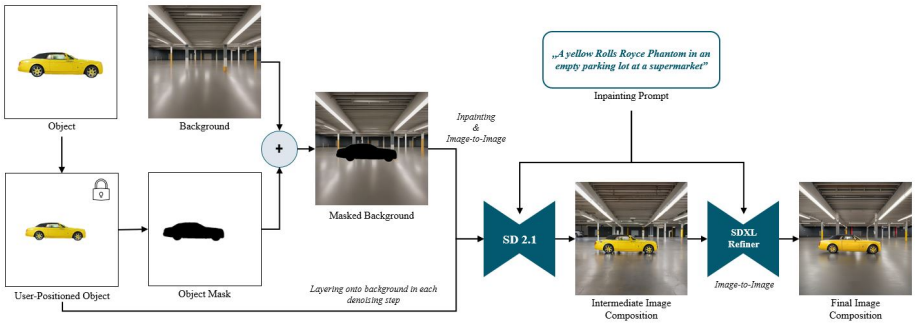

Fig. 2: The **InsertDiffusion Architecture** is designed to insert an object into a background while preserving key visual characteristics of the object. The object is positioned by the user, while an object-mask is created and composed with the background image. The masked background is passed to SD together with the original object. Using image-to-image and inpainting, the original image is layered onto the background for each denoising step. The resulting image composition is subsequently refined by a second diffusion model (SDXL).

3.1 Core Architecture

Our core architecture is inspired by RePaint's [17] resampling strategy and extends it by leveraging intermediate latent image compositions and using text-conditioned guidance to introduce a multimodal layer to contextually adapt the background. Given an isolated and user-positioned object on a white background $x^{(obj)}$, the object mask $m = mask(x^{(obj)})$ is obtained from $x^{(obj)}$ by applying a threshold i.e. by setting all pixels brighter than the threshold to 0 and all others to 1. Using an existing background image, the latent representation z_{bg} is obtained by passing it through the SD encoder. This latent representation is then noised according to the noise schedule. For a new background, the image is simply generated by the text-conditioned sampling with SD.

Intermediate Image Composition. The intermediate image composition produces a modified version of the background. In general, it can be computed by:

$$\hat{z}^{(comp)} = m \odot z^{(obj)} \quad +(1-m) \odot G(z^{(bg)}, m, z^{(obj)}, \tau_\theta(y)). \tag{1}$$

Whereby, $z^{(obj)}$ is a latent representation of the object image obtained by passing it through the SD encoder, $z^{(bg)}$ is a latent representation of the background image obtained using the SD encoder and G is a masked diffusion process. The term $m \odot z$ ensures that the object area is preserved in the latent representation. $(1-m) \odot G(z^{(bg)}, m, z^{(obj)}, \tau_\theta(y))$ updates the background image in the regions where $m = 0$. The generation is guided by the masked object and the CLIP-encoded text-prompt $\tau_\theta(y)$. To iteratively refine the background and allow for seamless object integration, the latent representation of the original object $z^{(obj)}$ is injected into the background for every denoising step. This is done to keep the object itself mostly unmodified. The update for each timestep, using the latent

diffusion model ϵ_θ, is:

$$\hat{z}_{t-1}^{(comp)} = m \odot z_{t-1}^{(obj)} + \quad + (1-m) \odot (\hat{z}_t^{(comp)} - \epsilon_\theta(\hat{z}_t^{(comp)}, \tau_\theta(y), t)). \quad (2)$$

Hereby, $z_t^{(obj)}$ is calculated by adding noise to the latent representation of the object according to the noise schedule and timestep. Noise is added according to the canonical formulation of the diffusion forward process [13] given by

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),\ \alpha_t = 1 - \beta_t,\ \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s \quad (3)$$

where all β_t are defined by the noise schedule. We use the default noise schedule for each model as set in the diffusers library [43]. The masked diffusion process G is obtained by iteratively applying Equation Eq. (2) from some initial timestep to $t = 0$. To insert the object into a given background we set the initialize $t = n$ and obtain $\hat{z}_n^{(comp)}$ by applying Equation Eq. (3) on a pasted composition $z^{(pasted)} = m \odot z^{(obj)} + (1-m) \odot z^{(bg)}$. To generate a completely new background we set $t = n = T$ which is equivalent to initializing $\hat{z}_T^{(comp)}$ from gaussian noise and iteratively applying Equation Eq. (2).

Refinement. The second step aims to refine the composed intermediate image by making it more consistent and modifying high-frequency image characteristics. The intermediate image composition $\hat{z}^{(comp)}$ obtained using Equation Eq. (1) is noised for n timesteps using Equation Eq. (3) to obtain z_n. Subsequent denoising steps are guided by the text-prompt and iteratively computed as follows:

$$z_{t-1} = z_t - \epsilon_\theta(z_t, t, \tau_\theta(y)), \text{ with } t = t_n, t_{n-1}, \ldots, t_0. \quad (4)$$

By using this refinement stage, both the background and the object undergo minor style changes. We accept a trade-off between object consistency and overall image quality and realism. We opt to not copy the initial object into the final image, as for example done by InpaintAnything [51], because this would deprecate seamless blending. In Section 4.5, we conduct an ablation study that compares the results with and without image refinement.

3.2 Optional Additions

To make our architecture more accessible and flexible, we provide additional functionalities to prepare the object for insertion. Again we utilize existing models that are available within Diffusers [43] and an implementation of Language Segment Anything (langSAM)[14, 20] found in Transformers [48].

Background Generation. If no existing background is provided, the user can utilize text-to-image models like SD [31] or SDXL [28] and generate a new background from a text prompt. We provide a prompt template that only requires the product-type, color, and place to be filled in by the user.

Colorization. Technical images often come as drawing- or CAD-like black-and-white representations. We observed that inserting such images severely deprecates the quality of the image composition. An example of this is given in **??**. Therefore, we provide a colorization scheme, visualized in Figure 3 based on inpainting with SD. For low-resolution images, we increase the resolution using SD upscaling. For an input image of 256×256 resolution we upscale it by a factor of 4 to 1024×1024. The upscaled image, a mask and a text prompt are passed to SDXL to perform colorization.

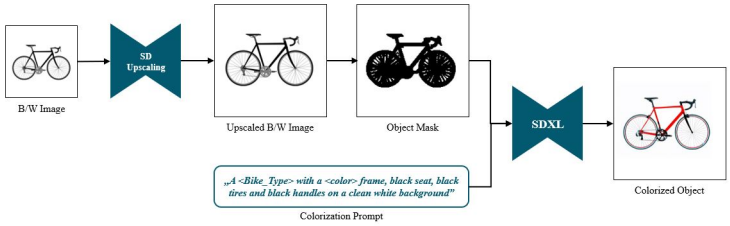

Fig. 3: **Image Colorization scheme** for black-and-white images. Given a mask of the object, SDXL [28] is prompted to color the object defined by the masked area. If the original image containing the object is of low resolution, we advise upscaling the object by using the functionality provided by Stable Diffusion.

Object Segmentation. Since not all object images are already separated from their original background, we include automatic object segmentation. For this task, we utilize langSAM [14]. The user only has to provide an approximate object category. The segmentation model is available in the Transformers-library on HuggingFace [48].

3.3 Implementation Details

Our model architecture is set up such that the components can be updated and replaced when more powerful models are released. For Intermediate Image Composition, we use SD-2.1 [31] as it provides stable image-to-image and inpainting capabilities. To generate the intermediate composition of the object and the background, we use 75 diffusion steps with a prompt-guidance strength of 15. For the second stage in the architecture, we use SDXL [28] with its image-to-image implementation. The prompt-guidance strength is 7.5 and we noise the intermediate image for 10 out of 50 steps before denoising it again. With the scaled linear scheduler, this corresponds to the image being noised by $\sim 20\%$. For colorization, we use SDXL [28] for a total of 30 steps with an image-to-image strength of 91%, and a prompt-guidance strength of 17. The upscaling is done using SD-1.5 [31]. We provide a discussion of our hyperparameters in the supplementary materials.

4 Experiments

Our experiments are twofold. For object injection into an existing background, we compare TF-Icon [16], AnyDoor [6] and our method. TF-Icon offers two configurations for inserting an object while maintaining style and domain and for inserting objects in arbitrary domains. For our evaluation, we use the same-domain configuration, as it provides significantly better results. ObjectDrop [47] and PrimeComposer [46] do not provide their code-base or a publicly available implementation to evaluate their methods independently. The standard SD [31] inpainting method shipped with the model has already been evaluated in the TF-Icon [16] paper and is outperformed by it. The same is true for Paint-by-Example [50], which is already outperformed by AnyDoor [6].

For generating a new background, we compare our method with ReplaceAnything [3] and the Background Replacement method inspired by a HuggingFace space provided by Shopify (SBR)[38]. At the time of writing this paper, ReplaceAnything has not released code. We therefore use their demo space on Huggingface for our experiments.

A high-level comparison of capabilities of relevant existing works and InsertDiffusion is provided in Table 1.

Table 1: **Capability Overview.** Our model (InsertDiffusion) provides a variety of functionalities compared to similar approaches. Besides being training-free, we allow for composition with both new and existing backgrounds, object consistency, the processing of CAD-like images and custom user-positioning of the object.

Model	Training-Free	New BG	Existing BG	Consistency	CAD-Images	Seamless Blending	Positioning	Text-Guidance
Poisson Blending [26]	✓	✗	✓	✓	✗	✗	✓	✗
SD-Inpaint [31]	✓	✗	✓	✗	✗	✗	✗	✓
InpaintAnything [51]	✓	✓	✓	✓	✗	✗	✓	✓
Any-Door [6]	✗	✗	✓	✗	✗	✓	✓	✗
TF-Icon [16]	✓	✗	✓	✗	✗	✓	✓	✓
SBR [38]	✓	✓	✗	✓	✗	✓	✗	✓
ReplaceAnything [3]	✗	✓	✗	✓	✓	✓	✗	✓
InsertDiffusion (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

4.1 Benchmarks

We evaluate our approach using two benchmark datasets. We derive the first from the benchmark used by TF-Icon [16]. Similar to their quantitative evaluation, we only use the Real-Real subset of their dataset to calculate metrics. Further, we filter their dataset by removing samples that already contain an object to be replaced in the target image as our method is not intended for object replacement.

The filtered benchmark contains 209 samples. The TF-Icon benchmark barely contains images from technical, design, and advertisement domains. Hence, we construct a second dataset to evaluate the capabilities of inserting technical and design products into backgrounds. We use bicycle images from the BIKED dataset [30], car images from Stanford-Cars [15] and catalog-images of consumer products from Amazon-Berkeley-Objects [7] and Products10K [2]. From each of the three categories, we select 20 samples randomly while manually labelling product types and color. For insertion into existing backgrounds we generate backgrounds using SDXL and assign backgrounds to objects at random. For the second task of inserting objects into a newly generated background, we only use the objects from our benchmark dataset and assign background prompts randomly.

4.2 Metrics

In our evaluation, we aim to assess the overall image quality and appeal of the resulting image composition, the alignment with the text-prompt and the geometric consistency of the inserted object compared to the ground-truth. We use the HPSv2-score [49] for overall image appeal, as it aims to replicate human preferences for natural and realistic images. To assess the alignment of the image composition with the text-prompt, we use the CLIP-score that measures the cosine similarity of the CLIP-embedded text and image [29]. For geometric consistency, we use LPIPS [53]. In addition to the automated metrics, we organize a user study with 15 participants to rate the overall image quality and appeal, the consistency with the text-prompt and the geometric consistency with the ground-truth of the composed image. We therefore randomly select 7 samples from each of our 3 benchmark categories and carry out the object-insertion task with the corresponding methods. We again compare our method to TF-Icon [16] and AnyDoor [6] for composition with an existing background with ReplaceAnything [3] and SBR [38] for composition with a newly generated background. To ensure objectivity, the human evaluation study is conducted blind with the shown examples being in random order.

4.3 Composition into Existing Background

In Fig. 4 we present some examples of the comparison of our approach with the existing alternatives TF-Icon [16] and AnyDoor [6]. The quantitative results and the results from the human evaluation study are summarized in Table 2.

In terms of image quality, our approach yields more appealing image compositions. The target object is embedded more realistically into the semantic framework of the existing background. The overall image appears more consistent. Our approach especially excels in visualizing fine-grained or beam-like structures. This is apparent with the bicycle images. Due to the adaptive masking approach, our method can handle empty spaces within an object and simultaneously preserve the structured geometry. While the quantitative evaluation using

Fig. 4: **Qualitative comparison with existing methods for insertion of product images into existing backgrounds**, including TF-Icon [16] and AnyDoor [6]. Our method improves seamless integration of the object into the background while preserving the geometry and structural integrity of the object.

HPSv2 [49] moderately hints towards the supremacy of our approach, the human evaluation study strongly favors our approach over the alternatives.

For the alignment of the composed image with the text-description, InsertDiffusion outperforms the alternatives across all benchmarks and metrics. While TF-Icon [16] is within range according to the CLIP-score, it falls noticeably short when evaluated by human annotators. AnyDoor [6] does not allow for the formulation of prompt and therefore consistently achieves the lowest score for both CLIP and human evaluation. The most significant gap again exists for structural objects like bicycles. We suspect that this is due to the bicycle geometries being severely altered, sometimes even rendered unrecognizable, by the alternative methods.

In analyzing the results for geometric consistency, we have to differentiate between the quantitative metrics and the human evaluation. For the quantitative LPIPS-score [53], all three models achieve similar results. On its own benchmark, TF-Icon [16] holds a 6% advantage over InsertDiffusion. For the remaining datasets, the scores are almost identical. AnyDoor [6] yields to better geometric consistency for the bicycle, car and consumer-product examples, according to the LPIPS-score. The human evaluation study paints a vastly different picture for geometric consistency. Averaged over our three benchmark sets, human annotators rate the geometric consistency of InsertDiffusion better by a factor of 2.4 over AnyDoor [6]. TF-Icon [16] performs reasonably well overall, but is outperformed by our approach by a factor of 2.97 for the bicycle samples.

We assume the reasons for the sharp difference between the quantitative results in LPIPS-score and the human evaluation to be twofold. The LPIPS-score measures the perceptual similarity of two images patches. It does not directly evaluate the structural composition of these patches and therefore most likely does not capture finer geometric details. By comparing solely the structural similarity, image fidelity and realism are not accounted for. This leads to unrealistic image compositions receiving a better LPIPS-score despite being visually and

Table 2: **Comparison results** for inserting an object into an existing background. "HPSv2", "CLIP" and "LPIPS" measure image appeal, text-alignment and geometric consistency. We compare our approach to TF-Icon [16] and AnyDoor [6].

Dataset	Model	CLIP (↑)	HPSv2 (↑)	LPIPS (↓)	Overall Appeal (↑)	Prompt Alignment (↑)	Geometric Consistency (↑)
TFI-Bench	*TF-Icon*	31.043	0.245	**0.589**	—	—	—
	AnyDoor	29.889	0.194	0.605	—	—	—
	Ours	**31.801**	**0.250**	0.624	—	—	—
Overall	*TF-Icon*	33.148	0.265	0.696	2.780	3.752	3.333
	AnyDoor	29.2982	0.224	**0.652**	1.905	2.762	1.695
	Ours	**34.997**	**0.287**	0.699	**3.410**	**3.790**	**3.790**
Bikes	*TF-Icon*	34.281	0.269	0.743	1.711	2.033	1.156
	AnyDoor	30.804	0.231	**0.709**	2.038	2.714	2.114
	Ours	**36.058**	**0.286**	0.757	**3.211**	**3.900**	**3.433**
Cars	*TF-Icon*	33.121	0.287	0.679	2.781	3.752	3.333
	AnyDoor	27.637	0.230	**0.654**	1.905	2.762	1.695
	Ours	**34.979**	**0.310**	0.672	**3.410**	**3.790**	**3.790**
Products	*TF-Icon*	31.987	0.238	0.647	2.743	3.248	1.895
	AnyDoor	29.453	0.213	**0.621**	2.343	2.895	2.257
	Ours	**33.955**	**0.267**	0.642	**3.571**	**4.162**	**3.695**

semantically unappealing. An example of this can be found in Figure 9 in the appendix. As a second reason we suspect that the annotators in our human evaluation study generally prefer more realistic image compositions. With AnyDoor [6], the object often appears to have just been pasted onto the background, while many objects get significantly altered by TF-Icon [16].

The objective of our work is to seamlessly visualize technical objects in different backgrounds. Therefore we need to preserve structural integrity as well as achieve realistic images. This is hard to capture in terms of pure quantitative metrics, which is why we deem the results of the human evaluation study important. The results confirm that InsertDiffusion reliably outperforms its alternatives for the given task.

4.4 Composition with Generated Background

To compare our approach with alternatives for inserting the product-image representations into generated backgrounds, we use the same evaluations as in the previous section. The quantitative results and the results from the human evaluation are summarized in Table 3. Since AnyDoor [6] and TF-Icon [16] do not provide for the option of novel background generation, we compare our approach to ReplaceAnything [3] and SBR [38]. Both methods are specifically designed to generate new backgrounds for the object to be inserted into.

In the quantitative analysis, our approach achieves superior results for human preference (HPSv2) and alignment with the text description (CLIP). On our benchmark dataset composed of bicycle, car and product images, we surpass ReplaceAnything by 8.67% and SBR by 16.60 % for the human preference score.

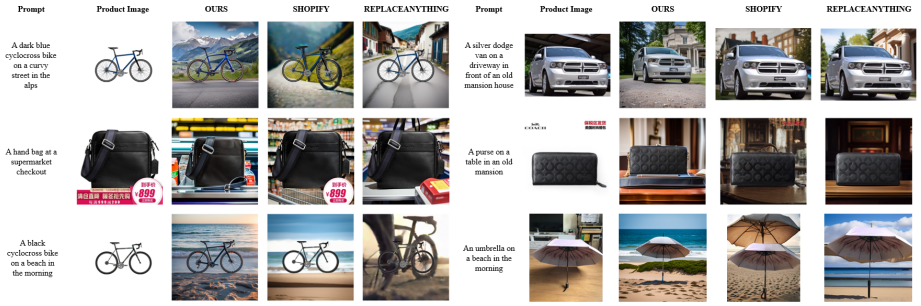

Fig. 5: **Qualitative comparison for insertion of product images into newly generated backgrounds**, including ReplaceAnything and SBR. Our method composes the object and the background in a more natural manner, being able to adapt the object to fit the background seamlessly while preserving its key geometric and semantic characteristics.

Table 3: **Comparison results** for inserting an object into a newly generated background. "HPSv2", "CLIP" and "LPIPS" quantitatively measure image appeal, text-alignment and geometric consistency. Overall appeal, prompt- alignment and geometric consistency are evaluated qualitatively in our human evaluation study on a scale of 1 to 5, 5 being the best. We compare our approach to SBR [38] and ReplaceAnything (ReplAny) [3].

Dataset	Model	CLIP (↑)	HPSv2 (↑)	LPIPS (↓)	Overall Appeal (↑)	Prompt Alignment (↑)	Geometric Consistency (↑)
Overall	*ReplAny*	31.070	0.265	**0.244**	3.057	3.600	**4.140**
	SBR	30.264	0.247	0.435	2.016	2.686	3.397
	Ours	**33.710**	**0.288**	0.403	**3.917**	**3.905**	3.498
Bikes	*ReplAny*	33.275	0.277	**0.213**	1.657	2.914	**3.486**
	SBR	31.449	0.245	0.527	1.610	2.467	2.762
	Ours	**33.810**	**0.296**	0.474	**4.181**	**3.971**	3.419
Cars	*ReplAny*	28.722	0.285	**0.280**	3.762	3.971	**4.610**
	SBR	28.914	0.270	0.438	2.152	2.829	3.619
	Ours	**33.837**	**0.301**	0.305	**3.990**	**4.086**	3.971
Products	*ReplAny*	30.977	0.236	**0.242**	**3.743**	**3.914**	**4.324**
	SBR	30.430	0.227	0.341	2.286	2.762	3.810
	Ours	**33.483**	**0.268**	0.430	3.705	3.752	3.210

For the CLIP-score, our approach holds an 8.50% advantage over ReplaceAnything and 11.39% over SBR. In terms of geometric consistency, ReplaceAnything performs best according to the LPIPS-score. The object geometries in the composed image are more consistent with the original. However, this consistency comes at the cost of sacrificing the quality of the overall image composition, as discussed before and shown in Figure 9. As shown in Fig. 5, some objects appear to be simply cut out and pasted onto the new background. This is especially true for bicycle images. In masking the objects adaptively and then allowing for marginal modifications, the compositions of our method look more realistic and seamless. The human evaluation study supports this conclusion. While the geometric consistency of ReplaceAnything is preferred by the evaluators, our approach generates vastly more appealing image compositions. Over the entire benchmark dataset, we achieve a 28.13% better evaluation score.

4.5 Ablations

To verify the efficiency of our approach, we perform a number of ablations. For the components of the InsertDiffusion architecture, we compare different model versions available within Diffusers. We also investigate the influence of the last refinement step, whether increased human interference in the generative pipeline leads to more appealing image compositions and image colorization.

InsertDiffusion Architecture. A significant ablation to the architecture is to leave out the refinement step where the intermediate image composition is noised and then denoised again using SDXL. To verify the usefulness of this additional step, we evaluate the results of our architecture with and without the refinement step on our benchmark dataset as well as on the TF-Icon benchmark. The results are summarized in Table 4 and visualized in Figure 6. The final images are more appealing when using the additional refinement step and show increased consistency with the text description. The intermediate compositions show an increased geometric consistency for some cases. This is somewhat expected since the refinement step adds noise to the object and then denoises it solely based on the guidance from the text-prompt.

Our modularized architecture (see Fig. 2) allows for the utilization of different versions of models from Diffusers. For our architecture, we find that using SDXL with the available inpainting function leads to worse results than using SD-2.1. For the intermediate composition, the generated backgrounds only contain faint and abstract structures. This is most likely an issue with the inpainting implementation. Furthermore, we observe that increasing the guidance scale reduces the quality and conistency of the inserted object. Using SDXL to synthesize the intermediate image composition does not allow for realistic visualizations. Examples generated using these ablated architectures are found in the supplementary material.

Additional User Interference. We investigate the effect of giving the user additional control by allowing to choose between 5 variations of the intermediate and the final image compositions. We find that additional interference has no

Fig. 6: Ablation for the image refinement. Top row shows the image composition before refinement, bottom row shows the refined image composition. Best viewed when zoomed in.

Table 4: Ablation for refinement stage. Minor stylistic changes of the object can occur due to the refinement.

Dataset	Metric	Ours	Ours w\o refinement
TFI Benchmark	*HPSv2 (↑)*	0.250	0.230
	CLIP (↑)	31.801	30.729
	LPIPS (↓)	0.623	0.611
Overall	*HPSv2 (↑)*	0.287	0.275
	CLIP (↑)	34.997	34.030
	LPIPS (↓)	0.699	0.701

significant impact on the overall appeal and the alignment with the text description. It does seem to have a small impact in improving the geometric consistency. More details are provided in the supplementary materials.

Colorization. For colorization, we compare masked SDXL, SDXL together with a ControlNet Sketch-to-Image adapter [52] and SDXL colorization after upsampling the input image. We observe that without upsampling, low guidance leads to the object not being colorized at all while high guidance leads to unwanted parts within the image being colorized. Upsampling the input image before passing it to a masked SDXL image-to-image transformation colorizes the geometry reliably.

5 Limitations and Future Work

A primary limitation of our approach is its dependence on adequate scaling and positioning of the inserted object. Our model cannot automatically detect where to place the object in a given background. Object misplacement can lead to unrealistic scenarios. Since we utilize pretrained latent diffusion models without finetuning them, our approach is limited by their generative capabilities. For example, with the current selection of models, we can not accurately generate or maintain text within images. The utilization of Diffusion Transformers (DiTs) might improve this capacity [25, 11]. Another limitation is the semantic and geometric consistency of the inserted objects, due to the final refinement step. Future research may explore approaches to ensure the consistency of the inserted objects by extracting targeted image features of the original object and injecting them into the final refinement step. Two promising approaches in the field of identity guidance are Readout Guidance [19] and InstantID [44].

6 Societal Impact

Our approach can enable improved visualizations in technical design processes and potentially lead to more user-centered experiences in digital product market-

ing. By eliminating the need for additional training and finetuning, we provide an efficient with a low skill barrier to create creative product visualizations.

However, this does not come without potential risks. It can be misused to create fake images of real objects, thereby contributing to misinformation or deception, which is a known issue for diffusion models for image generation [32]. The model might be used to create unethical content or violate privacy by placing humans or their personalized objects in compromising situations. Additionally, the automation of tasks in content creation might affect jobs in fields like photography, graphic design or marketing. By being based on SD, InsertDiffusion may inadvertently amplify biases present in the training data of SD [31, 10].

7 Conclusion

With InsertDiffusion, we present a novel, training-free approach for inserting objects into existing or newly generated backgrounds while preserving identity and structural integrity. Unlike previous methods requiring extensive fine-tuning or model-specific training, our approach leverages the inherent capabilities of state-of-the-art diffusion models. By adaptively masking the target area for the inserted object and using a step-wise combination of inpainting and image-to-image transformations, combined with a two-step refinement process, we achieve seamless compositions with the backgrounds and are able to visualize the objects in realistic scenes.

Our approach outperforms alternative methods in terms of image quality and alignment with textual descriptions and achieves on-par object consistency. It excels at visualizing technical and CAD-like images. The modular architecture of InsertDiffusion allows for easy adaptability, for example if more powerful diffusion-based models for image generation are released.

Bibliography

[1] Alex Andonian, Sabrina Osmany, Audrey Cui, YeonHwan Park, Ali Jahanian, Antonio Torralba, and David Bau. Paint by Word, March 2023.

[2] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset, 2020.

[3] Binghui Chen, Chao Li, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. ReplaceAnything as you want: Ultra-high quality content replacement, 2024.

[4] Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. VirtualModel: Generating Object-ID-retentive Human-object Interaction Image by Diffusion Model for E-commerce Marketing, May 2024.

[5] Ruidong Chen, Lanjun Wang, Weizhi Nie, Yongdong Zhang, and An-An Liu. Anyscene: Customized image synthesis with composited foreground. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8724–8733, June 2024.

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot Object-level Image Customization, July 2023.

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.02045.

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020.

[9] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021.

[10] Patrick Esser, Robin Rombach, and Björn Ommer. A note on data biases in generative models, 2020.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Neural Information Processing Systems*, December 2020.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023.

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-3022-7. doi: 10.1109/ICCVW.2013.77.

[16] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2294–2305, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00218.

[17] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models, August 2022.

[18] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence, May 2023.

[19] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features, 2024. URL https://arxiv.org/abs/2312.02150.

[20] Luca Medeiros. Language segment-anything, 2023. URL https://github.com/luca-medeiros/lang-segment-anything.

[21] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models, November 2023.

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models, March 2023.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024.

[24] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-Guided Diffusion with Inside-Outside Attention, March 2023.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[26] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pp. 313–318, New York, NY, USA, 2003. Association for Computing Machinery. ISBN

1581137095. doi: 10.1145/1201775.882269. URL https://doi.org/10.1145/1201775.882269.

[27] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Björn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. State of the Art on Diffusion Models for Visual Computing, 2023.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, July 2023.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, 2021.

[30] Lyle Regenwetter, Brent Curry, and Ahmed Faez. BIKED: A Dataset and Machine Learning Benchmarks for Data-Driven Bicycle Design. In *International Design Engineering Technical Conference & Computers and Information in Engineering*, 2021. doi: 10.1115/DETC2021-71681.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the CVPR*, 2022.

[32] Nega Rostamzadeh, Emily Denton, and Linda Petrini. Ethics and creativity in computer vision, 2021.

[33] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models, May 2022.

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, May 2022.

[35] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage Diffusion, August 2023.

[36] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation, 2023.

[37] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing, December 2023.

[38] Shopify. Shopify Image Background Replacement, 2023.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2022.

[40] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion, 2023.

[41] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR 2017*, 2017.
[42] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, 2022.
[43] Patrick von Platen, Patil Suraj, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, and Mishig Davaadorj. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2024.
[44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. URL https://arxiv.org/abs/2401.07519.
[45] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is All You Need for Image-to-Image Translation, May 2022.
[46] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. PrimeComposer: Faster Progressively Combined Diffusion for Image Composition with Attention Steering, March 2024.
[47] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. ObjectDrop: Bootstrapping Counterfactuals for Photorealistic Object Removal and Insertion, March 2024.
[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[49] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis, September 2023.
[50] 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, November 2022.
[51] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.
[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *International Conference on Computer Vision*. arXiv, 2023.
[53] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, April 2018.
[54] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. *IEEE International Conference on Computer Vision (ICCV) 2015*.

A Appendix

A.1 More Results

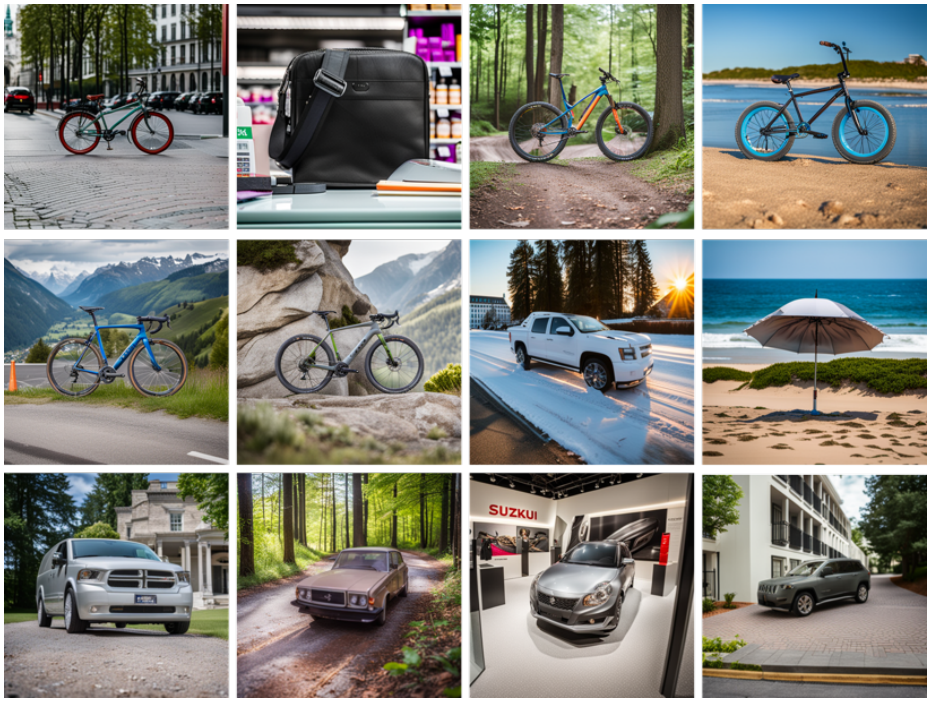

Fig. 7: More examples of achieving realistic object insertion into novel backgrounds using InsertDiffusion

A.2 Poisson Image Blending

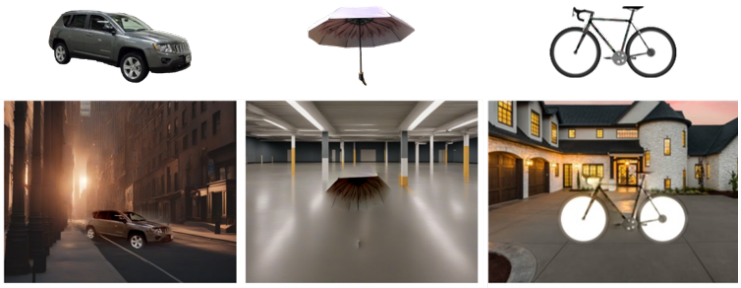

Fig. 8: Object insertion into given background using Poisson image editing [26]. Top row shows the object with its' background already removed and bottom row shows the merged image.

A.3 LPIPS Metric

Fig. 9: Analysis of the LPIPS metric for structural similarity. The LPIPS score favors the results produced by ReplaceAnything despite it failing to merge or harmonize the object with the background. Maintaining the identity of the object comes at the cost of image fidelity and sometimes produces unrealistic results. The image composition produced with InsertDiffusion performs worse in terms of LPIPS-score but produces a perceptually appealing and realistic image.

A.4 Human Evaluation Study

The quantitative metrics we employ to evaluate InsertDiffusion and existing alternatives do not capture the task entirely. We therefore conduct an evaluation study with human annotators who rate the results of the methods. Our study is conducted in a blind and randomized manner, meaning that the human evaluators do not know which result was obtained by which method and the presented images are in random order. We employed a total of 15 participants for the study.

Each participant rated seven images of each method. The prompts, reference objects, and composition backgrounds were held constant across methods. The participants rated the general appeal and realism of the composed image, its alignment with the given text prompt and the geometric consistency between the object in the reference image and in the composed image. The rating was given on a five point Likert scale where 1 denoted a very poor result and 5 a perfect result.

A.5 Inserting Humans into Backgrounds

InsertDiffusion has the purpose of inserting (technical) objects into different backgrounds. Our experiments on inserting images of humans into new scenes

show that, although the insertion is not entirely seamless, it certainly is possible. Minor alterations are visible in the composed image, especially in the area of the face of the human. Figure 10 shows two examples.

Fig. 10: Insertion of a human into new backgrounds.