Teleportraits: Training-Free People Insertion into Any Scene

Jialu Gao
Carnegie Mellon University
jialug@andrew.cmu.edu

K J Joseph
Adobe Research
josephkj@adobe.com

Fernando De La Torre
Carnegie Mellon University
ftorre@cs.cmu.edu

Abstract

*The task of realistically inserting a human from a reference image into a background scene is highly challenging, requiring the model to (1) determine the correct location and poses of the person and (2) perform high-quality personalization conditioned on the background. Previous approaches often treat them as separate problems, overlooking their interconnections, and typically rely on training to achieve high performance. In this work, we introduce a unified **training-free** pipeline that leverages pre-trained text-to-image diffusion models. We show that diffusion models inherently possess the knowledge to place people in complex scenes without requiring task-specific training. By combining inversion techniques with classifier-free guidance, our method achieves affordance-aware global editing, seamlessly inserting people into scenes. Furthermore, our proposed mask-guided self-attention mechanism ensures high-quality personalization, preserving the subject's identity, clothing, and body features from just a single reference image. To the best of our knowledge, we are the first to perform realistic human insertions into scenes in a training-free manner and achieve state-of-the-art results in diverse composite scene images with excellent identity preservation in backgrounds and subjects.*

1. Introduction

Human-centric personalized image content creation has received a lot of attention in recent years due to the increasing demand in commerce for customized experiences, including e-commerce advertising [9, 46], avatar creation [48, 56], and virtual try-on [7, 40, 55, 62]. Recent advances in large-scale text-to-image diffusion models [8, 17, 30, 37, 39, 42] have given rise to many customization methods that can generate images with the same individual in different scenes, poses, and styles. In this work, we study the problem of personalized human insertion into any scene, or "person teleportation": Given a scene image and a human reference image, how can we perform personalized human insertions into the background?

Figure 1. **Illustration of Teleportraits**. Teleportraits can insert humans into scenes, while maintaining high degree of affordance.

There are two major challenges to this problem: insertion and personalization. The first challenge is highly associated with the concept of affordances, proposed by J.J. Gibson [14] to describe the functional visual relationship between subjects and scenes. To seamlessly insert human subjects into a given background, global affordances reasoning determines the optimal placement, while local affordance understanding refines the subject's precise pose and action. Previous work on global affordances includes human pose and action estimation conditioned on input scenes [4, 25, 47, 50]. However, these methods typically rely on training with smaller, curated datasets containing ground-truth annotations, which limits their performance in diverse real-world scenes. Another line of research focuses on local affordances [6, 21, 57], aiming to seamlessly synthesize subjects within a given background based on a user-specified location. Nevertheless, such location information is not always available, and absence of global understanding can significantly constrain the effectiveness of local affordance reasoning. Most recently, Text2Place [32] is the first work to consider both levels of affordances, proposing the use of Score Distillation Sampling loss [35] to optimize a human mask parameterized by Gaussian blobs. An off-the-shelf inpainting model [34] is then employed to generate human poses at the predicted mask location. Still, the two levels of affordances are treated as separate problems, and test-time tuning is required for each individual scene.

The second challenge is achieving personalization in affordance-aware insertion, generating realistic poses for the person while preserving their facial and clothing fea-

tures. Current personalization methods can be classified based on whether per-subject optimization is required. Methods such as Textual Inversion [12], DreamBooth [38], and LoRA [18] fine-tune the generation model on reference images to capture visual features. Other methods avoid inference-time tuning by training a lightweight adapter on paired datasets to learn how to extract visual features directly from reference images [31, 48, 58]. However, most personalized generation techniques are conditioned solely on textual inputs. With the introduction of ControlNet [60], structural controls like depth maps and scribbles can also guide generation, yet no prior work has explored conditional personalization using input background images. While, in theory, text-to-image personalization methods could be combined with an inpainting model to enable such conditional personalization, as demonstrated in Text2Place [32], the quality remains limited since these personalization methods are not specifically trained for inpainting tasks.

In this paper, we propose a unified framework, termed **Teleportraits**, that addresses both challenges simultaneously, in a training free manner. We demonstrate that current large-scale text-to-image diffusion models inherently possess the semantic knowledge required for affordance-aware human insertion. Furthermore, the internal representations of these models can effectively capture and transfer the visual features of the subject for personalization, eliminating the need for additional training or test-time tuning.

Our proposed pipeline operates in three steps. Given a reference subject image and a background scene image, we first approximate the initial noise latents that can reconstruct the two images using inversion techniques. Then, starting from the background noise latent, we apply classifier-free guidance [16] to direct the model in generating a human within the background using text prompts, for example, *"a person running on the curved road"*. Finally, by leveraging the reference noise latent, we extract internal feature representations from the diffusion model during the self-attention layers. This allows the generated images to attend to the subject patch through our mask-guided self-attention mechanism, ensuring identity preservation between the subject in the provided reference image and the final generated output.

To demonstrate Teleportraits's ability to perform affordance-aware human insertion into diverse scenes, we first evaluate Teleportraits on the dataset proposed in Text2Place [32] alongside prior methods. The results show that Teleportraits outperforms existing approaches in semantically meaningful human insertion with perfect background preservation. Moreover, we introduce new metrics to assess personalization quality. The quantitative and qualitative results demonstrate that Teleportraits excels in generating humans with both global and local affordances, while effectively preserving the subject identity, including facial features, clothing, and body shapes.

In summary, our contributions are as follows: First, we propose a training-free method capable of inserting any person into any scene. Second, we extend the Semantic Human Placement setting introduced in Text2Place to incorporate full-body personalization and introduce new metrics for its evaluation.

2. Related Work

2.1. Subject-driven Image Generation

Subject-driven text-to-image generation focuses on preserving the visual features of any given object or person while performing text-conditioned generation. One approach to capturing a subject's visual characteristics is per-subject optimization, which typically involves learning a special token for each subject [12, 19] and/or optimizing the network parameters [22, 38, 59]. However, these methods are computationally expensive, as they require optimization for each subject, and they often suffer from overfitting to the reference images. To address these limitations, test-time tuning-free methods have been explored. These approaches leverage pre-trained image encoders to extract visual features and fine-tune the model to condition generation on these extracted features, enabling identity-preserving subject synthesis [15, 23, 26, 27, 31, 48, 49, 52–54, 56, 58]. Another direction to solve this problem is through training-free approaches. Consistory [45] extends self-attention by copying keys and values from the reference image and employs DIFT [43] feature injection to maintain subject consistency in story generation. DreamMatcher [29] proposes Appearance Matching Self-Attention, which warps value patches from the reference image's self-attention layers using a correspondence mapping to ensure accurate appearance alignment. MagicFace [51] focuses specifically on facial identity preservation, utilizing extended self-attention with keys and values from reference images, guided by aggregated semantic masks. While these methods enable personalization, they are limited to text-conditioned generation and cannot incorporate background images as additional conditioning for personalization.

2.2. Affordance-aware Subject Insertion

The concept of affordance, introduced by J.J. Gibson [14], has inspired extensive research in understanding scene-human affordances [4, 25, 47, 50]. In image generation, studies on local affordances focus on seamlessly blending a subject into a background at a user-specified location. Anydoor [6] constructs a dataset of objects and backgrounds with location annotations from videos and finetunes a diffusion model to place objects at designated positions. Similarly, Kulal et al. [21] build a dataset of humans and backgrounds from videos, training a model to in-

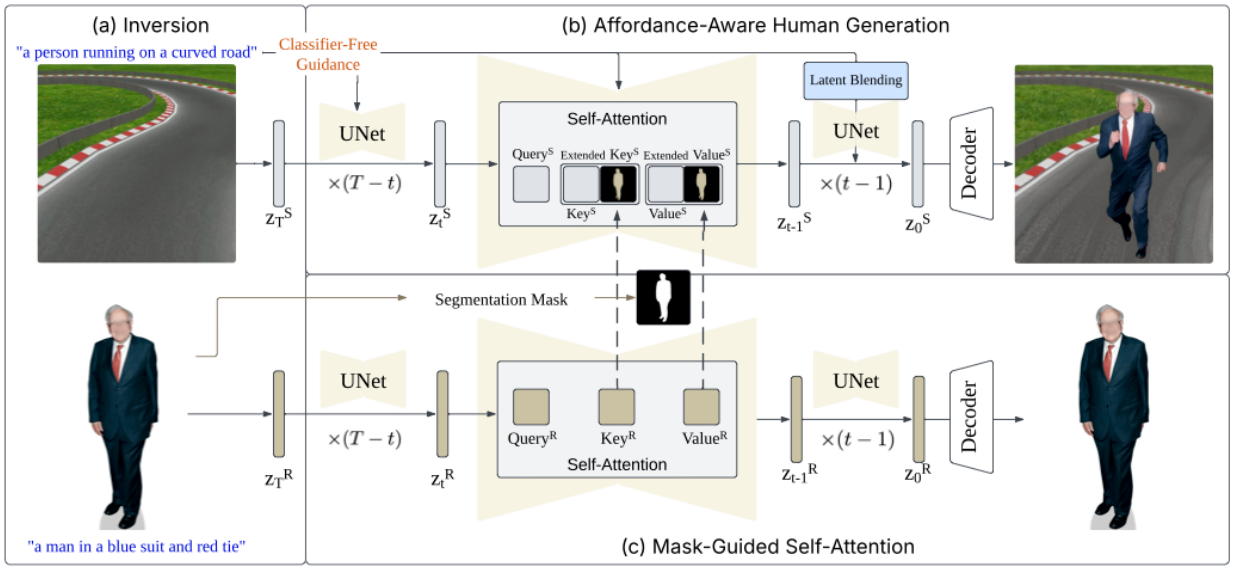


Figure 2. **Method overview.** Teleportraits consists of three steps: (a) Inversion, where we invert the input scene image and reference image into initial latent noise z_T^S and z_T^R. This allows Teleportraits to utilize the inherent semantic knowledge of diffusion models to place humans and use the hidden representation of diffusion models to perform personalization. (b) Affordance-Aware Human Generation. Starting from the inverted latent z_T^S of the scene image, Teleportraits uses classifier-free guidance to gradually guide the model to generate a human at reasonable locations with realistic poses following the text prompt. Latent blending is applied at later denoising steps to ensure background fidelity. (c) Mask-Guided Self-Attention. Teleportraits achieves personalization through an extended self-attention mechanism that additionally attend to the keys and values extracted from the recovered diffusion process that reconstruct the reference image.

sert people into predefined locations. While these methods successfully enable affordance-aware local subject insertion, they heavily depend on accurate insertion cues and curated datasets, limiting their effectiveness in diverse real-world scenes where such cues may not always be available. Studies on global affordances focus on estimating plausible human placement within a scene. Wang et al. [50] estimate plausible human poses given scenes by learning on extracted poses from sitcom videos. SmartMask [41] trains a diffusion model to predict fine-grained mask for subject insertion. Text2Place [32] leverages Score Distillation Sampling loss from DreamFusion [35] to optimize a subject mask parameterized by Gaussian blobs, with an off-the-shelf inpainting model used to generate the person within the mask. Although these methods can perform global affordance-aware subject insertion, they require either expensive test-time tuning for each scene, or rely on large-scale training which limits their ability to generalize to novel scenes. Moreover, they treat localization and human generation as separate problems, overlooking their interdependence, which can degrade the overall quality of the generated image. In contrast, Teleportraits unifies the two challenges and achieves training-free human insertion by leveraging the semantic understanding inherently captured by large-scale diffusion models.

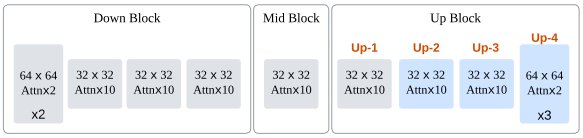


Figure 3. **SDXL architecture illustration.** SDXL consists of 70 attention layers, where each attention layer includes a cross-attention layer and self-attention layer. In Teleportraits, we apply self-attention based personalization on the up-2, up-3, and up-4 layers, as they determine the color, style, and texture details.

3. Preliminaries

3.1. Latent Diffusion Models

Latent Diffusion Models [37] are a family of diffusion models that use an autoencoder to project images onto the latent space and apply the standard diffusion process [17, 42]. The process begins with an initial Gaussian noise z_T and and undergoes a series of denoising steps, where at each timestep t, the model predicts $\epsilon(z_t, t)$ that will be used to compute z_{t-1} from z_t. At the final timestep $t = 0$, the model produces z_0, the final sampled image in latent space. In this work, we utilize the publicly available Stable Diffusion XL (SDXL) model for generation, where the model architecture is illustrated in Fig. 3.

3.2. Self-attention in Diffusion Models

Self-attention mechanism plays a crucial role in maintaining consistency in the style and structure of generation [3, 29]. A self-attention layer takes in a hidden feature $x_{in} \in \mathcal{R}^{d \times W \times H}$, where d is the feature dimension, and W and H are the resolution of the current attention layer, either 32 or 64 in SDXL model. The hidden feature x_{in} is then mapped to three intermediate representations denoted as Query $\mathbf{Q}$, Key $\mathbf{K}$, and Value $\mathbf{V}$ using linear projections $\mathcal{W}^Q$, $\mathcal{W}^K$, and $\mathcal{W}^V$. The self-attention operation is defined as follows:

$$x_{out} = softmax(\frac{Q^T K}{\sqrt{d}})V$$

3.3. Classifier-free Guidance

Classifier-free guidance is proposed by Ho et al. [16] to trade off controllability with sample fidelity. It samples the noise prediction twice, with and without the conditional text embeddings c, and amplifies the difference between them using a guidance weight w:

$$\hat{\epsilon} = \epsilon_\theta(x_t, t) + w \cdot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t))$$

4. Method

In this section, we present the key components of Teleportraits. Given an input scene image I^S and a generation prompt P^S describing the scene with human inserted, along with a reference image I^R and its corresponding text description P^R, Teleportraits aims to generate a human inside I^S that both aligns with the scene context described by P^S and resembles the person depicted in I^R.

Teleportraits consists of three steps, as shown in Fig. 2. Teleportraits first inverts the scene image I^S and the reference image I^R to obtain obtain their initial latent noise representations, z_T^S and z_T^R, respectively (Sec. 4.1). Then starting from z_T^S, Teleportraits leverages classifier-free guidance to direct the model in generating a human within the scene in accordance with the text prompt P^s (Sec 4.2). In the final step, Teleportraits introduces mask-guided self-attention, which transfers the visual features of the reference subject into the generated human, ensuring identity preservation while maintaining high-quality generation (Sec 4.3).

4.1. Inversion

The goal of inversion is to recover an estimated diffusion trajectory that can approximately reconstruct the input scene image I^S and the reference image I^R. Inspired by recent work in inversion in diffusion models [13, 28, 42], we adopt a similar approach to ReNoise-Inversion [13] and employ a fixed-point iteration strategy. Specifically, the images I^S and I^R are first encoded into the latent space, producing z_0^S and z_0^R. At each diffusion timestep $t = 0, 1, ..., T-1$, we want to estimate z_{t+1} with z_t. Since diffusion models are trained to predict z_t from z_{t+1} and not in the opposite direction, we initialize an estimate $z_{t+1}^{(0)}$ using DDIM inversion. Then, the estimated $z_{t+1}^{(0)}$ is passed through the UNet, which outputs a noise prediction $\epsilon_\theta^{(0)}$. This noise prediction $\epsilon_\theta^{(0)}$ is then used to renoise z_t, producing an updated estimate $z_{t+1}^{(1)}$. The same process is repeated across multiple iterations until an accurate estimation of the latent noise z_{t+1}. After going through all the timesteps, we obtain the estimated initial latent noise z_T.

Our experiments suggest that two iterations are sufficient to reconstruct the input images with high fidelity. As in ReNoise-Inversion, we disable classifier-free guidance during inversion to minimize accumulated error. However, unlike ReNoise-Inversion, we omit the noise averaging step in latent estimation to enhance numerical stability.

4.2. Affordance-aware Human Generation

The main idea of Teleportraits is to utilize the inherent semantic understanding in large-scale diffusion models to perform affordance-aware human insertion in a training-free manner. Once we recover the initial latent noise z_T^S for the input scene image, we want to use the text prompt P^S which describes a plausible and reasonable human insertion solution. By encouraging the model to follow P^S, we can achieve seamless human insertion into scenes. To this end, we start with the inverted latent noise z_T^S and utilize classifier-free guidance with a higher guidance weight of $w = 7.5$ to encourage the model to generate a person according to the prompt.

A problem with directly relying on text guidance to insert human into scenes is the final generated image may deviate from the original scene image, especially in the background area. To achieve high background fidelity during human generation, we adopt latent blending [2] to solve this problem. Specifically, we first perform an inference pass with classifier-free guidance to generate a human inside the scene image. While the background may change, the overall structure and layout of the scene remain faithful to the input I^S. Therefore, we use off-the-shelf segmentation model such as SAM [20] to obtain a foreground mask and a background mask. During the second inference pass, we apply latent blending using the two masks to ensure that the backgrounds are left intact and the human is being generated into the scene image.

4.3. Mask-guided Self-attention

The final part of Teleportraits is to perform personalized generation onto the scene image given reference image I^R. To achieve this, we propose mask-guided self-attention to use the hidden representation of diffusion models to transfer visual features of the subject, as depicted in Fig. 2(c).

As introduced in Sec. 3.2, at the core of the diffusion model is a UNet which includes 70 self-attention layers. From the latent noise z_T^R that reproduces the reference image, we can perform a forward diffusion pass and extract the keys and values patches at each self-attention layer. These patches, acting as a natural feature representation for the subject, will be used to transfer the visual identity feature of the subject onto the human being generated. During the generation process described in Sec. 4.2, we first retrieve the keys and values from the reference generation process. To ensure that these keys and values contain only visual information of the subject and not the background, we utilize segmentation models on the reference image I^R to obtain a subject mask and apply it on the retrieved keys and values. Then, the keys and values are concatenated with the original keys and values during affordance-aware human generation. In this way, the query patches can attend to both the keys and values from current generation and from the reference image, to allow seamless visual feature transfer.

Following previous works [1, 11], we only apply mask-guided self-attention on part of the up blocks, namely Up-2, Up-3, and Up-4 in Fig. 3, where the texture, style, and color of the image are being influenced the most.

The primary distinction between our mask-guided self-attention and the extended self-attention mechanism used in Consistory [45] is that we apply it solely to the conditional branch in classifier-free guidance, whereas Consistory applies it to both the conditional and unconditional branches. We find that applying it to both branches significantly impedes effective identity transfer. Additionally, Consistory requires an extra feature injection process to achieve consistent subject generation, while in Teleportraits, mask-guided self-attention alone is sufficient.

5. Experiments

We conduct extensive evaluations of Teleportraits to assess its ability to realistically insert humans into scenes. Given a scene image and a subject reference image, we prompt BLIP-2 [24] to generate a text description of human within the scene, such as "a person riding a bike on the street". We also generate a text summary of the subject, like "a man in a blue suit". During personalized generation, we replace the "a person" in the scene description with the subject's description to enhance identity preservation. We use SDXL as the default text-to-image model, employing a DDIM sampler with 50 inference steps. Mask-guided self-attention is applied throughout all inference steps, while latent blending is performed during the later denoising timesteps $t \in [10, 20]$ to achieve balance between background preservation and seamless subject integration. All experiments are conducted on a single NVIDIA V100 GPU with 32GB of memory.

Dataset. We evaluate Teleportraits on the dataset proposed in Text2Place [32], including 25 celebrities with 26 scenes. During evaluation, we insert each celebrity into all the 26 scenes, resulting in a total of 650 generated images. Notably, the celebrity dataset in Text2Place contains multiple reference images for each person. Since Teleportraits only requires one reference image, ideally full-body images as it would provide more information, we select one reference image per person as input for Teleportraits. Moreover, to ensure consistency in generation, we resize the reference images so that the human figures are roughly the same size, as shown in the Person column in Fig. 4.

Metrics. We first follow Text2Place and evaluate Teleportraits on these metrics for realistic human placement: **(1)** CLIP-T, which computes the cosine similarity between the CLIP [36] embeddings of the text prompt and the final generated image to measure prompt alignment. **(2)** Person generation, which uses SAM [20] to detect whether a human is being generated in the final image and calculate the percentage of successful human insertion. **(3)** Background preservation, which uses SAM to mask out the human subject and compute the LPIPS [61] score between the generated image and the original scene image to measure background fidelity. Furthermore, we include two metrics to measure identity preservation in personalization following prior works [38]: **(4)** CLIP-I, which calculates the cosine similarity between the CLIP Image embedding of the generated human image and the reference image. **(5)** DINO, which is the average cosine similarity between the ViTS/16 DINO [5] embeddings of the generated and reference human image.

VLM evaluation. As discussed in prior work [33], automated evaluations of personalization models can be misaligned with humans. For a more comprehensive evaluation, we adopt and extend the GPT-based evaluation protocol from [10] to assess subject identity preservation (**VLM-S**), text alignment (**VLM-T**), and background preservation (**VLM-BG**). Details of the evaluation pipeline and GPT prompts are provided in the Appendix. C.

Human evaluation. We perform human evaluation following DreamMatcher [35] with 51 users and 36 samples from the Text2Place dataset. Users rank the generated images according to subject identity, text alignment, and scene consistency. Additional details can be found in Appendix. D.

Baselines. We compare Teleportraits with multiple baselines to show its superior performance in both affordance-aware human placement and personalization. We first compare Teleportraits with **Text2Place** on the overall task of personalized human insertion into scenes. Since original Text2Place operates on multiple reference images, we also compare with Text2Place using a single reference image, which we call **Text2Place (single)**. Additionally, we compare Teleportraits's ability to perform affordance-aware human placement with Text2Place by using the human mask

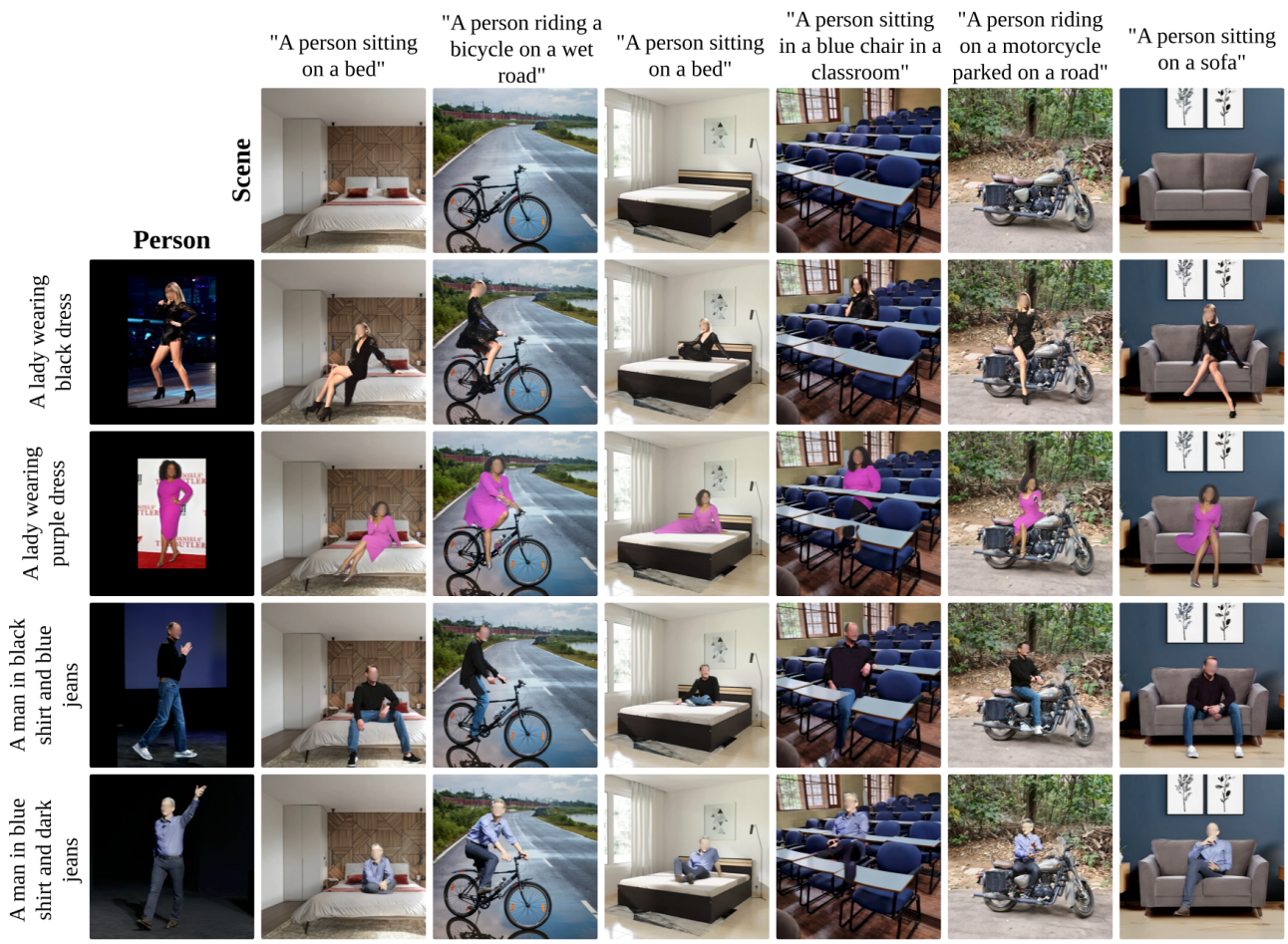


Figure 4. **Qualitative results.** Teleportraits can perform realistic human insertion into various indoor and outdoor scenes given just a single reference image. Results show that Teleportraits is able to reason the location and poses of the inserted human, while preserving the subject identity including hair style, clothing, and body shape.

extracted from our generated image as the mask for personalized inpainting in Text2Place, which is denoted as **Mask+Text2Place**. Lastly, we compare Teleportraits with a state-of-the-art, open-source zero-shot object insertion method, **AnyDoor** [6]. Since a mask is also required for Anydoor, we use the human mask extracted from our final generated image and measure Anydoor's performance on zero-shot personalized human insertion.

5.1. Qualitative Results

We first present the qualitative results of Teleportraits in Fig. 4. Our method can achieve realistic human insertion into various scenes following the text prompts, and can reason the correct human poses that interact with the objects in the scenes accordingly, such as riding on the bike or sitting in a chair. Moreover, our method can perform high-quality personalization given only a single reference image, perfectly maintaining the person identity, including hair style, clothing, and body features.

In Fig. 5, we compare Teleportraits with the baselines. Firstly, Text2Place sometimes fails to generate the most suitable mask blobs, leading to unsuccessful human insertion. In addition, Text2Place cannot successfully transfer clothing and body shape features, either with multiple reference images or with only a single reference image. In contrast, our method perfectly preserves the clothing details of the subjects without any test-time tuning. In Mask+Text2Place, we enhance Text2Place by using the bounding box of the generated person from Teleportraits as the subject mask for Text2Place's inpainting module. Results show that this leads to more successful inpainting, demonstrating that Teleportraits produces more accurate and affordance-aware human insertion. For Anydoor, we provide the extracted human bounding box from Teleportraits as input. While Anydoor preserves subject identity well, it fails to generate plausible human poses that mean-

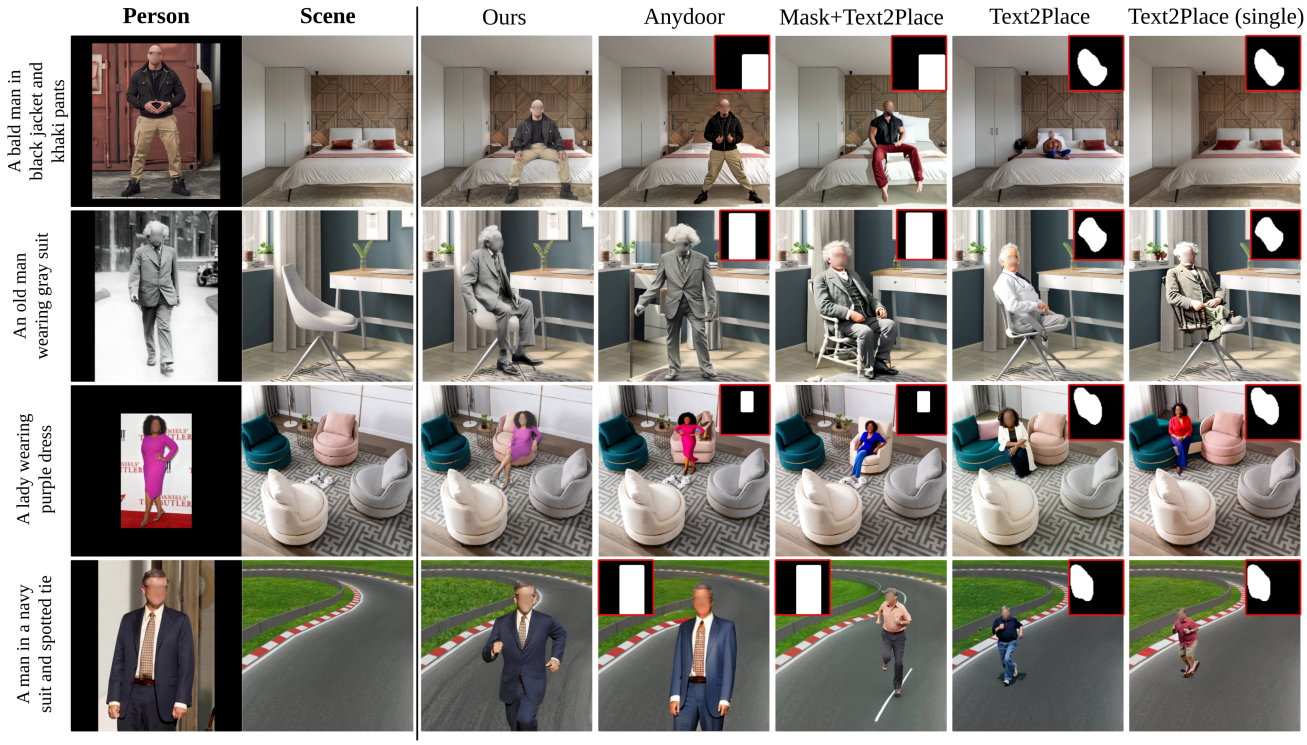


Figure 5. **Qualitative comparison with baselines.** Using the bounding box from the generated human in Teleportraits, Anydoor is able to insert human but fails to generate realistic human poses that interacts with the scene. Compared to Text2Place, Teleportraits can not only generate better location for human insertion, leading to better inpainting results, but also can preserve the human identity much better.

ingfully interact with surroundings. This is likely because Anydoor is trained on large-scale internet videos with objects, making it less effective for human-centric generation.

5.2. Quantitative Results

Next we present the quantitative evaluation results on Teleportraits and the baselines, with automated evaluation results in Table. 1 and human study results in Figure 6.

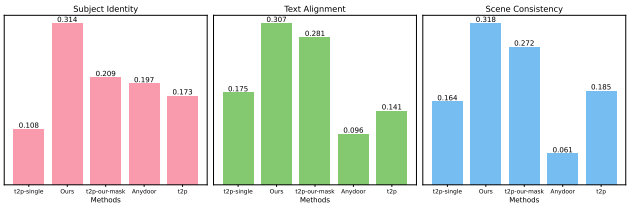


Figure 6. **Human Evaluation Results.** Following DreamMatcher [29], 51 users ranked 36 samples from the Text2Place dataset based on subject identity, text alignment, and scene consistency. Teleportraits achieves the highest scores across all aspects.

Teleportraits shows the best performance in text-to-image alignment and background preservation compared to all baselines, demonstrating its superior ability to semantically insert human into scenes while leaving the background almost unchanged. By replacing the predicted mask in Text2Place with ours, we are able to achieve higher human insertion success rate, demonstrating Teleportraits's state-of-the-art ability to accomplish the task of semantic human placement. Additionally, Teleportraits achieves a higher success rate in human insertion than Text2Place, suggesting that the pre-trained inpainting model used in Text2Place has worse performance compared to our training-free approach. This highlights the effectiveness of Teleportraits in handling human insertion tasks without relying on task-specific training, and emphasizes the advantage of using semantic knowledge within diffusion models to perform global editing. For personalization quality, Teleportraits achieves higher identity preservation scores compared to all baselines, except Anydoor. Anydoor's better performance in identity similarity is due to its tendency to copy-paste the human directly from the reference image into the scene, without adjusting the subject's pose or angle. As shown in Fig. 5, this results in a higher similarity between the reference image and the generated subject, but limits its ability to produce natural, dynamic human poses.

Overall, both the qualitative and quantitative results demonstrate Teleportraits's capability in solving the two challenges in human insertion into scenes: affordance-aware human insertion and conditional personalization given background scenes. Notably, Teleportraits achieves

	CLIP-T↑	Person (%)↑	LPIPS↓	CLIP-I ↑	DINO↑	VLM-S↑	VLM-T↑	VLM-BG↑
Text2Place	0.267	84.2	0.094	0.573	0.164	2.25	3.85	6.17
Text2Place (single)	0.267	86.2	0.093	0.567	0.180	1.93	3.79	6.00
Our Mask + Text2Place	0.269	91.8	0.075	0.572	0.177	2.18	4.17	6.02
Our Mask + AnyDoor	0.275	**100.0**	0.053	**0.681**	**0.447**	**4.05**	1.82	5.58
Ours	**0.287**	97.4	**0.025**	0.596	0.244	4.01	**4.93**	**6.29**

Table 1. **Quantitative results.** Compared to Text2Place, Teleportraits has better performance across all metrics. For affordance-aware human placement, Teleportraits can predict the location better, leading to high success-rate for Text2Place's inpainting pipeline to generate a human. While Anydoor can generate human more similar to the references, it falls short in following the text prompt, failing to generate semantically meaningful poses for the inserted human.

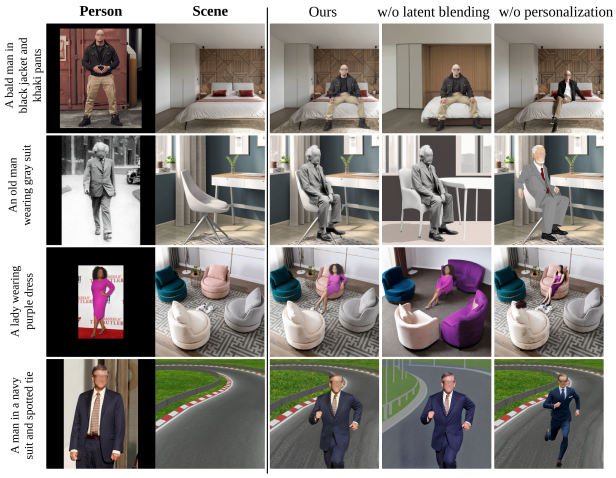


Figure 7. **Ablation study of Teleportraits.** Results show that latent blending plays an important role in background preservation, and our personalization module can successfully transfer detailed visual features of the reference subject into the generated image.

the task in a training-free manner, costing less than 1 min to generate a single image, while Text2Place requires per-subject optimization and can take up to 1 hour to generate a novel subject in a novel scene.

5.3. Ablations

We now present the ablation study on the following components in Teleportraits: (1) Latent blending during semantic human generation (2) Mask guided self-attention for personalization. The qualitative results are displayed in Fig. 7, and quantitative results are reported in Table. 2. More ablation results can be found in Appendix. B.

Latent blending plays a crucial role in preserving background fidelity. When removed, the overall structure of the image still resembles the input scene, but the detailed appearance of the background changed. This is because the structure is largely influenced by the initial noise [44]. Therefore, even under the influence of classifier-free guidance, when the generation starts from the same latent noise that reconstructs the scene image, it will automatically follow the scene layout and place humans at reasonable locations. However, the detailed appearance of the image can be largely influenced by the text prompt, thus latent blending is required to preserve the visual details of background.

	Ours	w/o Latent blending	w/o Personalization
CLIP-T↑	0.287	0.288	**0.289**
Person(%)↑	97.4	**98.9**	94.3
LPIPS↓	**0.025**	0.189	0.029
CLIP-I ↑	**0.596**	0.590	0.536
DINO↑	**0.244**	0.236	0.174

Table 2. **Ablation study on Teleportraits.** Quantitative results also demonstrate the role of latent blending in keeping the background unchanged. And metrics in subject similarity show that our personalization method greatly contributes to better subject identity preservation.

Personalization using mask-guided self-attention is crucial to preserving high-level and low-level subject identity. When generating without mask-guided self-attention, we can see that the model generates a human that roughly resembles the reference image, because the subject prompt already contains some high-level information about the appearance, such as the color of the clothing. However, text has limited granularity and thus cannot encode detailed visual features, such as the buttons on the jacket and the color of the spotted tie. This is where mask-guided self-attention comes in and performs detailed visual feature transfer from the reference image to the final generated image.

6. Conclusion

In this work, we present Teleportraits, a training-free method that builds on pre-trained diffusion models and successfully tackles the challenging problem of inserting humans into any scene with a single reference image. Teleportraits achieves affordance-aware human insertion by leveraging the semantic knowledge inherent in large-scale diffusion models, while transferring the visual features of the subject through a mask-guided self-attention process. We hope this work sheds light on the potential of training-free approaches that utilize the inherent knowledge of diffusion models to perform various image manipulation tasks.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis, 2023. 5

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

[3] 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, 2023. 4

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context, 2020. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization, 2024. 1, 2, 6

[7] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer, 2024. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[9] Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, et al. Towards reliable advertising image generation using human feedback. In *European Conference on Computer Vision*, pages 399–415. Springer, 2024. 1

[10] Jisu Nam et al. Visual persona: Foundation model for full-body human customization. *arXiv:2503.15406*, 2025. 5, 2

[11] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024. 5

[12] 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. *arXiv preprint arXiv:2208.01618*, 2022. 2

[13] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising, 2024. 4

[14] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979. 1, 2

[15] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization, 2024. 2

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2, 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1, 3

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[19] Jian Jin, Yang Shen, Zhenyong Fu, and Jian Yang. Customized generation reimagined: Fidelity and editability harmonized, 2024. 2

[20] 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, 2023. 4, 5

[21] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes, 2023. 1, 2

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 2

[23] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023. 2

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 5

[25] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[26] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 2

[27] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning, 2024. 2

[28] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 4

[29] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization, 2024. 2, 4, 7

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[31] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces, 2024. 2

[32] Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R. Venkatesh Babu. Text2place: Affordance-aware text guided human placement, 2024. 1, 2, 3, 5

[33] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation, 2025. 5

[34] 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. *arXiv preprint arXiv:2307.01952*, 2023. 1

[35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 1, 3, 5

[36] 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. 5

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 3

[38] 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. 2, 5

[39] 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. 1

[40] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*, 2024. 1

[41] Jaskirat Singh, Jianming Zhang, Qing Liu, Cameron Smith, Zhe Lin, and Liang Zheng. Smartmask: Context aware high-fidelity mask generation for fine-grained object insertion and layout control, 2023. 3

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1, 3, 4

[43] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. 2

[44] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024. 8

[45] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation, 2024. 2, 5, 1

[46] Haohan Wang, Wei Feng, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, and Jingping Shao. Generate e-commerce product background by integrating category commonality and personalized style. *arXiv preprint arXiv:2312.13309*, 2023. 1

[47] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes, 2021. 1, 2

[48] 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. 1, 2

[49] Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight, 2024. 2

[50] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms, 2018. 1, 2, 3

[51] Yibin Wang, Weizhong Zhang, and Cheng Jin. Magicface: Training-free universal-style human image customized synthesis, 2024. 2

[52] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation, 2023. 2

[53] Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, Lei Zhang, and Wangmeng Zuo. Masterweaver: Taming editability and face identity for personalized text-to-image generation, 2024.

[54] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention, 2023. 2

[55] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on, 2024. 1

[56] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds, 2023. 1, 2

[57] Jieteng Yao, Junjie Chen, Li Niu, and Bin Sheng. Scene-aware human pose generation using transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 2847–2855, New York, NY, USA, 2023. Association for Computing Machinery. 1

[58] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[59] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis, 2023. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 5

[62] Yuxuan Zhang, Qing Zhang, Yiren Song, Jichao Zhang, Hao Tang, and Jiaming Liu. Stable-hair: Real-world hair transfer via diffusion model, 2024. 1

Teleportraits: Training-Free People Insertion into Any Scene

Supplementary Material

A. Limitation

While Teleportraits has demonstrated state-of-the-art performance in the task of human insertion into scenes, there are some limitations to the method.

Firstly, Teleportraits performs the best with full-body images as reference, and will suffer from problems like low-quality personalization and disproportional human sizes if the reference image only contains the upper body, or only the face of the human (Fig. 8)

Figure 8. **Failure case 1.** When the reference image only contains a small part of the body, the personalized generation quality degrades.

Secondly, the quality of the generation is influenced by the text prompt, especially when the scene is complex or the person has many detailed visual characteristics to capture. For example, a shorter prompt like "a person sitting on the bed" will lead to worse result compared to a more detailed prompt like "a man wearing blue shirt and dark jeans sitting on the bed". Another example would be "a person sitting on the sofa" leads worse result compared to a more detailed prompt like "a person sits in the round sofa chair at one corner, surrounded by three empty chairs, top-down", on a scene containing multiple sofas captured from top-down view (Fig. 9). This is probably due to the bias in large-scale internet dataset that the diffusion model is trained on, but overall, for common scene images and people, the effort for prompt tuning is minimal.

B. More Ablation Results

Here we present more ablation studies on the hyperparameters used in Teleportraits.

Influence of classifier-free guidance scale. In Fig. 10, we present the effect of different classifier-free guidance scale has on the final generated images. With a guidance scale of 1, it is equivalent to disabling classifier-free guidance, and therefore only the scene image is reconstructed and

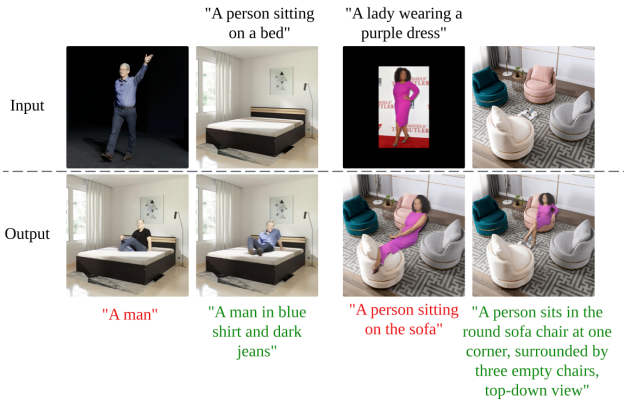


Figure 9. **Failure case 2.** The influence of text prompts with complex examples.

no human is being generated. With the guidance scale increasing, we can observe that the human being generated is getting clearer and clearer, taking up more space in the image. This is because a larger guidance scale will drive the generation more towards the direction of text prompt, where a human is included.

Influence of latent blending timesteps. In Fig. 11, we show how different latent blending timesteps influences the output images. We can observe that applying latent blending during earlier timestep results in more obvious changes in backgrounds. This is because diffusion models usually determine the structure and layout during early timesteps, and detailed appearances are determined during the later timesteps. When we move the t range to later timesteps, we can see that the background fidelity significantly increases. However, if we only apply latent blending right before the denoising process finishes, it may result in visual artifacts such as a glow surrounding the subject. Therefore, we choose to apply latent blending during $t \in [10, 20]$ in Teleportraits to achieve a balance between background preservation and overall image quality.

Influence of performing mask-guided on the unconditional branch. Here we compare our mask-guided self-attention mechanism with the one proposed in Consistory [45]. In particular, the main difference between our method and the one used in Consistory is that we are only applying the mask-guided self-attention on the conditional generation branch of classifier-free guidance. In contrast, Consistory applies it on both the conditional branch and unconditional branch during generation. We report the results in Fig. 12, which clearly indicates that applying mod-

Figure 10. **Influence of guidance-scale.** Results show that with a larger guidance scale, we can achieve better human insertion into scenes because the generation process will be guided more towards the text prompt, which describes the scene containing a human.

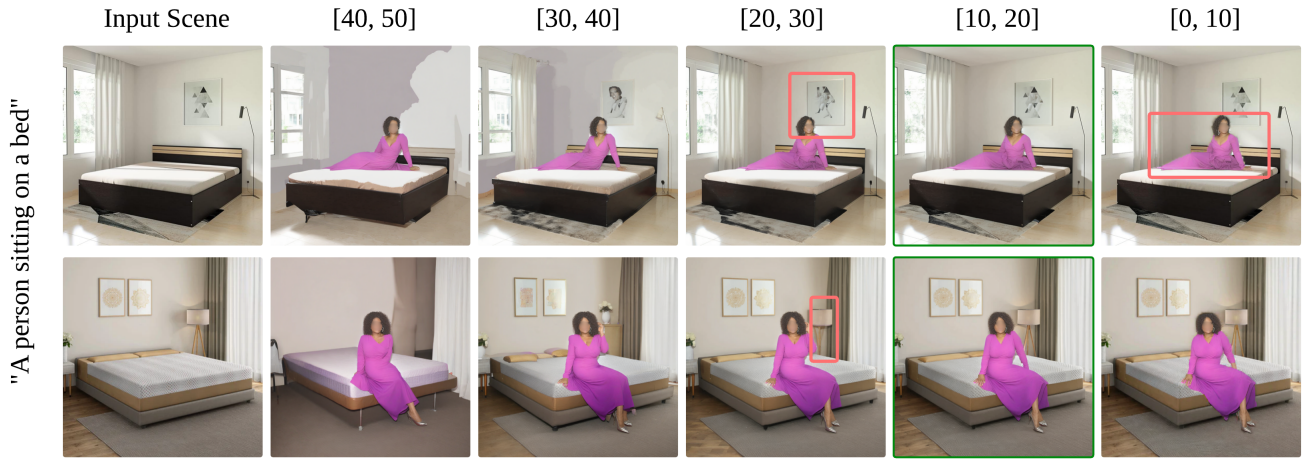


Figure 11. **Influence of latent blending timesteps**. We report results obtained by applying latent blending during $t \in [0, 10], [10, 20], [20, 30], [30, 40], [40, 50]$. The Denoising process starts from $t = 50$ and ends in $t = 0$, meaning that larger t indicates earlier diffusion steps, and smaller t represents later steps. Results show that applying latent blending during $t \in [10, 20]$ achieves a perfect balance between background preservation and seamless foreground blending.

ified self-attention on both conditional and unconditional branches during generation largely degrades the personalization quality, demonstrating Teleportraits's superior performance in transferring visual features from a single reference image into various scenes during human generation.

C. VLM Evaluation Details

Following the GPT evaluation protocol in [10], we designed three different prompts for evaluating Teleportraits's ability in subject identity preservation (Fig. 13), text alignment (Fig. 14), and background scene preservation (Fig. 15). The GPT model version is GPT-4o, and all evaluations are performed with a temperature of 0 and high image details.

D. Human Evaluation Details

We conducted a paired human preference study on subject fidelity, prompt alignment, and background fidelity, comparing Teleportraits to the baseline works as listed in Sec. 5 of the main paper. The results are summarized in Fig. 6 in the main paper.

We provide example questions of the user study. For subject fidelity, participants were presented with a reference

Figure 12. **Influence of whether applying self-attention feature transfer on the unconditional branch**. Results show that only applying mask-guided self-attention on the conditional branch as in Teleportraits can significantly increase the personalization performance, generating subjects highly similar to the reference.

image and several generated images using different methods, and were asked to rank the generated images according to which better represents the subject in the reference image, as shown in Fig. 16. For prompt alignment, the subjects were presented with the generated images alongside the text prompt used to generate these images, and were asked to rank the images according to which aligns best with the given prompt, as shown in Fig. 17. For background fidelity, the subjects were presented with the generated images alongs with the original scene image, and were asked to rank the images according to which aligns best with the original scene image, with an example shown in Fig. 18. A total number of 51 users responded to 36 ranking questions, resulting in a total of 1836 responses.

E. Implementation Details

E.1. Code Snippet

Task Definition

You will be provided with an image generated based on a reference image.

As an experienced evaluator, your task is to assess how well the appearance of the human subject is preserved in the generated image compared to the reference image, based on the scoring criteria.

Focus solely on the human subject. Regardless of whether the subject in the generated image differs in size, pose, action, or surroundings compared to the one in the reference image, your evaluation should prioritize the subject's visual appearance.

Scoring Criteria

Assess whether the human subject in the generated image remains consistent with the one in the reference image, focusing on the preservation of fine details across the following five visual features:

1. Clothing Types: Check whether the clothing types in the generated image match those in the reference image. This includes distinctions like short vs. long sleeves, short vs. long pants, and the presence of accessories.

2. Design: Evaluate whether the design of the subject's clothing in the generated image matches that in the reference image. This includes the pattern (e.g., floral, striped, or solid) and decorative elements (e.g., logos, zippers, or pockets). Focus on fine-grained details in the design.

3. Texture: Assess whether the texture of the fabrics worn by the subject in the generated image matches that in the reference image. This includes the material's appearance and quality. Focus on fine details that contribute to realism.

4. Color: Compare the primary colors of the subject's clothing and body in both images, considering hue, saturation, brightness, and overall color distribution.

5. Face Identity: Evaluate whether the subject's face in the generated image resembles the face in the reference image. It is acceptable for the subject in the generated image to have a different expression or pose than in the reference image. The focus should be on whether the facial identity aligns, without expecting an exact replica.

Scoring Range

You need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from 0 to 9:

- Very Poor (0): No resemblance. The generated image's subject has no relation to the reference. If no human is detected, assign a score of 0.
- Poor (1-2): Minimal resemblance. The subject falls within the same broad category but differs significantly in appearance.
- Fair (3-4): Moderate resemblance. The subject shows some likeness to the reference but has notable variances.
- Good (5-6): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
- Very Good (7-8): Very close resemblance. The subject of the generated image is similar to the reference, with few differences in details.
- Excellent (9): Near-identical resemblance. The subject of the generated image is virtually indistinguishable from the reference.

Input format

Every time you will receive two images, the first image is the generated image, and the second image is the referece image.

Please carefully review each image of the subject.

Output Format

[Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.'"

Figure 13. **GPT prompts for evaluating personalization quality.**

''' ### Task Definition
You will be provided with an image and a text prompt.
As an experienced evaluator, your task is to evaluate the semantic consistency between the image and the text prompt, focusing on human pose, human action, surroundings, composition and image quality, according to the criteria below.
Scoring Criteria
Assess how well the visual content of the image aligns with the text prompt based on the following five key aspects:
1. Human Pose: Assess whether the body pose of the human subject aligns with the pose described in the text (e.g., "stand" or "stretch out arms"). Focus on the subject's pose regardless of their size and position.
2. Human Action: Examine the action or movement of the human subject as described in the text prompt (e.g., "jogging," "climbing," or "walking"). Focus on the subject's action regardless of their size and position.
3. Surroundings: Evaluate whether the environment and background elements in the image are consistent with the text prompt. The surroundings should match the described context, including location, props, and overall atmosphere.
4. Composition: Assess how naturally the arrangement of the human subject in the generated image aligns with the description, considering variations in the subject's placement, position, and size.
5. Image Quality: Evaluate whether the overall image exhibits realistic fidelity, clarity, and visual appeal, avoiding an overly synthetic or artificial look.
Scoring Range
Based on these criteria, a specific integer score from 0 to 9 can be assigned to determine the level of semantic consistency:
- Very Poor (0): No correlation. The image does not reflect any of the key points or details of the text. If no human is detected, assign a score of 0.
- Poor (1-2): Weak correlation. The image addresses the text in a very general sense but misses most details and nuances.
- Fair (3-4): Moderate correlation. The image represents the text to an extent but lacks several important details or contains some inaccuracies.
- Good (5-6): Strong correlation. The image accurately depicts most of the information from the text with only minor omissions or inaccuracies.
- Very Good (7-8): Very strong correlation. The image captures nearly all relevant details from the text, with very few omissions or inaccuracies.
- Excellent (9): Near-perfect correlation. The image captures the text's content with high precision and detail, leaving out no significant information.
Input format
Every time you will receive an image and a text prompt.
Output Format
[Your Score]
You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.'''

Figure 14. **GPT prompts for evaluating prompt alignment.**

Task Definition

You will be provided with an image and reference scene image.

As an experienced evaluator, your task is to evaluate the scene consistency between the image and the original scene image, focusing on overall structure, visual details,

surroundings, composition and image quality, according to the criteria below.

Scoring Criteria

Focus solely on the background. The foreground object is a human, and you should only focus on the similarity of the background scene.

Your evaluation should prioritize the background scene's visual appearance compared to the original scene image, regardless of the human object.

If no human is detected in the generated image, assign a score of 0.

1. Background Structure: Assess whether the overall structure of the generated image aligns with the original scene image.

This includes evaluating the arrangement of elements and objects in depth, and perspective.

2. Background Visual Details: Examine the image for any visual details of the background that are missing, modified, or inaccurately represented.

Focus especially on the background elements around the foreground object with their details, such as textures, patterns, and colors.

3. Background Color Tone: Evaluate whether the color tone of the generated image matches the original scene image. Consider the overall color scheme and mood.

4. Composition: Assess how naturally the arrangement of foreground subject in the generated image aligns with the surrounding scene.

The surronding scene should be consistent with the reference image, and the foreground subject should be well integrated into the background.

5. Image Quality: Evaluate whether the overall image exhibits realistic fidelity, clarity, and visual appeal, avoiding an overly synthetic or artificial look.

You need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from 0 to 9:

- Very Poor (0): No resemblance. The generated image's background has no relation to the reference. However, if no human is detected, assign a score of 0.
- Poor (1-2): Minimal resemblance. The generated image's background differs significantly in appearance than the reference.
- Fair (3-4): Moderate resemblance. The generated image's background shows some likeness to the reference but has notable variances.
- Good (5-6): Strong resemblance. The generated image's background closely matches the reference with only minor discrepancies.
- Very Good (7-8): Very close resemblance. The generated image's background is very similar to the reference, with few differences in details.
- Excellent (9): Near-identical resemblance. The generated image's background is virtually indistinguishable from the reference.

Input format

Every time you will receive two images, the first image is a generated image, and the second image is the reference image.

Please carefully review each image of the background scene.

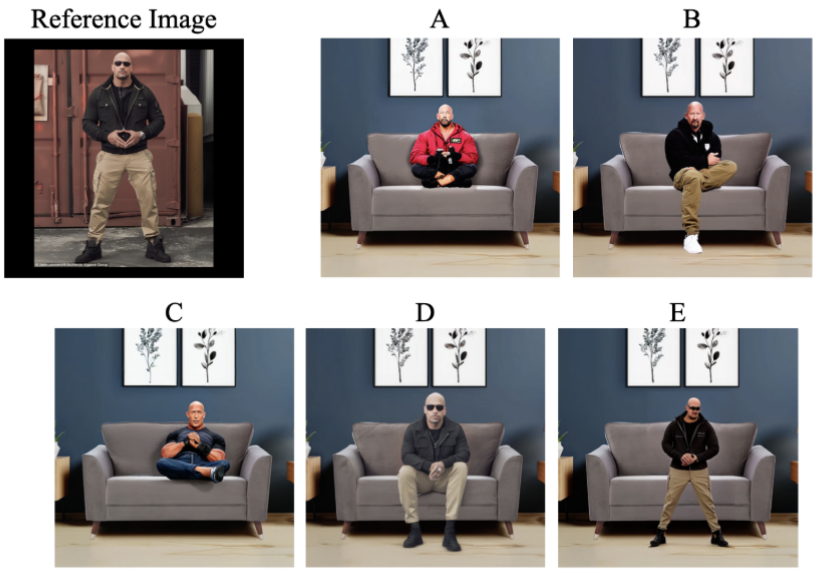
Output Format

[Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.'"

Figure 15. **GPT prompts for evaluating background fidelity during insertion.**

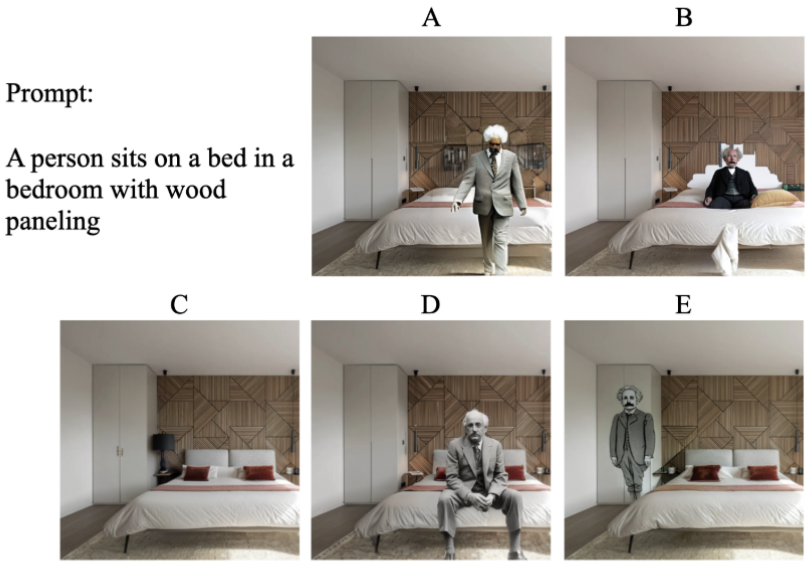
Please rank the generated images (A–E) based on **how closely they resemble** the **reference person**. Focus on both the person's identity (facial features) and clothing appearance. Rank 1 indicates the most similar and Rank 5 the least similar. *



	1	2	3	4	5
A	○	○	○	○	○
B	○	○	○	○	○
C	○	○	○	○	○
D	○	○	○	○	○
E	○	○	○	○	○

Figure 16. **Example questionnaire for evaluating subject fidelity.**

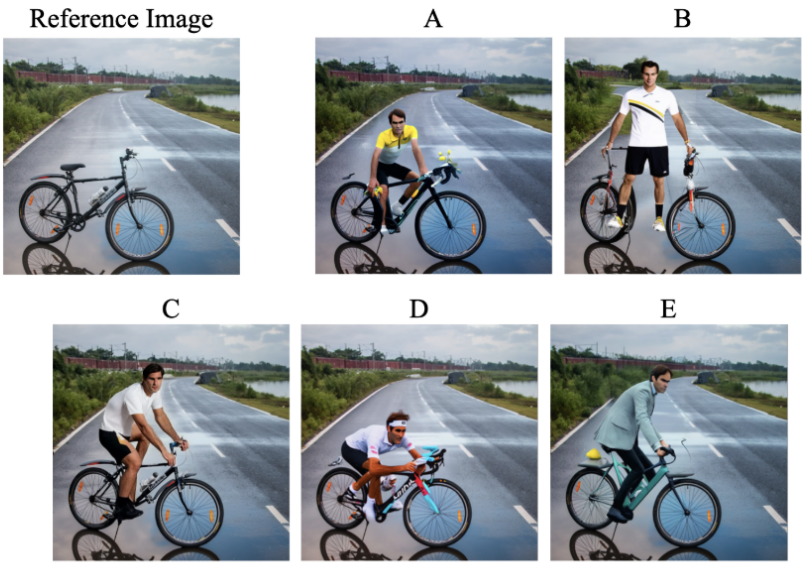
Please rank the generated images (A–E) based on **how well they match** the **given prompt**, with *
Rank 1 indicating the most similar and Rank 5 the least similar



	1	2	3	4	5
A	○	○	○	○	○
B	○	○	○	○	○
C	○	○	○	○	○
D	○	○	○	○	○
E	○	○	○	○	○

Figure 17. **Example questionnaire for evaluating prompt alignment.**

Please rank the generated images (A–E) based on the **quality of the human insertion** compared to the **reference image**. A good insertion should preserve the original scene without unnecessary modifications and maintain high visual quality. Rank 1 indicates the best overall insertion, and Rank 5 the worst. *



	1	2	3	4	5
A	○	○	○	○	○
B	○	○	○	○	○
C	○	○	○	○	○
D	○	○	○	○	○
E	○	○	○	○	○

Figure 18. **Example questionnaire for evaluating background fidelity.**

```
def mask_guided_attn():
    output_res = int(hidden_states.shape[1] ** 0.5)

    anchors_hidden_states = anchors_cache.input_h_cache[self.place_in_unet][self.attnstore.curr_iter]

    ref_mask = self.downsample_mask(          [anchors_cache.masks["ref_subject_mask"]],
        output_res=output_res, visualize=True,
        image=anchors_cache.masks["ref_image"], name="ref")

    anchors_hidden_states = anchors_hidden_states[:, ref_mask==1, :]
    anchors_keys = attn.to_k(anchors_hidden_states, *args)
    anchors_values = attn.to_v(anchors_hidden_states, *args)

    # original attn
    orig_query = attn.head_to_batch_dim(query).contiguous()
    orig_value = attn.head_to_batch_dim(value).contiguous()
    orig_key = attn.head_to_batch_dim(key).contiguous()

    hidden_states = xformers.ops.memory_efficient_attention(
        orig_query, orig_key, orig_value, op=self.attention_op, scale=attn.scale
    )
    subject_key = torch.cat([key.chunk(2, dim=0)[1], anchors_keys[1].unsqueeze(0)], dim=1)
    subject_value = torch.cat([value.chunk(2, dim=0)[1], anchors_values[1].unsqueeze(0)], dim=1)
    subject_key = attn.head_to_batch_dim(subject_key).contiguous()
    subject_value = attn.head_to_batch_dim(subject_value).contiguous()

    sim = torch.einsum("h i d, h j d -> h i j", query, subject_key) * attn.scale
    sim_gen, sim_refs = sim[..., :output_res**2], sim[..., output_res**2:]
    attn_map = sim.softmax(-1).to(subject_value.dtype)
    subject_output = torch.einsum("h i j, h j d -> h i d", attn_map, subject_value)

    uncond_hidden, cond_hidden = hidden_states.chunk(2)

    cond_hidden = subject_output

    hidden_states = torch.cat([uncond_hidden, cond_hidden], dim=0)

    if self.enable_cpu_offloading:
        anchors_hidden_states.to("cpu")
    return hidden_states

def classifier_free_guidance():
    if self.do_classifier_free_guidance:
        noise_pred_uncond, noise_pred_text = noise_pred.chunk(2)
        noise_pred = noise_pred_uncond + self.guidance_scale * (noise_pred_text - noise_pred_uncond)


def latent_blending():
    if blend_latents and ((i >= blend_t_range[0] and i <= blend_t_range[1])):
    print(f"blending latents at timestep: {i}")
    source_latents = all_latents[-(i+1)]
        if blend_mask is None:
            res_64_attnmap = self.attention_store.last_mask[64].reshape((1, 1, 64, 64)).float()
            resized_attnmaps = F.interpolate(res_64_attnmap, size=source_latents.shape[2], mode='nearest')
        else:
            mask = blend_mask
            mask = mask.reshape((1, 1, mask.shape[0], mask.shape[1])).float().to(device)
            resized_attnmaps = F.interpolate(mask, size=source_latents.shape[2], mode='nearest')

        mask_img = np.array(resized_attnmaps[0][0].cpu().detach())
        # blend the masks and latents
        resized_attnmaps = resized_attnmaps.repeat(1, 4, 1, 1)
        blended_latents = resized_attnmaps * latents + (1 - resized_attnmaps) * source_latents

        latents = blended_latents.half()
```

Figure 19. **Here are the code snippets of the core components in Teleportraits.**