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

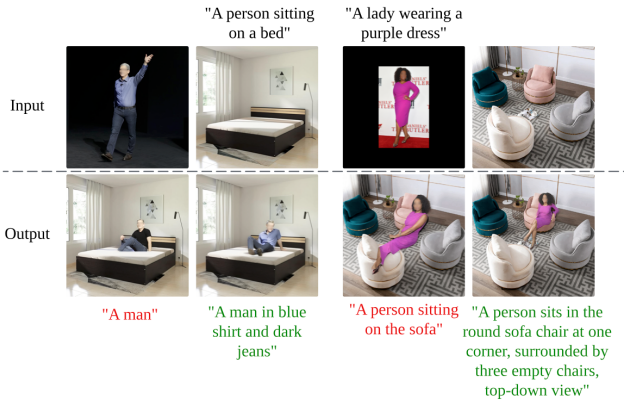


Figure 9. **Failure case 2.** The influence of text prompts with complex examples.

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.

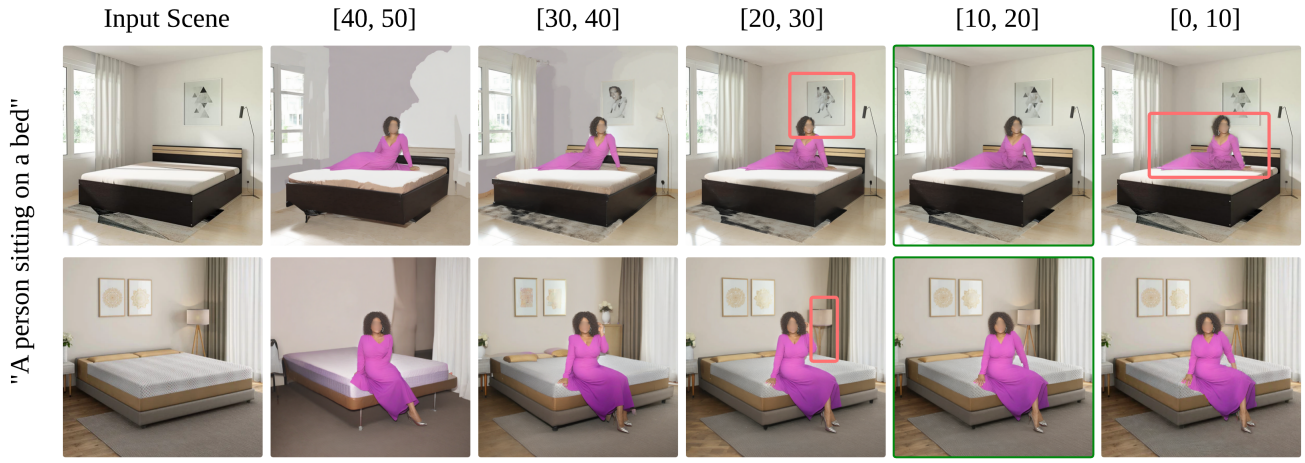


Figure 11. **Influence of latent blending timesteps**. We report results obtained by applying latent blending during $t \in [0, 10], [10, 20], [20, 30], [30, 40], [40, 50]$. The Denoising process starts from $t = 50$ and ends in $t = 0$, meaning that larger t indicates earlier diffusion steps, and smaller 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.

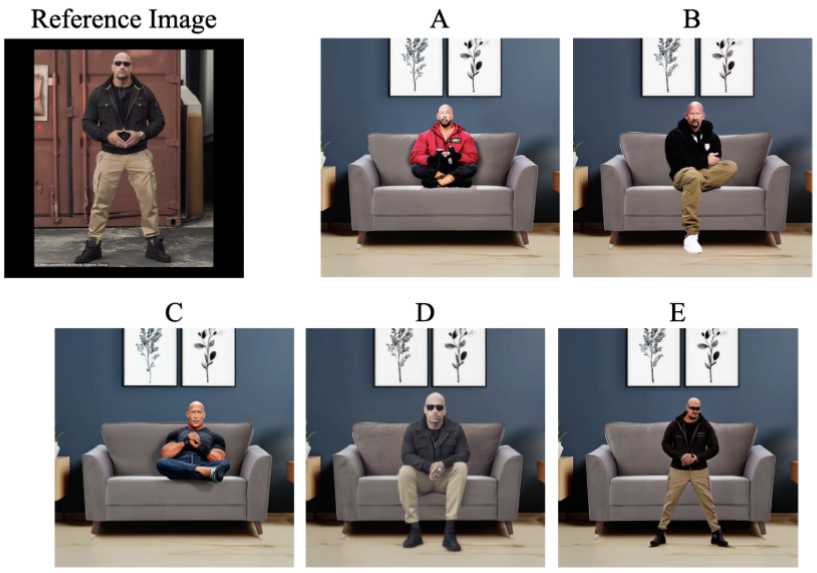
Output Format

[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.**

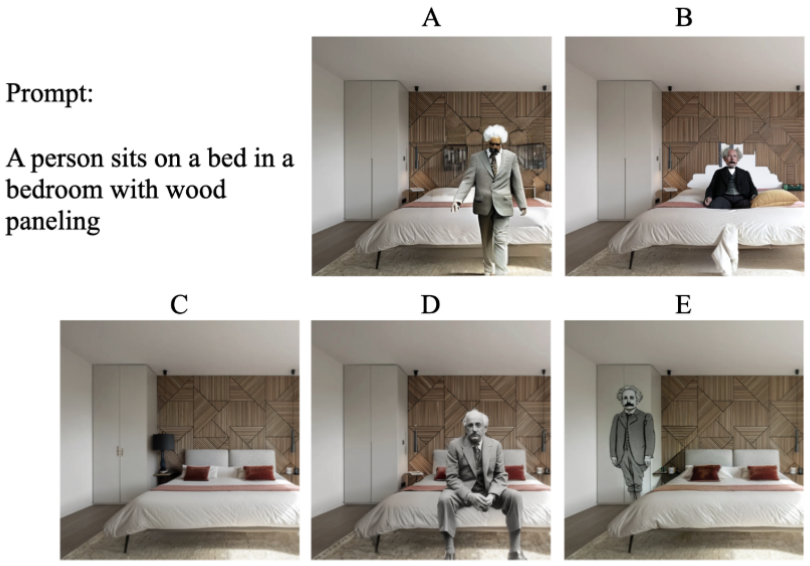
Please rank the generated images (A–E) based on **how closely they resemble** the **reference person**. Focus on both the person's identity (facial features) and clothing appearance. Rank 1 indicates the most similar and Rank 5 the least similar. *



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

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

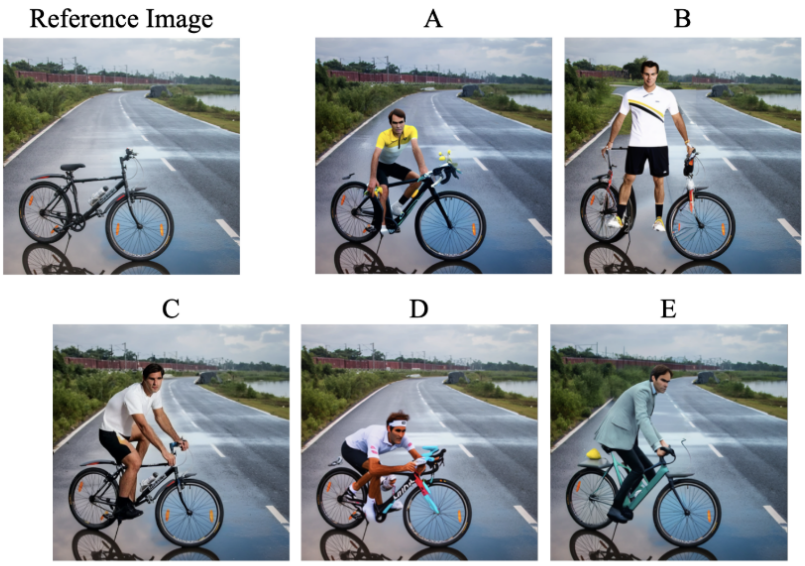
Please rank the generated images (A–E) based on **how well they match** the **given prompt**, with *
Rank 1 indicating the most similar and Rank 5 the least similar



	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. **Code snippets of the core components in Teleportraits.**