Putting People in Their Place: Affordance-Aware Human Insertion into Scenes
Supplementary Materials

Sumith Kulal[1] Tim Brooks[2] Alex Aiken[1] Jiajun Wu[1] Jimei Yang[3] Jingwan Lu[3]
Alexei A. Efros[2] Krishna Kumar Singh[3]
[1]Stanford University [2]UC Berkeley [3]Adobe Research

We discuss implementation details in Sec. 1. We present results on general segmentation masks instead of bounding boxes in Sec. 2, partial body completion in Sec. 3, and cloth swapping in Sec. 4. We highlight the diversity of poses predicted by our model in Sec. 5. We conclude with a discussion of failure cases in Sec. 6 and societal impact in Sec. 7.

1. Implementation details

We use the Stable Diffusion [2] architecture as our backbone and leverage their pre-trained weights as our network initialization. During the forward process, we use a linear noise schedule for 1000 noising steps in the interval $[0.00085, 0.0120]$. During the reverse process, at inference time, we use DDIM sampler [3] for 200 steps.

As in the Stable Diffusion architecure, our model uses the first stage VAE to encode $256 \times 256 \times 3$ image into $32 \times 32 \times 4$ latents. The denoising UNet has a convolution encoder that transforms the 9 channel input with noisy image, masked image, and mask to a 320 channel embedding. The multiplication factors of our UNet are $[1, 2, 4, 4]$. In the input blocks, the height and width get scaled down by the factor and the channel dimension gets scaled up by the same factor. The output blocks do the opposite. Self-attention and cross-attention with conditioning are present at layers 8×8, 16×16, and 32×32 with 8 attention heads.

1.1. Masking details

The configuration of the masking strategy used is:

- **Bounding box**: 30% of the time, we use randomly dilated (0 upto 20 pixels) person bounding box.
- **Larger boxes**: 20% of the time, we randomly sample a larger bounding box (5-20% larger in area) that contains the person bounding box.
- **Random boxes**: 15% of the time, we randomly sample a smaller bounding box (25-75% area) within the person bounding box.
- **Person segmentation**: 15% of the time, we use randomly dilated (0 upto 20 pixels) person segmentation masks.
- **Random scribbles**: 20% of the time, we use randomly generated scribble and brush masks as done in prior inpainting works [5, 6].

1.2. Augmentation details

We apply augmentation on the reference person alone. Given a reference person, we first apply color augmentations. We then mask and center the person. We then apply geometric augmentations. Our augmentation pipeline closely follows StyleGAN-ADA [1].

For color augmentations, we perform brightness, contrast, saturation, image-space filtering and additive noise with probabilities of 0.2, 0.2, 0.2, 0.1 and 0.1 respectively. For geometric augmentations, we perform isotropic scaling, rotation, anistropic scaling and cutout with probabilities 0.4, 0.4, 0.2 and 0.2.

2. Segmentation mask results

We present results on person segmentation masks as holes in Fig. 1 to demonstrate support for arbitrary shaped masks in addition to rectangular bounding boxes.

3. Partial body completion

We present results on partial human body completion in Fig. 2. Our model can recognize and synthesize partial human bodies in addition to full bodies.

4. Cloth swapping

In addition to partial bodies, our model can also be used for interactive editing such as swapping clothes as demonstrated in Fig. 3.

5. Diversity in predicted poses

Different initial noise maps produce different insertions of the reference person into the scene. We present such diverse insertions predicted by our model for the same input scene and reference person in Fig. 4.

Figure 1. **Qualitative results of conditional generation on segmentation masks.** In the first column, we show the scene image with a segmentation mask. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed by our insertion result. We can see that our inserted person follows the segmentation mask while being coherent with the scene.

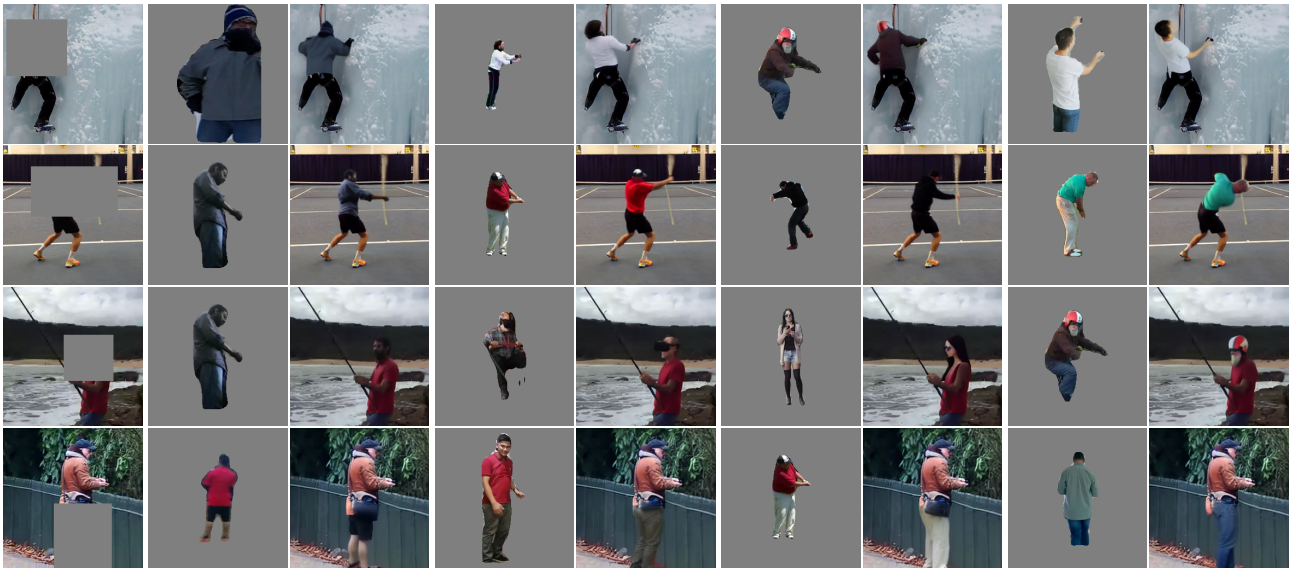

Figure 2. **Qualitative results of partial body completion.** In the first column, we show the scene image with only the partial body masked. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed by our insertion result. We can see that our inserted person is consistent with the visible partial body in terms of the pose while retaining its original appearance.

6. Failure cases

We present common failure modes of our model in Fig. 5.

- **Bad faces and limbs**: Our model often outputs poor face and limb structures (first and second row). This is a result of the first-stage VAE being unable to encode face and limb structures well. This is also a known issue in Stable Diffusion and an active area of development. Training pixel-based diffusion models or improving the auto-encoding quality of the first-stage VAE for humans might alleviate this issue to some degree. These improvements would directly translate to our model.
- **Lighting failure**: The harmonization of lightning of

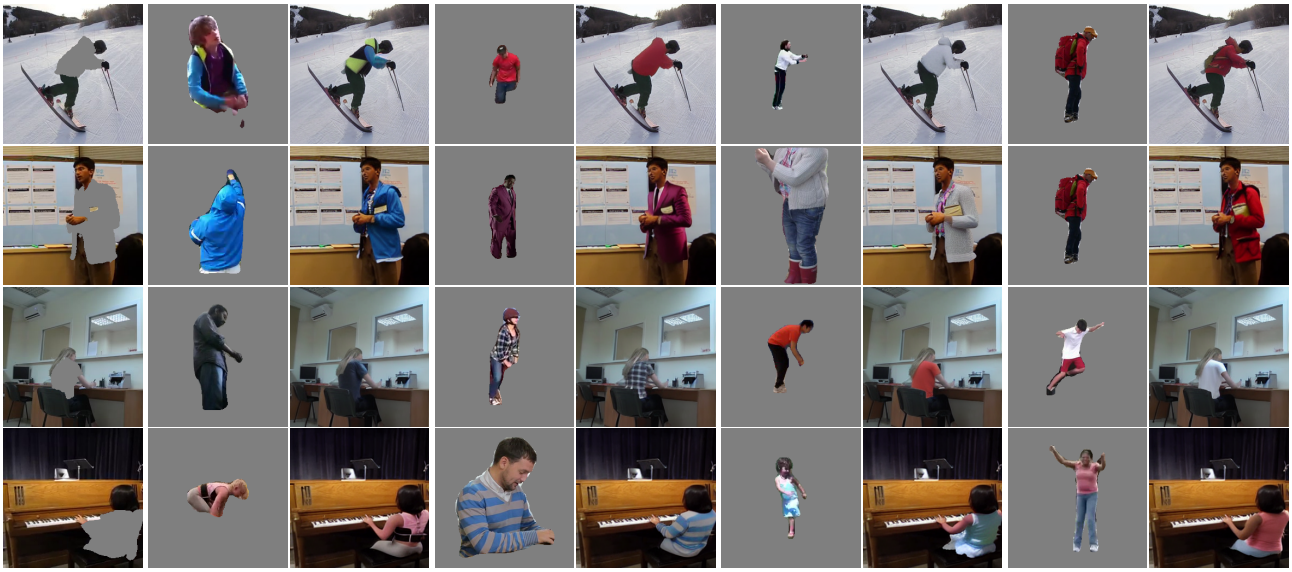

Figure 3. **Qualitative results of cloth swapping.** In the first column, we show the scene image with only the upper body cloth masked. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed by our insertion result. We can see that our generated result is successfully able to borrow the upper body cloth from the person to be inserted. Also, these cloth swaps were quite challenging due to differences in the pose, viewpoint, and scale between the person in the scene image and the person to be inserted.

Figure 4. **Diverse generation for same masked scene image and reference person to be inserted.** In the first column, we show the masked scene image, followed by the reference person to be inserted. After that, we show three different variations of the same person inserted in the scene image. Each variation corresponds to a different noise map during inference. We can see, we are able to compose the same person in the masked region in multiple meaningful ways.

the reference person when inserted in the scene fails at times (first row), if the difference is quite large.

- **Extreme poses**: Extreme input poses are sometimes not reposed (second row) and the model tries to retain the input pose.

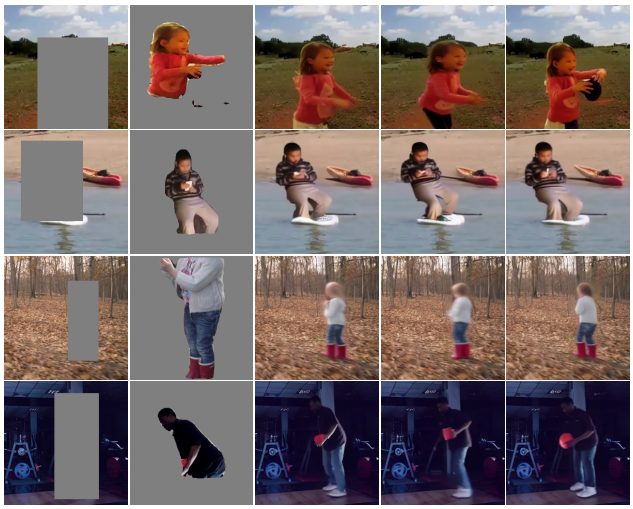

Figure 5. **Failure cases.** Our common failure modes are generating bad faces (row 1 & 2), poor lighning (row 1), extreme pose (row 2), blurry samples (row 3), and generating the object present in the reference person (row 4).

- **Blurry samples**: Our model outputs blurry samples at times (third row). Since our training data is primarily videos, we speculate this is due to the motion blur present in the video dataset.
- **Object failure**: If the reference person is interacting with a visible object in the input, the model also attempts to insert the object into the scene. This leads to artifacts (fourth row).

7. Societal impact

Our model presents a method for affordance-aware human insertion into scenes. We can also hallucinate humans given scene context and vice-versa. The presented research has implications for future work in computer vision, graphics, and robotics. However, our model can be misused to generate malicious content. Similar to Stable Diffusion, the samples generated by our model will be watermarked. Additionally, there's a line of research on detecting fake samples from generative models [4], which we encourage the use of. Since our model is trained on internet videos, it inherits several demographic biases present in the data. We believe the research contributions of this work outweigh the negative impacts. The model will be released after further evaluation and with safety guards.

References

[1] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 1

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable Diffusion. https://github.com/CompVis/stable-diffusion, 2022. 1

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[4] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 4

[5] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 1

[6] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 1