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

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

Abstract

We study the problem of inferring scene affordances by presenting a method for realistically inserting people into scenes. Given a scene image with a marked region and an image of a person, we insert the person into the scene while respecting the scene affordances. Our model can infer the set of realistic poses given the scene context, re-pose the reference person, and harmonize the composition. We set up the task in a self-supervised fashion by learning to re-pose humans in video clips. We train a large-scale diffusion model on a dataset of 2.4M video clips that produces diverse plausible poses while respecting the scene context. Given the learned human-scene composition, our model can also hallucinate realistic people and scenes when prompted without conditioning and also enables interactive editing. A quantitative evaluation shows that our method synthesizes more realistic human appearance and more natural human-scene interactions than prior work.

1. Introduction

A hundred years ago, Jakob von Uexküll pointed out the crucial, even defining, role that the perceived environment (*umwelt*) plays in an organism's life [64]. At a high level, he argued that an organism is only aware of the parts of the environment that it can affect or be affected by. In a sense, our perception of the world is defined by what kinds of interactions we can perform. Related ideas of functional visual understanding (what actions does a given scene afford an agent?) were discussed in the 1930s by the Gestalt psychologists [35] and later described by J.J. Gibson [21] as *affordances*. Although this direction inspired many efforts in vision and psychology research, a comprehensive computational model of affordance perception remains elusive. The value of such a computational model is undeniable for future work in vision and robotics research.

The past decade has seen a renewed interest in such computational models for data-driven affordance perception [15, 20, 24, 25, 67]. Early works in this space deployed a mediated approach by inferring or using intermediate semantic or 3D information to aid in affordance perception [24], while more recent methods focus on direct perception of affordances [15, 20, 67], more in line with Gibson's framing [21]. However, these methods are severely constrained by the specific requirements of the datasets, which reduce their generalizability.

To facilitate a more general setting, we draw inspiration from the recent advances in large-scale generative models, such as text-to-image systems [49, 50, 54]. The samples from these models demonstrate impressive object-scene compositionality. However, these compositions are implicit, and the affordances are limited to what is typically captured in still images and described by captions. We make the task of affordance prediction explicit by putting people "into the picture" [24] and training on videos of human activities.

We pose our problem as a conditional inpainting task (Fig. 1). Given a masked scene image (first row) and a reference person (first column), we learn to inpaint the person into the masked region with correct affordances. At training time, we borrow two random frames from a video clip, mask one frame, and try to inpaint using the person from the second frame as the condition. This forces the model to learn both the possible scene affordances given the context and the necessary re-posing and harmonization needed for a coherent image. At inference time, the model can be prompted with different combinations of scene and person images. We train a large-scale model on a dataset of 2.4M video clips of humans moving in a wide variety of scenes.

In addition to the conditional task, our model can be prompted in different ways at inference time. As shown in the last row Fig. 1, when prompted without a person, our model can hallucinate a realistic person. Similarly, when prompted without a scene, it can also hallucinate a realistic scene. One can also perform partial human completion tasks such as changing the pose or swapping clothes. We show that training on videos is crucial for predicting affordances and present ablations and baseline comparisons in Sec. 4.

To summarize, our contributions are:

- We present a fully self-supervised task formulation for learning affordances by learning to inpaint humans in

Project page: https://sumith1896.github.io/affordance-insertion.

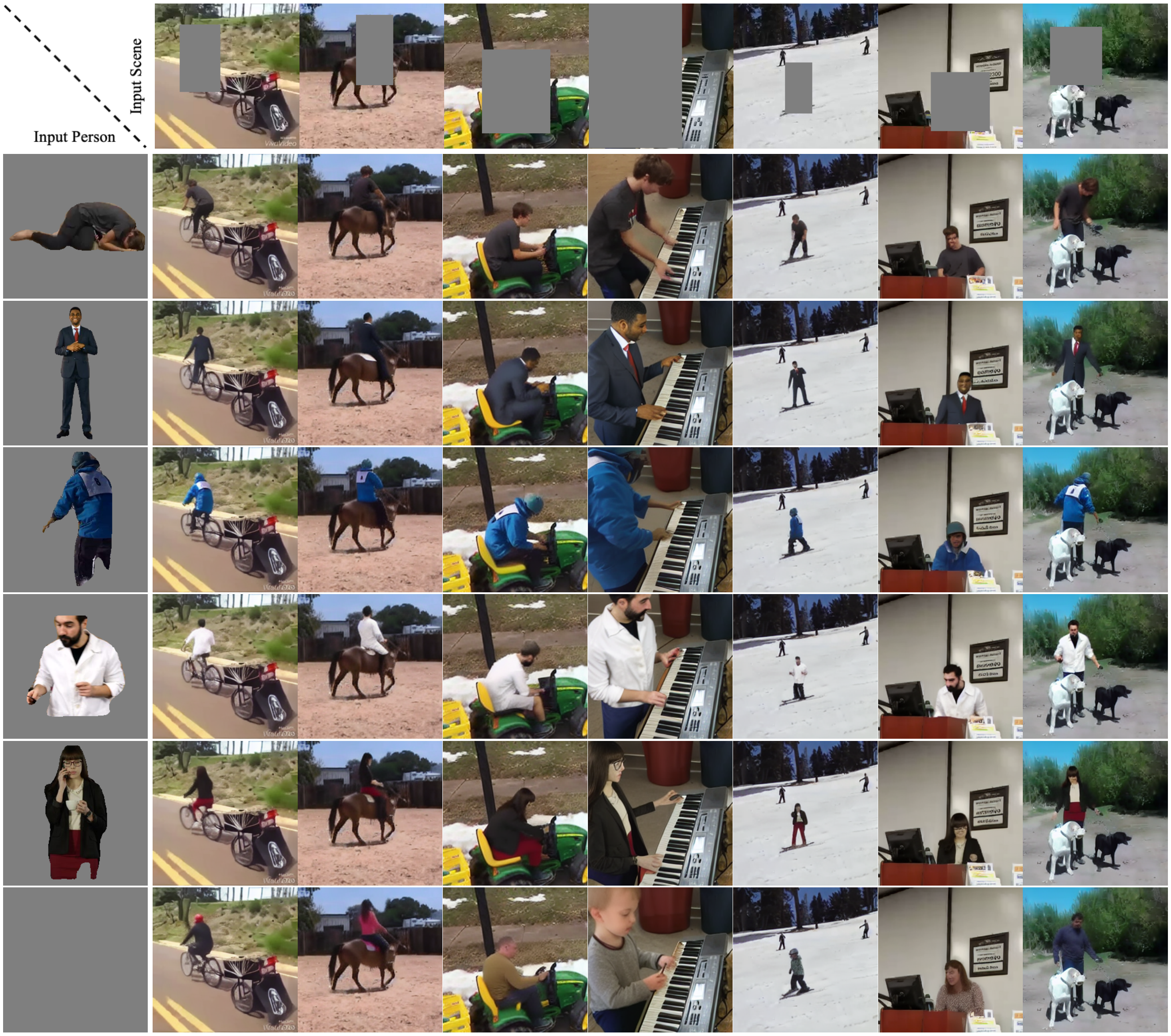


Figure 1. Given a masked scene image (first row) and a reference person (first column), our model can successfully insert the person into the scene image. The model infers the possible pose (affordance) given the scene context, reposes the person appropriately, and harmonizes the insertion. We can also partially complete a person (last column) and hallucinate a person (last row) when no reference is given.

masked scenes.

- We present a large-scale generative model for human insertion trained on 2.4M video clips and demonstrate improved performance both qualitatively and quantitatively compared to the baselines.
- In addition to conditional generation, our model can be prompted in multiple ways to support person hallucination, scene hallucination, and interactive editing.

2. Related Work

Scene and object affordances. Inspired by the work of J.J. Gibson [21], a long line of papers have looked into operationalizing affordance prediction [9,14,15,19,20,23,24, 33,38,67]. Prior works have also looked at modeling human-object affordance [12,22,36,69,76] and synthesizing human pose (and motion) conditioned on an input scene [10,37,65]. Several methods have used videos of humans interacting with scenes to learn about scene affordances [15, 19, 67]. For example, Wang et al. [67] employed a large-scale video dataset to directly predict affordances. They generated a dataset of possible human poses in sitcom scenes. However, their model relies on having plausible ground-truth poses for scenes and hence only performs well on a small number of scenes and poses. On the other hand, we work with a much larger dataset and learn affordances in a fully self-supervised generative manner. We also go beyond synthesizing pose alone and generate realistic humans conditioned on the scene. By virtue of scale, our work generalizes better to diverse scenes and poses and could be scaled further [60].

Inpainting and hole-filling. Early works attempted to use the information within a single image to inpaint masked regions by either diffusing local appearance [5, 8, 46] or matching patches [7, 17]. More recent works use larger datasets to match features [26, 47]. Pathak et al. [47] showed a learning-based approach for image inpainting for large masks, followed up by several recent works that use CNNs [32,40,68,71,72,74,75] and Transformers [6,18,73]. The most relevant works to ours are diffusion-based inpainting models [41, 50, 53]. Rombach et al. [50] used text to guide the diffusion models to perform inpainting tasks. Our task can also be considered as guided inpainting, but our conditioning is an image of a person to be inserted in the scene instead of text. The masking strategy we use is inspired by [72, 74].

Conditional human synthesis. Several works have attempted synthesizing human images (and videos) from conditional information such as keypoints [1,4,13,39,42,56,63], segmentation masks or densepose [2, 43, 66, 70], and driving videos [55, 62]. Prior reposing works do not consider scene context to infer the pose, since the target pose is explicitly given. Moreover, most of the reposing happens in simple backgrounds without semantic content. In contrast, our model conditions on the input scene context and infers the right pose (affordance) prior to reposing. Additionally, our model is trained on unconstrained real-world scenes in an end-to-end manner with no explicit intermediate representation, such as keypoints or 3D.

Diffusion models. Introduced as an expressive and powerful generative model [58], diffusion models have been shown to outperform GANs [16,30,45] in generating more photorealistic and diverse images unconditionally or conditioned by text. With a straightforward architecture, they achieve promising performance in several text-to-image [44, 49, 50, 54], video [29, 57], and 3D synthesis [48] tasks. We leverage ideas presented by Rombach et al. [50] which first encodes images into a latent space and then performs diffusion training in the latent space. We also use classifier-free guidance, introduced by Ho et al. [31], which improves sample quality by trading off against diversity.

3. Methods

In this section, we present details of our learning framework. Given an input scene image, a masked region, and a reference person to be inserted, our model inpaints the masked region with a photo-realistic human that follows the appearance of the reference person, but is re-posed to fit the context in the input scene. We use the latent diffusion model as our base architecture, described in Sec. 3.1. We present details on our problem formulation in Sec. 3.2, our training data in Sec. 3.3, and our model in Sec. 3.4.

3.1. Background - Diffusion Models

Diffusion models [30, 58] are generative models that model data distribution $p(x)$ as a sequence of denoising autoencoders. For a fixed time step T, the forward process of diffusion models gradually adds noise in T steps to destroy the data signal. At time T the samples are approximately uniform Gaussian noise. The reverse process then learns to denoise into samples in T steps. These models effectively predict $\epsilon_\theta(x_t, t)$ for $t = 1 \ldots T$, the noise-level at time-step t given the x_t, a noisy version of input x. The corresponding simplified training objective [50] is

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\Big], \qquad (1)$$

where t is uniformly sampled from $\{1, \ldots, T\}$ and c are the conditioning variables: the masked scene image and the reference person.

Latent diffusion models. As shown in Rombach et al. [50], we use an autoencoder to do perceptual compression and let the diffusion model focus on the semantic content, which makes the training more computationally efficient. Given an autoencoder with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, the forward process uses $\mathcal{E}$ to encode the image, and samples from the model are decoded using $\mathcal{D}$ back to the pixel space.

Classifier-free guidance. Ho et al. [31] proposed classifier-free guidance (CFG) for trading off sample quality with diversity. The idea is to amplify the difference between conditional and unconditional prediction during sampling for the same noisy image. The updated noise prediction is

$$\hat{\epsilon} = w \cdot \epsilon_\theta(x_t, t, c) - (w - 1) \cdot \epsilon_\theta(x_t, t), \qquad (2)$$

3.2. Formulation

The inputs to our model contain a masked scene image and a reference person, and the output image contains the reference person re-posed in the scene's context.

Inspired by Humans in Context (HiC) [9], we generate a large dataset of videos with humans moving in scenes and use frames of videos as training data in our fully self-supervised training setup. We pose the problem as a conditional generation problem (shown in Fig. 2). At training time, we source two random frames containing the same human from a video. We mask out the person in the first frame and use it as the input scene. We then crop out and center the human from the second frame and use it as the reference person conditioning. We train a conditional latent diffusion model conditioned on both the masked scene image and the reference person image. This encourages the model to infer the right pose given the scene context, hallucinate

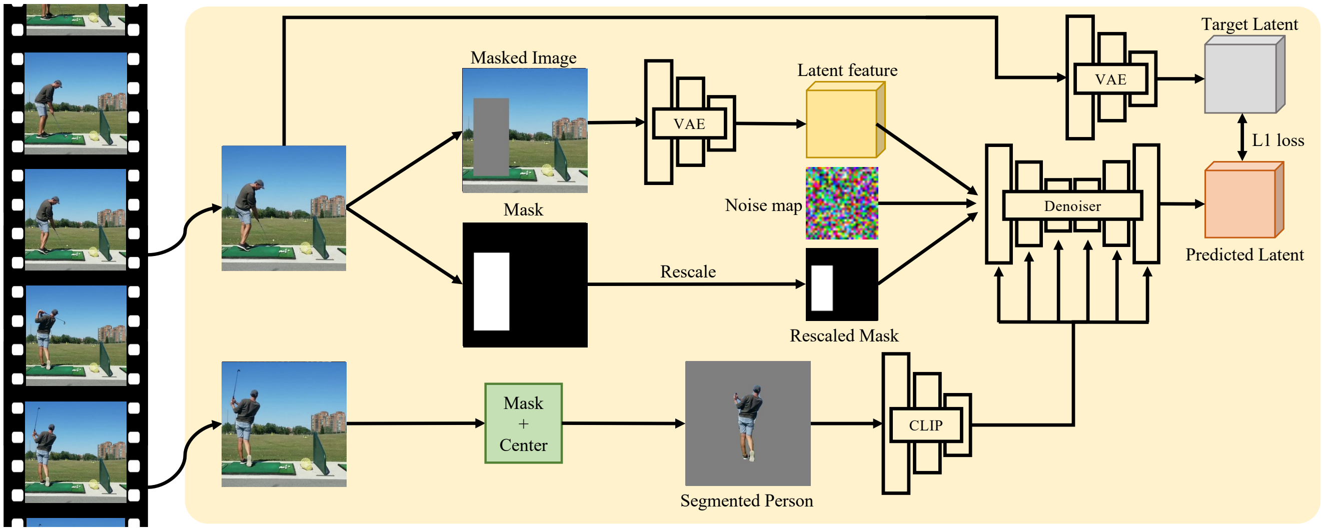


Figure 2. **Architecture overview.** We source two random frames from a video clip. We mask out the person in the first frame and use the person from the second frame as conditioning to inpaint the image. We concatenate the latent features of the background image and rescaled mask along with the noisy image to the denoising UNet. Reference person embeddings (CLIP ViT-L/14) are passed via cross-attention.

the person-scene interactions, and harmonize the re-posed person into the scene seamlessly in a self-supervised manner.

At test time, the model can support multiple applications, inserting different reference humans, hallucinating humans without references, and hallucinating scenes given the human. We achieve this by randomly dropping conditioning signals during training. We evaluate the quality of person conditioned generation, person hallucination and scene hallucination in our experimental section.

3.3. Training data

We generate a dataset of 2.4 million video clips of humans moving in scenes. We follow the pre-processing pipeline defined in HiC [9]. We start from around 12M videos, including a combination of publicly available computer vision datasets as in Brooks et al. [9] and proprietary datasets. First, we resize all videos to a shorter-edge resolution of 256 pixels and retain 256 × 256 cropped segments with a single person detected by Keypoint R-CNN [27]. We then filter out videos where OpenPose [11] does not detect a sufficient number of keypoints. This results in 2.4M videos, of which 50,000 videos are held out as the validation set, and the rest are used for training. Samples from the dataset are shown in Fig. 3. Finally, we use Mask R-CNN [27] to detect person masks to mask out humans in the input scene image and to crop out humans to create the reference person.

We briefly describe our masking and augmentation strategy and present more details in the supp. materials.

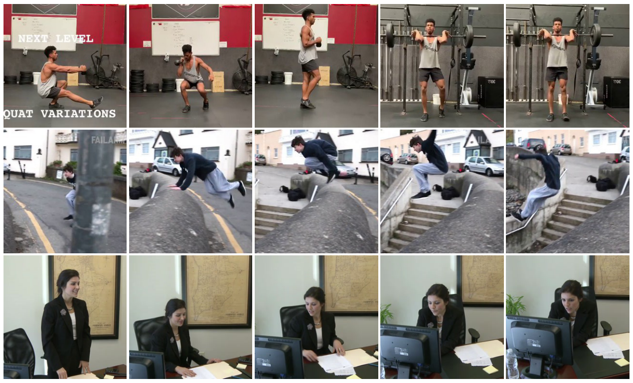


Figure 3. **Sample videos from our dataset.** Each row has five frames uniformly sampled from a video.

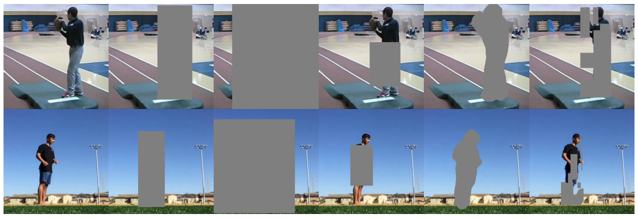


Figure 4. **Various masks used during training.** We use bounding boxes of the person, larger boxes around the person, smaller boxes, segmentation masks, and randomly generated scribbles.

Masking strategy. We apply a combination of different masks for the input scene image, as shown in Fig. 4. These contain bounding boxes, segmentation masks and random scribbles as done in prior inpainting works [72, 74]. This masking strategy allows us to insert people at different levels of granularity, i.e., inserting the whole person, partially completing a person, etc.

Augmentation strategy. We apply data augmentation to reference person alone (as shown in Fig. 5). We borrow

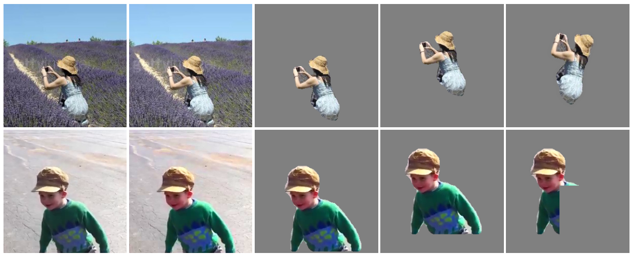


Figure 5. **Data augmentation used for the reference person.** We first apply color augmentations and image corruptions. We then mask and center the person, followed by geometric augmentations.

the augmentation suite used in StyleGAN-ADA [34]. We randomly apply color augmentations. We then mask and center the reference person. After this, we randomly apply geometric augmentations (scaling, rotation, and cutout). Color augmentations are important as, during training, the frames within the same video would usually have similar lighting and brightness. However, this may not be the case during inference, when we want to insert a random person in a random scene.

3.4. Implementation details

We train all models at 256 × 256 resolution. We encode these images using an autoencoder to a latent space of $32 \times 32 \times 4$ ($8\times$ downsample) resolution. The denoising backbone is based on time-conditional UNet [52]. Following prior diffusion inpainting works [50, 53], we concatenate the noisy image with the mask and the masked image. We pass the reference person through an image encoder and use the resulting features to condition the UNet via cross-attention. The mask and the masked image are concatenated as they are spatially aligned with the final output, whereas the reference person is injected through cross-attention as it would not be aligned due to having a different pose. We present ablations of different image encoders in our experiments. We also initialize our model with weights from Stable Diffusion's checkpoint [50].

At training time, to encourage better quality for the human hallucination task, we drop the person-conditioning 10% of the time. We also drop both masked image and person-conditioning 10% of the time to learn the full unconditional distribution and support classifier-free guidance. At test time, we use the DDIM sampler [59] for 200 steps for all our results.

4. Experiments

We present evaluations on a few different tasks. First, we show results on conditional generation with a reference person in Sec. 4.1. We also present ablations of data, architecture, and CFG in this section. We then present results on person hallucinations in Sec. 4.2 and scene hallucinations in Sec. 4.3 and compare with Stable Diffusion [50] and DALL-E 2 [49] as baselines. We present additional results in the supp. material along with a discussion of failure cases.

Table 1. Comparison of metrics for different ablations. First set are on data used for training, second set are on encoders and the final set are on model scaling and effects of pretraining. Metrics used are FID (lower is better) and PCKh (higher is better).

Method	FID ↓	PCKh ↑
Image (w/o aug)	13.174	8.321
Image (w/ aug)	13.008	10.660
Video (w/o aug)	12.103	15.797
VAE KL-8x (concat)	14.956	13.020
Small (400M, scratch)	12.366	15.095
Large (scratch)	11.232	15.873
Large (fine-tune)	10.078	17.602

Metrics. We primarily use two quantitative metrics. First is FID (Fréchet Inception Distance) [28], which measures realism by comparing the distributions of Inception [61] network features of generated images with real images. We measure FID on 50K images, unless specified as FID-10K, wherein we use 10K images. Second is PCKh [3], which measures accurate human positioning by computing the percentage of correct pose keypoints (within a radius relative to the head size). We use OpenPose [11] to detect keypoints of generated and real images.

4.1. Conditional generation

We evaluate the conditional task of generating a target image given a masked scene image and a reference person.

All our models were trained on 32 A100s for 100K steps with a batch size of 1024. We compute the metrics on the held-out set of 50K videos, by trying to inpaint the first masked frame for each video. We choose a reference person from a different video to make the task challenging and use the same mapping for all evaluations.

We present three sets of ablations. **Data.** We experiment with using different training data. We simulate image-only supervision by taking the masked scene image and the reference person from the same frame. We also ablate with data augmentations turned on and off. **Encoders.** We experiment with using the first-stage VAE features, passed in as concatenation instead of CLIP ViT-L/14 embeddings. **UNet.** We experiment with a smaller UNet (430M) compared to ours (860M). We also study the effects of initializing with a pre-trained checkpoint.

Quantitative results are shown in Tab. 1. We observe that image-only models (with or without augmentations) always underperform models trained on video data. This shows that videos provide richer training signal of the same person in different poses which cannot be replicated by simple augmentations. Augmentations, however, do help improve our results. CLIP ViT-L/14 features perform better than the

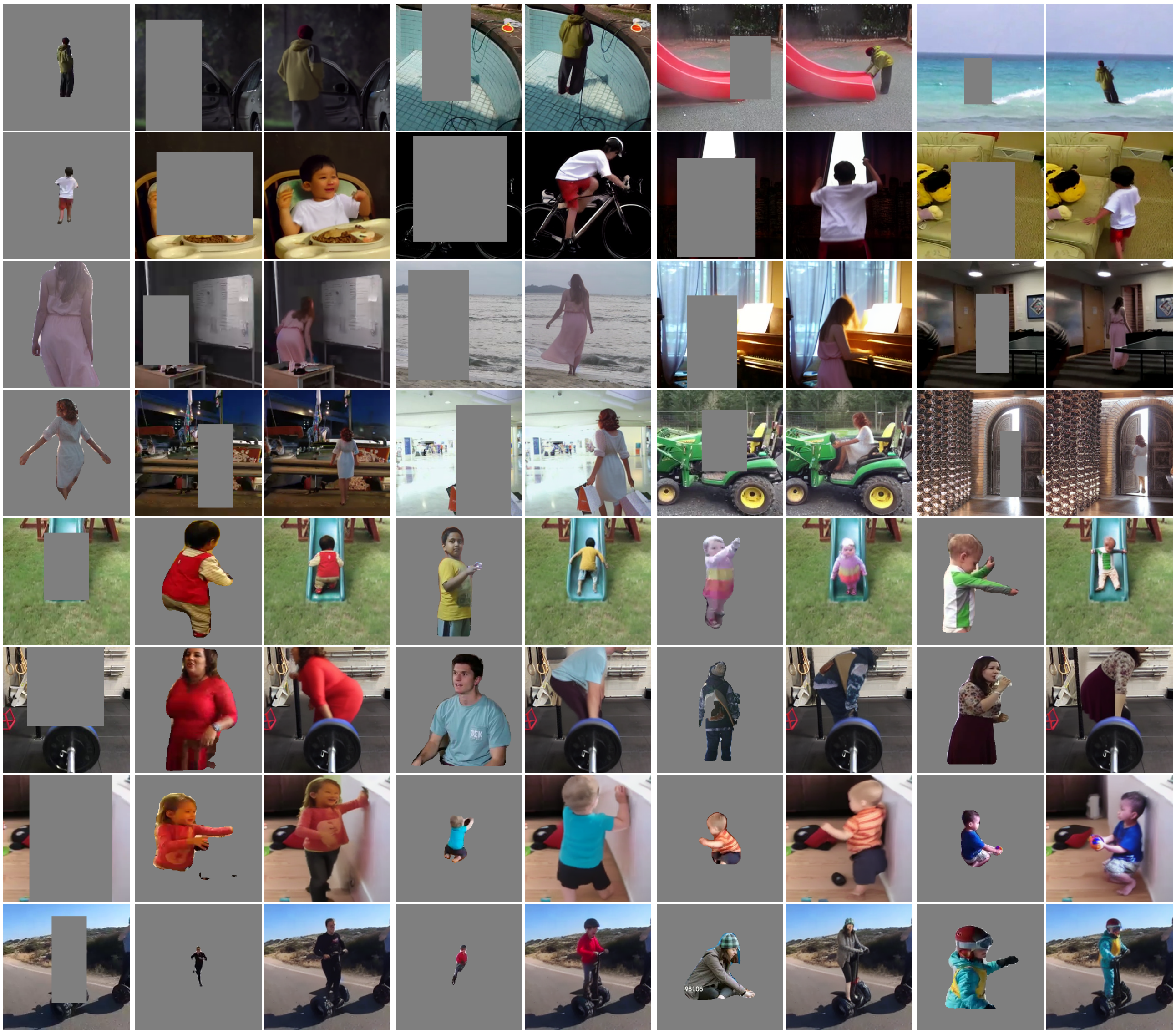
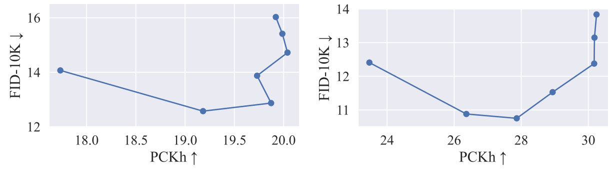


Figure 6. **Qualitative results of conditional generation.** In the top 4 rows, we show a reference person in the first column, followed by four pairs of masked scene image and corresponding result. In the bottom 4 rows, we show a masked scene image in the first column, followed by four pairs of reference person and corresponding result. Our results have the reference person re-posed correctly according to the scene.



(a) Person conditioning (b) Person hallucination

Figure 7. **Classifier-free guidance.** Effect of increasing CFG guidance scale. Evaluated at [1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0].

VAE features passed through concatenation. We also note that using a larger 860M UNet and initializing with Stable Diffusion checkpoints help with our model performance.

We present qualitative results for our best-performing model in Fig. 6. In the top four rows, we show how our model can infer candidate poses given scene context and flexibly re-pose the same reference person into various different scenes. In the bottom four rows, we also show how different people can coherently be inserted into the same scene. The generated images show the complex human-scene composition learned by our model. Our model also harmonizes the insertion by accounting for lighting and shadows.

Effect of CFG. We present the metric trend with varying CFG [31] guidance scales in Fig. 7a. In line with observations from text-to-image models [50, 54], our FID and PCKh both initially improve with CFG. At high values, the image

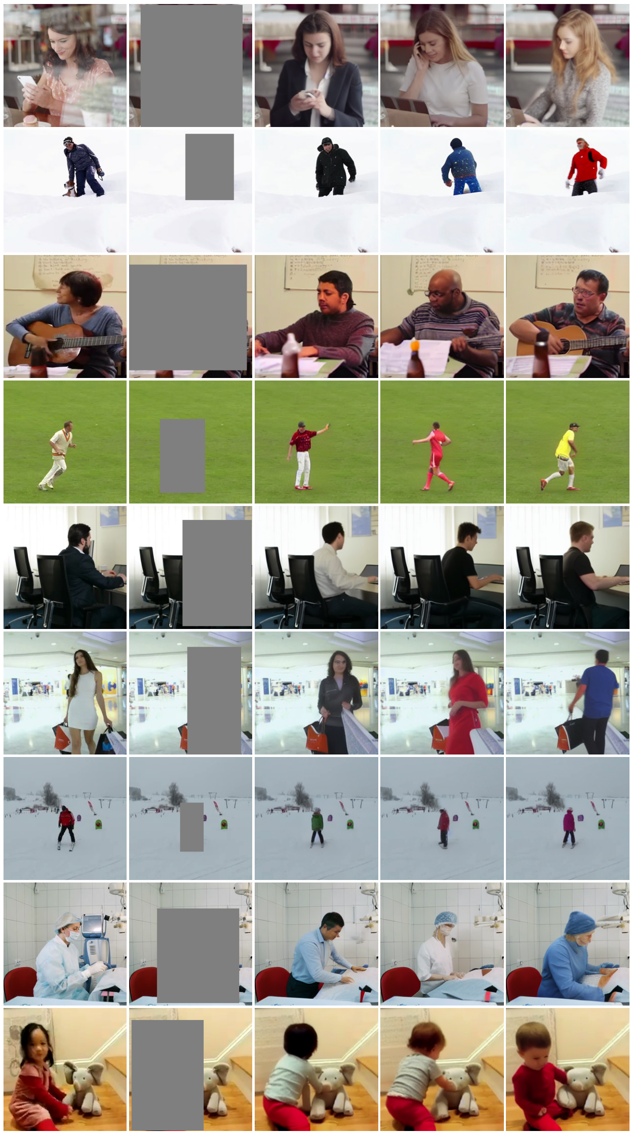


Figure 8. **Qualitative results of person hallucination.** From left to right, groundtruth image, masked scene image, 3 hallucinated persons in the scene. Our method can hallucinate plausible pose and appearance.

quality (FID) suffers. We perform CFG by dropping both the masked scene image and reference person to learn a true unconditional distribution. We observed that dropping only the reference person was detrimental to our model performance.

4.2. Person Hallucination

We evaluate the person hallucination task by dropping the person conditioning and compare with baselines Stable Diffusion [51] and DALL-E 2 [49]. We evaluate our model by passing an empty conditioning person. We evaluate quantitatively with Stable Diffusion (SD) with the following prompt: "natural coherent image of a person in a scene". For qualitative evaluation, we generate SD and DALL-E 2 results with the same prompt.

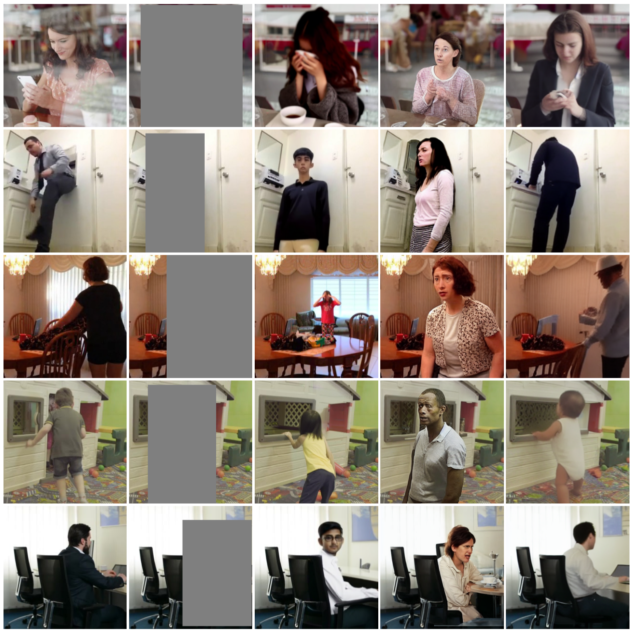


Figure 9. **Baseline comparisons for person hallucination.** From left to right, ground-truth, masked scene image, DALL-E 2 result, Stable Diffusion result and our result. Our model does the best job in hallucinating humans consistent with the context.

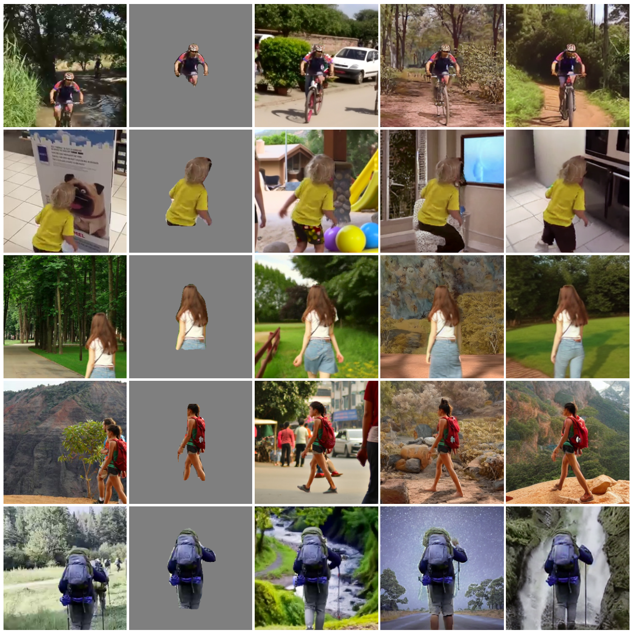


Figure 10. **Baseline comparisons for scene hallucination.** From left to right, ground-truth, reference person, DALL-E 2 result, Stable Diffusion result, and our result. Our model does the best job of hallucinating the scene consistent with the reference person.

We present qualitative results in Fig. 8 where our model can successfully hallucinate diverse people given a masked scene image. The hallucinated people have poses consistent with the input scene affordances. We also present quantitative results in Tab. 2. While Stable Diffusion does produce

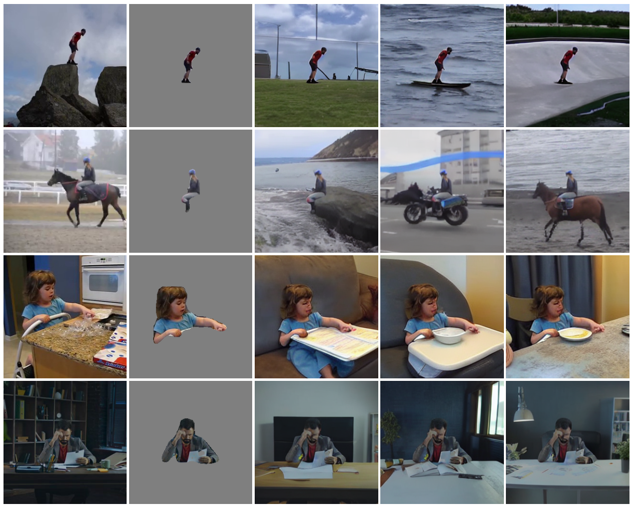


Figure 11. **Constrained scene hallucination.** From left to right, ground-truth, reference person, 3 hallucinated scene samples where the pose and location of the person is constrained. Our hallucinated scene has consistent affordances with the reference person and the reference person stays unchanged.

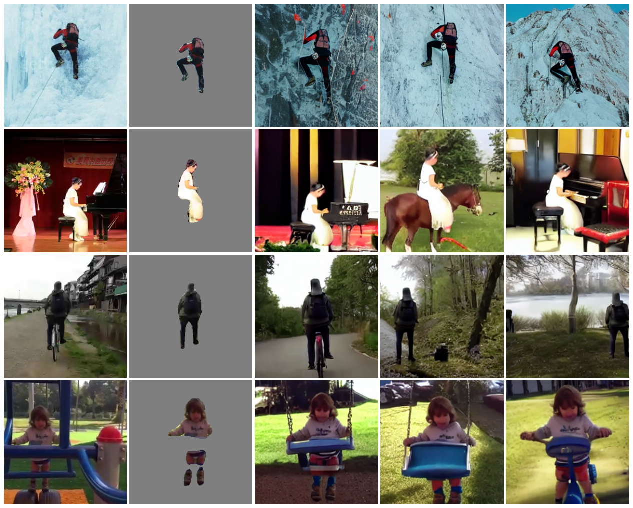


Figure 12. **Qualitative results of unconstrained scene hallucination.** Similar to Fig. 11 but the person here is not constrained on location and pose, hence they can change according to the hallucinated scene. As a result, we are able to generate our reference person in diverse poses while hallucinating different scenes.

high-quality results in some cases, it sometimes fails catastrophically such as hallucinating a person in incorrect poses, or completely ignoring the text conditioning. This is expected as Stable Diffusion's inpainting is trained to inpaint random crops with generic captions rather than inpainting with a consistent human, which is our objective.

We present qualitative baseline comparisons in Fig. 9, we observe that baseline models sometimes ignore the scene context while our model does better at hallucinating humans consistent with the scene.

Table 2. Comparison of metrics with Stable Diffusion for person and scene hallucination. For Stable Diffusion. we use the following prompt: "a natural coherent image of a person in a scene.".

Method	Person hall.		Scene hall.
	FID ↓	PCKh ↑	FID ↓
Stable Diffusion	19.651	0.023	44.687
Ours	8.390	23.213	20.268

Effect of CFG. The metric trend with varying CFG scales for person hallucination follows closely with the person conditioning trend, as shown in Fig. 7b. Both FID and PCKh initially improve, after which FID worsens.

4.3. Scene Hallucination

We evaluate two kinds of scene hallucination tasks. **Constrained:** For the constrained setup, we pass the reference person as the scene image. The model then retains the location and pose of the person and hallucinates a consistent scene around the person. **Unconstrained:** For the unconstrained setup, we pass an empty scene conditioning. Given a reference person, the model then simultaneously hallucinates a scene and places the person in the right location and pose. We evaluate the constrained setup quantitatively with SD with the same prompt as before. We also present qualitative samples from SD and DALL-E 2.

We present qualitative results of the constrained case in Fig. 11 and unconstrained case in Fig. 12. Quantitative comparisons are in Tab. 2. As hallucinating scenes is a harder task with large portions of the image to be synthesized, FID scores are generally higher with our model performing better. Some qualitative baseline comparisons are presented in Fig. 10. Compared to the baselines, our model synthesizes more realistic scenes while maintaining coherence with the input reference person.

5. Conclusion

In this work, we propose a novel task of affordance-aware human insertion into scenes and we solve it by learning a conditional diffusion model in a self-supervised way using video data. We show various qualitative results to demonstrate the effectiveness of our approach. We also perform detailed ablation studies to analyze the impacts of various design choices. We hope this work will inspire other researchers to pursue this new research direction.

Acknowledgement. We are grateful to Fabian Caba Heilbron, Tobias Hinz, Kushal Kafle, Nupur Kumari, Yijun Li and Markus Woodson for insightful discussions regarding data and training pipeline. This work was partly done by Sumith Kulal during an internship at Adobe Research. Additional funding provided by ONR MURI and NSF GRFP.

References

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, pages 219–233. Wiley Online Library, 2019. 3

[2] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 3

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 5

[4] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 3

[5] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 3

[6] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. *arXiv preprint arXiv:2209.00647*, 2022. 3

[7] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 3

[8] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 3

[9] Tim Brooks and Alexei A Efros. Hallucinating pose-compatible scenes. *ECCV*, 2022. 2, 3, 4

[10] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 2

[11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 4, 5

[12] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *arXiv e-prints*, pages arXiv–2012, 2020. 2

[13] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019. 3

[14] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2

[15] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012. 1, 2

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[17] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. 3

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[19] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. In *European Conference on Computer Vision*, pages 732–745. Springer, 2012. 2

[20] David F. Fouhey, Xiaolong Wang, and Abhinav Gupta. In defense of the direct perception of affordances, 2015. 1, 2

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

[22] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 2

[23] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1529–1536, 06 2011. 2

[24] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 1, 2

[25] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1

[26] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007. 3

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, 2017. 5

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[31] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 6

[32] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[33] Yun Jiang, Hema Koppula, and Ashutosh Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[34] 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. 5

[35] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 1935. 1

[36] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 2

[37] Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 491–500, 2002. 2

[38] 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 *CVPR*, 2019. 2

[39] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[40] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 3

[41] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[42] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017. 3

[43] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 123–138, 2018. 3

[44] 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. 3

[45] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[46] Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005. 3

[47] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3, 5, 7

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3, 5, 6

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable Diffusion. `https://github.com/CompVis/stable-diffusion`, 2022. 7

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[53] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3, 5

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3, 6

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[56] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 3

[57] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation

without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[58] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[60] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019. 3

[61] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 5

[62] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3

[63] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pages 3560–3569. PMLR, 2017. 3

[64] Jakob Von Uexküll. Environment [umwelt] and inner world of animals. *Foundations of comparative ethology*, pages 222–245, 1985. 1

[65] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes, 2020. 2

[66] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[67] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 1, 2

[68] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6721–6729, 2017. 3

[69] B. Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2010. 2

[70] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 3

[71] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 3

[72] 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. 3, 4

[73] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 3

[74] 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. 3, 4

[75] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Cm-gan: Image inpainting with cascaded modulation gan and object-aware training. *arXiv preprint arXiv:2203.11947*, 2022. 3

[76] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 2

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.

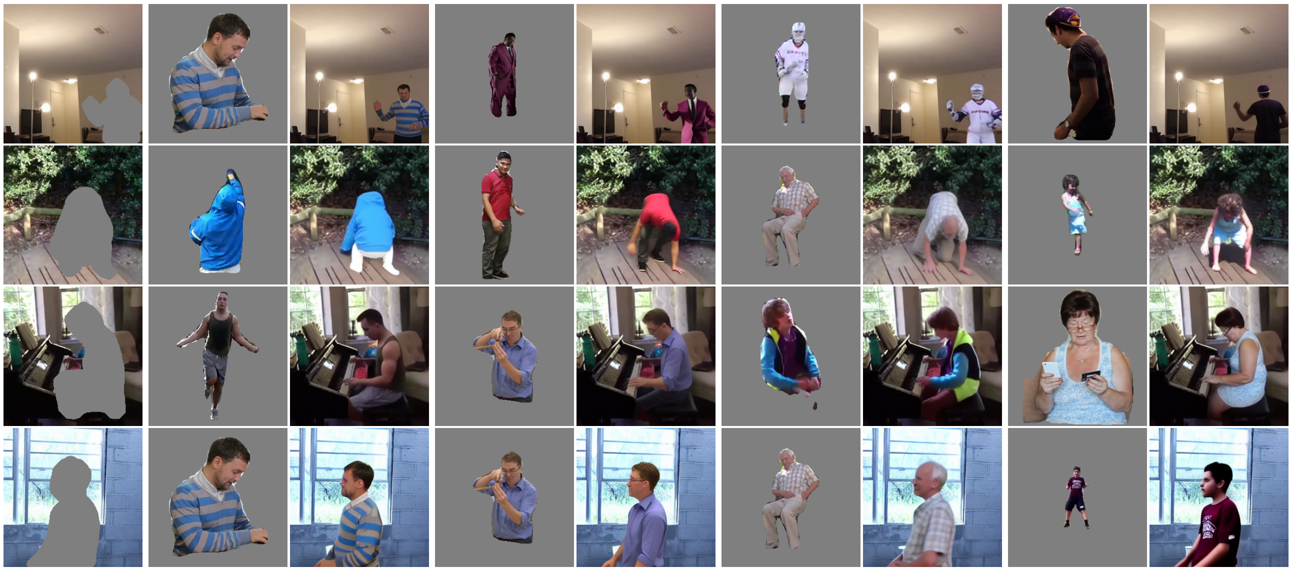


Figure 1. **Qualitative results of conditional generation on segmentation masks.** In the first column, we show the scene image with a segmentation mask. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed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

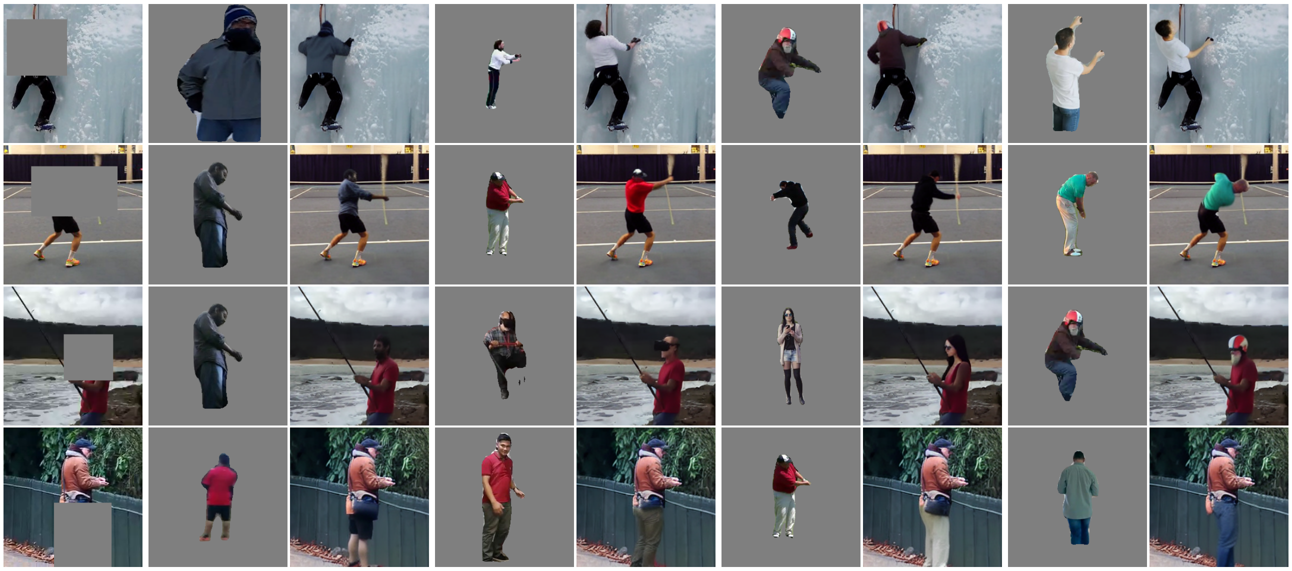


Figure 2. **Qualitative results of partial body completion.** In the first column, we show the scene image with only the partial body masked. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed by our 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

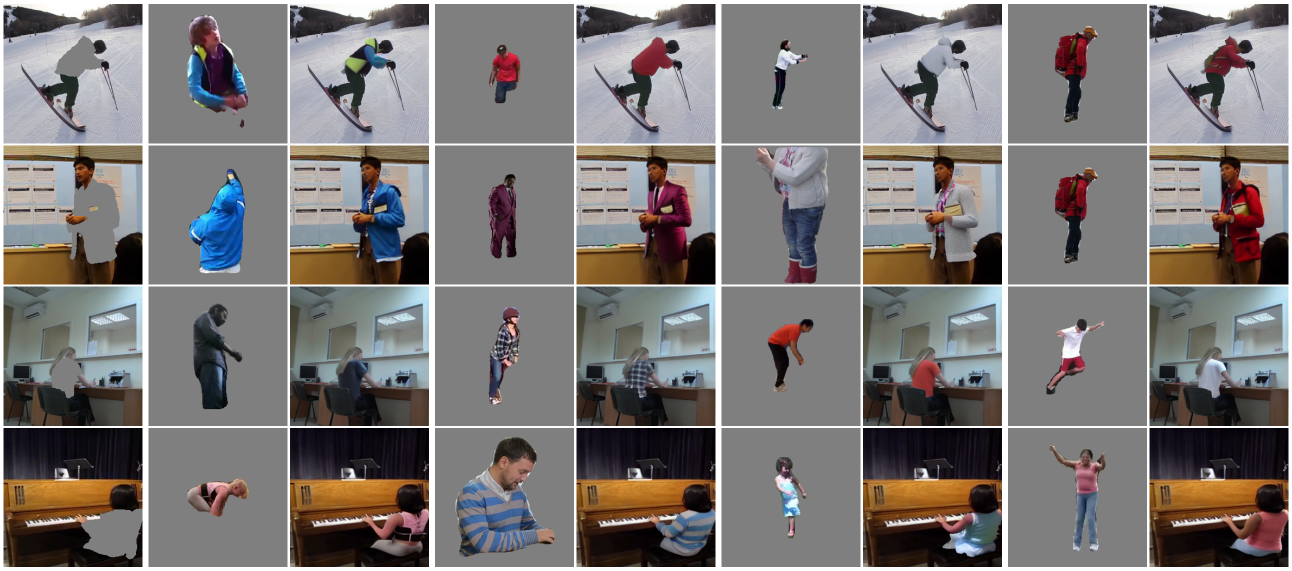


Figure 3. **Qualitative results of cloth swapping.** In the first column, we show the scene image with only the upper body cloth masked. After that, we show the results of inserting four different people in the scene image. For each result, we first show the person to be inserted followed by our insertion 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.

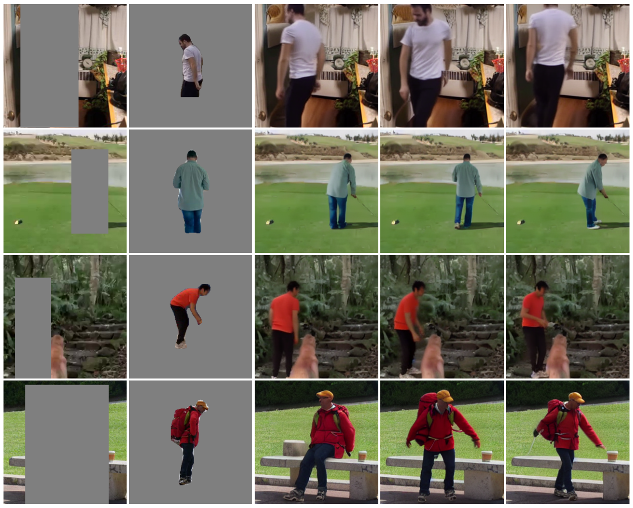


Figure 4. **Diverse generation for same masked scene image and reference person to be inserted.** In the first column, we show the masked scene image, followed by the reference person to be inserted. After that, we show three different variations of the same person inserted in the scene image. Each variation 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.

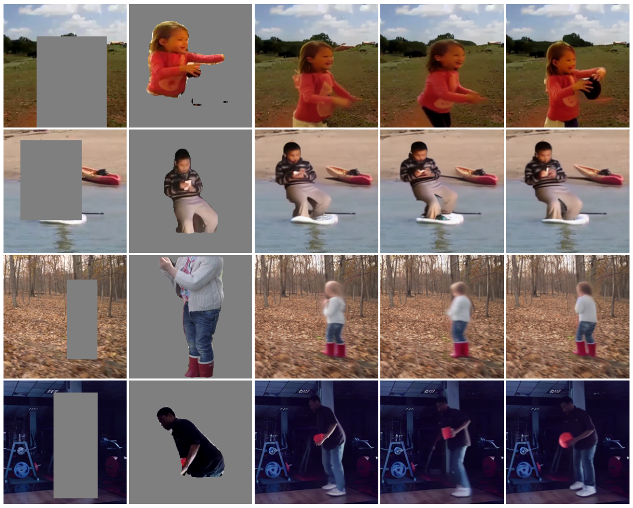


Figure 5. **Failure cases.** Our common failure modes are generating bad faces (row 1 & 2), poor lighning (row 1), extreme pose (row 2), blurry samples (row 3), and generating the object present in the reference person (row 4).

- **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

We present 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 can 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.

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