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. We conduct quantitative evaluation and show that our method synthesizes more realistic human appearance and more natural human-scene interactions when compared to prior work. This work will also appear as a full paper in CVPR 2023.

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 [34]. 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 [20] and later described by J.J. Gibson [10] 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 [6, 9, 13, 14, 36]. Early works in this space deployed a mediated approach by inferring or using intermediate semantic or 3D information to aid in affordance perception [13], while more recent methods focus on direct perception of affordances [6, 9, 36], more in line with Gibson's framing [10]. 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 [27, 28, 30]. 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" [13] 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 enforces 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.

Apart from 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.

To summarize, our contributions are:

- We present a fully self-supervised task formulation for learning affordances by learning to inpaint humans in masked scenes.
- We present a large-scale generative model for human insertion trained on 2.4M video clips and demonstrate improved performance compared to the baselines.
- In addition to conditional generation, our model can be

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.

prompted in multiple ways to support person hallucination and scene hallucination.

2. Related Work

Scene and object affordances. Inspired by the work of J.J. Gibson [10], a long line of papers have looked into operationalizing affordance prediction [1, 5, 6, 8, 9, 12, 13, 19, 23, 36]. Prior works have also looked at modeling human-object affordance [4, 11, 21, 37, 38] and synthesizing human pose (and motion) conditioned on an input scene [2, 22, 35]. Several methods have used videos of humans interacting with scenes to learn about scene affordances [6, 8, 36]. For example, Wang et al. [36] employed a large-scale video dataset to directly predict affordances. 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. By virtue of scale, our work generalizes better to diverse scenes and poses and could be scaled further [33].

Diffusion models. Introduced as an expressive and powerful generative model [32], diffusion models have been shown to outperform GANs [7, 17, 25] 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 [24, 27, 28, 30], video [16, 31], and 3D synthesis [26] tasks. We leverage ideas presented by Rombach et al. [28] 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. [18], a sampling trick that yields higher-quality samples by trading-off diversity.

3. Methods

We use the latent diffusion model as our base architecture. We present details on our problem formulation in Sec. 3.1, our training data in Sec. 3.2, and our model in Sec. 3.2.

3.1. 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) [1], 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 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,

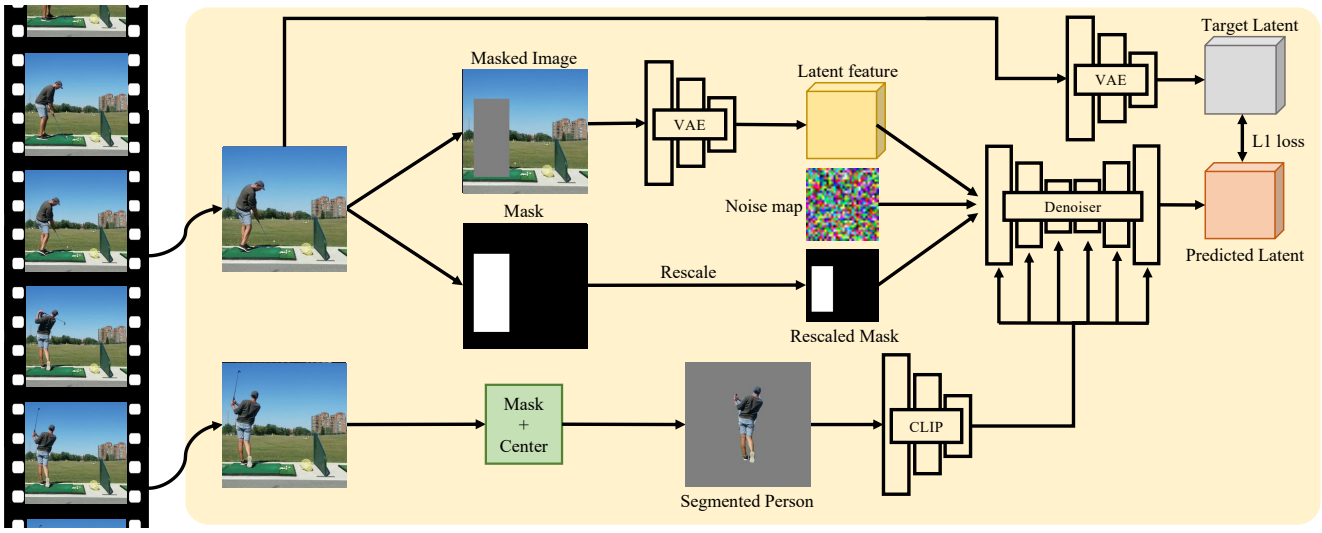

Figure 2. **Architecture overview.** We source two random frames from a video clip. We mask out the person in the first frame and use the person from the second frame as conditioning to inpaint the image. We concatenate the latent features of the background image and rescaled mask along with the noisy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

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.

3.2. 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 [1]. We start from around 12M videos, including a combination of publicly available computer vision datasets as in Brooks et al. [1] 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 [15]. We then filter out videos where OpenPose [3] 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. Finally, we use Mask R-CNN [15] to detect person masks to mask out humans in the input scene image and to crop out humans to create the reference person.

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 then present results on person hallucinations in Sec. 4.2 and scene hallucinations in Sec. 4.3 and compare with Stable Diffusion [28] and DALL-E 2 [27] as baselines.

4.1. Conditional generation

We evaluate the conditional task of generating a target image given a masked scene image and a reference person. We present qualitative results for our best-performing model in Fig. 3. In the top two 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 two 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.

4.2. Person Hallucination

We evaluate the person hallucination task by dropping the person conditioning and compare with baselines Stable Diffusion [29] and DALL-E 2 [27]. We evaluate our model by passing an empty conditioning person.

We present qualitative results in Fig. 4 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 present qualitative baseline comparisons in Fig. 5, we observe that baseline models sometimes ignore the scene context while our model does better at hallucinating humans consistent with the scene.

4.3. Scene Hallucination

For the scene hallucination task, 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. We evaluate the constrained setup SD and DALL-E 2 with the same prompt as before.

We present qualitative results of in Fig. 7. Some qualitative baseline comparisons are presented in Fig. 6. Compared to the baselines, our model synthesizes more realistic scenes while maintaining coherence with the input reference person.

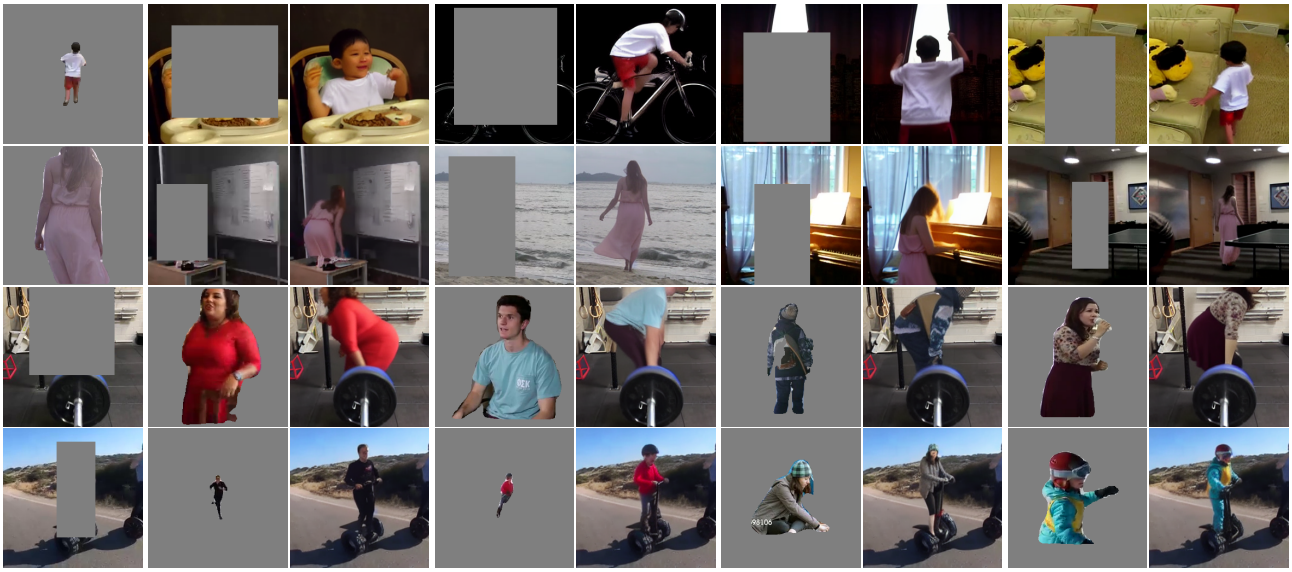

Figure 3. **Qualitative results of conditional generation.** In the top 2 rows, we show a reference person in the first column, followed by four pairs of masked scene image and corresponding result. In the bottom 2 rows, we show a masked scene image in the first column, followed by four pairs of reference person and corresponding result. Our results have the reference person re-posed correctly according to the scene.

Figure 4. **Qualitative results of person hallucination.** From left to right, GT image, masked scene image, 3 hallucinated persons.

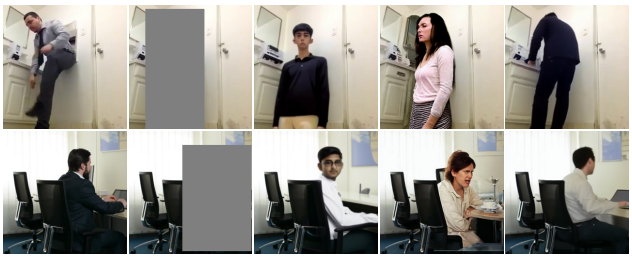

Figure 5. **Baseline comparisons for person hallucination.** From left to right, ground-truth, masked scene image, DALL-E 2 result, Stable Diffusion result and our result. Our model does the best job in hallucinating humans consistent with the context.

Figure 6. **Baseline comparisons for scene hallucination.** From left to right, ground-truth, reference person, DALL-E 2 result, Stable Diffusion result, and our result.

Figure 7. **Scene hallucination.** From left to right, ground-truth, reference person, three hallucinated scene samples.

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 hope this work will inspire other researchers to pursue this new research direction.

References

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

[2] 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

[3] 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. 3

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

[5] 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

[6] 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

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[8] 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

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

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

[11] 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

[12] 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

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

[14] 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

[15] 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. 3

[16] 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. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[19] 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

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

[21] 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

[22] 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

[23] 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

[24] 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. 2

[25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

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

[27] 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, 2, 3

[28] 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, 2, 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable Diffusion. `https://github.com/CompVis/stable-diffusion`, 2022. 3

[30] 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, 2

[31] 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. 2

[32] 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. 2

[33] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019. 2

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

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

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

[37] 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

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