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

Sumith Kulal\textsuperscript{1} \quad Tim Brooks\textsuperscript{2} \quad Alex Aiken\textsuperscript{1} \quad Jiajun Wu\textsuperscript{1} \quad Jimei Yang\textsuperscript{3} \quad Jingwan Lu\textsuperscript{3} \quad Alexei A. Efros\textsuperscript{2} \quad Krishna Kumar Singh\textsuperscript{3} \\
\textsuperscript{1}Stanford University \quad \textsuperscript{2}UC Berkeley \quad \textsuperscript{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-poses 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 (\textit{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 \textit{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 marked 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: \url{https://sumith1896.github.io/affordance-insertion}. 

17089
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.

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

**Inpainting and hole-filling.** Early works attempted to use

**Conditional human synthesis.** Several works have at-

**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.

3. Methods

In this section, we present details of our learning frame-
work. 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 au-
toencoders. 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 \( \epsilon_t(x_t, t) \) 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

\[
L_{DM} = E_{x, \epsilon \sim \mathcal{N}(0,1), t} \left[ || \epsilon_t - \epsilon_t(x_t, t) ||_2^2 \right].
\]  

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

\[
\dot{\epsilon} = w \cdot \epsilon_t(x_t, t, c) - (w - 1) \cdot \epsilon_t(x_t, t),
\]  

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 con-
tingent 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, 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 $256 \times 256$ 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.

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
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 $256 \times 256$ 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.

**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.

### Table 1. Comparison of metrics for different ablations.

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

**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.

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 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
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.

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
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


[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


