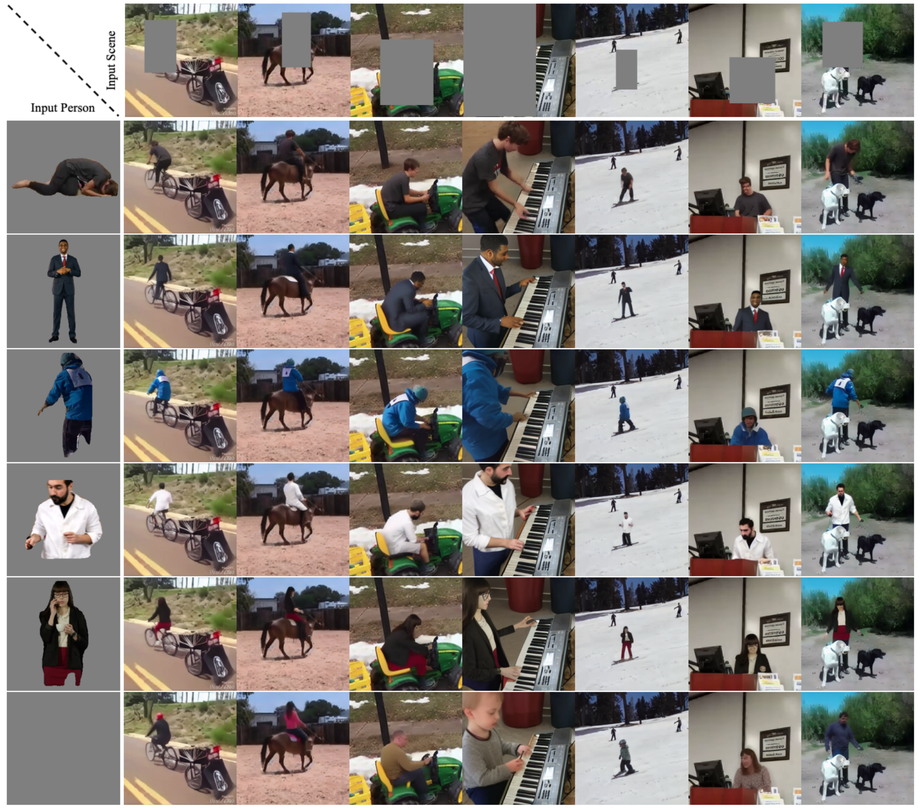
Putting People in Their Place:
Affordance-Aware Human Insertion into Scenes

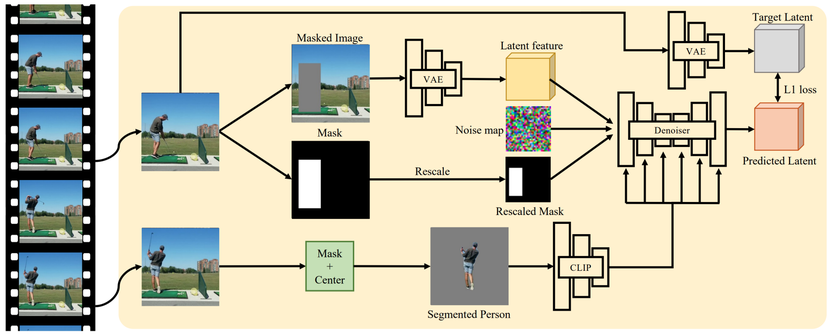
Sumith Kulal[1] Tim Brooks[2] Alex Aiken[1] Jiajun Wu[1]
Jimei Yang[3] Jingwan Lu[3] Alexei A. Efros[2] Krishna K. Singh[3]
[1]Stanford University, [2]UC Berkeley, [3]Adobe Research

Photo-realistic Affordance-Aware Human Insertion into Scenes



Inputs: person image (left) and scene image with marked region (top)
Outputs: realistic insertion of the person into the scene image
Method: large-scale diffusion model trained in a self-supervised fashion on videos
Data: 2.4 million videos of humans moving around in scenes.
Highlights: self-supervised learning, affordances, image synthesis and editing

Learning Architecture Overview



Quantitative Results

	FID	PCKh
Image (w/o aug)	13.17	8.32
Image (w/ aug)	13.01	10.66
Video (w/o aug)	12.10	15.80
Video (w/ aug)	10.08	17.60

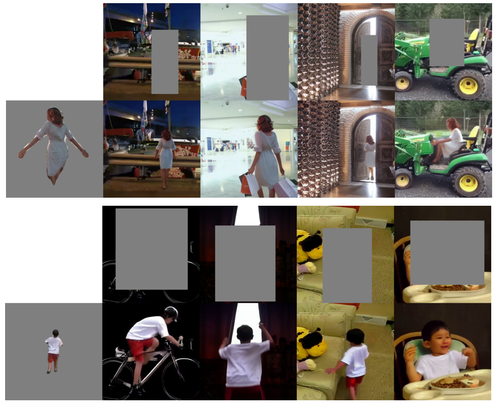
Video data is critical for this task, image only data even with aug performs poorly.

	FID	PCKh
Small (scratch)	12.37	15.10
Large (scratch)	11.23	15.87
Large (SD finetune)	10.08	17.60

Large-scale models are also critical. Initializing with Stable-Diffusion helps.

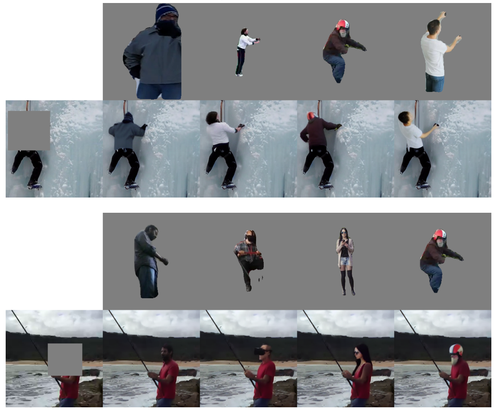
Qualitative Results

Same Person in Different Scenes



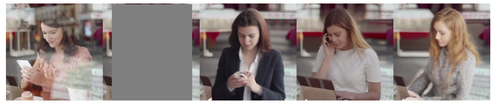
Different People in Same Scene

Partial Body Completion



Scene Hallucination

Person Hallucination



Cloth Swapping