

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

Sumith Kulal



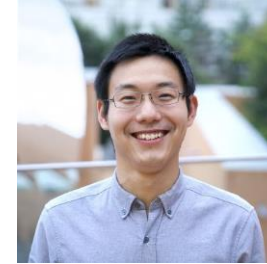
Tim Brooks



Alex Aiken



Jiajun Wu



Jimei Yang



Cynthia Lu



Alyosha Efros



Krishna Kumar Singh



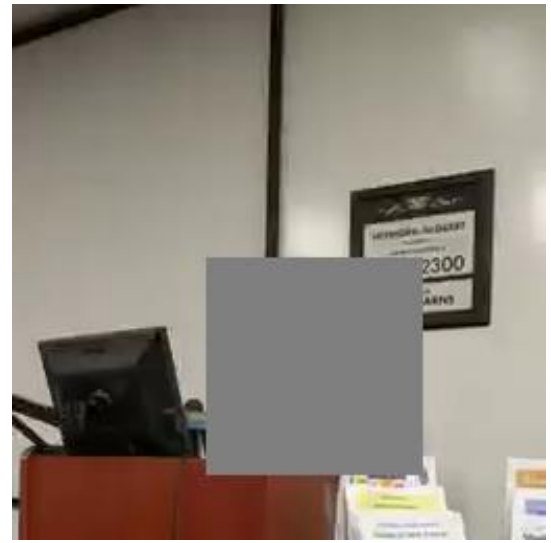
CVPR 2023

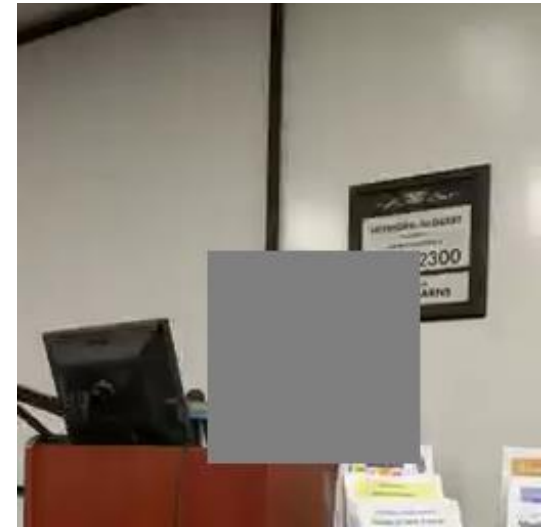


Adobe



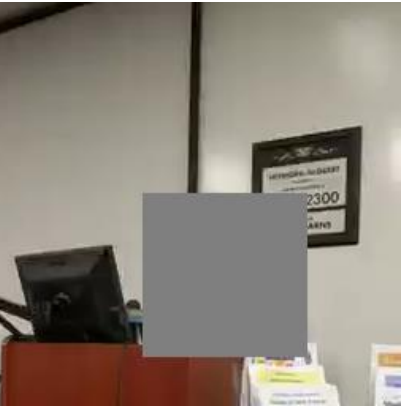
Stanford
University



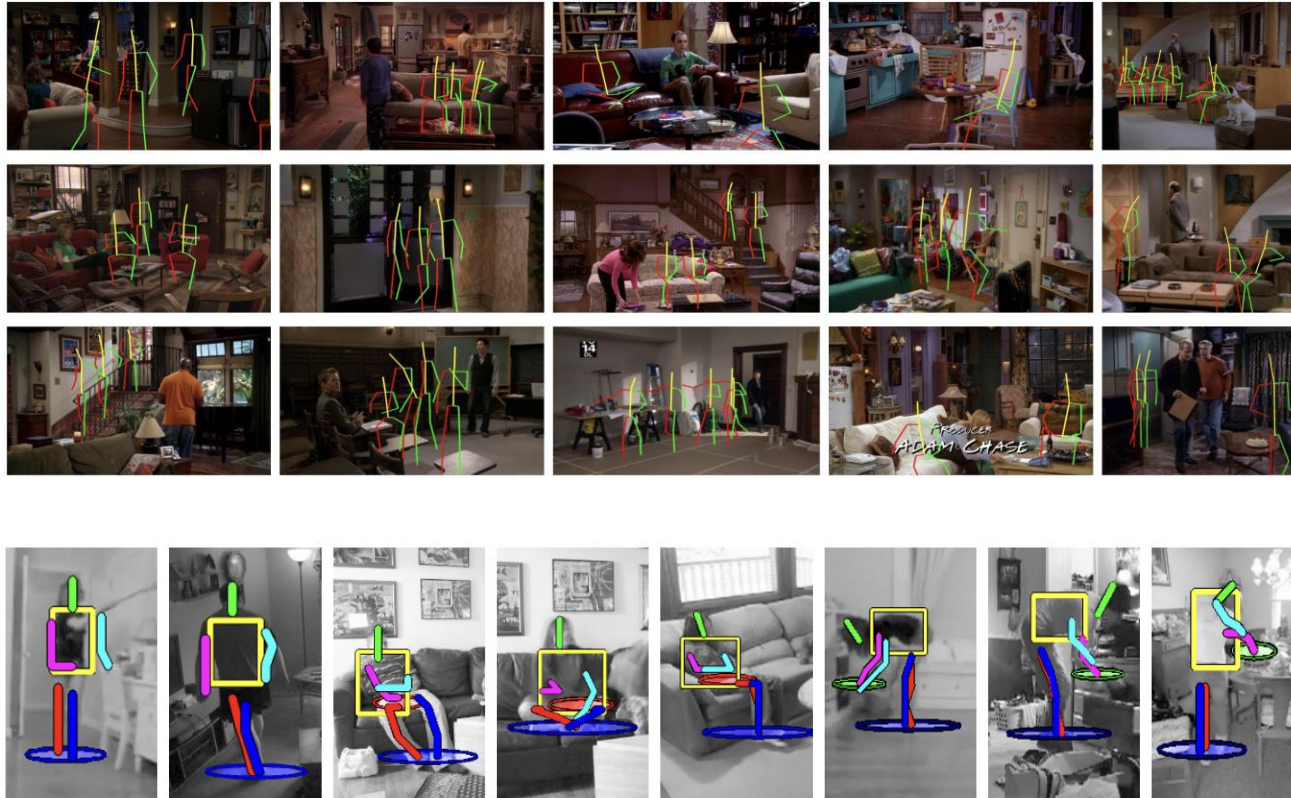


Input Scene

Input Person



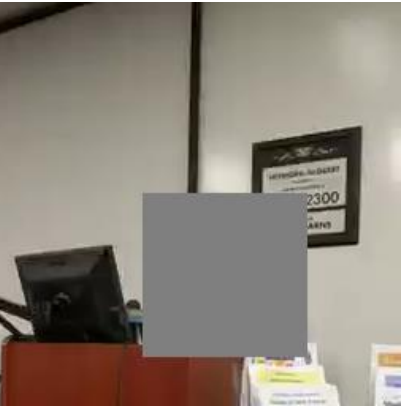
Prior work: Affordance Learning



People Watching: Human Actions as a Cue for Single View Geometry. Fouhey et al.
Scene semantics from long-term observation of people. Delaitre et al.
Binge Watching: Scaling Affordance Learning from Sitcoms. Wang et al.

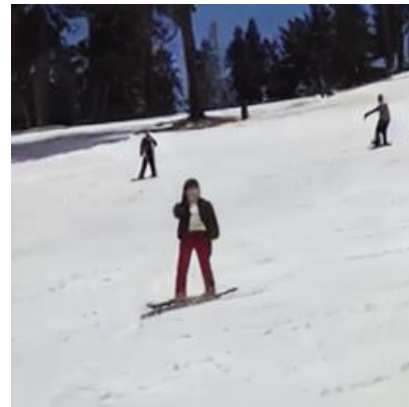
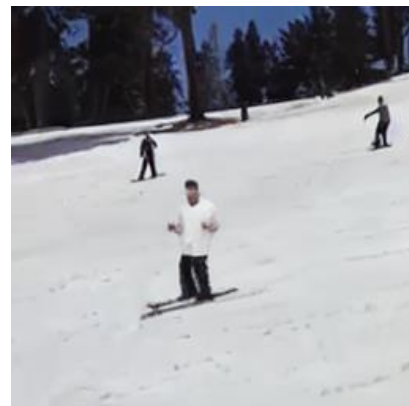
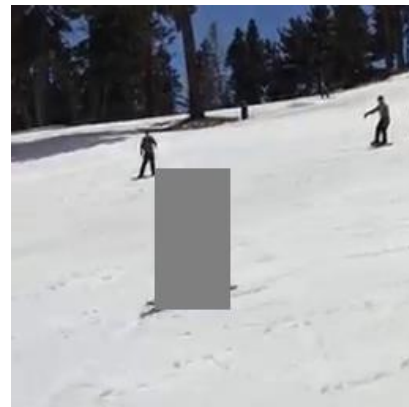
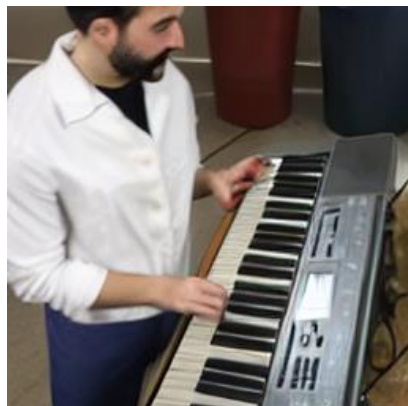
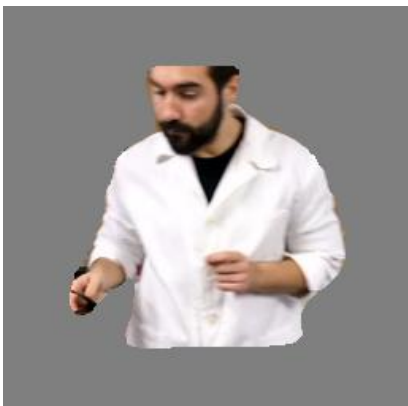
Input Scene

Input Person



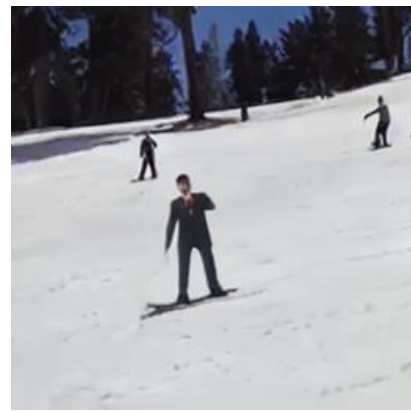
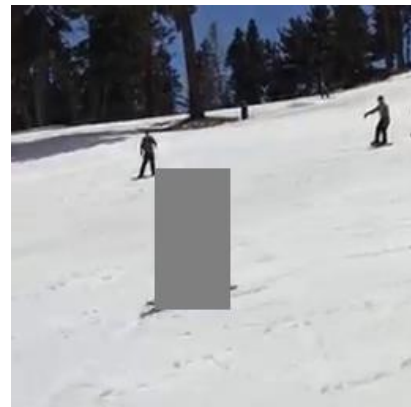
Input Person

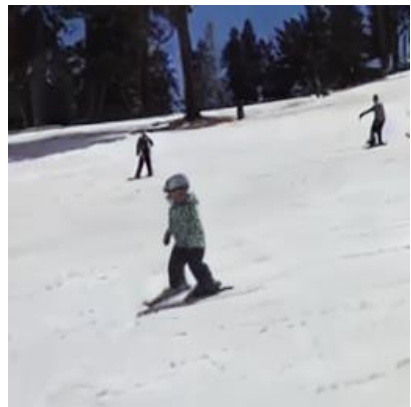
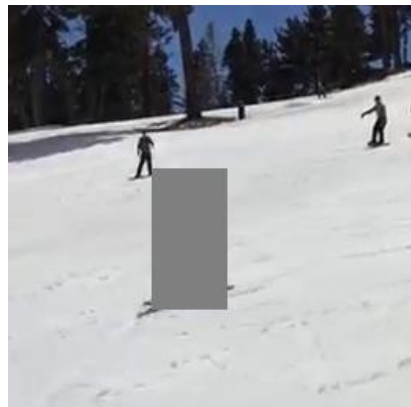
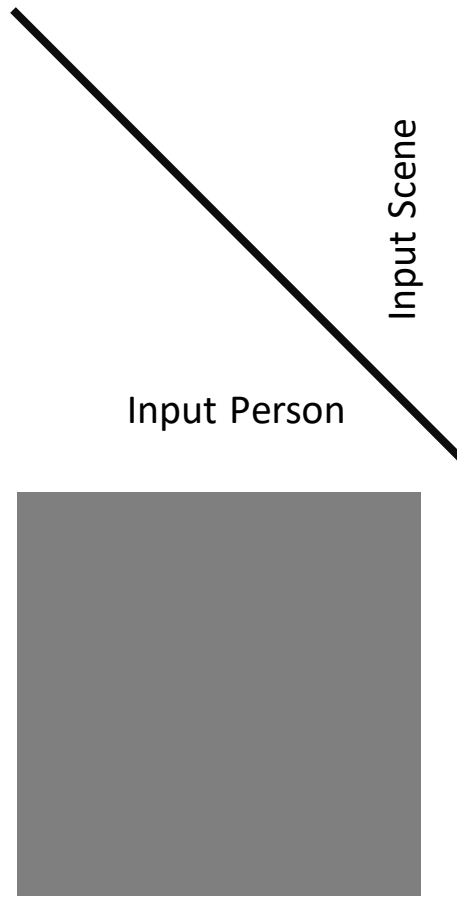
Input Scene



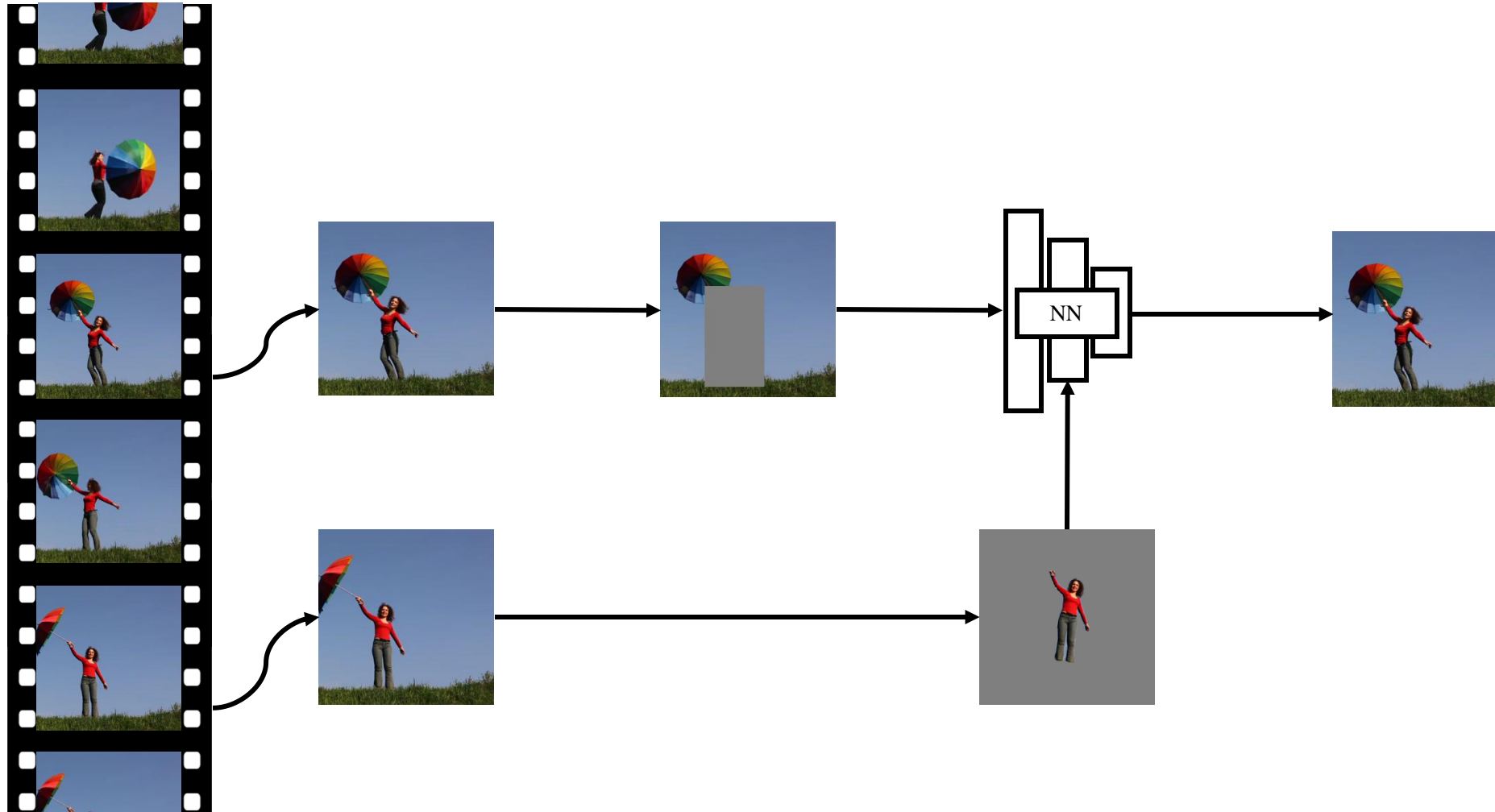
Input Person

Input Scene

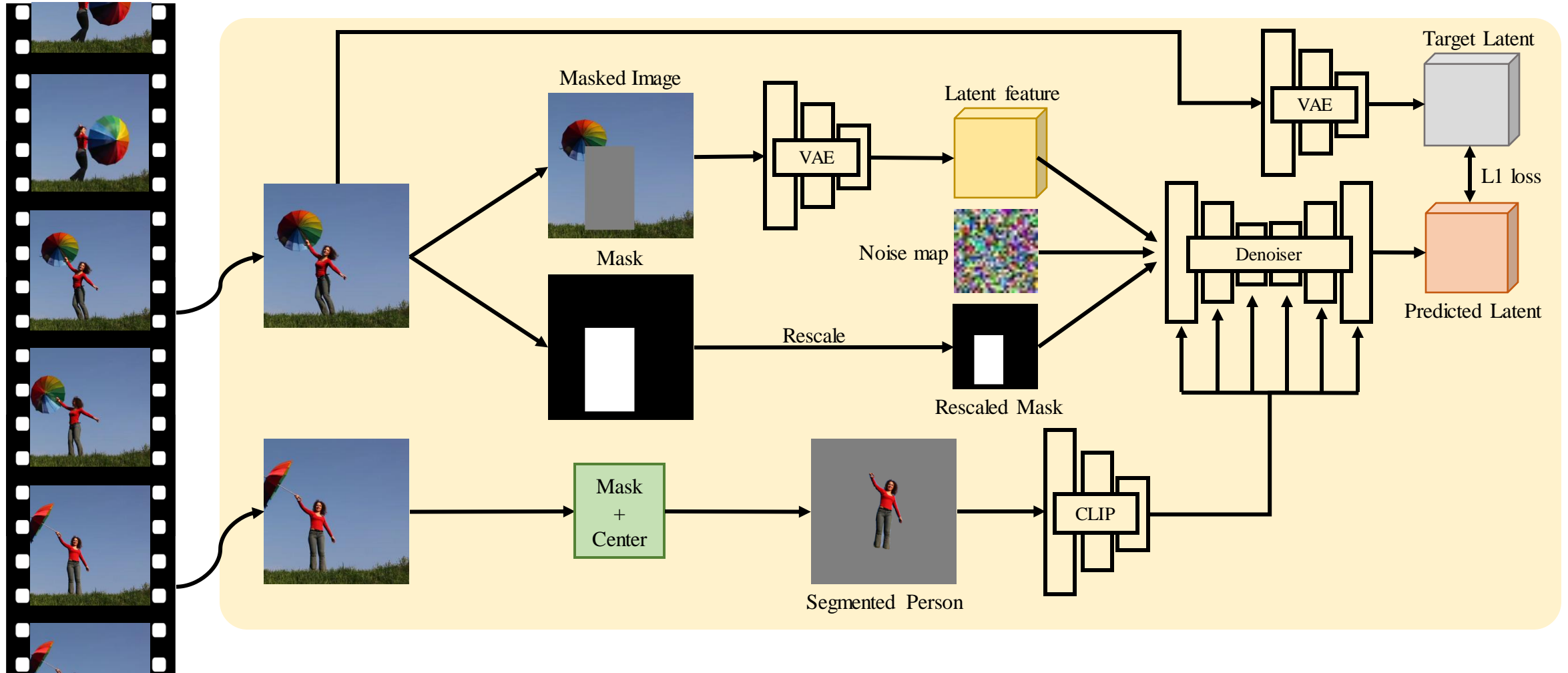




Core Idea



Learning Architecture

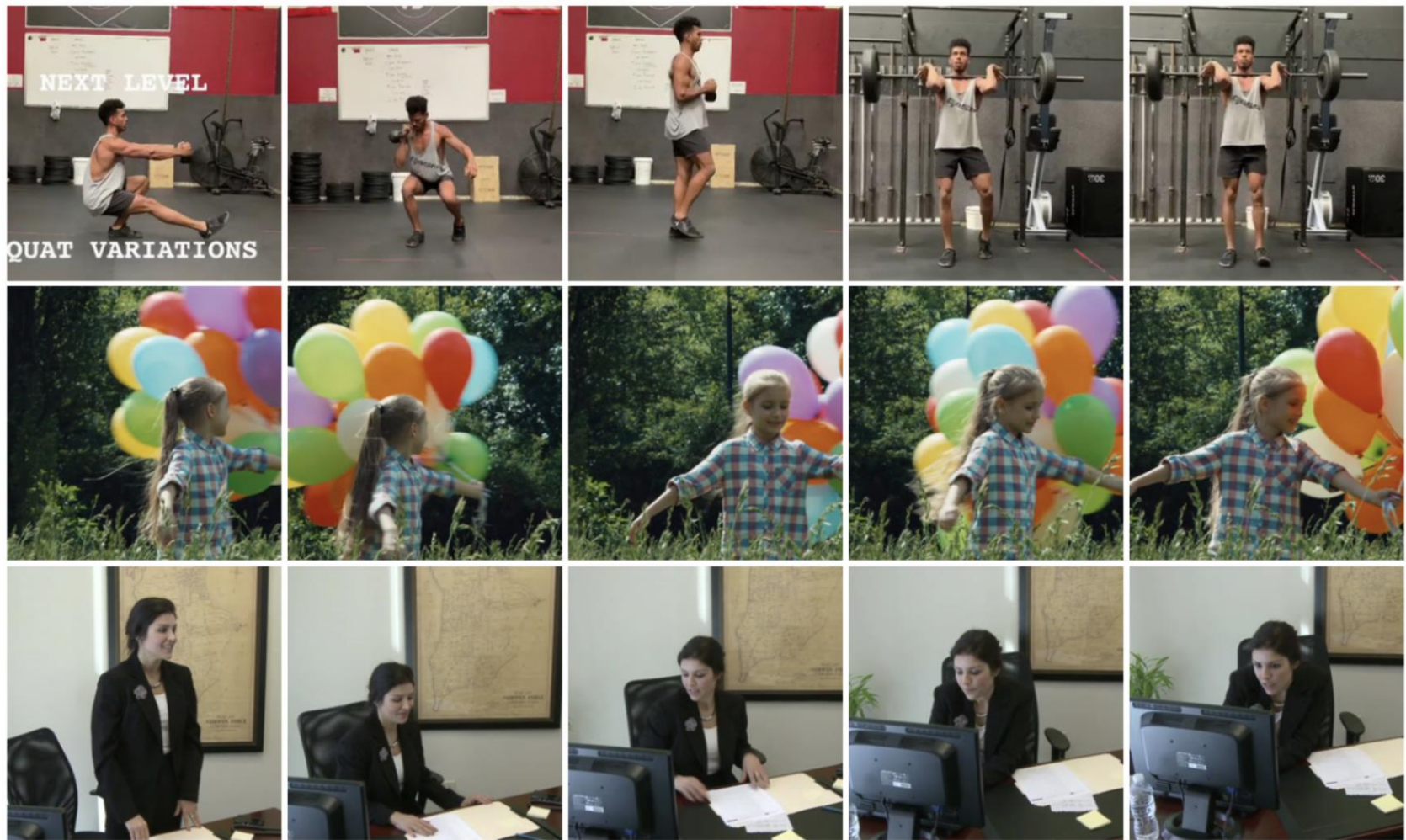


Dataset

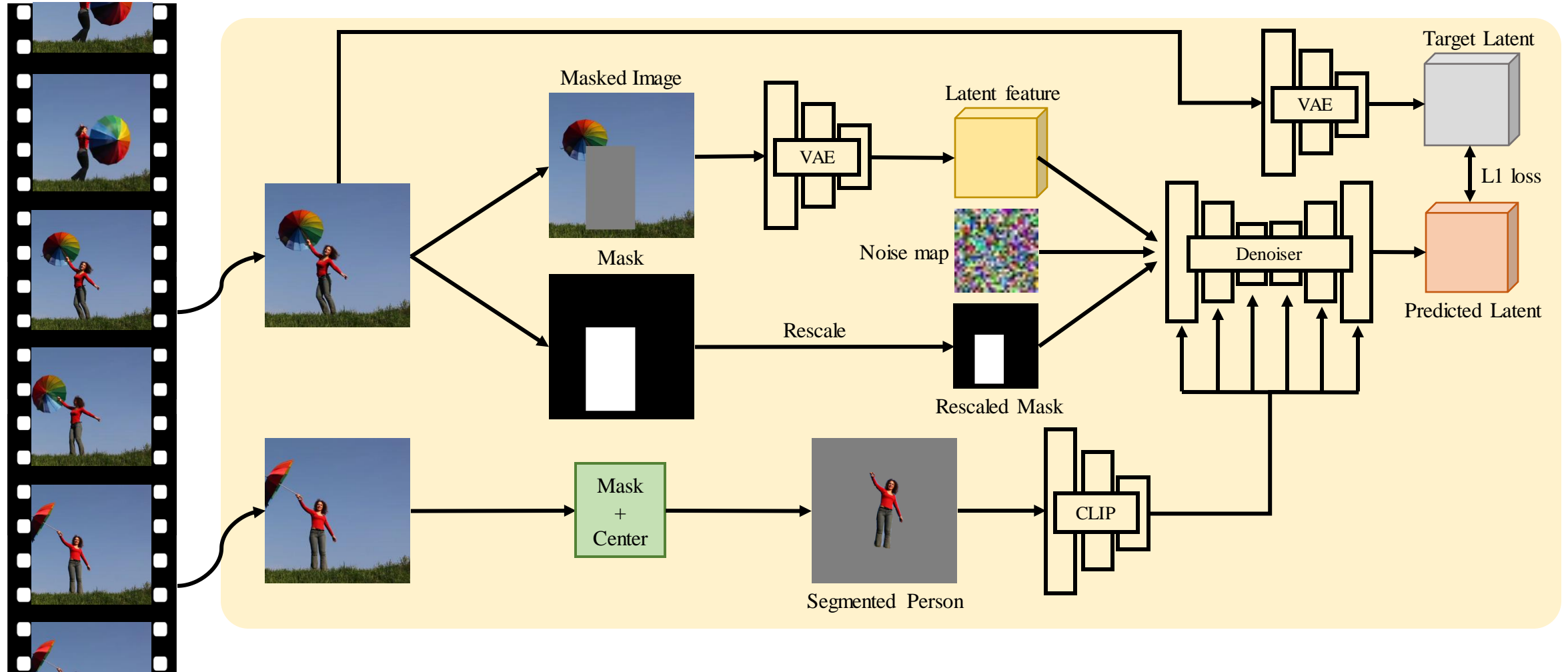


- Prepared a dataset of humans interacting with diverse scenes.
- Inspired by prior work, we search for 256 x 256 spatiotemporal segments with human presence.
- Our data source includes public computer vision datasets and internal Adobe data.
- Starting with 13 million videos, we ended up 2.4 million clips of interest after processing.

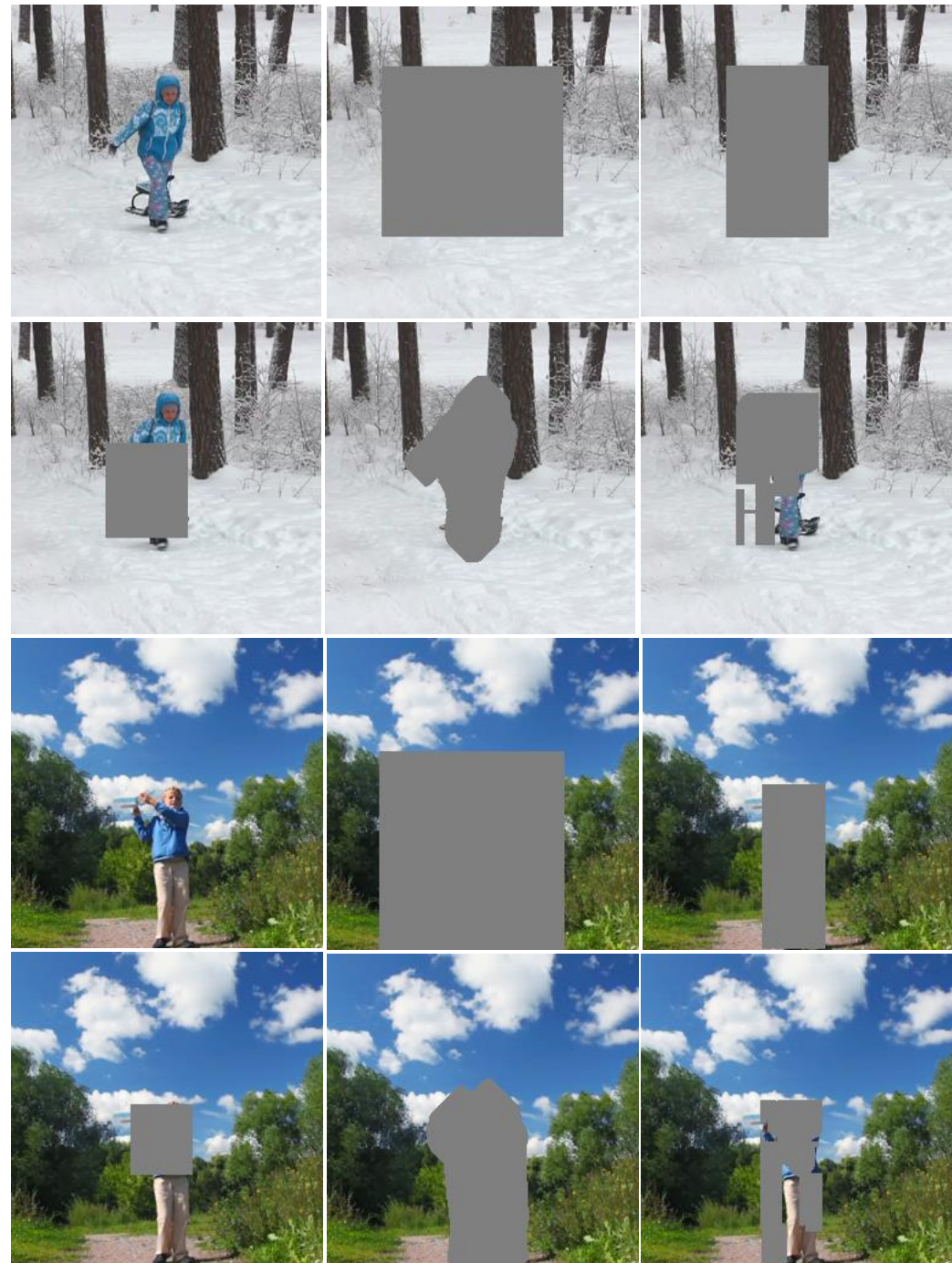
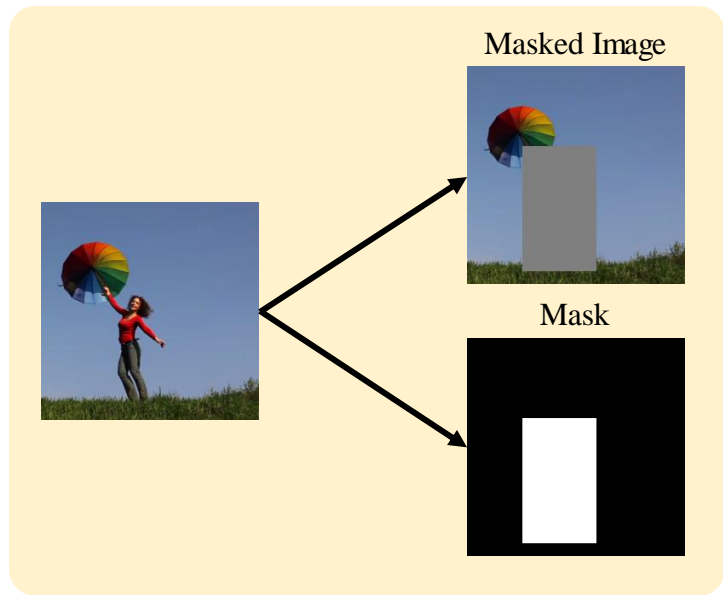
Dataset



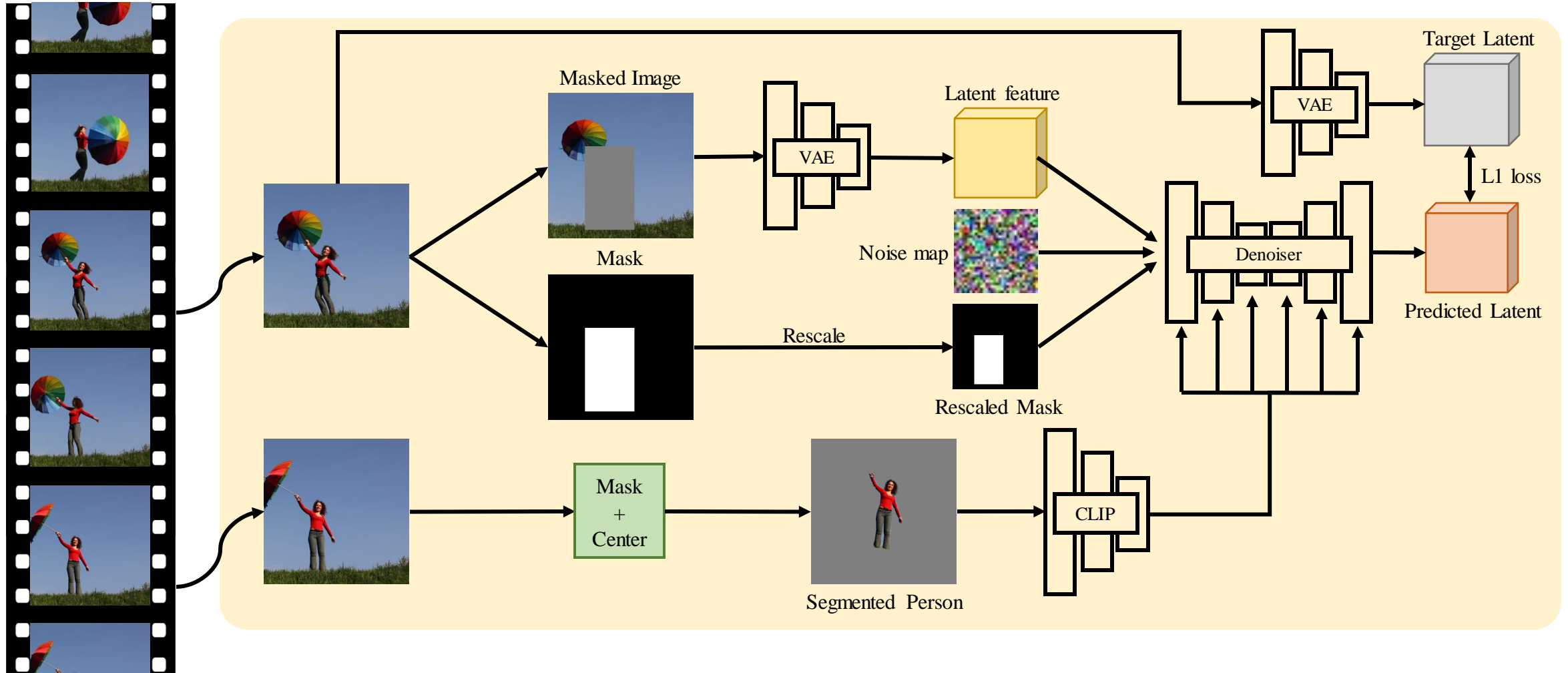
Learning Architecture



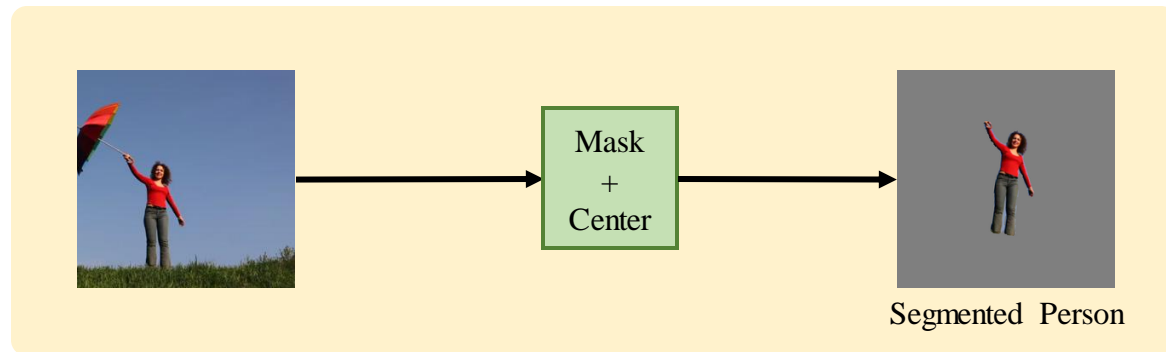
Masking strategy



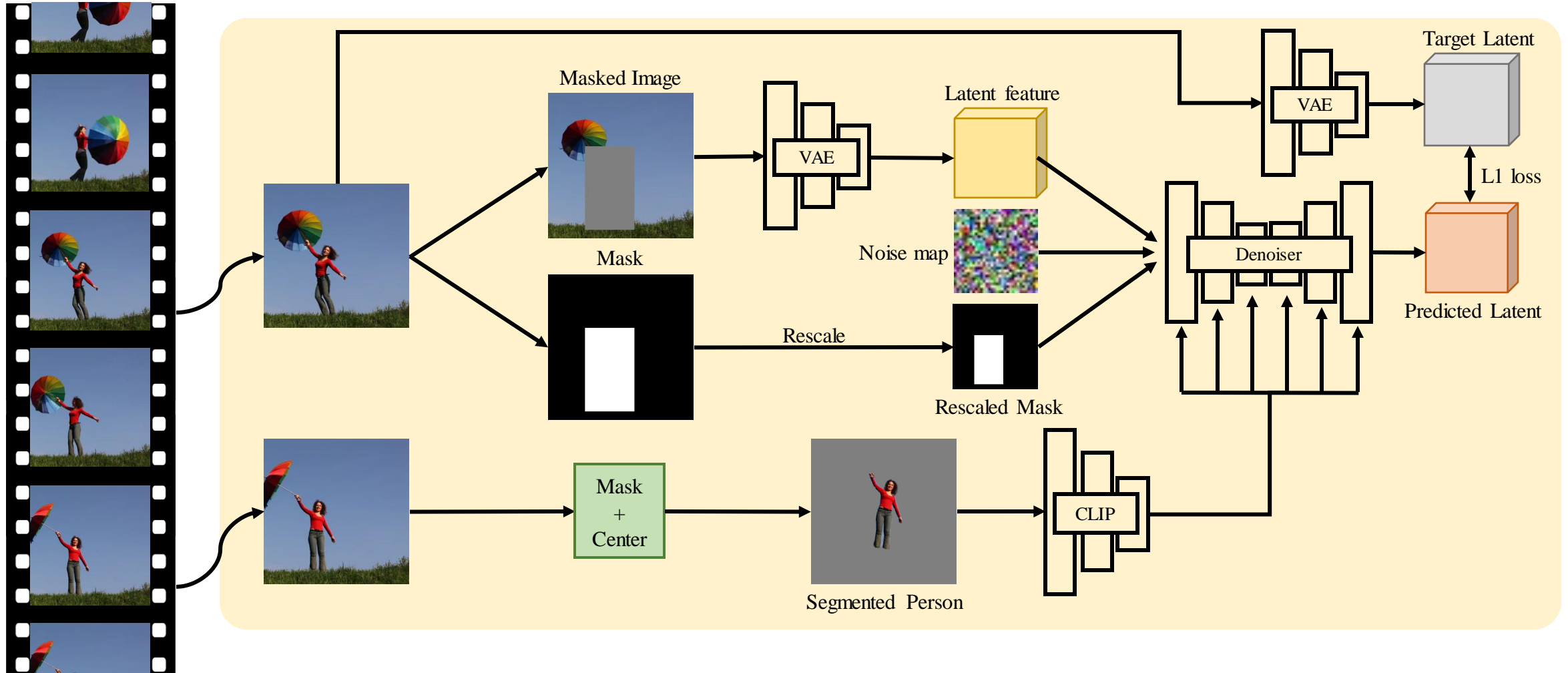
Learning Architecture



Reference person

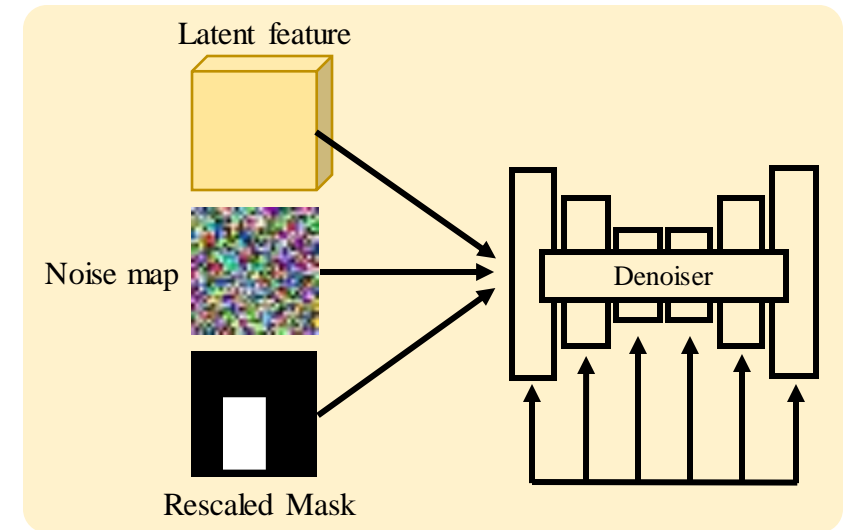


Learning Architecture

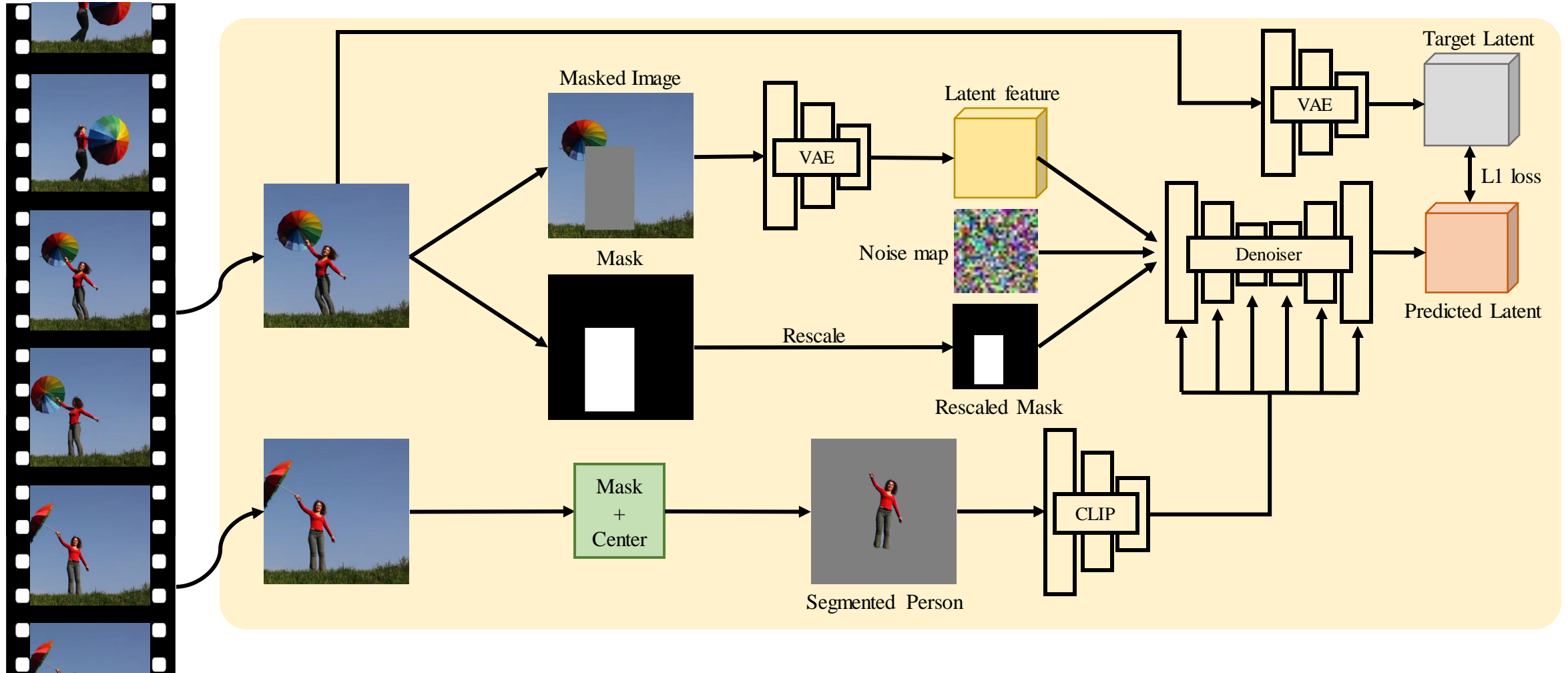


Denoising U-Net

- Follows standard distributed training procedures.
- Input scene passed through concatenation.
- Refer person passed through cross-attention
- Classifier-free guidance by dropping both the conditioning.

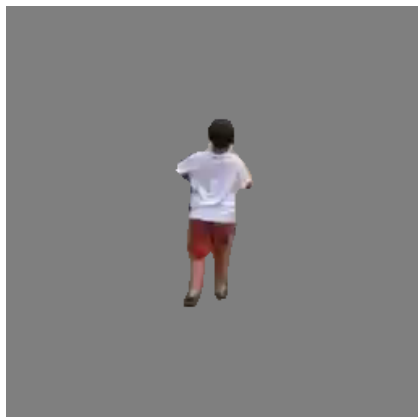


Learning Architecture

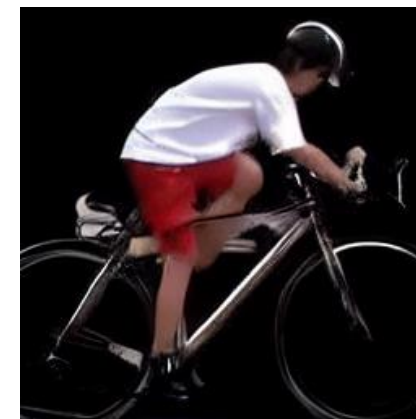
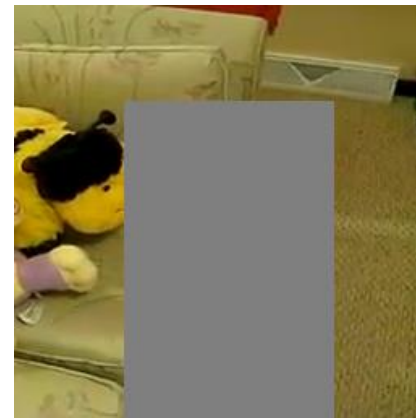
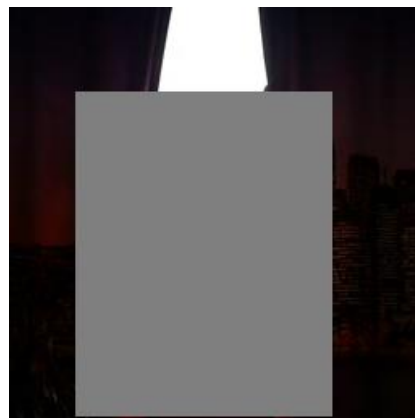
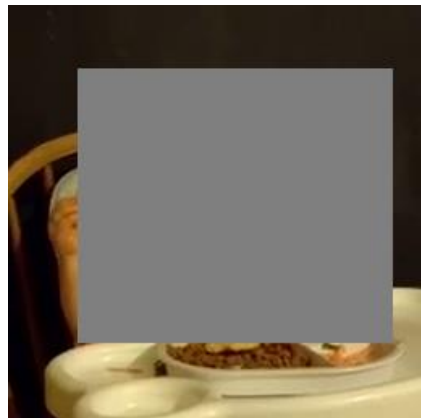


Person Conditioned Insertions

Scenes



Input



Samples

Person Conditioned Insertions

Scenes



Input



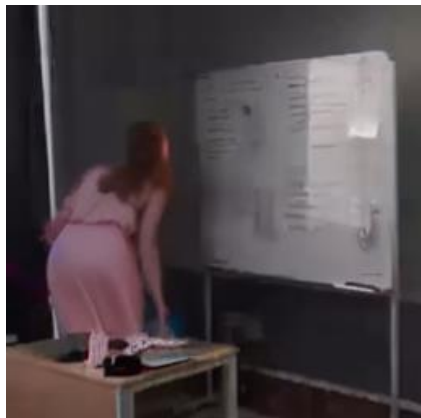
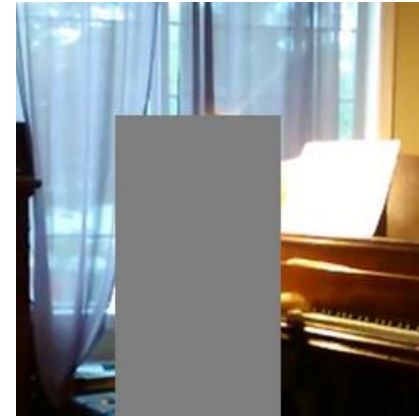
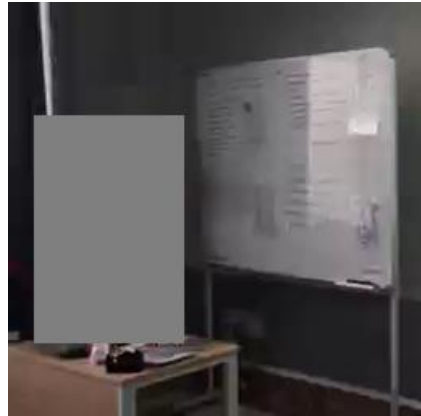
Samples

Person Conditioned Insertions

Scenes



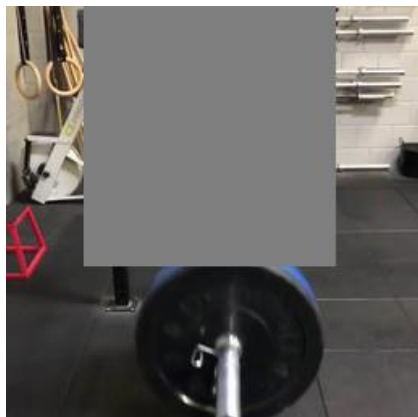
Input



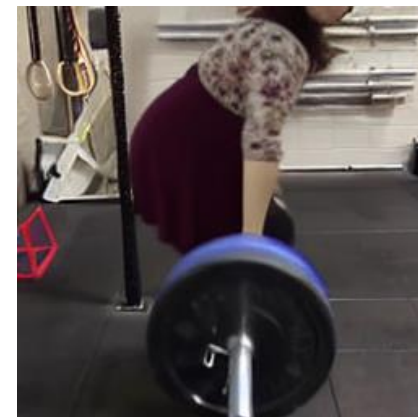
Samples

Person Conditioned Insertions

Reference



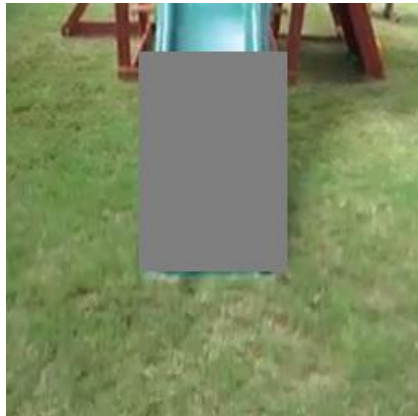
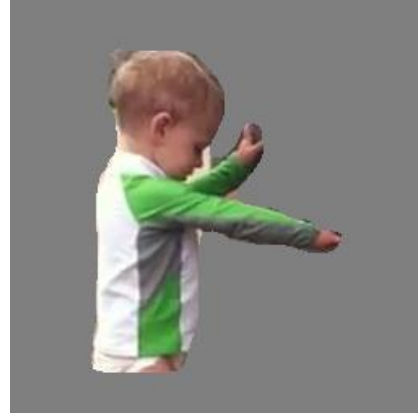
Input



Samples

Person Conditioned Insertions

Reference



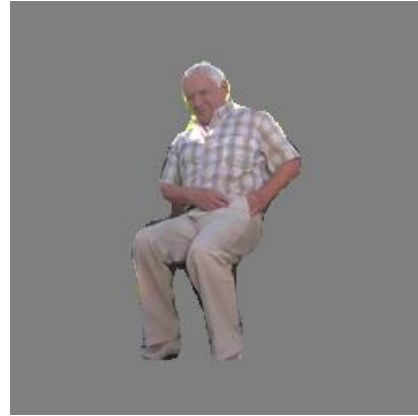
Input



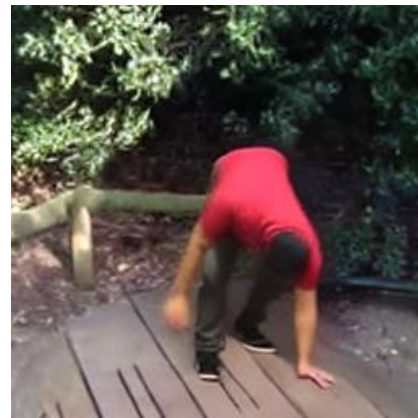
Samples

Person Conditioned Insertions

Reference



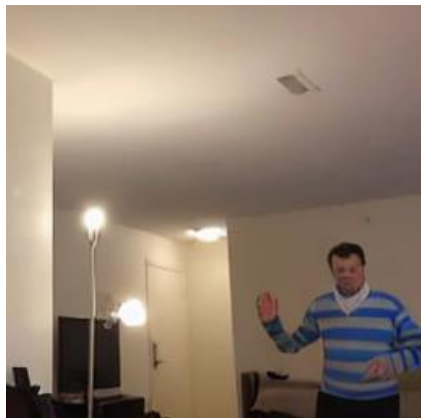
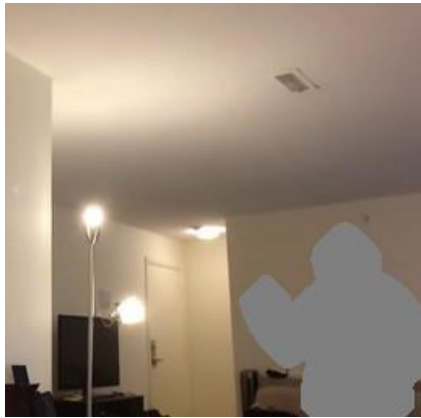
Input



Samples

Person Conditioned Insertions

Reference



Input

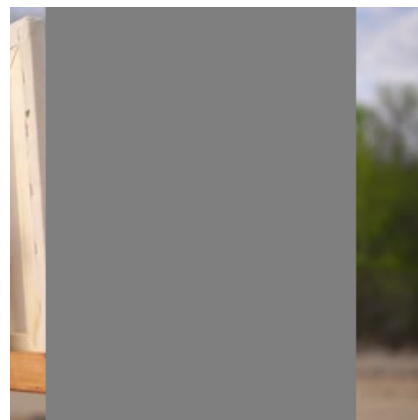
Samples

Architecture ablations - Data

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
Video (w/ aug)	10.078	17.602

- Video data is **crucial** for our task.
- Using image-only data even with augmentations doesn't work.

Person Hallucinations

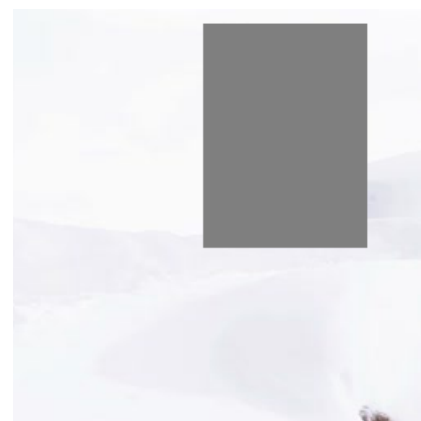
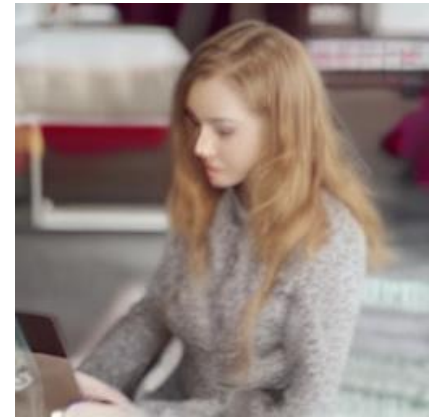
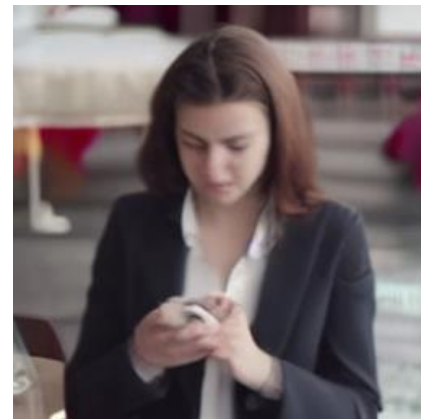
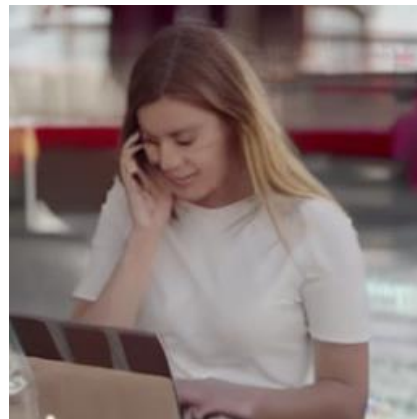
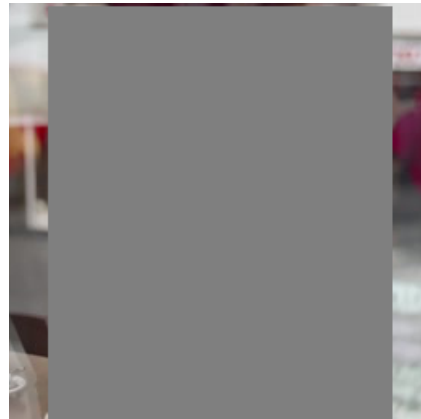
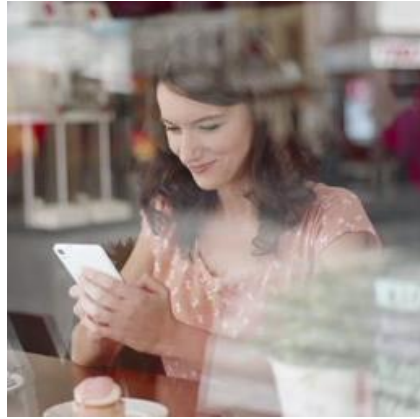


Ground-Truth

Input

Samples

Person Hallucinations

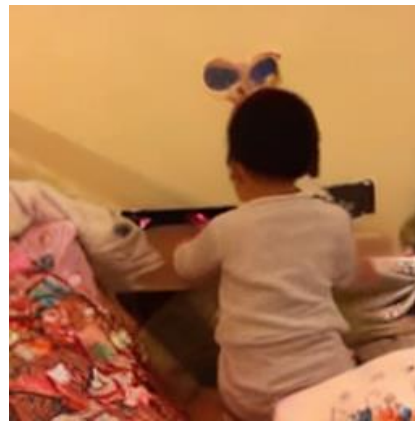
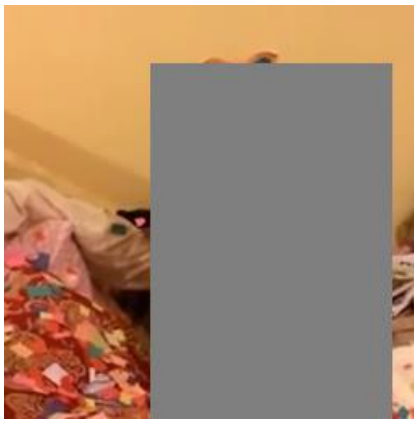
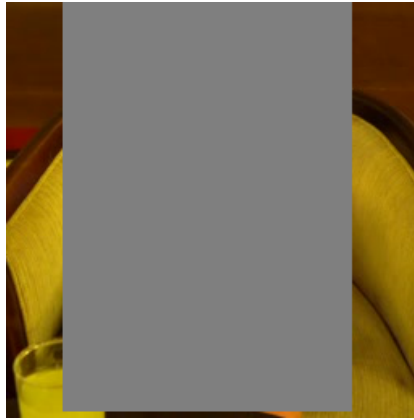


Ground-Truth

Input

Samples

Person Hallucinations



Ground-Truth

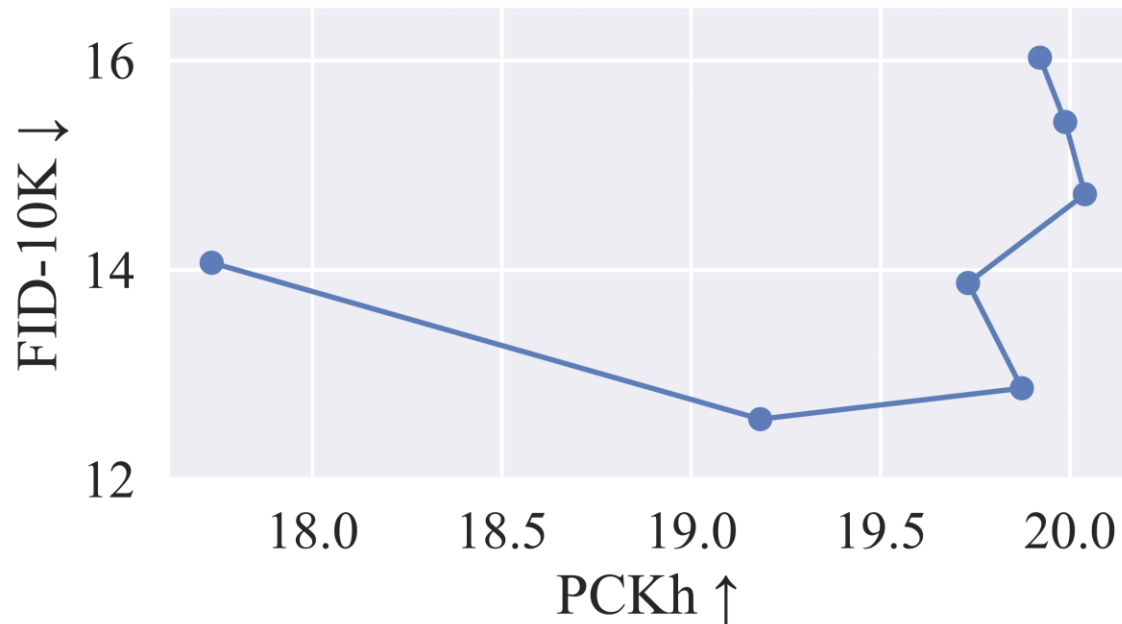
Input

Samples

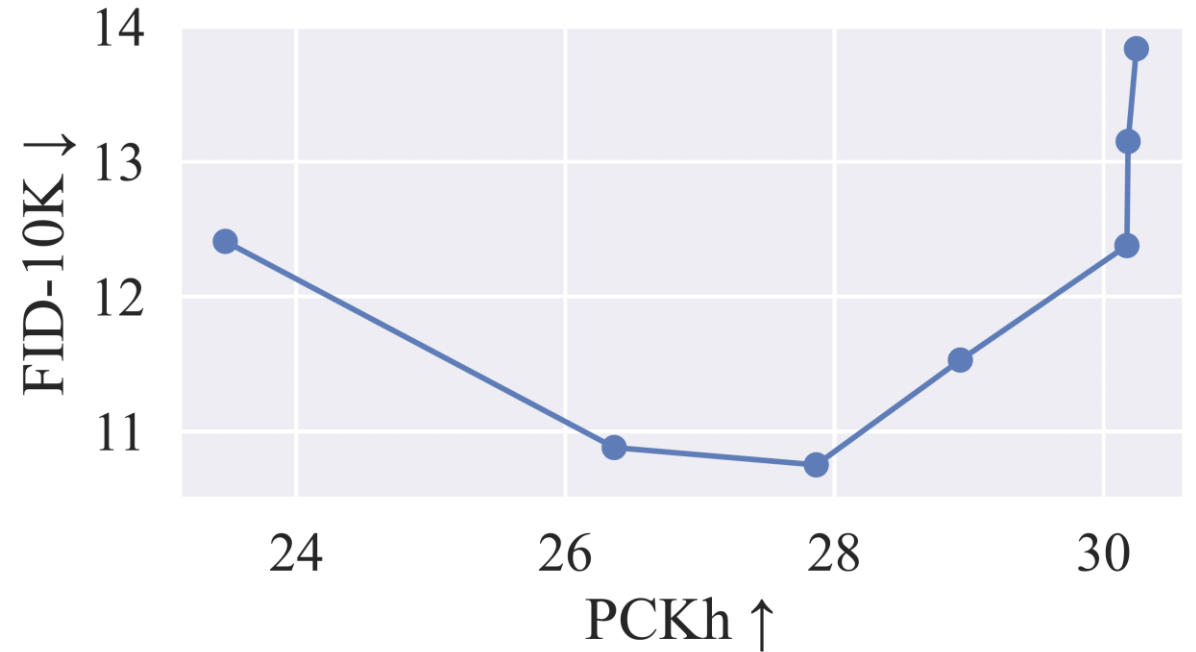
Classifier-Free Guidance

Effect of increasing CFG guidance scale.

Evaluated at scale values [1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0]

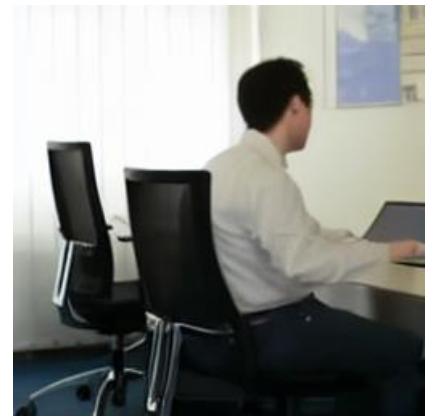
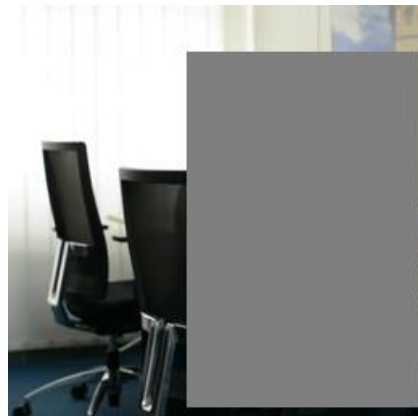
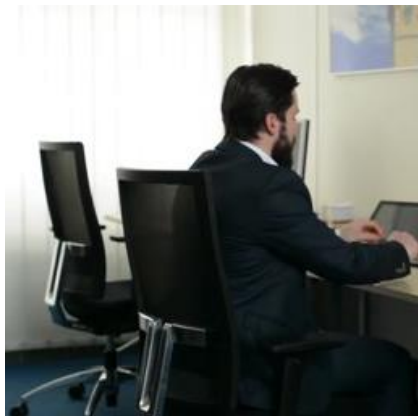
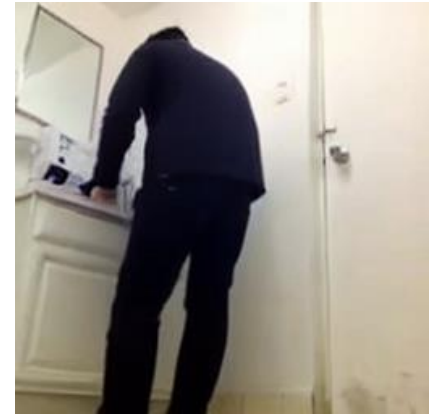
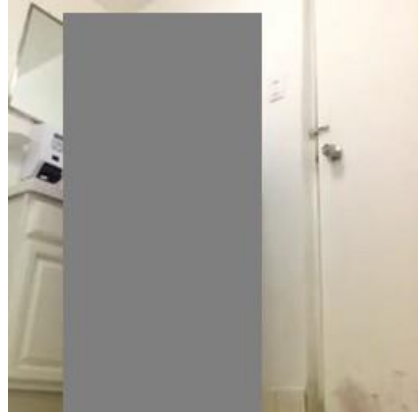


Person Conditioned



Person Hallucination

Person Hallucinations - Baselines



Ground-Truth

Input

DALL-E 2

Stable-Diffusion v1.5

Ours

Constrained Scene Hallucination

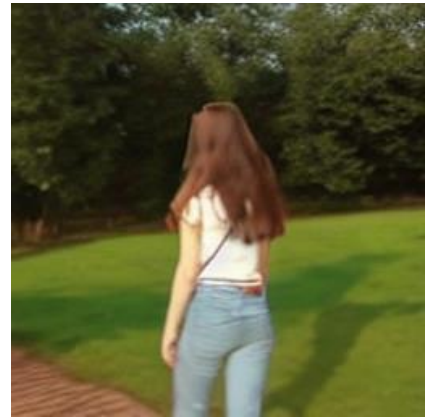
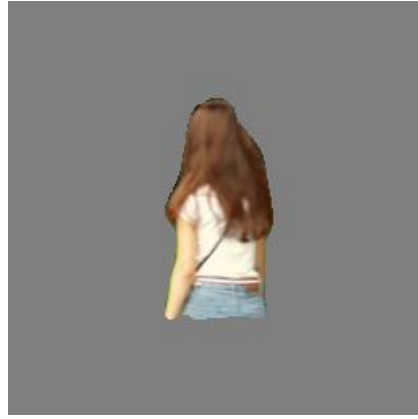


Ground-Truth

Input

Samples

Constrained Scene Hallucination - Baselines



Ground-Truth

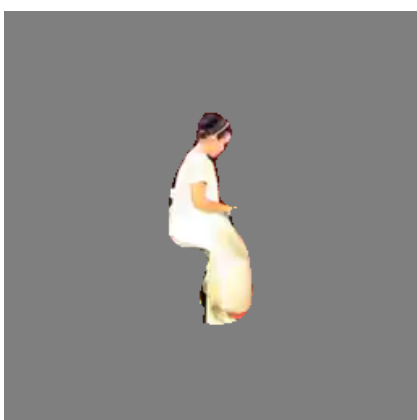
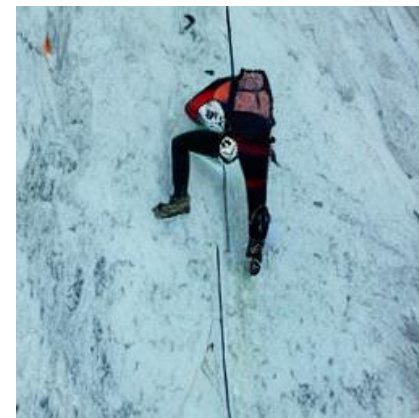
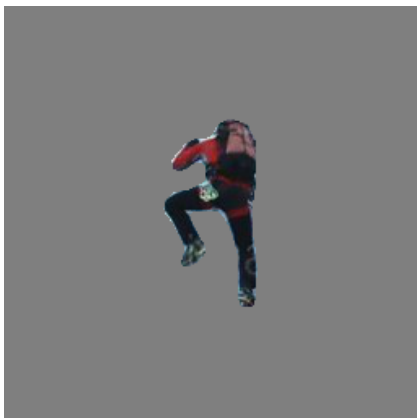
Input

DALL-E 2

Stable-Diffusion v1.5

Ours

Unconstrained Scene Hallucination



Ground-Truth

Input

Samples

Partial Body Completions

Reference

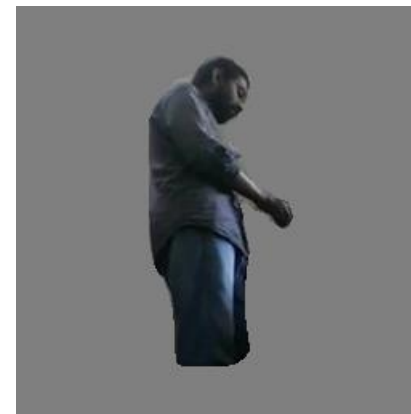
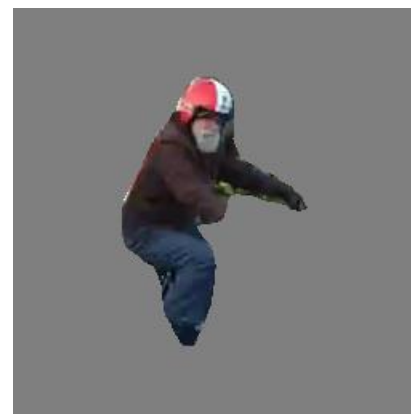
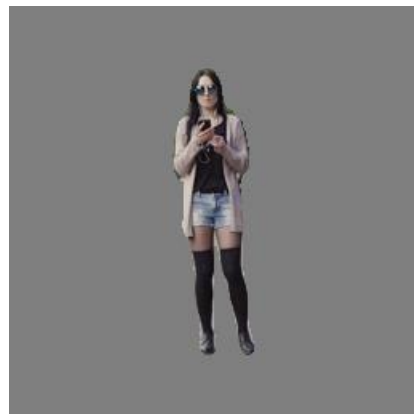


Input

Samples

Partial Body Completions

Reference



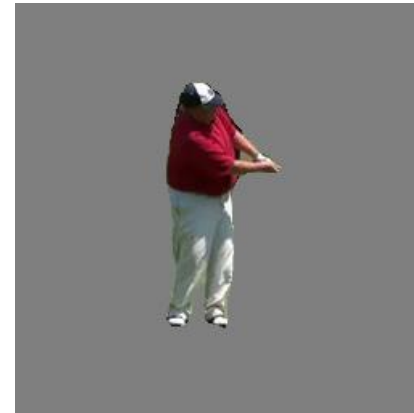
Input



Samples

Partial Body Completions

Reference



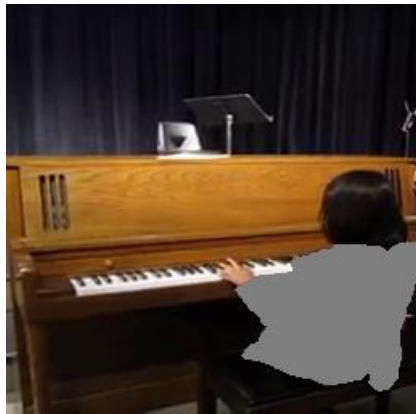
Input



Samples

Cloth Swapping

Reference clothes



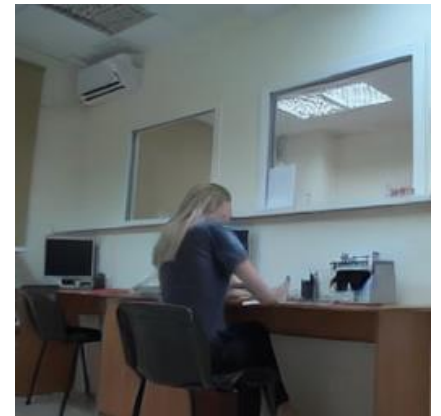
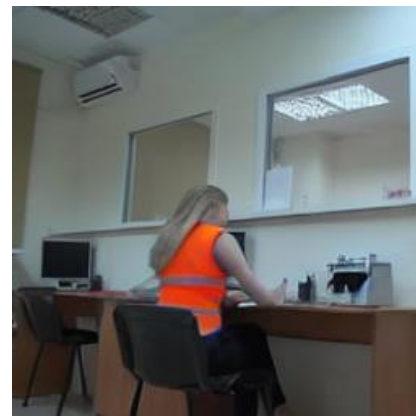
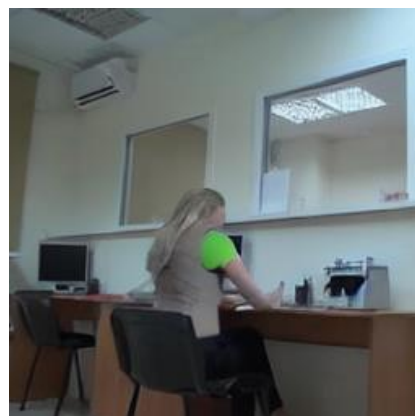
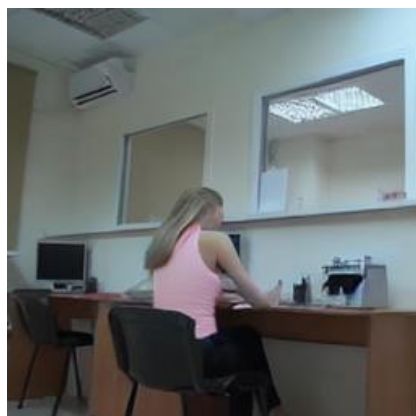
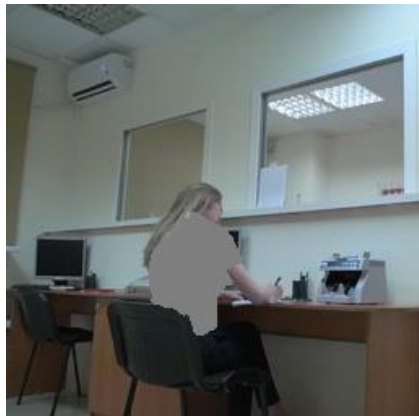
Input



Samples

Cloth Swapping

Reference clothes



Input

Samples

Failure Cases

- Faces & limbs are bad due to Stable Diffusion's VAE.
- Harmonization is off (top).
- Humans not fully reposed (bottom).



In summary

