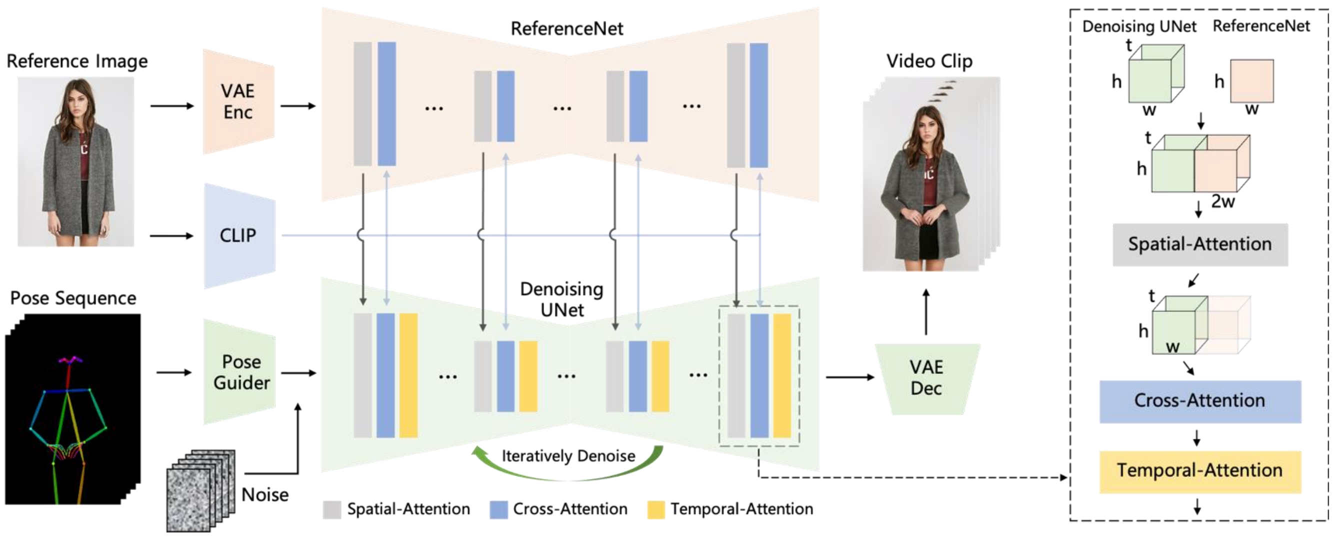
HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation

NeurIPS Dataset & Benchmark Track 2024

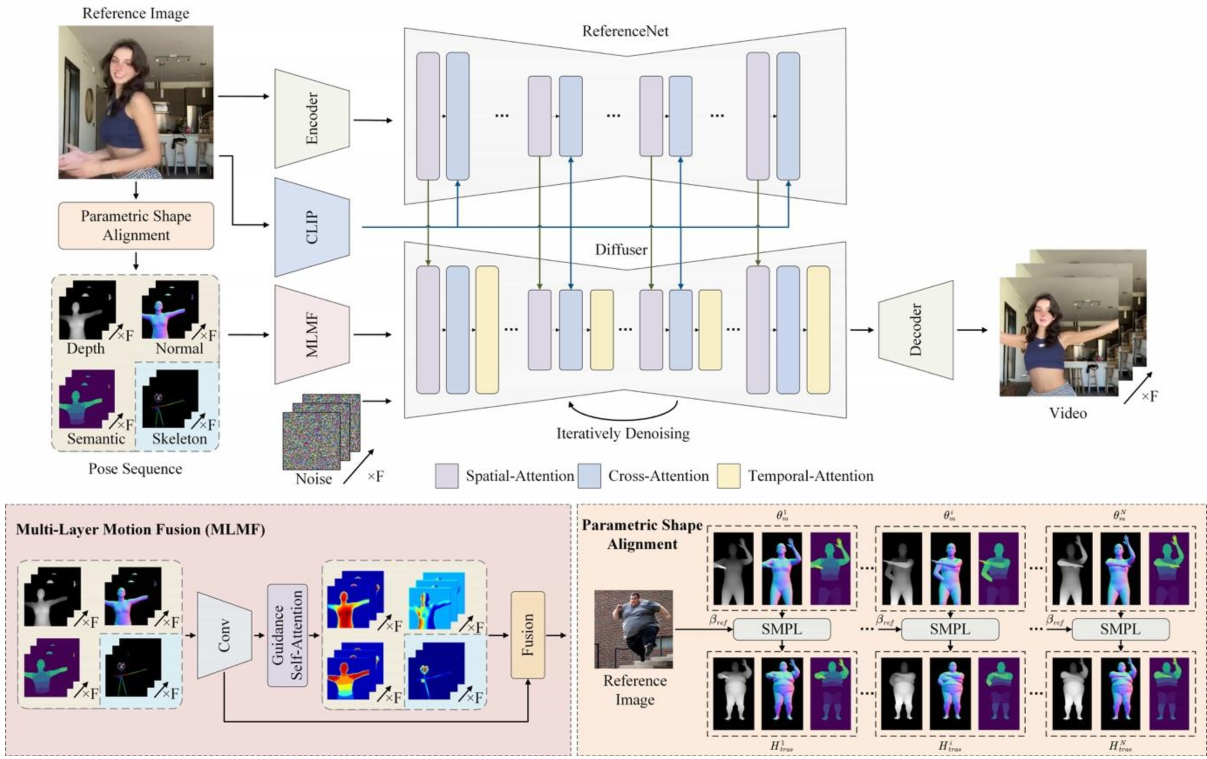
Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo,
Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, Dahua Lin
CUHK, Shanghai AI Lab, HKU

Project page: https://humanvid.github.io/

Pose-Guided Human Image Animation



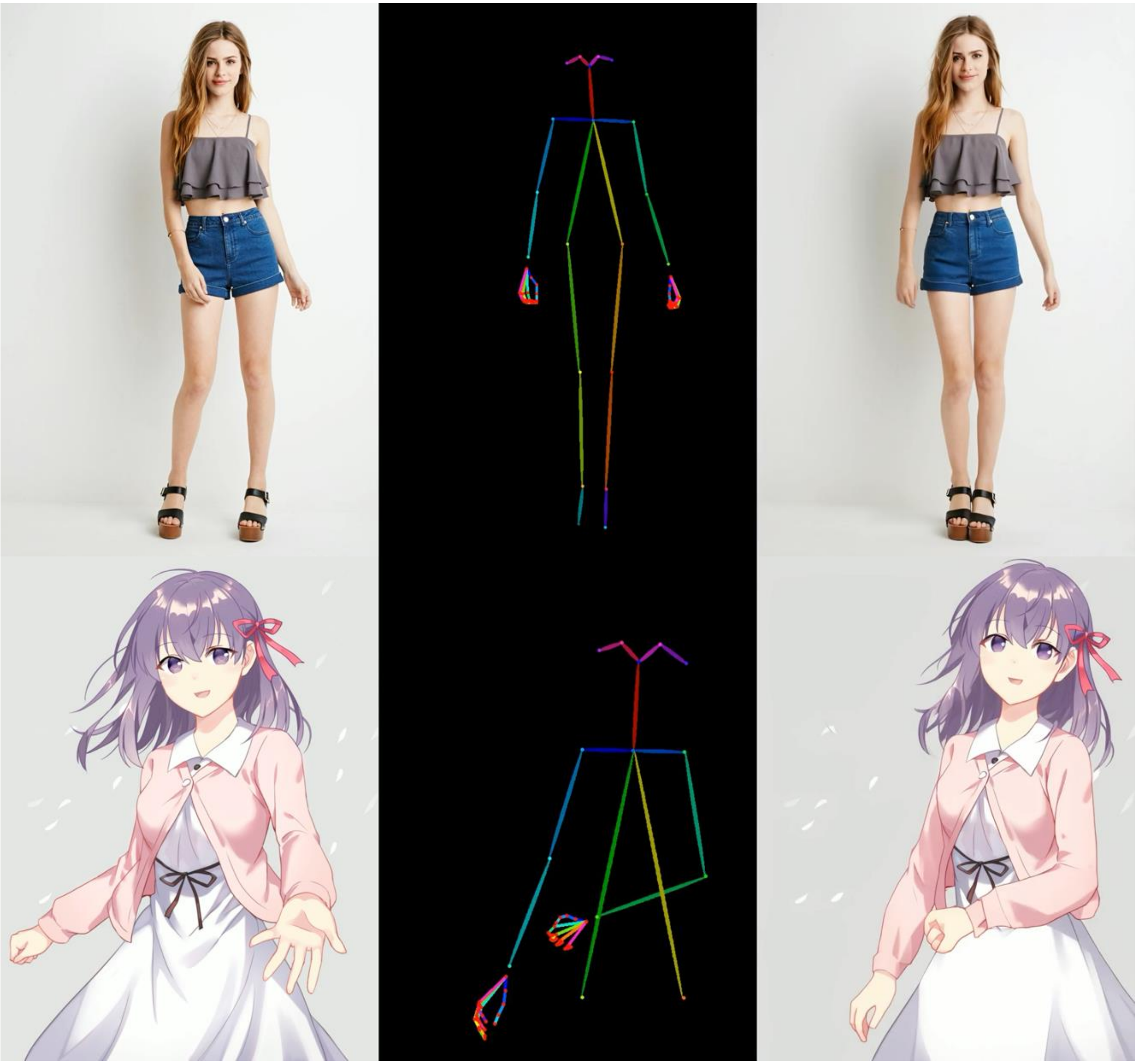
Animate Anyone [Hu et. al. 2024]



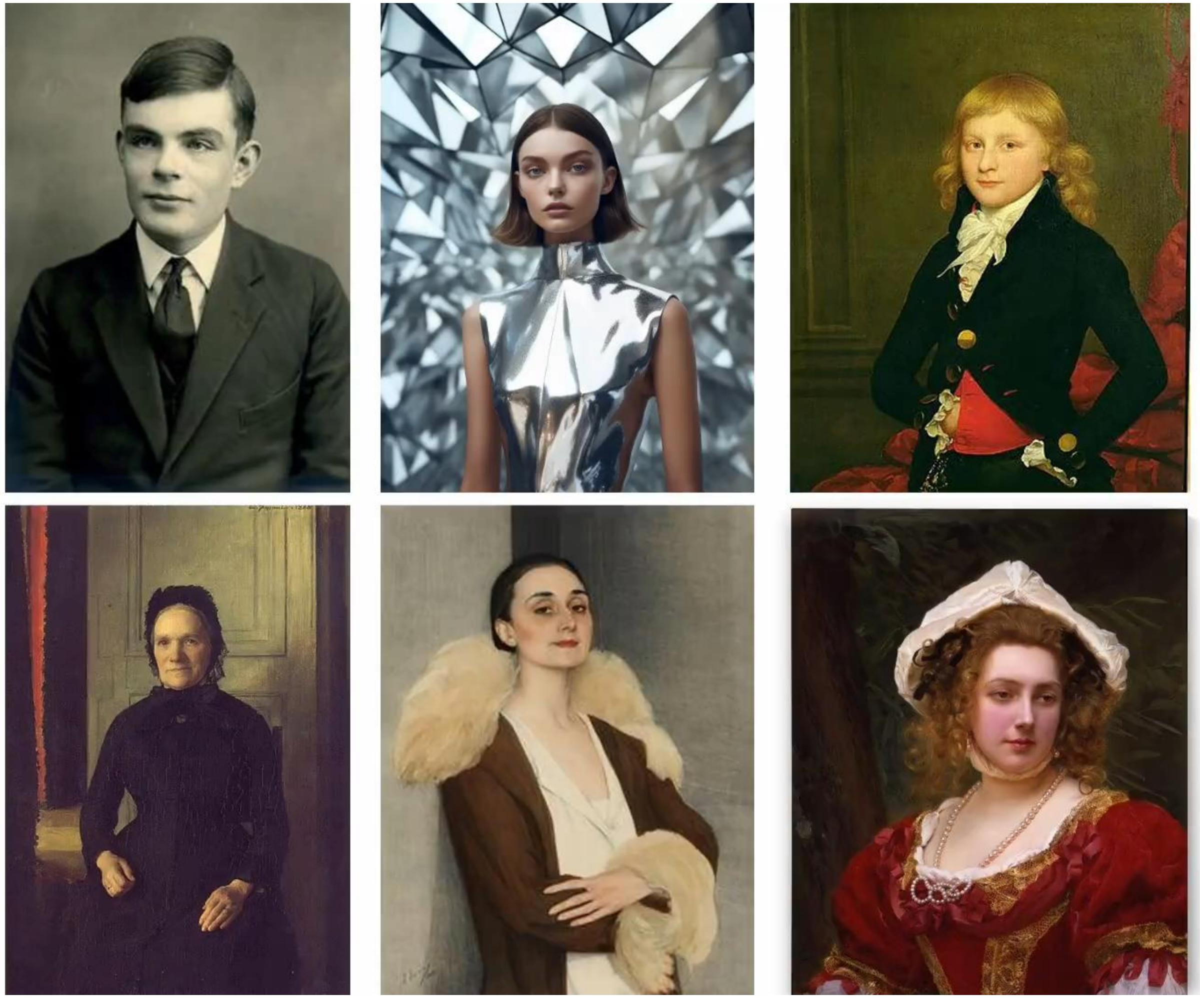
Champ [Zhu et. al. 2024]

Limitations: (1) Static-camera videos only; (2) No public training data.

Pose-Guided Human Image Animation



Animate Anyone [Hu et. al. 2024]

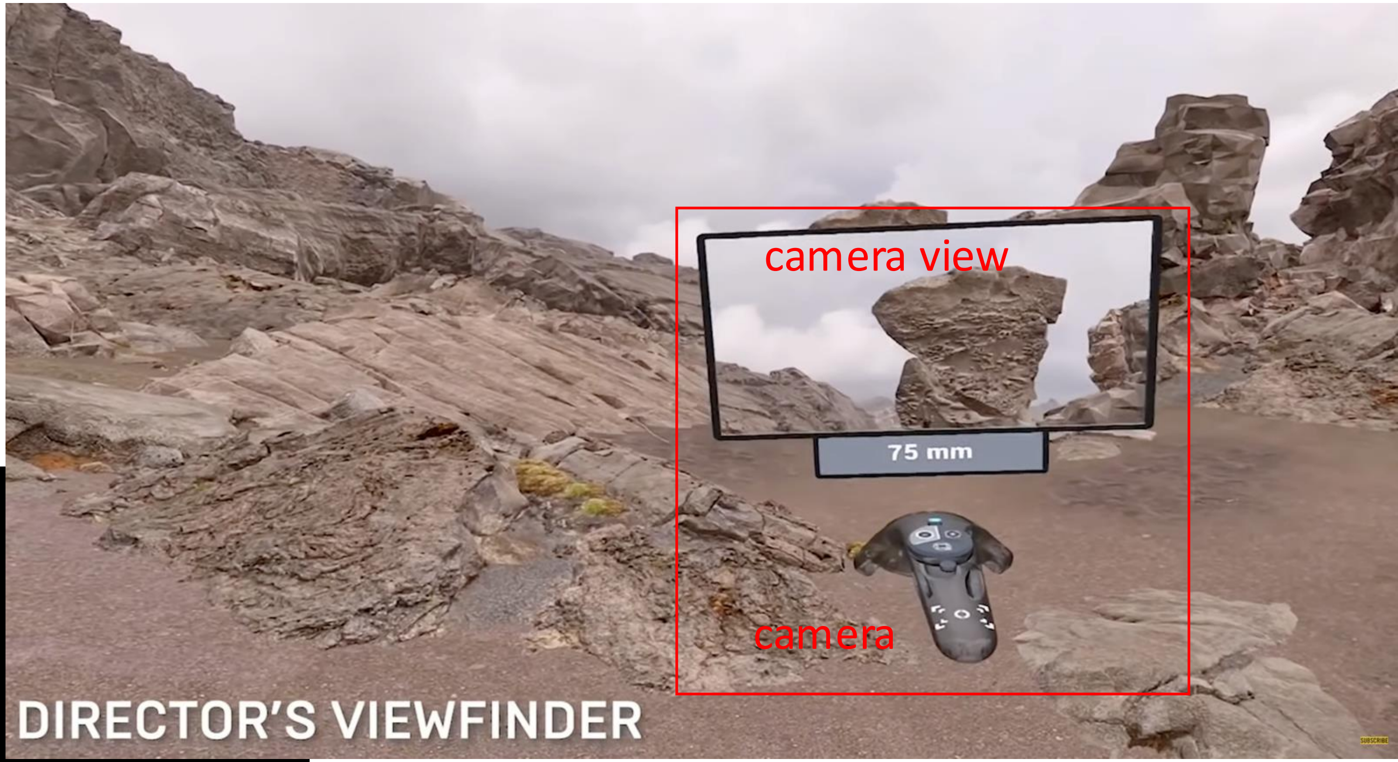


Champ [Zhu et. al. 2024]

How about Creating a Movie?

In addition to subject movement in Animate Anyone, movie creation requires more controllable abilities: Camera, Scene and Object.

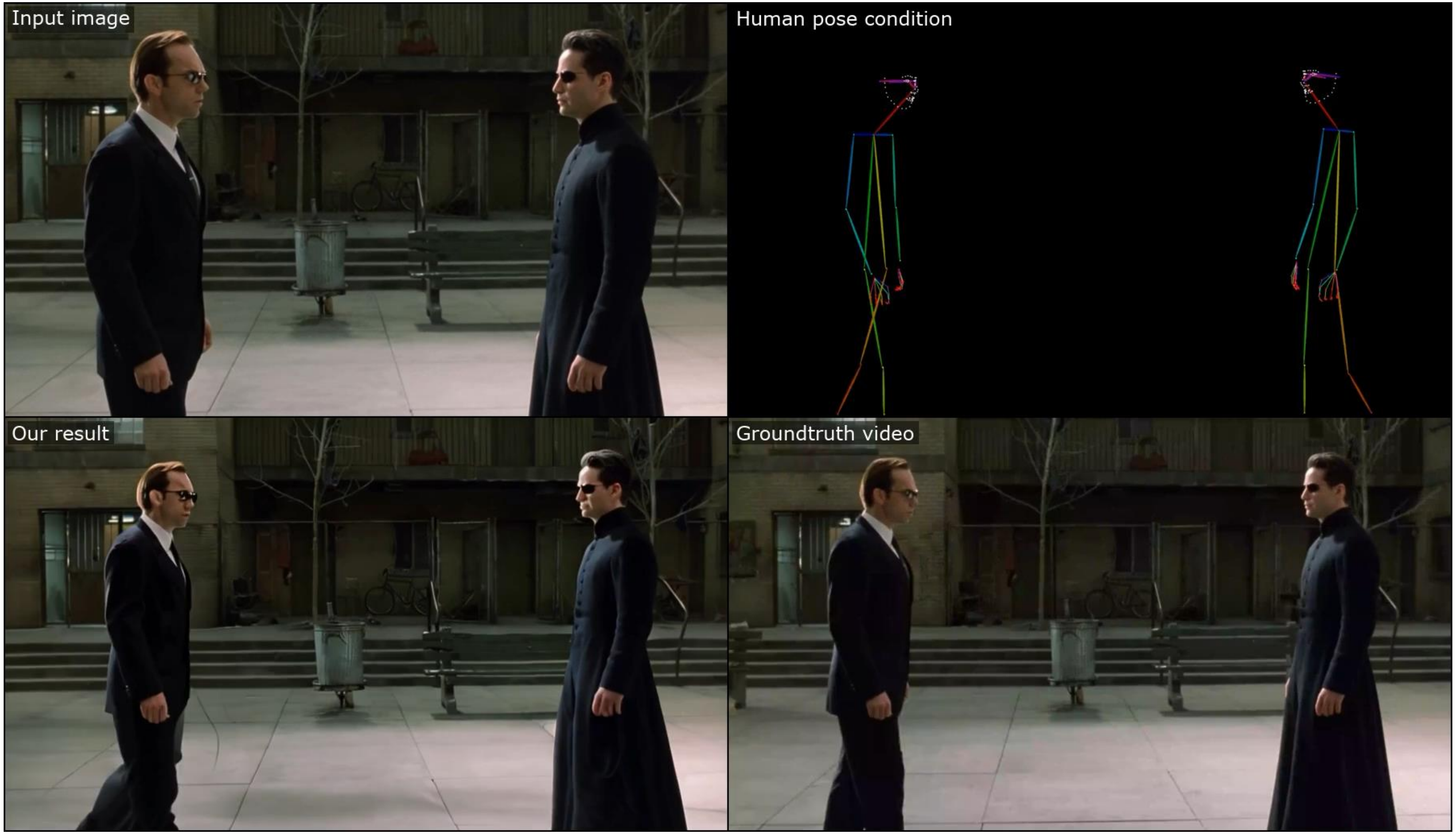
In this paper, we focus on controlling both human movement and camera movement in video diffusion models.



How are movies made?
https://www.youtube.com/watch?v=NnnuleYz8vU

Camera-controllable Human Image Animation

Challenges:
(1) Training data;
(2) Model design



Existing Video Datasets: WebVid10M [Bain et. al. 2021]

- Pros: Large scale, estimated #human videos > 300 K
- Cons:
  - Low resolution, only horizontal orientation, watermarks,
  - No camera annotations, some frames may not contains human appearance if the camera movement is large
  - There are a lot of human-object interactions, it is difficult to remove them

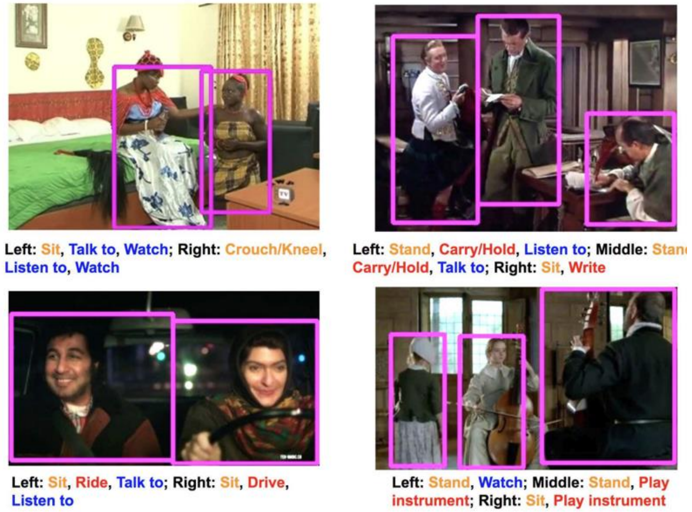
Good example: large human IoU, no objects

Bad example: many objects, no human face in some frames

Existing Video Datasets: Kinetics & AVA

- Pros: Large scale
- Cons:
  - Videos in AVA could have shot changes
  - High resolutions but low quality, humans may only occupy a very small region
  - Unstable camera poses without annotations, some frames may not contains human appearance if the camera movement is large
  - There are a lot of human-object interactions, it is difficult to remove them

Kinetics Dataset [Kay et. al. 2017]



AVA Dataset [Gu et. al. 2018]

Existing Video Datasets: UBC & TikTok

- Pros: Human videos, high resolution, good quality
- Cons:
  - Very small scale (500 for UBC, 340 for TikTok), hard to finetune video diffusion models on them
  - Static camera only
  - Limited diversity of background

TikTok Dataset [Jafarian et. al. 2021]

UBC Dataset [Zablotskaia et. al. 2019]

Existing Video Datasets: RealEstate10K

- Pros: Large-scale high-quality videos with annotation of cameras
- Cons:
  - There are no humans. Scene only.

RealEstate10K Dataset [Zhou et. al. 2018]

Summary of Existing Video Datasets

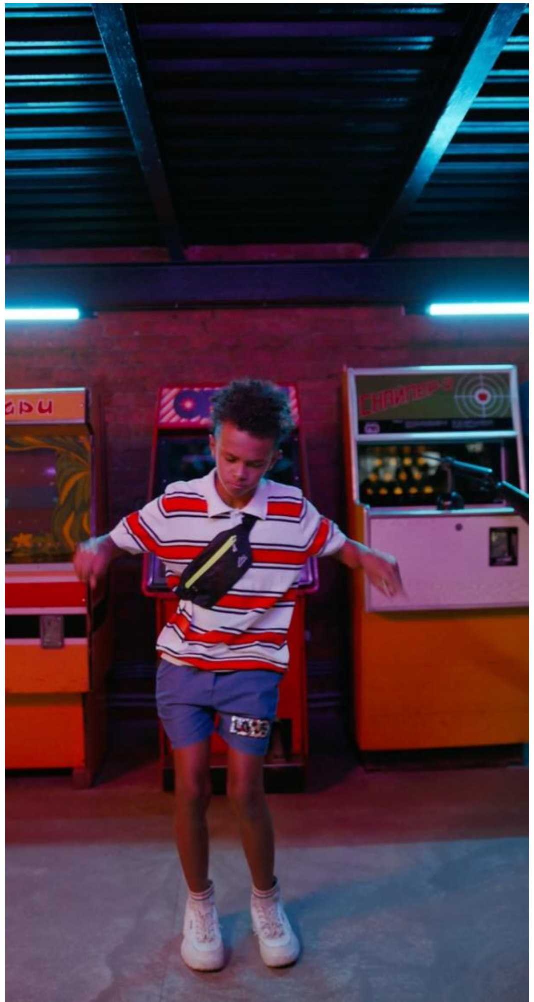
- No large-scale human video dataset with high-quality, high resolution videos
- No human video datasets have high-quality camera annotations
- It is even hard to reproduce the performance of methods like Animate Anyone and Champ due to their private datasets!

High-quality Human Video Dataset

- How to curate a high-quality human video dataset?
  - Static-camera videos only: hard to collect, cannot enable camera control
- Eliminate the static camera requirement in data collection:
  - Current methods (Animate Anyone & Champ) cannot modeling camera movements, so we cannot directly animate human images only condition on human pose: add a camera pose encoder!
  - Large scale Internet videos do not have accurate camera annotations and complex camera trajectories: add a synthetic part that ensure accurate and complex camera poses!
- Then, we could curate human videos from a large corpus according to:
  - There are humans: large IoU of human and small #human, by a human detector
  - Humans are prominent subjects: confidence of detected human pose keypoints

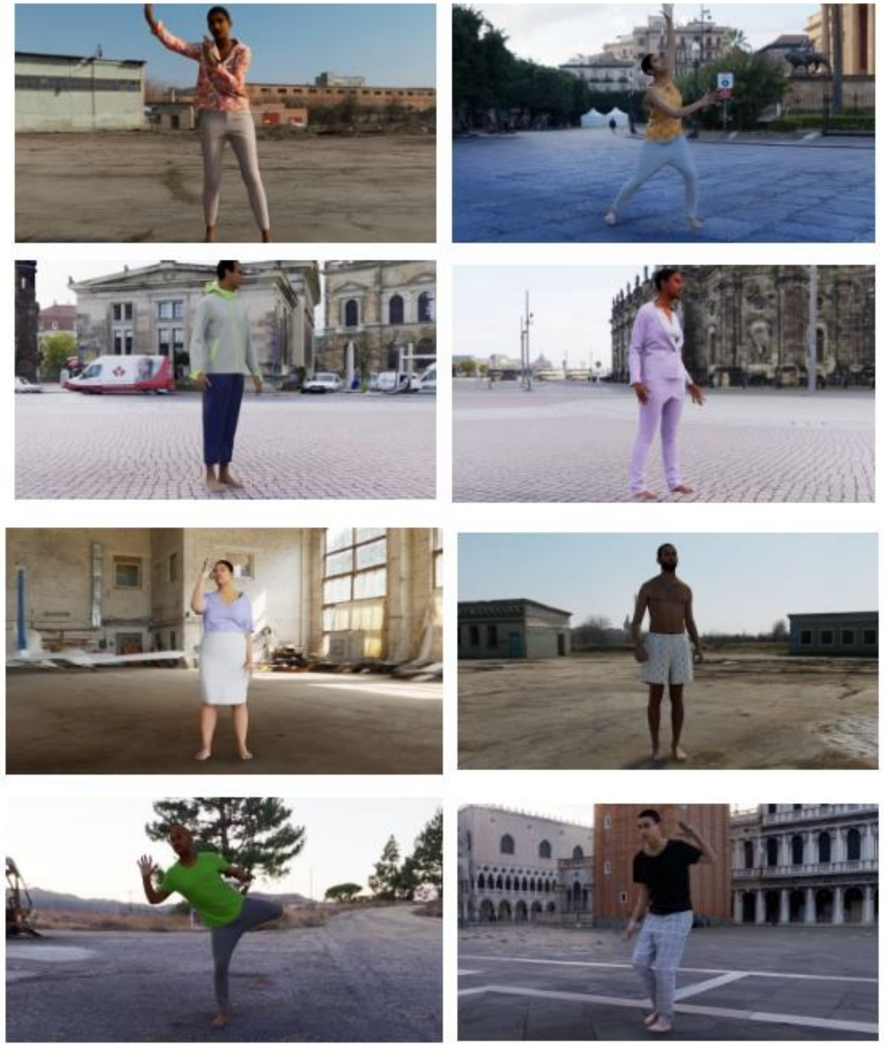
Internet Videos

- A great copyright-free internet video platform: pexels.com
- Some curated video examples



Synthetic Videos - Illustration

- However, we still need accurate camera annotations. -> Synthetic data!



Synthetic Videos – Camera Trajectories

Camera
trajectory,

2 keyframes

Camera
trajectory,

3 keyframes

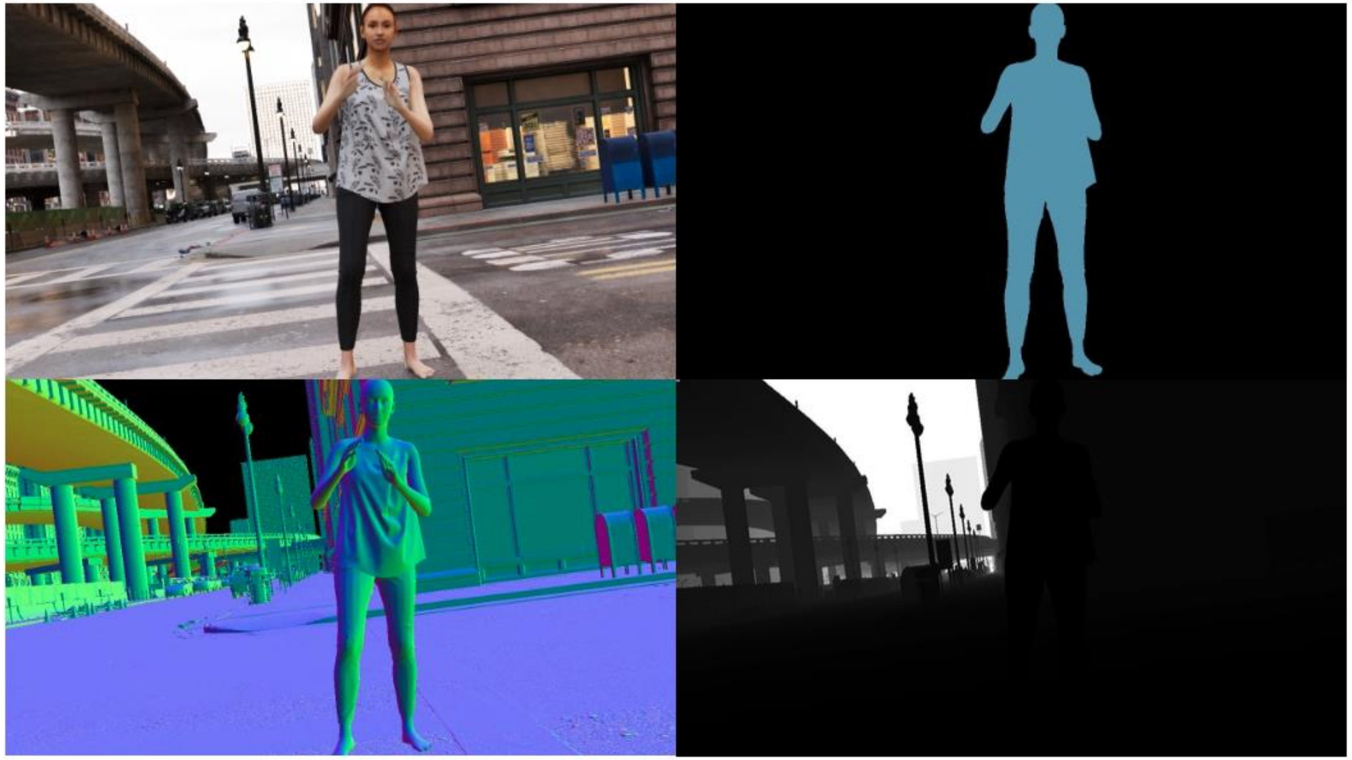
Camera
trajectory,

5 keyframes

Point
Trajectories

Synthetic Videos – Light Condition

Synthetic Videos – Depth/Normal



Synthetic Videos – Statistics

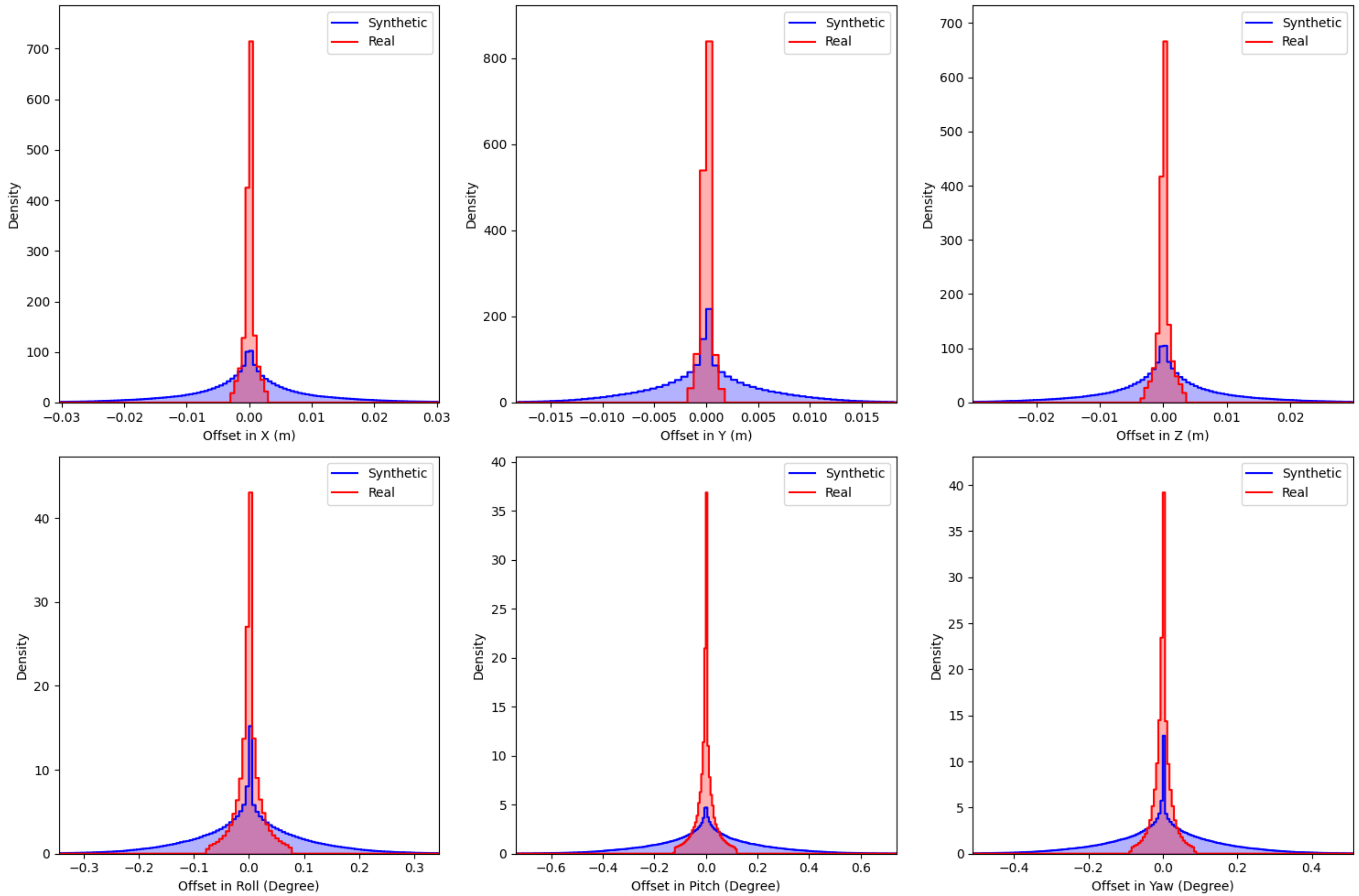
Table 1: Comparison of our Internet and synthetic data size with existing datasets.

Dataset	Clips	Frames	Resolution	Camera Pose	Human Pose
TikTok [29]	340	93k	604×1080	Static	Fitting
UBC-Fashion [80]	500	192k	720 × 964	Static	Fitting
IDEA-400 [37]	12k	2.5M	720P	Static	Fitting
Bedlam [10]	10k	1.5M	720P	Ground Truth	Ground Truth
Ours Real	20k	10M	1080P	Fitting	Fitting
Ours Synthetic (SMPL-X)	50k	8M	720P	Ground Truth	Ground Truth
Ours Synthetic (Anime)	25k	2M	1080P	Ground Truth	Ground Truth

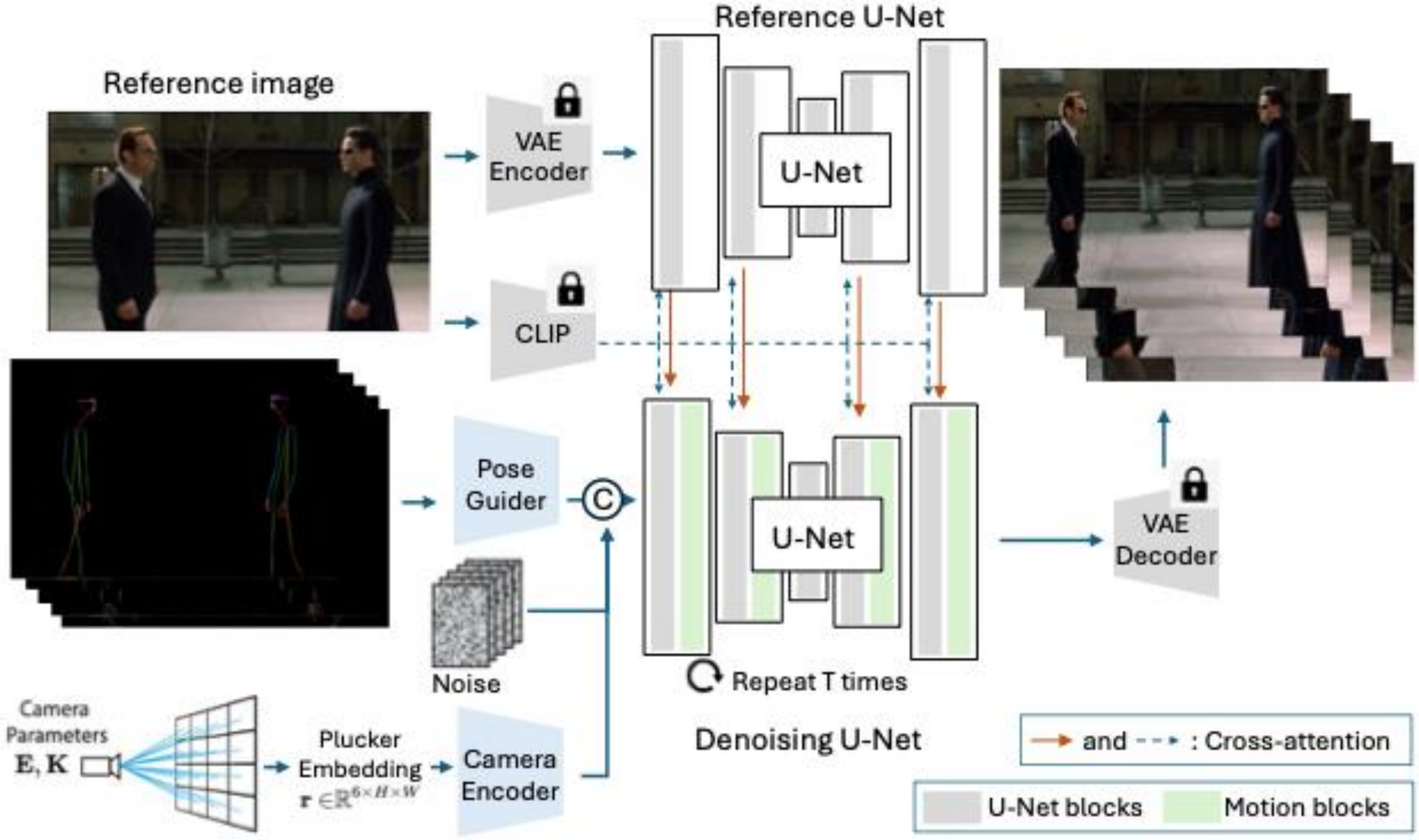
Table 2: Statistics of the diversity of appearance, motion and scene in HumanVid.

Dataset Split	#Subject	#Motion	#Scene	Avg. Clip Length
Internet videos	24,012	24,012	19,688 (= #video)	16.65s
Synthetic (SMPL-X)	271 (body shapes) × 100 (skin textures) × 1,691 (clothings)	2,311	100 (HDRIs) + 587 (3D scenes)	6.34s
Synthetic (Anime)	10K (anime assets)	40	100 (HDRIs) + 93 (3D scenes)	3.2s

Synthetic Videos – Camera Movement Stat.



Baseline Method - CamAnimate



Quantitative Results

TikTok Test Set	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	FID ↓
Animate Anyone [28]	0.752	16.971	0.288	935.6	52.26
Magic-animate [74]	0.748	17.890	0.270	876.0	56.84
Champ [88]	**0.778**	18.434	0.267	736.1	50.76
Ours	**0.778**	**18.762**	**0.247**	**691.8**	**41.35**

UBC-Fashion Test	SSIM↑	PSNR↑	LPIPS↓	FVD↓	FID↓
Magic-animate [74]†	0.602	6.663	0.552	1583.9	118.76
Animate Anyone [28]	0.914	23.163	0.069	345.4	33.77
Champ [88]	0.922	25.269	0.057	269.4	**27.35**
Ours	**0.929**	**25.921**	**0.049**	**256.6**	29.30

Static camera results

Landscape	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	FID ↓
Animate Anyone [28]	0.602	16.108	0.368	1248.4	97.74
Magic-animate [74]	0.543	15.567	0.361	1325.2	109.33
Champ [88]	0.653	15.028	0.426	1985.2	100.59
Ours (1× batch size)	0.641	18.008	0.309	960.1	77.73
Ours (4× batch size)	**0.672**	**19.534**	**0.275**	**732.7**	**46.06**

Portrait	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	FID ↓
Animate Anyone [28]	0.613	15.514	0.379	1254.1	88.70
Magic-animate [74]	0.621	16.091	0.341	1418.8	123.94
Champ [88]	0.669	16.021	0.360	1316.9	84.59
Ours (1× batch size)	0.675	18.081	0.309	816.5	75.67
Ours (4× batch size)	**0.678**	**18.939**	**0.303**	**792.2**	**54.02**

Moving camera results

Table 5: User study on videos of Tiktok dataset and our test set.

Method	Average Score	Top-1 Preference
Animate Anyone [28]	0.171	0.10
Magic-animate [74]	0.133	0.03
Champ [88]	0.256	0.14
Ours	**0.440**	**0.73**

Table 6: Comparison with original Animate Anyone trained without camera condition.

TikTok Test Set	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	FID ↓
Animate Anyone [28]	0.658	15.954	0.337	1133.1	53.65
Ours	**0.778**	**18.762**	**0.247**	**691.8**	**41.35**

Table 7: Comparison of training strategies on different data parts.

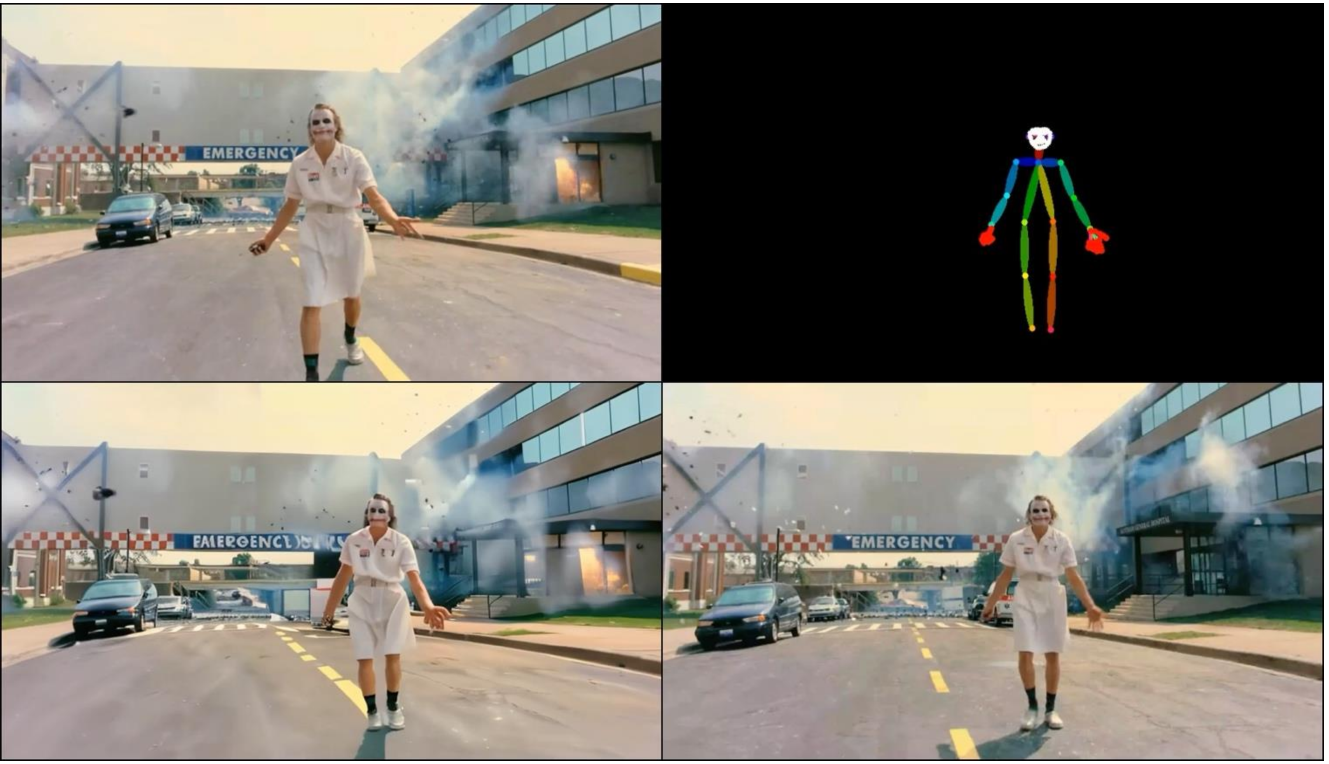
TikTok Test Set	Stage 1 w/ Syn. Data	Stage 2 w/ Syn. Data	SSIM↑	PSNR↑	LPIPS↓	FVD↓	FID↓
Variant 1	×	×	0.677	15.957	0.333	1066.9	53.08
Variant 2	✓	✓	0.734	17.339	0.287	980.3	56.32
Ours	×	✓	**0.778**	**18.762**	**0.247**	**691.8**	**41.35**

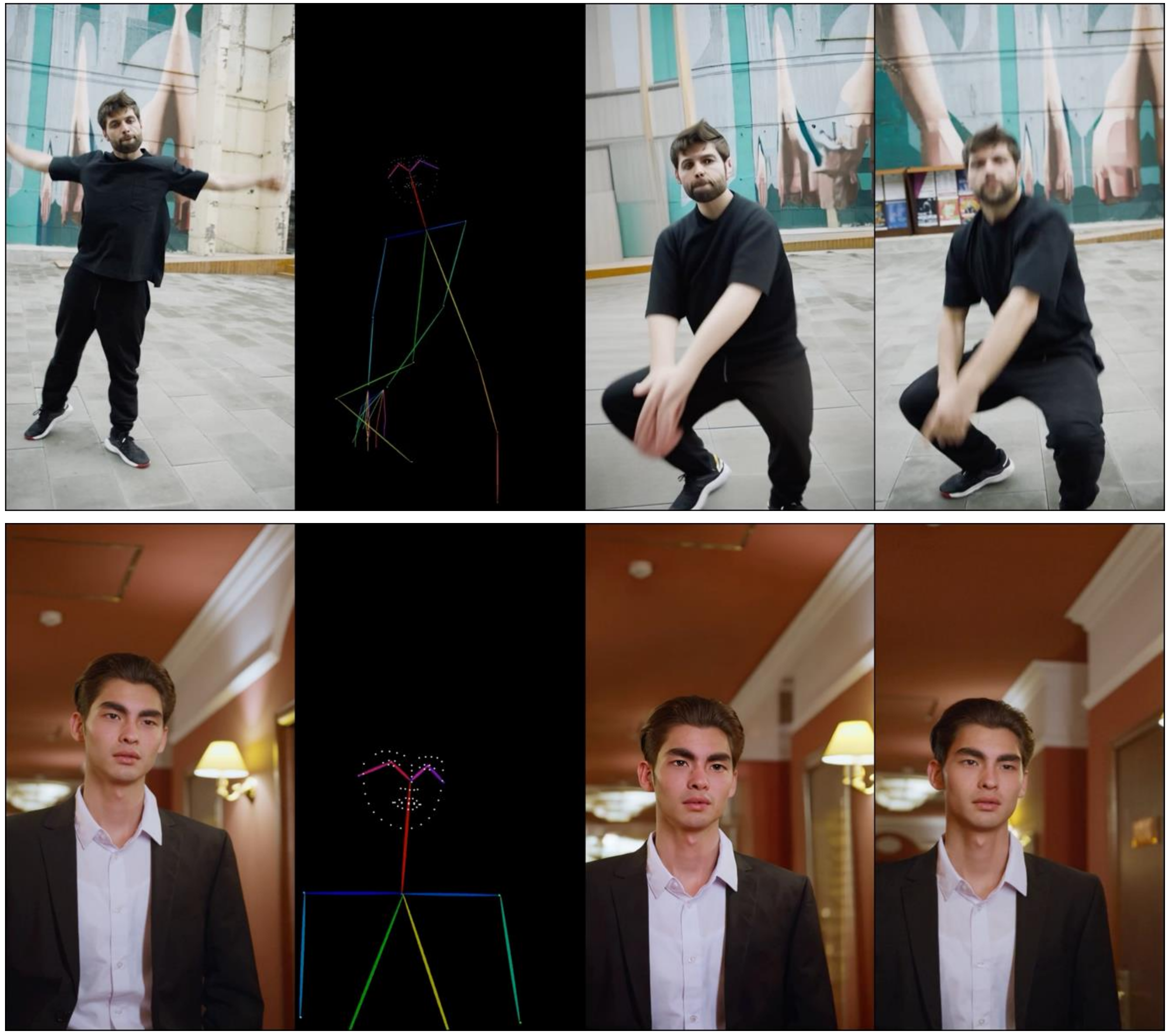
Qualitative Results

Figure 5: **Qualitative comparisons** with previous SOTA methods on the test set.

Figure 6: **Qualitative results** on in-the-wild videos with realistic camera movements.

Qualitative Results





Please visit our project page for more demos!

https://humanvid.github.io/

Thank you!