IDOL: Unified Dual-Modal Latent Diffusion for Human-Centric Joint Video-Depth Generation

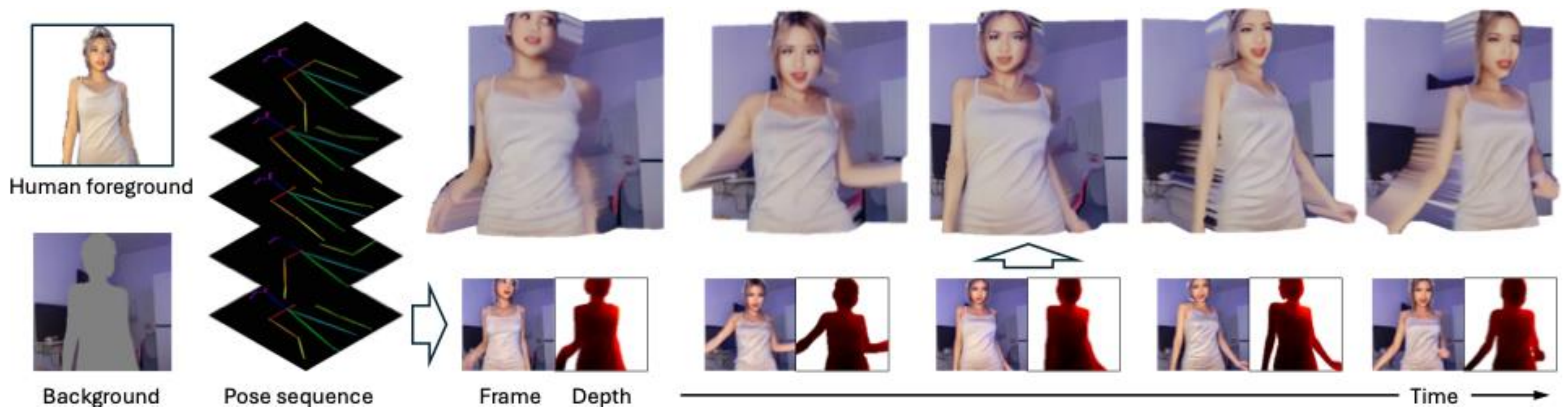
Yuanhao Zhai[1], Kevin Lin[2], Linjie Li[2], Chung-Ching Lin[2], Jianfeng Wang[2], Zhengyuan Yang[2], David Doermann[1], Junsong Yuan[1], Zicheng Liu[2], Lijuan Wang[2]

[1]State university of New York at Buffalo, [2]Microsoft

On the job market

yhzhai.github.io/idol

Motivation

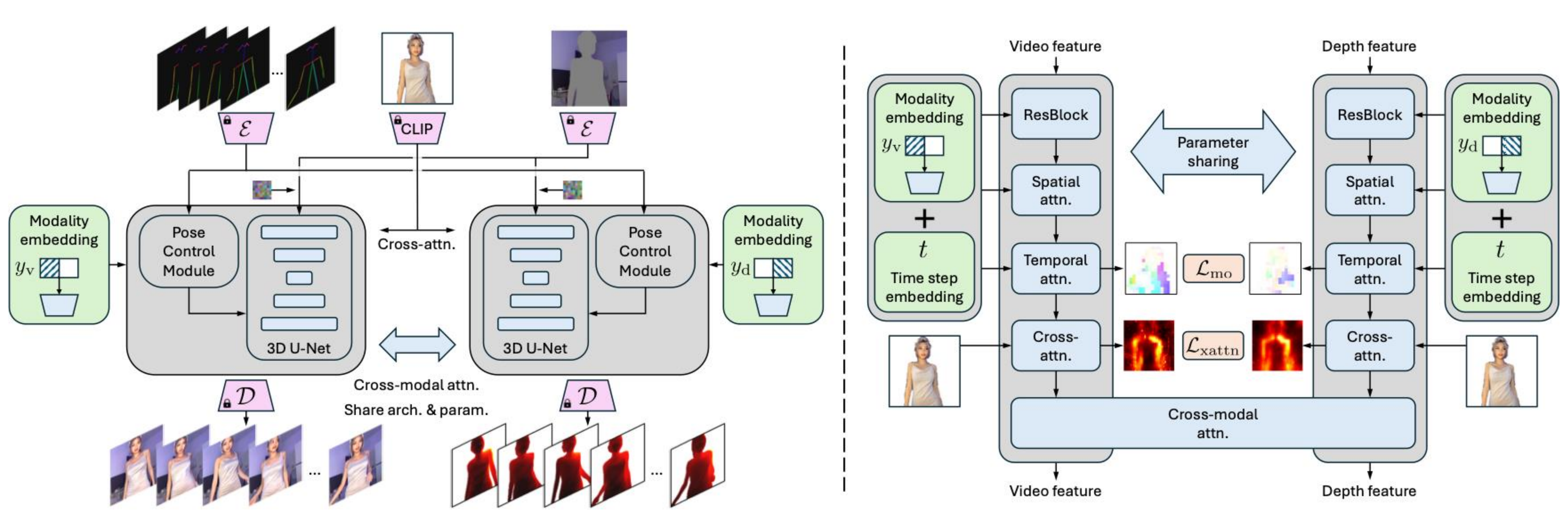


Given a human foreground image, an arbitrary background image, and a defined pose sequence, our IDOL generates **high-fidelity video** and the **corresponding depth maps**, which can be rendered as realistic 2.5D video.

Method

- Challenges
  - ❑ Video and depth are distinct modalities
  - ❑ Most generative methods focus on RGB contents
- Insight
  - ❑ Reframe depth generation as stylized image generation
  - ❑ Convert depth maps to colored heatmaps (RGB)

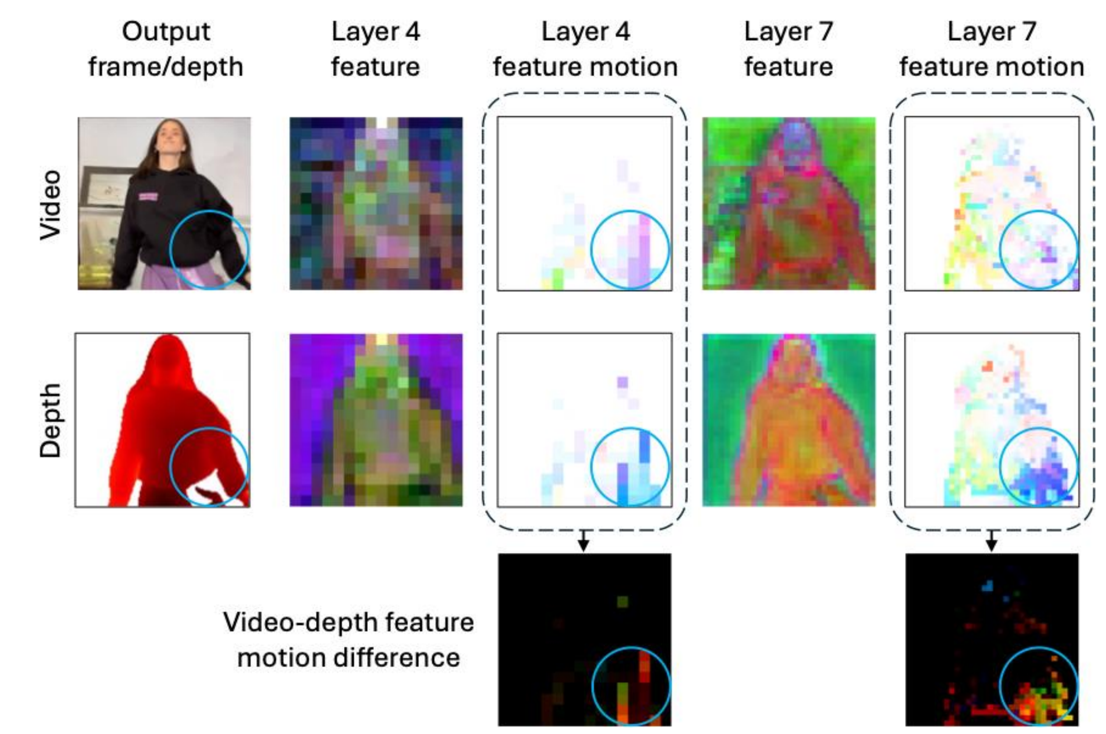
Unified dual-modal U-Net



- Video LDM backbone
  - ❑ 3D U-Net for video and depth denoising
  - ❑ Pose control via ControlNet
- Sharing U-Net for joint video-depth denoising
  - ❑ Parameter-efficient
  - ❑ Learnable modality embedding for denoising modality control
  - ❑ Implicit structural information learning
- Cross-modal attention
  - ❑ Explicit cross-modal information exchange
- Joint video-depth denoising objective

Method (cont'd)

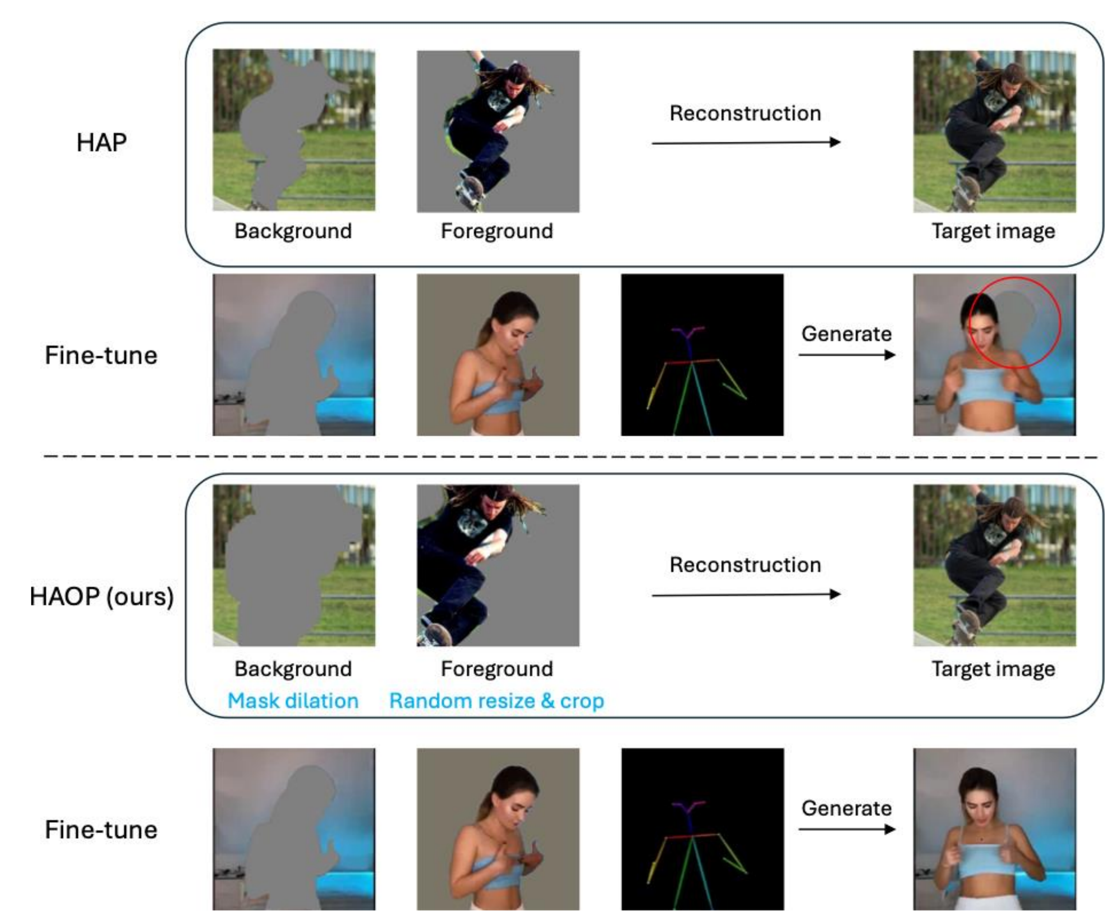
Learning video-depth consistency



Video and depth feature maps and motion fields visualization

- Video-depth inconsistency (blue circles) stem from mismatched feature motions
- Introduce motion consistency loss to align the feature motions
  - ❑ Minimize MSE between video and depth feature cost volume

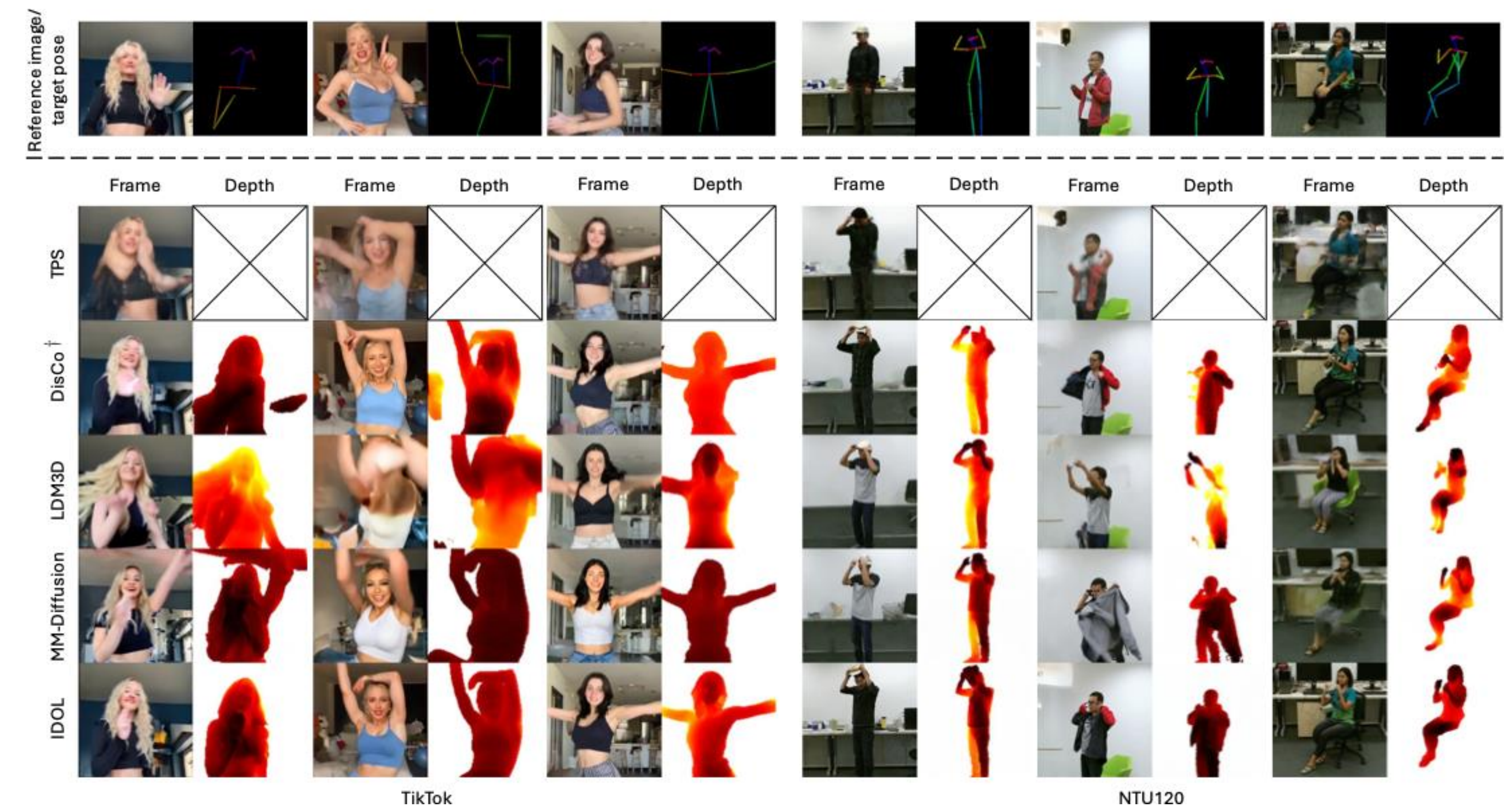
Human attribute outpainting pre-training



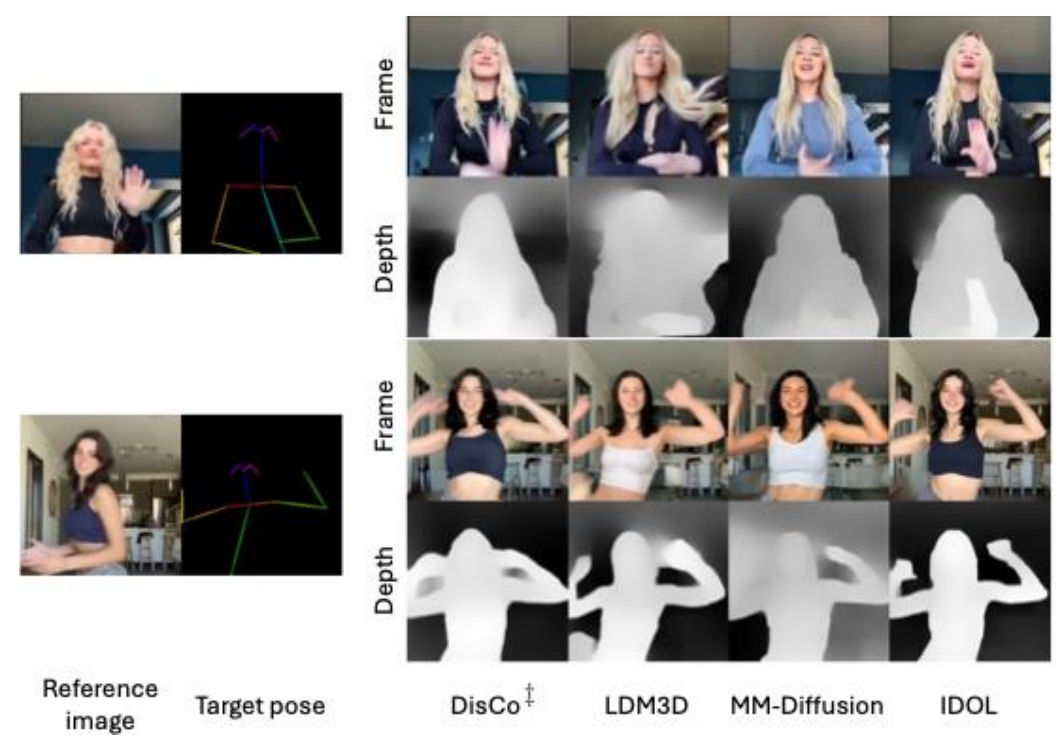
Comparison of HAP vs. HAOP

- HAP can produce background masks when the target post shifts (red circles)
- HAOP addresses this by filling in the masks

Experiments

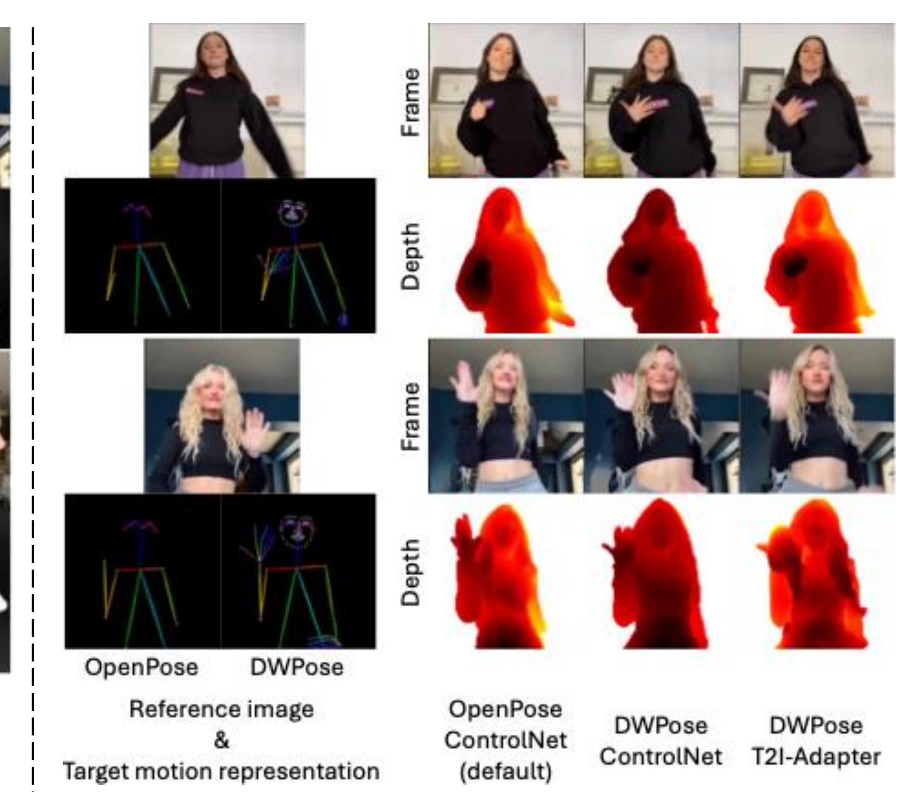


Better identity preservation & depth alignment



Generalization to different depth maps

- Human-centric depth and whole-frame depth
- Gray-scale and colored depth images



Generalization to different motion representations and pose control modules

Method	Motion control	TikTok Video FID-FVD↓	TikTok Video FVD↓	TikTok Depth L2↓	TikTok Image FID↓	NTU120 Video FID-FVD↓	NTU120 Video FVD↓	NTU120 Depth L2↓	NTU120 Image FID↓
FOMM [62]	Target video	38.36	404.31	-	85.03	40.34	1439.50	-	80.29
MRAA [63]	Target video	24.11	306.49	-	54.47	58.19	1441.79	-	97.07
TPS [79]	Target video	29.20	337.79	-	53.78	37.42	1339.86	-	<u>61.75</u>
DreamPose [30]	DensePose [19]	52.62	614.07	-	75.08	80.11	791.25	-	116.23
DisCo [68]	OpenPose [7]	<u>20.75</u>	<u>257.90</u>	0.0975†	<u>39.02</u>	<u>26.21</u>	<u>458.92</u>	<u>0.0371</u>†	68.53
LDM3D [64]	OpenPose [7]	45.30	553.03	0.0637	69.36	71.11	587.84	0.0650	120.74
MM-Diffusion [58]	OpenPose [7]	48.92	771.32	<u>0.0367</u>	68.47	58.44	504.05	0.0404	102.77
IDOL	OpenPose [7]	**17.86**	**223.69**	**0.0336**	**36.04**	**20.23**	**314.82**	**0.0317**	**50.70**

State-of-the-art video and depth generation performance