Towards Multiple Character Image Animation Through Enhancing Implicit Decoupling

Jingyun Xue[1,2*], Hongfa Wang[2,3*], Qi Tian[2*], Yue Ma[2,4], Andong Wang[2], Zhiyuan Zhao[2], Shaobo Min[2], Wenzhe Zhao[2], Kaihao Zhang[5], Heung-Yeung Shum[3,4], Wei Liu[2], Mengyang Liu[2†], Wenhan Luo[4‡]
[1]Sun Yat-sen University, [2]Tencent Hunyuan, [3]Tsinghua Univerisity, [4]HKUST, [5]Harbin Institute of Technology
*These authors contributed equally to this research, †Project leader, ‡Corresponding author

Overview

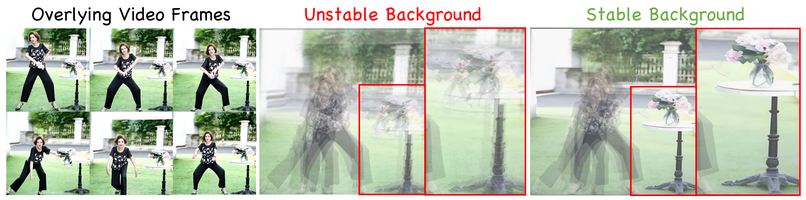
Motivation

Challenges of multiple character animation:

- Due to the prevalent camera shake in datasets, existing models are affected by the noise, leading to abrupt changes, flickering, and artifacts in the generated background.
- when generating animated videos with multiple characters, existing methods tend to produce chaotic character identities and erroneous occluded body parts.

Idea

- ✓ We enables the model to implicitly learn to decouple the background and characters.
- ✓ We carefully design guidance to enhance the implicit decoupling capability of the model.



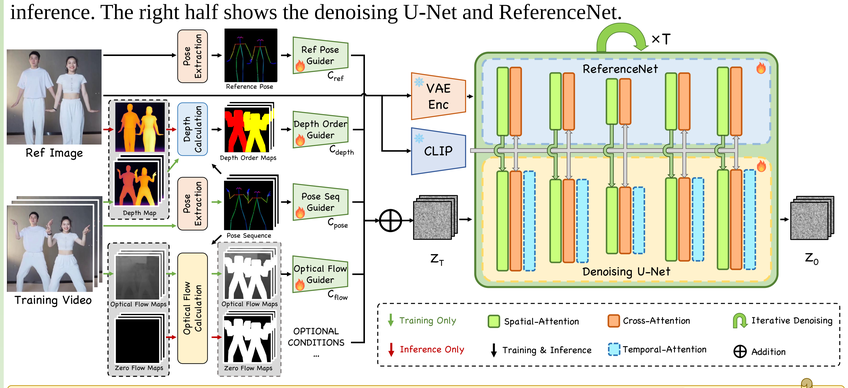
Contributions

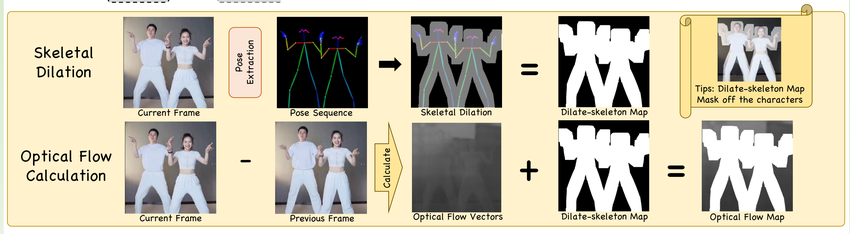
- ➢ We propose a multi-condition guided framework with multiple guiders for multiple-character image animation. Our framework effectively handles the multiple-character image animation.
- ➢ We enhance the implicit decoupling capability of the model. Technically, the *optical flow guider* decouples the background momentum to ensure background stability; the *depth order guider* provides multiple character positional information to address the occlusion among body parts; and *reference pose guider* introduces the source pose to better align the character with the target pose.
- ➢ We address the lack of a benchmark in multiple character image animation by introducing *Multi-Character Bench*, a new dataset containing about 4,000 frames for comprehensive evaluation.

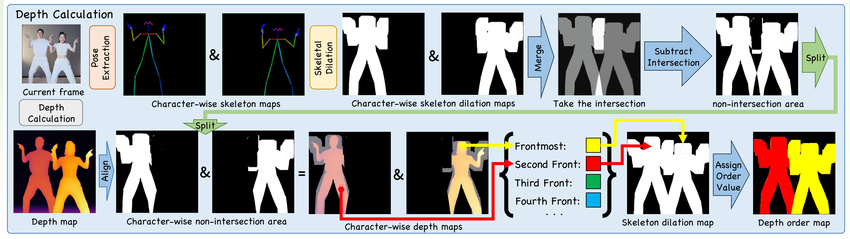
Method

Overview

The overview of the proposed framework. The left half illustrates the data flow of the multiple condition guiders, with green and black arrows denoting training data flow, and red and black arrows indicating inference data flow. The gray box represents different inputs for training and inference. The right half shows the denoising U-Net and ReferenceNet.







Experiments & Results

Quantitative comparisons

Table 1: Quantitative comparison on Tiktok dataset. The best and second-best results are indicated in red and blue respectively. AnimateAnyone† is trained on our noisy dataset. Ours* is trained on the Tiktok training set.

Method	FID↓	SSIM↑	PSNR↑	LPIPS↓	L1↓	FID-VID↓	FVD↓
MRAA (Siarohin et al., 2021)	54.47	0.672	29.39	0.296	3.21E-04	66.36	284.82
TPSMM (Zhao & Zhang, 2022)	53.78	0.673	29.18	0.299	3.23E-04	72.55	306.17
DreamPose (Karras et al., 2023)	79.46	0.509	28.04	0.450	6.91E-04	80.51	551.56
DisCo (Wang et al., 2023)	51.29	0.699	28.70	0.333	1.10E-04	61.41	379.56
DisCo+ (Wang et al., 2023)	48.29	0.713	28.78	0.320	1.03E-04	52.56	334.67
MagicAnimate (Xu et al., 2023)	32.09	0.714	29.16	0.239	3.13E-04	21.75	179.07
MagicPose (Chang et al., 2023)	25.50	0.752	29.53	0.292	0.81E-04	46.30	216.01
AnimateAnyone (Hu et al., 2023)	-	0.718	29.56	0.285	-	-	171.90
AnimateAnyone†	54.42	0.685	29.01	0.316	1.06E-04	47.93	236.28
Ours*	29.15	0.735	29.61	0.287	0.79E-04	35.28	153.47
Ours	27.70	0.760	29.70	0.272	0.73E-04	14.30	117.81

Table 2: Quantitative comparison on TED-talks dataset. The best and the second-best results are indicated in red and blue respectively. AnimateAnyone† is trained on our noisy dataset.

Method	FID↓	SSIM↑	PSNR↑	LPIPS↓	L1↓	FID-VID↓	FVD↓
MRAA (Siarohin et al., 2021)	50.36	0.762	31.90	0.266	0.50E-04	82.79	493.02
TPSMM (Zhao & Zhang, 2022)	23.71	0.771	32.30	0.252	0.49E-04	32.12	260.67
DisCo (Wang et al., 2023)	75.48	0.575	27.99	0.309	1.21E-04	66.18	393.04
DisCo+ (Wang et al., 2023)	63.28	0.596	28.12	0.300	1.11E-04	55.81	343.20
MagicAnimate (Xu et al., 2023)	41.58	0.529	28.28	0.310	1.73E-04	33.61	223.54
MagicPose (Chang et al., 2023)	23.39	0.723	30.08	0.236	0.81E-04	27.53	214.23
Moore-AnimateAnyone	25.93	0.710	30.99	0.310	0.46E-04	41.20	262.49
AnimateAnyone†	47.68	0.691	29.59	0.283	1.15E-04	30.07	241.76
Ours	18.21	0.779	30.88	0.198	0.46E-04	10.24	81.73

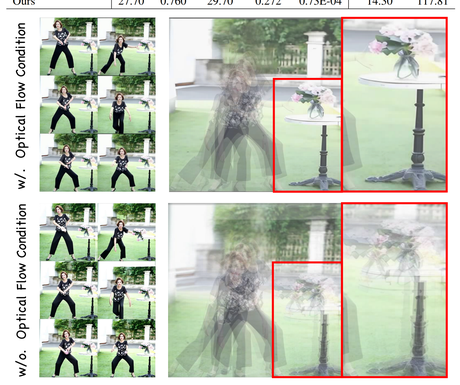
Table 3: Quantitative comparison on *Multi-Character* Bench. The best and the second-best results are indicated in red and blue respectively. AnimateAnyone† is trained on our noisy dataset.

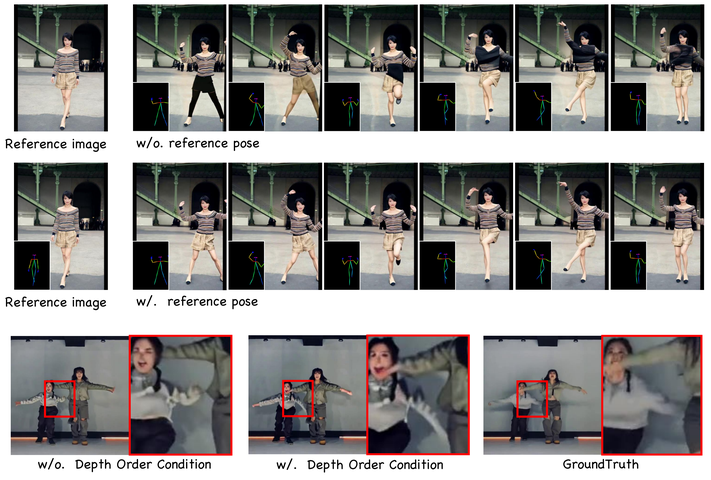
Method	FID↓	SSIM↑	PSNR↑	LPIPS↓	L1↓	FID-VID↓	FVD↓
DisCo (Wang et al., 2023)	77.61	0.793	29.65	0.239	7.64E-05	104.57	1367.47
DisCo+ (Wang et al., 2023)	73.21	0.799	29.66	0.234	7.33E-05	92.26	1303.08
MagicAnime (Xu et al., 2023)	40.02	0.819	29.01	0.183	6.28E-05	19.42	223.82
MagicPose (Chang et al., 2023)	31.06	0.806	31.81	0.217	4.41E-05	30.95	312.65
Moore-AnimateAnyone	33.04	0.795	31.44	0.213	5.02E-05	22.98	272.98
AnimateAnyone†	35.59	0.796	31.10	0.208	4.87E-05	22.74	236.48
Ours	26.95	0.830	31.86	0.173	4.01E-05	14.56	142.76

Ablation Study

Table 4: Quantitative ablation results on TikTok Dataset.

Method	FID↓	SSIM↑	PSNR↑	LPIPS↓	L1↓	FID-VID↓	FVD↓
w/o. All Conditions	54.42	0.685	29.01	0.316	1.06E-04	47.93	236.28
w/o. Optical Flow	52.94	0.710	28.21	0.318	0.96E-04	34.35	178.48
w/o. Ref. Pose	34.74	0.740	29.13	0.285	0.82E-04	19.58	139.32
w/o. Depth Order	27.43	0.754	29.98	0.270	0.79E-04	14.50	119.26
Ours	27.70	0.760	29.70	0.272	0.73E-04	14.30	117.81





Qualitative comparisons