Supplementary Materials: Follow-Your-Pose v2: Multiple-Condition Guided Character Image Animation for Stable Pose Control

Anonymous Authors

This appendix includes our supplementary materials as follows:

1 METHODOLOGY SUPPLEMENT

1.1 Further Details on Depth Guider

Algorithm 1: Pseudocode for depth map mask extraction

Input: Given a training video v of length N, where the i_{th} frame is denoted as v_i, and suppose that J characters on v_i. The skeleton extraction network is denoted as f_{s}. The expansion network is denoted as f_{e}. The depth extraction network is denoted as f_{d}. The average depth sorting (ascending order) operator is denoted as f_{sort}. The value assigned to r_j based on depth ranking is denoted as L_{r_j}. The depth guider is denoted as g_{dp}.

Initialize $c_{\text{depth},1,0}, \ldots, c_{\text{depth},N,0} = \mathbf{0}$.
Initialize an array C_R.
for $i = 1$ **to** N **do**
  $a_{i,1}, \ldots, a_{i,J} = f_{\text{e}}\left(f_{\text{s}}\left(v_i\right)\right)$
  for $j = 1$ **to** J **do**
    $m_{i,j} = a_{i,j} - \left(1 - \bigcup_{j \in \{1,\ldots,J\}}\left(a_{i,j}\right)\right)$
  $r_1, \ldots, r_J = f_{\text{sort}}\left(m_{i,1} \odot f_{\text{d}}\left(v_i\right), \ldots, m_{i,J} \odot f_{\text{d}}\left(v_i\right)\right)$
  for $r_j = r_1$ **to** r_J **do**
    $c_{\text{depth},i,r_j} = m_{i,r_j} \odot L_{r_j} + \left(\left(1 - m_{i,r_j}\right) \odot c_{\text{depth},i,(r_j-1)}\right)$
  $C_R[i] \leftarrow c_{\text{depth},i,r_J}$
$c_{\text{depth}} = g_{\text{dp}}\left(C_R\right)$
return c_{depth}

1.2 Baseline implementations

BC. Each agent policy π_i w.r.t. parameters θ_i is optimized by the following loss

$$\mathcal{L}_{\text{BC}}(\theta_i) = \mathbb{E}_{\tau_i, a_i \sim \mathbb{B}}[-\log(\pi_i(a_i \mid \tau_i))]. \tag{1}$$

1.3 Long Video Inference

We employ the overlap method for long video inference to maintain the consistency of long video. As illustrated in Fig. 1, We divide the pose sequence into multiple smaller segments for inference, with overlapping parts between adjacent segments. For the overlapping parts of two segments, we perform addition and averaging to generate temporal smoothing between the two segments. In this work, we perform inference on every 16 frames with a stride of 8 frames, and then stitch them together using an overlap of 8 frames.

1.4 Further Details on Skeletal Dilation

We use skeletal dilation as a mask to cover the character region in the “Optical Flow Maps” and “Depth Maps”. The purpose of masking off the character regions in “Optical Flow Maps” is to separate character motion from background motion. Therefore, the model can differentiate character motion that requires learning from background motion that needs to be removed. Additionally, the skeletal dilation of different characters is assigned different values in “Depth Maps” based on positions, which are utilized to differentiate various character regions. In summary, the skeletal dilation represents the character region in “Optical Flow Maps” and “Depth Maps”.

As illustrated in Fig 2, skeletal dilation is computed from pose sequence, one of the inputs during inference. We do not propose employing more precise regional information of characters during inference, such as semantic segmentation or 3D modeling graphs. Because they impose strong spatial constraints on the motion of the overall region of the character, inconsistent body shape or clothing in the inference image will lead to a decrease in effectiveness. This is primarily why we utilize pose sequence and skeletal dilation. Maybe one concern is that skeletal dilation often fails to cover the character region comprehensively. First, skeletal dilation represents only rough but not exact character regions, both during training and inference. This is aligned during both training and inference. Second, in the notable work of regional image animation [5], it is confirmed that models animate the objects represented by the mask region, rather than animating the mask region itself.

1.5 Data Flow

Here we will describe it in detail. In training, we use “Reference Image” and “Training Video” as input before the data processing. After the data processing, we obtain four conditions: “Reference Pose”, “Character Depth Maps”, “Pose Sequence”, and “Optical Flow

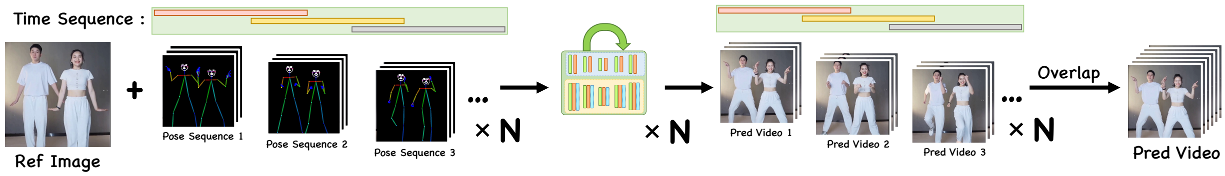

Figure 1: The pipeline for inferring long videos.

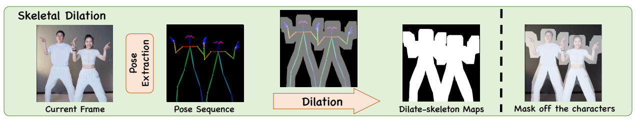

Figure 2: The pipeline of skeletal dilation.

Maps" from the input "Reference Image" and the "Training Video". Specifically, in data processing, we extract "Reference Pose" from "Reference Image"; extract "Pose Sequence" from "Training Video"; obtain "Depth Maps" from both the "Training Video" and "Dilate-skeleton Maps"; obtain "Optical Flow Maps" from both the "Training Video" and "Dilate-skeleton Maps". The detailed data processing can be found in the methodology section of the paper.

The main difference between inference and training is that using "Reference Image", and "Pose Sequence" instead of "Training Video", as input before the data processing, but getting the same four conditions after data processing. During the data processing, we directly utilize the input "Pose Sequence"; obtain "Depth Maps" from "Reference Image" and copy it into n frames instead of obtaining from "Training Video"; and obtain "Optical Flow Maps" from "Zero Tensor" because we want to generate videos with a background optical flow of 0, i.e., static background.

2 EXPERIMENTAL SUPPLEMENT

2.1 More Implementations

In the data processing, we utilize the DWPose [12] to extract pose sequence from videos, and PWC-Net [8] from the open-source toolbox MMFlow [2] to calculate optical flow vectors. Additionally, we use the Depth Anything [11] to extract depth maps from videos.

When conducting long video inference, we perform inference on every 16 frames with a stride of 8 frames, and then stitch them together using an overlap of 8 frames. Besides, We resize and center-crop the "Reference Image" and "Pose Sequence" to a uniform resolution of 896×640 pixels (512×512 pixels in comparative experiments). We apply DDIM [7] sampler for 50 denoising steps, with classifier-free guidance [3] scale of 1.5.

2.2 Unified Standard for Comparative Experiments

We noticed in the comparative experiment that not all approaches adhere to a uniform inference size and other inference details. As depicted in Table 1, there are primarily three inconsistent standards that may affect fair comparisons. Below, I will use Method

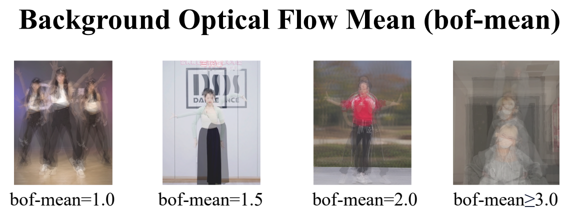

Figure 3: Videos with different background optical flow means.

Disco+ [9] as an example to illustrate how inconsistent standards affect the fair comparison of metrics, as shown in Table 2. Firstly, as shown in the comparison between the first and second rows of Table 2, different inference sizes result in different metrics. Table 1 shows the inconsistent inference sizes of each method. Secondly, some works resize without center-crop, which can result in significant differences in the test set, as illustrated in Table 1. Besides, compared to other methods, Disco has a bug in measurement, resulting in fewer video segments being sampled when calculating FID-VID and FVD. This will lead to a decrease in FID-VID and FVD of Disco, as shown in Table 2.

As methods with different standards result in different metrics, potentially leading to unfair comparisons, we standardize the inference sizes by center-cropping and resizing to 512×512, and we rectify the bug in the measurement of disco. Under this unified standard, we reevaluate methods that do not conform to this inference size and directly reference relevant statistical data from the original literature of methods that do comply. Most previous works directly reference inconsistent statistical data from other works for comparison, resulting in unfair comparisons. We are the first to conduct comparative work under a unified standard, and this is one of our notable contributions.

3 DATASET SUPPLEMENT

3.1 Detailed Information on Training Dataset

We collect 4,017 character videos totaling 2,013,628 frames as our training set. The data come from public videos on TikTok, YouTube, and other websites. The detailed composition of the training dataset is shown in Table 3.

Table 1: Inconsistent standards across all methods.

Method	Inference Size	Center-Crop	Uninterrupted Frames
MRAA [6]	384×384	×	✓
TPSMM [13]	384×384	×	✓
DreamPose [4]	512×640	×	✓
DisCo[9]	256×256	✓	×
MagicAnime[10]	512×512	✓	✓
MagicPose[1]	512×512	×	✓

Table 2: The ablation experiment on the inference standards of Disco+.

Inconsistent Standards	FID↓	SSIM↑	PSNR↑	LPIPS↓	L1↓	FID-VID↓	FVD↓
256×256, w/o uninterrupted frames	28.31	0.674	29.15	0.285	3.69E-04	55.17	267.75
512×512, w/o uninterrupted frames	48.29	0.713	28.78	0.320	1.03E-04	52.56	334.67
512×512, w/. uninterrupted frames	48.29	0.713	28.78	0.320	1.03E-04	47.73	312.49

Table 3: The detailed composition of the training dataset.

Source	Videos	Frames	Proportion
Tiktok	2,493	1,379,449	68.5%
YouTube	938	435,293	21.6%
Kuaishou	424	115,101	5.7%
Bilibili	162	83,785	4.2%

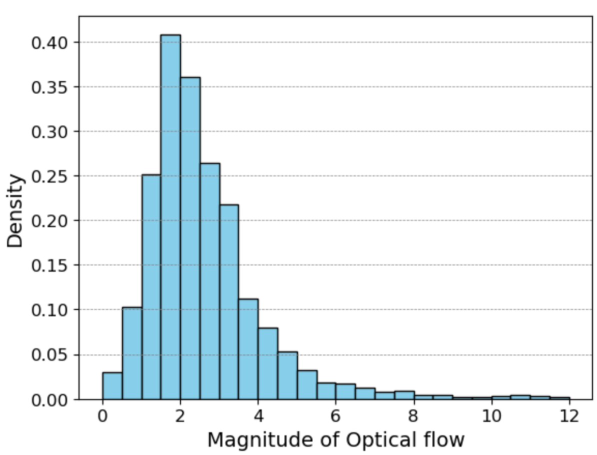

Figure 4: Histogram of the distribution of background optical flow mean in training set.

3.2 Analysis of Noisy Training Dataset

In our training dataset, there exists a substantial amount of noisy data, with the significant proportion is data with unstable backgrounds. We calculate the optical flow map of the background by using skeletal dilation map as mask to remove the character regions. After averaging the background optical flow map for each frame of a video, we obtain the background optical flow mean for each video. Fig. 3 shows videos with different background optical flow means. We can observe that only when the background optical flow mean is less than 1, the motion of the background of video is imperceptible to the human eye. Fig. 4 illustrates the distribution of the background optical flow mean in the training set. We can find that only 12% of the training set videos have a background optical flow mean less than 1. This indicates that at least 88% of the background unstable noise data in our training set is present, and it is necessary to incorporating background optical flow maps into the network for training.

3.3 Detailed Information on Multi-Character Dataset

We collect 20 multiple-character dancing videos, totally 3917 frames, from social media, named *Multi-Character*. Table 4 show the detailed sources of *Multi-Character*.

4 LIMITATION

In this work, we are dedicated to addressing the issues of background stability and character overlap. However, there are still several problems in pose-controllable character video generation that we have not resolved: First, similar to most approaches, our model struggles to generate highly refined facial and hand details. Second, our model also struggles to generate substantial swaying of long skirts, Hanfu, or other large-area clothing very well. Third, our model also faces challenges in handling complex multiple-character scenarios, such as those involving four or more characters, or extensive swapping of positions among characters.

5 ADDITIONAL VISUALIZATIONS

See the video in the supplementary materials, as well as Fig. 5 to Fig. 9 provided last in this document.

Table 4: The source of Multi-character dataset.

Video Name	Url	Timestamp
Daovm348PQQ_0	https://www.youtube.com/watch?v=Daovm348PQQ	00:10−00:14
Daovm348PQQ_1	https://www.youtube.com/watch?v=Daovm348PQQ	00:16−00:25
Daovm348PQQ_2	https://www.youtube.com/watch?v=Daovm348PQQ	00:47−00:53
Daovm348PQQ_3	https://www.youtube.com/watch?v=Daovm348PQQ	01:11−01:15
Daovm348PQQ_4	https://www.youtube.com/watch?v=Daovm348PQQ	01:16−01:21
Daovm348PQQ_5	https://www.youtube.com/watch?v=Daovm348PQQ	01:22−01:25
Daovm348PQQ_6	https://www.youtube.com/watch?v=Daovm348PQQ	02:02−02:09
Daovm348PQQ_7	https://www.youtube.com/watch?v=Daovm348PQQ	02:11−02:15
Daovm348PQQ_8	https://www.youtube.com/watch?v=Daovm348PQQ	02:16−02:20
HpFDXGAo25c_0	https://www.youtube.com/watch?v=HpFDXGAo25c	00:20−00:33
HpFDXGAo25c_1	https://www.youtube.com/watch?v=HpFDXGAo25c	00:38−00:43
HpFDXGAo25c_2	https://www.youtube.com/watch?v=HpFDXGAo25c	00:44−00:52
jx_VseYOi5A_0	https://www.youtube.com/watch?v=jx_VseYOi5A	00:32−00:37
jx_VseYOi5A_1	https://www.youtube.com/watch?v=jx_VseYOi5A	00:40−00:53
jx_VseYOi5A_2	https://www.youtube.com/watch?v=jx_VseYOi5A	01:02−01:07
jx_VseYOi5A_3	https://www.youtube.com/watch?v=jx_VseYOi5A	01:08−01:12
ka3BfUsvRqE_0	https://www.youtube.com/watch?v=ka3BfUsvRqE	00:21−00:27
ka3BfUsvRqE_1	https://www.youtube.com/watch?v=ka3BfUsvRqE	00:28−00:33
ka3BfUsvRqE_2	https://www.youtube.com/watch?v=ka3BfUsvRqE	02:56−03:05
ycInNCB8rbA_0	https://www.youtube.com/watch?v=ycInNCB8rbA	00:10−00:15

Figure 5: Additional Visualizations 1.

Figure 6: Additional Visualizations 2.

Figure 7: Additional Visualizations 3.

REFERENCES

[1] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. 2023. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052* (2023).

[2] MMFlow Contributors. 2021. MMFlow: OpenMMLab Optical Flow Toolbox and Benchmark. https://github.com/open-mmlab/mmflow.

[3] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[4] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 22623–22633.

[5] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.

[6] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings*

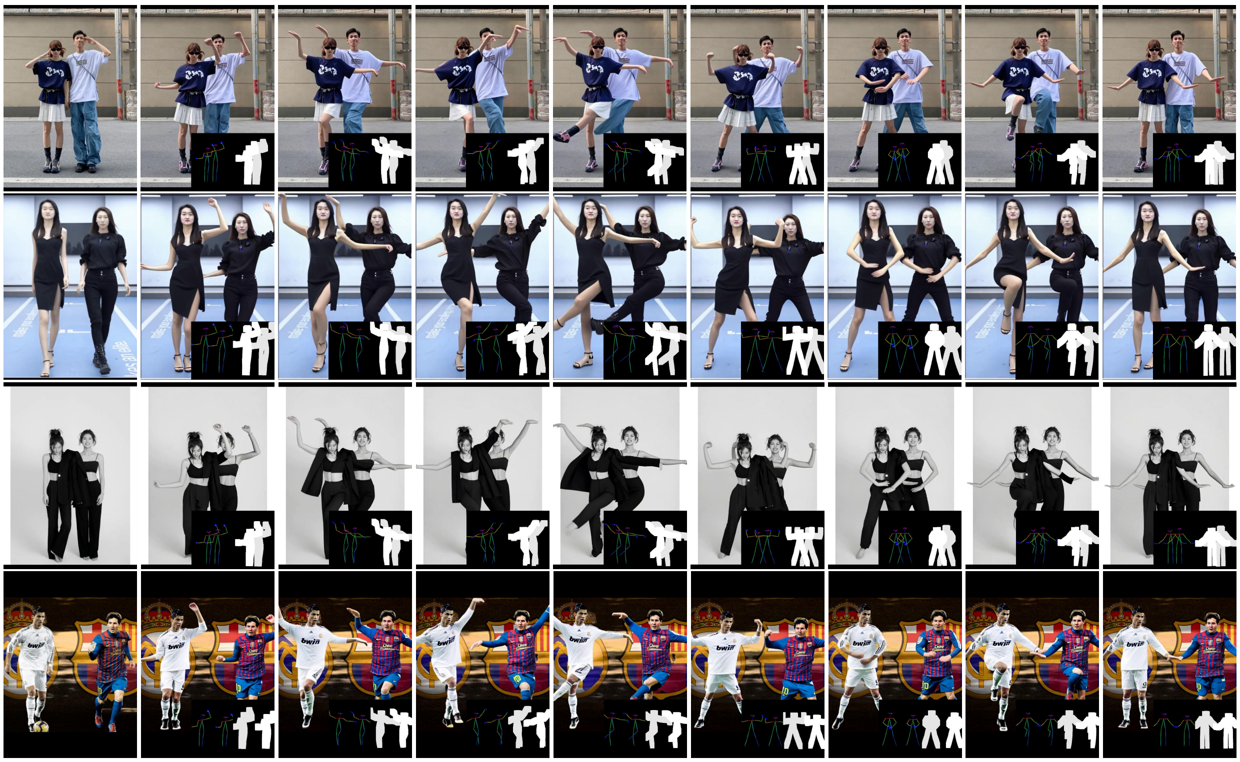

Figure 8: Additional Visualizations 4.

Figure 9: Additional Visualizations 5.

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13653–13662.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[8] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8934–8943.

[9] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints* (2023), arXiv–2307.

[10] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2023. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498* (2023).

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891* (2024).

[12] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.

[13] Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3657–3666.