DiffPerformer: Iterative Learning of Consistent Latent Guidance for Diffusion-based Human Video Generation

Supplementary Material

In this supplementary material, we present extra experiment details (Sec. A), user study (Sec. B), algorithm details (Sec. C), more comparison results (Sec. D), and more analysis on core components (Sec. E).

A. Experiment Details

A.1. Training Settings

We adopt AdamW as the optimizer to finetune the pose-guided diffusion model with batch size of 1. We use the center crop and randomly sample frames to construct the input for each iteration and the frame number in each iteration is 8. To preserve generation priors from the pretrained model while embedding the specific appearance in the temporal domain, we update the parameters of the temporal-related layers and freeze other layers when finetuning the pose-guided diffusion model. Besides, we adopt slide window to generate long videos. The window side is set to 16 and the stride is 8. For each video, we select 80% frames for training and 20% frames for testing.

A.2. Datasets

We conduct experiments on two datasets: (1) the Daily Captured Videos dataset and (2) the TikTok dataset [17].

Daily Captured Videos. To cover comprehensive motion in daily action, we propose the Daily Captured Videos dataset to record dynamic human appearance in daily life. We utilize a mobile phone (Samsung Note 20 Ultra) to record 15 videos of characters performing various actions, including self-occlusion movements such as squatting and rotation. Each video takes about one minute under 30 fps. Besides, we extract the control signals using openpose [4] and densepose [11], respectively. This dataset is a valuable resource for evaluating generation quality, with a particular emphasis on dynamic appearances.

TikTok Dataset. The TikTok dataset is from social media dance videos captured from the TikTok application. It includes more than 300 dance videos that capture a single person from TikTok dance challenge with various types of dances. Each video is 10-15 seconds with diverse dance motions without much blur. Similar to the proposed dataset, we estimate the keypoints and UV coordinates for each frame using openpose [4] and densepose [11], respectively.

B. User Study

We conduct a user study to evaluate the subjective perception of different methods. Specifically, we randomly select 15 videos from the Daily Captured Videos and TikTok datasets and invite 20 participants (10 males and 10 females) to attend the user study. For a fair comparison, the user study is conducted in the same environment (room, display and light). Then, we perform comparisons between the generated videos of all the methods. The results are presented in a random order to avoid subjective bias. For each video, participants are asked to give three separate scores ($1 \sim 5$) in terms of high fidelity, appearance consistency, and pose alignment. As shown in Fig. 10, our method is more preferred by human subjects.

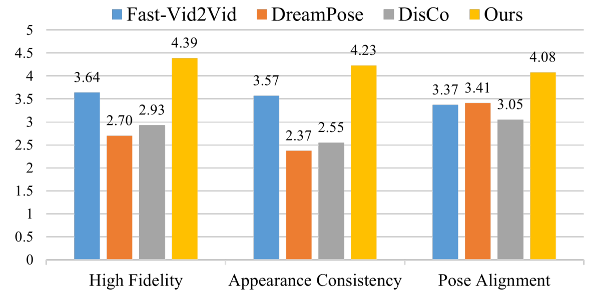


Figure 10. User Study. We obtain the highest scores.

Algorithm 1 Iterative joint optimization of the implicit video representation and the pose-guided diffusion results.

```
Input: implicit video representation model with initiated parameters $\phi$,
    pose sequence $P$, coordinates $(x, y, n)$, learning rate $l$, pose-guided
    diffusion model $\mathcal{D}_p$, finetuned autoencoder $\mathcal{E}$, reconstruction loss $\mathcal{L}_{\text{rec}}$
    and 3D-aware human flow loss $\mathcal{L}_{\text{flow}}$.
$Steps = 10000$
$V_s = \mathcal{C}\left(\mathcal{D}\left(\mathcal{H}\left(x, y, n\right)\right)\right)$
$V_p = \mathcal{E}_{\text{dec}}\left(\mathcal{D}_p\left(\mathcal{E}_{\text{enc}}\left(V_s\right), P\right)\right)$
for $step = 1, \ldots, Steps$ do
    $\mathcal{L}_{\phi}^{\text{total}} = \mathcal{L}_{\text{rec}}\left(V_p, V_s\right) + \mathcal{L}_{\text{flow}}\left(V_s, P\right)$
    $\phi \leftarrow \phi - l \nabla_{\phi} \mathcal{L}_{\phi}^{\text{total}}$
    if $step$ mod $2000 == 0$ then
        $V_p = \mathcal{E}_{\text{dec}}\left(\mathcal{D}_p\left(\mathcal{E}_{\text{enc}}\left(V_s\right), P\right)\right)$
    end if
    $V_s = \mathcal{C}\left(\mathcal{D}\left(\mathcal{H}\left(x, y, n\right)\right)\right)$
end for
Output: $V_p$
```

C. Algorithm Details

In this paper, we design an iterative joint optimization algorithm in Sec. 4 of the paper. The detailed algorithm is shown in Algorithm 1. In practice, the optimization iteration interval can be adjusted according to the video length and pose complexity. For some simple cases, once optimization is able to bring pleasure results.

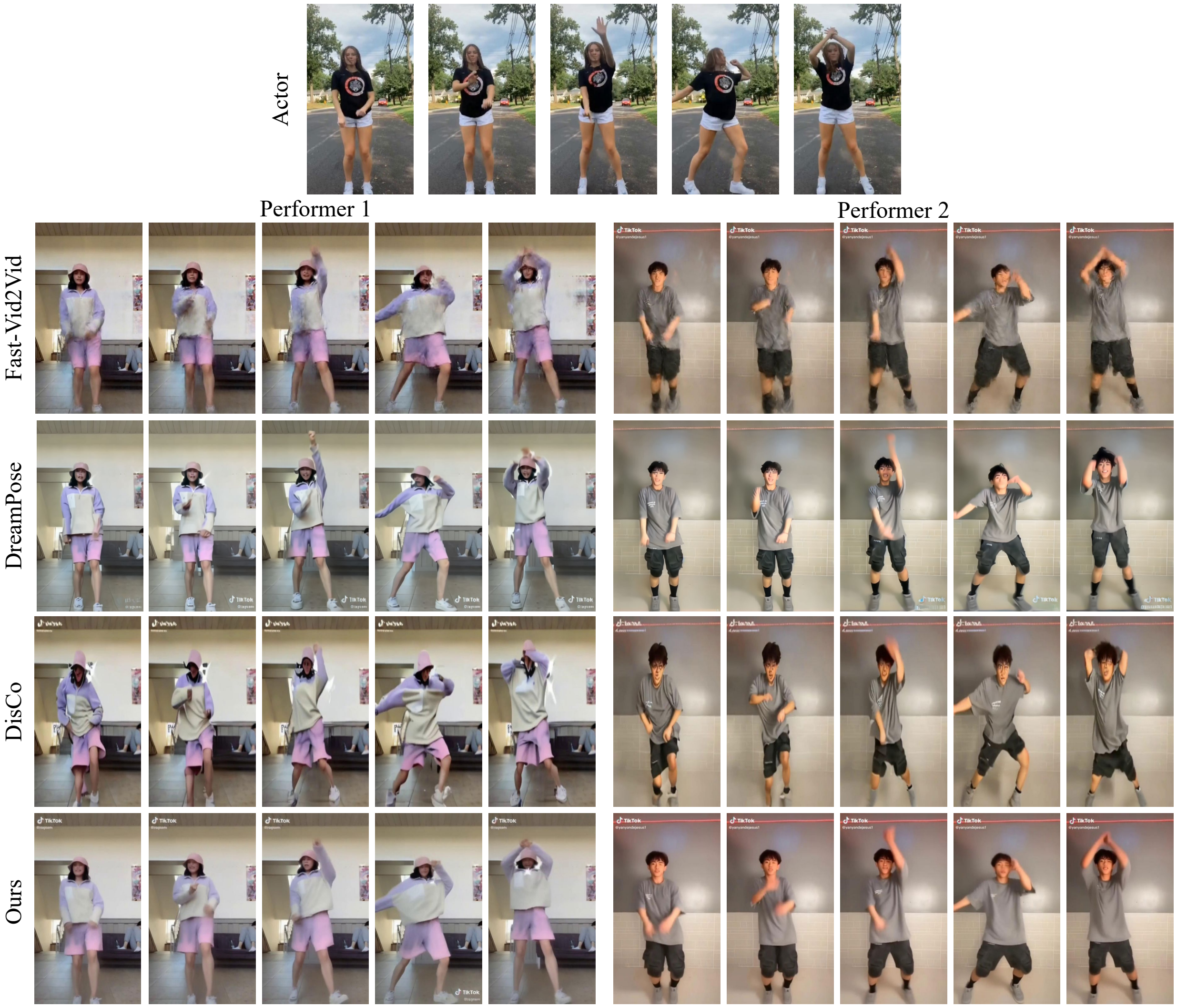


Figure 11. Qualitative comparisons of cross-identity driving with Fast-Vid2Vid [66], DreamPose [19] and DisCo [52] on the TikTok dataset. Zoom in for the best view.

D. More Comparison Results

We show more results in Figs. 11, 12 and 13. Figs. 11 and 12 show the cross-identity driving comparison results on the TikTok and Daily Captured Videos datasets. Fig. 13 shows the same-identity driving results on these two datasets. Note that although DisCo [52] trains the model on the TikTok dataset and achieves compelling results, it crops the video to 256 × 256 and focuses on the motion of the upper body. Therefore, when dealing with the poses on the whole body, it suffers from performance degradation. More dynamic results are shown in the video demo.

E. Further Analysis on Core Components

In this section, we provide more analysis on the core components (Implicit Video Representation and 3D-aware Human Flow) in the proposed method. Note that the dynamic results are shown in the video demo.

E.1. Implicit Video Representation

We further analyze the effect of implicit video representation (donated as IVR). As mentioned in Sec. 4.3, we add noise to the latent feature of the IVR results instead of at the image level. In order to verify the reasonableness of the latent guidance, we conduct additional experiment settings: (1) DiffPerformer without optimization (w/o Opt.), (2) optimization with the results of diffusion model (Image Opt.), (3) guiding the diffusion model at image-level as mentioned in Sec. 4.3 (Noisy Image Opt.), and (4) the complete framework of DiffPerformer (IVR Opt.). The results in Fig. 14 show that the output without optimization is temporally inconsistent and refining it using the diffusion model directly

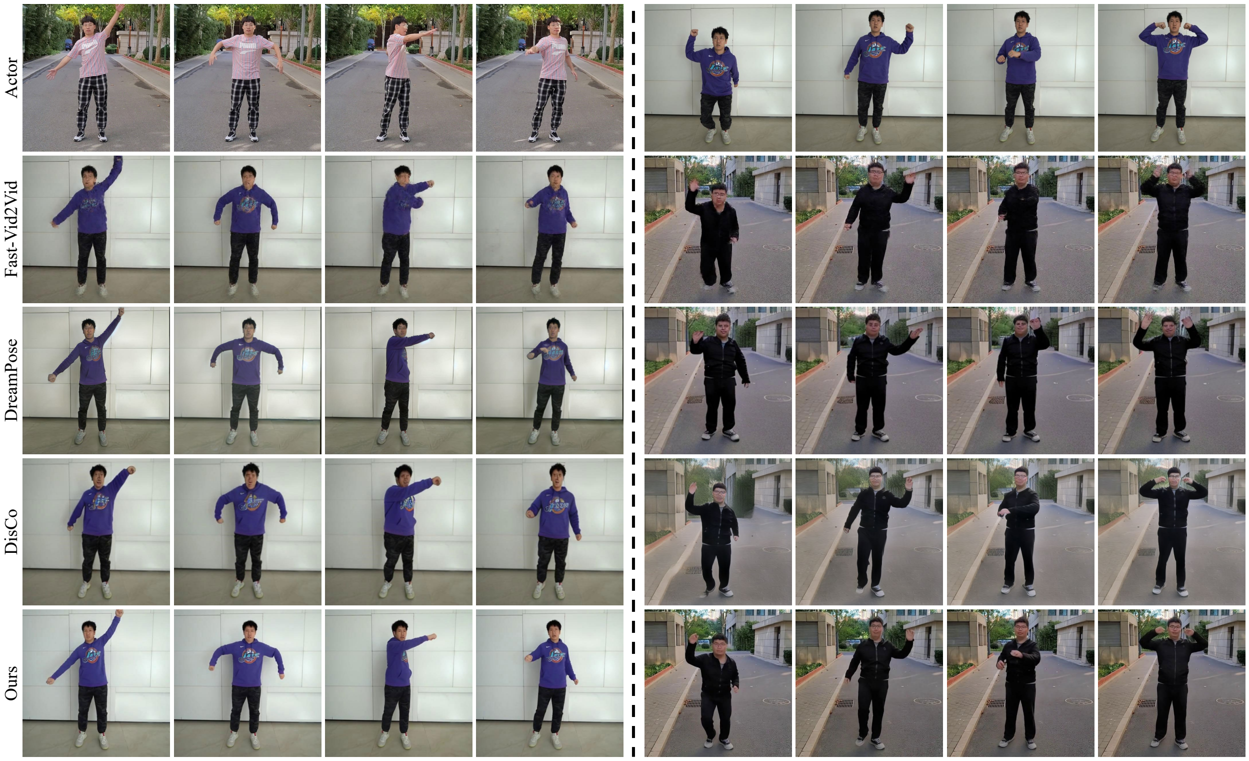


Figure 12. Qualitative comparisons of cross-identity driving with Fast-Vid2Vid [66], DreamPose [19] and DisCo [52] on the Daily Captured Videos dataset. Zoom in for the best view.

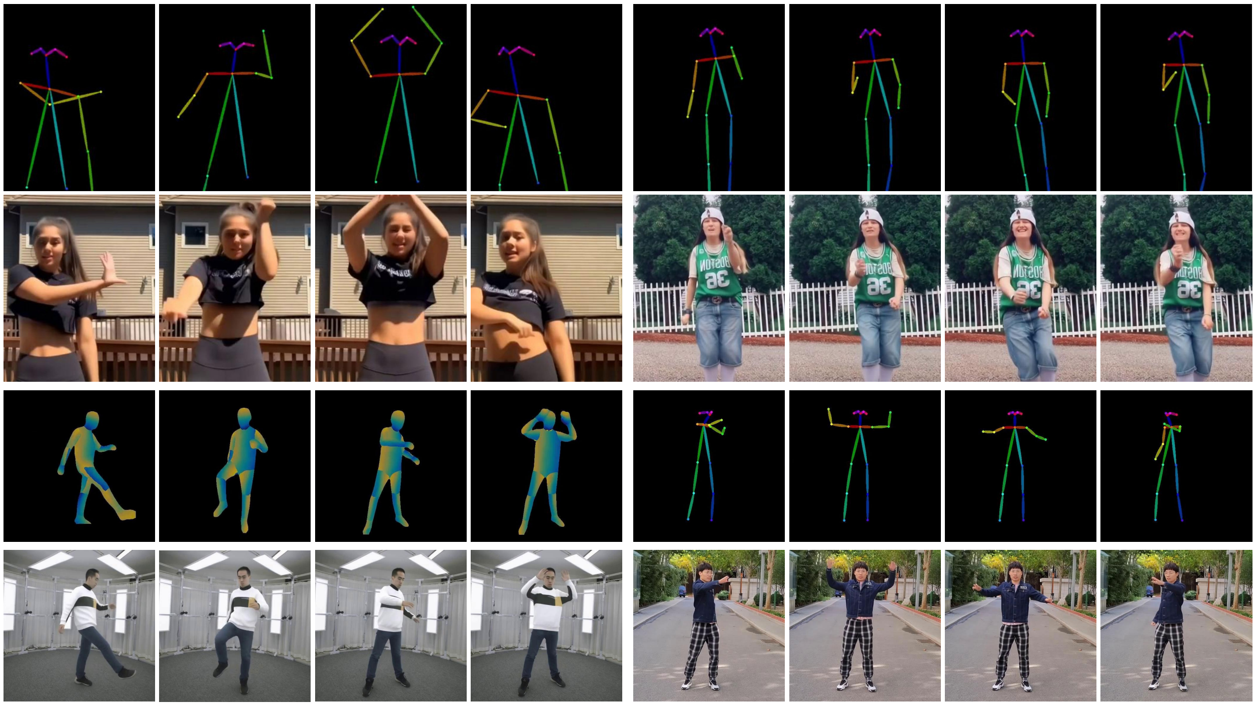
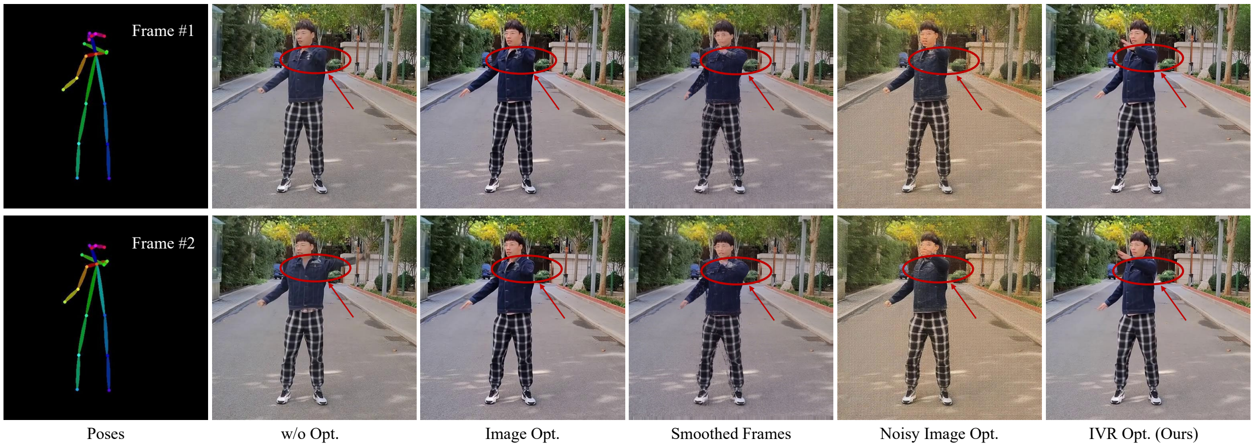


Figure 13. Our generated videos of self-identity driving on the TikTok and Daily Captured Videos datasets. Zoom in for the best view.

Table 2. Quantitative evaluation on test poses.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	L1 ↓	FVD ↓	FID-VID ↓
w/o Opt.	28.99	**0.70**	0.25	48.90	6.09E-5	24.81	375.45
Image Opt.	28.17	0.63	0.25	44.73	7.03E-5	25.11	432.22
Noisy-Image Opt.	27.35	0.45	0.41	43.12	8.51E-5	58.78	411.69
DiffPerformer (Ours)	**30.72**	0.69	**0.22**	**36.00**	**4.33E-5**	**22.32**	**254.39**



Poses w/o Opt. Image Opt. Smoothed Frames Noisy Image Opt. IVR Opt. (Ours)

Figure 14. Ablation study of the implicit video representation.

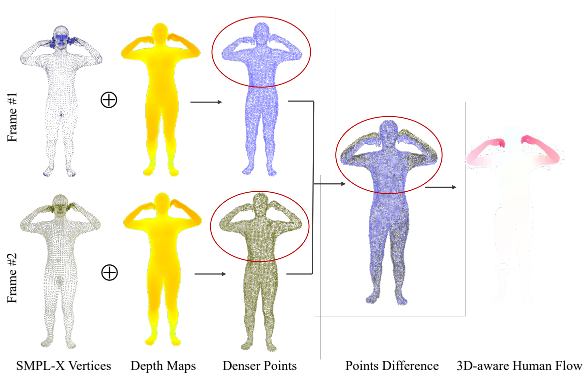
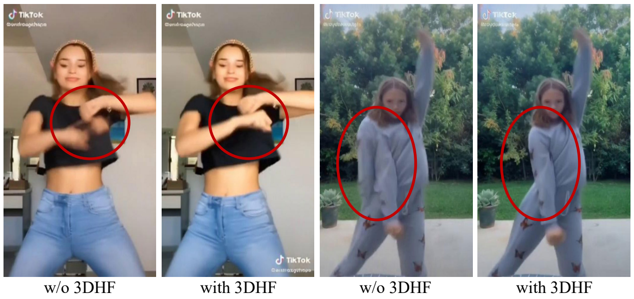


Figure 15. Illustration of 3D-aware Human Flow.

can not alleviate the issue but causes extra degradation such as color distortion. Furthermore, adding noise at the image level is unavailable to maintain temporal consistency. Even worse, it destroys the image content and produces unpleasing results. On the contrary, latent guidance not only preserves image content and style but also exploits the temporal domain properties of the smoothed video to generate a high-fidelity video with a consistent appearance. Table 2 provides the quantitative comparisons.

E.2. 3D-aware Human Flow

We propose a 3D-aware human flow (donated 3DHF) to build the correlation between motion and the specific character, as shown in Fig. 15. To exemplify the effectiveness of it, we visualize more ablation results in Fig. 16. We find that the results generated by DiffPerformer without 3DHF adhere to the appearance of the performer, but suffer from the misalignment between driving poses and appearance. We claim that the reasons are two-fold: (1) The joint optimization lacks constraints about the correspondence of poses and therefore cannot correct the mistakes during the generation. (2) The initialization of the implicit video representation provides fixed appearance guidance at the beginning of the generation, leading to the failure of pose alignment, especially in large-scale motion. 3DHF can effectively address the issue and help the diffusion model generate pose-alignment results.



w/o 3DHF with 3DHF w/o 3DHF with 3DHF

Figure 16. Ablation study of 3DHF.