Appendix for "DreamDance: Animating Human Images by Enriching 3D Geometry Cues from 2D Poses"

We present the video results at https://anonymous-dreamdance.github.io/. We also provide a static link in the supplementary material. Please double-click **index.html** in the **webpage** folder to view. We provide our code in the **code** folder.

1. Implementation Details

Our experiments are conducted on 8 NVIDIA A800 GPUs. To train the geometry diffusion model in the first stage, we initialize both the reference UNet and denoising UNet with Stable Diffusion v1.5. The training resolution is 256x384 and comprises 16 frames. A summarization is listed in Table 1. Note that we train this stage on low resolution, which makes it efficient when training for 3 steps and each step for 60,000 steps.

	Step1	Step2	Step3
Training Modules	all	geo. attn	temp. attn
Training Steps	60,000	60,000	60,000
Learning Rate	3e-5	2e-5	2e-5
Warm-up Steps	500	500	500

Table 1. Summarization of training details in Stage1.

As for the video diffusion model in the second stage, we initialize the model with SVD v1.1 and train the model for 50,000 steps with a batch size of 8. The training video is cropped and resized to the resolution of 512x768 and comprises 16 frames. We adopt Adamw as the optimizer and set the learning rate to 2e-5.

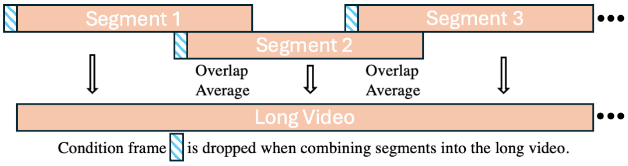

Figure 1. Long video generation.

2. Long Video Generation

Though our models are trained on a fixed length of frame numbers, they can be easily extended to generate long videos. As shown in Figure 1, during inference in each sampling step, we first generate each segment and then average the overlapped part to obtain a long segment. After iterative denoising, the process will finally generate a long video or condition maps.

3. Additional Results

We present more results on the generated results in Figure 2 and show the pseudo ground truth normal and depth in the last column.

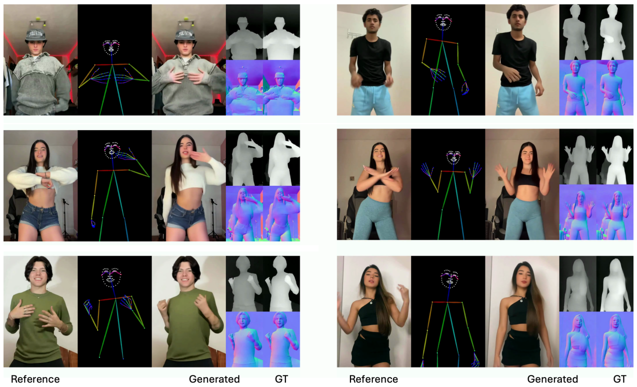

Figure 2. More results for generated results with pseudo ground truth.

We present more generated results with the generated normal and depth maps in Figure 3.

We provide additional results for comparison with baseline methods in Figure 4.

We provide more results from cross ID animation in Figure 5.

4. Limitations

Our method consists of two stages, where the generation results of the second stage are inherently influenced by those of the first stage. While we have implemented multiple

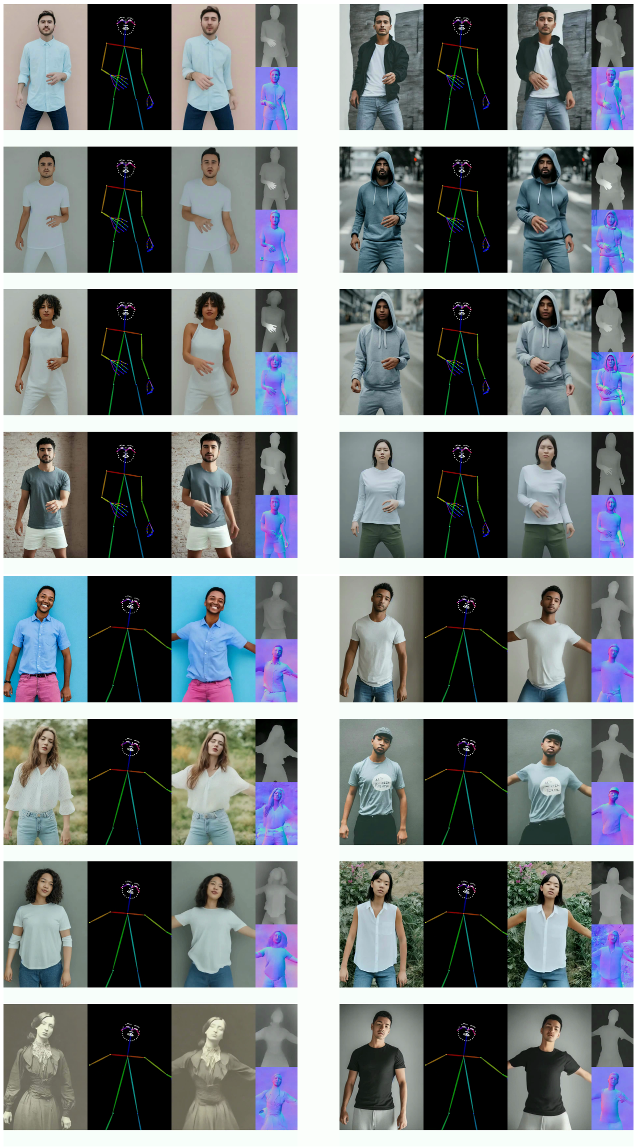

Figure 3. More results for generated results.

strategies to enhance the robustness and minimize error accumulation between the stages, rare and complete failure cases in Stage 1 could still have an impact on the performance of Stage 2.

5. Potential Negative Impacts

Our work may have potential negative impacts similar to those of other video generation methods. These may lead to copyright infringement, the proliferation of misleading content, and diminished creativity in video creation professions. Additionally, given the method's reliance on Internet-collected datasets, it raises concerns about data privacy and security.

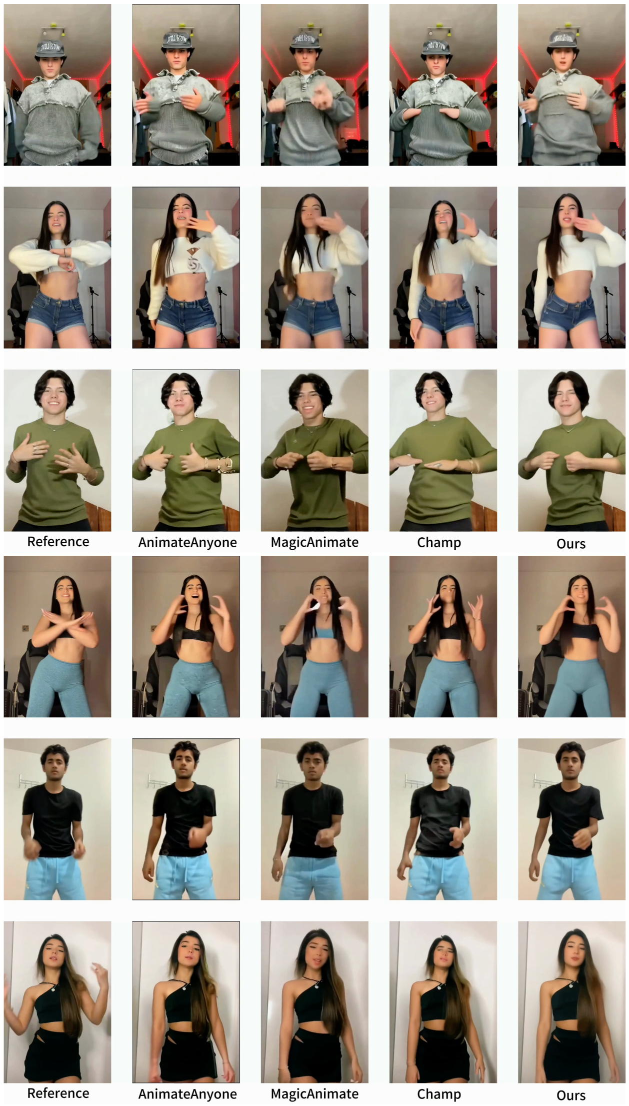

Figure 4. More results for comparison with baseline methods.

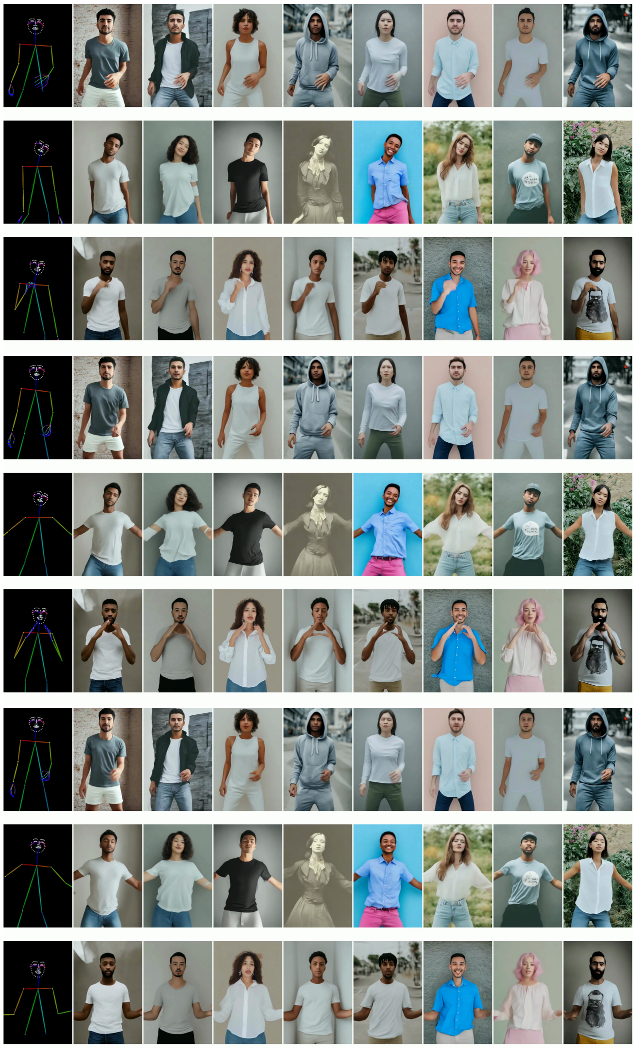

Figure 5. More results for Cross ID Animations.