Appendix - DISCO: Disentangled Control for Realistic Human Dance Generation

This appendix is organized as follows:

- Section A includes comprehensive analysis and comparison between our proposed DISCO and related works.
- Section B demonstrates how DISCO can be readily combined with subject-specific fine-tuning.
- Section C provides more qualitative and quantitative results to supplement the main paper.

A. Detailed Discussion on Related Work

We include additional discussions with the related visual-controllable image/video generation methods, especially the more recent diffusion-based models, due to the space limitation of the main paper. To fully (or partly) maintain the visual contents given a reference image/video, existing diffusion-based synthesis methods can be broadly divided into the following two categories based on their immediate applications:

Image/Video Editing Conditioned on Text. The most common approach for preserving specific image information is to edit existing images [14, 26, 38, 59] and videos [35, 42, 50, 68] with text, instead of unconditioned generation solely reliant on text descriptions. For example, Prompt-to-Prompt [14] control the spatial layout and geometry of the generated image by modifying the cross-attention maps of the source image. SDEdit [38] corrupts the images by adding noise and then denoises it for editing. DiffEdit [6] first automatically generates the mask highlighting regions to be edited by probing a diffusion model conditioned on different text prompts, then generates edited image guided by the mask. Another line of work requires parameter fine-tuning with user-provided image(s). For example, UniTune [61] tries to fine-tune the large T2I diffusion model on a single image-text pair to encode the semantics into a rare token. The editing image is generated by conditioning on a text prompt containing such rare token. Similarly, Imagic [26] optimizes the text embedding to reconstruct reference image and then interpolate such embedding for image editing.

For video editing, in addition to the Follow-your-pose [37] and Text2Video-Zero [27] discussed in the main text, Tune-A-Video [68] fine-tunes the SD on a single video to transfer the motion to generate the new video with text-guided subject attributes. Video-P2P [35] and FateZero [42] extend the image-based Prompt-to-Prompt to video data by decoupling the video editing into image inversion and attention map revision. However, all these methods are constrained, especially when the editing of the content cannot be accurately described by the text condition. We notice that a very-recent work Make-A-Protagonist [74] tries to introduce visual clue into video editing to mitigate this issue.

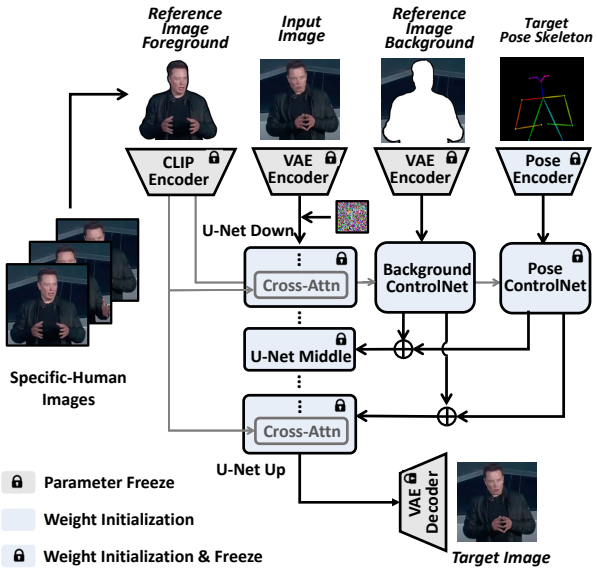


Figure 9. The model architecture for further subject-specific fine-tuning.

However, this approach, while innovative, is still met with limitations. On the one hand, it still struggles to fully retain the fine-grained human appearance and background details; on the other hand, it still requires a specific source video for sample-wise fine-tuning which is labor-intensive and time-consuming. In contrast, our DISCO not only readily facilitates human dance synthesis given any human image, but also significantly improves the faithfulness and compositionality of the synthesis. Furthermore, DISCO can also be regarded as a powerful pre-trained human video synthesis baseline which can be further integrated with various subject-specific fine-tuning techniques (see section B for more details).

Visual Content Variation. For preserving the visual prior, another line of work [8, 24, 44] directly feeds the CLIP image embedding into the diffusion model to achieve image/video variation. However, these approaches struggle to accurately control the degree as well as the area of the variation. To partially mitigate this problem, DreamBooth [47] and DreamMix [39] necessitate multiple images to fine-tune the T2I and T2V models for learning and maintaining a specific visual concept. However, the precise visual manipulation is still missing. In this paper, we propose a disentangled architecture to accurately and fully control the human attribute, background and pose for referring human dance generation.

B. Subject-Specific Finetuning

As mentioned in the main paper, our DISCO can be flexibly integrated with existing efficient fine-tuning techniques for even more fine-grained human dance synthesis. This is particularly beneficial when facing out-of-domain reference images, which appear visually different to the TikTok style

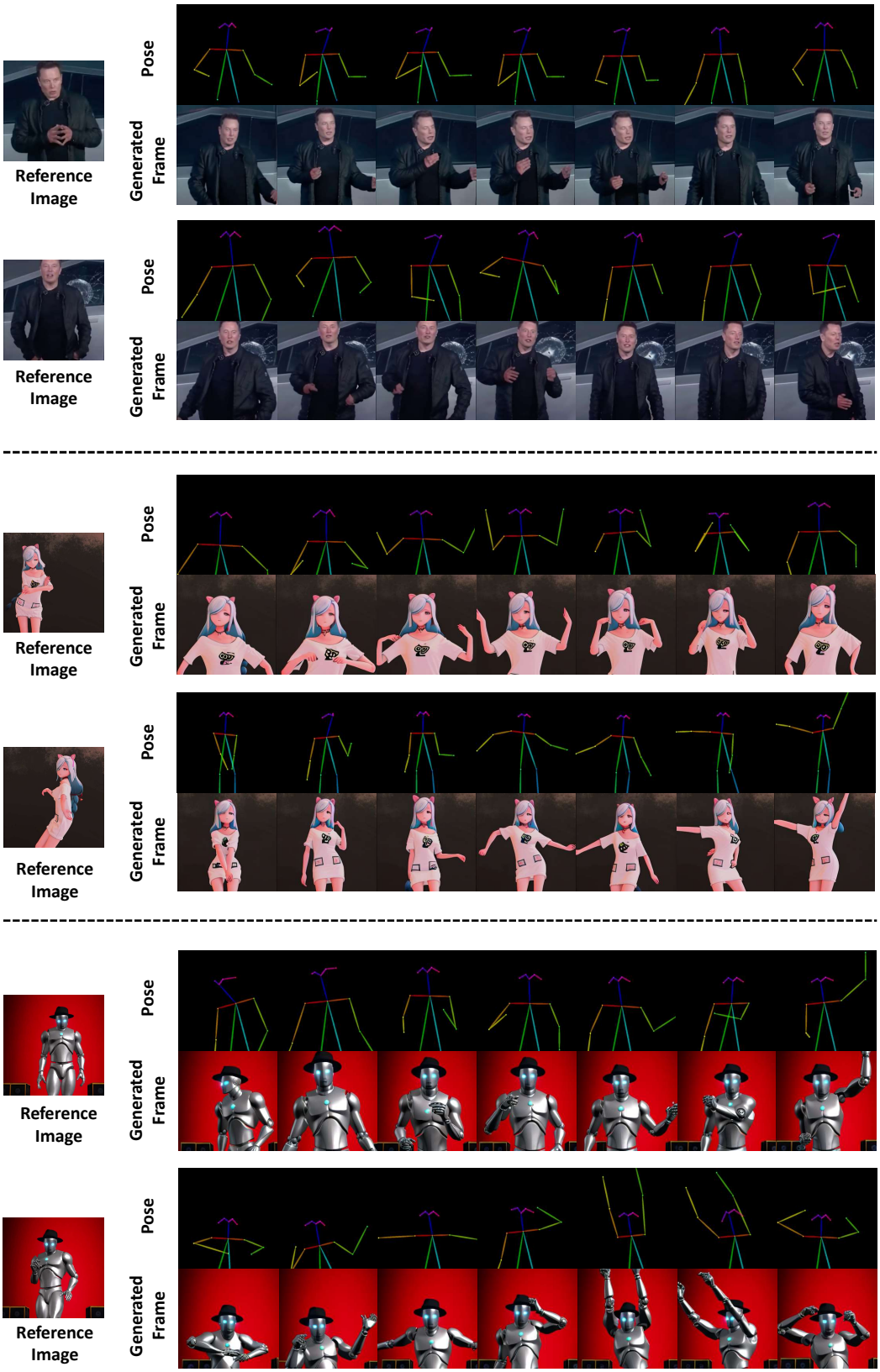


Figure 10. The synthesis frames for out-of-domain human subject after subject-specific fine-tuning guided by the pose sequence extracted from the TikTok dataset.

Table 6. Additional ablation results on architecture designs. "ControlNet (fg+bg)" and "Attention (fg+bg)" in the second block denote inserting the control condition of reference image (containing both foreground and background) via a single ControlNet or cross-attention modules. "HAP w/ pose" denotes adding pose ControlNet path with pose annotation into HAP. "ControlNet-Pose" Init. means initializing pose ControlNet of fine-tuning stage with the pre-trained ControlNet-Pose [71] checkpoint.

Method	FID ↓	SSIM ↑	PSNR ↑	LISPIS ↓	L1 ↓	FID-VID ↓	FVD ↓
DISCO	61.06	0.631	28.78	0.317	4.46E-04	73.29	366.39
DISCO + TikTok HAP	50.68	0.648	28.81	0.309	4.27E-04	69.68	353.35
Ablation on control mechanism w/ reference image (DISCO setting: ControlNet (bg) + Attention (fg))							
ControlNet (fg+bg, no SD-VAE)	83.53	0.575	28.37	0.411	5.35E-04	89.13	551.62
ControlNet (fg+bg)	65.14	0.600	28.57	0.355	4.83E-04	74.19	427.49
Attention (fg+bg)	80.50	0.474	28.01	0.485	7.50E-04	80.49	551.51
Ablation on HAP w/ pose (DISCO setting: HAP w/o pose)							
TikTok HAP w/ pose	51.84	0.650	28.89	0.307	4.16E-04	68.55	346.10
Ablation on initializing w/ pre-trained ControlNet-Pose (DISCO setting: initialize with U-Net weights)							
ControlNet-Pose Init.	62.18	0.633	28.37	0.320	4.46E-04	72.98	389.47
ControlNet-Pose Init.+TikTok HAP	55.81	0.641	28.69	0.316	4.43E-04	78.13	363.38

images. Figure 9 presents the framework for subject-specific fine-tuning, which is easily adapted from the framework presented in the main text (Figure 2a). Rather than utilizing a set of videos of different human subjects for training, subject-specific fine-tuning aims to leverage limited video frames of a specific human subject (*e.g.*, the video of Elon Mask talking about Tesla Model 3 in Figure 9 or even anime in Figure 10) for better dance synthesis. Compared to the standard fine-tuning, we additionally freeze the pose ControlNet branch and most parts of U-Net to avoid over-fitting to the limited poses in the subject-specific training video, only making the background ControlNet branch and the cross-attention layers in U-Net trainable. We also explored the widely-used LoRA [21] for parameter-efficient fine-tuning and observed similar generation results. As this is not the main focus of this paper, we leave other advanced techniques to future explorations along this direction.

Implementation Details. The model weights are initialized with the model fine-tuned on the general TikTok dancing videos. We train the model on 2 NVIDIA V100 GPUs for 500 iterations with learning rate $1e^{-3}$, image size 256×256 and batch size 64. The randomized crop is adopted to avoid over-fitting. The subject-specific training videos range from 3s to 10s, with relatively simple poses.

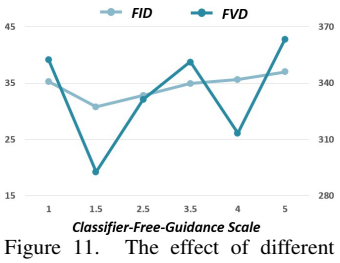


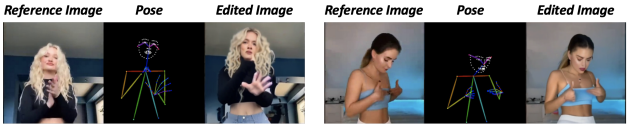
Figure 11. The effect of different classifier-free-guidance scale.

Qualitative Results. We test the subject-specific fine-tuning on various out-of-domain human subjects, including real-world celebrities and anime characters. After training, we perform the novel video synthesis with an out-of-domain reference image and a random dance pose sequence sampled in TikTok training set. As shown in Figure 10, upon additional fine-tuning, DISCO is able to generate dance videos preserving faithful human attribute and consistent background across an extensive range of poses. This indicates the considerable potential of DISCO to serve as a powerful pre-trained checkpoint.

C. Addition Results

C.1. Adaptable to Other Conditions

DISCO can be easily extended to face landmarks and hand gestures to provide more fine-grained control.



C.2. More Quantitative Results

Full Ablation Results. We show the full ablation results on architecture design in Table 6. In what follows, we focus on discussing results that are not present in the main text. In the first block of the table, we copy over the results from the full instances of DISCO, with or without HAP on TikTok Dance Dataset for reference.

As mentioned in the main text, we propose to use the pre-trained VQ-VAE from SD, instead of four randomly initialized convolution layers in the original ControlNet for encoding the background reference image. In the second block of the table, we ablate this design by comparing two models, (1) "ControlNet (fg+bg)", inserting the control condition of reference image (containing both foreground and

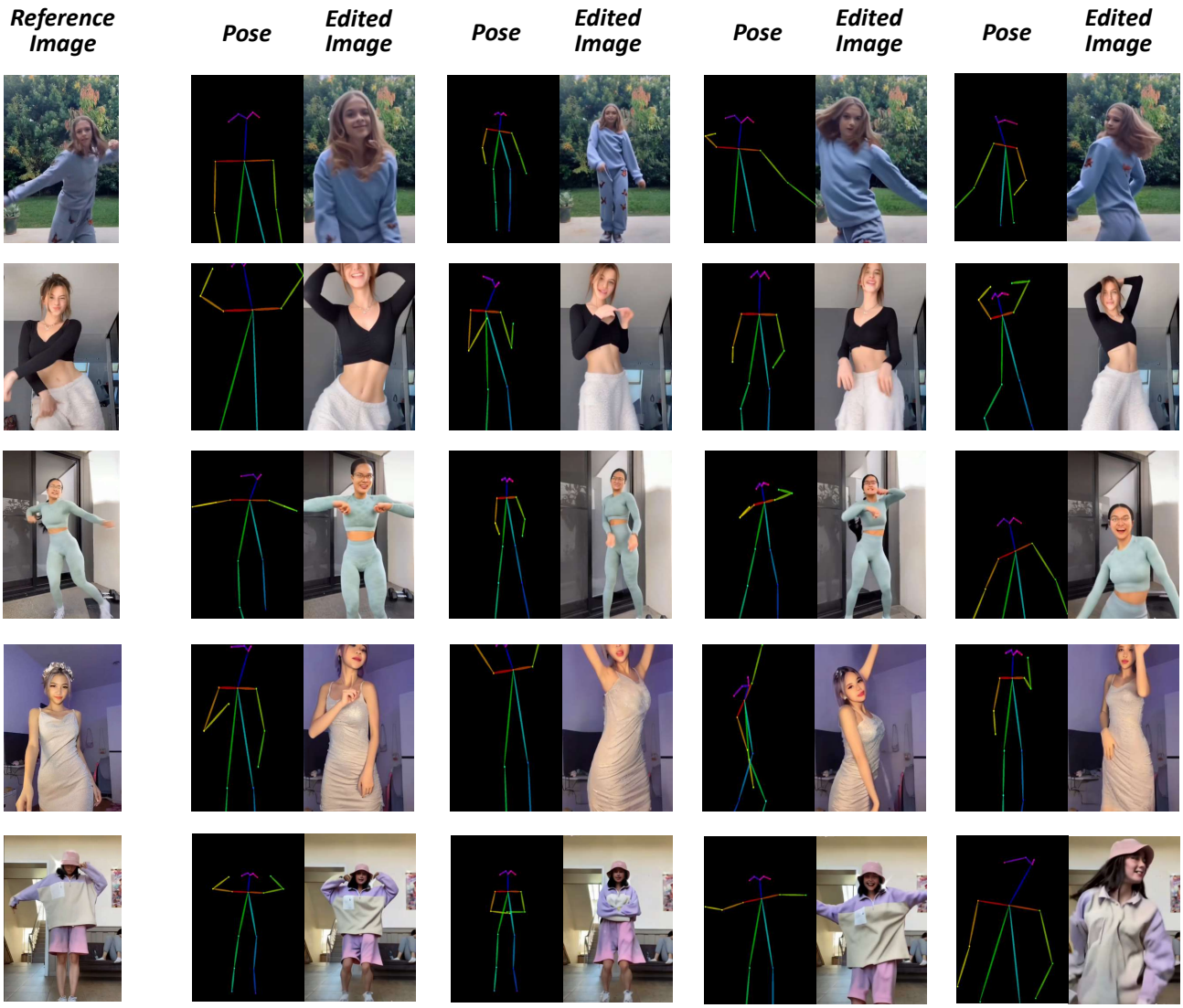


Figure 12. More visualizations for different image ratios (*i.e.*, vertical image) and various human views (*e.g.*, from close-view to full-body) of human image editing.

background) via a single ControlNet, with VQ-VAE encoding and (2) "ControlNet (fg+bg, no SD-VAE)", inserting the control condition of reference image via a single ControlNet with four randomly initialized convolution layers as the condition encoder. We note that the pre-trained VQ-VAE can produce a more descriptive dense representation of the reference image, contributing to better synthesis results (FID 65.14 v.s 83.59). In the third block of the table, we investigate whether adding pose condition into human attribute pretraining is beneficial to the downstream performance. We observe that integrating pose into HAP leads in similar results, but requires additional annotation efforts on pose estimation.

Last but not least, we examine on the initialization of the pose ControlNet branch. Specifically, we try to initialize from the pre-trained ControlNet-Pose checkpoint [71] during fine-tuning. The results are shown in the last block of Table 6. Without HAP, the performance is comparable to DISCO, but it gets much worse than DISCO when both are pre-trained with HAP. This is because that the ControlNet-Pose is pre-trained with text condition and can not fully accommodate referring human dance generation with the reference image condition. After HAP, such gap is further enlarged, leading to even worse results.

Effect of CFG Scale. In Figure 11, we show the effect of varying the classifier-free-guidance scale. We can find that scale of 1.5 gives the best quantitative results for both image-wise and video-wise fidelity.

Pre-trained Motion Prior. In addition to utilizing the appearance pre-training, we also investigate the potential of the motion prior learned from the large-scale general video. We inject motion prior from the

Figure 13. The qualitative comparison between different architecture designs for the video frame generation.

[User Study] Human Dance Generation

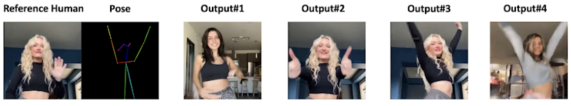
Task Description:
Given the reference human image and the reference pose， the model need to generate the image/video containing the same human in new pose.**Objective of the Study:**
Participants are asked to evaluate the quality of the generated image or video. Ratings should be given on a scale from 0 to 5, where 0 indicates a complete mismatch and 5 represents an ideal synthesis. The evaluation focuses on four key aspects:

1. **Human ID Alignment:** Assess how closely the details of the generated person align with those in the reference image. Closer alignment is considered better.
2. **Pose Alignment:** Evaluate whether the pose in the generated output matches the reference pose. Greater alignment is preferable.
3. **Background Alignment:** Determine the extent to which the background in the generated output matches the background in the reference image. More consistent backgrounds are favored.
4. **Overall Quality:** Provide a subjective overall rating that considers all the above aspects.

For further guidance, please refer to the provided rating examples.

Please rate the following generated image. Your rating should reflect how well the image meets the specified requirements.

Rating Examples (For Reference Only, No Response Required)



Human ID Alignment	0	5	5	1
Pose Alignment	0	0	5	4
Background Alignment	4	5	5	1
Overall Quality	0	2	5	2

Please rate the following generated image. *

	0	1	2	3	4	5
Human ID Alignment	○	○	○	○	○	○
Pose Alignment	○	○	○	○	○	○
Background Alignment	○	○	○	○	○	○
Overall Quality	○	○	○	○	○	○

Please rate the following generated video. *

	0	1	2	3	4	5
Human ID Alignment	○	○	○	○	○	○
Pose Alignment	○	○	○	○	○	○
Background Alignment	○	○	○	○	○	○
Overall Quality	○	○	○	○	○	○

Figure 14. The interface of our user study. We provide detailed instructions, rating examples, and rating instances for users.

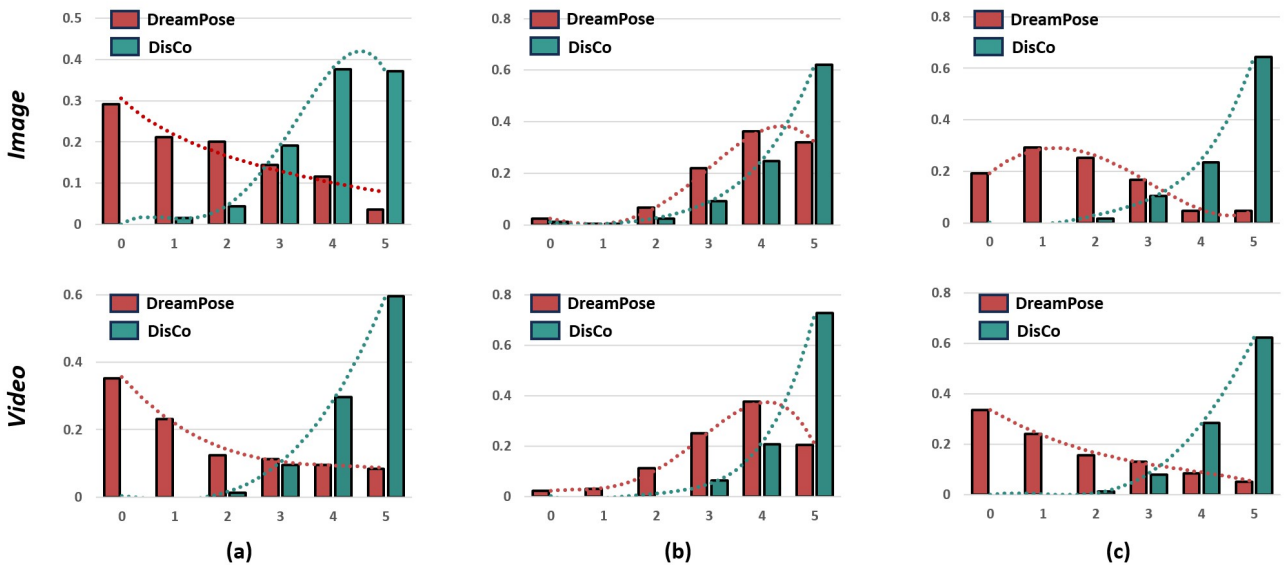


Figure 15. Full results of User Study: (a) human ID alignment; (b) pose alignment; and (c) background alignment score distribution on both synthesis images (above part) and videos (below part) of DreamPose and DISCO.

AnimateDiff (AD) [13] checkpoint pre-trained on the large-scale video and then fine-tune on TikTok data. Results in left table show superior temporal smoothness, making it worthwhile to further explore motion pre-training via video data together with attribute pre-training via image corpus.

Method	FID-VID ↓	FVD ↓
w/o. AD	20.75	257.90
w/. AD	18.96	241.55

C.3. Qualitative Results

DISCO can be easily adapted to different image size. For example, we show more qualitative results of human image editing in Figure 12 with image size of 256×384 to include more human body. Please note that most videos of the TikTok dataset are relatively close to the camera. However, we can see that DISCO can handle both partial and full human body synthesis even with large changes in viewpoints and rotations in the human skeleton. More results for video generation are shown in Figure 16.

Figure 13 compares the video synthesis results with baseline architectures, to supplement Figure 8 of the main text. With a sequence of continuous poses, we can discern more clearly that both ControlNet-only and Attention-only baseline fail to maintain the consistency of human attributes and background, leading to less visually appealing generations than our DISCO.

C.4. User Study

As indicated in the main paper, we conducted a user study involving 50 distinct individuals to compare our DISCO with its key counterpart, DreamPose [25] to evaluate the quality of the generated images and videos. The study contains 80 rating questions for 20 synthesis (10 images and 10 videos) in total. There are 5 images and 5 videos for both DISCO and DreamPose. In addition to the coarse-grained overall quality score adopted in DreamPose, we further divide the rating evaluation into three fine-grained aspects: (i) human ID alignment; (ii) pose alignment; and (iii) background alignment. For each aspect, the users are required to rate on a scale of 0 to 5, where 0 corresponds to an ideal corresponding (*e.g.*, the ground-truth image/video) to the reference input image and 0 for non-match at all. This results in 4000 responses in total. Please check Figure 14 for the user study interface and Figure 15 for the full results.

We can observe that, for both human ID and background, DISCO (green histogram) achieves clearly higher scores compared to DreamPose (red one). This indicates that the naive combination of different conditions (*e.g.*, DreamPose) suffers from inconsistent human attribute and unstable background, while DISCO can reconstruct the fine-grained human attributes and main steady backgrounds thanks to the proposed human attribute pre-training and well-designed disentangled control. As a much easier condition, pose control can be generally maintained for both DreamPose and DISCO. However, DISCO still demonstrates obvious superiority compared to DreamPose.

C.5. Ethical Concern

Despite the broad applicability of DISCO, it is crucial to consider the risks of misuse in creating deceptive media, address potential biases in training data, and ensure respect for intellectual property.

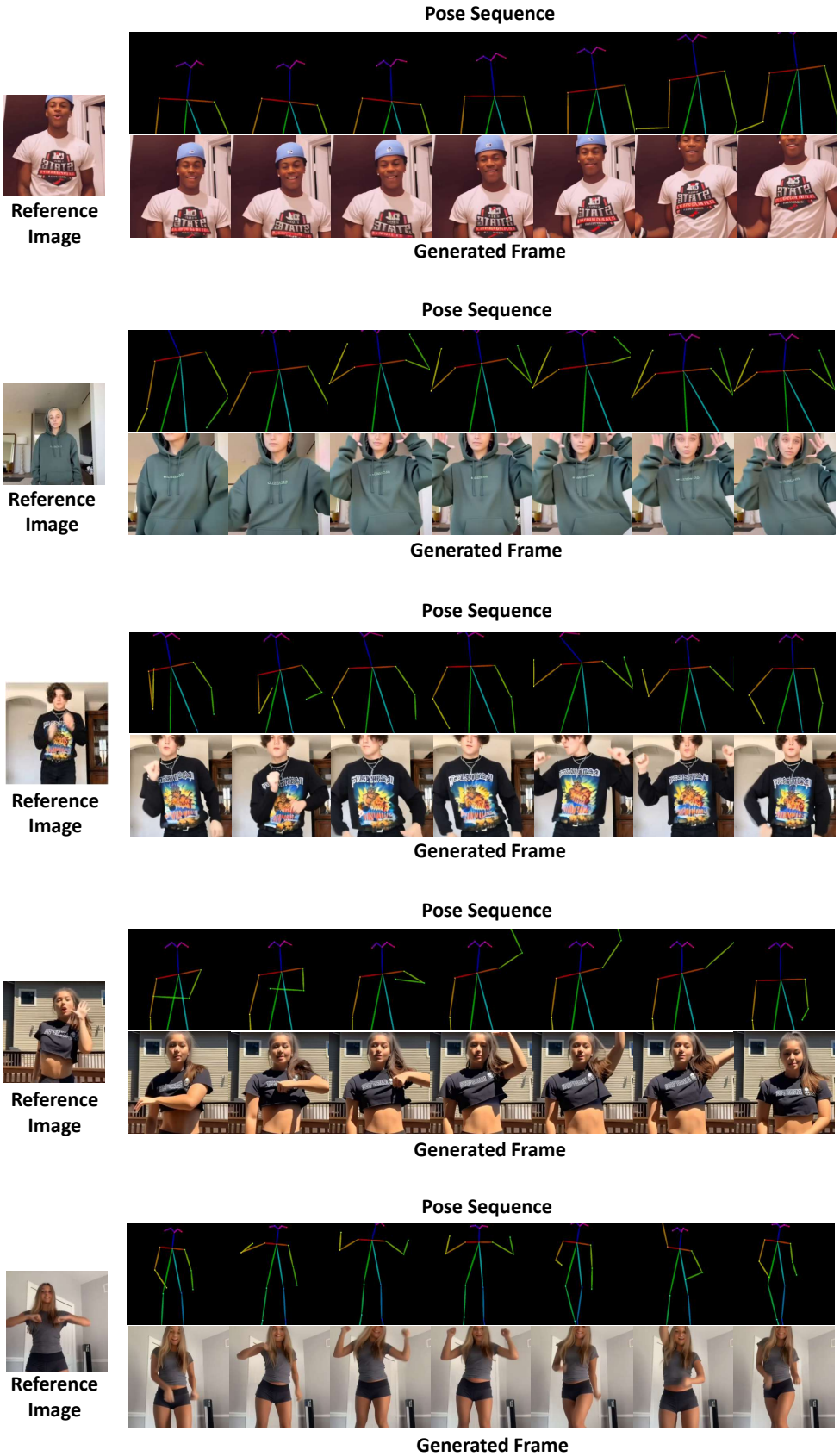


Figure 16. More qualitative examples for video generation.