Supplemental Materials Controllable Human Video Generation from Sparse Sketches

I. Implementation Details

A. Training Dataset Generation

For training data preparation, we initially fine-tune the Stable Diffusion (SD) [1] using a real video and then utilize a Scribble ControlNet [2] to introduce sketch guidance for generating a series of shape-variant images with customized appearances. For the sketch input to the ControlNet, directly extracting Canny edge maps from video frames will capture unwanted fine details of the texture. Thus, we first extract the parsing maps of the video via the existing method [3] and then regard the boundary of those parsing maps as the sketch images (Figure 1 (a)).

Since AvatarBooth [4] demonstrates that fine-tuning a SD model with full-body human images might lead to an imbalance between the representations of facial features and body clothes, we separately customize two models $\epsilon^{full}_{sd+control}$ and $\epsilon^{face}_{sd+control}$ for the full body appearance and facial appearance. The training processes of the two models are almost the same. Only the training data for the facial model is derived from cropped regions of the full-body images, ensuring focused training on facial attributes. During the synthetic reference image generation, we first produce the full-body image with the input of the sketch and text via $\epsilon^{full}_{sd+control}$. Since the facial region constitutes only a small fraction of the entire body image, it often leads to ambiguity in the output (Figure 1 (b) (II)). To enhance the face quality, we regenerate this region using $\epsilon^{face}_{sd+control}$ along with the cropped facial sketch. The facial output is then seamlessly integrated back into the original image (Figure 1 (b) (III)).

Fig. 1. (a) An example about sketch extraction. (b) Three generated reference results. (I) generated from fine-tuned ϵ^{face}_{sd} and ϵ^{full}_{sd} with the pretrained ControlNet. (II) generated from fine-tuned $\epsilon^{full}_{sd+control}$. (III) generated from fine-tuned full-body $\epsilon^{full}_{sd+control}$ and face $\epsilon^{face}_{sd+control}$

Specifically, we fine-tune $\epsilon^{face}_{sd+control}$ and $\epsilon^{full}_{sd+control}$ for 2,000 and 5,000 iterations separately with the LoRA rank of 16, a learning rate of 10^{-4}, and a batch size of 1. The fine-tuning procedure needs 30 minutes for $\epsilon^{full}_{sd+control}$ and 10 minutes for $\epsilon^{face}_{sd+control}$ on an NVIDIA RTX4090 GPU. During data generation, the shape-varying sketches are collected from the existing human full-body dataset [5], [6], and we use a DDIM sampler for 50 denoising steps. The seeds for all experiments are 42. We manually remove the low-quality synthetic images and ensure that the number of synthesized reference images corresponding to each video is more than 300. The examples of synthesized reference images are shown Figure 2.

B. Long Video Generation with Few Sketches

Considering computational constraints, the generation of long videos is commonly handled in segments. In our method, we employ a sliding mechanism [7] to partition an entire pose sequence into multiple segments with temporal overlap. These segments are then fused in the latent space to enhance the smooth concatenation of frames. When provided with multi-frame sketches, for each segment, we individually select the closest sketch for both the first and last frames within that specific segment. For instance, when considering the conditions $\{S_i, S_j, S_k\}$, the sketch inputs for the segment $\{P_{m:n}\}$ are as follows:

$$\begin{cases} (S_i, S_k), & |m-i| < |m-j| \& |n-k| < |n-j| \\ (S_i, S_j), & |m-i| < |m-j| \& |n-k| > |n-j| \\ (S_j, S_k), & |m-i| > |m-j| \& |n-k| < |n-j| \end{cases}, \quad (1)$$

where $|\cdot|$ is the absolute value between the position of two frames.

C. Training Strategy

To decrease the training time, we first temporarily forego the motion modules in the SSE and focus on training this model to restore the shape features between two arbitrary sketches. The input data comprises three selected frames $\{I_j, I_i, I_w\}$ with noise, the corresponding pose images $\{P_j, P_i, P_w\}$, two sketch images (S_j, S_w) ($j < i$ and $w > i$), and a reference image $\hat{I}_{ref}$ randomly chosen from our synthesized dataset. Until the model demonstrates satisfactory performance, we re-introduce the motion modules into the encoder, and initialize them with the pre-trained weights from [8]. During this stage, we mainly focus on training the temporal layers while fixing the weights for the rest of the network. The training data for this stage consists of an N-frame video clip $\{I_{1:N}\}$ alongside the corresponding conditions $(\{P_{1:N}\}, (S_1, S_N), \hat{I}_{ref})$. The losses for both stage are computed as the average across all frames.

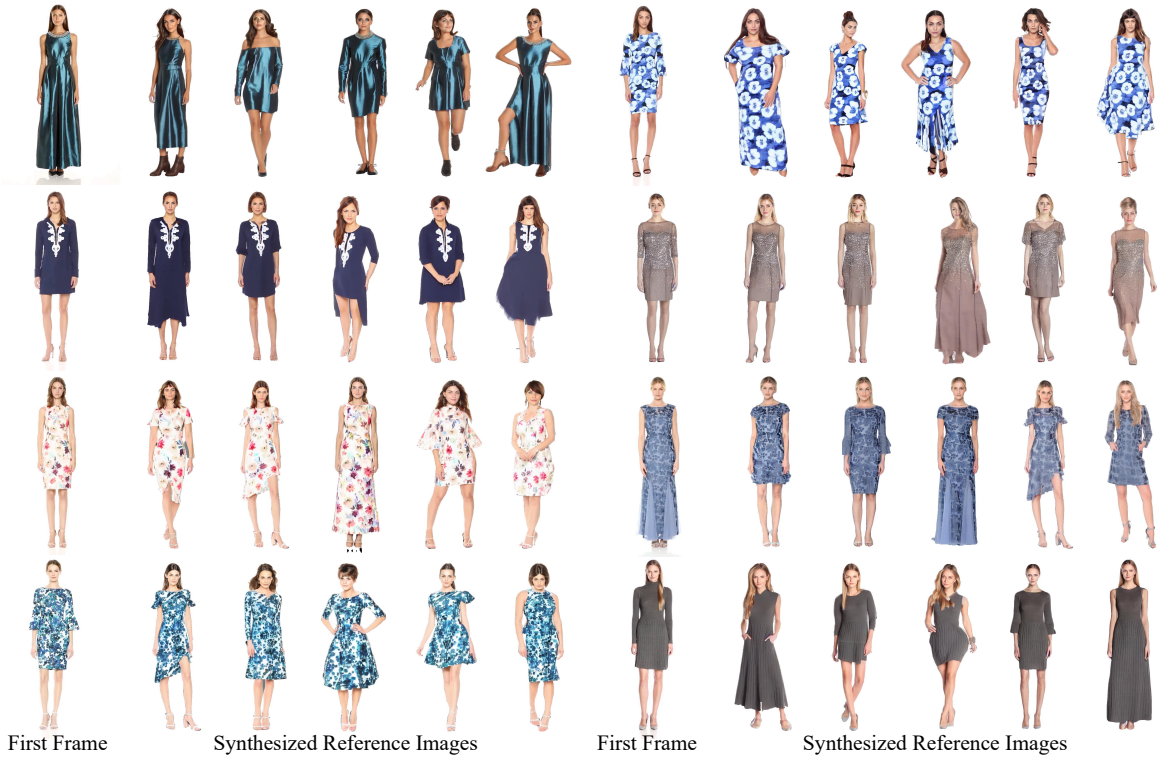

Fig. 2. Several examples of the synthetic reference training dataset.

Fig. 3. Seven paired examples of generated reference images for test videos in the UBC Fashion dataset [5]. Each pair includes the first frame of the video (Left) and the corresponding reference image (Right).

D. Details for Compared Methods

To facilitate a quantitative comparison about sketch controllability across different shape adapters, we prepare a shape-varying reference image for each video in the test dataset. Each reference image differs from the corresponding video in at least one aspect, such as hair length or collar type. Examples of these reference images are depicted in Figure 3. We plan to release all synthesized reference images upon the acceptance of the paper. For qualitative comparison, the reference images are selected from our synthesized shape-varying reference images and real full-body human datasets [5], [6], [9].

1) DreamPose [10]: Initially, we fine-tune DreamPose using their official implementation. To ensure a fair qualitative comparison, we utilize the edited reference image instead of the reference image and the DensePose sequence [11] as input. The edited image is the selected front-view sketch-conditioned frame from our results. The denoising steps are set to 100.

2) BDMM [12]: We directly employ the official checkpoints for BDMM, with inputs structured similarly to DreamPose. The DensePose sequence are substituted with the rendered SMPL images.

3) MagicAnimate [7]: We directly employ the official checkpoints for MagicAnimate, with inputs structured similarly to DreamPose. The denoising steps are set to 30.

4) Champ [8]: We directly employ the official checkpoints for Champ, with inputs structured similarly to DreamPose. The DensePose sequence are substituted with DWpose skeleton images. Additionally, the inputs also include the corresponding depth images, normal images and semantic maps. The denoising steps are set to 20.

5) SpCtrlN-V: This baseline combines the backbone SD model with two add-on modules, SparseCtrl [13] and the motion modules [14]. We first fine-tune the SD model with the LoRA layers. The fine-tuning details closely follow the data generation stage of our method (Section III-A in main paper). Subsequently, the two add-on modules incorporate the pre-trained models from SparseCtrl [13] and AnimateDiff [14]. Since the scribble encoder does not support pose inputs, the inputs only consist of the reference image (for fine-tuning) and sparse sketches. The denoising steps are set to 25.

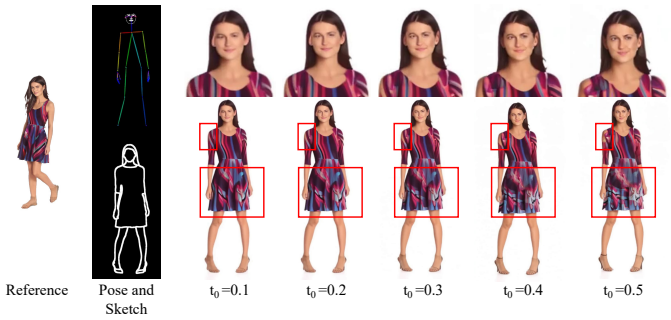

Fig. 4. The results after resampling stage with different parameters t_0.

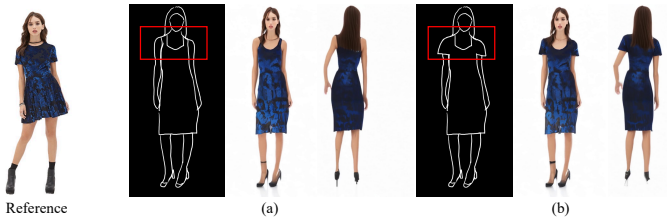

Fig. 5. An example of sketch geometry editing. Given the same reference image and pose sequence, (a) and (b) illustrate the corresponding results before and after editing, respectively.

6) CtrlN-V: Similar to SpCtrlN-V, CtrlN-V substitutes the SparseCtrl with Scribble ControlNet [2] and Skeleton ControlNet [2]. The fine-tuning process for SD remains consistent with SpCtrlN-V. Here, we integrate the pre-trained models from ControlNet [2] and AnimateDiff [14]. During inference, Scribble ControlNet receives sketch images for conditional frames and black images for unconditional frames as inputs, while Skeleton ControlNet utilizes skeleton images derived from the DensePose sequence [11]. The denoising steps are set to 30.

II. Addition Results

A. Disentangled Control for Shape and Appearance

Benefiting from the model design and our synthesized reference data, our model has the capability to modify the appearance while retaining its shape or modify the shape while retaining its texture. Figure 6 illustrates the human sequence generated solely by using different reference images but the same pose sequences and sketches. Similarly, inputting different sketches but the same reference can lead to varied shapes, as shown in Figure 7. Additionally, our method facilitates video shape editing through modifications of partial input sketches, as shown in Figure 5. The results demonstrate that our proposed method supports controllable generation of high-quality and diverse human videos.

B. Details for Ablation Study

1) The hyperparameter of Resampling Strategy: Our resampling strategy can effectively improve the visual quality of generated videos. An important parameter of the resampling stage is t_0, which control the degree of the perturbation. As t_0 increases, less input information in coarse canvas is retained and tends to cause discontinuities at the overlap region (Figure 4). Thus, considering the realism of results and the smoothness of the stitching areas after resampling stage, we choose the $t_0 = 0.3$.

III. Ethical Issues

Our approach has the potential to modify real portrait images and videos. However, utilizing the reference image for appearance control without altering the geometry could inadvertently affect the portrayal of gender and race, potentially compromising the ethical integrity of the original portrait. We strongly discourage the misuse of our technology for spreading false information or tarnishing someone's reputation. It is of utmost importance to exercise caution and prudence when employing this technology.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[3] L. Lin, Y. Gao, K. Gong, M. Wang, and X. Liang, "Graphonomy: Universal image parsing via graph reasoning and transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2504–2518, 2020.

[4] Y. Zeng, Y. Lu, X. Ji, Y. Yao, H. Zhu, and X. Cao, "Avatarbooth: High-quality and customizable 3d human avatar generation," *arXiv preprint arXiv:2306.09864*, 2023.

[5] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "Dwnet: Dense warp-based network for pose-guided human video generation," in *Proceedings of the British Machine Vision Conference*, 2019.

[6] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C. C. Loy, W. Wu, and Z. Liu, "Stylegan-human: A data-centric odyssey of human generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–19.

[7] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[8] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, "Champ: Controllable and consistent human image animation with 3d parametric guidance," in *European Conference on Computer Vision (ECCV)*, 2024.

[9] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion video synthesis with stable diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 680–22 690.

[11] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[12] W.-Y. Yu, L.-M. Po, R. C. Cheung, Y. Zhao, Y. Xue, and K. Li, "Bidirectionally deformable motion modulation for video-based human pose transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7502–7512.

[13] Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai, "Sparsectrl: Adding sparse controls to text-to-video diffusion models," *arXiv preprint arXiv:2311.16933*, 2023.

[14] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *International Conference on Learning Representations*, 2024.

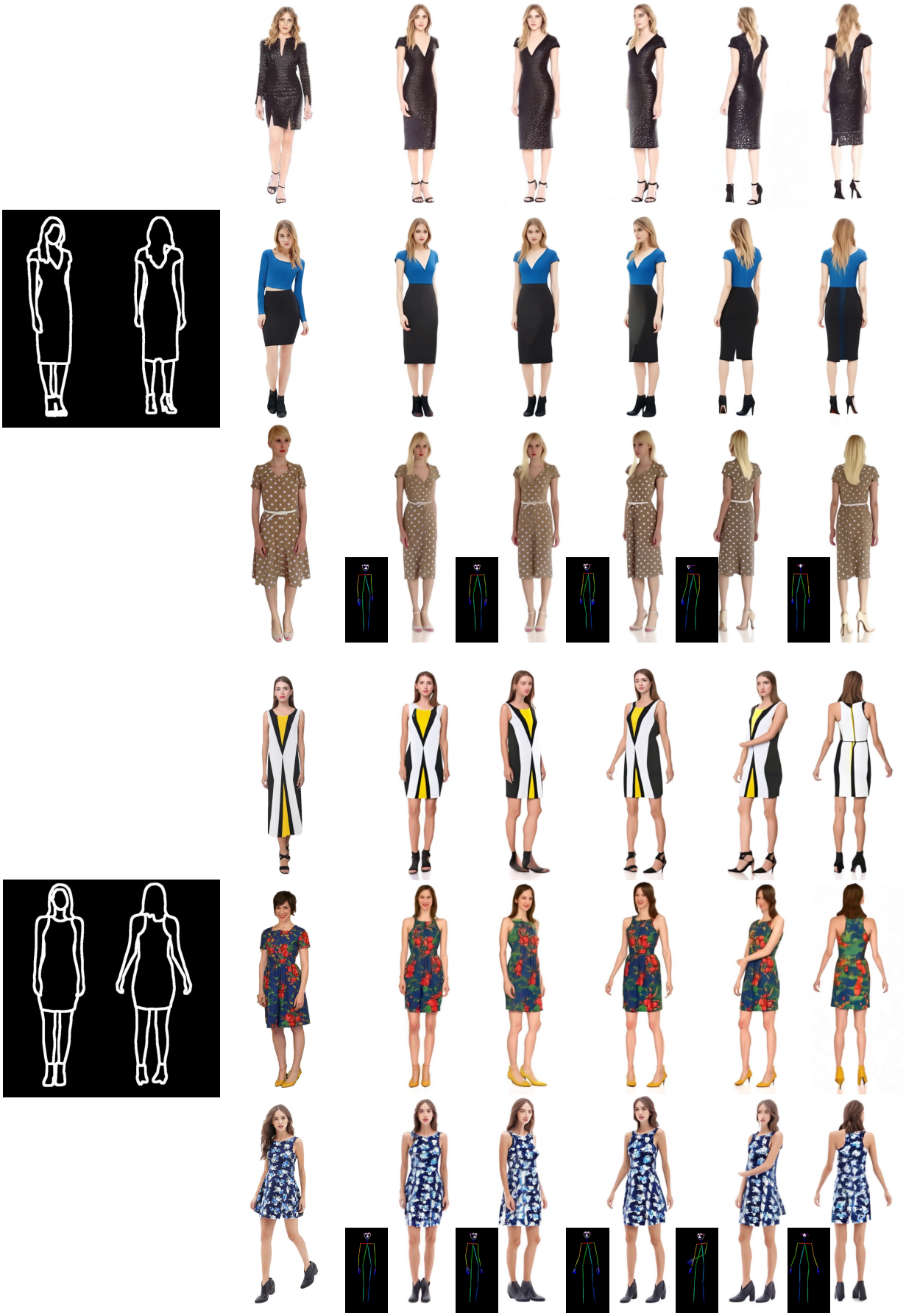

Fig. 6. More results for conditioning on the same pose sequence and sketches but different reference images. The reference images are from the SHHQ dataset [6].

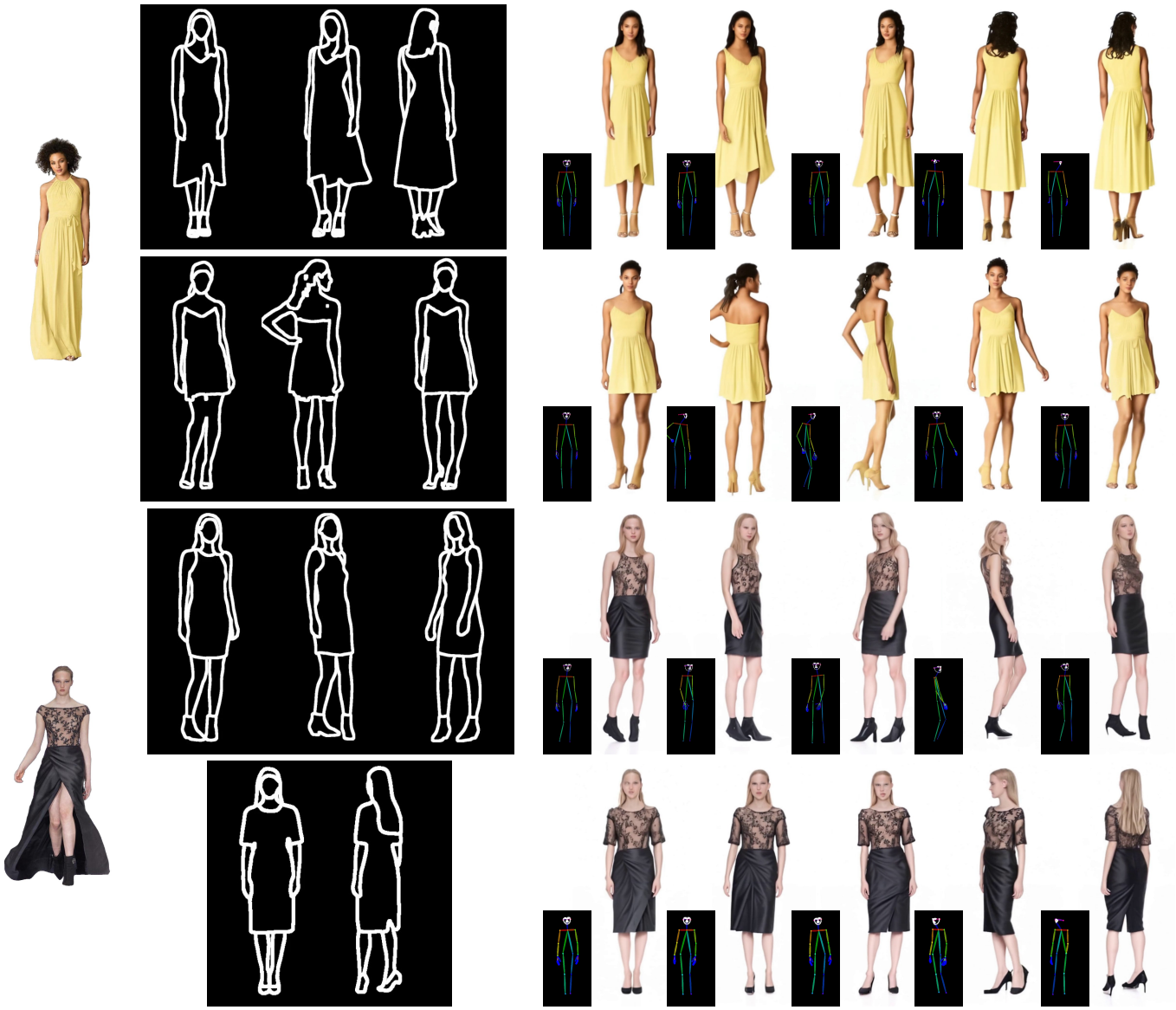

Fig. 7. More results for conditioning on the same reference image but different pose sequences and sketches. The reference images are from the SHHQ dataset [6].