Pose-Diverse Multi-View Virtual Try-on from a Single Frontal Image via Diffusion Transformer

Supplementary Material

Prompt for Gemini Flash 2.5 Image VTON

Our virtual try-on module is based on Gemini Flash 2.5 Image. We designed a structured multi-modal prompt consisting of the target person image, the garment image, and a text instruction. The exact prompt used in our implementation: Create a new image by combining the elements from the provided images. Take the [element from image 2] and place it with/on the [element from image 1]. Ensure the final image is a realistic depiction of a person wearing the garment. The output should be a single image that seamlessly integrates the clothing onto the person.

Dataset processing details

Our training data preparation involved processing five datasets: Renderpeople [6], THuman 2.1 [10], 2K2K [2], MVHumanNet [8], and Renderpeople(A-pose). The first three datasets contain complete 3D objects, which we projected into four canonical views using a camera positioned at 90-degree intervals. For the last two datasets, which do not provide 3D models directly, additional preprocessing was required to align them with our standardized rendering pipeline.

MVHumanNet: The MVHumanNet dataset consists of multi-view captured video frames, which are not in 3D object. To utilize it for our training pipeline, we performed a 3D reconstruction process for each subject. Specifically, we first applied COLMAP [7] to extract keypoints from the multi-view images and approximate the corresponding camera parameters for each view as supplied camera parameter in MVHumanNet was unreliable. We then used Splatfacto [4] to reconstruct the subject as a 3D Gaussian splatting. Finally, from the gaussian splat, we projected the reconstructed 3D models into 2D images from four fixed viewpoints (0°, 90°, 180°, 270°). These rendered views were then used as supervision for our multi-view diffusion training. The artifacts of the reconstructed 3D gaussian splattings were manually removed for better quality.

Renderpeople: Renderpeople offers rigged 3D human models. We rig all characters into a A-pose. We rendered images at fixed viewpoints (0°, 90°, 180°, 270°) with consistent camera parameters across identities. The obtained high-quality paired images are suitable for evaluating pose control and multi-view generation fidelity.

Inference on In-the-wild images. Our method requires multi-view OpenPose [1] skeletons as conditional input, which are not available for single frontal view in-the-wild images. To address this, we employ ECON [9] to obtain SMPL [5] and convert it into OpenPose. More specifically, we applied ECON to frontal human photos to generate a SMPL mesh, then projected the mesh to novel viewpoints. We extracted 2D human keypoints via OpenPose from these synthetic views to serve as pose guidance. This procedure enabled us to extend our pipeline to unconstrained internet images.

Inference time

	Inference Time
VTON360 (A100)	**1491.95s**
Ours first stage	**15s**
Ours second stage(multi-view, A100)	**8.7 s**
Ours second stage(multi-view, RTX3090)	**11.28 s**

Table 1. Inference time of our framework on RTX 3090 and baseline on A100.

Table 1 shows the inference time measured on a single RTX 3090 and a single A100. VTON360 takes 1491.95 seconds with an A100 (does not fit on RTX 3090), while our framework takes 15 seconds on the VTON stage and 8.7 seconds with an A100. (VTON360 does not fit on RTX 3090)

Additional qualitative results on complex garments

To further evaluate our method, we additionally conduct an experiment on complex garments as shown in Figure 1. We also extend our evaluation to a CustomHumans dataset [3]. Our proposed method successfully reflects complex patterns of input garments. For the other views (90°, 180°, 270°), our multi-view synthesis diffusion model generates complex patterns well followed by the frontal view.

Gemini Flash 2.5 multi-view generation

We also attempted to generate multi-view virtual try-on results using only Gemini 2.5 Flash Image. However, due to the lack of explicit control mechanisms such as pose conditioning, the model fails to produce consistent outputs. In particular, it struggles to generate results in the same pose as input image. Figure 2 shows the multi-view VTON results using Gemini Flash 2.5 with the given prompt. Gemini Flash 2.5 fails to maintain the posture of VTON input and view faithfulness.

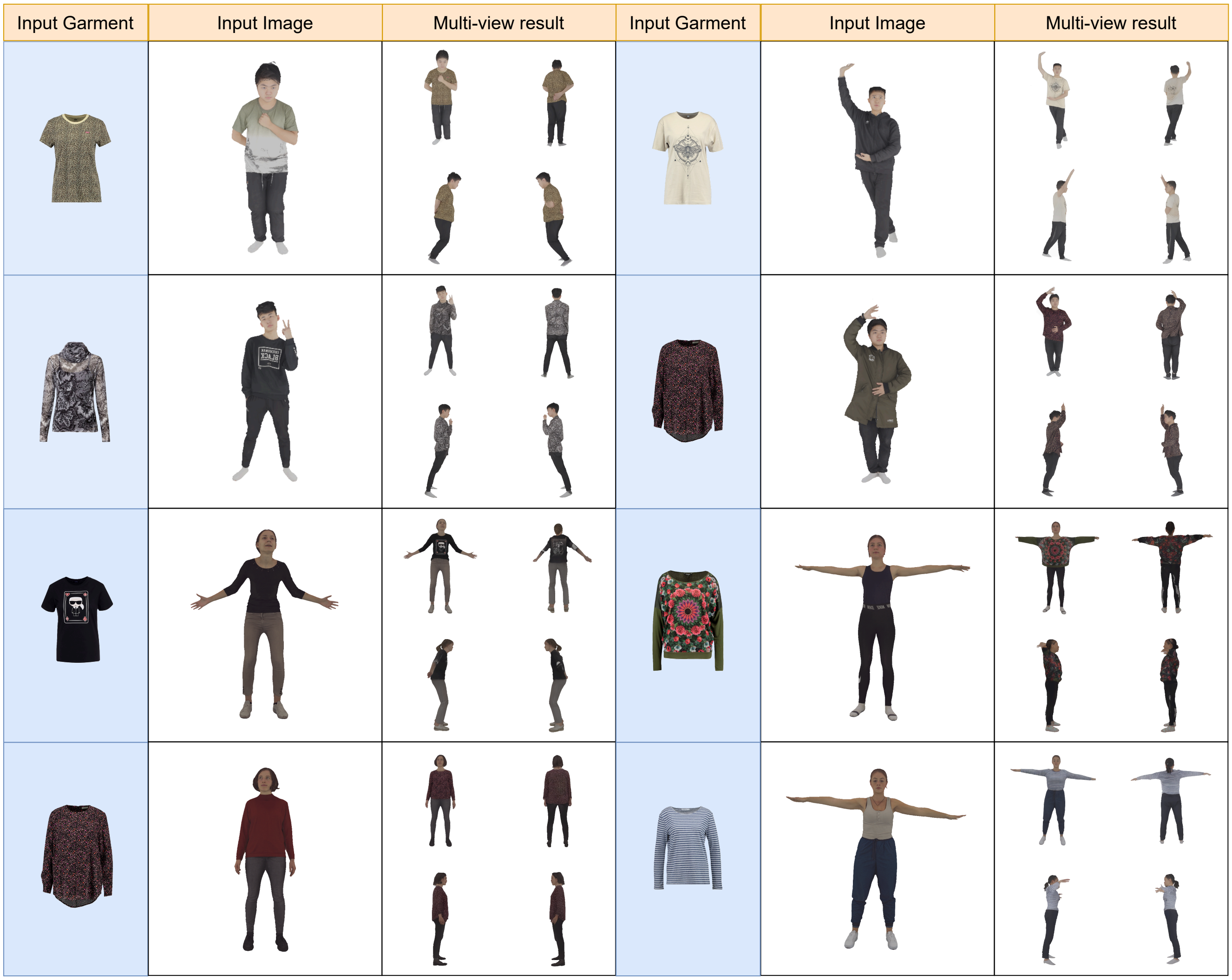

Figure 1. Additional inference result with complex garments.

Prompt 1: Use the person in this image to create a full-length character turnaround sheet from the front, right, back, and left sides with the same look and dress on a clean monochrome background.

Prompt 2: Using the person in this image, generate a full-body person turnaround sheet showing the same appearance and outfit against a clean solid background, including front, right side, back, and left side views.

Ethical considerations

Our system integrates Gemini Flash 2.5 Image for virtual try-on synthesis. We emphasize that Gemini's built-in safeguards automatically filter NSFW (not safe for work) content during inference. As such, our pipeline does not produce inappropriate or harmful outputs even when tested on in-the-wild images.

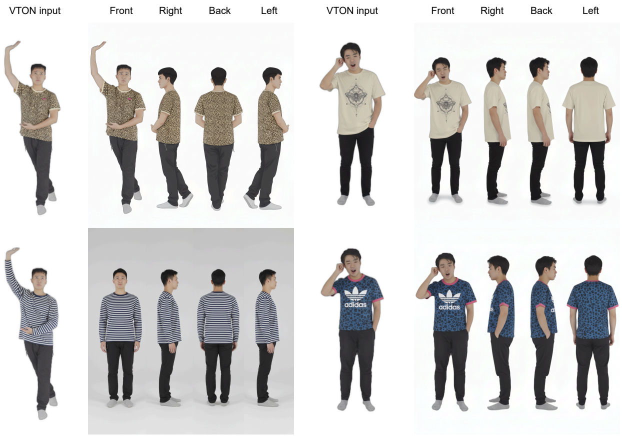

Figure 2. Gemini Flash 2.5 fails to generate Multi-view and maintain posture.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 1

[2] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1

[6] Renderpeople. Renderpeople 3d human model dataset, 2024. Accessed: 2025-05-20. 1

[7] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[8] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[9] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[10] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1