Supplementary Material for FashionMAC: Deformation-Free Fashion Image Generation with Fine-Grained Model Appearance Customization

Rong Zhang[1], Jinxiao Li[1], Jingnan Wang[1], Zhiwen Zuo[1*], Jianfeng Dong[1], Wei Li[2], Chi Wang[3], Weiwei Xu[3], Xun Wang[1*]

[1]Zhejiang Gongshang University
[2]Nanjing University
[3]Zhejiang University

zhangrong@zjgsu.edu.cn, {module8627, wangjingnan751}@gmail.com, zzw@zjgsu.edu.cn, {dongjf24, liweimcc}@gmail.com, wangchi1995@zju.edu.cn, xww@cad.zju.edu.cn, wx@zjgsu.edu.cn

In this supplementary material, we provide more details of the FashionMAC and additional results.

Garment-Centric Pose Predictor

Without any human structural priors, directly outpainting a deformed garment to synthesize a fashion model wearing it is very difficult. Therefore, we first design a garment-centric pose prediction model to generate the corresponding poses that fit the given garment.

Specifically, the garment-centric pose predictor is based on the LDM framework. Given an input fashion showcase image, we obtain the deformed garment image and the corresponding pose map from it by a pre-trained cloth segmentation model (Dabhi 2021) and a DensePose model (Güler, Neverova, and Kokkinos 2018), respectively. We then train a UNet denoiser ϵ_{θ_p} from scratch to generate the pose conditioned on the garment. Since the pose map and the garment image are spatially aligned, we simply concatenate the garment image's latent encoding z_c with the pose image's noised latent encoding z_{p_t} in the channel dimension as the input to the UNet denoiser ϵ_{θ_p} at timestep t. The objective function for training is as follows:

$$\mathcal{L}_{pose} = \mathbb{E}_{p,c,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_{\theta_p}(z_{p_t}, z_c, t)||_2^2]. \quad (1)$$

During the inference, thanks to the stochasticity of the reverse denoising process, we can utilize the garment-centric pose predictor to sample diverse pose maps that fit the given garment.

Dataset

To support our framework, we construct fine-grained text prompts and corresponding spatial masks for each training image of VITON-HD (Choi et al. 2021). For the prompt extraction, we adopt QWen 2.5VL (Yang et al. 2025) to generate descriptive attribute prompts for predefined semantic regions. We define 9 core region categories that are most relevant to fashion, including expression, skin, hair, mouth, etc. To extract the spatial region masks, we utilize the semantic segmentation maps from the VTON-HD dataset for the body region and employ additional segmentation tools SAM (Kirillov et al. 2023) and Bisenet (Yu et al. 2018) to obtain the hair and face region.

Implementation Details

We adopt a three-stage training strategy: 1) We train FashionMAC without RADA for $55,000$ iterations, using a batch size of 16 and learning rate of $2e-5$. 2) We add the RADA module and continue finetuning with ground-truth region masks for $6,000$ iterations, using the same learning rate. 3) We freeze the denoising model and RADA, and train the mask prediction head for 60 epochs with a learning rate of $1e{-}4$. In inference, we replace ground-truth masks with predicted ones and apply the chained mask injection strategy to propagate structural guidance across timesteps.

Results Conditioned on Virtual Faces

Our FashionMAC can generate fashion showcase images under the guidance of facial images. To avoid issues of portraiture rights of real humans, the input facial images can be obtained from portrait generation approaches such as StyleGAN (Karras, Laine, and Aila 2019). Fig. 1 demonstrates the results of FashionMAC conditioned on virtual faces generated by StyleGAN. The generated images successfully preserve the distinctive characteristics of the input faces, showcasing FashionMAC's ability to maintain high fidelity in face-conditioned image generation.

Qualitative Results of the Ablation Study

More qualitative results of the ablation study are shown in Fig. 2 to assess the effectiveness of the proposed Chained Mask Injection (CMI) strategy and Region-Adaptive Decoupled Attention (RADA) module. The ablation studies are conducted on our full model FashionMAC, model without CMI and model without CMI and RADA. These results further validate the effectiveness of CMI and RADA.

Experiments of the Mask Prediction

To enable accurate region-aware conditioning during the denoising process, we explore where to place the mask prediction head for optimal spatial guidance. We experimentally found that applying mask prediction on low-level encoder

*Corresponding author.

Figure 1: The results of our method conditioned on virtual facial images generated by StyleGAN.

features at early denosing timesteps tends to produce imprecise region masks. In contrast, decoder-based masks remain relatively stable and less affected by the timestep. The results are illustrated in Fig. 3. We attribute this to the noisy nature of early-step features and the limited capacity of shallow encoder blocks to extract meaningful spatial cues from them. Conversely, decoder features benefit from deeper hierarchical feature integration, which enhances their ability to localize fine-grained semantic regions. Based on this observation, we adopt a chained mask injection strategy that progressively predicts and injects masks at network layers across timesteps, providing stable and precise regional priors to guide generation.

More Results of the Fine-grained Customization

Fig. 4 and Fig. 5 present additional qualitative comparisons and diverse visual results of FashionMAC. These examples highlight the framework's superior capability for fine-grained appearance customization and its effectiveness in generating high-fidelity, diverse fashion showcase images.

Comparison with the Virtual Try-On Based Methods

To evaluate the performance of our method and virtual try-on (VTON) based methods on the task of fashion image generation, we conduct an experiment with a two-stage VTON-based pipeline CosmicMan+TPD. In the first stage, we generate a virtual fashion model to show the target garment using a state-of-the-art text-to-image human generation framework CosmicMan (Li et al. 2024). In the second stage, we utilize the VTON method TPD (Yang et al. 2024) to generate the showcase image in which the above virtual fashion model wears the target garment. This pipeline mimics the showcase image generation capabilities of our framework. The results reveal that FashionMAC achieves outstanding generation performance, excelling in preserving human structural features and garment fidelity. Fig. 6 demonstrates some generated showcase images with CosmicMan+TPD and our method, respectively.

Fashion Image Generation for Mannequins

In e-commerce, the retailers usually use mannequins to coarsely demonstrate the garments. Our FashionMAC can utilize mannequin garment images as input and generate realistic apparel showcase images. For an input mannequin image that wears a garment, we can obtain the target garment through a pretrained segmentation network. Then FashionMAC can generate the corresponding pose map and produce high-quality fashion images. Fig. 7 shows the results conditioned on mannequin images. The results indicate the robustness and generalization ability of our method.

Applicability of FashionMAC

The goal of FashionMAC is to generate high-quality garment showcase images for e-commerce retailers. The key points of the task are (i) preserving garment details and (ii) controllable model appearance customization. Given a fixed segmented garment, the diversity of body shapes and the poses of the synthesized models are limited. Yet, FashionMAC can accept photos of users or plastic mannequins wearing the garment with diverse body shapes and poses as inputs, and then customize the model appearances from these inputs, which are easy to obtain in real e-commerce scenarios. Currently, FashionMAC's performance across different body shapes and poses is constrained by limited dataset diversity. However, as FashionMAC does not require paired training data as in VTON, expanding training data to include a wider range of body shapes should substantially improve its generalization.

Limitation and Future Work.

Since current fashion datasets contain data biases, such as the predominance of young female fashion models, the range of controllable attributes achievable by our method trained on these datasets is limited. Expanding these datasets to include a broader variety of ages, body types, genders, cloth types and backgrounds would be a promising direction for future work, as it would enhance the versatility and controllability of our method in generating more diverse and representative outputs.

References

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Dabhi, L. 2021. Cloth Segmentation. https://github.com/levindabhi/cloth-segmentation. Accessed: 2024-11-15.

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, S.; Fu, J.; Liu, K.; Wang, W.; Lin, K.-Y.; and Wu, W. 2024. CosmicMan: A Text-to-Image Foundation Model for Humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6955–6965.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, X.; Ding, C.; Hong, Z.; Huang, J.; Tao, J.; and Xu, X. 2024. Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7017–7026.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Figure 2: The qualitative results of the ablation study.

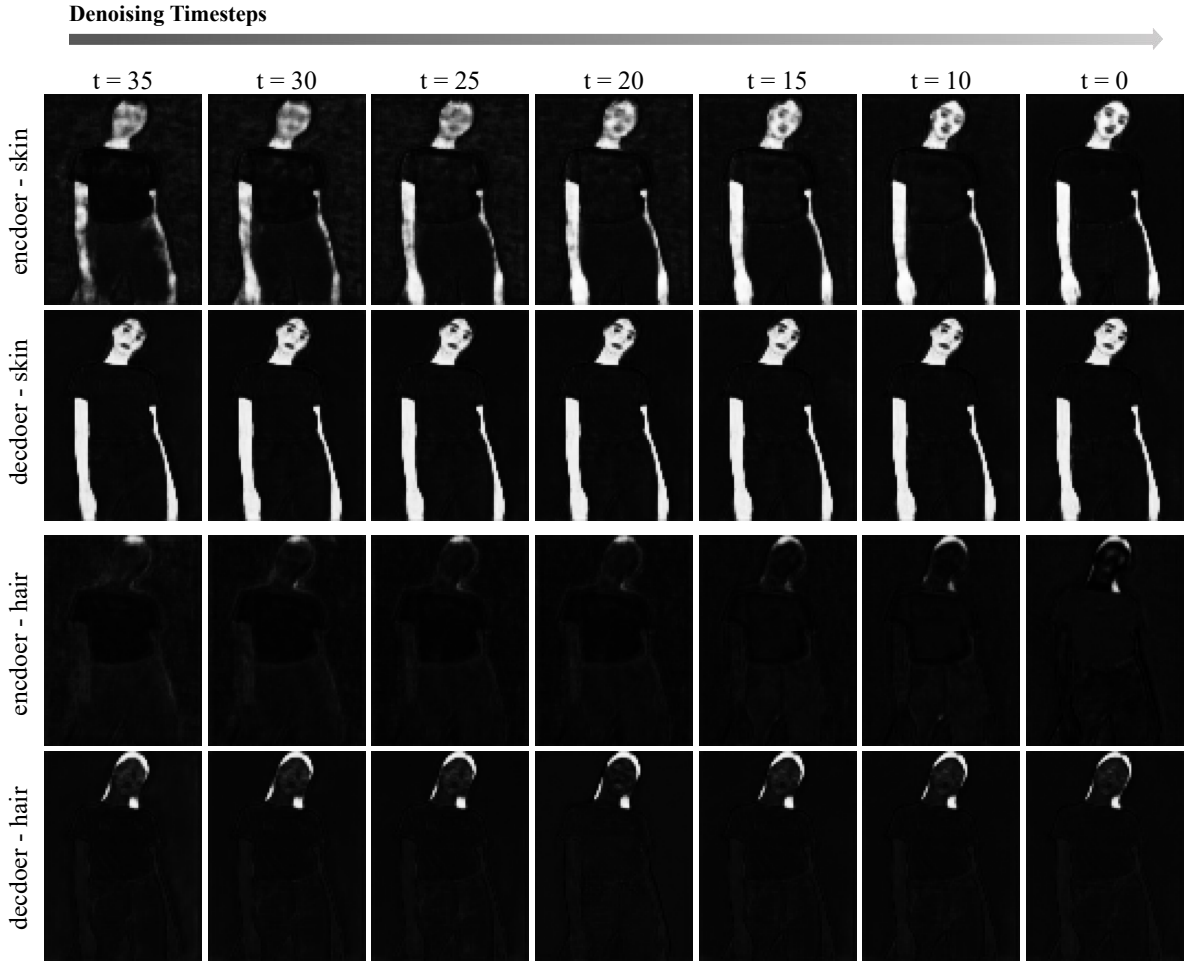

Figure 3: The masks predicted by different parts of the network at different timesteps.

Figure 4: Qualitative comparison with existing methods.

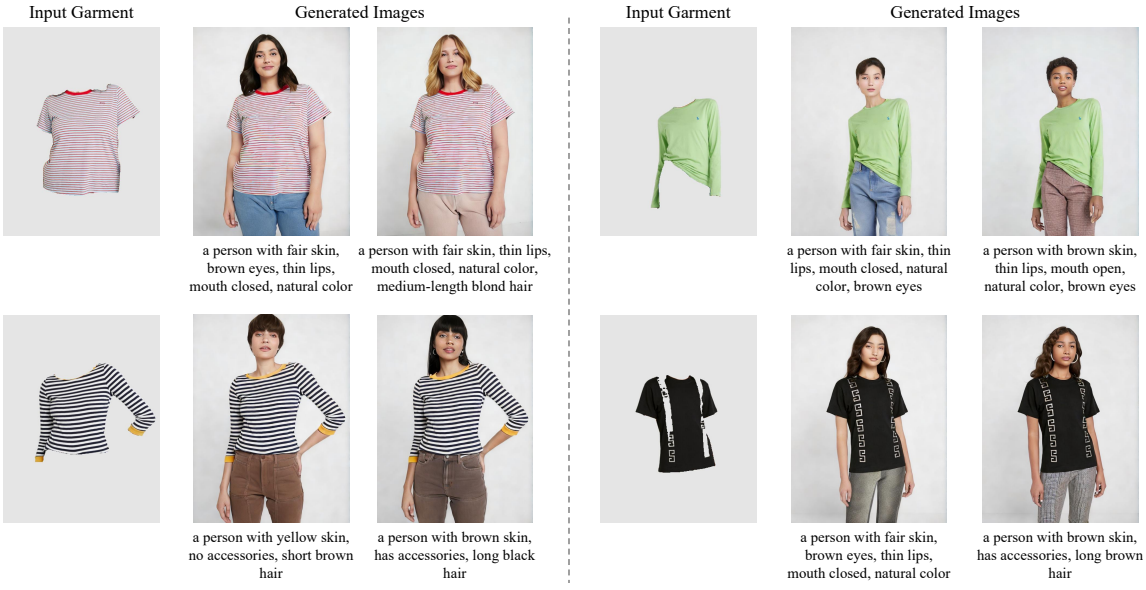

Figure 5: More Results of FashionMAC.

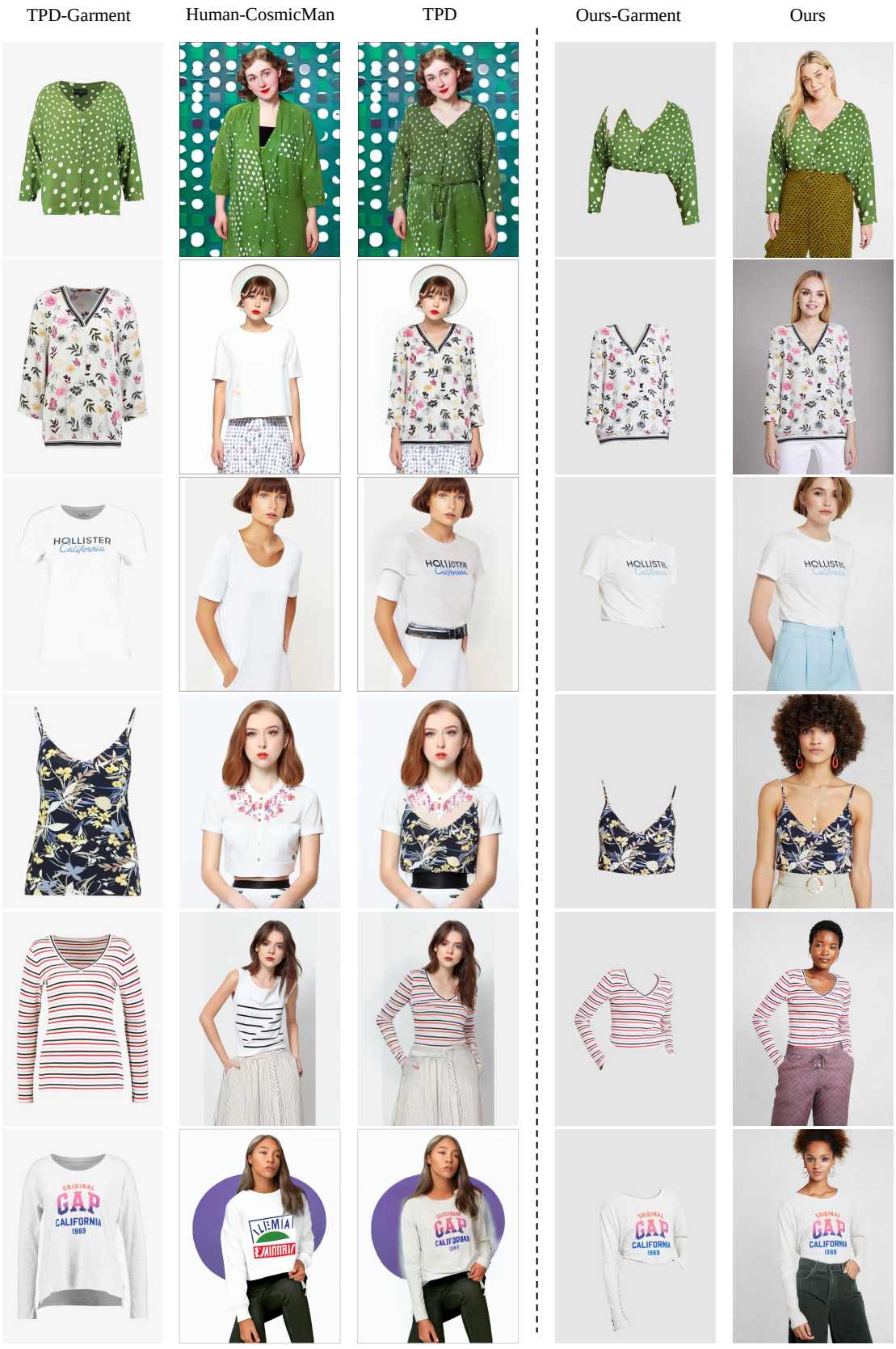

Figure 6: Qualitative comparison with CosmicMan+TPD.

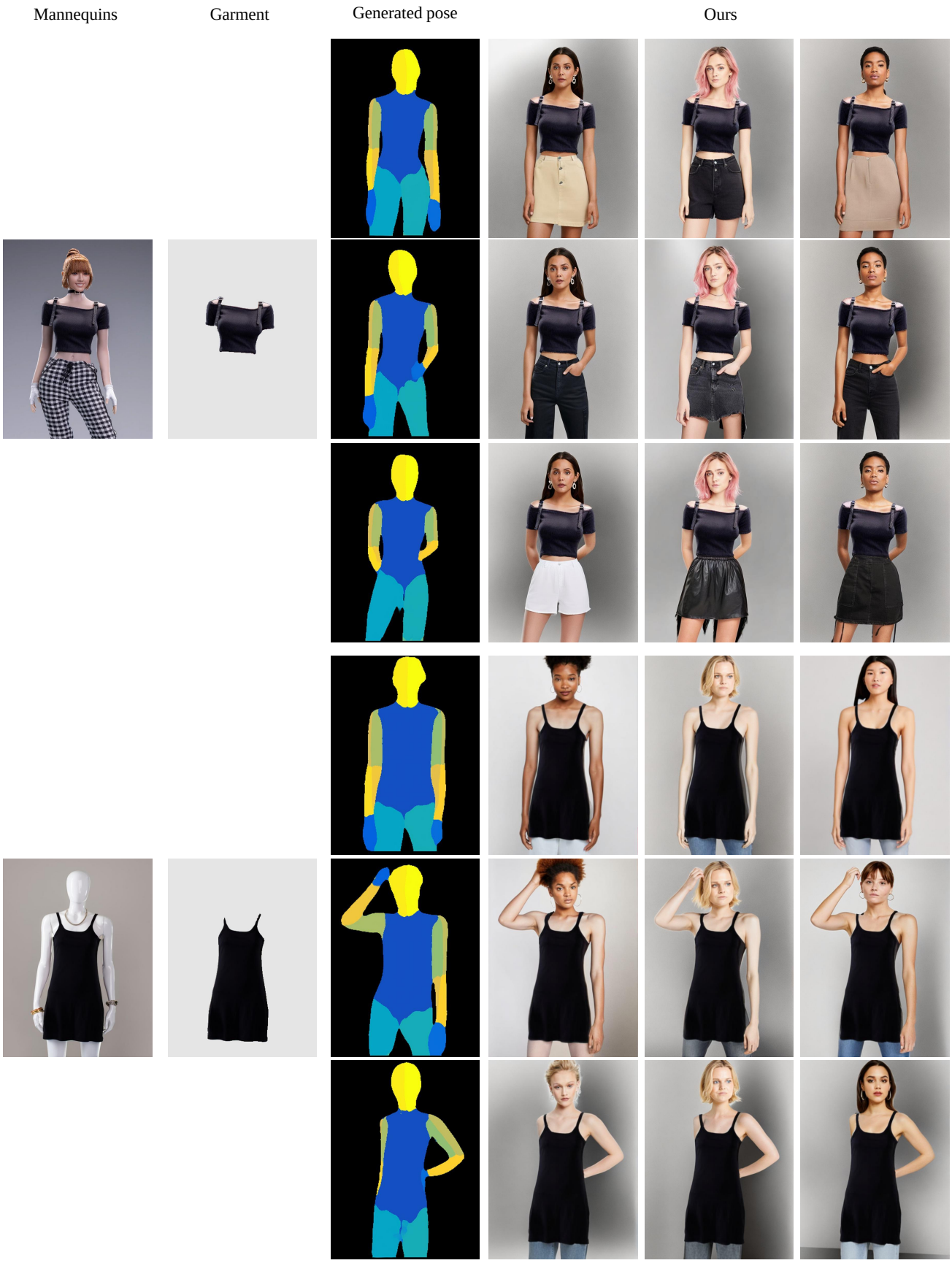

Figure 7: The results of our method conditioned on garments worn by a mannequin.