AnyDressing: Customizable Multi-Garment Virtual Dressing via Latent Diffusion Models

Supplementary Material

In the supplementary material, the sections are organized as follows:

- We provide more details regarding parameters, datasets and user study in Sec. A.
- We further prove the scalability of AnyDressing in Sec. B.
- We provide more ablation results in Sec. C.
- We provide more comparisons with baselines, more qualitative results in the wild and more applications in Sec. D.

A. Implementation Details

A.1. Detailed Parameters

In our experiments, we use SOTA large multi-modal model CogVLM [9] to caption the image. GarmentsNet requires only one step forward process before the multiple denoising steps in DressingNet, causing a minimal amount of extra computational cost. And we generate images for each test garment with the provided 7 text prompts. The hyper-parameters used in our experiments are set as follows:

- For the Dressing-Attention mechanism, we set the hyper-parameter $\lambda = 0.7$ during inference to get customized results.
- For the noisy timestep threshold discussed in the Garment-Enhanced Texture Learning (GTL) strategy, we set $\eta = 350$.
- The other hyper-parameters used in the experiment are as follows: $\lambda_1 = 0.01$, $\lambda_2 = 0.001$.

A.2. Datasets

To facilitate research on multi-garment virtual dressing, a dataset consisting of image triplets is necessary, with each triplet containing an upper garment image, a lower garment image, and a model image wearing the corresponding garments. However, existing in-shop garment to model pairs [2, 6] only contain a single reference garment. We leverage the public DressCode dataset along with a proprietary dataset to construct triplets, as illustrated in Fig. 1. Assuming we begin with the upper garment data, where we already have an in-shop upper garment and a model image wearing it, we employ human parsing techniques [4, 5] to roughly segment and extract the lower garment portion from the model image, using it as the corresponding lower garment image. At this stage, the triplet comprises an in-shop upper garment image, a cropped lower garment image, and a model image. Similarly, triplets derived from the lower garment data consist of a cropped upper garment image, an in-shop lower garment image, and a model image. Finally, we constructed 26,114 public triplets from Dresscode and 37,065 triplets from the proprietary dataset to train AnyDressing.

It is worth noting that our model has not encountered garment pairs in the form of (in-shop upper garment, in-shop lower garment) or (cropped upper garment, cropped lower garment) during training. Nevertheless, it exhibits strong robustness during inference, indicating that the model has effectively learned the proper way to combine upper and lower garments through training.

A.3. User Study

To compare with the baseline methods, we conduct a user study as part of the evaluation. The survey randomly presented 50 sets of generated results to each participant. A screenshot of the survey for a set of generated results is displayed in Fig. 2, which includes five images and four questions:

1. *Which result appears to have the highest consistency with reference garments?*
2. *Which result best matches the prompt '[prompt]'?*
3. *Which result appears to have the highest image quality?*
4. *Which result matches your best choice based on comprehensive considerations?*

For each set of results displayed in the survey, we ensured that their order was randomly shuffled to prevent bias. Responses where all answers had the same selection and responses with completely identical answers were considered invalid. Finally, we obtained a total of 40 valid surveys to evaluate the model.

B. Scalability of AnyDressing

To further validate the scalability of our designed GarmentsNet structure, we introduce more combinations of clothing items (hat, upper garment and lower garment), as illustrated in Fig. 4. As shown in Fig. 3, to train the model, we construct datasets using the same idea as introduced in Sec. A.2. Specifically, we select 18,059 pairs from the proprietary dataset that satisfies the model image containing the hat, and use the human parsing techniques to obtain the cropped hat image from the model image.

Notably, each additional garment condition requires only some newly added LoRA matrix $\triangle\hat{\mathbf{W}}$ in the Garment-Specific Feature Extractor (GFE) module. And it requires only a single forward pass (timestep $t = 0$) to encode the clothing before injecting features into the DressingNet, minimizing the additional computational time during both

the training and inference process. This experiment effectively demonstrates that our GarmentsNet can be extended to accommodate any number of clothing items. Additionally, thanks to our proposed Instance-Level Garment Localization (IGL) learning mechanism, AnyDressing can further prevent garment blending and enhance fidelity to customized text prompts.

C. More Ablation Study

In Fig. 5, we present additional visual results to validate the effectiveness of the Garment-Specific Feature Extractor (GFE) module and the Instance-Level Garment Localization (IGL) learning mechanism. We employ traditional ReferenceNet [3] to encode multiple garments concurrently and then incorporate them into the denoising U-Net similar to [1, 7] as our base model. As shown in Fig. 5, **Base** model encounters severe clothing confusion issues, resulting in the colors and patterns of multiple garments blending. In contrast, **Base+GFE** significantly reduces garment confusion and improves garment consistency, which is attributed to the multi-garment parallel processing design of our designed GFE module. **Base+GFE+IGL** shows better fidelity to the text prompts and further mitigates background contamination, which demonstrates IGL mechanism effectively constrains garment features to attend to the correct regions and avoid influencing other irrelevant regions in the synthetic images.

D. More Results

D.1. More Comparisons

As shown in Fig. 6-7, We provide more visual comparisons between our method and state-of-the-art baselines [1, 7, 8, 10]. It is clear from these comparisons that our method maintains superior consistency in clothing style and texture, and exhibits better text fidelity.

D.2. More Visual Results

As shown in Fig. 8-10, we provide more multi-garment virtual dressing results of AnyDressing in the wild. It can be observed that our method produces high-quality customized virtual dressing results for various types of garment combinations, while faithfully adhering to personalized text prompts. Experiments in complex scenarios demonstrate that AnyDressing significantly enhances the practical application of Virtual Dressing in e-commerce and creative design.

D.3. More Applications

Combined with ControlNet. Leveraging the capabilities of ControlNet, our model can generate personalized models guided by specific conditions. We present the OpenPose-guided generation results in Fig. 11.

Combined with IP-Adapter. Our model enables the generation of target individuals wearing specified garments integrated with the IP-Adapter. We utilize the ID preservation capability of FaceID [10] to provide an authentic virtual dressing experience. The visual results, as shown in Fig. 11.

Stylized Customization. Furthermore, by utilizing stylized base models or customized LoRAs, we can generate creative and stylized outputs while preserving the intricate details of the garments, as shown in Fig. 10 and Fig. 12.

References

[1] Weifeng Chen, Tao Gu, Yuhao Xu, and Chengcai Chen. Magic clothing: Controllable garment-driven image synthesis. *arXiv preprint arXiv:2404.09512*, 2024. 2

[2] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 1

[3] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2

[4] Zhenchao Jin. Sssegmenation: An open source supervised semantic segmentation toolbox based on pytorch. *arXiv preprint arXiv:2305.17091*, 2023. 1

[5] Zhenchao Jin, Xiaowei Hu, Lingting Zhu, Luchuan Song, Li Yuan, and Lequan Yu. Idrnet: Intervention-driven relation network for semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[6] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022. 1

[7] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinghui Tang. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*, 2024. 2

[8] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stablegarment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024. 2

[9] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[10] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 11

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 11

Figure 1. Examples of the **training dataset I**.

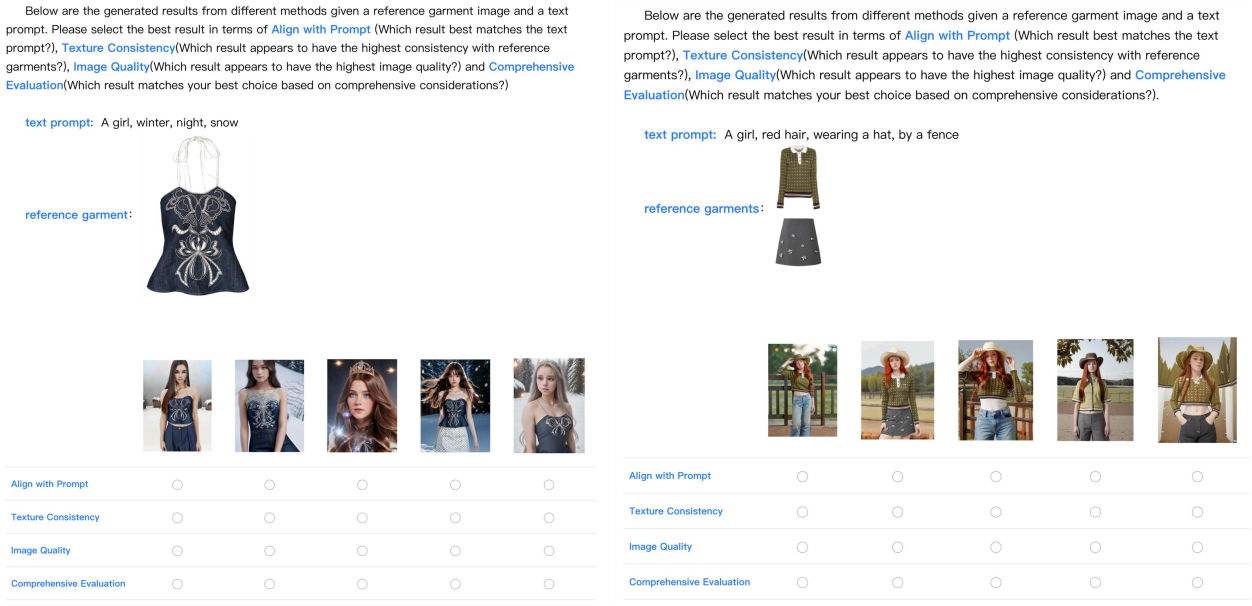

Figure 2. Screenshot of **user study**.

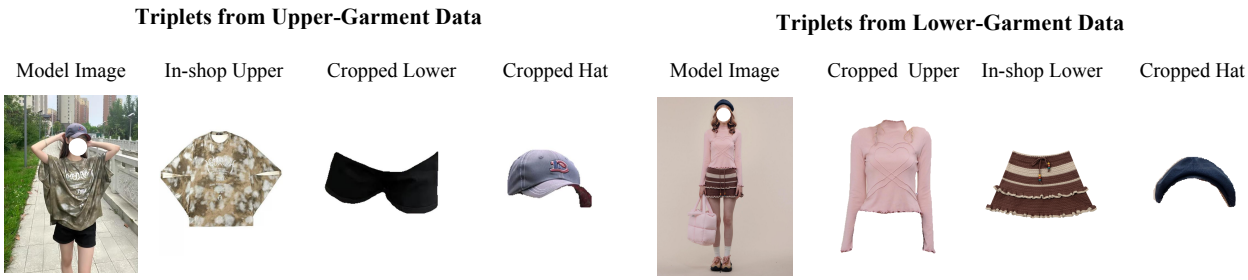

Figure 3. Examples of the **training dataset II**.

Figure 4. Qualitative results of **more combinations of clothing items**.

Figure 5. More **ablation results on GFE and IGL modules**.

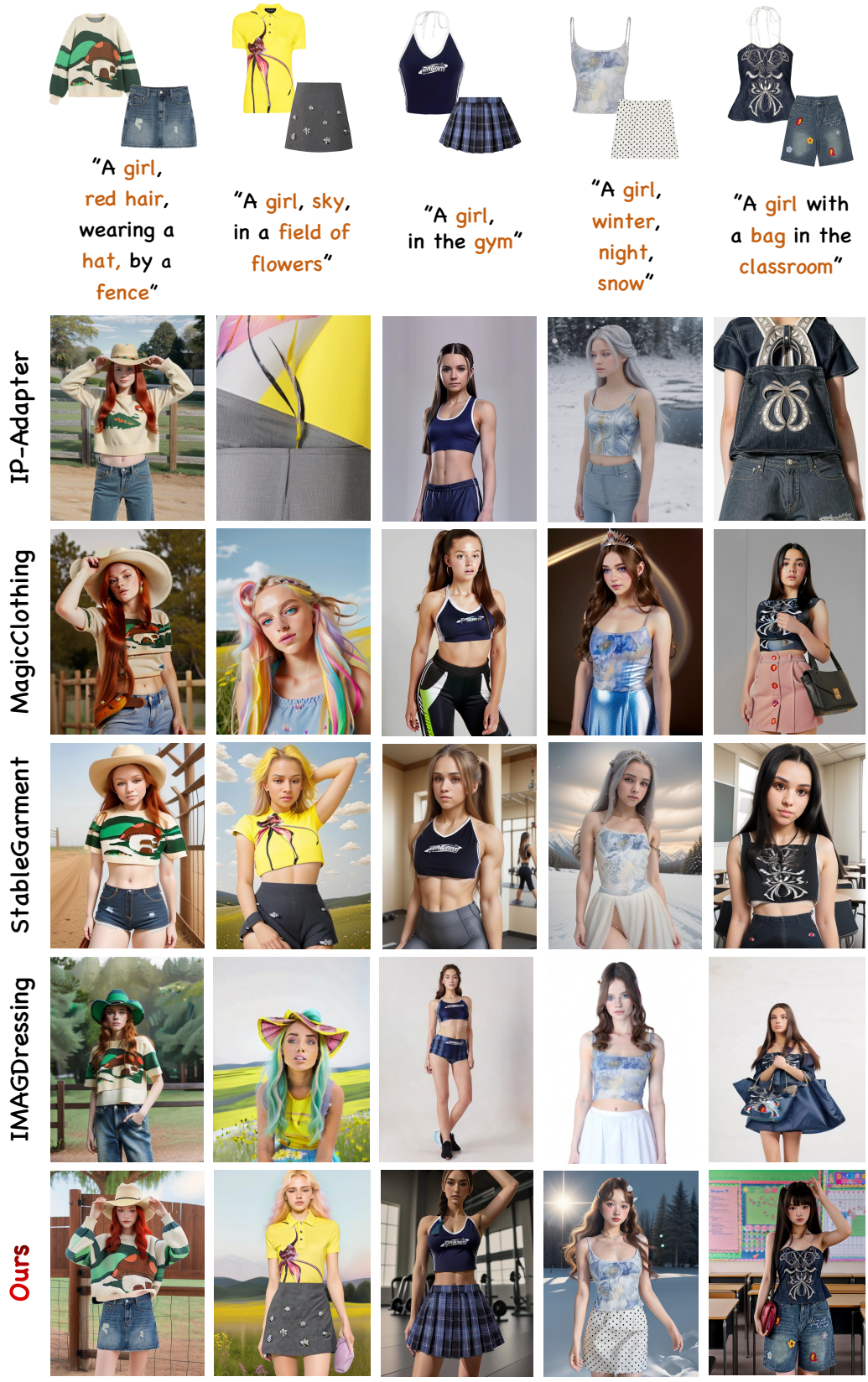

Figure 6. More **qualitative comparisons I**.

Figure 7. More **qualitative comparisons II**.

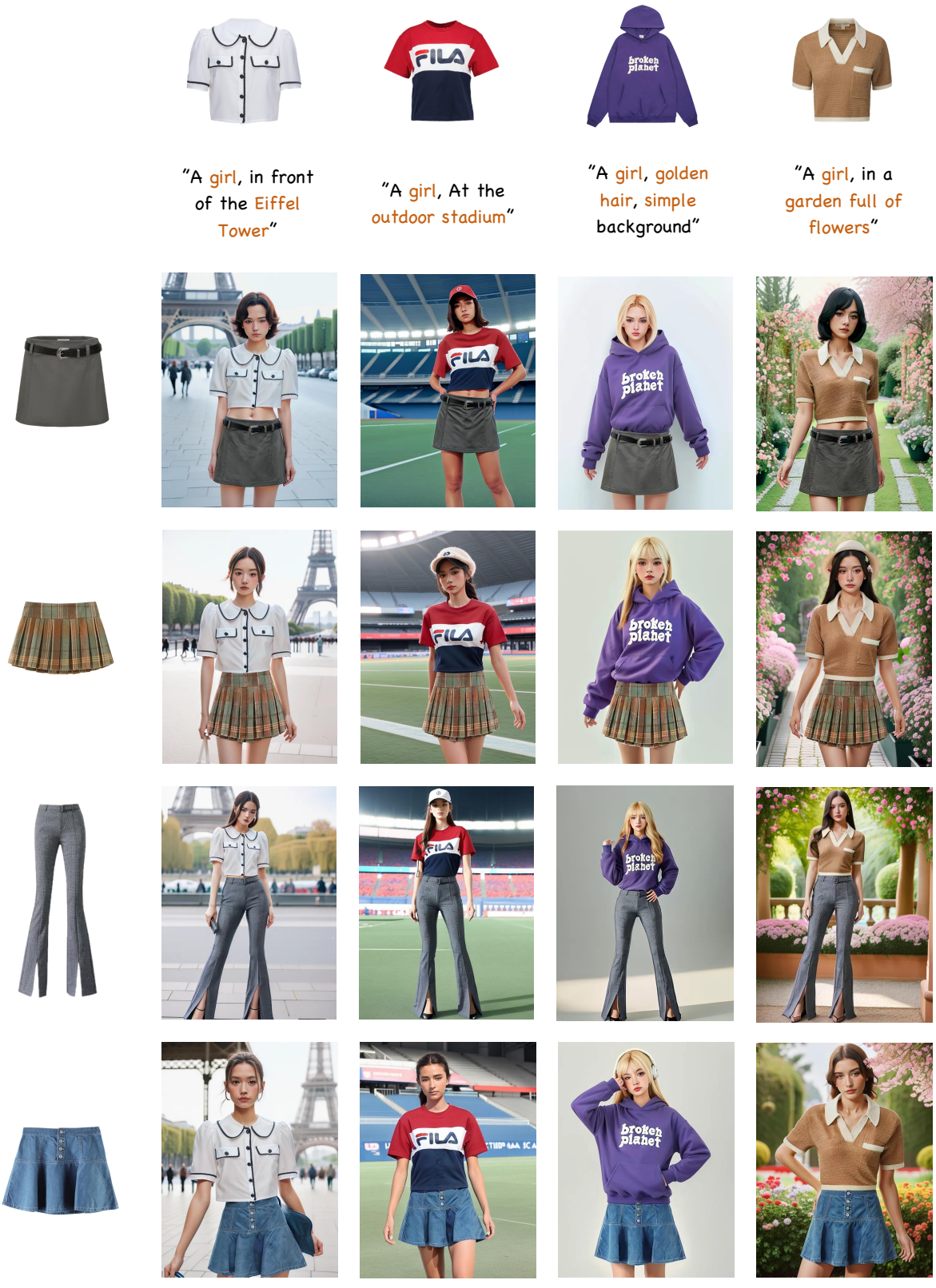

Figure 8. More **qualitative results I**.

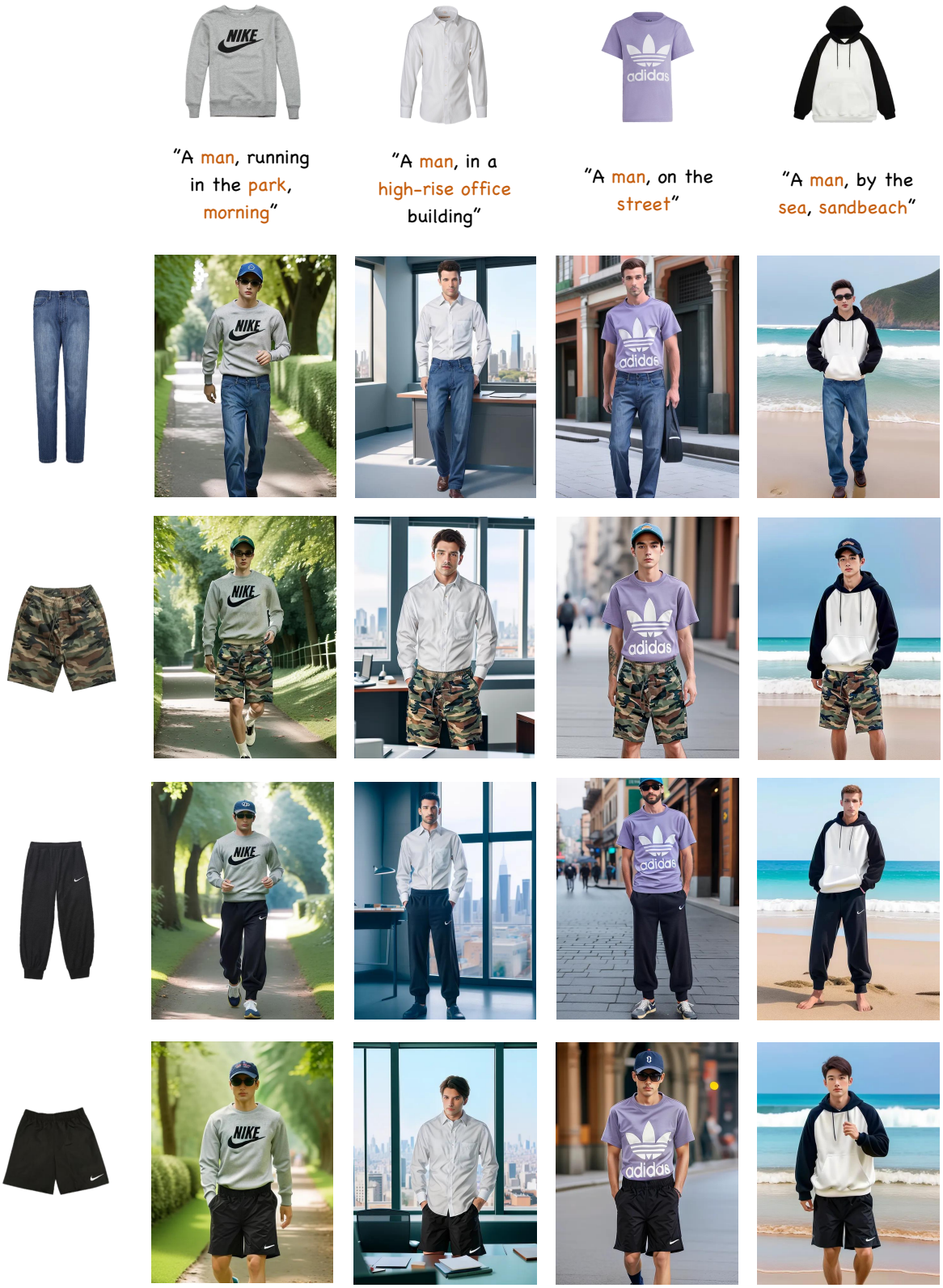

Figure 9. More **qualitative results II**.

Figure 10. More **qualitative results III**.

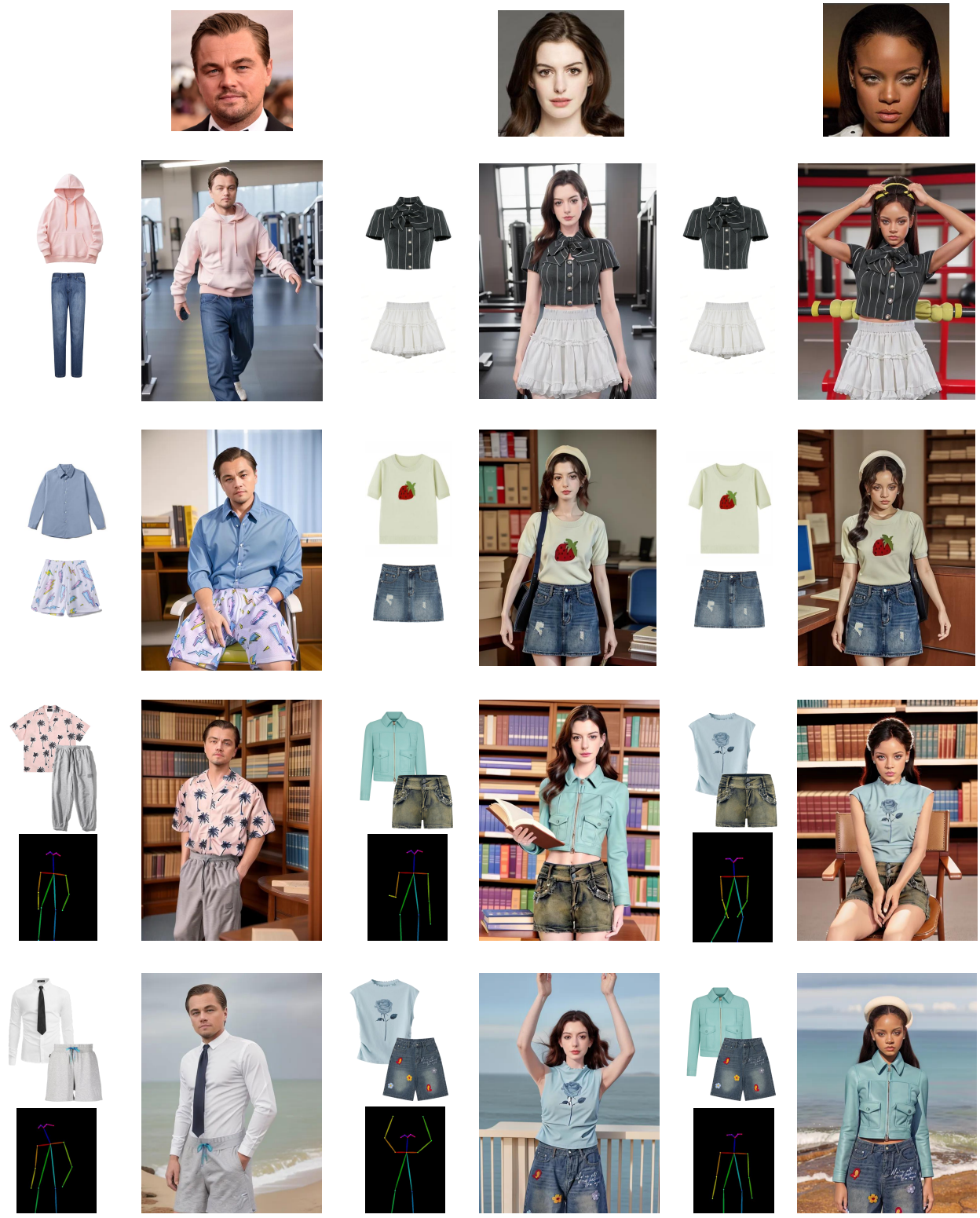

Figure 11. More results of **combining ControlNet [11] and FaceID [10]**.

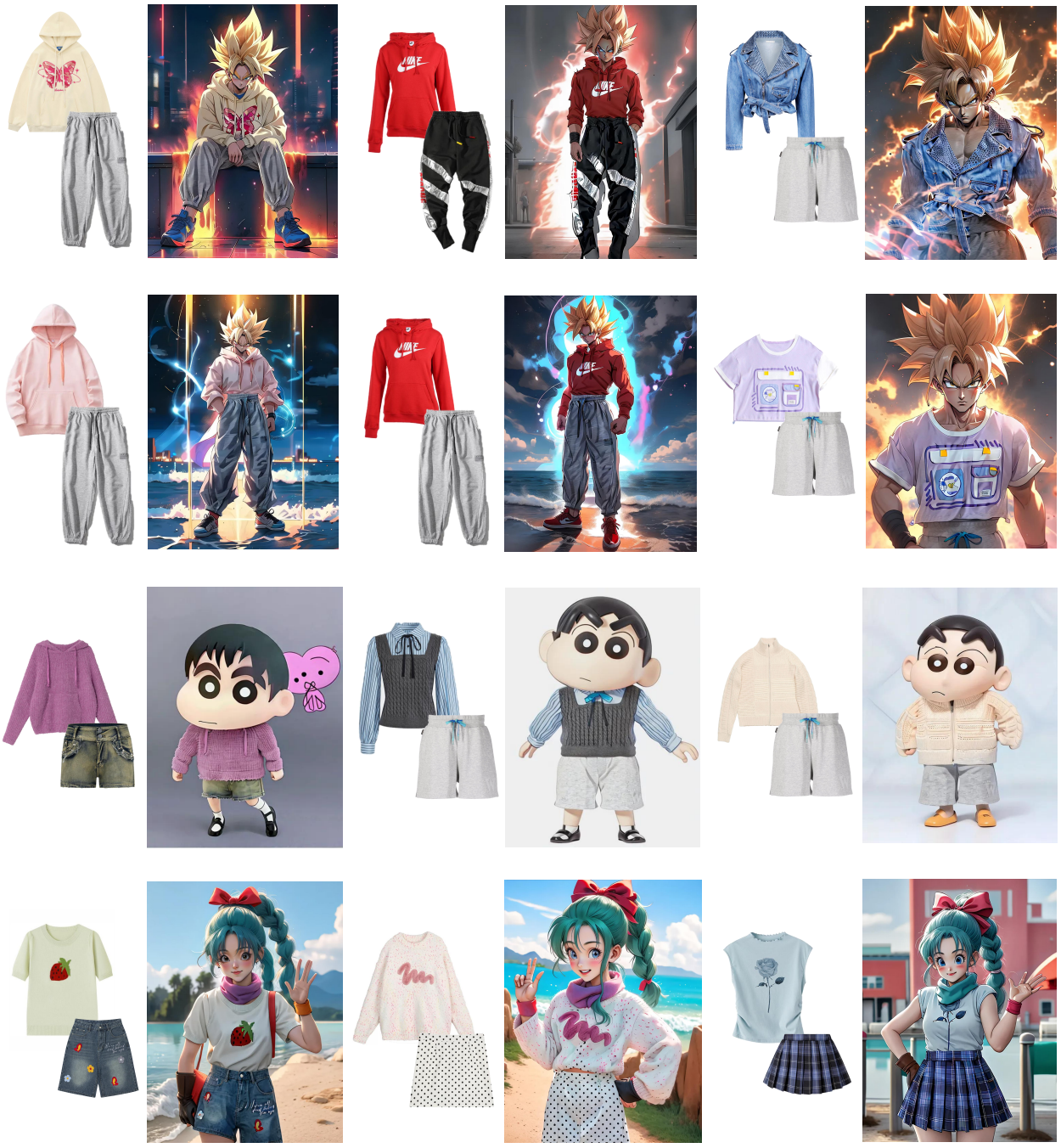

Figure 12. More results of **combining LoRAs**.