MOFA-VTON: More Fashion Possibilities with Fine-Grained Adaptations in Virtual Try-On

Supplementary Material

A. Implementation Details

In our experiments, we initialize the weights of Cloth-Net and Adapt-Net with the pre-trained weights of Paint-by-Example [4], and the weights of the CLIP [3] model are taken from the version of ViT-L/14. Our model is trained using the AdamW optimizer [2] with a learning rate of 1e-5. The training is conducted on paired images with a resolution of 512×384, and we adopt a batch size of 8 throughout the training process. For inference, the whole virtual try-on pipeline can be executed in approximately 5.7 seconds when running on a single NVIDIA A100 GPU.

B. User Study Details

To provide a more comprehensive evaluation of our proposed MOFA-VTON, we conduct a detailed user study focusing on diversity. In this study, diversity is further divided into four items: fitness, usability, functionality, and fineness. Specifically, fitness is designed as a purely subjective metric to assess users' general preference for diverse try-on results without being influenced by other factors, which measures whether the result overall meets user expectations. Usability evaluates the ease of operation and user interaction of each method. Therefore, no visual results are shown for this criterion, where participants are presented with textual descriptions of the interaction workflows and asked to select the method they considered more convenient and intuitive. Functionality evaluates whether additional or extended functions are achieved, where we conduct a region-wise evaluation by asking participants to score whether each method supports controllable try-on in four body regions (waist, arms, torso, and legs). Fineness assesses the level of fine-grained control achieved, where we explicitly visualize the interaction traces on the try-on results (e.g., control points before and after manipulation for COTTON [1], input curves for MOFA-VTON), allowing participants to judge the precision and granularity of layout control based on these visual cues.

C. Limitations

Since the training data for our model is entirely collected from real-world settings under normal and practical conditions, MOFA-VTON learns to predict clothing adaptations that conform to realistic scenarios. Therefore, when presented with unusual or extreme input curves, the generated output may not fully match the intended layouts. As illustrated in Figure 1, for certain clothing categories such as short-length tops or bodysuits, input curves positioned too low or too high provide ineffective guidance for consistent clothing adaptation.

Figure 1. MOFA-VTON struggles to accurately adjust clothing adaptation based on curves positioned at unrealistic locations.

Figure 2. MOFA-VTON enables adjustment of long sleeve length.

D. More Results

Additional Ablation Study. In the design of the region encoder, we regard the CLIP text embedding as a complementary enhancement, which is introduced as a high-level semantic prior to guide the layout prediction process. By

Figure 3. More qualitative results generated by MOFA-VTON.

Table 1. Quantitative results of ablation studies on the CLIP text embedding.

Method	Paired				Unpaired	
	FID (↓)	KID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)	KID (↓)
MOFA-VTON w/o CLIP text embedding	6.02	0.98	0.8861	0.0633	8.77	1.20
MOFA-VTON	**5.97**	**0.92**	**0.8870**	**0.0632**	**8.61**	**1.17**

incorporating multi-modal inputs, the network is able to more effectively capture spatial arrangements and clothing semantics. To assess its contribution, we construct an additional variant that ablates the CLIP text embedding from the region encoder. As shown in Table 1, there is a slight performance degradation when removing the text embedding, and we also observe that the training process became less stable, especially in the early stages. These results indicate that the text encoder benefits both overall performance and convergence stability.

Additional Visual Displays. To further demonstrate the fine-grained control capability of our MOFA-VTON, we pair each person image with various clothing options and assign a fixed hand-drawn curve to each group, generating multiple results with consistent clothing adaptations within the same group. As shown in Figure 5, MOFA-VTON effectively fits the target clothing of varying appearances to the human body and accurately adapts them according to the input curves, further highlighting its strong performance and robustness. In addition, more qualitative results are also provided in Figure 3.

Additional Function. MOFA-VTON also supports virtual try-on for other types of clothing such as pants and skirts. This can be achieved by modifying the mask expansion step of the mask construction strategy, where the expansion is performed from the highest point along the y-axis instead of the lowest. Therefore, we conduct multi-item try-on experiments, which are presented in Figure 4. These results demonstrate the generalization capability and adaptability of MOFA-VTON across diverse clothing categories. Moreover, although MOFA-VTON primarily focuses on the interaction between tops and bottoms, we are pleasantly surprised to discover that by slightly modifying the arm-occlusion mask, MOFA-VTON can further support adjusting the sleeve length of clothing, as shown in Figure 2. We speculate that this effect arises from the model learning strong control over clothing boundaries in the training phase, enabling it to generate realistic details (e.g., wrinkles) around these regions, and this capability generalizes to other regions as well.

Video Display. To better demonstrate the effectiveness and advantages of our method, we provide a carefully crafted video in our supplementary material, which outlines the key contributions and showcases detailed experimental results.

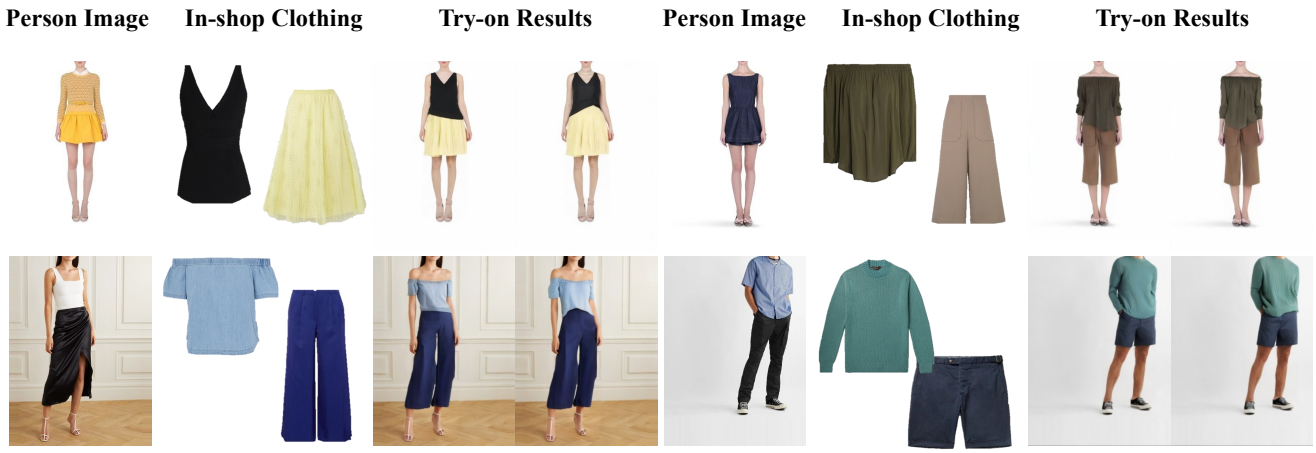

Figure 4. Multi-item try-on results generated by MOFA-VTON.

Figure 5. Each person image is paired with various clothing options to generate try-on results with consistent clothing adaptations.

References

[1] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7513–7522, 2023. 1

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–19, 2019. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1

[4] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023. 1