Supplementary Material for IMAGDressing-v1: Customizable Virtual Dressing

Fei Shen[1] , **Xin Jiang**[1] , **Xin He**[2], **Hu Ye**[3], **Cong Wang**[4], **Xiaoyu Du**[1], **Zechao Li**[1], **Jinhui Tang**[1*]

[1]Nanjing University of Science and Technology
[2]Wuhan University of Technology
[3]Tencent AI Lab
[4]Nanjing University

Supplementary Material

This supplementary material serves as an extension to the main paper, *IMAGDressing-v1: Customizable Virtual Dressing* (Shen et al. 2025). It provides additional insights into our methods, datasets, and experimental evaluations. We begin by detailing the implementation setup used in our experiments. We then offer a comprehensive introduction to the proposed IGPair dataset. In addition, we present extended experimental results, including a user study and further evaluations of IMAGDressing-v1 under various conditions. Lastly, we discuss ethical considerations and the broader societal impacts of our work. Our code and pre-trained models are publicly available at https://github.com/muzishen/IMAGDressing to facilitate future research and reproduction.

Implement Details

In our experiments, we initialize the clothing UNet's weights by inheriting the pre-trained UNet weights from Stable Diffusion v1.5 (Rombach et al. 2022) and fine-tune them. We follow the standard training strategies and hyperparameters for diffusion models. We use OpenCLIP ViT-large/14[1] and OpenCLIP ViT-bigG/14[2] as the image and text encoders. Our model is trained on paired images from the IGPair dataset at a resolution of 512×640. We employ the AdamW optimizer with a fixed learning rate of 5e-5, and the dropout probability for the condition c is set to 10%. The model is trained for 200,000 steps on 10 NVIDIA RTX 3090 GPUs with a batch size of 5. We use a constant warm-up for the first 2000 steps and accelerate training with DeepSpeed. During inference, we use the UniPC sampler to generate images, performing 50 sampling steps, and set the guidance scale w to 7.0.

Metric. We propose a comprehensive affinity metric index (CAMI) for evaluating virtual dressing (VD) tasks, which consists of the unspecified score (CAMI-U) and the specified score (CAMI-S). The reason for introducing CAMI-U and CAMI-S is to provide a more holistic and flexible evaluation framework for various VD scenarios.

CAMI-U evaluates image generation in scenarios where pose, face, and text are unspecified. It focuses on fundamental aspects of the generated clothing images, such as structure (S_s), texture (S_t), and keypoints (S_k). This metric is crucial for assessing the general quality and consistency of clothing items independent of specific contextual factors.

CAMI-S, on the other hand, extends CAMI-U by incorporating additional context-specific factors such as pose matching degree (S_p), facial similarity (S_f), and text-image matching degree (S_c). These additional metrics are essential when the generated images need to align with specific poses, faces, or text descriptions, which is critical for more personalized and contextually accurate virtual dressing applications.

This dual-metric approach ensures that the evaluation framework can handle both general and specific VD tasks, making it versatile and robust. Detailed implementation can be found in the code, which provides further insights into the metric calculations and their applications.

IGPair Dataset

The IGPair dataset is designed to meet the specific requirements of virtual dressing (VD) tasks. To ensure its broad applicability in research, we have made the dataset publicly accessible and easy to use. High-resolution images are crucial for capturing the details of garments and accurately generating virtual dressing effects; hence, we have paid special attention to image quality. Additionally, since virtual dressing tasks involve various scenes and styles, we have included diverse scenarios accompanied by detailed textual descriptions, allowing models to understand and generate virtual dressing effects in different styles. Compared to existing datasets, IGPair not only offers a larger number of image pairs but also includes multiple models wearing the same garment, providing more training data for model generalization. Moreover, the ultra-high resolution (exceeding $2K \times 2K$) images enhance the generated details, while the textual descriptions further improve the model's performance in understanding and generating clothing ensembles.

Source and Diversity: We collected images from the internet, covering various clothing styles such as casual, for-

*Corresponding Author.

[1]https://huggingface.co/openai/clip-vit-large-patch14
[2]https://huggingface.co/laion/CLIP-ViT-bigG-14-laion2B-39B-b160k

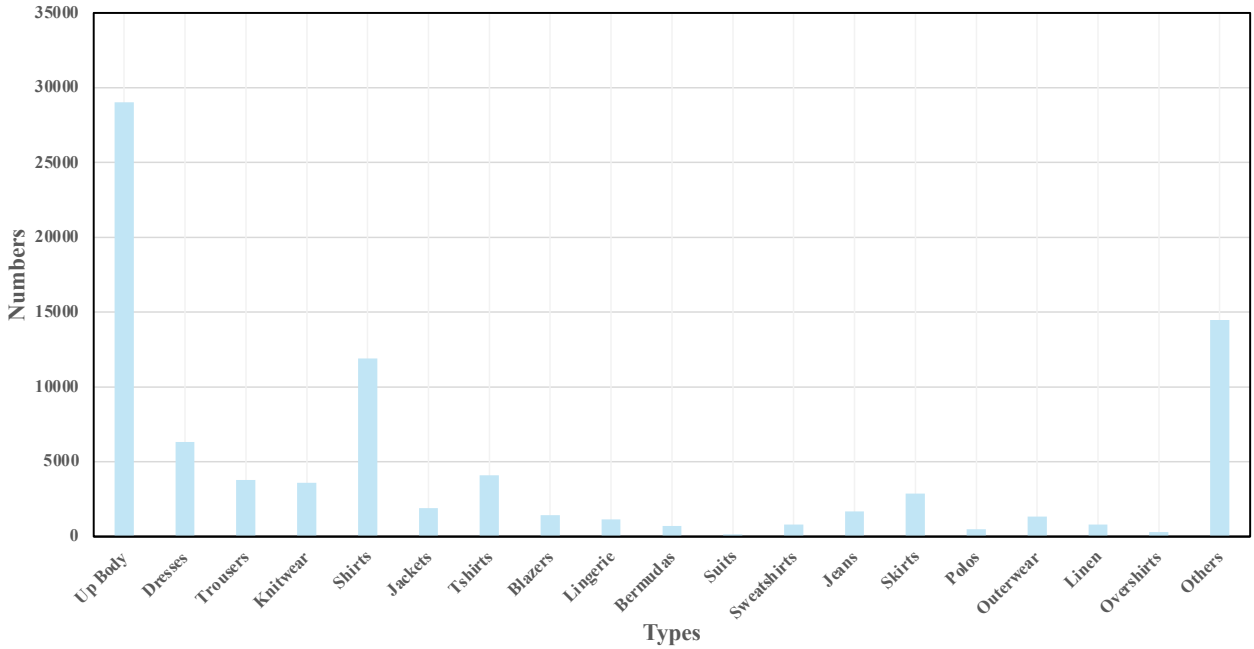


Figure 1: Distribution of images across different clothing types in the IGPair dataset.

mal, athletic, fashionable, and vintage. This ensures that the dataset is widely applicable and that models can adapt to different dressing scenarios and styles.

Image Quantity and Angles: We gathered 86,873 garment images, each paired with 2 to 5 images of models wearing the clothing from different angles. These multi-view images help models better understand and process the appearance of garments from various perspectives, enhancing the realism of virtual dressing effects. The total number of image pairs in the dataset is 324,857.

Automated and Manual Filtering: Initially, we used classifiers to differentiate between clothing and human model images, followed by a human pose estimator to select complete and usable images of models wearing the clothing. This automated filtering process significantly reduced manual labor while ensuring the basic quality of the data. Subsequently, manual verification ensured the final dataset's quality, guaranteeing the accuracy and consistency of all images.

Categorization and Annotation: As shown in Figure 1, we categorized the garments into 18 types to better control and train models on different clothing types. This classification method helps the models perform more accurately and consistently when generating specific types of clothing ensembles.

Additional Information Extraction: To further enrich the dataset, we used OpenPose (Cao et al. 2017) to extract key points of the human figure, DensePose(Güler, Neverova, and Kokkinos 2018) to compute dense poses for each reference model, and SCHP (Li et al. 2020) to generate segmentation masks for body parts and clothing items. This additional information is vital for enhancing the model's understanding of human poses and clothing details, making the virtual dressing effects more realistic and natural.

Text Description Generation: We used advanced models such as BLIP2-OPT-6.7B (Li et al. 2023), INTERNLM-XCOMPOSER2-VL-7B (Dong et al. 2024), LLaVA-V1.5-13B (Liu et al. 2024), and Qwen-VL-Chat (Bai et al. 2023) to generate detailed textual descriptions for the images. These textual descriptions further enhance the model's understanding of the garments and scenes, allowing it to better match the scene and clothing characteristics when generating virtual dressing effects.

Anonymization: All model images have been anonymized to ensure the dataset's privacy and security. This measure not only complies with ethical requirements but also ensures the legality and usability of the data. Through these steps, the IGPair dataset ensures high quality, diversity, and applicability in virtual dressing tasks, providing robust support for future research and model development.

User Study

The quantitative and qualitative comparisons in the main text demonstrate the significant advantages of IMAGDressing-v1 in generating results. Additionally, we conducted a user study involving 50 volunteers, where each participant answered 100 questions, with each question consisting of three sub-questions. Specifically, participants were asked to evaluate which of two images (one generated by the baseline model and the other by IMAGDressing-v1) performed better in the following aspects: 1) fidelity, 2) clothing identity, and

Table 1: Results of user study.

Metric	Fidelity (↑)	Clothing identity (↑)	Scene quality (↑)
Baseline	5.4	8.3	6.2
Ours	94.6	91.7	93.8

3) scene quality. The criteria for each aspect were as follows: As shown in Table 1, IMAGDressing-v1 overwhelmingly outperforms in human evaluations, demonstrating a strong preference in terms of fidelity, clothing identity, and scene quality.

- **Fidelity**: Choose the image that more closely resembles reality in terms of human form and color harmony.
- **Clothing identity**: Choose the image that better preserves the design, markings, and shape characteristics of the input clothing.
- **Scene quality**: Choose the image that better maintains consistency with the input text in terms of the scene.

Additional Results

In this section, we present the testing results of IMAGDressing-v1 across a variety of scenarios. As shown in Figure 2, we display the generated results for clothing with subtle logos and human faces. The results demonstrate that IMAGDressing-v1 can accurately reproduce logos with high fidelity. In Figure 3, the model's ability to handle more complex text-based jerseys is showcased, further highlighting its strong performance. Figure 4 illustrates the model's capability to control different scenarios through textual prompts. Figure 5 demonstrates its ability to simultaneously control face, pose, and clothing attributes. Finally, Figure 6 showcases the robustness of IMAGDressing-v1 in virtual dressing applications for cartoon characters. These results collectively emphasize the versatility and effectiveness of the model across various challenging use cases.

Limitation and Future Work

Artifact Issues in Detail Generation. Our method may produce artifacts when generating fine details, particularly in areas such as hands, especially when the hands occupy only a small portion of the image. This limitation can be addressed by using higher resolution and larger diffusion models, such as SD-XL. Moreover, incorporating localized refinement modules or hand-specific priors could further enhance detail fidelity in these challenging regions.

Challenges in Preserving Human Attributes. Preserving human attributes, such as facial features or posture, in occluded areas presents a significant challenge. To address this, methods can be designed to adjust human attributes during the generation of try-on images. Future work may leverage multi-view supervision or pose-conditioned priors to improve consistency in occluded or partially visible regions.

Initial Implementation of Text Control. While the method performs well in scene generation applications, its text control functionality is still in its early stages and not fully developed. Future work will explore the potential of enhancing text control through the use of large language models (LLMs). Aligning multimodal embeddings and jointly optimizing visual-language objectives may enable more precise and semantically rich control.

Ethics and Broader Impacts

This paper proposes a virtual dressing task and method that combines fixed clothing guidance with optional face and pose information. While this technology has the potential for misuse, such as generating misleading content, this risk is common to all character generation technologies. However, current research on identifying and preventing malicious attacks has made significant progress, which helps mitigate these risks. Our work provides crucial support for further research and external auditing in this field, while also balancing the value of the technology with the potential risks of open access. By carefully weighing the benefits of the technology against the risks of unrestricted access, we are committed to ensuring the safe and beneficial use of IMAGDressing-v1.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420*.

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.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3260–3271.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

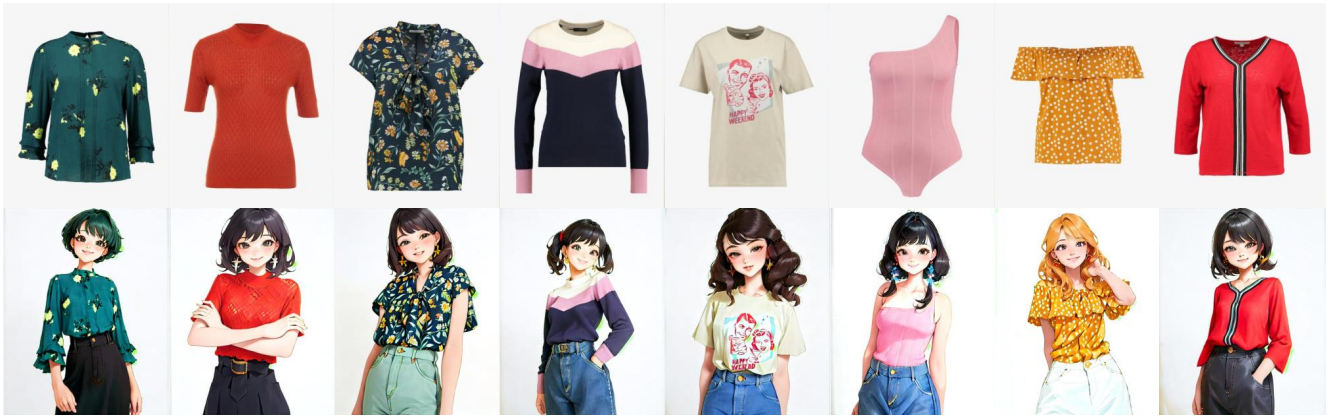
Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Figure 2: More results of IMAGDressing-v1 synthesizing person images given garment with logos and optional faces.

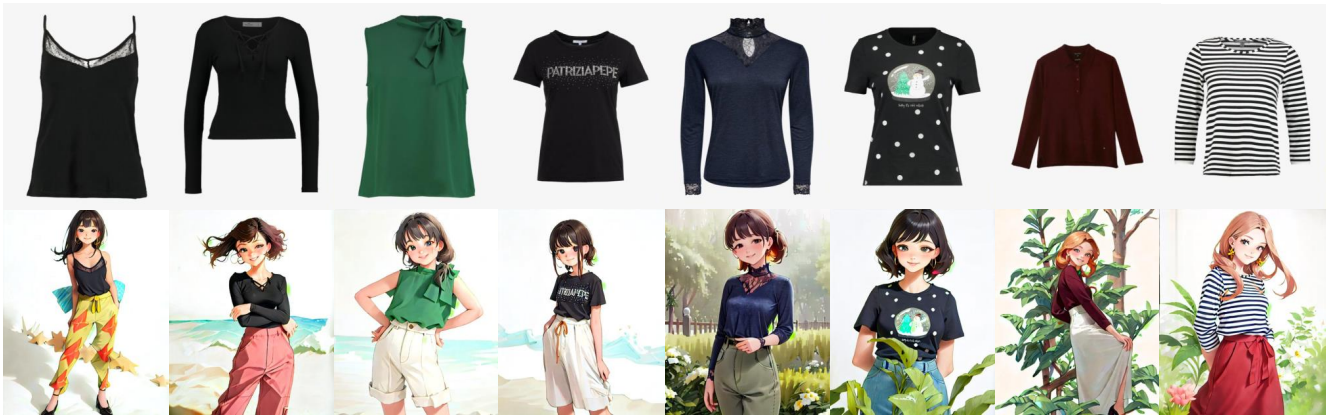
Figure 3: More results of IMAGDressing-v1 generating person images with complex logos (such as text and dense letters) on garment and optional faces.

Figure 4: More results of IMAGDressing-v1 synthesizing person images based on clothing and different text prompts.

Figure 5: More results of IMAGDressing-v1 synthesizing person images given garment with logos, and optional faces and pose.



A smiling woman



A young girl on the beach

A beautiful woman in the garden

Figure 6: More results of IMAGDressing-v1 synthesizing cartoon images based on clothing and different text prompts.