Any2AnyTryon: Leveraging Adaptive Position Embeddings for Versatile Virtual Clothing Tasks

Supplementary Material

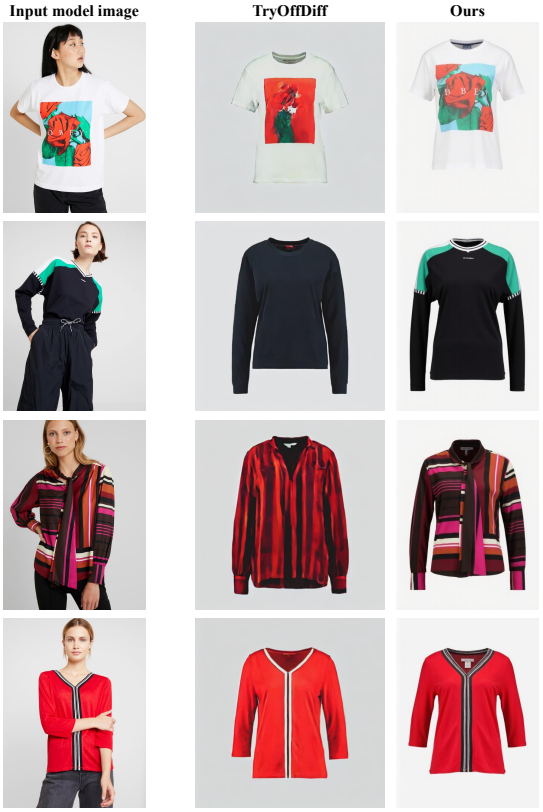

Figure 1. Additional qualitative comparison of garment reconstruction.

In this Supplementary Material, we provide the details of evaluation metrics in Section. 1, and in Section. 3, we provide more visualization of Any2AnyTryon generation results.

1. Validation Metrics

1.1. Garment Reconstruction

We follow the existing SOTA garment reconstruction method TryOffDiff [17] and leverage the full-reference metrics SSIM [18], MS-SSIM, and CW-SSIM to validate the alignment between the reconstructed garment and the ground truth, while utilizing the metrics LPIPS [19], FID [10], CLIP-FID, KID [2], and the Deep Image Structure and Texture Similarity (DISTS) [8] to evaluate the quality and fidelity of the generated images.

1.2. Virtual Try-on Generation

Our Any2AnyTryon supports both model-free VTON and VTON generation. In the experimental section, we conduct a quantitative comparison of both tasks. For model-free VTON generation, we follow MagiClothing [4] and use MP-LPIPS and CLIP-I to measure the consistency of the garments with the generated outfitted model results. To make our quantitative comparison more compelling, we introduce more recent benchmarks, DiffSim [16] and FFA [11], to further enhance the validity of the evaluation.

For VTON generation, for paired datasets with ground truth, we use LPIPS [19], SSIM [18], FID [10], and KID [2] to evaluate the quality and faithfulness of the VTON generation. For unpaired datasets without ground truth, we use FID [10] and KID [2] to validate the generation quality. We also include user study for VTON evaluation. We randomly sampled 15 cases from the test dataset and recruited 30 volunteers. Each volunteer was shown the outputs of GP-VTON, IDM-VTON, CatVTON, FitDiT and our method, and asked to select the best result. They received 66, 54, 77, 121, and 132 votes, respectively. Our method's try-on results received the highest score, further validating model effectiveness.

2. Implementation Details

2.1. Dataset Construction

The LAION-Garment dataset integrates several established virtual try-on (VTON) benchmarks through systematic curation: VITON-HD [5] (11,491 pairs), DressCode[14] Upper (13,563 pairs), DressCode Dresses (27,678 pairs), and DressCode Lower (8,689 pairs), DeepFashion2[9] Shop, LRVS-Fashion[12]. We removed test set images and low-quality samples from the dataset. To construct mask-free image pairs, we implement an automated inpainting pipeline combining multi-modal mask generation and context-aware garment replacement. Mask regions are generated through a hybrid approach utilizing CatVTON's Automasker [6] and SAM segmentation [15], deliberately preserving non-garment elements such as hairstyles and background structures. Garment replacement is executed via FLUX-ControlNet-Inpainting [1] under unconditional generation settings, followed by background consistency optimization through selective region repainting to maintain spatial coherence. Representative samples demonstrating this pipeline are provided in Fig.2. We create explicit instructions for each task. For better understanding, we pro-

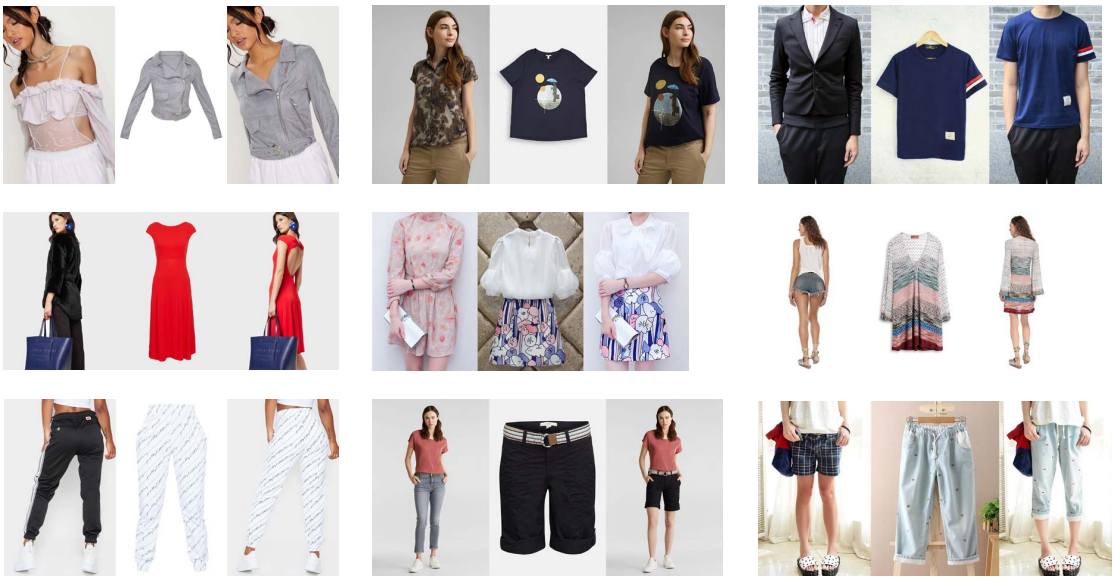

Figure 2. LAION-Garment examples.

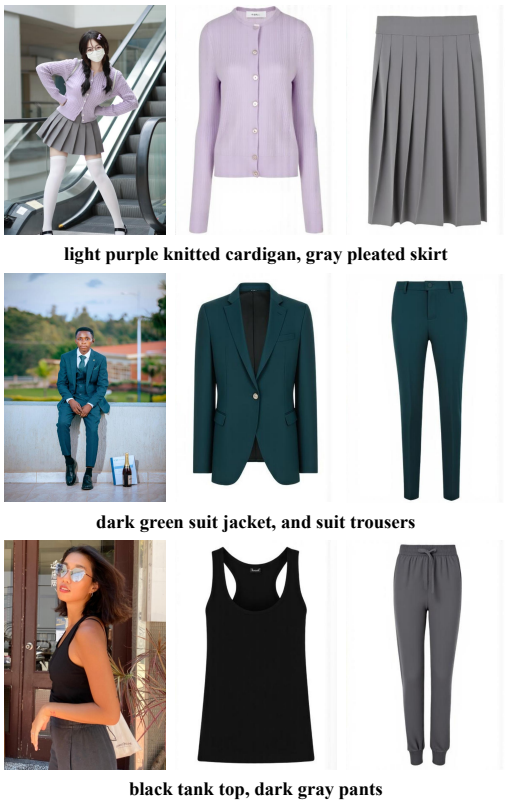

Figure 3. Additional garment reconstruction results in the wild.

vide an example for the Virtual Try-On task: the instruction template is "The set of three images display a model, a garment, and the model wearing the garment. <IMAGE1> shows a person wearing the garment. <IMAGE2> depicts the garment. <IMAGE3> illustrates <IMAGE1> with <IMAGE2> worn by the model". For the more challenging task of Virtual Try-on in layers, the instruction clearly specifies actions such as "drape the coat over the existing outfit".

We also systematically conduct human evaluation on dataset quality through five metrics: pose spatial alignment fidelity, garment-agnostic region preservation, textile pattern consistency, photorealism, and perceptual quality.

2.2. Training Protocol and Inference Configuration

The model employs Prodigy optimization [13] (weight decay=0.01) with text embedding dropout (p=0.1) for regularization. Architectural modifications focus exclusively on the base transformer through LoRA adapters (rank=256, alpha=256), while maintaining frozen VAE and text encoder. Inference follows the base model's native scheduler configuration with default steps.

For fair benchmarking against prior mask-based approaches, evaluation metrics are computed using models trained on fixed-resolution datasets. Although LAION-Garment contains variable-resolution samples to enhance mask-free generalization, conventional methods' architectural constraints (fixed input dimensions and explicit mask dependencies) prevent direct training on our dataset without compromising benchmark integrity.

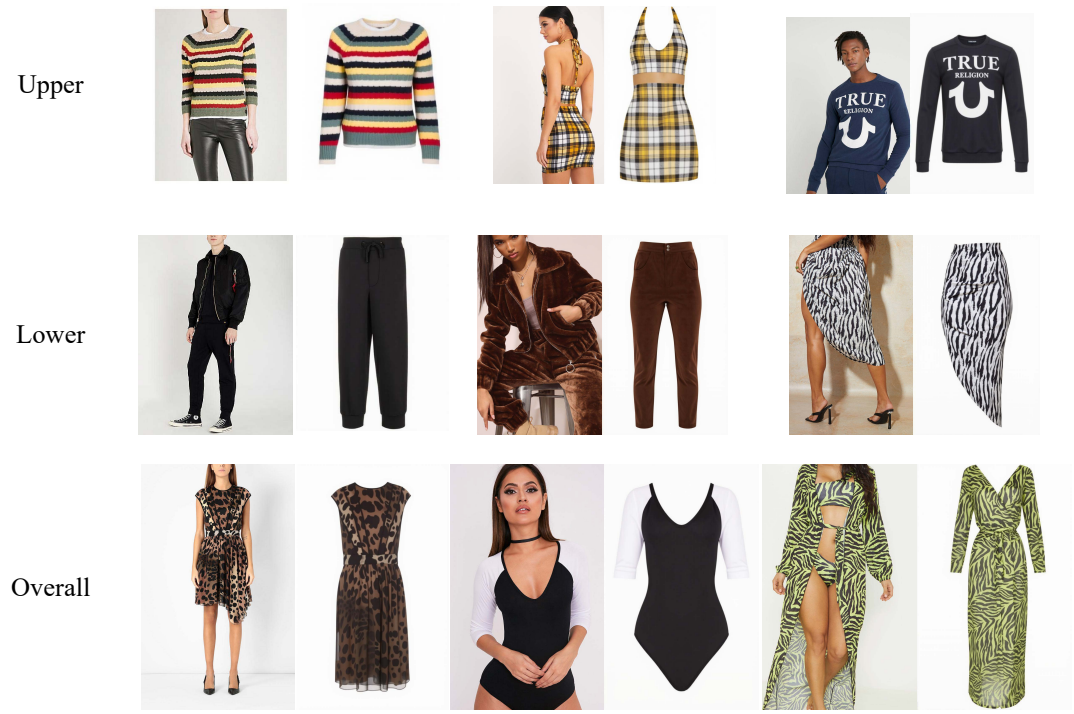

Figure 4. Additional garment reconstruction results in the shop.

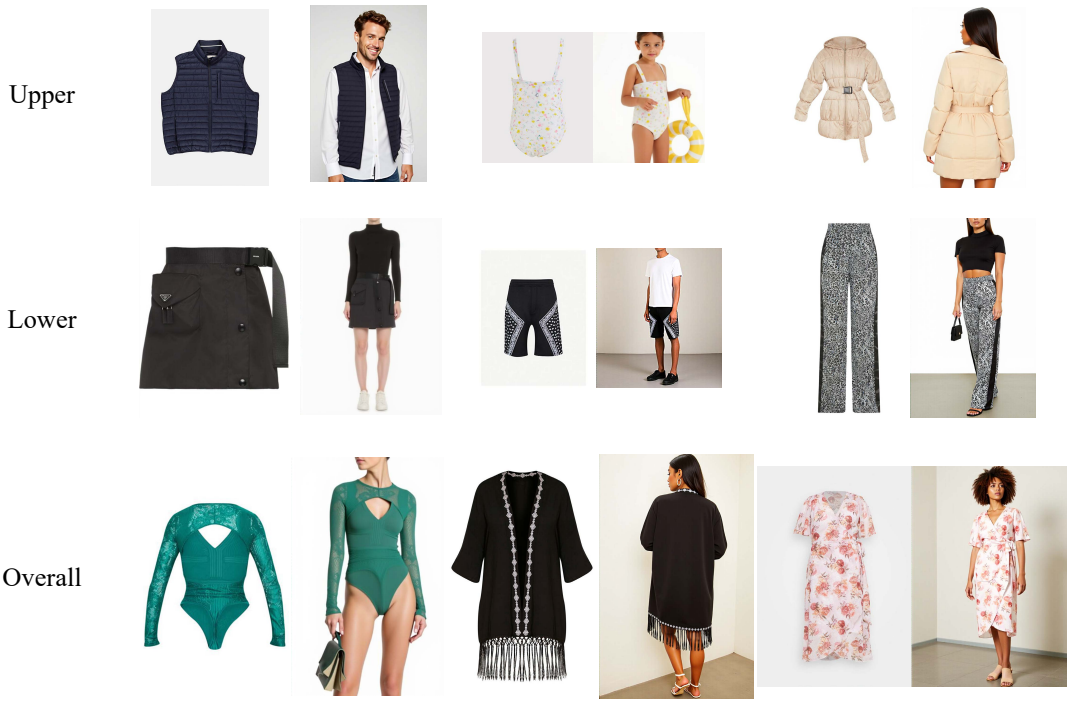

Figure 5. Additional model-free virtual try-on results in the shop.

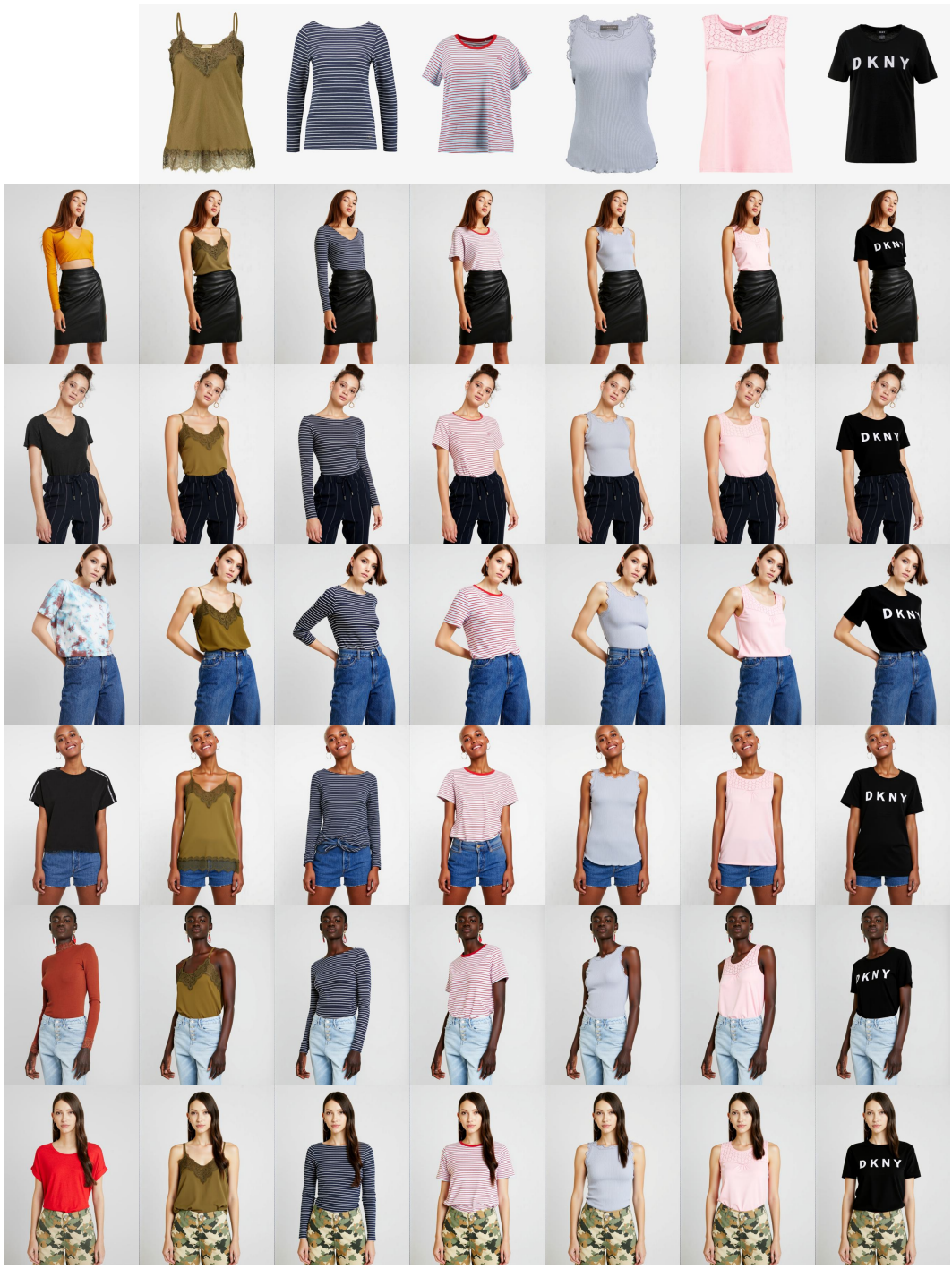

Figure 6. Additional virtual try-on results in the shop.

3. More VTON Generation Results

3.1. Model-free Virtual Try-on

In the main paper, we presented a quantitative comparison between Any2AnyTryon and other baseline methods for Model-free Virtual Try-on. To further showcase the high-quality generation results of Any2AnyTryon on the Model-free Virtual Try-on task, we provide additional visualizations of the generated outfitted model images in Fig. 5. To highlight the generated results, we use the user instruction "model in the shop." The results demonstrate that, whether it's for the upper garment, lower garment, or overall outfit

Figure 7. Sequential try-on.

Figure 8. Comparison with Wear-Any-Way and StreetTryOn.

Figure 9. Try-on results over multiple runs with different seeds.

change, our method can consistently achieve high-fidelity VTON generation.

3.2. Virtual Try-on

Table 1. Variability in virtual try-on numerical results.

Model	LPIPS↓	SSIM↑	FID↓	KID↓
paired	0.0898±0.0013	0.8434±0.0028	7.194±0.133	0.941±0.070
unpaired	-	-	8.901±0.229	1.095±0.100

In Fig. 6, we display more generation results of Any2AnyTryon in the Virtual Try-on task. We provide six different models and six garments with distinct styles, and our Any2AnyTryon produces 36 different, rational, high-quality outfitted results. This proves that our Any2AnyTryon can stably realize impressive VTON generation. Fig. 7 shows that model and garment maintain consistent during sequential try-on. Furthermore, Fig. 8 shows that our method surpasses both Wear-Any-Way [3] and StreetTryOn [7] in performance. In Tab. 1, we report the mean and standard deviation (mean ± std) to quantify the variability in virtual try-on numerical results. Fig. 9 illustrates diverse try-on outcomes generated using different random seeds. As demonstrated in Tab. 1 and Fig. 9, our model consistently produces high-quality try-on results across varying seeds, highlighting its robustness.

3.3. Garment Reconstruction

To further evaluate the quality of garment reconstruction, we provide additional qualitative comparison results in

Figure 10. Additional VTON in layers results in the wild.

Figure 11. Text-driven task switching.

Fig. 1. The garments reconstructed using our method in Any2AnyTryon preserve the details of the input models' garments better than those reconstructed by TryOffDiff, besides, we provide more garment generation in Fig. 4. Additionally, to further demonstrate the stability and generalization ability of Any2AnyTryon in garment reconstruction, we display results of garment reconstruction for in-the-wild model images in Fig. 3. The results show that our method can still generate impressive garment results even for more challenging inputs.

3.4. Virtual Try-on in Layers

We present Virtual Try-on in layers generation results for in-the-wild model images in Fig. 10. The results demonstrate that, even in complex scenarios such as models in the wild, our Any2AnyTryon still produce rational, high-quality outfitted model images, proving the effectiveness of our method.

3.5. Text Instruction

Fig. 11 demonstrates our method's capability to perform impressive try-off, editing and more tasks from same input image by leveraging different text prompts with appropriate position embeddings. The textual guidance significantly influences the output, facilitating precise subject extraction and controllable image editing.

4. Limitation

Just like most diffusion-based models, the model may show some unexpected outputs because the model inherit image generation ability from pretrained text-to-image model. Besides, the proposed methods may fail when encountering with extreme inputs like unrecognizable model, strange poses and low quality image.

References

[1] Alibaba. Flux-controlnet-inpainting, 2024. 1

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. arXiv preprint arXiv:1801.01401, 2018. 1

[3] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In European Conference on Computer Vision, pages 124–142. Springer, 2024. 5

[4] Weifeng Chen, Tao Gu, Yuhao Xu, and Arlene Chen. Magic clothing: Controllable garment-driven image synthesis. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 6939–6948, 2024. 1

[5] 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

[6] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886, 2024. 1

[7] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8235–8239, 2024. 5

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence, 44(5):2567–2581, 2020. 1

[9] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. CVPR, 2019. 1

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017. 1

[11] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Dan Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. Advances in Neural Information Processing Systems, 36:40853–40871, 2023. 1

[12] Simon Lepage, Jérémie Mary, and David Picard. Lrvs-fashion: Extending visual search with referring instructions. arXiv:2306.02928, 2023. 1

[13] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. arXiv preprint arXiv:2306.06101, 2023. 2

[14] 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

[15] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 1

[16] Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity, 2024. 1

[17] Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. arXiv preprint arXiv:2411.18350, 2024. 1

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 1

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 1