# DreamFit: Garment-Centric Human Generation via a Lightweight Anything-Dressing Encoder

**Ente Lin[1*], Xujie Zhang[2*], Fuwei Zhao[3], Yuxuan Luo[3], Xin Dong[3], Long Zeng[1†], Xiaodan Liang[2,4†]**

[1]Shenzhen International Graduate School, Tsinghua University
[2]Shenzhen Campus of Sun Yat-sen University
[3] ByteDance
[4]Sun Yat-sen University
linet22@mails.tsinghua.edu.cn, zhangxj59@mail2.sysu.edu.cn,
zhaofuwei.777@bytedance.com, luoyuxuan@bytedance.com, dongxin.1016@bytedance.com,
zenglong@sz.tsinghua.edu.cn, xdliang328@gmail.com

## Abstract

Diffusion models for garment-centric human generation from text or image prompts have garnered emerging attention for their great application potential. However, existing methods often face a dilemma: lightweight approaches, such as adapters, are prone to generate inconsistent textures; while finetune-based methods involve high training costs and struggle to maintain the generalization capabilities of pretrained diffusion models, limiting their performance across diverse scenarios. To address these challenges, we propose DreamFit, which incorporates a lightweight Anything-Dressing Encoder specifically tailored for the garment-centric human generation. DreamFit has three key advantages: (1) **Lightweight training**: with the proposed adaptive attention and LoRA modules, DreamFit significantly minimizes the model complexity to 83.4M trainable parameters. (2) **Anything-Dressing**: Our model generalizes surprisingly well to a wide range of (non-)garments, creative styles, and prompt instructions, consistently delivering high-quality results across diverse scenarios. (3) **Plug-and-play**: DreamFit is engineered for smooth integration with any community control plugins for diffusion models, ensuring easy compatibility and minimizing adoption barriers. To further enhance generation quality, DreamFit leverages pretrained large multi-modal models (LMMs) to enrich the prompt with fine-grained garment descriptions, thereby reducing the prompt gap between training and inference. We conduct comprehensive experiments on both $768 \times 512$ high-resolution benchmarks and in-the-wild images. DreamFit surpasses all existing methods, highlighting its state-of-the-art capabilities of garment-centric human generation.

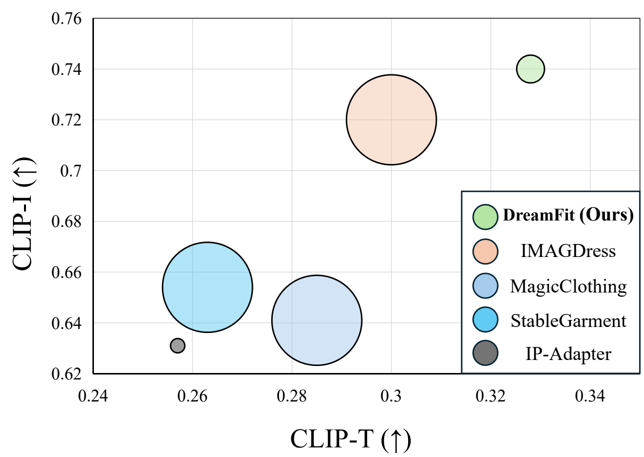


Figure 1: Performance comparison between baselines and our DreamFit. The circle size represents the number of trainable parameters, with larger circles indicating a higher parameter count. Higher CLIP-I and CLIP-T scores signify better alignment between the generated results and text descriptions. Our method not only achieves the best performance but also maintains much fewer training parameters.

## Introduction

Garment-centric human generation aims to synthesize high-quality images of stylized humans based on diverse combinations of text and image prompts, with a focus on accurately depicting the given garment details, as shown in Figure 2. This technology has gained significant traction in industries such as creative advertising and fashion design and sparks active exploration in the academic community (Wang et al. 2024b; Chen et al. 2024a).

Currently, this task faces two critical challenges: the resource-intensive process of full model fine-tuning and the discrepancy between training and inference text prompts. Specifically, finetune-based methods like StableGarment(Wang et al. 2024b) rely on a full copy of diffusion UNet as the garment encoder, namely the "ReferenceNet", which consumes high GPU memory and becomes intractable when scaling up, e.g., from SD1.5(Rombach et al. 2022a) to SDXL(Podell et al. 2023). Moreover, directly fine-tuning the entire UNet is parameter inefficient, requires more training time, and is prone to destroying the pretrained

[*]These authors contributed equally.
[†]Corresponding author

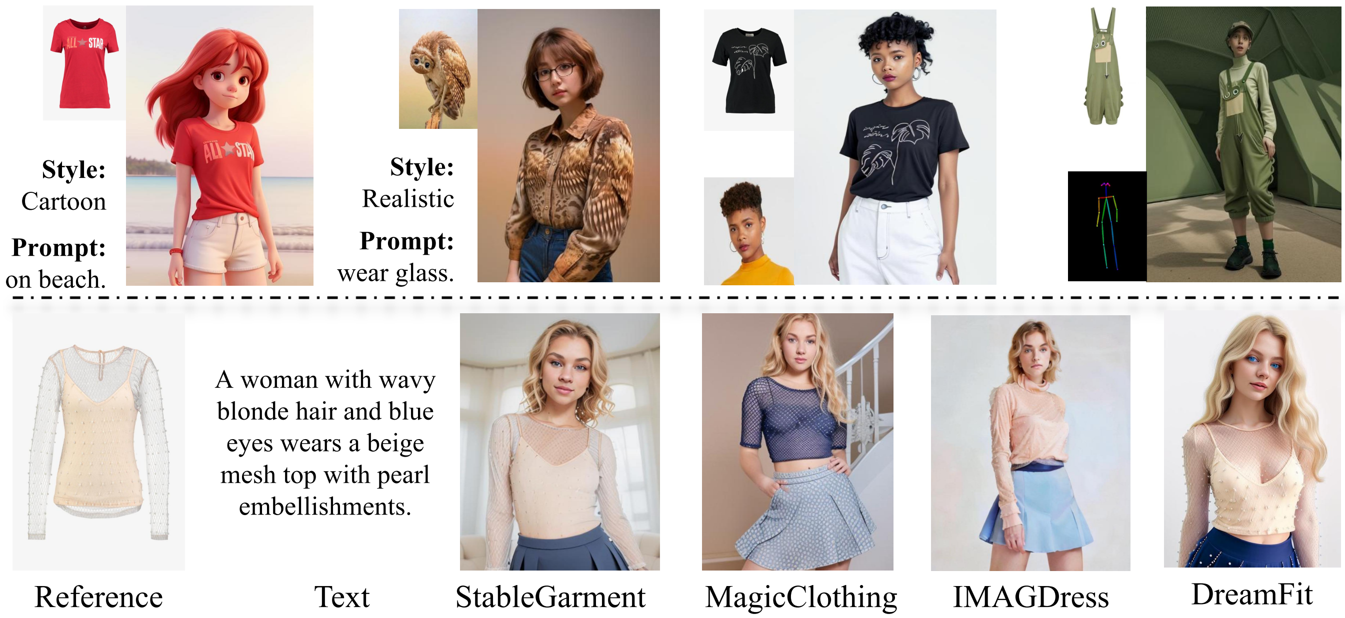


Figure 2: Garment-centric human generation results of our DreamFit: **TOP:** DreamFit can synthesize human images with varied styles, poses, and faces complying with the given clothing image and prompt. **Bottom:** DreamFit demonstrates superior performance compared to SOTA methods.

priors. As a result, the extracted garment feature is less descriptive, making it difficult to fully preserve the texture details (Figure 2 Bottom). Alternatively, some lightweight methods (Ye et al. 2023; Mou et al. 2024) attempt to bypass the bulky ReferenceNet by directly mapping image features into the CLIP (Radford et al. 2021) latent space to control the generation through cross attention. Despite training faster, they tend to overlook crucial image feature details. Therefore, the trade-off between training efficiency and visual detail preservation is a significant problem for garment-centric human generation.

Another challenge is the domain gap between text prompts for training and inference. To fully exploit the rich prior of pretrained T2I diffusion models, training prompts generated by Large Multimodal Models (LMMs) are typically comprehensive (Choi et al. 2024). However, in inference, users often resort to simplified text due to the difficulty of manually creating such rich prompts. This gap further complicates the garment-centric generation process.

To address these challenges, we introduce DreamFit, a novel framework specifically designed for garment-centric human generation. DreamFit leverages a lightweight Anything-Dressing Encoder, which is derived through the activation of LoRA layers within the denoising UNet. This design eliminates the need for resource-intensive components like ReferenceNet and effectively utilizes the rich priors of the diffusion model to extract fine-grained garment features. The extracted garment features are then integrated into the diffusion model through a novel adaptive attention mechanism, enabling high-quality and texture-consistent human image generation. Furthermore, to bridge the gap between training and inference text prompts, DreamFit incorporates LMMs into the inference pipeline. This integration minimizes discrepancies, ensuring that the generated images maintain high fidelity and consistent quality across diverse scenarios.

As shown in Figure 1, DreamFit beats all baselines at CLIP-I and CLIP-T metrics, showing the best text and texture consistency of generation results even with 10x fewer trainable parameters (83.4M v.s. 875M full model fine-tuning). Our main contributions are summarized as follows:

- We propose **DreamFit**, an efficient and dedicated framework for garment-centric human generation, enabled by a lightweight and plug-and-play Anything-Dressing Encoder. Our model addresses the inefficiencies and limitations of existing methods, particularly in aspects of model fine-tuning and computational efficiency.
- We integrate LMMs into the inference pipeline that effectively reduces the prompt gap between training and testing, enhancing the overall quality and fidelity of the generated results.
- Extensive experiments on open and internal benchmarks of $768 \times 512$ resolution verify the superiority of DreamFit, demonstrating state-of-the-art performance and robust generalization in diverse human generation tasks.

## Related Work

Human generation has witnessed rapid advancements, particularly with the emergence of controllable text-to-image generation technologies. Existing methods in this field often concentrate on either enhancing lightweight efficiency or enhancing texture consistency. However, these approaches

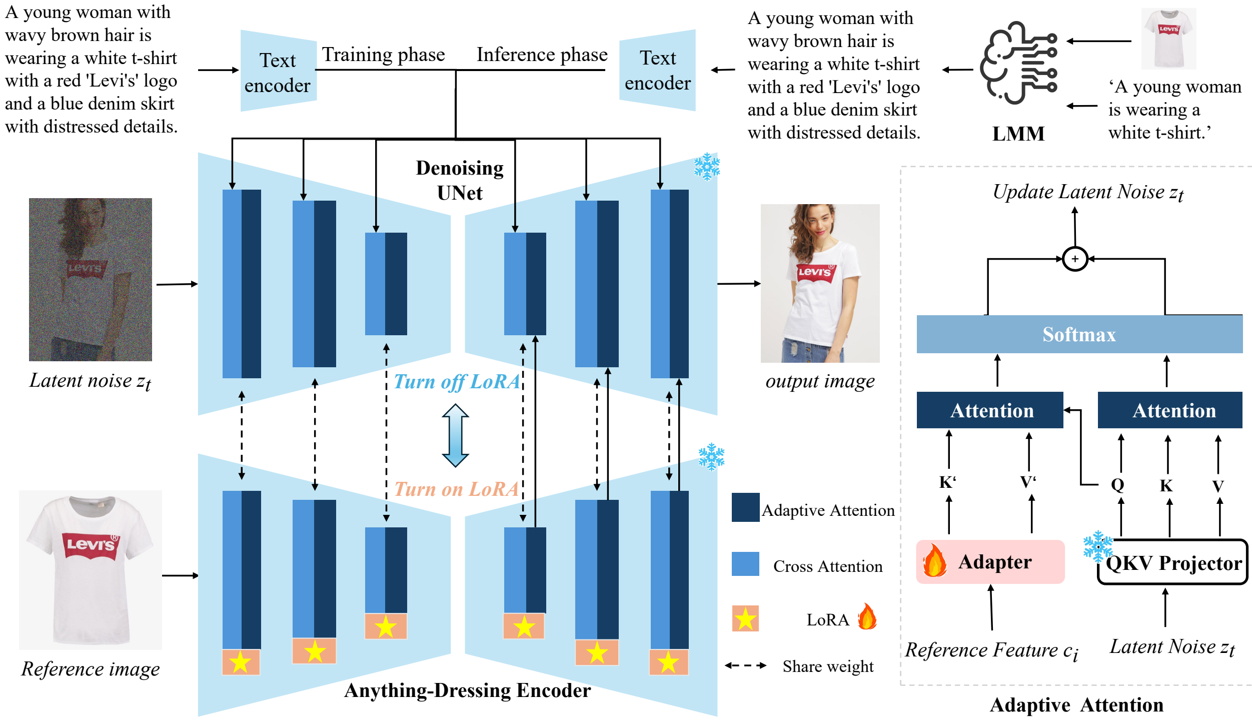


Figure 3: Overview of DreamFit. Our method constructs an Anything-Dressing Encoder utilizing LoRA layers. The reference image features are extracted by the Anything-Dressing Encoder and then passed into the denoising UNet via adaptive attention. Furthermore, we incorporate Large Multimodal Models (LMM) into the inference process to reduce the text prompt gap between the training and testing.

struggle to balance the trade-offs between computational requirements and image quality.

Lightweight methods like (Hu et al. 2021; Ye et al. 2023; Mou et al. 2024; Zhang, Rao, and Agrawala 2023; Ruiz et al. 2023; Chen et al. 2024b; Liu et al. 2024b; Han et al. 2024; Yang et al. 2024; Li et al. 2024; Xing et al. 2024) have been proposed as lightweight approaches to achieve controllable human generation. These methods typically align image features with text features, using a combination of text and visual prompts to guide the generation process. T2I-adapter (Mou et al. 2024) designed a series of convolutional layers to downsample and extract the reference image features, which are then fed into the denoising UNet via residual connections. IP-Adapter (Ye et al. 2023) introduces a sophisticated decoupled cross-attention mechanism to integrate image features into the denoising UNet. Although these methods excel in lightweight design, they often struggle to maintain high texture consistency, especially in intricate garment details. In contrast, our method not only retains the same level of lightweight efficiency but also ensures significantly improved texture consistency.

Additionally, several methods (Zhu et al. 2023; Morelli et al. 2023; xujie zhang et al. 2023; Wang et al. 2024a; Xu et al. 2024; Zhu et al. 2024; Kim et al. 2024; Zhang et al. 2024; Hu 2024; Cui et al. 2023) aiming for better texture consistency have chosen to utilize the UNet from Stable Diffusion as the image encoder to capture image prompt features. StableGarment (Wang et al. 2024b) finetunes a trainable copy of denoising UNet as the garment encoder and incorporates garment features using additive self-attention. MagicClothing (Chen et al. 2024a) proposes joint classifier-free guidance to obtain a trade-off between garment features and text prompts. While these approaches, which require extensive finetuning of pretrained models such as those based on Stable Diffusion, yield commendable results in terms of quality, they come with significant drawbacks. These methods are resource-intensive, demanding significant memory resources and exhibiting sluggish training procedures, thus limiting their scalability for real-world applications.

Our work builds on these developments by leveraging frozen large diffusion models (LDMs) for garment-centric human image synthesis. We introduce a lightweight, plug-and-play Anything-Dressing Encoder that addresses the inefficiencies and limitations of existing methods, offering a scalable and flexible solution for garment-centric human image synthesis.

## Methodology

### Preliminary

Stable Diffusion is a text-conditioned latent diffusion model (Rombach et al. 2022b). Given a latent feature $z_0$ encoded from the input image by a VAE (Kingma and Welling 2013), the forward diffusion process is firstly performed by adding noise according to a predefined noise scheduler

$\alpha_t$ (Ho, Jain, and Abbeel 2020):

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_0, (1-\alpha_t)I). \quad (1)$$

Then to reverse the diffusion process to generate new images, a noise estimator $\epsilon_\theta(\cdot)$ parameterized by an UNet is learned to predict the forward added noise $\epsilon$ with the objective function as follows:

$$\mathcal{L}_{\text{dm}} = E_{\boldsymbol{z}_0,\boldsymbol{\epsilon},\boldsymbol{c}_t,t}\left[||\epsilon - \epsilon_\theta(z_t,t,c_t)||^2\right], \quad (2)$$

Where $c_t$ is the text condition associated with the image latent $z$. In the Stable Diffusion model, each block of the UNet consists of cross-attention and self-attention layers. The cross-attention layer facilitates attention between the image feature query and the text condition embedding, while self-attention layer operates within the image feature space.

## DreamFit

The overall architecture of DreamFit is shown in Figure 3. Our model consists of two main components: a frozen denoising UNet with pretrained stable diffusion weights, and a set of trainable LoRA layers. DreamFit takes as input a reference garment image and a text prompt that describes the detail of the output image. The text prompt features, extracted by a frozen text encoder, are integrated into the denoising UNet via a cross-attention mechanism. The features of the reference image are extracted by a lightweight Anything-Dressing Encoder, which is obtained by turning on the LoRA layers in the denoising UNet. Once the LoRA layers are turned off, the lightweight Anything-Dressing Encoder reverts to the denoising UNet. The extracted reference features are then integrated into the denoising UNet through an adaptive self-attention mechanism. After multiple rounds of denoising, the denoising UNet generates an image that closely aligns with the reference image and text prompt. We elaborate the anything-dressing encoder and the adaptive attention in the following sections.

**Lightweight Anything-Dressing Encoder.** The core concept of our model is that the pretrained diffusion model possesses extensive prior knowledge, making it a potential robust feature extractor. However, the diffusion model is designed to process latent noise, rendering them less effective for extracting features from noise-free reference images. To address this, we utilize LoRA layers to extend the feature extraction capability of the diffusion model to noiseless reference images. Specifically, we incorporated the LoRA layers into the linear and convolutional layers of the denoising UNet, modifying the forward pass as follows:

$$h = \phi(x) + I(x) * \Delta W(x) \quad (3)$$

$$I(x) = \begin{cases} 0, & \text{if } x = z_t, \\ 1, & \text{if } x = c_i. \end{cases} \quad (4)$$

Here, $\phi$ denotes the linear or convolutional layers in the denoising UNet. $I(x)$ denotes a gate function and $\Delta W(x)$ denotes the added LoRA layer (Hu et al. 2021). $z_t$ and $c_i$ represent the latent noise and the extracted reference features at each block of denoising UNet, respectively. With this approach, we only need to train the lightweight LoRA layers instead of a bulky ReferenceNet. And the denoising UNet is now seamlessly transformed into an Anything-Dressing Encoder with retained pretrained weights (i.e., better generalization capabilities).

**Adaptive attention.** As shown in Figure 3, the reference image features are integrated into the denoising UNet through adaptive attention. During training, the weights of our denoising UNet the frozen, which can hinder the pretrained attention module from effectively capturing the relationship between reference image features and the latent noise. Inspired by the adapter mechanism (Ye et al. 2023), we introduced two trainable linear projection layers $\mathbf{W}'_k$, $\mathbf{W}'_v$, into the adaptive attention. These layers act as adapters to align the reference image features with latent noise. After applying cross-attention mechanisms, we add the cross-attention output to the self-attention output. The formulation of adaptive attention is:

$$\mathbf{z}_t^{\text{new}} = \operatorname{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} + \operatorname{Softmax}\left(\frac{\mathbf{Q}\left(\mathbf{K}'\right)^\top}{\sqrt{d}}\right)\mathbf{V}' \quad (5)$$

where $\mathbf{Q} = \mathbf{W}_q\mathbf{z}_t, \mathbf{K} = \mathbf{W}_k\boldsymbol{z}_t, \mathbf{V} = \mathbf{W}_v\boldsymbol{z}_t, \mathbf{K}' = \mathbf{W}'_k\boldsymbol{c}_i, \mathbf{V}' = \mathbf{W}'_v\boldsymbol{c}_i$, and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are frozen linear projection layers in the denoising UNet. To accelerate the convergence of attention modules, we initialize the $\mathbf{W}'_k$, $\mathbf{W}'_v$ with $\mathbf{W}_k$, $\mathbf{W}_v$. This allows us to seamlessly and lightweightly inject the reference image features into the denoising UNet, resulting in synthesized images that are highly consistent with the reference image.

**Training and inference** During the training phase, we exclusively optimize the LoRA layers and the trainable adapters in the adaptive attention. Dreamfit is trained on a dataset containing image-text pairs, with the text generated by Language Model Models (LMMs), such as CogVLM (Wang et al. 2023) and LLaVA1.5 (Liu et al. 2024a). To fully leverage the rich prior of the pretrained T2I diffusion model (Choi et al. 2024), we prepare comprehensive captions for each image (e.g., "woman wears a long-sleeved polo shirt..."). The training objective remains consistent with the original diffusion model:

$$L_{\text{LDM}} = E_{\boldsymbol{z}_0,\boldsymbol{\epsilon},\boldsymbol{c}_t,\boldsymbol{c}_i,t}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t,\boldsymbol{c}_t,\boldsymbol{c}_i,t\right)\right\|^2 \quad (6)$$

During the inference phase, users are expected to input a reference image and a text description related to the desired output image. However, users often struggle to accurately describe fine-grained details in images and tend to provide simple text, leading to a domain gap between user input text and training text. This gap can result in decreased consistency and quality of the generated images. To address this challenge, we propose utilizing LMMs to rewrite the user input text based on the reference images. This approach enriches the text prompt with fine-grained descriptions of the garment, thereby reducing the prompt gap between training and inference. We denote the rewritten text as $c'_t$. Then, we apply classifier-free guidance (Ho and Salimans 2022) during the denoising process to generate the human image:

$$\hat{\boldsymbol{\epsilon}}_\theta\left(\boldsymbol{z}_t,\boldsymbol{c}'_t,\boldsymbol{c}_i,t\right) = w\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t,\boldsymbol{c}'_t,\boldsymbol{c}_i,t\right) + (1-w)\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t,t\right) \quad (7)$$

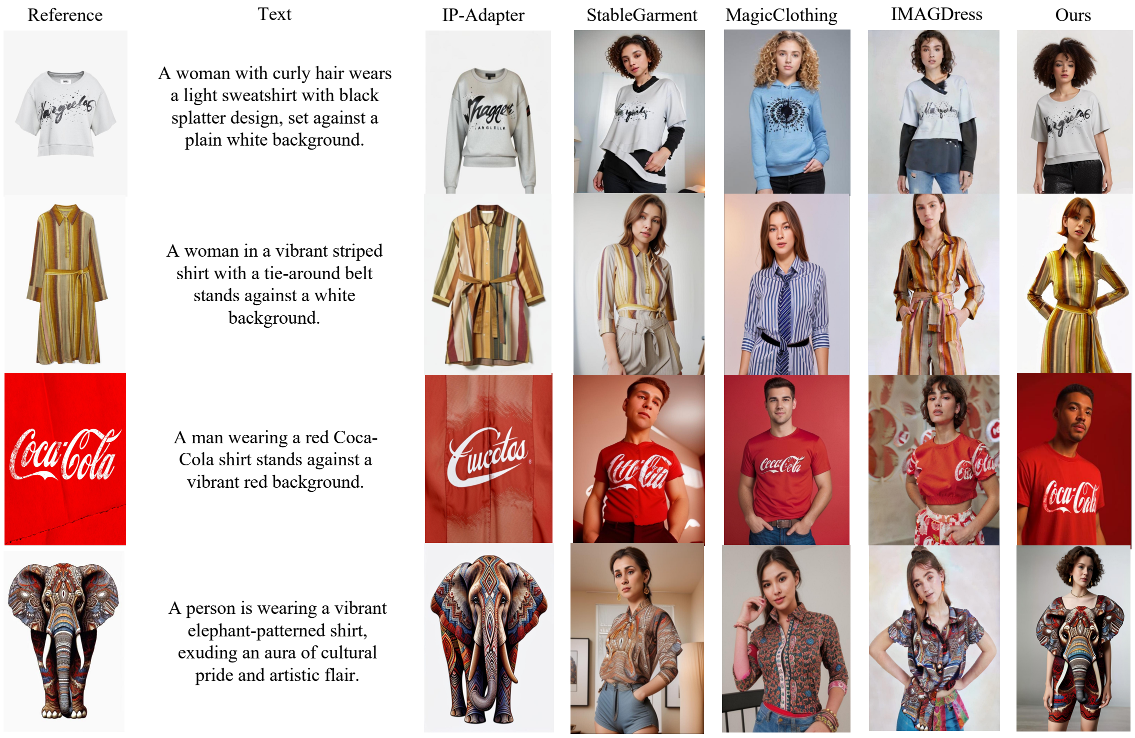


Figure 4: Qualitative comparison on the open and internal benchmarks. DreamFit demonstrates a distinct advantage in handling complex patterns and text. Please zoom in for more details.

# Experiments

## Datasets

To train Dreamfit, we collected approximately 500,000 garment-person image pairs from the internet and captioned them using large multi-modal models. For model evaluation, we introduce two garment-centric human generation benchmarks derived from public datasets and the Internet. The open benchmark is constructed using a subset of VITON-HD (Choi et al. 2021) and DressCode (Morelli et al. 2022) test sets. Specifically, we handpicked 200 diverse garments from these datasets encompassing various styles, colors, shapes, and textures. Each garment is associated with refined prompts generated through large multi-modal models. In addition to the open benchmark, we further developed a more challenging internal benchmark. To build the internal benchmark, we manually gathered 200 reference images from the internet, including garments with intricate textures and backgrounds, and non-garment images featuring animals, fruits, and patterns. This benchmark is designed to evaluate the robustness and generalization capability of the model across diverse reference image types.

## Implementation Details

The denoising UNet is initialized with the weights of SD1.5 and we use CLIP ViT-L/14 (Radford et al. 2021) as the text encoder. Our model was trained on paired images with a resolution of $768 \times 512$. We initialized the LoRA layers in the same manner as described in (Hu et al. 2021), with the LoRA rank set to 64. The training was conducted on 8 A800 (40G) GPUs for 90k steps, with a batch size of 4 per GPU. We utilized the AdamW optimizer with a fixed learning rate of 1e-4. To validate scalability, we also initialized the denoising UNet as SDXL and trained the model on 8 A100 (80G) GPUs for 90k steps with the same batch size. Thanks to the efficient model design, acceleration techniques such as FSDP or Deepspeed were not required for training. During inference, we use CogVLM (Wang et al. 2023) to refine the user input text. We use DDIM (Song, Meng, and Ermon 2020) sampler with 50 steps and set guidance scale $w$ to 7.5.

## Baselines

We selected four state-of-the-art methods, including StableGarment (Wang et al. 2024b), MagicClothing (Chen et al. 2024a), IMAGDress (Shen et al. 2024), and IP-Adapter (Ye et al. 2023). We directly use their released pretrained models for comparison. All experiments are conducted with a resolution of $768\times512$. For a fair comparison, we used the model initialized with the weights of SD1.5 in the experiment.

## Qualitative Results

The qualitative comparison is illustrated in Figure 4. The results demonstrate the superiority of our DreamFit over other baselines. Firstly, our method exhibits better texture and text consistency. Existing lightweight approaches, such as IP-Adapter, lose text and texture consistency when faced with reference images containing complex details.

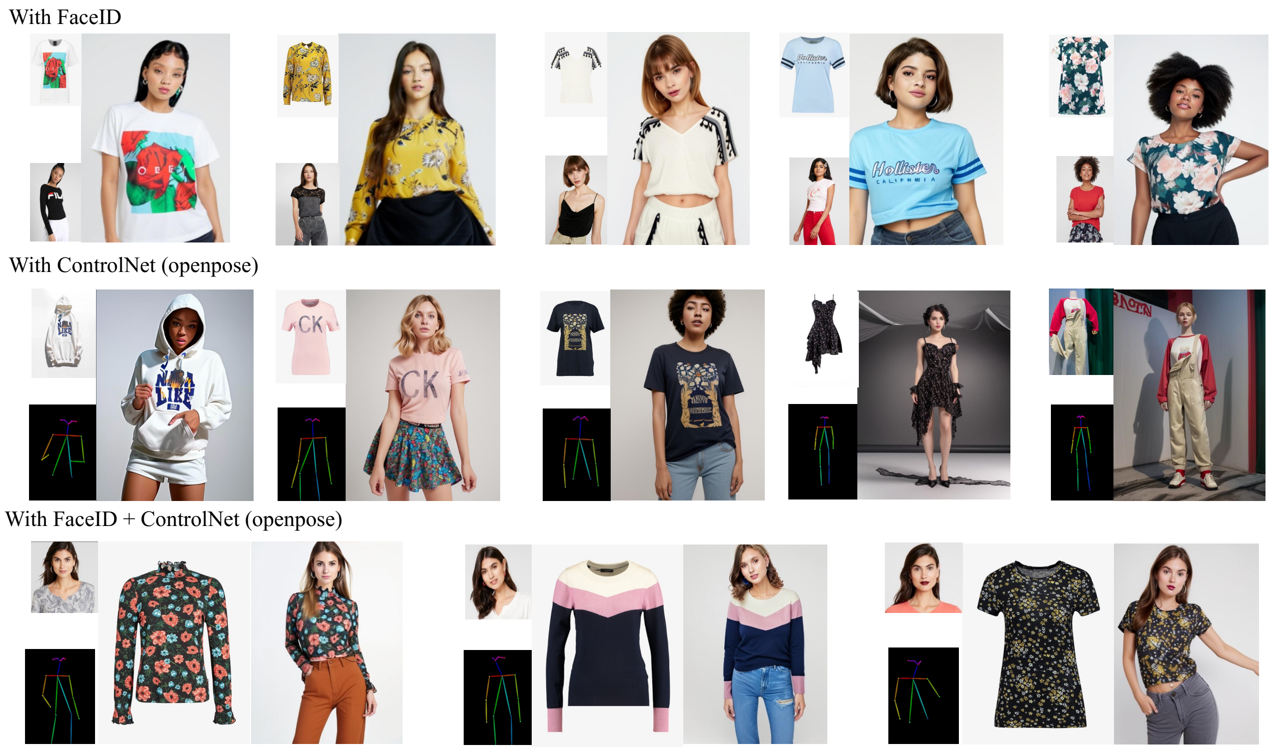


Figure 5: Plug-and-play results of DreamFit, our method can seamlessly integrate with community conditional control plugins.

While finetune-based methods like StableGarment, MagicClothing, and IMAGDress surpass IP-Adapter significantly, they still struggle to generate human images with fine-grained features that highly match the text prompts. Secondly, images generated by baselines generally exhibit issues such as blurriness (IMAGDress) and incorrect body proportions (third row of StableGarment column). In contrast, our method achieves the best quality and realism. To show the model's excellent compatibility, we provide additional results showcasing DreamFit integrated with community plugins in Figure 5.

## Quantitative Results

**Metric.** We quantitatively measure the generated results from three aspects: consistency, generation quality, and human preference. Regarding consistency, we employ CLIP-T and CLIP-I (Hessel et al. 2021) to evaluate text and texture consistency, respectively. We utilize Aesthetic score(AS) (Schuhmann et al. 2022) to evaluate generation quality. Several studies (Ku et al. 2023; Peng et al. 2024) have shown that GPT can produce evaluation results that are remarkably consistent with humans. Consequently, we employ GPT-4o (Achiam et al. 2023) as the automatic evaluation of human preference, denoted as Human-Aligned Score (HAS). For more details on the evaluation, please refer to the supplementary material.

**Evaluation.** We conducted experiments using both open and internal benchmarks to assess the performance of our model against baselines. The quantitative results are reported in


Figure 6: Qualitative results of the ablation study on network components.

Tab 1. Our DreamFit consistently outperforms all baselines across various metrics and benchmarks, showcasing its ability to produce results with superior text and texture consistency, exceptional generation quality, and strong alignment with human preferences. Furthermore, compared with fully finetune-based methods, like StableGarment, IMAGDress, DreamFit has the fewest trainable parameters while achieving the best performance. The quantitative result underscores DreamFit's resource efficiency and robustness.

## Ablation Study

**Effectiveness of Network Components** To evaluate the impact of different network components on the final per-

| Method | Trainable Parameters | Open Benchmark | | | | Internal Benchmark | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CLIP-T↑ | CLIP-I↑ | AS↑ | HAS↑ | CLIP-T↑ | CLIP-I↑ | AS↑ | HAS↑ |
| IP-Adapter | **22M** | 0.257 | 0.631 | 4.667 | 0.568 | 0.135 | 0.422 | 2.634 | 0.430 |
| StableGarment | 859M | 0.263 | 0.654 | 4.877 | 0.768 | 0.275 | 0.629 | 5.238 | 0.683 |
| MagicClothing | 875M | 0.285 | 0.641 | 5.007 | 0.443 | 0.320 | 0.583 | 5.462 | 0.595 |
| IMAGDress | 875M | 0.300 | 0.720 | 4.867 | 0.783 | 0.290 | 0.632 | 5.204 | 0.840 |
| Ours | 83.4M | **0.328** | **0.740** | **5.414** | **0.856** | **0.334** | **0.687** | **5.516** | **0.920** |

Table 1: Quantitative comparison of different methods on three metrics for both Open and Internal benchmarks.


Figure 7: Qualitative results of the ablation study on text prompts.

| Method | Trainable Parameters | CLIP-T↑ | CLIP-I↑ | AS↑ |
|---|---|---|---|---|
| Finetuning | 875M | 0.322 | 0.622 | **5.534** |
| Only LoRA | 66M | 0.322 | 0.686 | 5.250 |
| Only Adapter | 17.4M | 0.310 | 0.679 | 5.230 |
| Ours | 83.4M | **0.334** | **0.687** | 5.516 |

Table 2: Quantitative results of the ablation study on network components

| Method | CLIP-I ↑ | AS↑ | Human Study↑ |
|---|---|---|---|
| Fixed text | 0.588 | 5.283 | 0.03 |
| Simple text | 0.663 | 5.186 | 0.10 |
| Ours | **0.687** | **5.516** | **0.87** |

Table 3: Quantitative comparison of different text prompt configurations.

formance, we conducted an ablation study comparing various model configurations. We experiment with three setups: finetuning the entire Anything-Dressing Encoder (finetuning), only optimizing parameters in LoRA layers (only LoRA), and only optimizing parameters of adapters in adaptive attentions (only Adapter). The results are summarized in Tab 2 and the qualitative results are shown in Figure 6.

Despite having the highest number of trainable parameters, finetuning the entire Anything-Dressing Encoder exhibited subpar performance on text and texture consistency. We conjecture this is because finetuning can compromise the model's pre-trained prior, leading to a decline of its feature extraction capability. While the ”only LoRA” and ”only Adapter” configurations showed better texture consistency than finetuning, they performed inadequately for generation quality. In contrast, our DreamFit, which combines and optimizes both LoRA and Adapter, achieves superior results across all metrics.

**Impact of Text Prompts** To further explore the influence of text prompts, we conducted experiments with three different types of text inputs: (1) a fixed text prompt (Fix text), (2) a simple text prompt (Simple text) for simulating user input, and (3) text prompts rewritten by LMMs (Ours). The qualitative and quantitative results for each configuration are shown in Figure 7 and Tab 3.

The results demonstrate that text prompts rewritten by large multi-modal models yield superior quality outputs across all metrics, including notably higher scores in human study evaluations. This underscores the efficacy of incorporating advanced LMMs to enhance prompts with detailed descriptions of reference images, thereby narrowing the gap between training and inference prompts and boosting the quality and realism of the generated images.

## Conclusion

In this paper, we introduced **DreamFit**, a novel garment-centric human image generation framework designed to address the inefficiencies and limitations of existing methods. By leveraging a lightweight, plug-and-play Anything-Dressing Encoder based on LoRA layers, DreamFit significantly streamlines model complexity and memory usage, facilitating more efficient and scalable training procedures. Our approach integrates large multi-modal models into the inference process, effectively reducing the domain gap between training and inference text prompts and enhancing the overall quality and consistency of the generated images. Extensive experiments conducted on open and internal benchmarks demonstrate that DreamFit not only achieves state-of-the-art performance but also exhibits superior generalization capabilities across diverse scenarios.

## Acknowledgments

This work was sponsored by National Key Research and Development Program of China (Grant No. 2022YFB3303101), in part by Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001) and Doubao Fund.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, W.; Gu, T.; Xu, Y.; and Chen, C. 2024a. Magic Clothing: Controllable Garment-Driven Image Synthesis. *arXiv preprint arXiv:2404.09512*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

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*.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*.

Cui, A.; Mahajan, J.; Shah, V.; Gomathinayagam, P.; and Lazebnik, S. 2023. Street TryOn: Learning In-the-Wild Virtual Try-On from Unpaired Person Images. *arXiv preprint arXiv:2311.16094*.

Han, Y.; Zhu, J.; He, K.; Chen, X.; Ge, Y.; Li, W.; Li, X.; Zhang, J.; Wang, C.; and Liu, Y. 2024. Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control. *arXiv preprint arXiv:2405.12970*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, L. 2024. 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*, 8153–8163.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ku, M.; Jiang, D.; Wei, C.; Yue, X.; and Chen, W. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Li, S.; Yang, L.; Jiang, X.; Lu, H.; Di, Z.; Lu, W.; Chen, J.; Liu, K.; Yu, Y.; Lan, T.; et al. 2024. SwiftDiffusion: Efficient Diffusion Model Serving with Add-on Modules. *arXiv preprint arXiv:2407.02031*.

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

Liu, J.; Li, C.; Sun, Q.; Ming, J.; Fang, C.; Wang, J.; Zeng, B.; and Liu, S. 2024b. Ada-adapter: Fast Few-shot Style Personlization of Diffusion Model with Pre-trained Image Encoder. *arXiv preprint arXiv:2407.05552*.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*.

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2231–2235.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Peng, Y.; Cui, Y.; Tang, H.; Qi, Z.; Dong, R.; Bai, J.; Han, C.; Ge, Z.; Zhang, X.; and Xia, S.-T. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024. IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Wang, H.; Zhang, Z.; Di, D.; Zhang, S.; and Zuo, W. 2024a. MV-VTON: Multi-View Virtual Try-On with Diffusion Models. *arXiv preprint arXiv:2404.17364*.

Wang, R.; Guo, H.; Liu, J.; Li, H.; Zhao, H.; Tang, X.; Hu, Y.; Tang, H.; and Li, P. 2024b. StableGarment: Garment-Centric Generation via Stable Diffusion. *arXiv preprint arXiv:2403.10783*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Xing, J.; Xu, C.; Qian, Y.; Liu, Y.; Dai, G.; Sun, B.; Liu, Y.; and Wang, J. 2024. TryOn-Adapter: Efficient Fine-Grained Clothing Identity Adaptation for High-Fidelity Virtual Try-On. *arXiv preprint arXiv:2404.00878*.

Xu, Y.; Gu, T.; Chen, W.; and Chen, C. 2024. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.

xujie zhang; Li, X.; Kampffmeyer, M.; Dong, X.; Xie, Z.; Zhu, F.; Dong, H.; and Liang, X. 2023. WarpDiffusion: Efficient Diffusion Model for High-Fidelity Virtual Try-on. arXiv:2312.03667.

Yang, Y.; Wang, W.; Peng, L.; Song, C.; Chen, Y.; Li, H.; Yang, X.; Lu, Q.; Cai, D.; Wu, B.; et al. 2024. LoRA-Composer: Leveraging Low-Rank Adaptation for Multi-Concept Customization in Training-Free Diffusion Models. *arXiv preprint arXiv:2403.11627*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, X.; Lin, E.; Li, X.; Luo, Y.; Kampffmeyer, M.; Dong, X.; and Liang, X. 2024. MMTryon: Multi-Modal Multi-Reference Control for High-Quality Fashion Generation. *arXiv preprint arXiv:2405.00448*.

Zhu, L.; Li, Y.; Liu, N.; Peng, H.; Yang, D.; and Kemelmacher-Shlizerman, I. 2024. M&M VTO: Multi-Garment Virtual Try-On and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1346–1356.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.