ShoeFit: A New Dataset and Dual-image-stream DiT Framework for Virtual Footwear Try-On

Yuhan Li[1*†], **Zhiyu Jin**[2*], **Yifan Tong**[2], **Wenxiang Shang**[2], **Benlei Cui**[2],
Xuanhong Chen[1], **Hangcheng Zhu**[2], **Bingbing Ni**[1‡]
[1]Shanghai Jiao Tong University [2]Alibaba Group

(a) We illustrate the challenges in VFTON. (1) Subtle variations make it difficult to match data pairs. (2) Information needed by the target view is often not provided in the given product image. (3) Shoes often exhibit complex reflections and shadows that interact with environment lighting, leading to a confusion of their natural color.

(b) ShoeFit has demonstrated excellent preservation of the fidelity of footwear across real-world scenes.

Figure 1: VFTON presents three challenges, while ShoeFit gives a successful solution.

Abstract

Virtual footwear try-on (VFTON), a critical yet underexplored area in virtual try-on (VTON), aims to synthesize faithful try-on results given diverse footwear and model images while maintaining 3D consistency and texture authenticity. Unlike conventional garment-focused VTON methods, VFTON presents unique challenges due to (1) **Data Scarcity**, which arises from the difficulty of perfectly matching product shoes with models wearing the identical ones, (2) **Viewpoint Misalignment**, where the target foot pose and source shoe views are always misaligned, leading to incomplete texture information and detail distortion, and (3) **Background-induced Color Distortion**, where complex material of footwear interacts with environmental lighting, causing unintended color contamination. To address these challenges, we introduce MVShoes, a multi-view shoe try-on dataset consisting of 7305 well-annotated image triplets, covering diverse footwear categories and challenging try-on scenarios. Furthermore, we propose a dual-stream DiT architecture, **ShoeFit**, designed to mitigate viewpoint misalignment through Multi-View Conditioning with 3D Rotary Position Embedding, and alleviate background-induced distortion using the LayeredRefAttention which leverages background features to modulate

* Equal contributions.
†Work done during an internship at Alibaba.
‡Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

footwear latents. The proposed framework effectively decouples shoe appearance from environmental interferences while preserving high-quality texture detail through decoupled denoising and conditioning branches. Extensive quantitative and qualitative experiments demonstrate that our method substantially improves rendering fidelity and robustness under challenging real-world product shoes, establishing a new benchmark in high-fidelity footwear try-on synthesis. The dataset and benchmark will be publicly available upon acceptance of the paper.

1 Introduction

Virtual Footwear Try-On (VFTON) involves synthesizing product shoes onto human foot images, preserving their geometric structure and material properties, based on input images of both the source product and the target foot. VFTON offers significant commercial potential and academic value, with Nike's annual footwear sales reaching 33,427 million dollars (48), while shoes, as complex 3D multi-view objects (57), present unique challenges in research on controllable generation. Unfortunately, this field has not been sufficiently explored within the community.

There are three main challenges hindering the advancement of VFTON. (1) **Data Scarcity**: Due to the prevalence of footwear with similar appearances but subtle variations, as in Fig. 1 (a-1), such as slight differences in shoelaces, heels, and stripes, it is challenging to accurately match product shoes with models wearing the identical ones. Further exacerbating this issue, companies tend to keep data related to shoe try-ons confidential due to its clear commercial value (75; 19). As a result, there is currently no available open-source dataset for footwear try-ons. (2) ***Viewpoint Misalignment***: Unlike flat-lay 2D garment VTON, where both the target garment and human pose are captured from aligned front-facing perspectives, as in Fig. 1 (a-2), virtual footwear try-on inherently involves mismatched viewpoints between in-shop product images (source view) and target human poses (target view). Specifically, information needed by the target view is often not provided in the given product image due to 3D constraints, creating an information-deficient scenario that results in detail distortion and artifacts (42; 56). (3) ***Background-induced Color Distortion***: Unlike the diffuse reflection properties of knitted fabrics in clothing, which ensure visual color largely remains unaffected by the background, shoes often exhibit complex reflective properties that interact with ambient lighting (32), making them more susceptible to visual interference from background colors, as in Fig. 1 (a-3). This interaction between appearance and background leads to color contamination from surroundings and distorted material rendering in synthesized results. The above issues are first identified and explicitly proposed in this paper, which make it impossible for VFTON to share the same pipeline as VTON, thus presenting VFTON with unique challenges and research value.

To tackle these challenges, we introduce MVShoes, **the first multi-view footwear try-on dataset**. It comprises 7305 annotated high-resolution image triplets, each containing two product shoe images from different views and a corresponding try-on result, as shown in Fig. 2 (a). The dataset covers a diverse range of shoe categories and try-on scenarios, supported by a comprehensive and rigorous data cleaning pipeline to ensure accurate matching of identical product shoes with their human images.

Furthermore, we propose ShoeFit, a customized dual-image-stream DiT (51) architecture for VFTON, which generates high-fidelity try-on results through decoupled denoising and conditioning branches. ShoeFit is also equipped with two key improvements: (1) **Multi-View Conditioning**, incorporating reference concatenation (10) and corresponding 3D Rotary Position Embedding (3D RoPE). While a single source product view often fails to provide complete shoe texture features required for the target try-on view, the try-on results can significantly benefit from multiple source views, effectively reducing uncertainty in target surface rendering and minimizing texture distortion. (2)**LayeredRefAttention Module**. LayeredRefAttention leverages background features to modulate shoe features (28), decoupling the intrinsic shoe appearance from the surroundings. Additionally, the subsequent denoising attention suppresses background-related computations, effectively mitigating background color contamination. This layer-aware attention mechanism significantly enhances robustness in real-world footwear try-on scenarios. Our contributions are summarized as follows:

- We formally define the VFTON problem, reporting the unique issues of Data Scarcity, Viewpoint Misalignment and Background-induced color distortion across real-world scenarios.
- We curate the first shoe try-on dataset, containing 7305 pairs of high-quality multi-view garment-model samples, paving the way for subsequent footwear try-on studies.

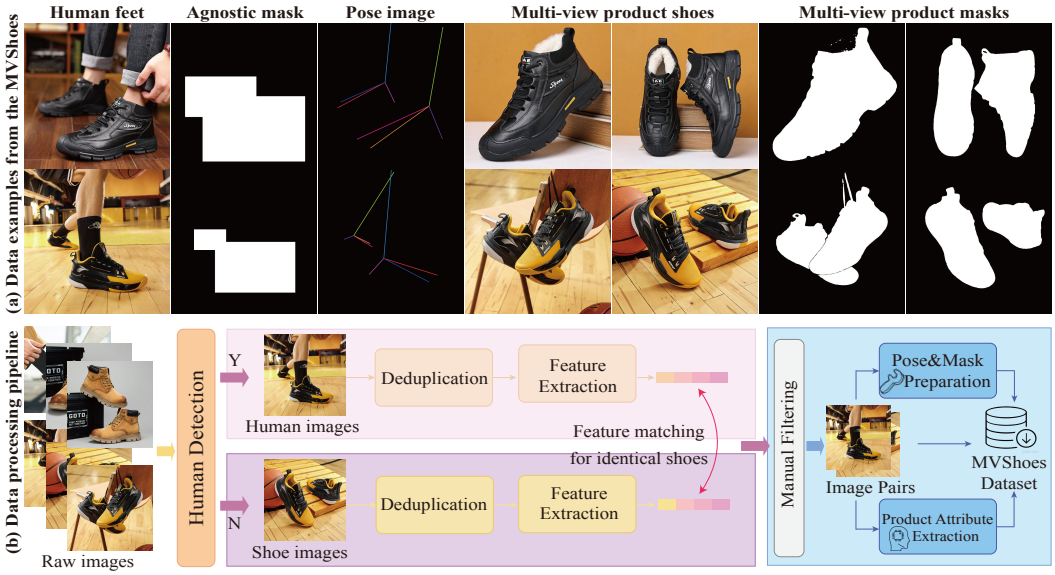


Figure 2: We illustrate examples of the MVShoes dataset in (a) and data processing pipeline in (b).

- We introduce a dual-stream architecture with Multi-View Conditioning and LayeredRefAttention modules to address the challenges in VFTON.
- Both quantitative and qualitative experiments validate the superiority of our approach.

2 Related Works

Image-based Virtual Try-on. The virtual try-on task is concerned with synthesizing images of a person donning the designated garment with appropriate fit (21), while retaining salient characteristics of the original garment and person, given a pair of images depicting a person and a target garment. To execute this task, numerous works (14; 26; 36; 4; 11; 20) have utilized Generative Adversarial Networks (GANs) (17) with two-stage strategy (37; 16; 66): (1) warping the clothing to the desired shape (5; 40) and (2) fusing the deformed clothing via try-on generator based on GAN. As significant progress in Text-to-Image diffusion models (22; 47; 25; 8) is witnessed in recent years, some works (6; 18; 45; 29; 38) have been motivated to incorporate pre-trained diffusion models (55; 52) as priors into virtual try-on task. Most recently, OOTDiffusion (65), IDM (9), and MMTryon (75) achieve garment feature extraction with a parallel U-Net and feed them through self-attention for enhanced integration. CatVTON (10) concatenates the garment and person images along the spatial dimension and feed them into a single UNet. Unfortunately, these methods serve for fabric clothing, intrinsically lack support for faithful try-on to 3D multi-view shoes with various materials, exhibiting severe color contamination from the background.

Controllable Diffusion Models. To attain conditional control in diffusion models, T2I-Adapter (46) and IP-Adapter (69) incorporate additional trainable modules to fuse condition features. Additionally, some investigations utilize various prompt engineering techniques (39; 68; 73) and implement cross-attention constraints (7; 64; 27; 74) to facilitate more controllable generation. Recently, the release of FLUX (33) and SD3 (15) has sparked renewed interest in DiT architectures (51) with subsequent conditional frameworks such as Flux.1 Fill (34) and Flux.1 Redux (35) achieving inpainting and object injection capabilities through enhanced encoder designs. However, these methods remain constrained by coarse-grained conditioning mechanisms, which fail to ensure generation fidelity—particularly in VFTON tasks with high geometric variability and stringent fidelity requirements.

3 MVShoes Dataset

We collect the first footwear try-on dataset which contains a total of 7305 annotated high-resolution image triplets, each consisting of two product shoe images from different views, and a corresponding try-on result. The dataset features diverse shoe categories and includes foot-specific pose landmarks and shoes-agnostic masks to facilitate precise alignment.

3.1 MVShoes Collection

All images are sourced from the Internet and include a diverse range of model scenarios such as full-body models, half-body models, foot close-ups, as well as a comprehensive variety of shoe categories, effectively meeting practical application needs. Initially, we collected approximately 25,000 raw shoe product image sets, each consisting of 2 to 5 images per shoe, including standalone product images and images of people wearing the shoes from different perspectives.

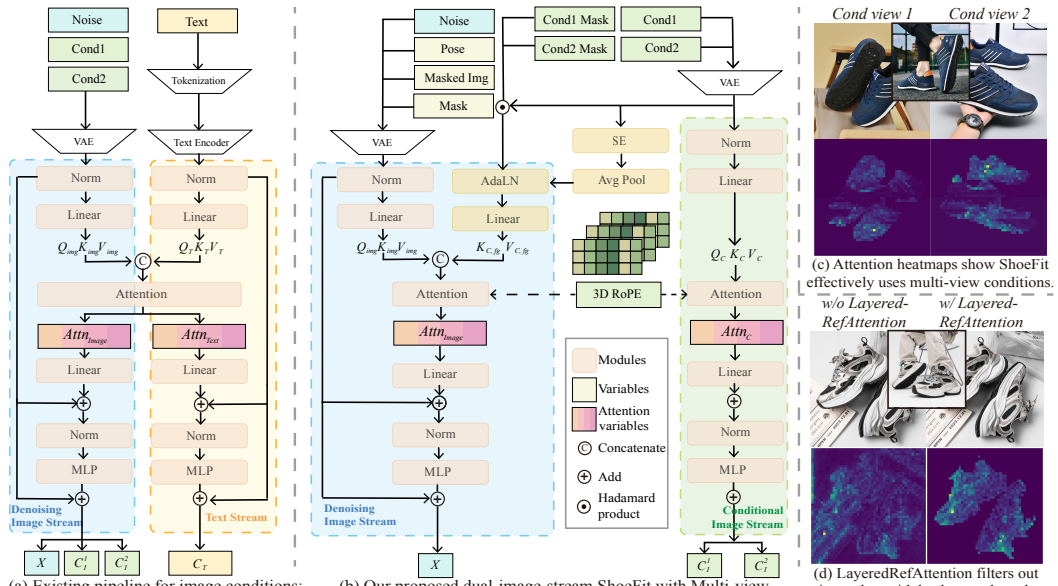
Data Processing. We first employ Qwen-VL (3) to differentiate between product footwear images (without humans) and try-on images featuring human models. Subsequently, we use the CLIP model (53) to extract image features and set a similarity threshold to eliminate duplicate images within the same raw image set. Next, we segment the shoe region in each image, match the DINO (43) features and assess their similarity to filter those data pairs with different footwear colors and styles. This process results in the formation of preliminary human-footwear try-on triplets, followed by manual filtering to eliminate any errors or oversights. We then remove visually blurred images and compile statistics on shoe categories and try-on scenarios. Finally, following the method proposed in DWPose (67) and Grounding DINO (43), we extract foot poses and shoe-agnostic masks. This process culminates in the creation of high-resolution, category-comprehensive try-on triplets. The data processing pipeline is illustrated in Fig. 2 (b).

3.2 Dataset Statistic

Rich Scenes. In conventional VTON tasks (37; 12), model images typically consist of full-body or upper-body views. However, given that footwear occupies a relatively small proportion of the body in VFTON tasks, full-body and half-body models are less effective in showcasing shoe details. To address this, we simulate the distribution of footwear model scenes observed on e-commerce platforms and curate 4 model scenarios: top-down foot close-up, horizontal foot close-up, half-body model, and full-body model, as illustrated in Fig. 3, where horizontal foot close-ups constitute the largest proportion. In such cases, accurate preservation of side details is particularly challenging, especially when compounded by viewpoint misalignment.

Figure 3: The category distribution of MVShoes.

Diverse Footwear Categories. We also prioritize the comprehensiveness of footwear categories. By annotating shoe attributes (2), we establish a two-level product category label set and ensure that its distribution aligns with everyday use, as depicted in Fig. 3. Notably, we observe that environmental lighting often affects the visual appearance of footwear, where reflections introduce background color contamination, particularly in categories such as boots, leather shoes, and high heels.



(a) Existing pipeline for image conditions: integrate image conditions by concatenation

(b) Our proposed dual-image-stream ShoeFit with Multi-view Conditioning and LayeredRefAttention Module

(c) Attention heatmaps show ShoeFit effectively uses multi-view conditions.

(d) LayeredRefAttention filters out mismatches with background patches, thereby preventing color confusion.

Figure 4: We illustrate existing and our DiT framework comparison in (a)-(b), and heatmaps visualization for component validation of Multi-view Conditioning and LayeredRefAttention in (c)-(d).

4 Method

An overview of the ShoeFit is presented in Fig. 4 (b). Given a human image X and target shoe images $C_I^{(n)}$, ShoeFit is aimed to generate an authentic try-on image. The backbone of ShoeFit employs the FLUX.1-Lite (33), with a customized dual-image-stream architecture in Sec. 4.1: denoising image stream and a copy of it as as conditioning branch to replace text modality. The features of human images and multi-view shoes are then integrated through Reference Concatenation and LayeredRefAttention modules, which are described in Sec. 4.2 and 4.3.

4.1 Dual-image-stream DiT Framework

The original FLUX (33) is a text-to-image model composed of a series of stacked MM-DiT blocks. Earlier works (60; 61) directly concatenate Image/Text condition tokens with noisy image tokens such as $[X; C_I; C_T]$, relying on a single denoising stream to finish condition extraction and denoising, while the other deal with text modality, as shown in Fig. 4(a). However, for VFTON, the generated try-ons are primarily determined by the given shoe images rather than limited semantic prompts. Thus, we remove the text stream from FLUX, achieving approximately 40% parameter savings, including those for T5 (54) and CLIP models (53). Instead, we replicate a portion of the denoising stream's architecture and weights to initialize the conditioning image stream, establishing a dual-image-stream framework. This design decouples feature preservation of shoe details from the high-quality generation process, allocating additional effective model capacity to enhance overall fidelity performance compared to direct token concatenation approaches (60; 61).

4.2 Multi-view Conditioning

In VFTON, the information about shoes from the try-on view remains incomplete due to view mismatch and 3D-to-2D projection occlusion (63). The inherent misalignment between source product images and target human images leads to unfaithful and physically implausible try-on results. Observing that e-commerce platforms often provide multiple-view product images (*e.g.*, front, side, top views), we introduce a multi-view shoe image dataset and propose a multi-view conditional injection mechanism via Reference Concatenation and 3D RoPE. The additional perspectives comprehensively

encompass more aspects of the shoe, as illustrated in Fig. 4 (c), significantly reducing uncertainty during generation and delivering high-fidelity and commercially viable synthetic results.

Reference Concatenation. Inspired by CatVTON (10), we adopt a simple yet effective multi-view injection strategy: concatenate multi-view conditioning shoe images along the spatial dimension as: $C_I^{cat} = [C_I^1, C_I^2, ...C_I^N] \in \mathbb{R}^{b \times N \cdot l \times c}$, where b s the batch size, l denotes the number of image tokens per conditioning image, and c is the feature channel. The concatenated multi-view latent C_I^{cat} is then fed into the conditioning image stream for feature extraction and interaction with the target human image. Concurrently, the denoising stream receives 4 concatenated components with $16 + 16 + 1 + 16$ channels: the noisy human latents X_t, the latent shoe-agnostic masked image $E(X_{masked})$, the resized agnostic mask X_m, and the foot pose landmarks $E(X_{pose})$, where $E(\cdot)$ represents VAE (30) encoding. The effectiveness of Reference Concatenation stems from the global attention interactions in attention layers. Intuitively, during the attention, each *Query* patch in the human image accesses and matches the most semantically relevant *Key-Value* pairs from the full set of concatenated multi-view product features, as illustrated in Fig. 4 (c). As a result, our method ensures consistent alignment between all regions in the target pose and their corresponding details in the product images, thereby achieving high-fidelity virtual try-on.

3D RoPE. However, naive Reference Concatenation introduces position misalignment in the relative positional relation of conditioning images. FLUX employs an essentially 2D RoPE scheme (58) to encode spatial relative relationships along the height and width dimensions, where the inner product in attention depends solely on the token embeddings x_m, x_n and their relative distances $m - n$:

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n). \tag{1}$$

When multi-view conditioning images are naively concatenated, discontinuous semantic jumps emerge at the concatenation boundaries, where tokens with small relative distances may correspond to semantically distant regions (*i.e.*, different views of the footwear).

To address this issue, we extend the 2D RoPE to three dimensions. Specifically, $E_{posID} = \text{concat}[E_{posID}^n; E_{posID}^h; E_{posID}^w] \in \mathbb{R}^{H \times W \times 3}$, where the first channel E_{posID}^n serves as a *index mask* to separate different conditioning views, while the second and third channels $E_{posID}^{\{h,w\}}$ encode spatial positions within each image along the height and width dimensions, respectively. Within each channel of $E_{posID}^{\{n,h,w\}}$, frequency encoding is performed given a base frequency of $\theta = 10000$.

$$\omega_d = \frac{1}{\theta^{2d/D}}, \text{for } d = 0, 1, ..., \frac{D}{2} - 1, \quad v_{\{n,h,w\}} = E_{posID}^{\{n,h,w\}} \cdot \omega_d \in \mathbb{R}^{H \times W \times \frac{\{D_n, D_h, D_w\}}{2}}, \tag{2}$$

where hype-parameters D_n, D_h, D_w is predefined as $16, 56, 56$. Subsequently, we obtain positional rotation matrices R_{1D} for each of the n, h, w channels internally. By concatenating these matrices along D dimension, we construct 3D RoPE matrix $R_{3D} \in \mathbb{R}^{H \times W \times \frac{(D_n + D_h + D_w)}{2} \times 2 \times 2}$ as follow:

$$R_{1D}^{\{n,h,w\}} = \begin{bmatrix} cos(v_{\{n,h,w\}} & -sin(v_{\{n,h,w\}}) \\ sin(v_{\{n,h,w\}}) & cos(v_{\{n,h,w\}}) \end{bmatrix}, \quad R_{3D} = \text{concat}[R_{1D}^n; R_{1D}^h; R_{1D}^w] \tag{3}$$

4.3 LayeredRefAttention Module

We observe that the appearance of the synthesized shoes is highly susceptible to contamination from the background colors of the conditioning images. This phenomenon arises from the complex reflective material of footwear interacting strongly with environmental lighting, termed **Background-induced Color Distortion**. Specifically, the issue arises when *Query* patches from the human image inadvertently match *key-value* pairs originating from the background regions of the conditioning image, as illustrated in Fig. 4 (d). A naive solution involves masking out background regions during attention, but this would ***neglect critical illumination cues embedded in the background***, which are essential for ***accurately inferring the shoe's intrinsic appearance and texture*** (32).

To address this challenge, we propose a layer-aware attention module called **LayeredRefAttention** to explicitly differentiate between background and foreground content in the conditioning images, as illustrated in Fig. 4 (b). By leveraging background features to modulate foreground shoe features, our method decouples the intrinsic shoe appearance from its background context (28). Subsequent denoising stream attention computations are restricted to interactions between the foreground shoe content and the target human image, effectively eliminating background illumination effects while

Table 1: Quantitative comparison on MVshoes. ShoeFit significantly surpasses all baselines.

View Num.	Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	DISTS ↓
1-View	FLUX.1-Fill (34)	0.725	0.201	24.44	2.221	0.1293
	FLUX.1-Fill-Redux (35)	0.724	0.199	24.67	2.086	0.1274
	OOTDiffusion-SDXL (65)	0.724	0.194	22.60	2.345	0.1114
	CatVTON-Flux-Lite (10)	0.745	0.185	22.71	3.327	0.1179
	ShoeFit-1V	<u>0.772</u>	<u>0.159</u>	<u>21.07</u>	<u>0.944</u>	0.1059
2-View	CatVTON-Flux-Lite-2V (10)	0.756	0.171	22.64	3.318	<u>0.1040</u>
	ShoeFit-2V	**0.780**	**0.149**	**20.18**	**0.485**	**0.0941**

avoiding background color contamination. Specifically, for $n \in \{1, ..., N\}$ conditioning view $C_I^{(n)}$, we employ a Squeeze-and-Excitation (SE) block (24) followed by global average pooling over spatial dimensions to compute channel-specific weights:

$$P^{(n)} = AvgPool(SE(C_I^{(n)}) \in \mathbb{R}^{b \times c}. \tag{4}$$

Subsequently, we employ two linear layers $F(\cdot)$ to extract background modulation parameters and modulate foreground shoe features, filtering irrelevant environmental lighting and background reflection to ensure faithful material preservation of shoes as:

$$\beta_{scale}^{(n)}, \beta_{shift}^{(n)} = F(P^{(n)}), \quad C_{fg}^{(n)} = (1 + \beta_{scale}^{(n)}) \cdot (LN(C_I^{(n)} \odot M_{fg}^{(n)})) + \beta_{shift}^{(n)}, \tag{5}$$

where $LN(\cdot)$ means layer normalization and $M_{fg}^{(n)}$ represents the binary shoe masks for conditioning product image. The attention variables $\{K, V\}_{c,fg}^{(n)}$ extracted from N views of modulated foreground conditioning feature are concatenated with the corresponding $\{Q, K, V\}_{img}$ from human image for subsequent denosing-stream attention:

$$Q_{C,fg}^{(n)}, K_{C,fg}^{(n)}, V_{C,fg}^{(n)} = QKV(C_{fg}^{(n)}), \tag{6}$$

$$Attn_{image} = \text{softmax}(\frac{Q_{img} \cdot \text{concat}(K_{img}; K_{C,fg}^{1,...,N,\top})}{\sqrt{d}}) \cdot \text{concat}(V_{img}; V_{C,fg}^{1,...,N}). \tag{7}$$

Meanwhile, the conditioning-stream branch performs attention computation using unmasked conditioning latents as:

$$Q_C^{(n)}, K_C^{(n)}, V_C^{(n)} = QKV(LN(C_I^{(n)})), \quad Attn_C = \text{softmax}(\frac{Q_C^{1,...,N} \cdot K_C^{1,...,N,\top}}{\sqrt{d}}) \cdot V_C^{1,...,N}. \tag{8}$$

To clarify our contributions, we omit the distillation strength normalization and final residual connections in the attention layer of FLUX.1-Lite (33) that remain unchanged from the original implementation. In summary, LayeredRefAttention effectively enables material-background decoupling by modulating foreground shoe features to recover intrinsic color properties, resulting in texture-preserving footwear try-on synthesis, as illustrated in Fig. 4 (d).

5 Experiments

5.1 Experimental Setup

Implementation Details. We employ the FLUX.1-Lite (33) as the backbone and customize a dual-image-stream architecture as in Sec. 4.1. Our experiments are carried out on MVShoes at a resolution of 768×768, with 6305 pairs for training and 1000 pairs for testing. The model is trained for 7 days on 8 80GB-A100 GPUs with DeepSpeed (1) ZeRO-2 to reduce memory usage, at a batch size of 4. All parameters of DiT are trainable, using an AdamW optimizer (44) with a constant learning rate of $3e-5$. At inference time, we run ShoeFit on a single A10 GPU for 25 steps. The data augmentation follows the same protocol as in StableVTON (29). Please refer to Appendix for more details.

Figure 5: Comparisons on the hard scenes, where ShoeFit delivers high-fidelity and rubost results.

Figure 6: Ablation study. We highlight the improvements by all three components of method in red.

Evaluation Protocols. Following evaluation in prior VTON works (45), we measure reconstruction accuracy by LPIPS (72), SSIM (62) and DISTS (13) in a paired setting given ground truth images, and authenticity of unpaired synthesized images by FID (50) and KID (59) without ground truth.

Baselines. As a subtask of inpainting, we first select two widely used inpainting FLUX variants as baselines: FLUX.1-Fill (34) and FLUX.1-Fill-Redux (35). The former involves concatenating product images with human images on the spatial dimension, similar to CatVTON (10). The latter enhances the conditioning stream with SigLip (70) to incorporate product shoe images. Additionally, we compare ShoeFit to two state-of-the-art methods in the VTON field: OOTDiffusion (65) based on the SDXL (52) and CatVTON (10) built upon the FLUX.1-Lite. All models are implemented as

originally deployed for single-view product shoe conditional generation. Considering CatVTON's capability to easily extend to multi-image injection, we further train a version of CatVTON to simultaneously inject two product views, the same as ShoeFit, denoted as CatVTON-FLUX-Lite-2V. Please refer to Appendix for more implementation details.

5.2 Qualitative Results

Fig. 5 provides a qualitative comparison between ShoeFit and the baselines on MVShoes, addressing challenges such as complex structures, background interference, and difficult target poses. For a fair comparison, we report the results of the single-view conditioning version of all methods. Notably, we observe that OOTDiffusion (65) tends to exhibit extreme color variations, while FLUX.1-Fill-Redux (35) and FLUX.1-Fill (34), possibly affected by the limitation of only LoRA (23) being trainable (r=64), often produce visually plausible results that do not strictly adhere to the input conditions. In contrast, ShoeFit accurately generates structural details of the shoes (rows 1 and 2), decouples environmental lighting (row 3), avoids background confusion (row 4), and demonstrates robustness to uncommon poses (row 5). Please refer to the Appendix for more results.

5.3 Quantitative Results

For a fair comparison, we report the results of one/two product view conditioning generation separately in Tab. 1. FLUX.1-Fill-Redux (35) and FLUX.1-Fill (34) achieve the worst FID and DISTS scores, while OOTDiffusion (65) and CatVTON (10) show complementary performance in unpaired and paired evaluations. ShoesFit, benefiting from the additional specialized parameters provided by the Dual-image-stream DiT framework and background decoupling enabled by the LayeredRefAttention, significantly surpasses all baselines under both single-view and multi-view conditional generation.

5.4 Ablation Study

To validate our technical contributions, we define a Vanilla Model that uses FLUX.1-Lite as the backbone, deploying both a text-stream and an image-stream as in Fig. 4 (a). Multi-view Conditioning and LayeredRefAttention Module are removed, using only one footwear image as conditioning.

Dual-image-stream DiT Framework. Since the text-stream only provides high-level semantics, its utility is limited in VFTON, which demands high image fidelity. Therefore, we first remove the text-stream and replicate a portion of the denoising stream's architecture and weights to initialize the conditioning image stream, effectively decoupling the conditional feature extraction from the denoising process, while keeping the Vanilla Model's other settings unchanged(denoted as "+ D-i-s"). As shown in Tab. 2 and Fig. 6, the vanilla model often generates artifacts and incorrect structures facing high-frequency textures due to the limitation in effective network capacity and the weakness in coupling of conditional images and noise. This demonstrates that the Dual-image-stream DiT framework plays a fundamental role in enhancing the model's expressive capabilities.

LayeredRefAttention Module. Observing the visual contamination from background and environmental lighting, we further integrate the LayeredRefAttention into the "+ D-i-s" version, denoted as "+ D-i-s + LRA", which effectively decouples the intrinsic appearance of the shoe from its surroundings. As illustrated in Tab. 2 and Fig. 6, the model without LayeredRefAttention often fails to accurately reproduce the natural colors of the shoe where environmental lighting or shadows of product images are prominent, resulting in outputs that are mixed with environmental light or shadows. Additionally, mismatches with background patches in Attention can introduce colors confusion from the background content, such as slippers being contaminated by floor colors in Column 3 of Fig. 6. In contrast, the one with "LRA" effectively avoids such color contamination, delivering superior try-on results.

To investigate the efficacy of simple background masking, we conducted an ablation study utilizing standard attention while removing the background from the product shoe images. Visually, this naive approach exacerbates color artifacts, particularly for highly saturated hues. We hypothesize that this occurs because removing the background discards crucial illumination cues embedded in the product image, which are essential for the model to infer the shoe's intrinsic material properties and texture. Given that the training set contains inherent color variations between in-studio product shots and on-foot images due to disparate lighting, forcing the model to operate without background context introduces lighting ambiguity. Consequently, the model struggles to maintain color constancy, leading to degraded try-on performance and reduced visual fidelity.

Table 2: Quantitative ablation study of each component in Section of Method.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	DISTS ↓
Vanilla Model	0.745	0.185	22.71	3.327	0.1179
+ D-i-s	0.761	0.174	21.57	1.493	0.1038
+ D-i-s + Standard Attn (w/ mask)	0.756	0.181	21.70	1.698	0.1056
+ D-i-s + LRA (w/ mask)	0.772	0.159	21.07	0.944	0.1059
+ D-i-s + LRA + MC (w/o 3D RoPE)	0.778	0.152	20.18	0.695	0.0971
+ D-i-s + LRA + MC **(Full)**	**0.780**	**0.149**	**20.18**	**0.485**	**0.0941**

Multi-view Conditioning System. While a single source product view often fails to capture the complete texture of footwear, try-on results can benefit from multiple source views, thereby effectively reducing uncertainty in generation. Thus, we introduce the Multi-view Conditioning pipeline using Reference Concatenation and 3D RoPE in the "+ D-i-s + LRA" version, denoted as "+ D-i-s + LRA + MC". As demonstrated in Tab. 2 and Fig. 6, the model without "MC", which uses only one product shoe image as input (yellow box), often generates unrealistic speculations and artifacts in unseen areas due to insufficient conditional information. In contrast, ShoeFit benefits from the Multi-view Conditioning system, producing details that are more faithful to the input across different views. Furthermore, we also conduct an ablation study upon 3D RoPE's impact in Tab. 2, demonstrating the performance gain we achieve using the proposed RoPE without incurring any additional computational overhead. Please refer to the Appendix for more ablation results.

5.5 Inference Time and Memory Usage

We evaluated the efficiency of our proposed enhancements by measuring inference time and memory footprint on an NVIDIA H20 GPU (with batch size of 1, 25 steps, and torch.bfloat16 precision). The results, presented in Table 3, reveal that the ShoeFit model with our Dual-image-stream (D-i-s) architecture substantially outperforms CAT-VTON in both speed and memory efficiency. This stems from removing the text stream and its computationally expensive T5 and CLIP encoders.

Table 3: Comparison of inference time and GPU memory usage across different architectures.

Method	**Inference Time per Image**	**GPU Memory Usage**
Cat-VTON (image + text streams)	22.27 seconds	35.5 GB
ShoeFit (Dual-image-stream, D-i-s)	17.44 seconds	22.0 GB
ShoeFit + D-i-s + LayeredRefAttention	21.15 seconds	28.5 GB

6 Conclusion

In conclusion, we formally define the VFTON problem and curate the first shoe try-on dataset, MVShoes, containing 7305 pairs of high-quality multi-view garment-model samples. Our proposed framework addresses the critical challenges of viewpoint misalignment and background-induced color distortion by leveraging a dual-stream architecture with Multi-View Conditioning and LayeredRefAttention modules. Through a comprehensive evaluation on the proposed MVShoes dataset, we demonstrate significant improvements in detail fidelity and texture consistency across diverse scenarios. This indicates that our method effectively mitigates background interference while preserving fine-grained shoe details across views, establishing a strong baseline for future VFTON research.

Broader impacts. With the ability to synthesize images, arises the risk that ShoeFit might be used for inappropriate purposes such as producing media that breaches intellectual property rights or privacy norms. Because of these risks, we strongly advocate for the conscientious use of this technology.

Limitation and future work. ShoeFit demonstrates substantial improvements in generating high-fidelity footwear try-on images across diverse scenes. However, certain limitations persist. Firstly, the model occasionally struggles with accurately rendering small logos and intricate text, particularly critical for e-commerce sellers. Secondly, we hope to introduce explicit 3D geometric and material priors in the future to achieve more robust multi-view representation and refined visual fidelity.

Acknowledgements

This work is supported by the Science and Technology Commission of Shanghai Municipality under research grant No. 25ZR1401187.

References

[1] Deepspeed. https://github.com/microsoft/DeepSpeed

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

[3] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966

[4] Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision (2022)

[5] Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence (1989)

[6] Chen, M., Chen, X., Zhai, Z., Ju, C., Hong, X., Lan, J., Xiao, S.: Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. arXiv preprint (2024)

[7] Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)

[8] Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)

[9] Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for virtual try-on. arXiv preprint arXiv:2403.05139 (2024)

[10] Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)

[11] Chopra, A., Jain, R., Hemani, M., Krishnamurthy, B.: Zflow: Gated appearance flow-based virtual try-on with 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[12] Davide, M., Matteo, F., Marcella, C., Federico, L., Fabio, C., Rita, C.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[13] Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)

[14] Dong, H., Liang, X., Zhang, Y., Zhang, X., Shen, X., Xie, Z., Wu, B., Yin, J.: Fashion editing with adversarial parsing learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8120–8128 (2020)

[15] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)

[16] Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems (2014)

[18] Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia (2023)

[19] Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., Liu, J.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. arXiv preprint arXiv:2501.15891 (2025)

[20] Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)

[21] Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[22] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (2020)

[23] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

[24] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)

[25] Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)

[26] Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1745–1753 (2019)

[27] Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15988–15998 (2023)

[28] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)

[29] Kim, J., Gu, G., Park, M., Park, S., Choo, J.: 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 (2024)

[30] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[31] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

[32] Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)

[33] forest labs, B.: Flux.1-dev. `https://github.com/black-forest-labs/flux` (2024), `https://github.com/black-forest-labs/flux`

[34] forest labs, B.: Flux.1-fill-dev. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev` (2024), `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`

[35] forest labs, B.: Flux.1-redux-dev. https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev (2024), https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev

[36] Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5549–5558 (2020)

[37] Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision (2022)

[38] Li, Y., Zhou, H., Shang, W., Lin, R., Chen, X., Ni, B.: Anyfit: Controllable virtual try-on for any combination of attire across any scenario. arXiv preprint arXiv:2405.18172 (2024)

[39] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)

[40] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)

[41] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)

[42] Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36**, 22226–22246 (2023)

[43] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2024), https://arxiv.org/abs/2303.05499

[44] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)

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

[46] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: 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. vol. 38, pp. 4296–4304 (2024)

[47] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

[48] NIKE, I.: Nike, inc. annual report. https://www.sec.gov/Archives/edgar/data/320187/000032018724000044/nke-20240531.htm (2024), https://www.sec.gov/Archives/edgar/data/320187/000032018724000044/nke-20240531.htm

[49] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)

[50] Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[51] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)

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

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

[54] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023), https://arxiv.org/abs/1910.10683

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

[56] Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)

[57] Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)

[58] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[59] Sutherland, J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference for Learning Representations (2018)

[60] Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024)

[61] Wang, H., Peng, J., He, Q., Yang, H., Jin, Y., Wu, J., Hu, X., Pan, Y., Gan, Z., Chi, M., et al.: Unicombine: Unified multi-conditional combination with diffusion transformer. arXiv preprint arXiv:2503.09277 (2025)

[62] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)

[63] Wikipedia: 3d projection. https://en.wikipedia.org/wiki/3D_projection (2025), https://en.wikipedia.org/wiki/3D_projection

[64] Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)

[65] Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779 (2024)

[66] Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)

[67] Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation (2023), https://arxiv.org/abs/2307.15880

[68] Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023)

[69] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)

[70] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343

[71] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

[72] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[73] Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X.: Controllable text-to-image generation with gpt-4. arXiv preprint arXiv:2305.18583 (2023)

[74] Zhang, X., Song, D., Zhan, P., Chen, Q., Xu, Z., Luo, W., Zhang, K., Liu, A.: Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. arXiv preprint arXiv:2408.06047 (2024)

[75] Zhang, X., Lin, E., Li, X., Luo, Y., Kampffmeyer, M., Dong, X., Liang, X.: Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. arXiv preprint arXiv:2405.00448 (2024)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made. And the claims match theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the limitations part in the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are accompanied by appropriate proofs and proper citations. Every claim is substantiated by empirical evidence or supported by referenced literature.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our contribution is a new dataset, a novel model architecture along with enhancement modules. All parameters and operational steps are detailed in the Experimental Section and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We introduce the first usable dataset in this area. Should our paper be accepted, we will release the dataset and benchmark to the public. While the code for the paper is not open-sourced, we have provided comprehensive instructions necessary for replication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: We have provided all the training and test details. Please refer to the "Experimental Setup" part in the main paper and other related sections in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We adhered to the common evaluation practices established by prior work in the similar field. Error bars are not reported because it would be too computationally expensive. We believe that the metrics reported in main paper prove the efficacy of the model we proposed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the sufficient information on the computer resources for each experiment in "Experimental Setup" part.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the broader impacts part in the main paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We propose a comprehensive and rigorous data cleaning pipeline to ensure safe and clean image pairs as described in the Section of MVShoes Dataset.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We follow the instructions by the creators of each asset. We also cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a detailed description of the collection and processing procedures for the new dataset, along with comprehensive reporting on its statistical characteristics and structure in the Section of MVShoes Dataset. Examples are included in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

- In Section A, we provide a more detailed discussion about the limitations of ShoeFit.
- In Section B, we provide additional details about the MVShoes datasets, including data processing and data statistics.
- In Section C, we provide preliminary knowledge about the FLUX (33) model, as well as how LayeredRefAttention modules are applied in the Single Stream DiT Block.
- In Section D, we provide additional details about the experimental setup, including baseline training, data augmentation, and hyperparameters.
- In Section E, we present a multitude of ShoeFit generated images, including more ablation results, more comparisons with the baselines, with additional results displayed in challenging scenarios.

A Limitations and Future Work

We propose ShoeFit, a dual-stream DiT framework that addresses the critical challenges of viewpoint misalignment and background-induced color distortion in VFTON by Multi-View Conditioning and LayeredRefAttention modules. However, certain limitations persist. Firstly, similar to the common constraints faced by existing generative models, the model occasionally struggles to accurately render small logos and intricate text due to their small portion in images and high-frequency variations. We illustrate this limitation in Fig. 7. These small logos are particularly critical for e-commerce sellers as they often convey brand information that consumers care about. Therefore, we plan to develop a detailed supplementary conditioning method in future research to pre-detect and enhance the injection of such patterns, fundamentally addressing this issue. Secondly, we aim to introduce explicit 3D geometric and material priors in the future to achieve more robust multi-view representation and refined visual fidelity.

Figure 7: Similar to the common constraints faced by existing generative models, ShoeFit occasionally struggles to accurately render small logos and intricate text due to their small portion in images and high-frequency variations.

B MVShoes Dataset Supplement

In the main body of the paper, we provide a brief description of the data processing pipeline and a rough visual presentation of the dataset statistics in terms of Rich Scenes and Diverse Footwear Categories due to space constraints. In this section, we offer more detailed information on the data processing pipeline to ensure reader comprehension. Additionally, we present the specific distribution statistics of MVShoes in a quantitative table format.

Data Processing Details. Given a dataset comprising raw images of various shoes and human models, we employ Qwen2.5-VL (3), which operates with 7 billion parameters, to distinguish between footwear images and human model images. The prompting framework utilized is as follows: "Analyze the provided image and determine whether it depicts a model image or a shoe image. A model image is defined as one that portrays a human model's lower body or legs adorned with shoes, while a shoe image solely comprises images of shoes without any human presence. Assign a value of '1' if it qualifies as a model image and '0' if it does not. "

Following this classification, we extract image features from the shoe images using the CLIP (53) model, implementing a similarity threshold of 0.9 to effectively eliminate duplicate images within the dataset. Subsequently, we perform segmentation by SAM (31) to isolate the shoe regions in each image, aligning the DINO (49) features to filter out shoes-model data pairs exhibiting inner similarities greater than 0.8.

This procedure facilitates the initial construction of human-footwear try-on triplets, which are then subjected to manual filtering to address any potential errors or oversights. Additionally, we exclude visually blurred images and compile statistical analyses pertaining to shoe categories and try-on scenarios. In accordance with methodologies established in DWPose (67) and GroundingDINO (43), we extract foot poses and shoe masks, ultimately resulting in the generation of high-resolution, category-comprehensive try-on triplets. All models referenced above are open-source, and their respective URLs are provided as follows:

- Qwen2.5-VL-7B (3): https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
- SAM (31): https://github.com/facebookresearch/segment-anything
- CLIP (53): https://github.com/openai/CLIP
- DINO (49): https://github.com/facebookresearch/dinov2
- DWPose (67): https://github.com/IDEA-Research/DWPose
- GroundingDINO (43): https://github.com/IDEA-Research/GroundingDINO

Dataset Statistics. The dataset features diverse shoe categories and rich human scenes. We present the quantitative distribution statistics of MVShoes in Tab. 4 and Tab. 5. We also provide more visual samples from MVShoes datasets in Fig. 8.

Pose Estimation for Shoe Try-On. Traditional open-source pose estimation methods (e.g., OpenPose, DWPose) in VITON pipeline, struggle to capture lower-body/foot poses in VFTON, which often only include partial leg shots or full lower-body views. To address this, we fine-tune the dwpose model on specially collected datasets containing diverse lower-body and close-up leg poses. This enables our model to generate accurate pose descriptions for various leg postures, significantly enhancing its generalization capability for shoe-related applications.

Generalizable Foot Mask Design. Given the diversity of shoe types (e.g., sneakers, boots, sandals) and their varying leg coverage, we propose a leg-pose-aware mask generation strategy. Unlike VTON's body-centric masks, our method combines leg pose information with bounding box expansion from shoe detection to create adaptive masks. This allows seamless cross-shoe category try-on capabilities, for example, a model wearing long boots or flip-flops can be transformed into sneakers while maintaining anatomical alignment and realistic occlusion. As illustrated in the first row in Fig. 11, this approach lays the foundation for robust, universal shoe try-on systems.

However, our mask generation strategy is not without its limitations. A failure case is presented in the second row of Fig. 5(b). Here, the region of the rear shoe is both heavily occluded and minimally visible, which leads to the failure of our keypoint and shoe detection modules. Consequently, this rear shoe area is erroneously classified as background, and only the front foot is designated as the inpainting region. This incomplete mask ultimately results in an unsuccessful virtual try-on. We contend that this is an inherent challenge for all methods reliant on an explicit masking strategy, and developing more robust solutions to handle such severe occlusion cases remains an important direction for our future work.

Shoe Expansion Phenomenon. The volumetric expansion of footwear observed during the try-on process is a naturally occurring phenomenon. In real-world datasets, it is common to find that shoes with stiffer tongues tend to bulge outwards when worn, resulting in a slight expansion of the shoe's overall volume. This effect is illustrated in Fig. 5, where the third column clearly shows a prominent bulge in the tongue area, leading to body expansion. This expansion occurs stochastically in real-world data pairs, as exemplified in the third row of Fig. 1(a), where the shoe upper in the second column exhibits a slightly thicker contour. Consequently, by effectively fitting to the training data distribution, our model learns to replicate this real-world phenomenon, accurately rendering the corresponding deformations inherent to the try-on process.

Table 4: Shoe Category Statistics. We report the sample count for each subcategory, with its proportion indicated in parentheses.

Primary Categories	**Subcategory**
Casual Shoes: 3158 (43.23%)	Lifestyle Casual Shoes: 803 (10.99%) Sneakers: 1428 (19.55%) Canvas Shoes: 141 (1.93%) Children Casual Shoes: 99 (1.36%) Men Casual Shoes: 687 (9.40%)
Athletic Shoes: 558 (7.64%)	Running Shoes: 322 (4.41%) Basketball Shoes: 121 (1.66%) Training Shoes: 45 (0.62%) Football Shoes: 70 (0.96%)
High-Heel Shoes: 895 (12.25%)	Classical High-heel Shoes: 340 (4.65%) Ladies Casual Shoes: 401 (5.49%) Mary Jane Shoes: 154 (2.11%)
Boots: 1045 (14.31%)	Ankle Boots: 416 (5.69%) Chelsea Boots: 56 (0.77%) High Boots: 153 (2.09%) Snow Boots: 183 (2.51%) Martin Boots: 237 (3.24%)
Sandals: 279 (3.82%)	Flip Flops: 183 (2.51%) Beach Sandals: 36 (0.49%) Strap Sandals: 60 (0.82%)
Dress Shoes: 389 (5.33%)	Loafers: 209 (2.86%) Formal Leather Shoes: 180 (2.46%)
Slippers: 981 (13.43%)	Clogs: 82 (1.12%) Thong Slippers: 18 (0.25%) House Slippers: 881 (12.06%)

Table 5: Human scene statistics.

Scene Types	Number of samples	Percentage
Top-down Foot	1719	23.53%
Horizontal Foot	4330	59.27%
Half-body Model	833	11.40%
Full-body Model	423	5.79%

C Method Supplement

C.1 Preliminary

FLUX.1 Our ShoeFit is an extension of Stable Diffusion 3 (55) and FLUX.1 (33), which are the most commonly used text-to-image diffusion models based on Flow Matching(41) and DiT(51). FLUX.1 (33) employs a variational autoencoder (30) (VAE) that consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ to enable image representations in the latent space. It is also equipped with rotary positional embeddings (RoPE)(58) and denoise-text-stream attention layers to improvement performance. FLUX.1 implements an actual two-dimensional RoPE scheme for encoding spatial positions in the latent space:

$$\omega_d = \frac{1}{\theta^{2d/D}}, \text{for} D = 0, 1, ..., D/2 - 1, \tag{9}$$

where θ is typically set to 10000. The position encoding applies a rotation matrix:

$$\begin{bmatrix} \cos(\omega_d \cdot \mathbf{pos}) & -\sin(\omega_d \cdot \mathbf{pos}) \\ \sin(\omega_d \cdot \mathbf{pos}) & \cos(\omega_d \cdot \mathbf{pos}) \end{bmatrix} \tag{10}$$

This rotation is applied to query and key vectors in the attention mechanism, enabling the model to capture relative positional relationships in the latent space.

Flow Matching Flow matching (41) aligns the flow of information between noise ϵ and data distributions by optimizing a velocity field u_t, which progressively converts noise into data over time. This technique ensures that the generative model maps the noise distribution to the actual data distribution in a structured manner. The text encoders (53) τ_θ are employed to deal with the given text prompt y. The flow matching loss is defined as follows:

$$\mathcal{L} = E_{t,p_t(z|\epsilon),p(\epsilon),y}\left[\|v_\Theta(z,t,\tau_\theta(y)) - u_t(z|\epsilon)\|^2\right]. \tag{11}$$

In this context, $v_\Theta(z,t,\tau_\theta(y))$ signifies the conditional velocity field determined by the weights of the neural network, while $u_t(z|\epsilon)$ represents the vector field created by the model to delineate the probabilistic trajectory between the noise and actual data distributions. The symbol E stands for the expectation, which involves either integration or summation over time t, latent variables z, conditions y, and noise ϵ. This expectation computes the mean of the squared differences for all conditions, ensuring that the model's performance is evaluated over numerous instances to yield a dependable estimate of its generative capability.

C.2 LayeredRefAttention in Single-stream DiT Block

The original FLUX is a text-to-image model composed of a series of stacked MM-DiT (double-stream) blocks, followed by a series of stacked single-stream DiT blocks. Due to space constraints, the framework of the LayeredRefAttention module shown in the main text is its structure within the double-stream DiT blocks. Here, we provide how it is used within a single-stream DiT block in Fig. 10. Similar to the way in double-stream blocks, we employ a Squeeze-and-Excitation (SE) block (24) followed by global average pooling over spatial dimensions to compute channel-specific weights:

$$P^{(n)} = AvgPool(SE(C_I^{(n)}) \in \mathbb{R}^{b\times c}. \tag{12}$$

Subsequently, we employ two linear layers $F(\cdot)$ to extract background modulation parameters and modulate foreground shoe features, filtering irrelevant environmental lighting and background reflection to ensure faithful material preservation of shoes as:

$$\beta_{scale}^{(n)}, \beta_{shift}^{(n)} = F(P^{(n)}), \quad C_{fg}^{(n)} = (1+\beta_{scale}^{(n)}) \cdot (LN(C_I^{(n)} \odot M_{fg}^{(n)})) + \beta_{shift}^{(n)}, \tag{13}$$

where $LN(\cdot)$ means layer normalization and $M_{fg}^{(n)}$ represents the binary shoe masks for conditioning product image. The subsequent operations are then performed as in regular single-stream DiT blocks.

D Implementation Supplement

D.1 Baseline Training Details

We here provide detailed descriptions of the training processes for the baseline models, focusing specifically on the training configurations. All models are trained on MVShoes at a resolution of 768×768, utilizing 6305 pairs for training and 1000 pairs for testing.

For Flux.1 Fill (34), we implement the training and inference pipeline by concatenating footwear images with human model images. In the trainable components, we apply LoRA (23) with a rank of 64 to all attention modules in the model and a Flux ControlNet (71) comprising 6 single layers and 6 double layers, injecting the poses at every denoising step. The model is trained for 6 days on 8 80GB-A100 GPUs using DeepSpeed ZeRO-2, with a batch size of 4. We utilize the AdamW optimizer, setting a constant learning rate of 3e-5 for training, and operate

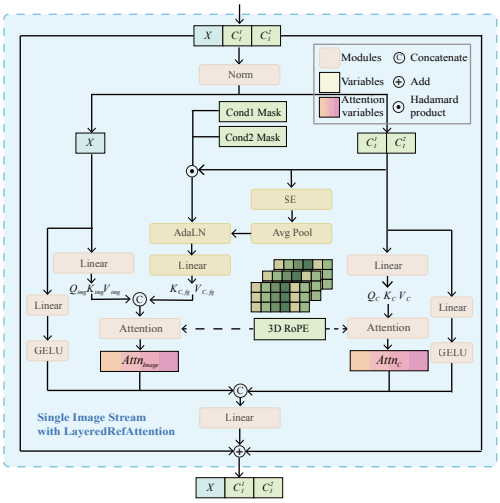


Figure 10: We provide how LayeredRefAttention is used within a single-stream DiT block.

the model on a single A100 GPU for 25 steps during inference.

For Flux.1 Redux (35), we adopt the same training settings as those used for FLUX.1-Fill. Additionally, we incorporate two linear layers for projecting SigLip (70) features of the product shoe images, training these two layers concurrently with the attention LoRAs and ControlNet. The parameters of SigLip remain frozen throughout the training process. We retain the same inference settings as employed in FLUX.1-Fill.

For OOTDiffusion (65), we execute the training and inference pipelines based on the official code. We also implement ControlNet to facilitate pose injection. The model is trained for 6 days on 4 80GB-A100 GPUs with DeepSpeed ZeRO-2, at a batch size of 8. During inference, the model is run on a single A10 GPU for 25 steps.

For CatVTON-Flux-Lite-2V, we maintain the text stream for FLUX-lite and concatenate footwear images with human model images. The model undergoes training for 7 days on 12 80GB-A100 GPUs utilizing DeepSpeed ZeRO-2, with a batch size of 2. All other training and inference configurations align with those of ShoeFit-2V.

D.2 Data Augmentation

We have implemented data augmentation techniques that could potentially enhance the model's generalization ability as well as its fidelity performance. Specifically, the data augmentation operations include (a) horizontal flipping of images, (b) resizing footwear and human figures through padding (up to 10% of the image size), (c) randomly adjusting the image's hue within a range of -5 to +5, and (d) randomly adjusting the image's contrast within a specified range (between 0.8 and 1.2 times the original contrast). Each of these operations occurs independently with a 50% probability. Moreover, these operations are simultaneously applied to both the footwear and model images.

D.3 LayeredRefAttention Hyperparameters

In the LayeredRefAttention layer, we primarily introduced a new Linear layer and an SE block. The Linear layer for the foreground follows the same dimensions as other linear layers within the module, which is the *hidden size*. The SE block takes input features with the same dimensions of *hidden size*, and internally, we use a compression rate of $reduction = 4$ for SE Block, as Linear$(channels, channels/4)$, ReLU$()$, Linear$(channels/4, channels)$.

E More visual results

Fig. 9 provides more results about the ablation study. We highlight the improvements by all three components of the method in red.

Fig. 11 provides more results on MVShoes for comparisons between the baselines and ShoeFit. For a fair comparison, we report the results of the single-view conditioning version of all methods. Our method substantially improves rendering fidelity and robustness under challenging real-world product shoes, establishing a new benchmark in high-fidelity footwear try-on synthesis.

Fig. 12 provides more results on MVShoes for inspection to demonstrate that ShoeFit synthesizes high-fidelity and detail-faithful try-on results.

Fig. 13 details two key limitations of our model. First, in the early training phase, the model has not yet mastered spatial reasoning, resulting in structurally distorted outputs (row (a)). Second, generating precise masks for tall footwear remains a challenge. As shown in the row (b), inaccurate masks for items like boots can cause significant visual artifacts or fitting errors.

Fig. 14 illustrate more results on out-of-distribution (OOD) samples not present in the MVShoes dataset. The high-quality try-on results demonstrate the model's notable stability and strong generalization capabilities.

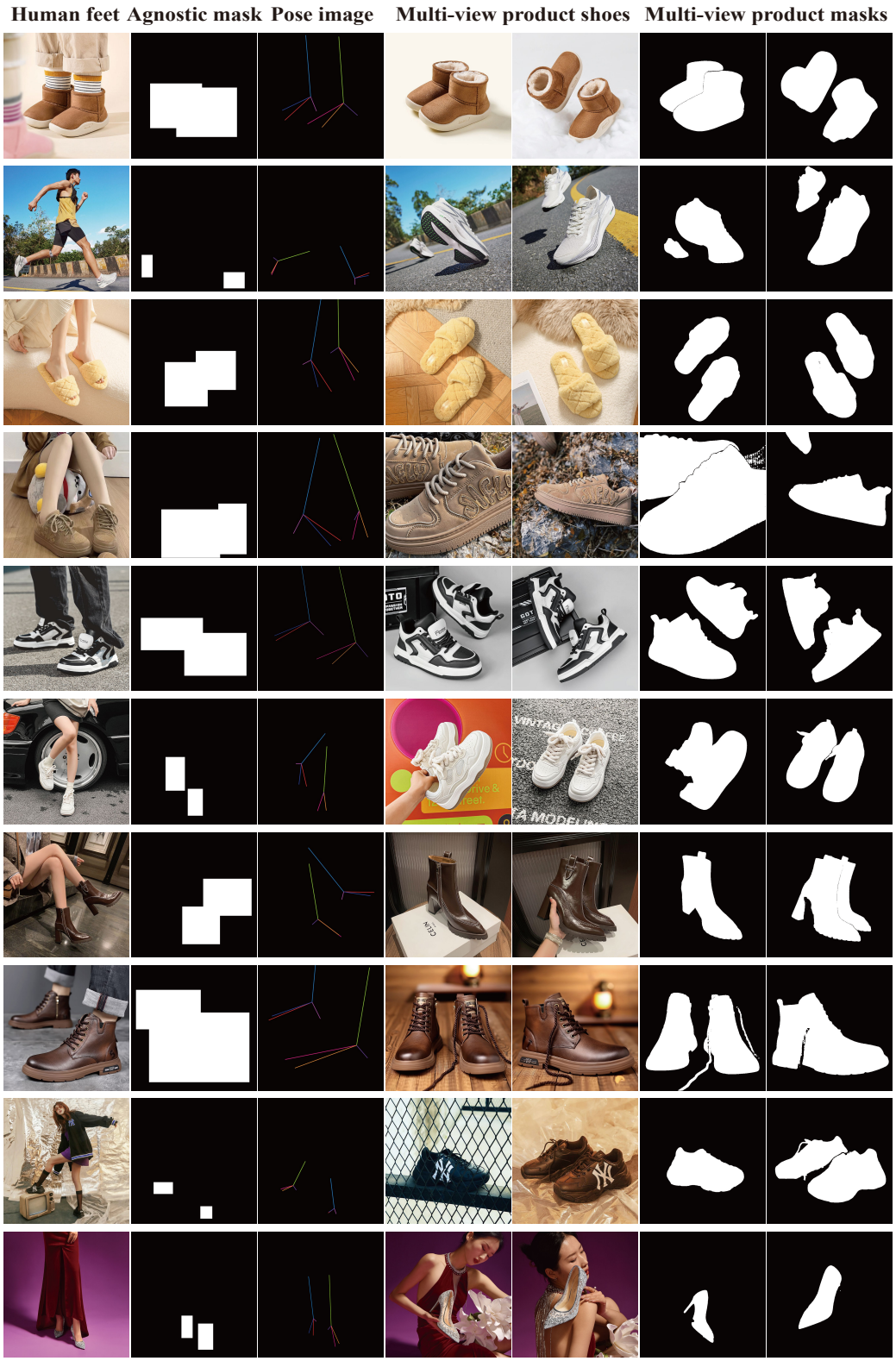


Figure 8: We provide more visual samples from MVShoes datasets.

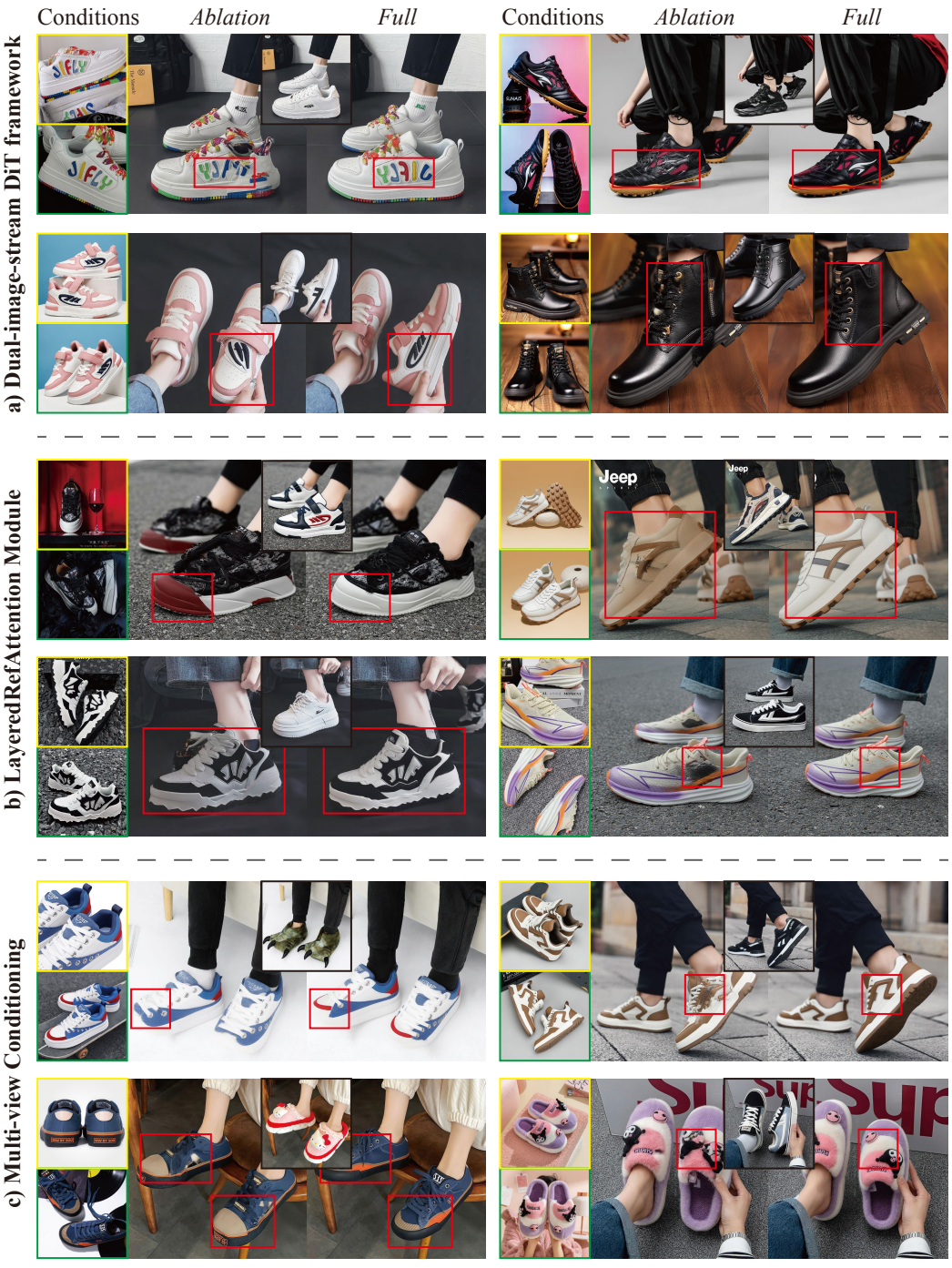


Figure 9: More visual results of the ablation study. We highlight the improvements by all three components of the method in red. Best viewed when zoomed in.

Figure 11: More visual comparisons on the MVShoes by ShoeFit. Best viewed when zoomed in.

Figure 12: More visual results on the MVShoes by ShoeFit. Best viewed when zoomed in.

Figure 13: More visual results on two primary failure modes of our method: (a) structural failure observed during early training stages, and (b) artifacts arising from inaccurate masks in long footwear scenarios.

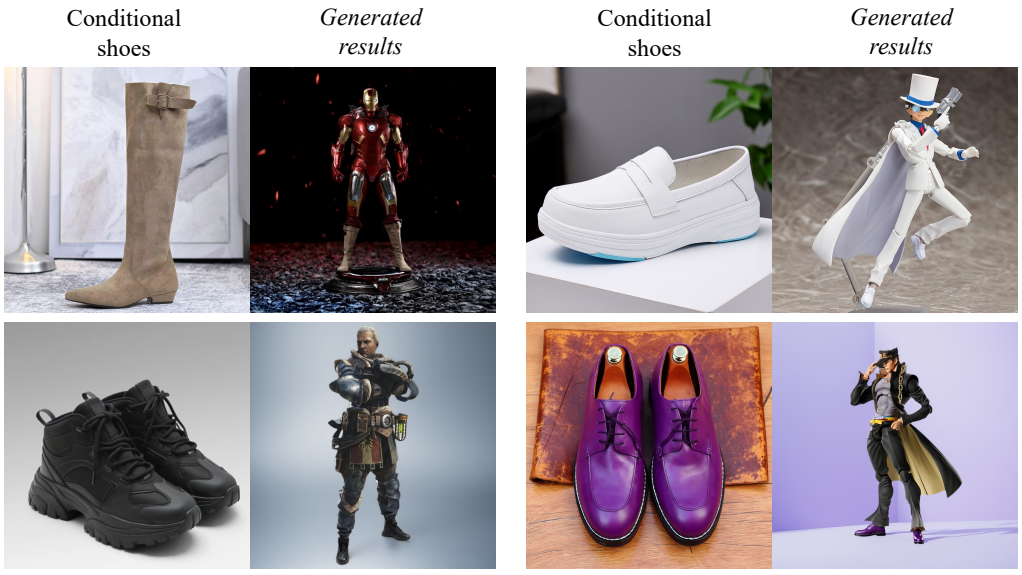


Figure 14: More visual results on out-of-distribution (OOD) samples not present in the MVShoes dataset. The high-quality try-on results demonstrate the model's notable stability and strong generalization capabilities.