HF-VTON: High-Fidelity Virtual Try-On via Consistent Geometric and Semantic Alignment
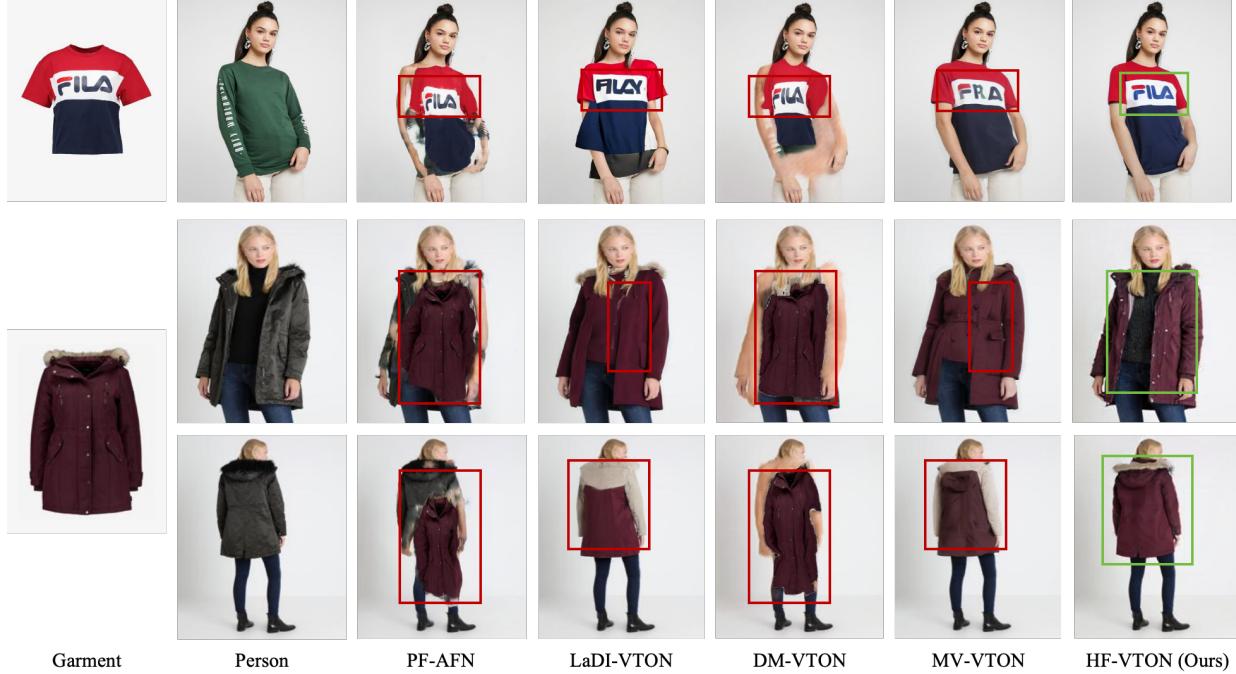
Ming Meng[1,*]   Qi Dong[1,*]   Jiajie Li[1]   Zhe Zhu[2]   Xingyu Wang[3]   Zhaoxin Fan[3,4,†]
Wei Zhao[1,†]   Wenjun Wu[3]

[1]School of Data Science and Media Intelligence, Communication University of China, Beijing, China
[2]Samsung Research America, USA
[3]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, School of Artificial Intelligence, Beihang University, Beijing, China
[4]Hangzhou International Innovation Institute, Beihang University, Beijing, China

mengming@cuc.edu.cn, zcdq@cuc.edu.cn, lijiajie@cuc.edu.cn,
ajex1988@gmail.com,20377014@buaa.edu.cn,zhaoxinf@buaa.edu.cn,
zhao_wei@cuc.edu.cn,wwj09315@buaa.edu.cn

Fig. 1. Our method (HF-VTON) outperforms existing state-of-the-art methods (e.g., PF-AFN [14], LaDI-VTON [23], DM-VTON [50], MV-VTON [34]) on two datasets: VITON-HD [1](single pose) and SAMP-VTONS (Pose 1 and Pose 2). Red boxes highlight errors, such as issues with logo details, geometric alignment, and texture fidelity, while green boxes emphasize the superior results produced by our approach.

Abstract—Virtual try-on technology has become increasingly important in the fashion and retail industries, enabling the generation of high-fidelity garment images that adapt seamlessly to target human models. While existing methods have achieved notable progress, they still face significant challenges in maintaining consistency across different poses. Specifically, geometric distortions lead to a lack of spatial consistency, mismatches in garment structure and texture across poses result in semantic inconsistency, and the loss or distortion of fine-grained details diminishes visual fidelity. To address these challenges, we propose HF-VTON, a novel framework that ensures high-fidelity virtual try-on performance across diverse poses. HF-VTON consists of three key modules: (1) the Appearance-Preserving Warp Alignment Module (APWAM), which aligns garments to human poses, addressing geometric deformations and ensuring spatial consistency; (2) the Semantic Representation and Comprehension Module (SRCM), which captures fine-grained garment attributes and multi-pose data to enhance semantic representation, maintaining structural, textural, and pattern consistency; and (3) the Multimodal Prior-Guided Appearance Generation Module (MPAGM), which integrates multimodal features and prior knowledge from pre-trained models to optimize appearance generation, ensuring both semantic and geometric consistency.

Additionally, to overcome data limitations in existing benchmarks, we introduce the SAMP-VTONS dataset, featuring multi-pose pairs and rich textual annotations for a more comprehensive evaluation. Experimental results demonstrate that HF-VTON outperforms state-of-the-art methods on both VITON-HD and SAMP-VTONS, excelling in visual fidelity, semantic consistency, and detail preservation. The code and dataset will be made publicly available upon acceptance: https://github.com/mmlph/HF-VTON/.

Index Terms—Virtual Try-On, Consistency Alignment, High-Fidelity, Multimodal Feature Fusion

I. INTRODUCTION

VIRTUAL try-on (VTON) has established itself as a cornerstone task within the realms of computer vision and virtual reality, driven by its profound potential to reshape industries such as e-commerce, fashion retail, and virtual human modeling. By seamlessly synthesizing garments onto human bodies in a highly realistic and natural manner—while rigorously preserving the fine-grained details of body pose, shape, and appearance—VTON addresses a fundamental challenge in personalized digital fashion visualization. This task not

*These authors contributed equally to this work.
†Corresponding author.

Figure caption row: Garment | Person | PF-AFN | LaDI-VTON | DM-VTON | MV-VTON | HF-VTON (Ours)

only promises to redefine the online shopping experience by offering unprecedented levels of interactivity and immersion but also serves as a benchmark for advancing core computer vision problems, including human parsing, pose estimation, and image synthesis. As the demand for intelligent, customer-centric virtual solutions accelerates, VTON stands at the forefront of research, unlocking new opportunities for practical deployment and fueling innovation in consumer-centric AI systems.

Existing virtual try-on methods can be broadly categorized into three main paradigms, each employing distinct modeling approaches to tackle this challenging task. GAN-based methods (e.g., CP-VTON [35]) exploit adversarial training between a generator and a discriminator to learn the mapping of garment regions, thereby enhancing the realism of the generated images. While these methods have achieved notable success in improving visual fidelity, they are prone to mode collapse and often exhibit structural misalignments when dealing with complex garment designs or diverse pose variations. U-Net-based methods (e.g., PF-AFN [14]), on the other hand, utilize encoder-decoder architectures with skip connections to reconstruct images, delivering strong texture preservation and enhanced garment details. However, these methods lack the flexibility to model non-rigid garment deformations, limiting their ability to handle intricate clothing shapes and movements. More recently, diffusion model-based methods (e.g., DM-VTON [50]) have emerged as a powerful alternative, leveraging iterative denoising processes to generate high-fidelity images. This approach has demonstrated state-of-the-art performance in terms of stability and detail retention, setting a new benchmark for VTON tasks. Despite the remarkable achievements of these above mentioned methods in improving visual quality and realism, significant challenges remain, particularly in generating spatially consistent images.

Specifically, as shown in Figure 1, existing VTON methods face three key consistency challenges: geometric alignment, semantic matching, and fine-grained detail preservation. First, geometric consistency requires garments to align accurately with the human body across diverse poses, yet current methods often fail to handle complex deformations, leading to unnatural distortions [55], [58]. Second, semantic consistency demands stable preservation of garment attributes such as style, color, and texture during pose transitions, but existing approaches frequently result in mismatches that undermine reliability [13], [23]. Lastly, fine-grained consistency focuses on retaining intricate garment details like wrinkles and textures, which are often lost or distorted under pose variations, reducing realism and user engagement [12], [35]. Though these challenges are typically addressed individually, there is a lack a unified framework to resolve them simultaneously, which hampers the stability and quality of cross-pose virtual try-on results. Moreover, the absence of large-scale, accurately annotated cross-modal datasets further exacerbates these challenges, hindering progress in addressing consistency issues comprehensively [11].

To address the aforementioned consistency challenges, we propose HF-VTON, a high-fidelity virtual try-on framework that systematically tackles geometric, semantic, and fine-grained consistency. The framework integrates multiple modules that work collaboratively to precisely align garments with the human body while preserving fine-grained garment details. Specifically, it includes the Appearance-Preserving Warp Alignment Module (APWAM), which leverages multi-scale appearance representation extraction and deformable

flow estimation networks to handle non-rigid garment deformations, ensuring geometric consistency across complex poses. The Semantic Representation and Comprehension Module (SRCM) constructs a structured garment attribute framework and utilizes large-scale image understanding models to enhance semantic consistency, ensuring accurate matching between garments and body poses. Finally, the Multimodal Prior-Guided Appearance Generation Module (MPAGM) integrates geometric conditions and multimodal semantic information to optimize fine-grained detail generation, preserving realistic garment features in high-fidelity virtual try-on images.

We evaluate HF-VTON on two datasets: VITON-HD (single pose) and SAMP-VTONS (cross-pose with Pose 1 and Pose 2). The results show that HF-VTON significantly outperforms state-of-the-art methods in visual fidelity, semantic consistency, and fine-grained detail preservation. Our approach handles complex garment deformations and diverse pose variations with superior accuracy, achieving consistent and realistic virtual try-on results across multiple poses. These findings demonstrate the effectiveness of our framework and its potential for real-world virtual try-on applications. Our contribution can be summarized as:

- We propose the HF-VTON framework, which systematically addresses the consistency challenges in virtual try-on from three key aspects: geometric alignment, semantic understanding, and fine-grained detail preservation.
- We design three novel modules to tackle these challenges: the Appearance-Preserving Warp Alignment Module for precise garment-body alignment across complex poses, the Semantic Representation and Comprehension Module for enhancing semantic consistency through a structured Garment Attribute Structure, and the Multimodal Prior-Guided Appearance Generation Module for preserving fine-grained garment details and generating high-fidelity virtual try-on images.
- We conduct extensive experiments on the VITON-HD and SAMP-VTONS datasets, demonstrating that HF-VTON significantly outperforms state-of-the-art methods in terms of visual fidelity, consistency, and detail preservation, validating its effectiveness for real-world virtual try-on applications.

II. RELATED WORK

The advancement of image-based virtual try-on technology has been a key focus in recent research. As a crucial component for realistic garment fitting and human interaction, it offers valuable insights for applications such as personalized fashion recommendations and virtual fitting rooms. In this context, we discuss the core approaches that are most relevant to our work, including Generative Adversarial Networks (GANs), U-Net architectures, and Diffusion Models, which have significantly contributed to improving the quality of virtual try-on tasks.

A. GANs-Based Virtual Try-On

Generative Adversarial Networks (GANs) [63] have played a critical role in virtual try-on technology, particularly in enhancing image realism and try-on effectiveness. Through adversarial training between the generator and discriminator, GANs can generate detailed and texture-rich garment images that naturally align with the human body.As a notable variant of GANs, StyleGAN [53] excels in high-fidelity image editing [3] and efficient image compression [4]. In recent years,

StyleGAN has been widely applied to virtual try-on tasks, including cross-domain garment translation, high-resolution image generation, and multimodal conditional control, significantly improving the visual quality of try-on images and laying the foundation for personalized garment recommendations and virtual digital avatar applications. As a milestone work, N. Jetchev et al. introduced the Conditional Analogy Generative Adversarial Network (CAGAN [54]), which, for the first time, learns the relationship between paired images in the training data through adversarial training and deep convolutional networks, generating plausible image pairs that follow learned relationships, even if unseen during training set. Ruiyun Yu et al. proposed the VTNFP [55] model, which adopts a three-stage deformation-segmentation-fusion architecture. This model deforms garments to fit the target pose, predicts body segmentation maps, and fuses the garments with the human image, utilizing GANs to achieve fine-grained virtual try-on image synthesis. The same year, Hyug Jae Lee et al. introduced the LA-VITON [56] virtual try-on network, which combines geometric matching and try-on modules, optimizing garment deformation through grid spacing consistency loss and occlusion handling techniques, and seamlessly generating the final try-on image using GANs. Han Yang et al. proposed the ACGPN [57] model, which introduces semantic layout for the first time, combining adaptive content generation and retention schemes, and enhancing thin-plate spline stability with second-order difference constraints to improve the handling of complex garment textures, generating high-quality virtual try-on images through adversarial training. Zhenyu Xie et al. proposed the GP-VTON [58] model, which combines local flow and global analysis deformation modules with dynamic gradient truncation strategies, addressing semantic information preservation and texture distortion issues under complex inputs, generating semantically consistent and complete garment images.

Despite significant advancements in virtual try-on using GANs, challenges remain in handling complex garment deformations, detail preservation, and diverse poses. Particularly, the quality of generated images is sensitive to imbalanced data distributions, and the hyperparameter tuning process is complex and costly. Consequently, researchers have turned to alternative approaches to overcome these limitations. U-Net-based virtual try-on methods, with their advantages in detail preservation, contextual integration, and local feature capture, have emerged as another important research direction in this field.

B. U-Net-Based Virtual Try-On

U-Net [59], introduced by Olaf Ronneberger et al., is initially designed for medical image segmentation. Its distinctive encoder-decoder architecture with skip connections has significantly advanced virtual try-on tasks. U-Net effectively captures fine-grained details, preserves key human features, and handles garment deformation, ensuring that generated images appear both natural and realistic. Xintong Han et al. proposed the VITON model [46], which introduces a garment-independent human representation to improve garment-body alignment. The model uses U-Net to synthesize a reference image conditioned on the target garment and human representation, further refining the coarse result with a refinement network to produce more natural try-on images. In the same year, Bochao Wang et al. introduced CP-VTON [35], building on VITON by incorporating a geometric matching module

that learns thin-plate spline transformations to adapt garments, followed by a try-on module that generates a composite mask to merge deformed garments with rendered images, reducing boundary artifacts and enhancing realism. As virtual try-on technology evolves, researchers continue to innovate in model diversity and practicality. Jin Hyun-woo et al. proposed Versatile-VTON [10], which combines explicit garment transformation networks and probabilistic models to enable versatile try-on for various garment types.In addition, Liu et al. proposed SPATT [5], which integrates U-Net into a spatial-aware texture transformer framework for high-fidelity garment transfer, demonstrating strong performance in preserving fine-grained garment details and handling pose variations.

While U-Net-based models excel at preserving details and alignment, their dependence on high-quality input images and difficulty in modeling complex garment deformations limit their robustness. To address these challenges, researchers have turned to diffusion models, which demonstrate greater robustness and stability in generating high-resolution images and preserving details, particularly in handling complex garment deformations and varied poses.

C. Diffusion Model-Based Virtual Try-On

In virtual try-on technology, diffusion models have emerged as a powerful alternative to GANs and U-Net, demonstrating exceptional performance in generating high-quality images, preserving details, and ensuring training stability.Their generalization ability has also been demonstrated in tasks such as image fusion [6], low-light image enhancement [7], and high color fidelity preservation [8], further validating their potential for high-fidelity virtual try-on. Particularly, diffusion models have shown significant advantages in handling complex garment deformations and diverse poses, paving new directions for the development of virtual try-on techniques. Inspired by random processes and physical diffusion phenomena, Jascha Sohl-Dickstein et al. first introduced the diffusion process into generative models [28], proposing a generative framework based on diffusion processes. Jonathan Ho et al. developed the Denoising Diffusion Probabilistic Model (DDPM [64]), which progressively restores noise data through a Markov chain, enhancing generative capabilities. Subsequently, Alexander Quinn Nichol et al. [27] optimized the model architecture and training strategy, reducing diffusion steps and significantly accelerating the generation process. Robin Rombach et al. [65] introduced the Latent Diffusion Model (LDM), which transfers the diffusion process to a lower-dimensional latent space, significantly reducing computational costs, accelerating generation speed, and improving computational efficiency and image quality. LDMs have been widely applied in the virtual try-on field. Building on this, Luyang Zhu et al. proposed the TryOnDiffusion architecture [29], which implicitly deforms garment through a cross-attention mechanism and integrates garment deformation with the human image as a unified process, effectively preserving garment details while handling pose and body shape variations. Junhong Gou et al. introduced the DCI-VTON model [13], which employs a two-stage process: first, the deformation module preserves local garment details, then the deformed garment is combined with the human image and noise is added as input to the diffusion model, ensuring that the generated image retains more detail. Additionally, Jeongho Kim et al. introduced StableVITON [36], which combines latent diffusion models to learn semantic correspondences in virtual try-on tasks. By incorporating ControlNet [37], this model leverages additional input conditions

to enhance the accuracy and flexibility of virtual try-on, offering a new solution for efficient and personalized virtual try-on. Mehmet Saygin Seyfioglu et al. proposed the DTC model [26], which enhances image conditioning within latent diffusion models. By integrating detailed reference image features into the latent feature map and utilizing perceptual loss, this model further preserves details, balancing inference speed and high-fidelity detail retention.

Recent advancements in diffusion models have demonstrated significant potential for text-guided image generation in virtual try-on applications. By iteratively reducing noise to synthesize high-fidelity images conditioned on textual inputs, diffusion models enable personalized try-on image generation aligned with user descriptions. Building upon Latent Diffusion Models (LDM), Davide Morelli et al. proposed LaDI-VTON [23], which integrates LDM with textual inversion to enhance image generation for image-based virtual try-on tasks. To mitigate reconstruction artifacts, LaDI-VTON introduces an augmented autoencoder with learnable skip connections, preserving details outside inpainting regions, and employs a forward textual inversion module for conditional refinement, ensuring texture consistency. Yuhao Xu et al. developed the OOTDiffusion model [25], which preprocesses input data to generate human body masks, encodes them into low-dimensional latent variables via a VAE encoder, and combines these variables with Gaussian noise as inputs to the denoising network. Concurrently, a U-Net architecture learns garment-specific details to synthesize realistic try-on images. Rui Wang et al. introduced StableGarment [24], which incorporates a garment encoder, adaptive self-attention layers, and a try-on ControlNet to stably preserve fine-grained textile textures while generating stylized images, achieving precise virtual try-on results. These diffusion-based approaches highlight the growing need for fine-grained control and multimodal conditioning in VTON tasks—a direction we pursue with the multimodal prior-guided generation module in HF-VTON.

D. Challenges in Existing Virtual Try-On Methods

Despite significant advancements in image quality and detail preservation, existing virtual try-on (VTON) methods still face several consistency challenges when handling complex garment deformations and diverse poses: **(i) Dependence on Input Quality and Geometric Alignment:** Existing methods heavily rely on high-quality input images and precise geometric alignment, which perform poorly in handling non-rigid garment deformations and pose variations, leading to generated images that lack detail and semantic consistency; **(ii) Multimodal Data Processing and Semantic Consistency:** Although diffusion models show advantages in image generation stability and detail preservation, many approaches still fail to effectively address the challenges of multimodal data processing, garment-body alignment optimization, and maintaining semantic consistency throughout the generation process; **(iii) Detail Preservation and Distortion:** Existing methods often suffer from distortion or information loss when processing complex garment patterns and details, especially when integrating fine-grained garment features and multimodal data. To address these consistency challenges, we propose the HF-VTON framework, which integrates multiple modules to achieve unified handling of geometric, semantic, and detail consistency.

III. METHODOLOGY

A. Overview of the HF-VTON Framework

The HF-VTON framework addresses key challenges in high-fidelity virtual try-on (VTON), focusing on maintaining consistency across geometric alignment, semantic understanding, and fine-grained detail preservation. It comprises three core modules: the Appearance-Preserving Warp Alignment Module (APWAM), the Semantic Representation and Comprehension Module (SRCM), and the Multimodal Prior-Guided Appearance Generation Module (MPAGM). These modules operate in concert to enable precise garment fitting, photorealistic image generation, and robust consistency across diverse poses and garment types, as illustrated in Figure 2.

B. Appearance-Preserving Warp Alignment Module (APWAM)

The Appearance-Preserving Warp Alignment Module (AP-WAM) is designed to achieve precise geometric alignment and preserve appearance details between garment and target human body image in high-fidelity virtual try-on tasks. The module consists of two key components: the Multi-Scale Appearance Representation Extraction (MRE) and the Deformable Flow Estimation Network (DFEN).

Multi-Scale Appearance Representation Extraction (MRE) extracts fine-grained appearance details and high-level semantic features from garment and person images, critical for geometric alignment and visual consistency in high-fidelity virtual try-on tasks. To address garment deformation and appearance variations, MRE utilizes a pyramid convolution architecture that progressively captures multi-scale features, from local details to global shape information of both the garment and the person body. The first part of MRE, the **Pyramid Feature Extraction Network (PFEN)**, extracts multi-level features from the input garment image I_c and person image I_p, along with 3D human keypoint information I_{dense} and person segmentation results S_p. These inputs provide multi-modal data, enabling the extraction of fine-grained texture features F_l, such as wrinkles, stretching, and material textures. These low-level features serve as a foundation for capturing high-level semantic features F_h that represent the overall shape of the garment and body structure. The PFE helps the model understand the garment and person body at various scales, supporting the subsequent appearance flow estimation and geometric alignment. To enhance the fine-grained details of the garment image, MRE integrates I_{dense} and S_p, ensuring precise spatial alignment between garment and person images. This spatial information improves geometric accuracy and ensures a better fit between the garment and person body. In the fusion stage, the low-level and high-level features F_l and F_h are combined into a unified multi-scale representation F_{final}, with weights controlled by hyperparameters α and β. This process preserves both fine-grained details and semantic information, providing accurate features for the subsequent appearance flow estimation and geometric alignment.

Deformable Flow Estimation Network (DFEN) is designed to refine the appearance flow between garment and the person body, addressing complex deformations. It consists of N cascaded flow networks, each progressively enhancing the flow accuracy. At each layer, DFEN takes as input the previous layer's flow p_{i-1}, the current pyramid features c_i, and calculates the correlation loss $\text{corr}(p_i, c_i)$ to ensure semantic consistency. These features are then passed through a deformable convolution module to predict the flow residual Δflow_i:

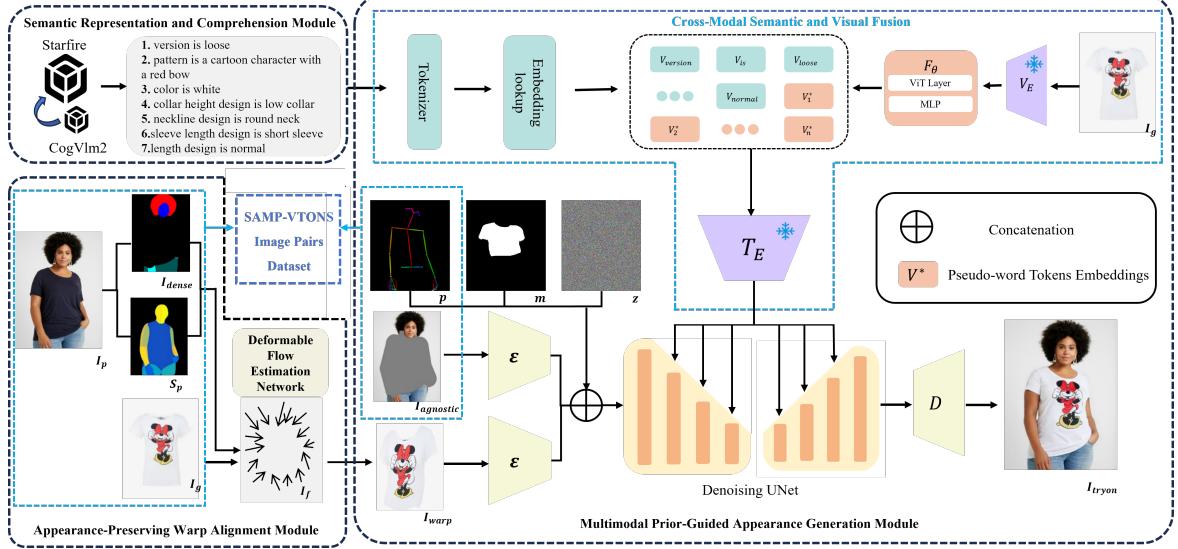$$\Delta\text{flow}_i = \text{DeformConv}(p_i, c_i, \text{corr}(p_i, c_i))$$

Fig. 2. Overview of the proposed HF-VTON framework for high-fidelity virtual try-on. HF-VTON integrates three core modules to achieve consistency across geometric alignment, semantic understanding, and fine-grained detail preservation. The Appearance-Preserving Warp Alignment Module (APWAM) ensures accurate garment-body alignment in various poses by capturing both local and global geometric features. The Semantic Representation and Comprehension Module (SRCM) improves semantic consistency by enhancing garment attribute understanding and aligning them with body poses. The Multimodal Prior-Guided Appearance Generation Module (MPAGM) combines geometric priors with semantic and visual cues to generate high-fidelity virtual try-on images, maintaining garment structure and appearance consistency across different garment types and poses.

Deformable convolution introduces dynamic sampling by learning offsets Δp_n, enabling the network to adapt to local deformations such as wrinkles and stretching. This flexibility allows the network to better handle non-rigid garment deformations compared to traditional convolution with fixed kernels. The mathematical formulation of deformable convolution is:

$$y(p_i) = \sum_{n \in R} w(p_n) \cdot x(p_i + p_n + \Delta p_n)$$

where p_i is the flow field of the current layer, p_n is the sampling position of the convolution kernel, Δp_n is the learned offset, enabling the convolution kernels to adjust based on the input's geometric changes. The refined flow Δflow_i is then added to the previous flow p_{i-1}, resulting in the updated flow p_i:

$$p_i = p_{i-1} \oplus \Delta \text{flow}_i$$

By progressively refining the flow and incorporating deformable convolution, DFEN accurately captures garment deformations across different poses, improving the visual quality and alignment for high-fidelity virtual try-on.

C. Semantic Representation and Comprehension Module (SRCM)

The Semantic Representation and Comprehension Module (SRCM) aims to improve the semantic understanding and representation of garments in virtual try-on. Existing approaches generally focus on low-level visual feature alignment but often overlook the semantic structuring of garment attributes and their alignment with human poses. To address these limitations, SRCM integrates three key components: Semantic Attribute Structuring (SAS), Dual-Model Collaborative Text Generation (DMTG), and Semantic-Aligned Multi-Pose Virtual Try-On Dataset Construction (SAMP-VTONS). This integration enhances semantic understanding in virtual try-on and ensures consistent alignment between garments and human poses across various poses.

Semantic Attribute Structuring for Garments (SAS) module provides a systematic and structured representation of upper-body garment attributes. Building on the hierarchical attribute tree from the FashionAI dataset [42], SAS refines and extends the attribute representation for garments. While FashionAI offers basic attribute classification for upper and lower body garments, it lacks detailed coverage of fine-grained garment features such as fit, pattern, and color. To address this, SAS introduces three additional attribute categories—fit, pattern, and color and refines garment attributes into seven main categories: **Fit** (slim, loose, straight), **Pattern**, **Color**, **Neckline Design** (low, mid, high), **Collar Design** (V-neck, deep V-neck, round neck, square neck, irregular), **Sleeve Length** (sleeveless, short sleeve, long sleeve), and **Shirt Length** (high waist, normal, long, extra-long). This extension provides a more comprehensive and precise representation of the appearance and design elements of the garment, allowing accurate textual descriptions for virtual tests and improving semantic understanding and precision.

Dual-Model Collaborative Text Generation for Garments (DMTG) addresses key challenges in the generation of garment text for virtual tests, specifically in terms of precision, speed, and compliance. While commercial models (e.g., ChatGPT-4.0 and Wenxin Yiyan) offer high accuracy, their API costs limit practical deployment. Open-source models (e.g., CogVlm2) provide better customization but suffer from slower generation speeds. In contrast, production-grade APIs (e.g., Starfire) excel in speed but are constrained by content safety filters, which prevent the generation of text descriptions for images containing sensitive content such as flags, logos, or exaggerated garment patterns. To mitigate these issues, we propose a dual-model collaborative strategy, using the Starfire model as the primary generator for fast response times, while leveraging CogVlm2 as a secondary model to handle edge cases and provide more detailed text descriptions. Additionally, we implement a regular expression parser and

pruning mechanism to remove API metadata and non-semantic tokens, further enhancing the precision and efficiency of the generated descriptions.

Semantic-Aligned Multi-Pose Virtual Try-On Dataset Construction (SAMP-VTONS) is a critical component of our proposed virtual try-on framework, addressing the limitations of existing datasets in terms of pose diversity and cross-modal semantic alignment. Existing datasets, such as VITON-HD [1] and Dress Code [9], predominantly focus on single-pose images and fail to adequately capture the adaptability and deformation of garment across different poses, limiting the performance and accuracy of virtual try-on models in multi-pose scenarios. While the FashionTryOn [11] dataset provides two poses per subject, making it suitable for multi-pose virtual try-on tasks, it still lacks sufficient pose diversity and cross-modal semantic alignment. To address these issues, we construct the SAMP-VTONS (Semantic-Aligned Multi-Pose Virtual Try-On) dataset, which incorporates multi-pose images with textual annotations. Based on FashionTryOn, the dataset employs DensePose [43] and OpenPose [44] for accurate pose estimation, ensuring precise alignment of each garment image with human images in various poses. Next, we use the Human Parse [45] method for fine-grained segmentation of the human body, labeling the distinct body parts, while the Parse Agnostic method [1] removes identity information and non-try-on areas, minimizing background interference in the virtual try-on process. Finally, our DMTG generates accurate textual descriptions for each garment image, forming a semantically aligned triplet of image, pose, and text. The SAMP-VTONS dataset consists of 21,104 training images and 7,487 test images. Each sample includes: human pose images, garment images, garment mask images, garment attribute textual descriptions, pose estimates (DensePose and OpenPose), human body part segmentation (Human Parse), and identity-agnostic processing (Human Agnostic). By enhancing pose diversity and achieving semantic alignment across images, poses, and texts, it provides superior data support for the training and evaluation of multi-pose virtual try-on models.

D. Multimodal Prior-Guided Appearance Generation Module (MPAGM)

The Multimodal Prior-Guided Appearance Generation Module (MPAGM) is a pivotal component designed to enhance the quality and accuracy of virtual try-on by effectively integrating geometric and multimodal geometric priors with advanced generative techniques. Existing methods typically struggle with insufficient integration of textual, visual, and geometric features, limiting their ability to generate realistic garment fittings with fine-grained details. To address these shortcomings, MPAGM introduces three tightly integrated submodules: Geometric-Conditioned Multimodal Fusion (GCMF), Cross-Modal Semantic and Visual Fusion (CSVF), and Dual-Conditioned Guidance Appearance Generation (DGAG).

Geometric-Conditioned Multimodal Fusion integrates garment deformation features, identity-agnostic human representations, and structured pose information to provide spatially consistent geometric priors for the diffusion model to enhance the accuracy and consistency of virtual try-on image generation. First, we perform feature encoding on the target garment deformation features I_{warp} using the pre-trained Stable Diffusion Encoder E_{SD}, extracting multi-scale features and producing the encoded deformation feature E_{warp}:

$$E_{warp} = E_{SD}(I_{warp}) \in \mathbb{R}^{c_e \times h \times w}, \quad h = \frac{H}{8}, w = \frac{W}{8}, c_e = 4$$

where c_e is the number of channels in the latent space and h, w are the downsampled spatial dimensions. Next, we perform similar encoding on the identity-agnostic human features $I_{agnostic}$ using E_{SD}, yielding the encoded identity-agnostic feature $E_{agnostic}$. To incorporate spatial constraints, we utilize the binary mask m and pose map p, which provide explicit guidance for the region of interest in the human body. These constraints allow the generation model to focus on the relevant parts of the image, reducing unwanted artifacts. Additionally, we introduce the noise latent variable $z \sim \mathcal{N}(0, I)$. We concatenate the aforementioned five input features into a single representation space $\Gamma_{gp} = [E_{warp}, E_{agnostic}, m, p, z]$ providing rich geometric priors for the diffusion model.

Cross-Modal Semantic and Visual Fusion integrates semantic and visual representations by constructing a joint embedding space, aligning and merging high-quality textual descriptions with garment visual features to ensure semantic consistency and fine-grained detail preservation. Initially, the target garment image $C \in \mathbb{R}^{3 \times H \times W}$ is passed through the CLIP visual encoder V_E to extract visual features E_{vis}, capturing key garment attributes such as texture, shape, and color. Then E_{vis} are processed by the network F_θ, consisting of a Vision Transformer (ViT) and a Multilayer Perceptron (MLP), mapping the extracted features to the CLIP text space and generating the predicted pseudo-word embedding V_{pseudo}. Concurrently, the garment's high-quality textual description, generated by the SRCM module, is processed through Tokenizer and Embedding Lookup to produce the text embedding vector $T_{feature} = \{T_{version}, T_{pattern}, T_{color}, T_{collar}, T_{neckline}, T_{sleeve}, T_{length}\}$. By aligning and transforming the respective feature spaces, textual and visual representations are effectively fused, yielding a joint embedding $Y \in \mathbb{R}^{d_{joint}} = \text{Concat}(V_{pseudo}, T_{feature})$. Finally, Y is passed to the CLIP text encoder T_E, generating the final text embedding $E_{text} \in \mathbb{R}^{d_{text}}$. This embedding serves as the conditioning signal for the diffusion model, guiding the generation of the virtual try-on image.

Dual-Conditioned Guidance Appearance Generation uses the denoising UNet as the core network, taking as input the geometric condition information Γ_{gp} from the GCMF and the semantic visual information E_{text} from the CSVF. During each denoising step, E_{text} serves as the guiding signal, ensuring that the generated image aligns with the target textual description. As the de-noising process progresses, the network gradually synthesizes the virtual try-on image I_{trvon} by iteratively integrating both image and text features. To optimize the generation process, the model trains the denoising network ϵ_θ by minimizing the following loss function:

$$L_{DM} = \mathbb{E}_{I_g, p, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(I_g, t, \tau_\theta(Y))\|^2 \right] \quad (1)$$

where t represents the diffusion timestep, $\tau_\theta(Y)$ is the conditional encoding of the text description Y, and ϵ_θ is the denoising network. The core objective of this loss function is to minimize the difference between the noisy image I_g and the predicted network output $\epsilon_\theta(I_g, t, \tau_\theta(Y))$, while accounting for the integration of image and text feature spaces. By optimizing this loss, the model is able to generate high-fidelity images that closely align with the input textual descriptions. It effectively guides the model to generate images consistent with both geometric features and semantic information. This dual-conditioned guidance improves the precision of the generated virtual try-on images, ensuring that the output matches the expected garment style, fit, and appearance.

IV. Experiments

A. Experimental Setup

Datasets. We conduct extensive experiments on the widely used VITON-HD benchmark and our newly proposed SAMP-VTONS dataset to comprehensively evaluate the robustness of HF-VTON across various challenging scenarios. (i) The VITON-HD dataset, collected by Choi et al. [1], is specifically designed for virtual try-on tasks and includes 13,679 front-facing images of women paired with corresponding upper-body garment images. It provides pose annotations such as OpenPose and DensePose, as well as garment masks for accurate fitting. Although the original dataset does not include textual descriptions, we generate them using our DMTG module from SRCM. The dataset is split into 11,647 training samples and 2,032 test samples, with a resolution of 786×1024, resized to 384×512 for consistency in our experiments. (ii) The SAMP-VTONS dataset is an extension of the FashionTryOn dataset, consisting of 21,104 training samples and 7,487 test samples. Each sample includes two images in different poses, accompanied by 200,137 paired textual descriptions. The images are of resolution 384×512, designed to evaluate multi-pose garment fitting and semantic alignment tasks. Both datasets are essential for evaluating the performance of HF-VTON and comparing it with existing methods, with VITON-HD focusing on pose-annotated garment fitting and SAMP-VTONS offering a multi-pose garment fitting scenario.

Evaluation Metrics. To thoroughly assess the performance of our model, we employed multiple evaluation metrics. Specifically, we use the Structural Similarity Index (SSIM) [2] to quantify image similarity in terms of brightness, contrast, and structural consistency. To capture perceptual similarity, we use Learned Perceptual Image Patch Similarity (LPIPS) [60], which reflects the human visual system's sensitivity to image differences. Additionally, we incorporate Fréchet Inception Distance (FID) [61] and Kernel Inception Distance (KID) [62] to assess the quality of the generative model. These metrics compare the feature distributions of generated images against real images, providing a comprehensive evaluation of the model's capability in generating high-quality outputs.

Implementation Details. The experiments are conducted on a system with an A800 GPU (80GB of memory) and a Linux-based environment, using Python 3.8, PyTorch 2.0.1, and CUDA 11.8 for both training and testing. The core modules, APWAM, CSVF, and DGAG, are implemented as follows: For APWAM, which generates deformed garments as input for synthesis, we use the Adam optimizer [47] to train the deformation network for 100 epochs, with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is set to 0.00005, decaying linearly to zero from epoch 50. The loss function hyperparameters are set to $\lambda_1 = 0.2$, $\lambda_2 = 0.01$, and $\lambda_3 = 6$. For CSVF, within the MPAGM, we generate 16 pseudo-word embeddings and traine the model for 150,000 steps with a learning rate of 0.00001. The optimizer is AdamW, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay set to 0.01. The OpenCLIP ViT-H/14 model [48], pre-trained on the LAION-2B dataset [49], serves as the visual encoder. In the DGAG component, we train for 150,000 iterations, randomly masking text, deformed garment, and pose inputs with a probability of 0.2 to enhance robustness. During inference, a non-classifier guided technique [51] is applied to ensure high authenticity and consistency in the generated images under multiple constraints. Lastly, following [52], we utilize a fast variant of multicondition nonclassifier guidance, optimizing

Method	Paired				Unpaired	
	SSIM↑	LPIPS↓	KID↓	FID↓	KID↓	FID↓
VITON-HD Dataset						
PF-AFN(CVPR) [14]	0.845	0.159	1.641	26.11	2.083	30.37
VITON-HD(CVPR) [1]	0.856	0.081	0.404	11.69	0.445	13.52
LaDI-VTON(ACM MM) [23]	0.846	0.103	1.123	6.29	0.265	11.08
DM-VTON(ISMAR) [50]	0.850	0.151	1.335	21.85	1.596	24.82
MV-VTON(AAAI) [34]	0.873	0.060	0.135	6.76	0.427	12.32
HF-VTON (Ours)	0.852	0.050	0.024	5.51	0.196	10.12
SAMP-VTONS Dataset						
PF-AFN(CVPR) [14]	0.843	0.175	3.766	43.92	4.590	49.31
LaDI-VTON(ACM MM) [23]	0.852	0.101	0.382	9.26	0.771	13.10
DCI-VTON(ACM MM) [13]	0.886	0.073	0.189	5.47	0.307	8.15
DM-VTON(ISMAR) [50]	0.844	0.181	3.158	35.97	3.704	39.72
MV-VTON(AAAI) [34]	0.869	0.096	0.220	6.10	0.263	7.38
HF-VTON (Ours)	0.892	0.043	0.084	3.35	0.256	6.92

computational efficiency by ensuring that the final results' complexity is independent of the number of input constraints.

B. Comparisons with State-of-the-art Methods

We conduct a comprehensive evaluation of HF-VTON against state-of-the-art virtual try-on methods on the VITON-HD and SAMP-VTONS datasets. In the VITON-HD dataset, we first compare HF-VTON with the baseline method VITON-HD [1], followed by comparisons with PF-AFN, LaDI-VTON, DM-VTON [50] and MV-VTON [34], ensuring consistent training protocols and evaluation metrics across all methods. Next, in the SAMP-VTONS dataset, we compare HF-VTON with five advanced methods: PF-AFN [14] (geometric alignment), DCI-VTON [13] (diffusion-based synthesis), LaDI-VTON (latent diffusion modeling), DM-VTON (multimodal fusion) and MV-VTON (multiview rendering), which represent different technical paradigms in virtual try-on research.

Quantitative Comparison. Table I presents the comprehensive quantitative evaluation of HF-VTON on the VITON-HD and SAMP-VTONS datasets, covering paired and unpaired settings. In the paired setting on the VITON-HD dataset, HF-VTON achieves an FID of 5.51 and a KID of 0.024, representing a 12.4% improvement over LaDI-VTON (6.29) and an 18.5% improvement over MV-VTON (6.76). KID is reduced by 82.1% compared to MV-VTON (0.135 → 0.024), demonstrating the superiority of HF-VTON in image quality and consistency of the generative distribution. With an LPIPS of 0.050, HF-VTON outperforms LaDI-VTON (0.103) and MV-VTON (0.060), indicating enhanced preservation of texture detail. In the unpaired setting, HF-VTON maintains robust performance with an FID of 10.12 and a KID of 0.196, surpassing LaDI-VTON (FID: 11.08, KID: 0.265) and MV-VTON (FID: 12.32, KID: 0.427), highlighting its strong generalizability to pose variations and data distribution shifts.

On the single-pose scenario of the SAMP-VTONS dataset, HF-VTON achieves the highest SSIM of 0.892 and the lowest LPIPS of 0.043, outperforming DCI-VTON (SSIM: 0.886, LPIPS: 0.073) and MV-VTON (SSIM: 0.869, LPIPS: 0.096), excelling in structure preservation and texture fidelity. In the unpaired setting, HF-VTON further reduced the FID to 6.92 and KID to 0.256, surpassing MV-VTON (FID: 7.38, KID: 0.263) and DCI-VTON (FID: 8.15, KID: 0.307), further validating its superior image quality, structural restoration, and detail retention in standard single-pose tasks.

Qualitative Comparison. Figure 3 presents a qualitative

Fig. 3. Qualitative comparison results on the VITON-HD dataset are presented in the following columns: Garment image, Person image, and results from five state-of-the-art virtual try-on methods. Red boxes highlight errors, while green boxes emphasize the result of ours.

comparison of HF-VTON with state-of-the-art virtual try-on methods on the VITON-HD dataset. The figure includes garment images (first column), person images (second column), and the results of six competing methods (columns 3 to 8). Red boxes highlight areas with poor performance, while green boxes emphasize the advantages of HF-VTON in detail restoration and alignment. Existing methods exhibit significant alignment errors, particularly in the shoulder and sleeve regions, resulting in unnatural garment-body fitting. In contrast, HF-VTON accurately aligns the garment with the body, ensuring a smooth transition of garment contours and poses. When handling complex textures such as stripes and patterns, existing methods often suffer from blurriness or distortion, especially in detail preservation. HF-VTON, however, maintains clear texture retention, particularly excelling in preserving pattern integrity and the natural folds of fabric. Overall, HF-VTON significantly improves image quality and structural consistency through its representation consistency framework, showcasing notable advantages in complex garment deformation, fine-grained detail fidelity, and precise garment-body alignment. In the single-pose scenario of the SAMP-VTONS dataset, HF-VTON also demonstrates significant advantages over existing methods, as shown in Figure 4. LaDI-VTON exhibits noticeable texture distortion, leading to a mismatch between the garment and the original image. Other methods suffer from considerable alignment errors in critical areas such as the shoulders and upper arms, resulting in unnatural garment-body fitting. HF-VTON achieves precise alignment between the garment and the body, particularly in challenging regions, ensuring a natural transition. While MV-VTON performs relatively well, it still exhibits discrepancies in texture detail preservation, especially in garment detail restoration. HF-VTON excels in maintaining fine details, particularly in handling complex garment deformations and texture fidelity, surpassing current methods.

C. Robustness Analysis

To evaluate the robustness of HF-VTON in multi-pose scenarios, we further analyze its performance on the SAMP-VTONS dataset, comparing single-pose (Pose 2) and multi-pose (Pose 1 + Pose 2) settings. The experiments cover both paired and unpaired settings, with results presented in Table II. The LPIPS difference between single-pose and multi-pose scenarios for HF-VTON is only $\Delta 0.0015$ ($0.04 \rightarrow 0.0415$), significantly lower than DCI-VTON ($\Delta 0.0095$) and MV-VTON ($\Delta 0.0175$), indicating superior texture consistency across poses. This demonstrates HF-VTON's ability to effectively handle pose variations while maintaining high texture fidelity and image quality. Additionally, the SSIM difference is reduced by 31.7% compared to DCI-VTON (2.05%), showing that HF-VTON preserves structural consistency in multi-pose scenarios, preventing reconstruction distortions or misalignment due to pose changes. These results highlight HF-VTON's robust adaptability and stability under diverse poses.

The visual comparison in Figure 5 demonstrates the robustness of HF-VTON on the SAMP-VTONS dataset across different poses (Pose 1 and Pose 2) and garment types. The solid blue boxes highlight the person images under Pose 2, while the dashed blue boxes correspond to the results of the comparison methods under Pose 2. The results show that HF-VTON excels in geometric alignment, particularly in Pose 2, where other methods exhibit significant alignment errors in critical regions such as the shoulders and upper arms, leading to unnatural garment-body fitting. For instance, LaDI-VTON suffers from misalignment in the shoulder area, while DM-VTON and MV-VTON also fail to align the garment properly with the body. In contrast, HF-VTON achieves precise alignment between the garment and the body, especially in difficult-to-align areas like the shoulders and upper arms, ensuring smooth transitions in garment contours and pose. Moreover, HF-VTON demonstrates superior performance in handling complex garment deformations (such as wrinkles and stretching). When dealing with wrinkled garments (e.g., the first and second rows in the figure), HF-VTON effectively
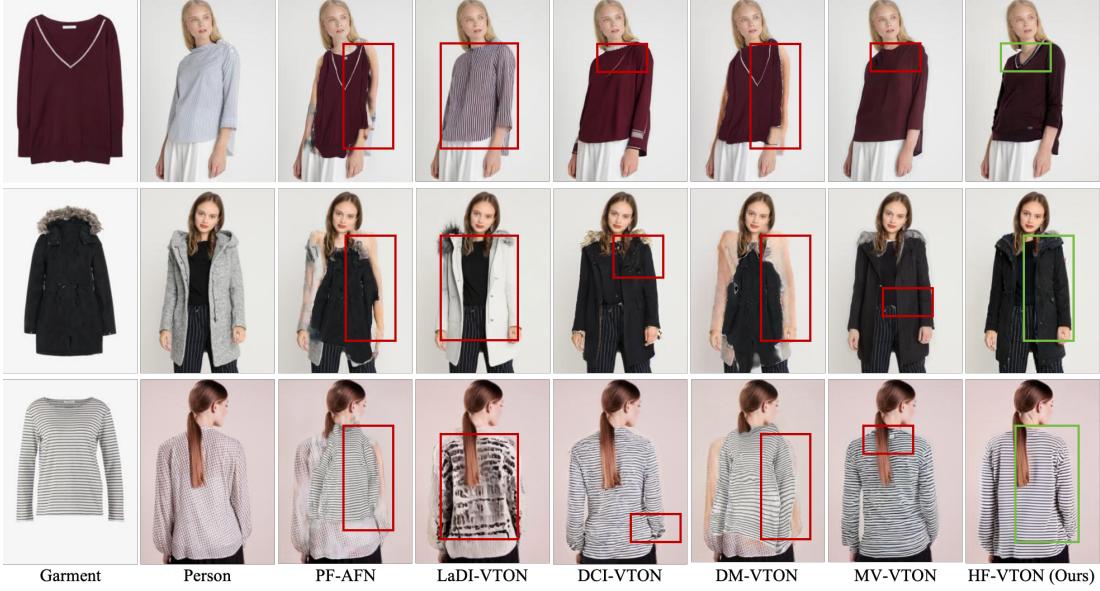
Fig. 4. Qualitative comparison results on the SAMP-VTONS dataset are presented in the following columns: Garment image, Person image, and results from five state-of-the-art virtual try-on methods. Red boxes highlight errors, while green boxes emphasize the result of ours.

TABLE II
ROBUSTNESS ANALYSIS OF STATE-OF-THE-ART VIRTUAL TRY-ON METHODS ON THE SAMP-VTONS DATASET WITH TWO POSES (POSE 1 CORRESPONDS TO THE SAMP-VTONS DATASET AS SHOWN IN TABLE 1, AND POSE 2). THE METRICS USE ↑/↓ TO INDICATE THAT HIGHER/LOWER VALUES CORRESPOND TO BETTER PERFORMANCE. THE TOP TWO RESULTS ARE HIGHLIGHTED IN THE FOLLOWING ORDER: RED FOR THE BEST RESULT AND GREEN FOR THE SECOND-BEST RESULT.

Method	Pose 2						Pose 1+Pose 2					
	Paired				Unpaired		Paired				Unpaired	
	SSIM↑	LPIPS↓	KID↓	FID↓	KID↓	FID↓	SSIM↑	LPIPS↓	KID↓	FID↓	KID↓	FID↓
PF-AFN(CVPR) [14]	0.860	0.186	6.207	64.17	6.576	65.42	0.8515	0.1805	4.9865	54.045	5.583	57.365
LaDI-VTON(ACM MM) [23]	0.895	0.102	0.449	8.928	1.348	18.81	0.8735	0.1015	0.4155	9.094	1.0595	15.955
DCI-VTON(ACM MM) [13]	0.927	0.054	0.172	5.44	0.202	6.31	0.906	0.0635	0.1805	5.455	0.2545	7.23
DM-VTON(ISMAR) [50]	0.864	0.191	5.282	54.45	5.535	55.05	0.854	0.186	4.22	45.21	4.6195	47.385
MV-VTON(AAAI) [34]	0.890	0.061	0.200	5.30	0.309	7.42	0.8795	0.0785	0.21	5.7	0.286	7.4
HF-VTON	0.920	0.040	0.082	3.73	0.198	5.86	0.906	0.0415	0.083	3.54	0.227	6.39

preserves the natural texture details of the garment, while other methods show poor texture retention, resulting in blurring or distortion.

D. Ablation Study

Through ablation studies, we systematically evaluate the contributions of the APWAM, SRCM, and MPAGM modules in enhancing image quality, geometric alignment, and texture fidelity.

Effectiveness of APWAM. To evaluate the effectiveness of APWAM, we conduct comprehensive ablation experiments on the VITON-HD and SAMP-VTONS datasets, analyzing performance under different settings. The experimental setups include: (1) baseline model (BS w/o), (2) multi-scale appearance representation extraction with deformable convolution (MRE w/), (3) deformable flow estimation network (DFEN w/), and (4) deformable convolution in both MRE and DFEN (M+F w/). Each metric corresponds to two columns representing the results for the VITON-HD and SAMP-VTONS datasets, respectively. As shown in Table III, the DFEN w/ configuration outperforms the others across several metrics. In terms of SSIM, DFEN w/ achieves 0.849 and 0.872 on the two datasets, respectively, showing improvement over the BS. For the LPIPS metric, DFEN w/ DC scores 0.066 and 0.083, lower than the BS, indicating improved detail fidelity. On the KID metric,

DFEN w/ achieves 0.14 on SAMP-VTONS, outperforming the BS and other settings, confirming its advantage in reducing generative distribution discrepancies. Regarding FID, DFEN w/ scores 6.93 and 5.3 on VITON-HD and SAMP-VTONS, respectively, showing a significant reduction in image distortion and better alignment between garment and person. As shown in Figure 6, in the absence of deformable convolution (BS w/o), the alignment between the garment and the person is significantly insufficient, especially in critical areas such as the chest and abdomen, where noticeable misalignments degrade the naturalness of the try-on effect. With the introduction of deformable convolution into MRE and DFEN, the alignment improves notably. Particularly in the DFEN configuration, the garment's contours and details are better preserved, and the handling of wrinkle areas is more accurate.

Effectiveness of SRCM. Based on the APWAM, we further validate the effectiveness of SRCM through ablation experiments on the VITON-HD dataset with different text configurations. The experimental setups include: no text (w/o Text), raw LaDI-VTON text guidance (w/ LaDI_Text), and our proposed text representation based on SAS+DMTG (w/ Proposed_Text). The results, shown in Table IV, demonstrate that w/ Proposed_Text achieves the best performance across all four quantitative metrics, particularly showing significant improvements in KID and FID. Specifically, the KID for w/

Fig. 5. Robustness analysis results on the SAMP-VTONS dataset with two poses (Pose 1 and Pose 2) are presented in the following columns: Garment image, Person image, and results from five state-of-the-art virtual try-on methods. The blue solid box highlights the person image for Pose 2, while the blue dashed box indicates the corresponding results from the comparison methods for Pose 2.

TABLE III
ABLATION STUDY ON THE PERFORMANCE OF DEFORMABLE
CONVOLUTIONS (DC) IN APWAM ON THE VITON-HD AND
SAMP-VTONS DATASETS: EFFECTS OF MULTI-SCALE APPEARANCE
REPRESENTATION EXTRACTION (MRE) AND DEFORMABLE FLOW
ESTIMATION NETWORK (DFEN). THE METRICS USE ↑/↓ TO INDICATE
THAT LARGER/SMALLER VALUES CORRESPOND TO BETTER
PERFORMANCE. THE TOP TWO RESULTS ARE HIGHLIGHTED IN THE
FOLLOWING ORDER: RED FOR THE BEST RESULT AND GREEN FOR THE
SECOND-BEST RESULT.

Settings	SSIM↑		LPIPS↓		KID↓		FID↓	
BS w/o	0.817	0.862	0.101	0.097	0.418	0.179	12.31	5.89
MRE w/	0.848	0.865	0.067	0.095	0.075	0.171	6.99	5.83
DFEN w/	0.849	0.872	0.066	0.083	0.072	0.140	6.93	5.30
(M+F) w/	0.849	0.863	0.067	0.097	0.074	0.178	6.94	5.89

Fig. 6. Qualitative comparison of APWAM effectiveness on the VITON-HD and SAMP-VTONS datasets. From left to right: garment images, person images, and the results from four different settings. Red boxes indicate errors, while green boxes highlight the improved results.

TABLE IV
QUANTITATIVE COMPARISON OF SRCM EFFECTIVENESS ON THE
VITON-HD DATASET. THE METRICS USE ↑/↓ TO INDICATE THAT
LARGER/SMALLER VALUES CORRESPOND TO BETTER PERFORMANCE.
THE BEST RESULTS ARE HIGHLIGHTED IN RED.

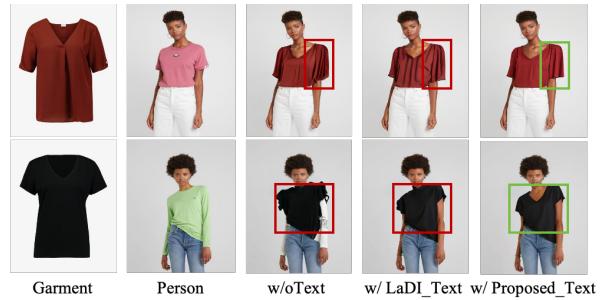| Settings	SSIM↑	LPIPS↓	KID↓	FID↓
w/o Text	0.849	0.066	0.072	6.934
w/ LaDI_Text	0.850	0.061	0.086	6.826
w/ Proposed_Text	0.851	0.060	0.064	6.516

Fig. 7. Qualitative comparison of SRCM effectiveness on the VITON-HD dataset. From left to right: garment images, person images, and results for different settings: w/o Text (baseline), w/ LaDI_Text (raw text), and w/ Proposed_Text (SAS+DMTG). Red boxes indicate errors, while green boxes highlight the improved results.

proposed text representation.

The visual results in Figure 7 demonstrate the effectiveness of the SRCM framework with different text settings. In the baseline (w/o Text), misalignments between the garment and body are evident, particularly around the shoulder and chest, resulting in an unnatural fit. Using raw LaDI_Text (w/ LaDI_Text) shows some improvement, but artifacts remain, especially around the garment edges. In contrast, the proposed text representation (w/ Proposed_Text) significantly improves alignment, especially in the shoulder and chest regions, as highlighted by the green boxes. These visual improvements are consistent with the quantitative results, where our method outperforms others in reducing image distortion and enhancing garment-body consistency, demonstrating the effectiveness of our proposed representation in improving both visual quality and fit accuracy.

Effectiveness of MPAGM. Building upon the APWAM and SRCM, we further validate the effectiveness of MPAGM through an ablation study on the VITON-HD dataset with

Proposed_Text is 0.064, which represents a 25.6% reduction compared to w/ LaDI_Text and an 11.1% reduction compared to w/o Text, indicating that our proposed text representation effectively reduces the generative distribution discrepancy, enhancing the consistency and stability of image generation. In terms of FID, w/ Proposed_Text yields a result of 6.516, which is 4.5% lower than w/ LaDI_Text and 6.0% lower than w/o Text, demonstrating that the SAS+DMTG-based text representation effectively reduces image distortion and improves the alignment between garment and the human body. These results highlight the significant improvement in image quality and consistency provided by the SRCM through the

TABLE V
QUANTITATIVE COMPARISON OF MPAGM EFFECTIVENESS ON THE
VITON-HD DATASET. THE METRICS USE ↑/↓ TO INDICATE THAT
LARGER/SMALLER VALUES CORRESPOND TO BETTER PERFORMANCE.
THE BEST RESULTS ARE HIGHLIGHTED IN RED.

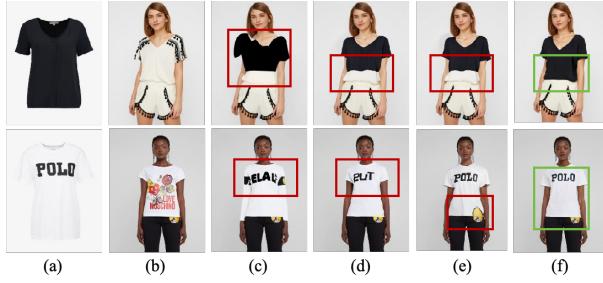| Settings	SSIM↑	LPIPS↓	KID↓	FID↓
only w/ APWAM	0.849	0.066	0.072	6.83
only w/ SRCM	0.847	0.053	0.036	5.72
APWAM + SRCM (LaDI_Text)	0.851	0.050	0.027	5.57
APWAM + SRCM (Proposed_Text)	**0.852**	**0.050**	**0.024**	**5.51**

Fig. 8. Qualitative comparison of MPAGM effectiveness on the VITON-HD dataset. From left to right: garment image, person image, and results of four settings—only w/ APWAM, only w/ SRCM, APWAM + SRCM (LaDI_Text), and APWAM + SRCM (Proposed_Text). Red boxes indicate errors, while green boxes highlight the improved results.

different settings. The settings include: only APWAM, only SRCM, APWAM + SRCM (LaDI_Text), and APWAM + SRCM (Proposed_Text). As shown in Table V, APWAM + SRCM (Proposed_Text) outperforms other configurations, particularly in KID and FID. Specifically, KID is reduced to 0.024, showing a 66.7% decrease compared to only APWAM (0.072) and 33.3% compared to only SRCM (0.036). For FID, it achieves 5.51, a 19.4% reduction from only APWAM (6.83) and 3.7% from only SRCM (5.72). These results demonstrate MPAGM's effectiveness in reducing generative distribution discrepancies and image distortion, leading to better garment-person alignment.

Figure 8 illustrates the impact of different configurations on garment-body alignment and texture fidelity. In the configurations using only APWAM (c) or only SRCM (d), clear misalignments are observed, particularly in the shoulder region, along with noticeable distortions in garment textures such as logos. When APWAM and SRCM are combined (e), the alignment improves, especially around the shoulder and chest areas, resulting in more natural fitting, though artifacts persist near the hemline. With the proposed text representation (f), both alignment and texture quality are further enhanced. The regions highlighted in green exhibit more accurate and natural fitting, confirming the effectiveness of our method in improving overall image quality and structural consistency.

V. DISCUSSION

A. Insights on Consistency

This paper introduces a comprehensive framework for high-fidelity virtual try-on, addressing key challenges in geometric, semantic, and fine-grained consistency. The proposed approach integrates multiple modules to ensure accurate garment fitting and realistic visual rendering across diverse poses. A central insight of this work is the effective management of consistency. The Appearance-Preserving Warp Alignment Module (APWAM) uses a multi-scale architecture to capture both local and global geometric features, ensuring spatial alignment while preserving fine-grained garment details, such as wrinkles and material deformation. This capability is essential for improving visual accuracy, particularly when handling complex fabric deformations.

The Semantic Representation and Comprehension Module (SRCM) enhances semantic consistency by extending traditional garment attribute structures with fine-grained features. It refines garment descriptions by integrating multi-pose datasets, ensuring precise semantic alignment between garments and human body representations across poses. This integration not only improves garment fitting consistency but also facilitates the generation of accurate textual descriptions of garments, further supporting the alignment between visual and semantic representations.

Finally, the Multimodal Prior-Guided Appearance Generation Module (MPAGM) combines geometric priors with semantic and visual cues to optimize the virtual try-on process. This multimodal integration ensures the generation of realistic images that preserve both garment structure and appearance, maintaining consistency across a wide range of garment types and poses.

B. Limitations and Future Directions

Despite significant progress, several limitations remain, primarily related to consistency across varying poses and garment types. One major challenge is the model's performance in handling extreme pose variations and intricate garment textures. While the framework performs well with standard poses and common garment types, it struggles to accurately model complex, non-rigid fabric deformations and rare garment styles. Future work should focus on improving the model's robustness to handle more complex and diverse garment deformations while maintaining consistency in alignment across a broader range of garment types. Another limitation stems from the framework's dependence on large-scale labeled datasets and precise pose estimation. While high-quality pose information and extensive labeled data are crucial for optimal performance, such resources may not be readily available in real-world scenarios. To mitigate this, future research could explore unsupervised or semi-supervised learning techniques to reduce reliance on labeled data and improve the model's generalization ability. Additionally, although the framework demonstrates promising results, computational efficiency remains a critical challenge, particularly for real-time applications. Future efforts should focus on optimizing the framework to achieve faster inference times without compromising consistency or accuracy. Techniques such as model compression, distillation, and hardware acceleration could be investigated to enhance processing speed, enabling real-world deployment.

Looking ahead, several promising directions for future research include extending the framework to other garment categories, such as accessories, footwear, and outerwear, where unique characteristics may require specialized consistency handling. Moreover, integrating additional sensor modalities, such as depth sensing or multi-view cameras, could offer richer data for more accurate garment fitting and pose alignment. Finally, improving real-time performance and reducing computational overhead will be key to the broader adoption of virtual try-on systems in commercial applications.

VI. CONCLUSION

This paper presents a high-fidelity virtual try-on framework that focuses on achieving consistency in geometric alignment, appearance, and semantic understanding of garments.

By integrating the Appearance-Preserving Warp Alignment Module (APWAM), the Semantic Representation and Comprehension Module (SRCM), and the Multimodal Prior-Guided Appearance Generation Module (MPAGM), the framework addresses key challenges in preserving geometric consistency across varying poses, ensuring semantic consistency between garments and body poses, and retaining fine-grained garment details. Extensive experiments demonstrate that the proposed framework significantly outperforms existing methods in terms of visual quality, semantic alignment, and garment fitting accuracy, showcasing its capability to handle complex garment deformations and generate realistic virtual try-on images. However, challenges remain in handling extreme poses and highly intricate garment structures. Future work will focus on improving robustness, enhancing computational efficiency, and extending the framework to a broader range of garment types. Overall, the proposed framework establishes a new standard for virtual try-on technology, offering a solid foundation for future research and real-world applications, particularly in enhancing consistency across diverse poses and garment types.

REFERENCES

[1] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14131–14140.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] T. Wei *et al.*, "E2Style: Improve the efficiency and effectiveness of StyleGAN inversion," *IEEE Trans. Image Process.*, vol. 31, pp. 3267–3280, 2022.

[4] Q. Mao et al., "Scalable face image coding via stylegan prior: Toward compression for human-machine collaborative vision," *IEEE Trans. Image Process.*, vol. 33, pp. 408–422, 2024.

[5] T. Liu et al., "Spatial-aware texture transformer for high-fidelity garment transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 7499-7510, 2021.

[6] Y. Shi, Y. Liu, J. Cheng, Z. J. Wang and X. Chen, "VDMUFusion: A versatile diffusion model-based unsupervised framework for image fusion," *IEEE Trans. Image Process.*, vol. 34, pp. 441-454, 2025.

[7] C. -Y. Chan, W. -C. Siu, Y. -H. Chan and H. Anthony Chan, "AnlightenDiff: Anchoring diffusion probabilistic model on low light image enhancement," *IEEE Trans. Image Process.*, vol. 33, pp. 6324-6339, 2024.

[8] J. Yue, L. Fang, S. Xia, Y. Deng and J. Ma, "Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705-5720, 2023.

[9] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2231–2235.

[10] H.-W. Jin and D.-O. Kang, "Versatile-VTON: A versatile virtual try-on network," in *Proc. IEEE Int. Conf. Consumer Electron.-Asia (ICCE-Asia)*, Busan, South Korea, 2023, pp. 1–4.

[11] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie, "Virtually trying on new clothing with arbitrary poses," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, Nice, France, 2019, pp. 266–274.

[12] X. Han, X. Hu, W. Huang, and M. R. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 10471–10480.

[13] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, "Taming the power of diffusion models for high-quality virtual try-on with appearance flow," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7599–7607.

[14] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8485–8493.

[15] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 764–773.

[16] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1316–1324.

[17] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5802–5810.

[18] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 833–842.

[19] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16515–16525.

[20] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, Jul. 2021, pp. 8821–8831.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents,"2022, *arXiv:2204.06125*.

[22] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Mach. Learn.*, PMLR, Jun. 2016, pp. 1060–1069.

[23] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "LaDI-VTON: Latent diffusion textual-inversion enhanced virtual try-on," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 8580–8589.

[24] R. Wang *et al.*, "StableGarment: Garment-centric generation via stable diffusion," 2024, *arXiv:2403.10783*.

[25] Y. Xu, T. Gu, W. Chen, and A. Chen, "OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 9, Art. no. 9, Apr. 2025.

[26] M. S. Seyfioglu, K. Bouyarmane, S. Kumar, A. Tavanaei, and I. B. Tutar, "Diffuse to choose: Enriching image conditioned inpainting in latent diffusion models for virtual try-all," 2024,*arXiv:2401.13795*.

[27] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, PMLR, Jul. 2021, pp. 8162–8171.

[28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, PMLR, Jun. 2015, pp. 2256–2265.

[29] L. Zhu *et al.*, "TryOnDiffusion: A tale of two unets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 4606–4615.

[30] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022, *arXiv:2208.01618*.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1.8, pp. 9, 2019.

[32] I. Han, S. Yang, T. Kwon, and J. C. Ye, "Highly personalized text embedding for image manipulation by stable diffusion," 2023, *arXiv:2303.08767*.

[33] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.(ICML)*, PMLR, Jul. 2021, pp. 8748–8763.

[34] H. Wang, Z. Zhang, D. Di, S. Zhang, and W. Zuo, "MV-VTON: Multiview virtual try-on with diffusion models," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 7, Art. no. 7, Apr. 2025.

[35] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.

[36] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, "StableVITON: Learning semantic correspondence with latent diffusion model for virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 8176–8185.

[37] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3836–3847.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, 2016, pp. 694–711.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition,"2015,*arXiv:1409.1556*.

[40] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 248–255.

[41] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comput. Vis. (IJCV)*, vol. 106, no. 2, pp. 115–137, Jan. 2014.

[42] Alibaba Cloud Tianchi, 2022, "Tianchi FashionAI Dataset," [Online]. Available: https://tianchi.aliyun.com/dataset/136948.

[43] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7297–7306.

[44] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7291–7299.

[45] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3260–3271, Jun. 2022.

[46] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7543–7552.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[48] M. Wortsman *et al.*, "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7959–7971.

[49] C. Schuhmann *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 25278–25294, Dec. 2022.

[50] K.-N. Nguyen-Ngoc, T.-T. Phan-Nguyen, K.-D. Le, T. V. Nguyen, M.-T. Tran, and T.-N. Le, "DM-VTON: Distilled mobile real-time virtual try-on," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Sydney, Australia, 2023, pp. 695–700.

[51] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022,*arXiv:2207.12598*.

[52] O. Avrahami *et al.*, "SpaText: Spatio-textual representation for controllable image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 18370–18380.

[53] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.

[54] N. Jetchev and U. Bergmann, "The conditional analogy GAN: Swapping fashion articles on people images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2017, pp. 2287–2292.

[55] R. Yu, X. Wang, and X. Xie, "VTNFP: An image-based virtual try-on network with body and clothing feature preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10511–10520.

[56] H. J. Lee, R. Lee, M. Kang, M. Cho, and G. Park, "LA-VITON: A network for looking-attractive virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.

[57] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7850–7859.

[58] Z. Xie et al., "GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23550–23559.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.

[62] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," 2021, *arXiv:1801.01401*.

[63] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014.

[64] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 6840–6851.

[65] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.

Jiajie Li is an undergraduate student at Data Science and Media Intelligence college, Communication University of China, China. Her current research interest is comuputer vision.

Zhe Zhu is currently a Research Engineer at Samsung Research America. He got his Ph.D. from the Department of Computer Science and Technology, Tsinghua University in 2017. He received his bachelor's degree from Wuhan University in 2011. His research interests are in computer vision and computer graphics.

Wang Xinyu is from the School of Artificial Intelligence, Beihang University. My field is the sparsification and quantization of large - model inference. By optimizing model structure and parameter storage, I aim to boost inference efficiency and cut computational costs.I'm keen on quantitative trading algorithms, using AI to build trading strategies and analyze financial data for investment chances. Also, I focus on AI4education, exploring ways to enhance education with AI.

Zhaoxin Fan received his Ph.D. degree from the School of Information, Renmin University, China in 2024. He has also conducted research at Carnegie Mellon University and Hong Kong University of Science and Technology. He is currently an Assistant Researcher in the Institute of Artificial Intelligence, Beihang University. His research interests include multi-modal large language models (LLMs), computer vision, and embodied AI.

Ming Meng is currently a lecturer at the Communications University of China. She received her Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University in 2022. Her main research interests include mixed reality (MR), computer vision, and computer graphics, with a particular focus on scene structure recovery and modeling and driving of virtual digital humans.

Wei Zhao is currently an associate professor at Communication University of China(CUC). She received her PhD degree from the Communication and Information System major at CUC. Her main research interests include intelligent audio video processing, virtual reality technology, Accessible Communication and Interaction Technologies.

Qi Dong is a master student at the School of ata Science and Media Intelligence, Communication University of China, China.Her current research interests are comuputer vision, Digital Image Processing.

Wenjun Wu received the PhD degree in computer-science from Beihang University, in 2001. He was employed with Argonne National Laboratory as a research scientist working on grid computing, cloud computing, media collaboration, etc., until 2012. He is currently employed with Beihang University as a professor. His research interests include crowdsourcing, machine learning, cloud computing, eScience, and cyber infrastructure.