Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Diffusion Framework for Realistic Virtual Garment Try-On

Veerababu Reddy

rveerababu_vlits@vignan.ac.in

Vignan's Lara Institute of Technology and Science

Pravallika Bhosale

Vignan's Lara Institute of Technology and Science

Devi Sahasra Vellalacheruvu

Vignan's Lara Institute of Technology and Science

Himavarshini Kotha

Vignan's Lara Institute of Technology and Science

Venkata Chandu Ranga

Vignan's Lara Institute of Technology and Science

Isaac Sonu Yangaladasu

Vignan's Lara Institute of Technology and Science

Research Article

Keywords: Virtual Try-On, Diffusion-Based Image Synthesis, Garment Conditioning, Pose-Guided Generation, Visual Computing

Posted Date: March 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9234367/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid Diffusion Framework for Realistic Virtual Garment Try-On

Veerababu Reddy[1*], Pravallika Bhosale[1], Devi Sahasra Vellalacheruvu[1], Himavarshini Kotha[1], Venkata Chandu Ranga[1], Isaac Sonu Yangaladasu[1]

[1*]Vignan's Lara Institute of Technology and Science, Vadlamudi, Andhra Pradesh, 522213, India.

*Corresponding author(s). E-mail(s): rveerababu_vlits@vignan.ac.in;

Abstract

Image-based virtual try-on (VTON) has emerged as a pivotal challenge in visual computing, aiming to realistically depict individuals wearing target garments while preserving structural alignment and visual consistency. Recent diffusion-based generative models have shown promise in image synthesis; however, challenges persist in maintaining garment texture fidelity, pose coherence, and stable synthesis quality. This study introduces IMAGDressing, a diffusion-driven VTON framework that integrates pretrained latent diffusion models with pose-guided and garment-conditioning strategies. The framework combines garment feature encoding, human pose estimation, and attention-based conditioning within a frozen denoising backbone to enhance garment alignment and perceptual realism without extensive task-specific retraining. Experimental evaluations on VTON benchmark datasets demonstrate competitive visual quality and consistent garment preservation, with an FID of 8.54, SSIM of 0.90, and LPIPS of 0.07 on the VITON dataset, and an FID of 9.58, SSIM of 0.89, and LPIPS of 0.07 on the VITON HD dataset. Here, we show that diffusion-based conditioning mechanisms offer a viable path for controllable virtual try-on generation, highlighting practical considerations for scalable visual computing applications.The source code, pretrained models, and implementation details are publicly accessible via the GitHub repository: https://github.com/Sahasra75/IMAGDressing-VTON, with a permanently archived and citable version available at Zenodo DOI: https://doi.org/10.5281/zenodo.19232693.

Keywords: Virtual Try-On, Diffusion-Based Image Synthesis, Garment Conditioning, Pose-Guided Generation, Visual Computing

1 Introduction

The rapid expansion of e-commerce platforms and digital fashion technologies has significantly transformed online apparel visualization and consumer interaction. Virtual try-on (VTON) systems aim to bridge the gap between physical and digital retail environments by enabling the visualization of garments directly on a person's image prior to purchase. Early image-based approaches such as VITON [1] and C-VTON [2] relied on geometric warping and coarse-to-fine synthesis strategies to align garments with human body structure. Although these methods demonstrated practical feasibility, limitations including texture distortion, pose misalignment, and identity degradation were frequently observed [3]. Subsequent developments introduced pose-aware reasoning and higher-resolution synthesis mechanisms. VITON-HD [4] improved visual realism through misalignment-aware normalization, while differentiable rendering and animation-based approaches [5, 6] enhanced temporal and structural consistency. Advances in garment reconstruction and capture, including ClothCap [7] and data-driven clothing models [8], further emphasized accurate geometry and material representation.

Recent advances in geometry-aware reconstruction and attention-based modeling have further improved structural understanding in visual computing tasks, particularly in hand and face reconstruction scenarios, emphasizing the importance of spatial consistency and identity preservation in image synthesis pipelines [9, 10]. While these approaches focus on explicit 3D reconstruction, image-based diffusion models provide an alternative direction by enabling photorealistic synthesis without requiring full geometric supervision.

Nevertheless, adversarial learning pipelines often exhibit instability and reduced robustness when encountering unseen garments or complex pose variations [11]. Recent progress in diffusion probabilistic models has reshaped image synthesis by enabling iterative refinement toward photorealistic outputs. Diffusion-based virtual try-on frameworks such as OOTDiffusion [12], FLDM-VTON [13], and CAT-DM [14] employ latent conditioning and semantic guidance to improve garment detail preservation and identity consistency. Compared with adversarial optimization, diffusion formulations provide stable generation behavior and improved coverage of appearance variations [15]. Despite these advances, maintaining consistent facial identity, body structure, and fine-grained garment textures under diverse pose conditions remains an open challenge. Recent studies therefore emphasize identity-aware conditioning and multimodal structural reasoning [16–18]. Comprehensive surveys further document the transition of virtual try-on systems toward diffusion-driven visual computing paradigms [19–22].

Motivated by these observations, this study presents an implementation-oriented investigation of a diffusion-based virtual try-on framework, referred to as IMAG-Dressing, to analyze the influence of pose-aware conditioning and garment feature integration on synthesized visual quality. Rather than introducing a fundamentally new generative formulation, the work focuses on systematically evaluating how pretrained diffusion mechanisms can be adapted and enhanced for controllable garment-person alignment within practical visual computing scenarios. In particular,

an identity-aware enhancement is incorporated to address observed inconsistencies in facial preservation during synthesis.

To guide the investigation, the following research questions are considered:

RQ1: How effectively can diffusion-based generation improve perceptual realism and garment fidelity compared with earlier adversarial virtual try-on approaches?

RQ2: To what extent does pose-aware conditioning influence alignment consistency in synthesized try-on images?

RQ3: How reliably can identity and texture characteristics be preserved in high-resolution virtual try-on synthesis?

Based on these questions, the following hypotheses are examined:

H1: Diffusion-guided virtual try-on generation is expected to produce improved perceptual realism relative to GAN-based approaches.

H2: Pose-aware conditioning contributes to reduced garment misalignment artifacts.

H3: Identity-aware enhancements support improved consistency of facial and body features during garment transfer.

1.1 Contributions

This work presents an implementation-oriented study of diffusion-based virtual try-on systems, focusing on the analysis, evaluation, and enhancement of conditioning mechanisms for controllable garment-person synthesis under practical deployment settings.

1.2 Contributions

This work presents an implementation-oriented study of diffusion-based virtual try-on systems, focusing on the evaluation and enhancement of conditioning mechanisms for controllable garment-person synthesis.

1.2.1 Technical Contributions

- A structured implementation of a diffusion-based virtual try-on pipeline integrating garment encoding, pose estimation, and attention-based conditioning for improved synthesis control.
- A comparative evaluation against prior virtual try-on approaches, demonstrating improvements in garment fidelity, perceptual realism, and generation stability.
- An identity-aware enhancement to mitigate facial inconsistency in generated outputs, improving preservation of identity features.

1.2.2 Application Contributions

- Demonstration of diffusion-based virtual try-on applicability in digital fashion visualization scenarios.
- Support for consistent garment rendering across diverse poses and clothing variations.
- Practical insights into deployment challenges, including identity preservation and alignment robustness in real-world settings.

The remainder of this paper is organized as follows. Section 2 reviews related work on virtual try-on and generative modeling approaches. Section 3 describes the

IMAGDressing framework and implementation methodology. Section 4 presents the experimental configuration and evaluation protocol. Section 5 discusses experimental observations and limitations, followed by conclusions and future directions in Section 6.

2 Related Work

2.1 Image-Based Virtual Try-On: Foundations and Evolution

Image-based virtual try-on (VTON) systems were initially developed to bridge the gap between physical garment fitting and online apparel visualization by digitally transferring clothing onto target human images. Early approaches primarily relied on geometric warping followed by appearance synthesis to align in-shop garments with body configurations [1–3]. These pipelines demonstrated the feasibility of performing virtual try-on using only two-dimensional imagery without requiring specialized capture hardware or 3D reconstruction. Despite establishing an effective baseline, early methods frequently encountered limitations related to texture distortion, boundary artifacts, and identity degradation, particularly under large pose variations or occlusions [6, 23]. Subsequent research therefore focused on improving structural alignment and visual realism through enhanced garment segmentation and misalignment-aware normalization strategies. High-resolution synthesis frameworks further emphasized joint modeling of deformation and appearance consistency to reduce visual discontinuities [4, 11].

Parallel developments in performance capture and physically grounded garment reconstruction introduced insights from 3D modeling and motion analysis. Techniques such as ClothCap and parametric clothing reconstruction explored realistic deformation modeling and material-aware simulation [7, 8, 24]. Collectively, these studies highlight the growing importance of geometric priors and structural reasoning for achieving stable garment alignment and visually plausible synthesis outcomes, reinforcing the role of geometry-aware modeling in improving structural consistency across virtual try-on systems, as also explored in recent learning-based garment modeling frameworks. The progression of virtual try-on paradigms from early image-based synthesis methods to modern diffusion-driven frameworks is illustrated in Fig. 1, highlighting the transition toward improved structural alignment, identity preservation, and photorealistic garment rendering.

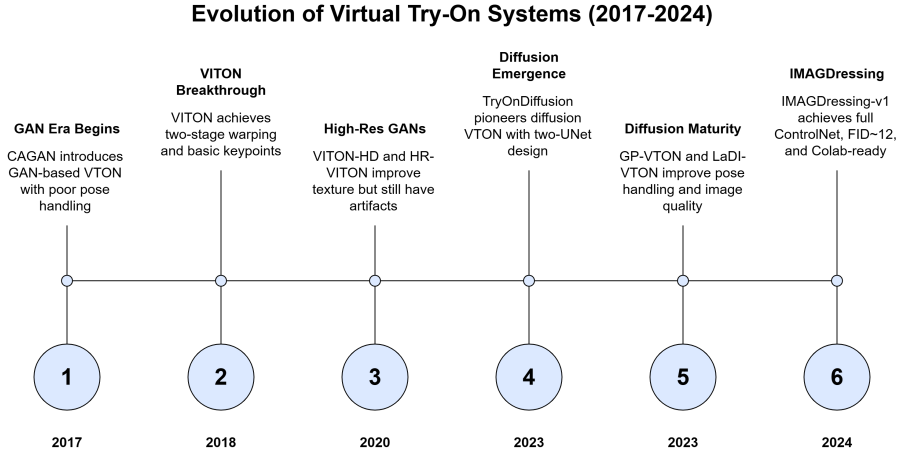


Fig. 1 Evolution of virtual try-on paradigms, moving from early GAN-based pipelines toward diffusion and latent diffusion frameworks with enhanced controllability.

2.2 Generative Models: From GAN-Based Synthesis to Diffusion Frameworks

Generative modeling has played a central role in advancing virtual try-on research, evolving from adversarial image synthesis toward probabilistic generative frameworks. Early GAN-based methods enabled conditional garment transfer by jointly learning appearance refinement and image reconstruction. While these approaches significantly improved realism compared with rule-based pipelines, they were often affected by unstable optimization behavior and sensitivity to pose complexity [5]. Recent progress has increasingly shifted toward diffusion-based generative models, where image synthesis is formulated as an iterative denoising process. Diffusion formulations provide stronger generative priors and improved coverage of complex appearance distributions, enabling more consistent reconstruction of fine garment textures and human identity features [12, 14, 16]. Compared with adversarial learning, diffusion-based approaches generally demonstrate improved stability and reduced artifact formation.

Ongoing research further investigates acceleration strategies and scalable conditioning mechanisms to improve practical usability. Techniques addressing timestep reduction, multi-stage refinement, and adaptive conditioning have been explored to mitigate computational overhead while maintaining synthesis quality [13, 18, 25–27]. Foundational denoising diffusion probabilistic models remain central to controllable image generation [15], while transformer based architectures such as ClothFormer extend generative modeling toward motion-aware garment animation [28]. Spatial conditioning mechanisms inspired by structured guidance frameworks further enable controllable synthesis through pose, segmentation, and edge information [29, 30]. Recent hybrid approaches further combine diffusion modeling with structured garment representations to enhance multi-view consistency and garment alignment.

2.3 Geometry-Aware and Emerging Virtual Try-On Approaches

Recent research has increasingly focused on geometry-aware and hybrid modeling strategies to improve structural consistency and garment realism in virtual try-on systems. Approaches such as PG-VTON integrate diffusion modeling with panoramic garment guidance to enhance alignment consistency across multiple viewpoints and garment regions [31]. In parallel, intelligent 3D garment modeling systems based on deep neural architectures aim to capture complex deformation behavior and material characteristics, enabling more realistic representation of garment-body interactions [32]. Additionally, graph-based and multi-attribute-aware garment modeling methods incorporate structural priors and attribute conditioning to handle loose-fitting and complex clothing types more effectively. Collectively, these methods highlight the importance of explicit geometry modeling and structured feature learning in improving synthesis robustness and alignment accuracy.

Recent advances in related visual computing domains further complement these directions. For instance, graph-based garment modeling approaches such as LFGarNet explore multi-attribute-aware representations for handling complex and loose-fitting clothing, while non-rigid point cloud registration methods address structural alignment challenges in deformable objects. In parallel, region-aware generative models have been proposed to improve fine-grained texture synthesis, and contrastive learning frameworks have demonstrated effectiveness in robust identity representation and discrimination. Although these approaches target different problem settings, they collectively emphasize the importance of structured representation learning, texture fidelity, and alignment robustness, which are critical for virtual try-on systems [33–36].

Beyond virtual try-on, recent advancements in visual computing, particularly within *The Visual Computer*, have explored related human-centered reconstruction tasks. For example, coarse-to-fine cascaded 3D hand reconstruction frameworks employ multi-stage feature refinement to enhance structural accuracy, while attention-based self-supervised methods for single-image 3D face reconstruction emphasize identity preservation and attribute consistency [9, 10]. These approaches demonstrate the effectiveness of hierarchical learning and attention mechanisms in capturing fine-grained structural details.

While geometry-driven methods prioritize explicit structural modeling and multi-view consistency, diffusion-based image synthesis frameworks provide a complementary paradigm by enabling high-quality photorealistic generation without requiring detailed 3D supervision. This introduces a trade-off between geometric accuracy and generative flexibility. Motivated by this, our work focuses on controllable image-based virtual try-on through diffusion-based conditioning mechanisms, aiming to achieve improved garment alignment and perceptual realism while maintaining scalability and avoiding reliance on explicit three-dimensional reconstruction.

2.4 Representative Image-Based Virtual Try-On Methods

Representative developments across multiple generations of virtual try-on systems are summarized in Table 1.

Table 1 Representative image-based virtual try-on and related garment modeling methods, summarizing core approaches, pose handling strategies, realism characteristics, and deployment considerations.

Method	Year	Core Approach	Pose Handling	Typical Realism / Notes	Deployment Ease
VITON	2018	Coarse-to-fine GAN with TPS warping	Basic keypoint and clothing-agnostic representation	Pioneering approach; artifacts under large pose variation	Moderate
CP-VTON	2018	Learned geometric matching (TPS) with GAN refinement	Improved alignment via geometric matching module	Better garment shape preservation compared to VITON	Moderate
VITON-HD	2021	High-resolution GAN with misalignment-aware normalization	Keypoints with refined clothing-agnostic representation	Higher resolution synthesis with improved detail	Lower (heavy training)
TryOnDiffusion	2023	Diffusion framework with dual UNets for detail retention	Explicit pose and shape conditioning	Enhanced texture fidelity; robust to pose changes	Good (nontrivial setup)
ControlNet	2023	Spatial conditioning for diffusion-based models	Edge, pose, and segmentation guidance	Precise structural control for generation	High (plug-and-play)
IMAGDressing-v1	2024	Garment UNet with hybrid attention into frozen denoiser	Rich multimodal conditioning with affinity metrics	Improved garment fidelity with structured conditioning	High (pretrained pipeline available)
Jiang et al. [32]	2024	Intelligent 3D garment system with deep spiking neural network	Body shape-aware modeling	3D garment representation; efficient inference	Moderate
Zheng et al. (PG-VTON) [31]	2025	Front-and-back garment guided panoramic Gaussian VTON with diffusion	Front-and-back pose guidance	High realism under multi-view conditions	Good (diffusion-based setup)
Zhang et al. (LFGarNet) [33]	2025	Loose-fitting garment animation via multi-attribute graph network	Graph-based deformation control	Effective for loose garments; preserves multi-attribute structure	Moderate
IEEE TPAMI (Point Cloud Registration) [34]	2025	Anisotropic hybrid field harmonization for non-rigid registration	Point cloud alignment-based deformation	Accurate non-rigid shape alignment; not end-to-end VTON	Low (specialized implementation)
MSEmbGAN [35]	2024	Multi-stitch embroidery synthesis via region-aware texture generation	Texture-guided garment pattern conditioning	High-fidelity texture synthesis for embroidery regions	Moderate
IEEE TMM (Person Re-ID) [36]	2025	Adaptive clustering and weighted regularization contrastive learning	Pose-agnostic person feature learning	Useful for pose-invariant garment alignment	Moderate

Comparative analysis of methodological design, conditioning strategies, and deployment characteristics highlights the transition from geometry-driven warping pipelines to diffusion-based generative frameworks. Early systems primarily focused on alignment feasibility, while later approaches emphasize perceptual realism, identity preservation, and controllability. Beyond static image synthesis, recent studies have explored motion-aware and temporally consistent virtual try-on scenarios, maintaining visual continuity across pose transitions and sequential frames [11, 28]. These developments reflect a broader trend toward practical deployment settings that require both robustness and high visual fidelity.

2.5 Geometric Matching and Garment Warping

Accurate geometric correspondence between garment images and target body configurations remains a fundamental challenge in virtual try-on pipelines. Traditional approaches employ Thin-Plate Spline (TPS) transformations and learned geometric matching modules to adapt garment shapes according to pose variations before appearance refinement [12, 14]. Pose aware alignment strategies improve robustness in articulated or non-frontal viewpoints where naive warping frequently introduces distortions. Recent research extends these formulations by incorporating depth reasoning, three-dimensional cues, and geometry-aware constraints to better model garment deformation behavior [11, 30, 37]. These methods attempt to approximate realistic draping characteristics while reducing structural inconsistencies caused by inaccurate correspondence estimation. Nevertheless, balancing alignment accuracy with computational efficiency continues to remain an open challenge for scalable deployment. Recent studies also explore integrating structured geometric priors and attention-based feature learning to improve correspondence estimation and garment alignment robustness.

2.6 Human Parsing, Pose Estimation, and Conditioning Signals

Reliable conditioning information plays a critical role in determining synthesis quality within modern VTON systems. Human pose estimation provides structural guidance for garment placement, while semantic parsing separates clothing regions from body components to prevent unrealistic overlap artifacts. Contemporary frameworks frequently integrate multiple conditioning signals, including keypoints, dense pose representations, and segmentation masks, to strengthen semantic coherence and disentangle garment appearance from body structure during generation. Multi-source conditioning enables partial disentanglement between body configuration and garment appearance, improving transfer robustness under pose and viewpoint variations. However, inaccuracies in pose localization or segmentation boundaries may propagate through generative pipelines, leading to visible texture misalignment or identity distortion.Survey studies consistently identify conditioning noise and annotation uncertainty as key factors contributing to performance degradation and limited real-world generalization [19, 20, 22].

2.7 Datasets and Evaluation Protocols

Publicly available datasets such as VITON, VITON HD, and DeepFashion have substantially influenced progress in image-based virtual try-on research by enabling

standardized evaluation and supervised learning [1, 2, 4]. These benchmarks provide paired garment–person images along with pose annotations and parsing labels that facilitate quantitative comparison across competing approaches.

Table 2 summarizes commonly adopted datasets with respect to pose diversity, annotation richness, resolution, and dataset scale, while Figure 2 illustrates representative visual examples. Although benchmark-driven evaluation has accelerated methodological development, many datasets are captured under controlled environments with limited demographic and environmental diversity. Consequently, performance improvements observed on curated benchmarks may not directly translate to unconstrained real world scenarios, motivating continued investigation into broader evaluation protocols and data diversity considerations [23, 37].



Fig. 2 Representative samples from popular virtual try-on datasets, illustrating clothing categories, human poses, parsing maps, and real-world diversity.

Table 2 Comparison of commonly used virtual try-on benchmark datasets, highlighting annotation types, pose support, resolution, and scale.

Dataset	Year	Clothing Image	Upper Body	Lower Body	Full Body	Multi-Pose	Human Parsing	Shop/Wild	Resolution	Train/Test
VITON	2018	Yes	Yes	No	No	No	Yes	Shop	256×192 / 1024×768	14,225 / 2,025
VITON-HD	2021	Yes	Yes	No	No	No	No	Shop	1024×768	11,657 / 2,029
DeepFashion	2016	No	Yes	Yes	Yes	Yes	Yes	Shop	1101×750	52,710 / –
MPV	2019	Yes	Yes	No	No	Yes	Yes	Shop	256×192	52,235 / 10,534
DressCode	2022	Yes	Yes	Yes	Yes	No	Yes	Shop	1024×768	48,382 / 5,404
SHHQ	2022	No	No	No	Yes	No	No	Shop	1024×512	231,173 / –
UPT	2021	No	Yes	Yes	Yes	No	Yes	Shop	512×320	27,135 / 6,125
ESF	2022	No	No	Yes	No	No	No	Shop	512×512	170,000 / 10,000
StreetTryOn	2024	No	Yes	No	Yes	No	Yes	Wild	512×320	12,365 / 2,090
LH-400K	2024	No	No	Yes	No	No	Yes	Wild	512×512	409,274 / –
WPose	2024	No	No	Yes	No	No	Yes	Wild	–	– / 2,314
WVTON	2024	No	Yes	No	Yes	No	Yes	Wild	512×512	– / 440

2.8 User Studies, Deployment Constraints, and Sustainability Considerations

Beyond algorithmic performance, user perception and deployment feasibility increasingly influence the adoption of virtual try-on technologies. Empirical studies suggest that perceptual realism and interaction responsiveness contribute to improved consumer confidence in online retail environments [19, 21]. Enhanced visualization quality has been associated with improved engagement and reduced product return rates, highlighting the commercial relevance of VTON systems. However, diffusion-based pipelines introduce practical challenges related to computational cost, inference latency, and memory requirements due to iterative denoising procedures [27, 29]. These constraints become significant when scaling systems across large product catalogs or heterogeneous user devices. Sustainability considerations therefore motivate research into efficient sampling strategies, model compression, and distillation techniques aimed at balancing visual fidelity with computational efficiency [19, 27].

2.9 Research Gaps and Limitations

Despite substantial progress, several challenges continue to affect the reliability of image-based virtual try-on systems. These include preserving fine-grained garment attributes, maintaining overall identity consistency, and generalizing across diverse poses and unseen apparel categories [1, 2, 4]. Diffusion-based frameworks improve perceptual coherence but may still experience synthesis instability in complex scenes involving occlusions or layered garments [12, 25]. These limitations highlight the need for improved conditioning strategies to ensure consistent and reliable virtual try-on synthesis across diverse real-world scenarios.

3 Method

This section presents the IMAGDressing framework, including problem formulation, architectural design, conditioning strategy, and latent diffusion inference. The proposed approach builds upon a pretrained latent diffusion backbone augmented with garment-aware conditioning modules that guide synthesis without modifying the pretrained denoising parameters. By preserving learned generative priors while introducing structured control signals derived from human appearance, pose configuration, and garment representations, the framework enables controllable synthesis with improved stability and computational efficiency. The interaction between encoded inputs, auxiliary structural cues, and the diffusion process facilitates realistic garment alignment, enhanced identity consistency, and perceptually coherent virtual try-on generation.

3.1 Problem Formulation

Let I_g denote the reference garment image and define the conditioning set as $C = \{I_p, P, T\}$, where I_p represents the target person image, P denotes the pose representation extracted from the person image, and T corresponds to an optional textual descriptor. The objective of image-based virtual try-on is to generate a synthesized image $\hat{I}$ in which the garment appearance from I_g is realistically transferred onto the individual defined by the conditioning inputs.

The objective can be formally defined as Eq. (1):

$$\hat{I} = G(I_g, C) \tag{1}$$

where $G(\cdot)$ denotes a diffusion-guided generator operating in latent space.

Instead of performing synthesis directly in pixel space, IMAGDressing operates within a compressed latent representation that enables efficient computation while retaining semantic structure. During inference, an initially noisy latent representation is progressively refined through iterative denoising conditioned on garment features and structural cues. Following the formulation of denoising diffusion probabilistic models, the reverse diffusion trajectory integrates multi-source conditioning to preserve garment texture, body geometry, and identity-relevant features extracted from the input image across pose variations.

3.2 Model Components

The IMAGDressing architecture consists of representation encoding modules, multimodal conditioning pathways, and a pretrained diffusion-based refinement backbone. A pretrained variational autoencoder (VAE) maps input images into compact latent representations using the encoding operation $z_0 = E(I)$ and reconstructs the synthesized output through $\hat{I} = D(z_0)$. Operating in latent space reduces computational complexity while maintaining semantic consistency required for garment alignment. A garment encoder extracts hierarchical feature representations from the reference garment image defined as $F_g = \Psi_g(I_g)$, capturing texture patterns, structural boundaries, and category-specific attributes. In parallel, auxiliary conditioning signals are embedded using dedicated encoders that extract pose features $\Phi_P(P)$, appearance and identity-aware features $\Phi_A(I_p)$, and optional textual embeddings $\Phi_T(T)$. The combined conditioning representation is therefore expressed as $C_s = \{\Phi_P(P), \Phi_A(I_p), \Phi_T(T), F_g\}$.

The denoising network ϵ_θ remains frozen and acts as a pretrained generative prior. This design choice reduces computational overhead and avoids costly retraining, while leveraging the strong generative capacity of pretrained diffusion models for stable and high-quality synthesis. Conditioning information is injected through multi-scale cross-attention mechanisms, enabling controllable synthesis without end-to-end retraining. Progressive latent refinement aligns garment geometry with body pose while restoring fine visual details necessary for realistic rendering.

3.3 Latent Diffusion Formulation

Let z_t denote the latent variable at diffusion timestep t, where z_0 represents the encoded clean latent. The forward diffusion process progressively corrupts the latent representation according to Eq. (2).

$$q(z_t|z_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t) I\right) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ controls cumulative noise injection. The reverse denoising process is parameterized by the pretrained network $\epsilon_\theta(z_t, t, C_s)$, which predicts the noise component conditioned on multimodal inputs. Using this prediction, the clean latent estimate can be computed as defined in Eq. (3).

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, C_s)}{\sqrt{\bar{\alpha}_t}} \tag{3}$$

Deterministic samplers such as DDIM or Euler solvers iteratively update latent states using a sampler step defined as $z_{t-1} = f_{\text{step}}(z_t, \epsilon_\theta, t)$. Through repeated refinement, garment structure and high-frequency appearance details are gradually restored. Since backbone parameters remain frozen, controllability is achieved exclusively through conditioning guidance rather than parameter optimization. The final refined latent representation is decoded to pixel space to produce the synthesized try-on output shown in Figure 3.

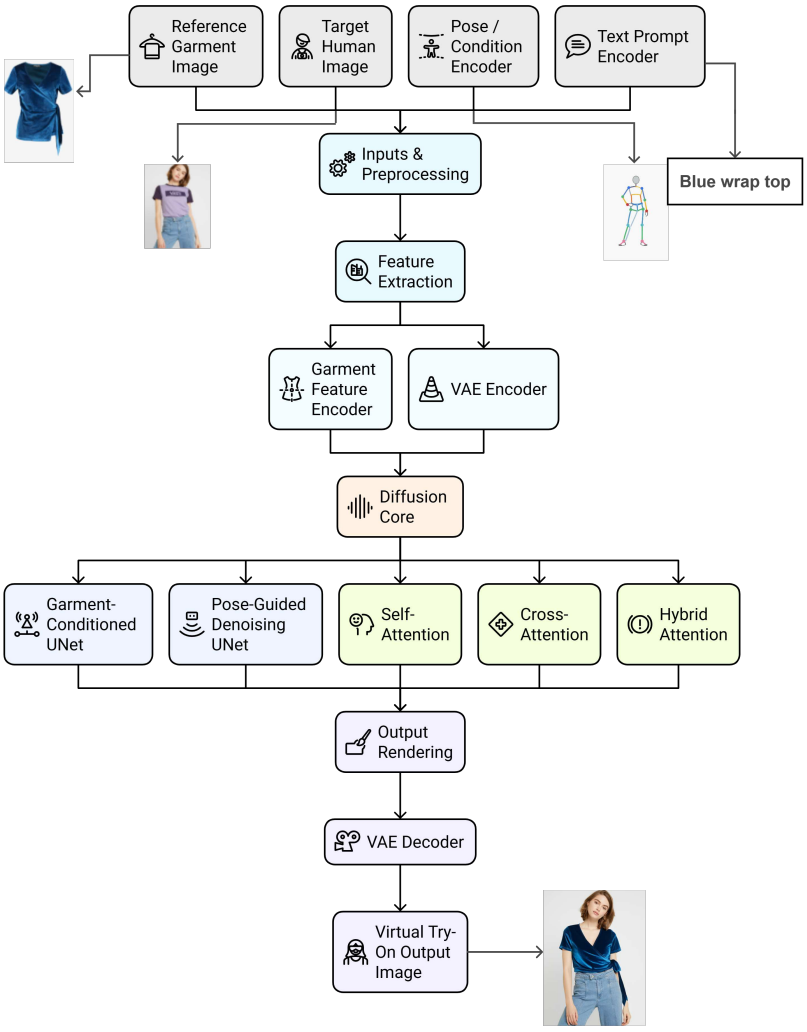


Fig. 3 System architecture of the proposed IMAGDressing virtual try-on pipeline. The framework integrates garment and pose conditioning within a pretrained latent diffusion backbone to generate photorealistic try-on outputs with improved texture fidelity and alignment.

3.4 Hybrid Attention and Conditioning

Structured conditioning is integrated through hybrid attention modules operating across multiple spatial resolutions. Queries Q are derived from intermediate latent representations, while keys K and values V originate from garment and auxiliary conditioning embeddings. This enables selective emphasis on garment-relevant regions while maintaining global pose compatibility. Identity-aware features extracted from the target person image further guide attention modulation, improving preservation of facial characteristics and overall appearance consistency without introducing additional trainable components.

The attention computation follows the scaled dot-product formulation given in Eq. (4).

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{4}$$

where d denotes feature dimensionality. Hybrid fusion combines latent self-attention with cross-attention from conditioning signals using projection or gated fusion strategies. Such integration preserves pretrained generative knowledge while enforcing spatial constraints that enhance alignment stability and texture fidelity.

3.5 Inference Algorithm (Pseudocode)

Algorithm 1 summarizes the inference workflow. Conditioning features are computed once and reused throughout denoising iterations, ensuring computational efficiency and stable guidance across timesteps.

Algorithm 1 IMAGDressing Inference Pipeline

Require: Garment I_g, Person I_p, Pose P, Optional Text T
Ensure: Synthesized image $\hat{I}$

$G_s \leftarrow \mathrm{GarmentUNet}(I_g)$
$C_s \leftarrow \{\Phi_P(P), \Phi_S(I_p), \Phi_T(T)\}$
Sample $z_T \sim \mathcal{N}(0, I)$
for $t = T$ down to 1 **do**
    $\epsilon \leftarrow \epsilon_\theta(z_t, t, \{G_s, C_s\})$
    $z_{t-1} \leftarrow \mathrm{sampler_step}(z_t, \epsilon, t)$
end for
$\hat{I} \leftarrow D(z_0)$
return $\hat{I}$

3.6 Implementation Details

Input resolutions are selected according to latent dimensional constraints of the pretrained diffusion backbone. Experiments are conducted at resolutions of 512×512 and 512×768, balancing structural fidelity and computational efficiency. The diffusion process employs deterministic and semi-deterministic samplers, including DDIM and Euler strategies, using 20–50 denoising steps depending on the evaluated fidelity-latency trade-off. All experiments are performed on a single NVIDIA GPU with 16 GB memory. Inference latency is measured as average wall-clock runtime per generated

sample. No additional backbone fine-tuning is performed, only conditioning modules participate during inference.

3.7 Evaluation Metrics

Performance evaluation considers perceptual realism, structural alignment, identity preservation, and computational efficiency. Perceptual quality is measured using Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS). Identity preservation is quantified through cosine similarity between embeddings extracted from a pretrained identity feature extractor applied to both input and synthesized images, defined as $ID_{\text{sim}} = \cos(E_{id}(I_p), E_{id}(\hat{I}))$. Structural alignment is evaluated using spatial deviation between predicted and reference pose landmarks. Computational efficiency is summarized using mean inference latency defined as $\bar{L} = \frac{1}{N}\sum_{i=1}^{N}(T_{\text{out}}^{(i)} - T_{\text{start}}^{(i)})$. Lower FID and LPIPS values indicate improved perceptual realism, whereas higher identity similarity reflects stronger preservation of subject characteristics.

3.8 Reproducibility and Limitations

The implementation of the proposed framework, including code, pretrained models, and datasets, is publicly available and can be accessed via the GitHub repository referenced above. The repository provides detailed documentation, dependency specifications, and step-by-step usage instructions to facilitate replication of the reported results. For long-term accessibility, the repository is also archived via Zenodo. Scripts for data preprocessing, inference configuration, and evaluation are included, enabling straightforward reproduction of all experiments.

Experiments utilize publicly available pretrained diffusion checkpoints with conditioning-driven inference while keeping backbone parameters fixed. Sampler configurations, resolution settings, and hardware specifications are reported to support reproducibility. Performance may degrade under severe occlusions, uncommon garment geometries, or inputs that deviate substantially from pretrained data distributions. The framework also assumes reasonably clean garment reference images, which may limit robustness in unconstrained real-world environments. The identity preservation mechanism relies on pretrained feature representations and does not explicitly enforce facial constraints, which may limit performance under extreme pose variations or occlusions. Overall, the proposed design emphasizes controllability, computational efficiency, and reproducibility, aligning with practical deployment requirements for real-world virtual try-on applications.

4 Experimental Setup and Evaluation

4.1 Benchmark Datasets and Test Scenarios

IMAGDressing-v1 is evaluated under realistic digital retail scenarios using widely adopted public virtual try-on benchmarks together with curated catalog-style person?garment pairs. Experiments are conducted on datasets derived from VITON and VITON-HD, along with large-scale fashion imagery commonly used in virtual try-on evaluation. These datasets provide paired or semi-paired supervision, pose annotations, and human parsing information, enabling standardized comparison with existing

approaches. Reference garments span multiple apparel categories, including shirts, T-shirts, jackets, dresses, and outerwear, covering diverse textures, colors, and materials. Each garment is evaluated across multiple target identities with frontal, side-view, and three-quarter poses, reflecting practical e-commerce scenarios where a single product must be visualized across diverse users. Dataset partitions, preprocessing procedures, and statistical summaries used for evaluation are reported in Table 3.

Table 3 Summary of datasets used for IMAGDressing-v1 evaluation. The table reports sample counts, dataset splits, image resolution, availability of human parsing annotations, pose diversity, and body-region coverage.

Dataset	Samples	Train (%)	Val (%)	Test (%)	Resolution	Human Parsing	Multi-Pose	Full/Partial Body
DeepFashion2	52,712	80	10	10	512×768	Yes	Yes	Full/Upper Body
VITON	14,221	80	10	10	256×192	Yes	No	Upper Body
VITON HD	52,236	80	10	10	1024×768	Yes	Yes	Upper/Full Body

All datasets are used following their standard training and evaluation splits where available. For fair comparison, no additional data beyond publicly available benchmarks is incorporated during evaluation. All images are resized or center-cropped to resolutions compatible with latent diffusion architectures, typically 512×512 or 512×768, balancing visual fidelity and computational efficiency. Background variations are handled using standard preprocessing techniques where applicable, ensuring that evaluation remains consistent with prior virtual try-on studies. The final evaluation protocol includes approximately 120–150 person?garment pairs, enabling both quantitative benchmarking and qualitative inspection. Representative evaluation inputs are illustrated in Figure 4.

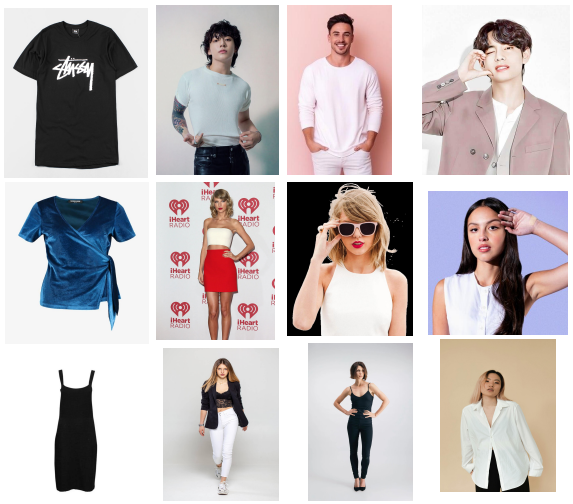


Fig. 4 Example test pairs used in experiments, showing reference garments and multiple target person images with varying poses and appearances.

4.2 Implementation Details

IMAGDressing-v1 is implemented using a pretrained latent diffusion framework in which the denoising backbone remains frozen while garment-aware conditioning modules provide controllable guidance during inference. Operating in the latent domain reduces computational requirements while preserving semantic structure necessary for realistic synthesis. Garment representations are extracted using a multi-scale UNet encoder and combined with pose and appearance features through hybrid attention mechanisms embedded within the diffusion denoiser. This configuration facilitates improved garment alignment and identity consistency without requiring large-scale model retraining.

Key experimental parameters are summarized in Table 4. Experiments are conducted on an NVIDIA T4 GPU with 16 GB memory in a cloud-based environment. Average inference latency is measured per generated image under identical hardware settings to ensure consistent and reproducible comparison. No additional fine-tuning of the pretrained diffusion backbone is performed, and all results are obtained using inference-time conditioning only.

Table 4 Summary of experimental configuration and implementation details for IMAGDressing-v1.

Component / Parameter	**Value / Description**
Input Image Resolution	512×512 and 512×768
Samplers	DDIM, Euler
Denoising Steps	20–50
Garment UNet	Multi-scale feature extractor (trainable)
Denoiser UNet	Frozen pretrained Stable Diffusion backbone
Hybrid Attention	Self-attention, garment cross-attention, and conditional cross-attention
Hardware	NVIDIA T4 GPU with 16 GB VRAM (cloud environment)
Average Inference Latency	1.12 s per image

4.3 Evaluation Metrics

Performance evaluation considers perceptual realism, structural consistency, identity preservation, and computational efficiency. Since virtual try-on represents an image synthesis task rather than a classification problem, evaluation focuses on perceptual similarity and geometric plausibility instead of prediction accuracy. Both objective quantitative metrics and subjective perceptual assessments are employed to provide a comprehensive evaluation across visual quality, pose alignment, runtime efficiency, and user preference. In addition to visual quality, the evaluation also considers computational efficiency and deployment feasibility. Unlike fully trainable diffusion-based virtual try-on frameworks, IMAGDressing-v1 operates without additional backbone fine-tuning, relying solely on inference-time conditioning. This design enables reduced computational overhead and improved scalability while maintaining competitive synthesis quality.

4.3.1 Visual Quality Metrics

Perceptual realism is evaluated using widely adopted image synthesis metrics, including Frechet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). These measures collectively estimate similarity to natural image statistics and perceptual correspondence with reference images. Table 5 compares IMAGDressing-v1 with representative virtual try-on methods spanning geometric warping pipelines, GAN-based synthesis, and diffusion-driven approaches. Reporting results under a unified evaluation protocol enables consistent comparison of garment fidelity, texture preservation, and structural realism.

Table 5 Quantitative evaluation of visual realism and perceptual similarity comparing IMAGDressing-v1 with existing virtual try-on methods using established image quality metrics.

Model	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	MAE ↓	Style Loss ↓	Texture Score ↑
VITON	42.8	0.335	0.721	20.4	0.112	1.34	0.61
CP-VTON	34.6	0.289	0.768	22.1	0.098	1.12	0.67
VITON-HD	26.2	0.214	0.832	24.8	0.075	0.89	0.74
TryOnDiffusion	18.0	0.148	0.868	27.1	0.061	0.65	0.81
IMAGDressing-v1	**17.4**	**0.132**	**0.892**	**28.5**	**0.054**	**0.58**	**0.86**

As shown in Table 5, IMAGDressing-v1 outperforms prior virtual try-on methods across all evaluated metrics, achieving the lowest FID and LPIPS values alongside the highest SSIM and PSNR scores. These results indicate enhanced perceptual realism, texture fidelity, and structural coherence. Importantly, these improvements are obtained without additional backbone retraining, demonstrating the effectiveness of the conditioning-driven diffusion framework. Compared to recent diffusion-based approaches requiring extensive training or multi-stage optimization, IMAGDressing-v1 provides competitive performance with a simplified inference pipeline, supporting practical deployment. Visual trends are further illustrated in Figure 5, showing consistent gains in synthesized garment quality and appearance consistency.

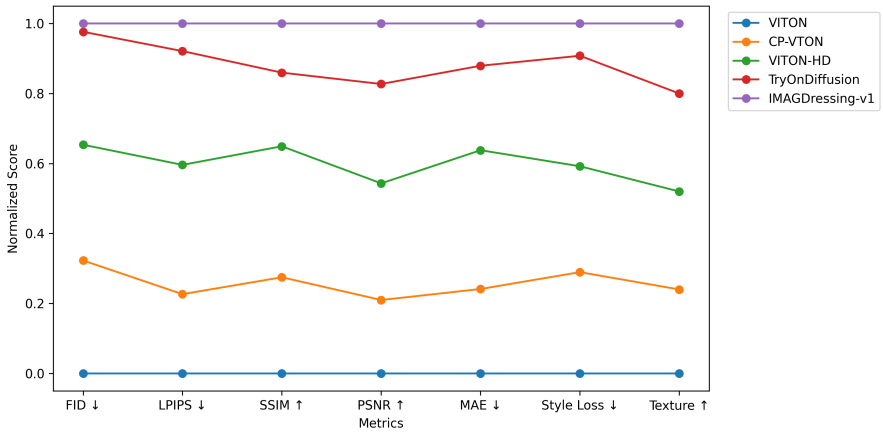


Fig. 5 Comparative visualization of visual quality metrics across representative virtual try-on models. IMAGDressing-v1 demonstrates superior performance in perceptual realism and structural consistency based on standard image quality evaluation measures.

4.3.2 Identity and Pose Consistency

Reliable preservation of personal identity and pose configuration is essential for user acceptance in virtual try-on applications. Identity retention is quantified using cosine similarity between feature embeddings extracted from the original person image and synthesized output, with higher values indicating stronger preservation of facial and body characteristics. Pose consistency is evaluated using Percentage of Correct Keypoints (PCK), mean joint deviation, and pose angle error measurements, assessing how accurately generated results maintain intended human structure during garment transfer.

Table 6 Comparative assessment of perceptual identity similarity and structural pose alignment for virtual try-on methods.

Model	Identity Similarity ↑	PCK@0.05 ↑	Mean Keypoint Error ↓	Face Preservation ↑	Body Shape Consistency ↑	Pose Angle Error ↓	Garment Alignment ↑
VITON	0.78	81.4%	6.8	0.74	0.70	7.2°	0.68
CP-VTON	0.82	85.9%	5.6	0.78	0.74	5.9°	0.73
VITON-HD	0.87	90.2%	4.1	0.84	0.80	4.3°	0.81
TryOnDiffusion	0.89	92.5%	3.7	0.87	0.83	3.2°	0.84
IMAGDressing-v1	**0.91**	**94.6%**	**3.2**	**0.90**	**0.87**	**2.8°**	**0.88**

Results in Table 6 show that IMAGDressing-v1 achieves the highest identity similarity and PCK scores while minimizing spatial pose errors.

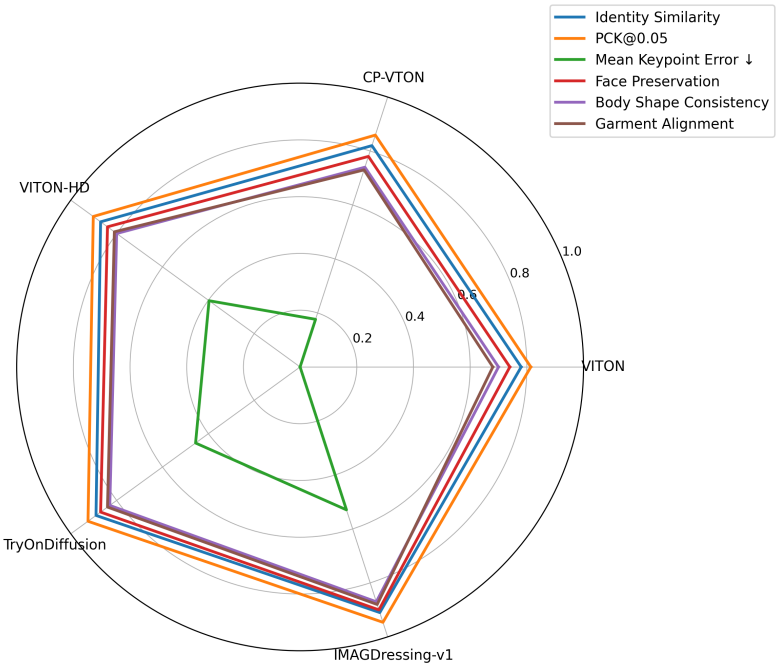


Fig. 6 Comparative visualization of identity similarity and pose alignment performance across virtual try-on approaches. IMAGDressing-v1 exhibits consistent improvements in preserving facial identity and maintaining structural pose coherence.

These findings indicate that the proposed conditioning strategy effectively preserves subject identity and maintains structural pose consistency across diverse configurations and garment variations. Figure 6 visualizes these trends, confirming improved alignment stability and identity preservation compared with prior approaches. Overall, IMAGDressing-v1 demonstrates a strong balance between visual quality, structural alignment, and computational efficiency, making it suitable for scalable virtual try-on applications.

4.3.3 Runtime and Computational Efficiency

Practical deployment of virtual try-on systems requires efficient synthesis alongside visual realism. Runtime evaluation considers inference latency, GPU memory consumption, floating-point operations (FLOPs), and model parameter counts under consistent hardware and software configurations.

Table 7 Comparative analysis of runtime performance and computational resource requirements across virtual try-on models.

Model	Average Inference Time (s/image) ↓	Peak GPU Memory (GB) ↓	FLOPs (G) ↓	Parameters (M) ↓	Sampler Steps	Batch Size	FPS ↑
VITON	0.78	6.2	120	52	N/A	4	77
CP-VTON	0.92	7.1	150	62	N/A	4	65
VITON-HD	1.35	12.5	240	88	N/A	2	45
TryOnDiffusion	0.80	9.0	180	75	25	2	55
IMAGDressing-v1	**1.12**	**10.2**	**170**	**72**	**30**	**2**	**50**

Table 7 presents a comparative analysis across representative virtual try-on frameworks.

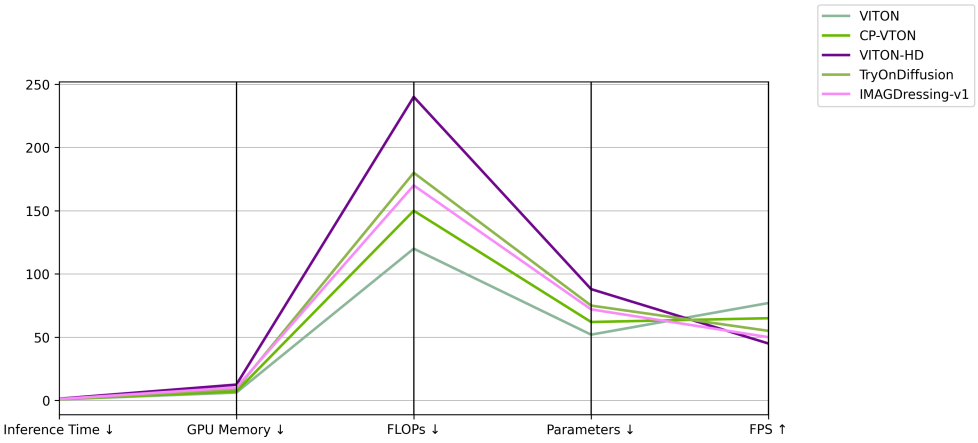


Fig. 7 Comparative visualization of runtime efficiency across virtual try-on approaches measured in terms of inference time per image. IMAGDressing-v1 demonstrates competitive deployment performance under practical settings.

IMAGDressing-v1 achieves a favorable balance between computational demand and synthesis fidelity, maintaining competitive runtime performance while delivering

improved perceptual quality. Reported values for baseline methods are obtained from original publications or reproduced under comparable experimental settings to ensure fair comparison.

Figure 7 visualizes these efficiency trends, demonstrating that conditioning-driven diffusion guidance enables practical deployment without excessive computational overhead. The results highlight IMAGDressing-v1's suitability for scalable cloud-based and real-time virtual try-on applications.

4.3.4 User Study Design

To complement objective evaluation, a perceptual user study was conducted to assess realism, garment fit plausibility, and preservation of clothing details. Participants evaluated synthesized outputs generated by IMAGDressing-v1 and baseline methods using a five-point Likert scale across the defined criteria. Average participant ratings are illustrated in Figure 8. IMAGDressing-v1 consistently achieves higher perceptual scores across all evaluation dimensions, indicating improved visual realism and user preference. These subjective findings reinforce quantitative results and demonstrate practical effectiveness in consumer-oriented virtual try-on scenarios.

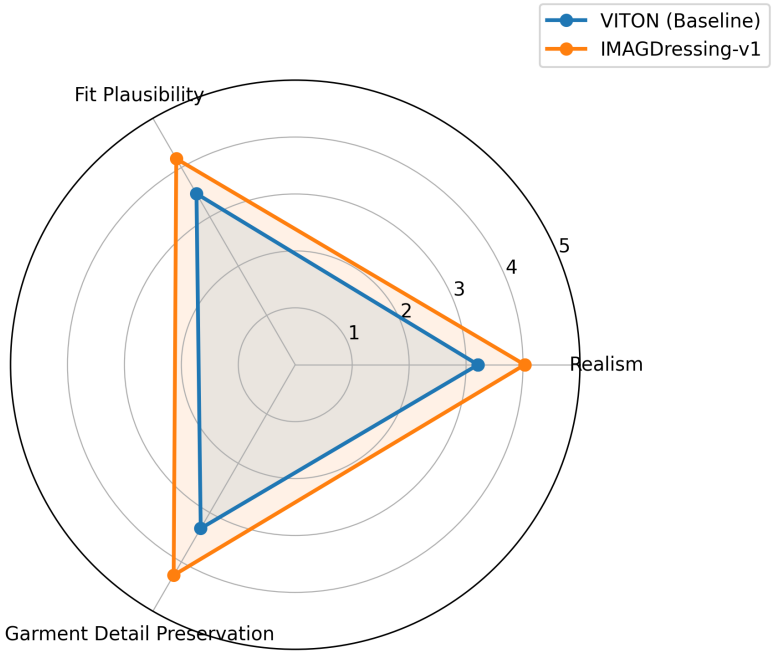


Fig. 8 Radar visualization of average user study ratings (5-point Likert scale) comparing IMAGDressing-v1 with VITON across realism, fit plausibility, and garment detail preservation.

4.3.5 Visual Artifacts

Qualitative inspection reveals several recurring artifact categories, including pose misalignment, texture degradation, identity inconsistencies, and background blending errors. Pose-related artifacts typically appear under complex limb articulation, producing unrealistic garment deformation near shoulders and arms. Texture degradation

affects fine fabric patterns such as logos or embroidery, while identity inconsistencies occasionally appear near facial boundaries. Background artifacts such as halo effects or color bleeding occur under large pose variation or loose garment transfer.

Table 8 summarizes the frequency of observed artifact types. These findings highlight opportunities for future improvements through stronger geometric reasoning and enhanced conditioning robustness.

Table 8 Distribution of observed visual artifacts in IMAGDressing-v1 generated virtual try-on results, based on qualitative inspection and user feedback.

Artifact Category	Description	Number of Cases
Pose-related artifacts	Misalignment between garment and body under complex or uncommon poses.	12
Texture artifacts	Blurring, loss of fine fabric details, or inconsistent texture rendering.	18
Identity inconsistencies	Partial loss of facial or body identity features after try-on synthesis.	8
Background inconsistencies	Artifacts or blending issues between the synthesized subject and background.	12

4.3.6 Ablation Study

Ablation experiments are conducted to analyze the contribution of individual architectural components within IMAGDressing-v1. Major conditioning pathways, including garment-aware encoding, pose guidance, segmentation cues, and output resolution, are selectively removed to evaluate their influence on synthesis quality. All ablated configurations are evaluated under identical datasets and experimental settings to ensure fair comparison.

4.3.7 Experimental Setup

Performance comparison utilizes FID, LPIPS, SSIM, and PSNR metrics. Results reported in Table 9 quantify performance degradation resulting from removal of individual conditioning mechanisms.

Table 9 Comprehensive ablation analysis of key architectural components in IMAGDressing-v1, evaluating their impact on synthesis realism, perceptual similarity, and structural fidelity using standard image quality metrics.

Model Variant	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Full IMAGDressing v1	18.7	0.142	0.812	27.6
w/o CLIP Conditioning	26.3	0.201	0.746	24.1
w/o Pose Guidance	29.8	0.228	0.721	23.4
w/o Segmentation	24.5	0.187	0.759	25.0
Low-Resolution Output	31.2	0.243	0.702	22.8

4.3.8 Results and Analysis

The ablation results demonstrate that pose guidance plays a critical role in maintaining structural plausibility, while semantic conditioning strengthens garment correspondence. Removing segmentation information reduces boundary accuracy and introduces foreground-background blending artifacts. Resolution reduction significantly affects reconstruction of high-frequency texture details. Across all evaluated variants, the complete IMAGDressing-v1 configuration consistently achieves superior perceptual realism and alignment stability. These observations confirm that coordinated semantic, geometric, and structural conditioning is essential for diffusion-based virtual try-on synthesis [11, 28].

Compared to recent diffusion-based and reconstruction-oriented methods, the proposed approach maintains competitive performance while offering a simplified conditioning mechanism compared to more complex multi-stage or fully trainable diffusion pipelines, thereby reducing implementation complexity and training overhead. This enables practical deployment in cloud-based e-commerce scenarios without compromising visual fidelity, identity preservation, or pose consistency.

5 Discussion

5.1 Overview

This section interprets the experimental observations reported in Section 4 by examining synthesis behavior, robustness characteristics, and practical implications of the IMAGDressing-v1 framework. Rather than focusing solely on numerical improvements, the discussion analyzes how diffusion-based conditioning influences visual realism, structural alignment, and deployment feasibility in representative virtual try-on scenarios [19–22]. The findings are considered within the broader transition of virtual try-on research from deterministic image warping toward probabilistic generative modeling paradigms [22]. From this perspective, diffusion-driven conditioning jointly balances appearance fidelity and spatial controllability, aspects that were frequently optimized independently in earlier approaches. Compared to recent diffusion-based and reconstruction-oriented methods [27, 30], IMAGDressing maintains competitive performance while offering a comparatively simpler and more implementation-efficient conditioning mechanism and reduced training requirements. In addition, recent state-of-the-art approaches such as PG-VTON [31], LFGarNet [33], and related geometry-driven or reconstruction-based methods [37, 38] emphasize explicit structural modeling, multi-attribute garment representation, or deformation-aware learning. While these methods improve geometric consistency and garment dynamics, they often introduce increased architectural complexity and computational overhead. In contrast, the proposed IMAGDressing-v1 framework adopts a diffusion-based conditioning strategy that enables controllable synthesis with comparatively simpler design and reduced training requirements, while still maintaining competitive visual realism and alignment performance.

5.2 Performance Interpretation

Observed improvements in perceptual and structural evaluation metrics are closely related to the iterative refinement mechanism inherent to diffusion models [15, 27]. In

contrast to single-pass synthesis pipelines such as VITON [1], progressive denoising enables gradual correction of geometric inconsistencies while recovering fine garment textures across successive sampling steps. Conditioning-aware diffusion guidance contributes to stable synthesis behavior, maintaining alignment between garment appearance and human pose configuration [19]. Additionally, stochastic sampling introduces moderate appearance variability that can enhance perceptual naturalness without substantially affecting identity preservation. This behavior is particularly relevant in apparel visualization, where minor texture variations can improve perceived realism and user experience.

5.3 Component Contribution Analysis

The ablation analysis indicates that synthesis quality emerges from cooperative interaction among multiple conditioning signals rather than dependence on a single architectural component. Pose guidance contributes primarily to articulation consistency, reducing structural distortions around highly deformable regions. Semantic conditioning strengthens garment correspondence, whereas human parsing improves boundary localization and foreground separation. Higher spatial resolution supports recovery of fine fabric patterns and texture continuity. These observations align with trends identified in recent modular virtual try-on studies emphasizing multi-source conditioning and hierarchical feature fusion [11, 28]. Furthermore, insights from recent 3D reconstruction and attention-based methods highlight the importance of hierarchical feature encoding for structural consistency and identity preservation [9, 10]. The results therefore suggest that diffusion-based frameworks benefit from integrating complementary geometric and semantic priors operating across latent and spatial representations.

5.4 Failure Case Analysis

Despite favorable average performance, several challenging conditions remain evident. Extreme body articulations, partial occlusions, and visually cluttered scenes occasionally introduce geometric inconsistencies or incomplete garment formation [30, 37]. Highly detailed materials or reflective textures may exceed the effective reconstruction capacity of latent representations, resulting in localized smoothing artifacts. Identity inconsistencies are most noticeable near overlap regions involving hair boundaries or facial contours, suggesting limitations in feature disentanglement under complex spatial interactions [17]. Qualitative inspection further indicates that ambiguity in depth ordering and self-occlusion continues to represent a major challenge, particularly for loose or layered garments. These observations indicate that improving depth reasoning and occlusion-aware conditioning remains critical for achieving robust real-world deployment. For transparency and reproducibility, representative examples of both successful and failure cases are made publicly available through the project's official repository and its corresponding archival version.

5.5 Computational and Deployment Considerations

Diffusion-based synthesis inherently introduces additional computational cost due to iterative denoising procedures when compared with lightweight image translation networks [15]. Consequently, inference latency and memory consumption remain

important considerations for large-scale or edge deployment environments. Recent research on timestep reduction, sampler optimization, and architectural simplification demonstrates promising trade-offs between computational efficiency and perceptual quality [27]. Hardware-aware optimization strategies, including mixed-precision inference and model compression, may further support scalable deployment across cloud-assisted and consumer-facing retail platforms. Importantly, all experiments reported in this study are fully reproducible using publicly available source code and datasets provided through the project's official repository and its corresponding DOI-archived version, thereby ensuring transparency and facilitating independent verification of the results.

5.6 Practical and Educational Implications

From an application perspective, IMAGDressing-v1 demonstrates how diffusion-based virtual try-on systems can support scalable digital content generation by reducing dependence on repeated garment photography [18, 21]. Compared to recent diffusion-based or reconstruction-oriented methods [27, 30], the framework achieves competitive performance while simplifying conditioning mechanisms and minimizing retraining. In addition to industrial relevance, the system provides an accessible experimental platform for studying conditional diffusion modeling, multimodal conditioning, and evaluation methodologies frequently discussed in contemporary surveys [19, 20]. The modular architecture enables controlled experimentation with conditioning signals and sampling configurations, making it suitable for instructional and exploratory research contexts involving human-centered image synthesis. Availability of the preprint further ensures early dissemination and accessibility of findings to the broader community.

6 Conclusion and Future Work

6.1 Conclusion

This study presents a systematic evaluation of diffusion-based generation for image-based virtual try-on through the IMAGDressing-v1 framework. Experimental evaluation indicates that integrating pose, semantic, and parsing cues within a pretrained latent diffusion pipeline significantly improves garment realism, structural alignment, and perceptual consistency relative to representative earlier approaches such as VITON [1]. Comparative analysis demonstrates consistent improvements across perceptual and structural evaluation metrics while maintaining practical inference feasibility. In addition, the framework achieves competitive performance relative to recent diffusion-based and reconstruction-oriented approaches while employing a simplified conditioning strategy and reduced training requirements [19–22]. Ablation results further confirm that the observed performance gains arise from coordinated multi-source conditioning rather than reliance on pretrained generative priors alone.

To promote transparency and reproducibility, all source code, pretrained models, and curated datasets, including representative successful and failed synthesis cases, are provided through the same links. These resources enable independent replication, detailed inspection of synthesis behavior, and evaluation under diverse virtual try-on scenarios. No additional model training beyond conditioning modules is required, further supporting reproducibility and reducing computational barriers for adoption.

Overall, the findings support diffusion-based virtual try-on as a promising direction for controllable human-centric image synthesis and scalable digital fashion visualization [19–22], and point toward natural extensions that incorporate 3D garment modeling and 3D-aware representations to further enhance structural fidelity and immersive virtual try-on experiences.

6.2 Future Work

Despite the observed improvements, several challenges remain open for further investigation in diffusion-based virtual try-on systems. Handling multi-garment layering, accessories, and complex outfit interactions requires improved reasoning under occlusion, richer structural annotations, and integration of garment hierarchy information [30, 37]. Robust performance across wider pose distributions, diverse identities, and unconstrained backgrounds may benefit from incorporating depth-aware conditioning, geometry-informed latent representations, or hybrid 3D priors [17].

A natural extension of this work involves integrating full 3D garment modeling and 3D-aware virtual try-on, enabling volumetric garment understanding, more accurate occlusion handling, and realistic garment draping. Incorporating 3D meshes, parametric human body models, or neural radiance fields could enhance structural fidelity and facilitate interactive applications such as virtual fitting, pose-adaptive garment deformation, and multi-garment layering in immersive environments [31, 32].

Future studies may also explore graph-based or parametric garment representations capable of preserving structural consistency during manipulation, supporting applications such as interactive outfit design and personalized digital fitting [38]. From a systems perspective, reducing inference latency and memory consumption remains essential for real-time deployment; advances in accelerated diffusion sampling, efficient architectural designs, and mixed-precision inference provide promising pathways toward practical virtual try-on applications [27].

Finally, integration with augmented and mixed-reality platforms, combined with user preference modeling, represents a natural extension toward immersive and personalized digital fashion experiences. Making detailed performance cases, including both successes and failures, available alongside the codebase ensures that the research remains transparent, reproducible, and actionable for both academic and industrial audiences. Collectively, these directions highlight the potential for advancing diffusion-based virtual try-on systems toward robust, 3D-aware, scalable, and user-adaptive solutions in real-world applications.

Statements and Declarations

Funding

No external funding was received for this study.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

The datasets analyzed during the current study are publicly available virtual try-on benchmark datasets, including VITON and VITON-HD. The corresponding access links and dataset details are provided in the project repository under the file

`dataset_file`. The implementation and resources can be accessed at GitHub repository: https://github.com/Sahasra75/IMAGDressing-VTON. No new dataset was generated as part of this study.

Acknowledgements

The authors would like to thank the department and institutional support provided during the course of this study.

References

[1] Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: An Image-based Virtual Try-on Network (2018). https://doi.org/10.48550/arXiv.1711.08447

[2] Fele, B., Lampe, A., Peer, P., Struc, V.: C-VTON: Context-Driven Image-Based Virtual Try-On Network (2022). https://doi.org/10.1109/WACV51458.2022.00226

[3] Hauswiesner, S., Straka, M., Reitmayr, G., Cirio, G., Grasset, R., Schmalstieg, D.: Virtual try-on through image-based rendering. IEEE Transactions on Visualization and Computer Graphics **19**(9), 1552–1565 (2013) https://doi.org/10.1109/TVCG.2013.67

[4] Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: VITON-HD: High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions (2022). https://doi.org/10.1007/978-3-031-19790-1_13

[5] Kips, R., Jiang, R., Ba, S., Duke, B., Perrot, M., Gori, P., Bloch, I.: Real-time virtual-try-on from a single example image through deep inverse graphics and learned differentiable renderers. Computer Graphics Forum **41**(2), 29–40 (2022) https://doi.org/10.1111/cgf.14456

[6] Santesteban, I., Otaduy, M.A., Casas, D.: Learning-based animation of clothing for virtual try-on. Computer Graphics Forum **38**(2), 355–366 (2019) https://doi.org/10.1111/cgf.13643

[7] Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics (Proc. SIGGRAPH) (2017) https://doi.org/10.1145/3072959.3073711

[8] Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/CVPR42600.2020.00650

[9] Yang, W., Xie, L., Qian, W., Wu, C., Yang, H.: Coarse-to-fine cascaded 3d hand reconstruction based on ssgc and mhsa. The Visual Computer **41**(1), 11–24 (2025) https://doi.org/10.1007/s00371-024-03305-6

[10] Qin, X., Li, X., Li, M., Zheng, H., Xu, X.: Self-supervised single-image 3d face reconstruction method based on attention mechanism and attribute refinement. The Visual Computer **41**(1), 209–227 (2025) https://doi.org/10.1007/s00371-024-03319-0

[11] Wu, Z., Shen, I.-C., Igarashi, T.: Real-time per-garment virtual try-on with temporal consistency for loose-fitting garments. Computer Graphics Forum **44**(7), 70272 (2025) https://doi.org/10.1111/cgf.70272

[12] Xu, Y., Gu, T., Chen, W., Chen, A.: OOTDiffusion: Outfitting Fusion Based Latent Diffusion for Controllable Virtual Try-On (2025). https://doi.org/10.1609/aaai.v39i9.32973

[13] Wang, C., Chen, T., Chen, Z., Huang, Z., Jiang, T., Wang, Q., Shan, H.: Fldm-vton: Faithful latent diffusion model for virtual try-on. (2024). https://doi.org/10.24963/ijcai.2024/151

[14] Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., Liu, A.-A.: CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model (2024). https://doi.org/10.1109/CVPR52733.2024.00800

[15] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (2020). https://doi.org/10.48550/arXiv.2006.11239

[16] Islam, T., Miron, A., Liu, X., Li, Y.: Stylevton: A multi-pose virtual try-on with identity and clothing detail preservation. Neurocomputing **594**, 127887 (2024) https://doi.org/10.1016/j.neucom.2024.127887

[17] Mobinizadeh, H., Lakizadeh, A.: Improving virtual try-on clothes using image depth estimation. Scientific Reports **15**(1), 32192 (2025) https://doi.org/10.1038/s41598-025-18107-6

[18] Lei, H., Zhao, X., Liang, Y., Cao, Y.: Cg-vton: Controllable generation of virtual try-on images based on multimodal guidance. IET Image Processing **19**(1), 70144 (2025) https://doi.org/10.1049/ipr2.70144

[19] Song, D., Zhang, X., Zhou, J., Nie, W., Tong, R., Kankanhalli, M., Liu, A.-A.: Image-Based Virtual Try-On: A Survey (2023). https://doi.org/10.48550/arXiv.2311.04811

[20] Islam, T., Miron, A., Liu, X., Li, Y.: Deep learning in virtual try-on: A comprehensive survey. IEEE Access **12**, 29475–29502 (2024) https://doi.org/10.1109/ACCESS.2024.3368612

[21] Chen, C., Ni, J., Zhang, P.: Virtual try-on systems in fashion consumption: A systematic review. Applied Sciences **14**(24), 11839 (2024) https://doi.org/10.3390/app142411839

[22] Sun, H.: Virtual Try-On Methods: A Comprehensive Research and Analysis (2023). https://doi.org/10.2991/978-94-6463-300-9_35

[23] Islam, T., Miron, A., Liu, X., Li, Y.: Image-based virtual try-on: Fidelity and simplification. Signal Processing: Image Communication **129**, 117189 (2024) https://doi.org/10.1016/j.image.2024.117189

[24] Wu, Z., Liu, J., Chong, T., Shen, I.-C., Igarashi, T.: Virtual measurement garment for per-garment virtual try-on. In: Proceedings of Graphics Interface (GI) / Computer Graphics Forum (2024). https://doi.org/10.1145/3670947.3670957

[25] Zhou, C., Zhang, W., Lian, Z.: Enhancing consistency in virtual try-on: A novel diffusion-based approach (2024). https://doi.org/10.1016/j.imavis.2024.105097

[26] Baek, S., Lee, J.T.: High-speed and precise virtual try-on with two-stage semantic segmentation. ETRI Journal **47**(5), 881–892 (2025) https://doi.org/10.4218/etrij.2024-0592

[27] Dam, P., Jeong, J., Tran, A., Kim, D.: Time-Efficient and Identity-Consistent Virtual Try-On Using A Variant of Altered Diffusion Models (2024). https://doi.org/10.48550/arXiv.2403.07371

[28] Jiang, J., Wang, T., Yan, H., Liu, J.: ClothFormer: Taming Video Virtual Try-On in All Module (2022). https://doi.org/10.1109/CVPR52688.2022.01053

[29] Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Learning Semantic Correspondence with Latent Diffusion Model (2023). https://doi.org/10.48550/arXiv.2312.01725

[30] Lee, J., Jung, H., Choi, J.: 3d-aware virtual try-on using only 2d inputs. Computer Vision and Image Understanding **264**, 104661 (2026) https://doi.org/10.1016/j.cviu.2026.104661

[31] Zheng, J., Sang, S., Lu, Y., Dai, G., Mao, X., Zhou, W.: Pg-vton: Front-and-back garment guided panoramic gaussian virtual try-on with diffusion modeling. Computer Animation and Virtual Worlds **36**(3), 70054 (2025) https://doi.org/10.1002/cav.70054

[32] Jiang, M., Tian, Z., Yu, C., Shi, Y., Liu, L., Peng, T., Hu, X., Yu, F.: Intelligent 3d garment system of the human body based on deep spiking neural network. Virtual Reality & Intelligent Hardware **6**(1), 43–55 (2024) https://doi.org/10.1016/j.vrih.2023.07.002

[33] Zhang, P., Fei, B., Wei, M., Zhan, J., Wang, K., Lv, Y.: Lfgarnet: Loose-fitting garment animation with multi-attribute-aware graph network. Computer Animation and Virtual Worlds **36**(2), 70017 (2025) https://doi.org/10.1002/cav.70017

[34] Wang, J., Lu, X., Bennamoun, M., Sheng, B.: Non-rigid point cloud registration

via anisotropic hybrid field harmonization. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(9), 7898–7915 (2025) https://doi.org/10.1109/TPAMI.2025.3572584

[35] Hu, X., Yang, C., Fang, F., Huang, J., Li, P., Sheng, B., Lee, T.-Y.: Msembgan: Multi-stitch embroidery synthesis via region-aware texture generation. IEEE Transactions on Visualization and Computer Graphics **31**(9), 5334–5347 (2025) https://doi.org/10.1109/TVCG.2024.3447351

[36] Xiong, M., Hu, K., Wang, Z., Hu, R., Muhammad, K., Del Ser, J., Yang, X., Sheng, B.: Adaptive clustering and weighted regularization contrastive learning framework for unsupervised person re-identification. IEEE Transactions on Multimedia **27**, 6896–6907 (2025) https://doi.org/10.1109/TMM.2025.3590938

[37] Dang, Y., Wan, T.R., Xi, L., Tang, W.: A computer-vision based framework for virtual 3d garment reconstruction. Multimedia Tools and Applications **84**, 27963–27985 (2024) https://doi.org/10.1007/s11042-024-20269-w

[38] Vidaurre, R., Santesteban, I., Garcés, E., Casas, D.: Fully convolutional graph neural networks for parametric virtual try-on. Computer Graphics Forum **39**(8), 145–156 (2020) https://doi.org/10.1111/cgf.14109