UniFit: Towards Universal Virtual Try-on with MLLM-Guided Semantic Alignment

Wei Zhang[1,*], **Yeying Jin**[2,*], **Xin Li**[3], **Yan Zhang**[4], **Xiaofeng Cong**[5], **Cong Wang**[6], **Fengcai Qiao**[7], **Zhichao Lian**[1,†]

[1]School of Cyber Science and Engineering, Nanjing University of Science and Technology
[2]Tencent
[3]University of Science and Technology of China
[4] ByteDance
[5] Southeast University
[6] University of California, San Francisco
[7] National University of Defense Technology
zwplus_pro@njust.edu.cn, jinyeying@u.nus.edu, xin.li@ustc.edu.cn, yanzhang.yz@bytedance.com
cxf_svip@163.com, supercong94@gmail.com, fcqiao@nudt.edu.cn, newlzcts@njust.edu.cn

Abstract

Image-based virtual try-on (VTON) aims to synthesize photorealistic images of a person wearing specified garments. Despite significant progress, building a universal VTON framework that can flexibly handle diverse and complex tasks remains a major challenge. Recent methods explore multi-task VTON frameworks guided by textual instructions, yet they still face two key limitations: (1) semantic gap between text instructions and reference images, and (2) data scarcity in complex scenarios. To address these challenges, we propose UniFit, a universal VTON framework driven by a Multimodal Large Language Model (MLLM). Specifically, we introduce an MLLM-Guided Semantic Alignment Module (MGSA), which integrates multimodal inputs using an MLLM and a set of learnable queries. By imposing a semantic alignment loss, MGSA captures cross-modal semantic relationships and provides coherent and explicit semantic guidance for the generative process, thereby reducing the semantic gap. Moreover, by devising a two-stage progressive training strategy with a self-synthesis pipeline, UniFit is able to learn complex tasks from limited data. Extensive experiments show that UniFit not only supports a wide range of VTON tasks, including multi-garment and model-to-model try-on, but also achieves state-of-the-art performance.

Introduction

Image-based virtual try-on (VTON) aims to synthesize photorealistic images of a person wearing specified garments, with broad applications in e-commerce. Recent methods, such as OOTD (Xu et al. 2025), IMAGDressing (Shen et al. 2025), and TryoffDiff (Velioglu et al. 2024), have achieved impressive results. However, most existing approaches are task-specific, such as single-garment try-on.

To address this limitation, recent studies (Guo et al. 2025; Zhang et al. 2024) have begun to explore multi-task VTON frameworks guided by textual instructions. As illustrated in Figure 1(a), a text encoder (e.g., CLIP (Radford et al. 2021) or T5 (Raffel et al. 2020)) extracts task-relevant cues which condition the generative model to extract relevant visual features from the reference images. However, relying only on abstract language for guidance often results in poor grounding in concrete visual details (e.g., texture or logo shape), causing a semantic gap. This gap prevents the model from faithfully integrating task-relevant visual features from the reference images, often resulting in outputs with low fidelity and weak controllability. Moreover, limited data for

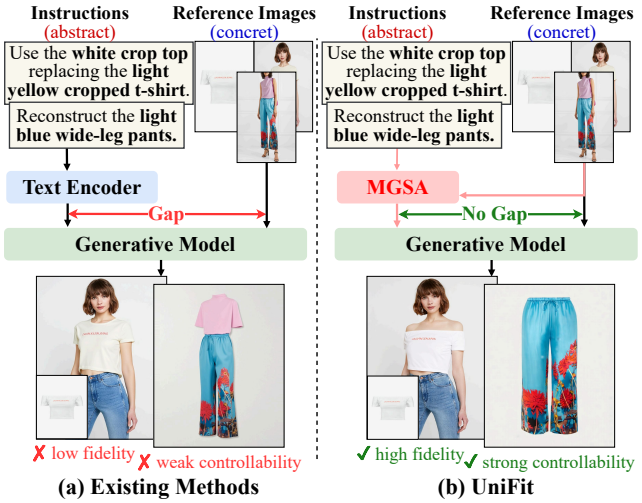

Figure 1: Motivation of UniFit: (a) Existing instruction-guided VTON methods process text and images separately, resulting in a semantic gap. (b) Our UniFit introduces an MLLM-Guided Semantic Alignment Module (MGSA), which integrates textual and visual inputs to produce coherent and explicit semantic guidance for the generative model, effectively bridging the semantic gap.

*These authors contributed equally.
†Corresponding Author

Method	Single-garment try-on	Model-free try-on	Garment reconstruction	Multi-view try-on	Multi-garment try-on	Model-to-model try-on
AnyFit (Li et al. 2024)	✓	-	-	-	✓	-
CatVTON (Chong et al. 2024)	✓	-	-	-	-	✓
MV-VTON (Wang et al. 2025)	✓	-	-	✓	-	-
Any2AnyTryon (Guo et al. 2025)	✓	✓	✓	-	-	-
UniFit (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of VTON functionalities achieved by UniFit and previous methods. UniFit is capable of handling multiple VTON tasks, including multi-view try-on, multi-garment try-on, and model-to-model try-on, surpassing previous methods.

complex scenarios in public datasets further restricts current methods from supporting advanced tasks such as model-to-model and multi-garment try-on (see Table 1).

To address these challenges, we propose UniFit, an instruction-guided, universal VTON framework that integrates a Multimodal Large Language Model (MLLM) with a Diffusion Transformer (DiT) (Peebles and Xie 2023). First, to align textual instructions with reference images, we introduce a novel MLLM-Guided Semantic Alignment Module (MGSA), which leverages the MLLM and a set of learnable queries to fuse multimodal inputs. Supervised by a semantic alignment loss, MGSA captures cross-modal semantic relationships and generates coherent, explicit guidance for the DiT, effectively bridging the gap between text and vision. To further enhance generation quality, we introduce a spatial attention focusing loss for the DiT, which regularizes the cross-attention maps between the reference images and the generated output. This encourages the DiT to focus on the task-relevant areas (e.g., the garment area to be replaced), enabling faithful transfer of fine-grained visual details.

To address data scarcity towards universal VTON, we propose a two-stage progressive training strategy with a self-synthesis pipeline. In the first stage, a Base Model is trained on publicly available datasets to learn foundational tasks such as model-free try-on and garment reconstruction. Then, the Base Model serves as a data synthesizer, generating high-quality paired samples for advanced tasks. For instance, it leverages its learned garment reconstruction ability to extract disentangled upper and lower garments from a single image, creating training samples for multi-garment try-on. In the second stage, the model is fine-tuned on a composite dataset of both real and synthesized data to support more advanced tasks. This training strategy enables UniFit to gradually master a wide range of VTON tasks, effectively overcoming the data limitations of existing datasets. Our contributions are summarized as follows:

- We propose UniFit, a universal VTON framework capable of handling a diverse range of tasks, from standard single-garment try-on to complex multi-garment and model-to-model try-on.
- We propose an MLLM-Guided Semantic Alignment Module (MGSA) to bridge the semantic gap between textual instructions and reference images. By imposing a semantic alignment loss, MGSA can capture cross-modal relationships and generate explicit guidance for the generative process. Additionally, we introduce a spatial attention focusing loss to regularize the DiT's attention, ensuring faithful transfer of fine-grained visual details.
- We devise a two-stage progressive training strategy with a self-synthesis pipeline, which effectively overcomes data scarcity for advanced VTON tasks, significantly enhancing the model's ability to handle multi-garment and model-to-model try-on tasks.

Method

Overview We propose UniFit, a universal virtual try-on framework capable of generating corresponding try-on results based on textual instructions and reference images. As illustrated in Figure 2, the overall architecture consists of three main components: an MLLM-Guided Semantic Alignment Module (MGSA) to associate textual instructions with visual content, a VAE encoder (Kingma and Welling 2013) for extracting low-level visual features, and a Diffusion Transformer (DiT) serving as the generative backbone. The generation pipeline begins with two parallel streams. The MGSA leverages a pre-trained MLLM (Qwen2-VL(Wang et al. 2024)) and a set of learnable queries to capture the semantic relationships between textual instructions and reference images, producing a high-level semantic representation of the target image, denoted as T_q. This representation serves as explicit semantic guidance for the subsequent generation process. Simultaneously, the VAE encoder processes the reference images to extract fine-grained visual features, denoted as $r = \{r_1, \ldots, r_n\}$, which provide rich visual cues for generation. We then concatenate T_q, the noisy latent z_t of the target image, and the reference tokens r to form the DiT input $[T_q; z_t; r_1; \ldots; r_n]$. During the iterative denoising process, the DiT, guided by T_q, effectively integrates the fine-grained visual information in r to progressively refine z_t into the final output image.

To enhance pose consistency and detail fidelity, we incorporate two auxiliary components. First, inspired by Catv2ton (Chong et al. 2025), we employ a Pose Block (a trainable copy of the first DiT block) to extract pose features, which are then injected into the second DiT block's input via element-wise addition to enforce pose alignment. Second, a spatial attention focusing loss regularizes the DiT's attention maps, ensuring the model precisely focuses on relevant regions when transferring details. During training, we jointly optimize a specific set of components: the MGSA

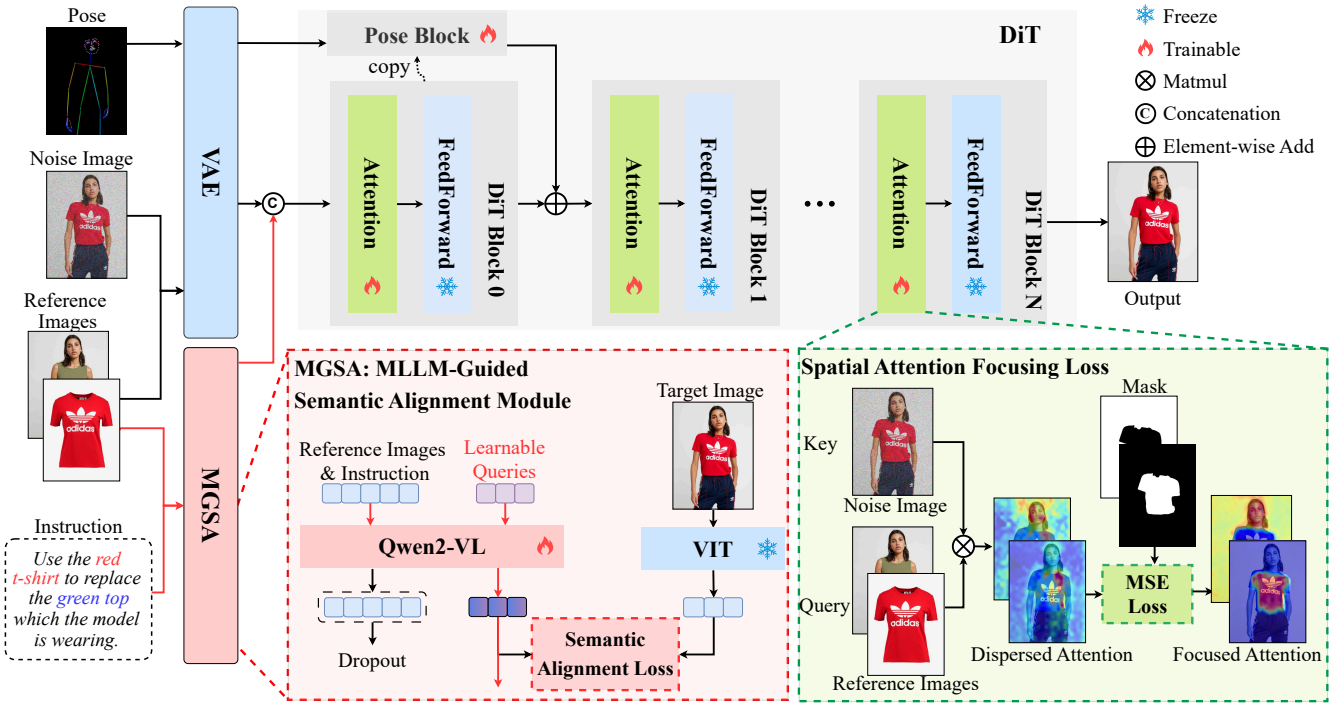

Figure 2: **Overview of UniFit.** UniFit consists of three main components: the MGSA module (red), the DiT (gray), and the VAE encoder (blue). The MGSA encodes multimodal inputs into coherent semantic guidance. The VAE extracts low-level visual features from reference images. The DiT generates the output image conditioned on the semantic guidance and low-level visual features. Additionally, a spatial attention focusing loss (green) supervises the attention maps of the DiT, encouraging the model to focus on the most task-relevant regions (e.g., the try-on area in single-garment try-on task).

(including the MLLM and learnable queries), the attention layers within the DiT, and the Pose Block, while keeping other parts like the VAE and the DiT's FeedForward layers frozen. The overall training objective is a combination of three components: (1) a flow matching loss, (2) the semantic alignment loss, and (3) the spatial attention focusing loss.

MLLM-Guided Semantic Alignment Module

A key challenge in instruction-guided multi-task VTON frameworks is bridging the semantic gap between abstract textual instructions and concrete visual content (see Figure 1). To address this issue, we propose the MLLM-Guided Semantic Alignment Module (MGSA) (see Figure 2 red), which leverages a pre-trained Multimodal Large Language Model (Qwen2-VL) to jointly process both textual and visual inputs. To enable the MGSA to effectively capture the semantic relationships between these inputs and generate a coherent and explicit semantic guidance for the subsequent generation process, we introduce two core components: a set of learnable queries and a semantic alignment loss.

As illustrated in Figure 2, the MGSA jointly processes reference image tokens and textual instruction tokens via the pre-trained Qwen2-VL. However, the resulting token sequence is typically long, introducing significant redundant information and computational overhead for the downstream DiT. To mitigate this, we introduce a set of learnable queries, $T_q \in \mathbb{R}^{N_q \times D_q}$, where N_q is the number of queries and D_q denotes their dimensionality. These queries are appended to the end of Qwen2-VL's input sequence and serve as information aggregators. Through Qwen2-VL's causal attention mechanism, they distill task-relevant signals from the extensive multimodal sequence into a compact representation.

To ensure that the representation T_q is semantically meaningful and aligned with the target output, we introduce a semantic alignment loss, $\mathcal{L}_{\text{align}}$, which aligns T_q with the ground-truth visual representation of the target image. Specifically, we extract a set of visual tokens $T_v \in \mathbb{R}^{N_v \times D_v}$ from the target image using a frozen VIT (Dosovitskiy et al. 2020), and set $N_q = N_v$. We then enforce a token-wise alignment between them using cosine similarity:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N_v} \sum_{n=1}^{N_v} \cos\left(T_{v,n}, \text{MLP}(T_{q,n})\right), \qquad (1)$$

where $T_{v,n}$ and $T_{q,n}$ represent the n-th tokens from the visual and query embeddings, respectively. A lightweight MLP projection is used to match the feature dimensions of T_q and T_v. The learnable queries and semantic alignment loss enable the MGSA to fuse multimodal inputs into a coherent semantic representation that captures task intent and visual cues. This representation provides explicit guidance to the DiT, allowing it to aggregate relevant features and generate high-quality outputs aligned with the instructions.

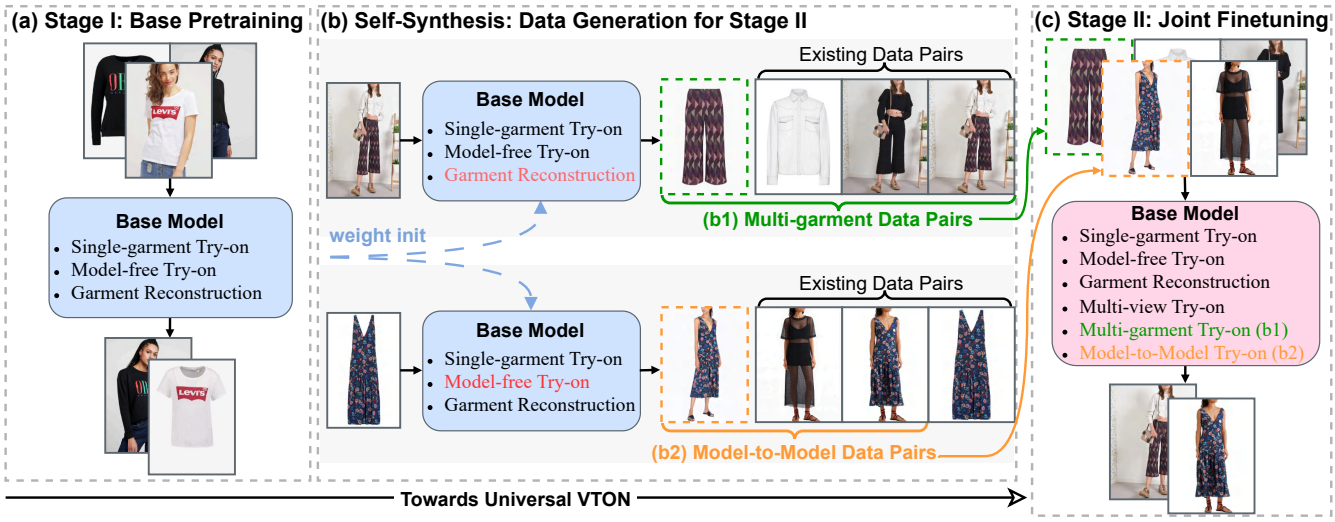

Figure 3: **Towards Universal VTON**: Two-Stage Progressive Training Strategy of UniFit with Self-Synthesis. **(a) Stage I:** A Base Model is trained on basic tasks using public datasets. **(b) Self-Synthesis:** The trained model is then used to generate pseudo-paired data for complex scenarios. **(b1)** For multi-garment try-on, we reconstruct garments from full-body images. **(b2)** For model-to-model try-on, we synthesize new person images conditioned on given garments. **(c) Stage II:** The model is fine-tuned on a composite dataset of both real and synthesized samples, enabling generalization to a wide range of VTON tasks.

Spatial Attention Focusing Loss

Although the MGSA provides strong high-level semantic guidance, the DiT's attention still tends to be dispersed across irrelevant areas in both the reference image and the generated output. This dispersion often degrades fine details and introduces visual artifacts. As illustrated in the bottom-right corner of Figure 2 (see green), the initial cross-attention between the garment image and the output is scattered rather than concentrated on the intended try-on area. To mitigate this issue, we introduce a spatial attention focusing loss inspired by DreamO (Mou et al. 2025). The loss regularizes the cross-attention maps, encouraging the model to focus on regions that are truly critical for the task.

In detail, we first compute the cross-attention map $AttnMap \in \mathbb{R}^{l_{r_i} \times l_{z_t}}$, using the reference tokens as queries and the output tokens as keys, where l_{r_i} and l_{z_t} denote their respective sequence lengths. Then, for try-on tasks, we average the map over the reference-token axis to obtain an output-centric response map $M \in \mathbb{R}^{l_{z_t}}$, which highlights the target areas expected to receive garment features or the model's appearance features. For garment-reconstruction tasks, we average over the output-token axis, producing a reference-centric map $M \in \mathbb{R}^{l_{r_i}}$ that pinpoints where to extract garment details in the reference image. Specifically, the model-to-model try-on task requires both extracting a specific garment from the reference image and transferring it to the try-on region of the generated output. Therefore, we compute and supervise both attention response maps simultaneously. Finally, we align each task-specific map M with a ground-truth spatial mask M_{target} via an MSE loss:

$$\mathcal{L}_{\text{focus}} = \frac{1}{N_R \times N_L} \sum_{j=1}^{N_L} \sum_{i=1}^{N_R} \left\| M_i^j - M_{\text{target},i} \right\|_2^2, \quad (2)$$

where M_i^j denotes the response map for the i-th reference image in the j-th attention layer, $M_{\text{target},i}$ is the ground-truth spatial mask corresponding to the i-th reference image, and N_R, N_L represent the number of reference images and attention layers, respectively. This targeted supervision enforces accurate spatial correspondence, enabling high-fidelity transfer of fine-grained details and preventing visual artifacts in the synthesized images.

Progressive Training via Self-Synthesis

Public VTON datasets such as VITON-HD (Choi et al. 2021) and DressCode (Morelli et al. 2022) have greatly advanced the open-source research community, yet they also exhibit inherent limitations. They provide only image pairs of a single garment and its corresponding try-on result, which introduces two challenges. First, such pairings compel many methods to depend on garment masks. Second, and more critically, they lack sufficient data to support complex tasks such as multi-garment and model-to-model try-on. Recent mask-free approaches mitigate the first issue by synthesizing triplet training samples via image inpainting, but data scarcity for complex tasks remains a bottleneck. To address this, we propose a two-stage progressive training strategy with a self-synthesis pipeline (Figure 3). Throughout training, reference-model images are generated by inpainting, thereby removing any reliance on garment masks.

Stage I: Base Pretraining. In the first stage (Figure 3(a)), we focus on pretraining a Base Model on three fundamental tasks: single-garment try-on, garment reconstruction, and model-free try-on. This pretraining, conducted on standard datasets such as VITON-HD and DressCode, equips the model with strong consistency-preserving capabilities.

Self-Synthesis for Complex Tasks. Next, we use the

Method	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓
TryOffDiff	**0.792**	0.337	0.227	21.40
Any2AnyTryon	0.762	0.367	0.231	13.57
Ours	0.775	**0.281**	**0.202**	**12.58**

Table 2: Quantitative comparison of garment reconstruction on the VITON-HD dataset. Best results are in **bold**.

Figure 4: Qualitative comparison of garment reconstruction.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
CatVTON	0.888	0.075	9.128	1.130
FitDiT	**0.895**	0.067	9.326	0.913
Any2AnyTryon	0.839	0.088	8.965	0.981
Ours	0.883	**0.065**	**8.799**	**0.702**

Table 3: Quantitative comparison of single garment try-on on the VITON-HD dataset.

Figure 5: Qualitative comparison of single-garment try-on.

Base Model as a data synthesizer to generate training samples for complex VTON tasks (Figure 3(b)). Specifically:

- **For Multi-garment Try-on (Figure 3(b1)):** We leverage the model's garment reconstruction ability to extract high-quality tops or bottoms from full-body images, thus creating new training samples for this task.
- **For Model-to-model Try-on (Figure 3(b2)):** We utilize the model-free try-on capability to synthesize new person images conditioned on existing garments, thereby expanding data support for the model-to-model task.

To guarantee the fidelity of the synthesized data, all synthesized samples undergo a two-step filter: DreamSim (Fu et al. 2023) for perceptual similarity, followed by a consistency check with Qwen2.5-VL-7B (Bai et al. 2025).

Stage II: Joint Finetuning. In the second stage (Figure 3(c)), we finetune the Base Model on a composite dataset that merges real images with synthetic data. This joint training enables UniFit to generalize across six VTON tasks, by adding supervision for multi-garment, model-to-model, and multi-view try-on, alongside the three foundational tasks. For the multi-view task, we construct training pairs from the MVG (Wang et al. 2025) dataset and a manually annotated subset of 2,000 image pairs from IG-Pairs (Shen et al. 2025), each containing multi-view try-on results and the corresponding front–back garment references. This training strategy not only mitigates the data scarcity problem but also simplifies the learning process by leveraging the visual priors acquired during Base Model pretraining.

Experiment

Experimental Setup

Datasets Our training incorporates a combination of public and self-synthesized data. For the three foundational tasks (single-garment try-on, model-free try-on, and garment reconstruction), we utilize the VITON-HD and DressCode datasets, which provide approximately 59K training pairs for each task. For the advanced tasks, we leverage our self-synthesis pipeline: for multi-garment try-on, we generate 10K data pairs by reconstructing garments from DressCode; for model-to-model try-on, we synthesize 30K data pairs from both VITON-HD and DressCode. For the multi-view try-on task, we use a combination of the MVG and IG-Pairs datasets, totaling 12K samples. Textual instructions for all tasks are produced with Qwen2.5-VL-7B. Following existing mask-free methods, we employ Flux.1 Fill (Labs 2024) to synthesize triplet training samples.

Training Details The MGSA is based on Qwen2-VL-2B. During training, we freeze its first 14 transformer layers and fine-tune the remaining 14. We directly reuse the ViT embedded in Qwen2-VL-2B to extract target image features; this ViT processes 756×504 inputs and outputs 486 visual tokens. Accordingly, we set the number of learnable queries to 486, each with a dimensionality of 1,536 to match the hidden size of Qwen2-VL-2B. For the DiT backbone, we experiment with StableDiffusion-3.5 Medium (Esser et al. 2024). Our training process is divided into two stages. We first pretrain the model on the three foundational tasks for 120K

Figure 6: Qualitative comparison of multi-garment try-on.

Method	CLIP-AS ↑	CLIP-I ↑	MP-LPIPS ↓
IMAGDressing-v1	**4.96**	0.880	0.107
Any2AnyTryon	4.95	0.843	0.127
Ours	4.91	**0.914**	**0.078**

Table 4: Quantitative comparison for model-free try-on on the VITON-HD dataset.

steps. The training resolution is $1,024\times768$ (or 768×576 for model-free try-on), with a batch size of 16. We then conduct joint fine-tuning on all six tasks for another 80K steps. The resolutions are 1024×768 for most tasks, and 768×576 for model-free and multi-view try-on, with the same batch size of 16. Throughout all training stages, we use the AdamW optimizer with a learning rate of 4×10^{-5}.

Comparative Experiments

Garment Reconstruction We first evaluate our method on the garment reconstruction task, which involves reconstructing a flattened garment image from an image of a model wearing the target garment. To assess performance, we adopt four standard metrics: SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), DISTS (Ding et al. 2020), and FID (Heusel et al. 2017). We compare UniFit with TryOffDiff and Any2AnyTryon on VITON-HD and DressCode. The results, shown in Figure 4 and Table 2, clearly show the superiority of our approach. Qualitatively, TryOffDiff often fails to preserve essential visual attributes such as color and texture, while Any2AnyTryon—an instruction-guided multi-task VTON framework—frequently struggles to comply with task instructions. For instance, as seen in the second row of Figure 4, it fails to reconstruct the garment specified by the textual instruction. In contrast, UniFit accurately follows the textual instructions to reconstruct garments that closely match those worn by the model in the reference image, both in shape and fine-grained patterns. Quantitatively, UniFit achieves state-of-the-art performance on the VITON-HD dataset, significantly surpassing both baselines

Figure 7: Qualitative comparison of model-to-model try-on.

across multiple perceptual and distributional metrics. The strong improvements over Any2AnyTryon further highlight UniFit's ability to generate high-fidelity and semantically aligned results under textual instruction guidance.

Single-Garment Try-On We compare our method with CatVTON, FitDiT (Jiang et al. 2024), and Any2AnyTryon on the single-garment try-on task. As reported in Table 3, UniFit outperforms all baselines across multiple metrics, with notable improvements in FID and KID (Bińkowski et al. 2018), demonstrating superior generation quality. Figure 5 provides qualitative comparisons, where UniFit not only generates higher-quality and more realistic outfitted model images but also ensures that the garments worn by the model align closely with the input garments.

Model-Free Try-On Model-free try-on refers to garment-driven model generation and can be viewed as a specific sub-task of subject-driven image generation. We evaluate UniFit on this task using the VITON-HD dataset and compare it against IMAGDressing-v1 (Shen et al. 2025) and Any2AnyTryon. Three metrics are used for evaluation: CLIP-I (Radford et al. 2021), MP-LPIPS (Chen et al. 2024), and CLIP-AS (Schuhmann et al. 2022). Among them, MP-LPIPS is particularly effective at measuring the visual consistency between the input garment and the garment rendered on the generated model. As shown in Table 4, UniFit achieves superior performance in both MP-LPIPS and CLIP-I. These results indicate that our method not only produces high-quality model images but also better preserves the visual appearance of the input garments, showing strong consistency and realism.

Multi-garment Try-on Most existing VTON methods are restricted to single-garment scenarios and lack the capability to compose and render combinations of tops and bottoms. Due to the absence of open-source multi-garment baselines, we conduct a qualitative comparison with two closed-source tools—OutfitAnyone (Sun et al. 2024) and Kolors-VTON (Team 2024). As illustrated in Figure 6, our method shows superior ability in preserving the shape and details of garments. For instance, Kolors-VTON fails to maintain the skirt texture in the first row and introduces noticeable distortion in the third-row outfit. OutfitAnyone often alters the garment structure, such as the pants in the second row.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
MV-TON	0.930	**0.062**	37.09	**3.23**
Ours	**0.935**	0.072	**35.62**	3.85

Table 5: Quantitative results on the MVG dataset for multi-view try-on.

Figure 8: Qualitative comparison of multi-view try-on.

In contrast, UniFit produces more realistic and coherent try-on results, faithfully retaining the garment appearance and shape across various combinations.

Model-to-model Try-on Model-to-model Try-on aims to transfer specific garments from a source model to a target model. This task demands both accurate comprehension of the instruction and precise localization of the garments to be transferred. While existing methods such as CatVTON rely on explicit garment masks to guide the transfer, UniFit operates in a fully instruction-driven manner without requiring such auxiliary inputs. As shown in Figure 7, qualitative results on the DressCode dataset indicate that UniFit can reliably interpret textual instructions, accurately identify the target garments, and generate high-fidelity try-on results with improved coherence and realism.

Multi-view Try-on Multi-view Try-on aims to generate realistic try-on results by utilizing both the front and back views of the garment. MV-VTON is the first to explore this task, introducing a pose-aware hard-selection and soft-selection mechanism to guide the aggregation of garment features across views. In contrast, our method capitalizes on the cross-modal reasoning capability of the MGSA module, which enables efficient semantic integration of multi-view garment features guided by instructions. As shown in Figure 8, UniFit effectively fuses visual cues from different viewpoints to produce high-quality try-on results. Notably, even under challenging viewing angles (e.g., the side view in the second row), our method maintains garment realism and structural consistency. Furthermore, as shown in Table 5, UniFit achieves superior fidelity and perceptual quality in multi-view settings.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
w/o MGSA	0.851	0.098	9.133	1.053
w/o $\mathcal{L}_{\text{align}}$	0.863	0.074	8.937	0.951
w/o $\mathcal{L}_{\text{focus}}$	0.872	**0.069**	8.870	0.835
Ours (Stage I)	**0.887**	0.071	**8.813**	**0.785**

Table 6: Quantitative results of ablation study on the VITON-HD dataset.

Figure 9: Qualitative results of ablation study on the VITON-HD dataset.

Ablation Study

To demonstrate the effectiveness of our key components—the MGSA and the spatial attention focusing loss—we conduct an ablation study on the single-garment try-on task. To ensure fair comparison, all ablation studies are performed on our Base Model trained only through Stage I, thus avoiding any potential biases introduced by the self-synthesis process. As shown in Table 6 and Figure 9, we evaluate three variants: (1) w/o MGSA, where we replace the MGSA module (including $\mathcal{L}_{\text{align}}$) with a T5 text encoder; (2) w/o $\mathcal{L}_{\text{align}}$, where the semantic alignment loss is removed; and (3) w/o $\mathcal{L}_{\text{focus}}$, where the spatial attention focusing loss is removed. The results clearly show that both ablated models exhibit a noticeable degradation in performance compared to the full model, which confirms the effectiveness and necessity of our proposed components.

Conclusion

In this paper, we present UniFit, an instruction-guided universal VTON framework capable of flexibly handling a diverse range of complex tasks. UniFit effectively addresses two core challenges faced by existing instruction-guided multi-task VTON frameworks: (1) the semantic gap between textual instructions and reference images, and (2) the data scarcity for complex scenarios. Our solution centers on two key innovations. First, an MLLM-Guided Semantic Alignment Module leverages a multimodal large language model to capture semantic relationships between instructions and reference images. It produces coherent and explicit semantic guidance for the generative process, effectively bridging the semantic gap. Second, a two-stage progressive training strategy with a self-synthesis pipeline enables UniFit to learn advanced tasks from limited data, overcoming the limitations of current datasets. Extensive experiments show that UniFit achieves state-of-the-art performance across six VTON tasks, from single-garment try-on to complex multi-garment and model-to-model try-on scenarios.

Acknowledgments

This work was partially supported by the 2024 Jiangsu Province Frontier Technology Research and Development Project (Grant No. BF2024071), the Jiangsu Provincial Science and Technology Major Project (Grant No. BG2024042), and the Qing Lan Project.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Chen, W.; Gu, T.; Xu, Y.; and Chen, A. 2024. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6939–6948.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

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

Chong, Z.; Zhang, W.; Zhang, S.; Zheng, J.; Dong, X.; Li, H.; Wu, Y.; Jiang, D.; and Liang, X. 2025. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Guo, H.; Zeng, B.; Song, Y.; Zhang, W.; Zhang, C.; and Liu, J. 2025. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jiang, B.; Hu, X.; Luo, D.; He, Q.; Xu, C.; Peng, J.; Zhang, J.; Wang, C.; Wu, Y.; and Fu, Y. 2024. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*.

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

Labs, B. F. 2024. FLUX: Official Inference Repository for FLUX.1 Models. Accessed: 2024-11-12.

Li, Y.; Zhou, H.; Shang, W.; Lin, R.; Chen, X.; and Ni, B. 2024. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *Advances in Neural Information Processing Systems*, 37: 83164–83196.

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

Mou, C.; Wu, Y.; Wu, W.; Guo, Z.; Zhang, P.; Cheng, Y.; Luo, Y.; Ding, F.; Zhang, S.; Li, X.; et al. 2025. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Sun, K.; Cao, J.; Wang, Q.; Tian, L.; Zhang, X.; Zhuo, L.; Zhang, B.; Bo, L.; Zhou, W.; Zhang, W.; et al. 2024. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*.

Team, K. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*.

Velioglu, R.; Bevandic, P.; Chan, R.; and Hammer, B. 2024. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. *arXiv preprint arXiv:2411.18350*.

Wang, H.; Zhang, Z.; Di, D.; Zhang, S.; and Zuo, W. 2025. Mv-vton: Multi-view virtual try-on with diffusion models.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7682–7690.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xu, Y.; Gu, T.; Chen, W.; and Chen, A. 2025. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8996–9004.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Lin, E.; Li, X.; Luo, Y.; Kampffmeyer, M.; Dong, X.; and Liang, X. 2024. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*.