SEEING MORE DETAILS: MULTI-REFERENCE VIRTUAL TRY-ON WITH FINE-GRAINED GARMENT REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual try-on (VTON) has achieved remarkable progress with the advancement of diffusion models. However, most existing methods focus on single-image guidance under *in-shop* settings, often producing unsatisfactory results when handling large pose variations, complex occlusions, or fine-grained texture transfer. To address this limitation, we propose a mask-free VTON method, termed **MR-VTON**, which leverages **M**ulti-**R**eference garment images captured from diverse viewpoints and wearing poses to preserve the garment details and realistic try-on appearance, particularly in *in-the-wild* scenarios. The proposed model employs a dual-branch architecture to dynamically adapt reference image selection and fine-grained feature aggregation. Specifically, the first branch introduces a router-based garment selection mechanism that adaptively selects a single image as the primary reference to anchor garment identity features, with minimal occlusion and maximum clothing information. The second branch introduces a target pose-guided feature aggregation module that integrates garment details and viewpoint information from multiple reference images, thereby enhancing spatial consistency and fine-grained detail preservation. The features from both branches are concatenated and injected into a Diffusion Transformer architecture as reference features, facilitating VTON across diverse viewpoints and poses. Furthermore, we construct an evaluation benchmark for in-the-wild multi-reference VTON, named **MRG-4K**, to fill the gap in this emerging task. Extensive experiments on the proposed benchmark and zero-shot academic benchmarks demonstrate that our method consistently outperforms existing approaches and achieves state-of-the-art performance in both single-reference and multi-reference try-on tasks.

1 INTRODUCTION

Virtual try-on aims to synthesize realistic images of a target person wearing a reference garment while preserving garment appearance, person identity, and pose. Owing to its broad applications in online shopping and digital human generation, it has attracted increasing attention. Compared with traditional GAN Goodfellow et al. (2020)-based methods Xie et al. (2021; 2022); Lewis et al. (2021); Xie et al. (2023); Men et al. (2020), recent diffusion-based approaches Chong et al. (2025); Xu et al. (2025); Cui et al. (2025); Jiang et al. (2024); Zhang et al. (2024) have substantially improved visual fidelity and demonstrated stronger preservation of garment textures and human body structure.

Nevertheless, several challenges remain in existing methods. First, most mask-guided or parsing-guided methods Wang et al. (2025a); Xu et al. (2025); Han et al. (2018; 2019) are dependent on the accuracy of the segmentation. Imprecise masks frequently lead to unnatural wearing results or even failure in garment transfer. Second, subject-driven image-editing methods Raisinghani (2025); OpenAI (2026); Wu et al. (2025); Team et al. (2026a), which rely on context guidance, tend to produce more realistic results but often fail to maintain consistent garment identity. Finally, existing methods impose strict requirements on input garment and exhibit poor robustness to in-the-wild scenarios, where occlusion, complex poses, or large viewpoint discrepancies relative to the target person often lead to severe performance degradation. Considering that ideal white-background product images are often difficult for users to obtain in real-world applications, in-the-wild garment images offer broader practical applicability. However, they also introduce greater challenges due to complex

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

backgrounds, occlusion, diverse model poses, and large viewpoint discrepancies. Under such conditions, relying on a single reference image is often insufficient to capture the complete garment structure and recover fine-grained appearance details.

To address these limitations, we propose MR-VTON, a multi-reference virtual try-on method that leverages complementary garment information from multiple reference images. It aims to overcome the incomplete representation of a single reference image and improve both global garment identity and fine-grained detail preservation, especially in challenging in-the-wild scenarios. However, variations in pose, viewpoint, scale, and occlusion may introduce redundant or conflicting cues, reducing the consistency of the synthesized output. To address this challenge, we develop a multi-reference garment guidance strategy that exploits complementary information while suppressing unreliable or conflicting signals. As a result, MR-VTON produces more coherent and high-fidelity try-on results with stronger generalization to diverse real-world scenes. Specifically, we adopt a dual-branch architecture to separately handle optimal reference selection and multi-reference garment feature aggregation. To establish a reliable garment identity anchor and reduce interference from inconsistent references, we design an adaptive main reference selection branch. It aggregates garment features from each reference through attention pooling and uses a learnable gate to select the most informative image. By favoring references with less occlusion, clearer structures, and richer details, the branch suppresses noise and provides a stable foundation for subsequent multi-reference fusion. To retrieve pose-compatible garment details from multiple references and ensure accurate adaptation to the target person, we design a pose-guided feature aggregation branch. Rather than directly fusing all reference features, this branch formulates the target pose as the query and the reference poses as structural keys within a multi-layer cross-attention framework, enabling selective retrieval of garment content that is spatially compatible with the target body layout. Moreover, 2D positional encoding and segment-aware embeddings are introduced to preserve spatial relationships and distinguish information from different references. This design effectively aggregates complementary fine-grained details while maintaining spatial consistency under large pose variations and complex body configurations. Finally, the features from both branches are concatenated and injected into the Qwen-Image-Edit model Wu et al. (2025) in the form of reference features, enabling high-quality multi-reference VTON synthesis.

Meanwhile, to support this task setting, we curate Multi-Reference Garment (MRG-4K), an in-the-wild multi-reference VTON evaluation dataset. MRG-4K comprises 1,296 samples and 4,505 images, capturing diverse appearances of the same garment under varying human poses, viewpoints, focal lengths, backgrounds, and real-world scenes. Extensive experiments on our proposed benchmark and three existing benchmarks, including VITON-HD Choi et al. (2021), DressCode Morelli et al. (2022), and StreetTryOn Cui et al. (2025), demonstrate that our method outperforms existing approaches and achieves state-of-the-art performance in both single-reference and multi-reference virtual try-on settings. Qualitative and quantitative results further validate the effectiveness of the proposed framework in terms of accurate reference utilization, fine-grained garment detail representation, and robustness to pose variations.

In summary, our contributions are as follows:

- Beyond single-reference approaches, we propose a multi-reference guided VTON method, namely MR-VTON, to effectively exploit the complementary information across multiple reference images for high-fidelity try-on.
- Our MR-VTON presents a dual-branch feature fusion architecture to dynamically select the optimal reference image and effectively fuse features across multiple references, enabling the capture of garment appearance from diverse viewpoints and with fine-grained details.
- For evaluation, we first construct a multi-reference VTON benchmark, namely MRG-4K, to fill the gap for this task, where our method clearly outperforms existing state-of-the-art approaches. Furthermore, zero-shot evaluations are conducted on multiple academic benchmarks demonstrate the strong generalization of our MR-VTON.

2 Related Work

Mask-guided or Parsing-guided Virtual Try-On. Image-based virtual try-on aims to synthesize a realistic image of a target person wearing a reference garment while preserving both garment

Table 1: Comparison of existing VTON benchmarks with our MRG-4K. MRG-4K provides multiple reference images and high-resolution images. Each sample consists of multiple reference garment images, a target person image, and the corresponding ground-truth image.

Dataset	#Refs.	Scene	Resolution	Samples
VITON-HD	1	In-shop	1024×768	2032
DressCode	1	In-shop	1024×768	5400
StreetTryOn	1	In-the-wild	∼0.9K	5937
MV-VTON	2	In-shop	∼1K	1009
MRG-4K (Ours)	**2–5**	**In-the-wild**	**1K–2K**	**1296**



Figure 1: Examples of the MRG-4K benchmark.

details and person identity. Early methods Han et al. (2018); Wang et al. (2018); Han et al. (2019); Lee et al. (2022); Xie et al. (2022; 2023); Yang et al. (2024b) typically decompose this task into sequential stages of human parsing, explicit garment warping, and image composition. With the advent of diffusion models, subsequent works improve realism and garment fidelity by replacing or integrating explicit warping with generative feature alignment Zeng et al. (2024); Yang et al. (2024a). Specifically, Morelli et al. (2023); Kim et al. (2024) incorporate garment conditions into latent diffusion models, while Zhu et al. (2023); Xu et al. (2025); Choi et al. (2024) employ specialized networks or attention mechanisms to enhance garment preservation. More recently, FitDiT Jiang et al. (2024) adopts a diffusion transformer-based architecture for high-fidelity fitting, and Zhang et al. (2025); Chong et al. (2025) further advances mask-free generation in unconstrained scenes. Beyond single-garment try-on, He et al. (2025); Wang et al. (2025a); Zhang et al. (2024) explore multi-garment and multi-view settings to achieve more realistic and fine-grained try-on results.

Context-guided Visual Try-on. Recently, with the advancement of image generation models, VTON has been supported as a form of image editing task Wang et al. (2025b); Liu et al. (2025); Team et al. (2025); Seedream et al. (2025); Team et al. (2026b); OpenAI (2026). By leveraging generative priors, these methods eliminate the reliance on segmentation masks, demonstrating strong adaptability to in-the-wild scenarios while producing more realistic results Wu et al. (2025); Team et al. (2026a). However, these models struggle to preserve fine-grained garment consistency, person identity, and pose consistency. In practice, multi-reference scenarios also suffer from confusion among garment details across different items.

3 Multi-Reference Garment Dataset

For the proposed VTON task, we curate the Multi-Reference Garment dataset, comprising a training set of 86,627 samples with 431,650 images and a test set of 1,296 samples with 4,505 images, the latter of which is referred to as **MRG-4K**. The data are collected from e-commerce product detail pages and real try-on video frames captured from corresponding live streaming. We apply random garment try-on edits to real try-on images using state-of-the-art image editing models Raisinghani (2025); OpenAI (2026); Wu et al. (2025) to construct triplet-wise try-on samples consisting of a target person image, multiple reference garment images, and the corresponding try-on result, as shown in Fig. 1. During training, the edited image is used as input, while the sampled video frame serves as the ground truth. To avoid data leakage, we split the training and test dataset based on data source rather than individual images, strictly ensuring that samples from the same product or video only appear in one split. Furthermore, we perform identity-level deduplication and data filtering based on visual similarity to remove near-duplicate samples. Detailed data construction pipelines and primary dataset statistics are provided in the **Supplementary Material**. As summarized in Tab. 1, compared with existing VTON datasets Choi et al. (2021); Morelli et al. (2022); Wang et al. (2025a); Cui et al. (2025), MRG-4K offers more diverse references, more flexible conditioning and higher-quality images.

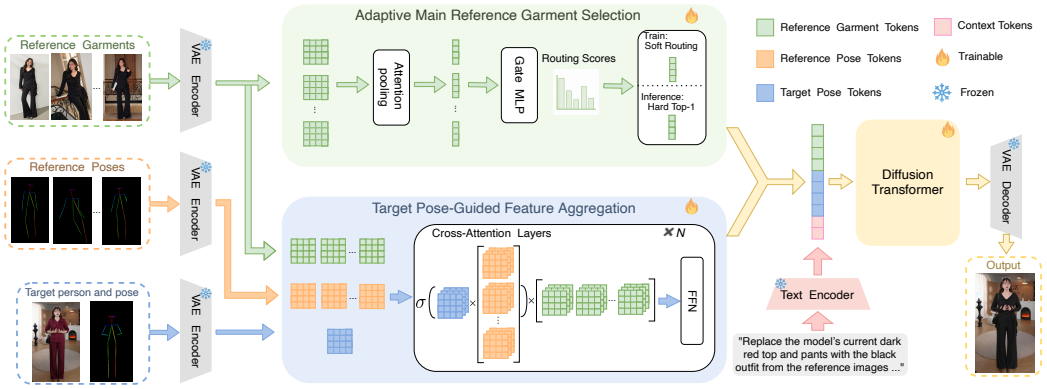


Figure 2: Overview of MR-VTON, which combines adaptive main reference selection with target pose-guided feature aggregation to preserve global garment identity and recover pose-aligned fine-grained details for high-fidelity virtual try-on.

4 METHOD

We design MR-VTON for the multi-reference garment try-on task based on Qwen-Image-Edit Wu et al. (2025). As illustrated in Fig. 2, MR-VTON coordinates high-fidelity try-on generation through a dual-branch architecture. Specifically, the target person image, multiple reference garment images, and pose representations are first encoded by a shared VAE encoder and patchified into token sequences. These tokens are then processed via two complementary branches: (i) the *Adaptive Main Reference Garment Selection* performs token attention pooling and adaptive routing to extract a globally informative reference as a garment identity anchor; (ii) concurrently, the *Target Pose-Guided Feature Aggregation* module employs target pose tokens as queries, reference pose tokens as keys, and reference content tokens as values to retrieve pose-aligned details through multi-layer cross-attention with segment embeddings and 2D RoPE. Finally, the outputs of both branches are concatenated and injected into the Diffusion Transformer (DiT) Peebles & Xie (2023) backbone for final try-on synthesis.

4.1 ADAPTIVE MAIN REFERENCE GARMENT SELECTION

The first branch aims to select the most representative reference image from multiple candidates at the global level. Since each reference latent contains not only garment information but also numerous distracting factors, such as varying model poses, body shapes, and complex background, directly comparing raw token sequences yields highly unreliable results. We therefore decompose the selection process into *intra-image token scoring* and *inter-image expert routing*.

4.1.1 TOKEN ATTENTION POOLING

For each reference image I_i, let its patchified latent tokens be

$$\mathbf{X}_i \in \mathbb{R}^{N_i \times D}, \tag{1}$$

where N_i is the number of tokens and D is the token dimension. To obtain a compact image-level representation, we apply a token attention pooling operator. Specifically, each token is first normalized and scored by a lightweight Multi-Layer Perceptron (MLP):

$$\mathbf{s}_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{X}_i)) \in \mathbb{R}^{N_i}. \tag{2}$$

The token weights are then normalized by softmax:

$$\mathbf{a}_i = \mathrm{softmax}(\mathbf{s}_i), \tag{3}$$

and the pooled image-level feature is computed as

$$\mathbf{f}_i = \sum_{n=1}^{N_i} \mathbf{a}_{i,n} \mathbf{x}_{i,n} \in \mathbb{R}^D. \tag{4}$$

This operation acts as an intra-image expert mechanism that identifies which latent tokens are most relevant for garment identity representation within each reference.

4.1.2 INTER-IMAGE GATING

Given the pooled features $\{\mathbf{f}_i\}_{i=1}^{K}$ from K references, we design an inter-image gating network to further estimate their relative importance. Specifically, a learnable MLP maps each pooled feature to a scalar score:

$$g_i = \mathrm{GateMLP}(\mathbf{f}_i), \tag{5}$$

followed by softmax normalization:

$$w_i = \frac{\exp(g_i)}{\sum_{j=1}^{K} \exp(g_j)}. \tag{6}$$

w_i denotes the routing weight of the i-th reference. To avoid biased routing at the early stage of training, we zero-initialize the last layer of the gate network, thereby enforcing an initially uniform routing distribution.

Since different reference images may have different token lengths, we align all token sequences to a shared length N^* via adaptive average pooling $\tilde{\mathbf{X}}_i \in \mathbb{R}^{N^* \times D}$. During training, we adopt soft routing to preserve differentiability:

$$\mathbf{Z}^{(1)} = \sum_{i=1}^{K} w_i \tilde{\mathbf{X}}_i. \tag{7}$$

During inference, we use hard top-1 routing:

$$i^* = \arg\max_i w_i, \qquad \mathbf{Z}^{(1)} = \tilde{\mathbf{X}}_{i^*}. \tag{8}$$

The resulting latent $\mathbf{Z}^{(1)}$ serves as a global garment identity anchor for downstream generation.

4.2 TARGET POSE-GUIDED FEATURE AGGREGATION

While the first branch captures the overall garment identity, it cannot fully exploit fine-grained complementary details distributed across different references. To address this limitation, our second branch performs target pose-guided retrieval and fusion of local appearance features from all references. The key idea is to employ the target pose as a structural query, enabling the model to retrieve pose-consistent content from multiple references.

4.2.1 POSE AND CONTENT TOKEN CONSTRUCTION

We use OpenPose Cao et al. (2019) to extract the person's pose representation. Let $\mathbf{P}^q \in \mathbb{R}^{N_q \times D}$ denote the patchified latent tokens of the target pose, and let $\{\mathbf{P}_i^r\}_{i=1}^{K}$ and $\{\mathbf{C}_i^r\}_{i=1}^{K}$ denote the pose tokens and content tokens of the K reference images, respectively. The target pose tokens are used as the query sequence, i.e., $\mathbf{Q} = \mathbf{P}^q$, while the reference pose tokens are concatenated to form the key sequence, $\mathbf{K} = \mathrm{Concat}(\mathbf{P}_1^r, \mathbf{P}_2^r, \ldots, \mathbf{P}_K^r)$, and the corresponding reference content tokens are concatenated to form the value sequence, $\mathbf{V} = \mathrm{Concat}(\mathbf{C}_1^r, \mathbf{C}_2^r, \ldots, \mathbf{C}_K^r)$. If the total token lengths of $\mathbf{K}$ and $\mathbf{V}$ are different, we align $\mathbf{V}$ to the same length as $\mathbf{K}$ via adaptive average pooling.

4.2.2 SEGMENT EMBEDDING AND 2D ROPE

To enhance structural awareness, we inject two types of auxiliary positional cues into the pose-guided branch. First, we introduce a learnable segment embedding to distinguish the target pose from different reference poses. Specifically, the target pose is assigned segment index 0, while each reference pose is assigned a distinct segment index. This allows the model to explicitly trace the source of each token after concatenation. Second, we employ 2D rotary position embedding (RoPE) Su et al. (2024) to encode the spatial layout of the patchified latent grid. Given a pose token sequence corresponding to a spatial grid of size $H \times W$, 2D RoPE generates position-dependent sinusoidal rotations along the horizontal and vertical axes. These embeddings are applied to the query and key projections in attention, enabling the model to better preserve spatial relationships when aligning target pose regions to reference pose regions.

4.2.3 MULTI-LAYER POSE-GUIDED CROSS-ATTENTION

After incorporating segment embeddings and 2D RoPE, we apply multi-layer cross-attention to retrieve pose-aligned appearance features from the references. At the l-th layer, the target pose tokens $\mathbf{Q}^{(l)}$, reference pose tokens $\mathbf{K}^{(l)}$, and reference content tokens $\mathbf{V}^{(l)}$ are first normalized and linearly projected to obtain $\hat{\mathbf{Q}}^{(l)}$, $\hat{\mathbf{K}}^{(l)}$, and $\hat{\mathbf{V}}^{(l)}$, respectively. The 2D RoPE is applied to the projected queries and keys to preserve spatial relationships. The attention output is then computed as

$$\mathbf{A}^{(l)} = \mathrm{softmax}\left(\frac{\hat{\mathbf{Q}}^{(l)}\left(\hat{\mathbf{K}}^{(l)}\right)^{\top}}{\sqrt{d}}\right)\hat{\mathbf{V}}^{(l)}, \tag{9}$$

where d denotes the dimension of each attention head. The query is updated through residual attention and a feed-forward network:

$$\tilde{\mathbf{Q}}^{(l)} = \mathbf{Q}^{(l)} + \mathrm{Proj}(\mathbf{A}^{(l)}), \qquad \mathbf{Q}^{(l+1)} = \tilde{\mathbf{Q}}^{(l)} + \mathrm{FFN}\Big(\mathrm{LN}(\tilde{\mathbf{Q}}^{(l)})\Big). \tag{10}$$

Through L layers of processing, the final output is denoted as $\mathbf{Z}^{(2)} \in \mathbb{R}^{N_q \times D}$. This branch thus establishes a pose-guided retrieval mechanism: the target pose determines *where* to attend, the multi-view reference poses determine *which regions are structurally relevant*, and the reference content tokens provide *what appearance information should be transferred*.

4.3 MULTI-BRANCH FEATURE FUSION

The two branches provide complementary information for multi-reference virtual try-on. The first branch produces a globally selected garment latent $\mathbf{Z}^{(1)}$ that captures the overall garment identity, while the second branch generates a pose-guided latent $\mathbf{Z}^{(2)}$ that aggregates local appearance details aligned with the target pose. We concatenate them to obtain the final reference condition, $\mathbf{Z} = \mathrm{Concat}(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)})$, which is injected into the Qwen-Image-Edit backbone together with the target image tokens to synthesize the final try-on result.

4.4 DISCUSSIONS WITH MV-VTON METHOD

Although both MV-VTON and our method tackle virtual try-on with multi-reference garment guidance, they differ substantially in task setting, data construction, and technical design. First, MV-VTON focuses on the *in-shop* scenario and constructs a dataset covering five fixed garment viewpoints, whereas our method targets the more challenging *in-the-wild* setting, where reference images are sourced from real-world scenarios encompassing diverse viewpoints, backgrounds and model poses. Second, architecturally, MV-VTON takes front and back garment images as input and leverages CLIP Radford et al. (2021) features along with multi-scale encoder representations to guide garment feature fusion within a U-Net architecture Rombach et al. (2022). In contrast, our method adopts a pose-guided re-weighting strategy for cross-reference feature aggregation, and injects the most informative garment feature into a DiT-based image editing model via feature concatenation.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

To evaluate the effectiveness of our method on the multi-reference try-on task, we conduct experiments and comparisons on the proposed MRG-4K benchmark. Furthermore, given that our method is also capable of handling single-reference try-on, we conduct additional zero-shot evaluations on VITON-HD Choi et al. (2021), DressCode Morelli et al. (2022), and StreetTryOn Cui et al. (2025), without any dataset-specific fine-tuning, to assess the generalization ability of our method.

Following previous works Ge et al. (2021); Lin et al. (2026); Wang et al. (2025c;a), we evaluate our method using FID Heusel et al. (2017), KID Bińkowski et al. (2018), LPIPS Zhang et al. (2018), DiffSim Song et al. (2025), and PSNR Hore & Ziou (2010). For the unpaired setting, pixel-wise reconstruction metrics are not applicable. Therefore, we report only FID, denoted as FID_u.

Table 2: Quantitative comparison on MRG-4K. ↓/↑ denotes lower/higher is better. Best and second-best results are shown in **bold** and underline. * denotes multi-reference garment images.

Category	Method	FID↓	KID↓	LPIPS↓	DiffSim↑	PSNR↑
Mask-guided	OOTDiffusion	54.87	220.14	0.2644	0.7242	15.35
	CatVTON	37.90	57.90	0.2578	0.7300	17.15
	MV-VTON	35.36	79.28	0.2471	0.7352	17.54
	FitDiT	41.33	104.21	0.2658	0.7375	16.07
Context-guided	Qwen-Image-Edit-2511	30.89	40.88	0.2913	0.7301	15.26
	FireRed-Image-Edit-1.1	24.84	26.47	0.2367	0.7855	17.37
	GPT-Image-2	24.86	21.66	0.2840	0.7714	17.15
	Nano Banana Pro	23.27	21.99	0.2472	0.7854	17.65
	Qwen-Image-Edit-2511*	38.60	77.86	0.4372	0.6452	13.55
	FireRed-Image-Edit-1.1*	56.58	230.56	0.4162	0.6386	15.72
	GPT-Image-2*	30.55	38.10	0.3736	0.7109	15.47
	Nano Banana Pro*	23.37	22.67	0.2563	0.7830	17.52
	Ours	**21.79**	**16.47**	**0.2213**	**0.8014**	**18.14**

Table 3: Quantitative comparison on academic benchmarks under zero-shot evaluation, including VITON-HD, DressCode, and StreetTryOn across different VTON scenarios.

Category	Method	VITON-HD	DressCode	Shop2Street	Model2Model	Model2Street	Street2Street	
		FID_u↓	FID_u↓	FID_u↓	FID_u↓	FID_u↓	Dress↓	Top↓
Mask-guided	OOTDiffusion	9.56	8.05	35.18	20.16	43.80	64.08	52.24
	CatVTON	21.96	8.27	48.86	23.82	51.33	50.75	45.59
	MV-VTON	18.04	16.78	–	–	–	–	–
	FitDiT	10.15	5.88	41.03	12.64	41.55	45.24	48.39
	StreetTryOn	–	–	34.05	12.16	34.45	34.16	33.04
Context-guided	Qwen-Image-Edit-2511	12.16	9.32	47.52	9.14	58.08	34.03	38.28
	FireRed-Image-Edit-1.1	16.58	7.99	49.10	9.07	56.23	31.04	36.17
	Nano Banana Pro	**9.20**	6.18	36.03	9.10	58.95	33.40	32.82
	GPT-Image-2	10.10	4.81	34.84	9.73	49.29	33.64	31.40
	Ours	9.28	**4.69**	**31.33**	**9.02**	**31.31**	**28.00**	**30.24**

We compare our method against two baseline categories: mask-guided VTON models (OOTDiffusion Xu et al. (2025), CatVTON Chong et al. (2025), MV-VTON Wang et al. (2025a), FitDiT Jiang et al. (2024), StreetTryOn Cui et al. (2025)) and context-guided image editing models (Qwen-image-edit-2511 Wu et al. (2025), FireRed-Image-Edit-1.1 Team et al. (2026a), GPT-image-2 OpenAI (2026), Nano Banana Pro Raisinghani (2025)). Since all mask-guided methods except MV-VTON support only single-reference try-on, we use the optimal reference for each MRG-4K sample. Specifically, the best reference image is selected via Gemini-2.5-Pro Comanici et al. (2025) and manual verification based on three criteria: minimal occlusion, pose similarity to the target model, and completeness of garment display. For context-guided methods, which support both single-reference and multi-reference inputs, we evaluate both settings on the MRG-4K benchmark, where results obtained with multiple reference images are marked with $*$.

5.2 Evaluation on MRG-4K

5.2.1 Quantitative and Qualitative Evaluation

We first compare with state-of-the-art methods on MRG-4K. As shown in Tab. 2, our method achieves the best performance across all evaluation metrics. The improvement stems from the complementary roles of our two branches: the main-reference selection branch provides a stable garment identity anchor, while the pose-guided aggregation branch retrieves pose-compatible details from multiple references. This design enables better garment preservation and detail recovery in challenging in-the-wild scenarios. As illustrated in Fig. 3, existing methods often suffer from distorted garment structures, blurred textures, inaccurate body alignment, and noticeable boundary artifacts, particularly under large pose variations and complex backgrounds. In contrast, our method consistently preserves fine-grained garment details in both multi-reference and single-reference settings, demonstrating stronger robustness across diverse try-on scenarios. Moreover, while maintaining high garment consistency, it also better preserves the target person's original appearance and pose, as evidenced by the more accurate preservation of the hand gesture in the second row.



Figure 3: Qualitative comparisons on MRG-4K dataset.



Figure 4: Qualitative comparisons on StreetTryOn dataset.

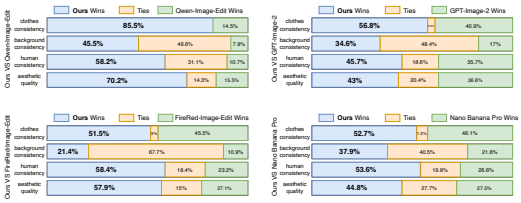


Figure 5: Human evaluation of MR-VTON vs. Qwen-Image-Edit, GPT-Image-2, FireRed-Image-Edit, and Nano Banana Pro.

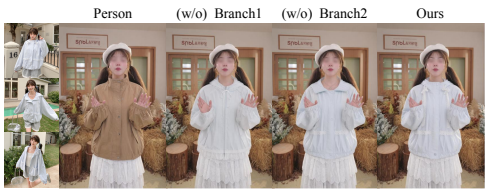


Figure 6: Visualization of the effect of the multi-branch architecture.

5.2.2 User Study

To further validate the effectiveness of our method, we conduct a blind user study with 20 participants on 100 randomly sampled samples from MRG-4K, comparing our method against each context-guided baseline individually. For each pairwise comparison, annotators assign a preference label of win, lose, or tie across four evaluation dimensions: Clothes Consistency, Background Consistency, Human & Pose Consistency and Aesthetic Quality. As shown in Fig. 5, our method achieves significant improvements over the counterparts in all dimensions, especially in clothes consistency and aesthetic quality.

5.3 Zero-shot Generalization to Other VTON benchmarks

Furthermore, we perform zero-shot evaluation on other VTON benchmarks. As shown in Tab. 3, our method achieves competitive or superior performance against both mask-guided and context-guided baselines on VITON-HD, DressCode, and StreetTryOn. Specifically, on the StreetTryOn benchmark, our method yields an FID of 31.31 in the challenging Model-to-Street setting and a Dress FID of 28.00 in the Street-to-Street setting. Notably, all these evaluations are conducted in a zero-shot manner, without any dataset-specific fine-tuning, demonstrating the strong generalization capability and robustness of our method across diverse try-on scenarios. As illustrated in Fig. 4, our method produces more realistic and visually coherent try-on results than existing baselines.

Table 4: Ablation study of multi-reference aggregation. Single-Reference* denotes the setting in which a single reference image is selected by Gemini Comanici et al. (2025).

Method	FID ↓	KID ↓	LPIPS ↓	DiffSim ↑	PSNR ↑
Single-Reference*	25.45	22.80	0.2401	0.7740	17.78
Multi-Refs-Concat	26.02	22.99	0.2530	0.7685	17.61
Ours	**21.79**	**16.47**	**0.2213**	**0.8014**	**18.14**

Table 5: Ablation study on hard and soft routing in the first branch during inference. Soft Routing denotes using soft routing during inference as well.

Method	FID ↓	KID ↓	LPIPS ↓	DiffSim ↑	PSNR ↑
w/ Soft Routing	25.70	19.62	0.2446	0.7659	17.47
w/ Hard Routing	**23.67**	**19.08**	**0.2315**	**0.7815**	**18.11**

Table 6: Ablation study of the two complementary branches in our framework.

Branch 1	Branch 2	FID ↓	KID ↓	LPIPS ↓	DiffSim ↑	PSNR ↑
×	×	30.89	40.88	0.2913	0.7301	15.26
✓	×	23.67	19.08	0.2315	0.7815	18.11
×	✓	24.14	21.33	0.2391	0.7798	18.03
✓	✓	**21.79**	**16.47**	**0.2213**	**0.8014**	**18.14**

5.4 Ablation Studies

Effect of Multi-Reference Strategy. This ablation study validates both the motivation for incorporating multiple reference images and the effectiveness of our proposed multi-reference aggregation strategy. Each model is independently trained and evaluated for each specific setting on MRG-4K. The *Single-Reference** model uses only the most informative reference garment image selected by a vision-language model Comanici et al. (2025) based on occlusion, pose similarity and garment visibility. The *Multi-Refs-Concat* model directly aggregates all reference features via concatenation. As shown in Tab. 4, our selective fusion achieves the best results across all metrics, demonstrating that it effectively preserves complementary garment details while avoiding redundant or conflicting information.

Effect of Hard Routing in Branch 1. To validate the effect of different training and inference strategies, we adopt soft routing during training and switch to hard routing at inference, after the Gate-MLP in the first branch. As shown in Tab. 5, soft routing for both training and inference introduces feature ambiguity among multiple references, leading to inferior performance. In contrast, hard routing for inference consistently outperforms soft routing across all evaluation metrics, demonstrating the benefit of selecting a single dominant reference during inference.

Effect of Multi-branch Architecture. The setting without both branches corresponds to the original Qwen-Image-Edit-2511 baseline. The first branch performs global reference selection to establish a reliable garment identity anchor, while the second branch employs pose-guided cross-attention to retrieve complementary details from multiple references. As shown in Tab. 6, while using either branch individually improves upon the baseline, combining both yields the best overall performance across all metrics. As illustrated in Fig. 6, without the first branch, cues from different references may be indiscriminately fused, resulting in mixed attributes and inconsistent garment identity. Conversely, relying solely on the first branch may leave local details (such as the collar) incomplete due to occlusions or limited viewpoints in the selected reference. Combining both branches successfully preserves global garment identity while recovering complementary local details.

6 Conclusion

We propose a multi-reference VTON task and introduce a dual-branch architecture with a primary reference selection strategy and a target pose-guided fusion mechanism, enabling richer garment detail capture and more consistent try-on results. We further construct MRG-4K as an academic benchmark for this task. Extensive experiments demonstrate leading performance across single-reference, multi-reference, in-the-shop, and in-the-wild scenarios, even surpassing closed-source models.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, code assistance and debugging, and for generating edited images during the construction of our training dataset, as described in the paper. We did not use generative AI tools to generate research conclusions or experimental results. All AI-assisted text and code were manually reviewed by the authors, and the assisted code was tested for correctness. All generated images used in the dataset were further subject to automatic and manual quality filtering. The experimental results, analyses, and claims were independently verified by the authors. We take full responsibility for the final content of this work, including any text, code, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available e-commerce product images and real try-on video frames to construct a virtual try-on dataset. We apply identity anonymization and data filtering to reduce privacy risks and do not intentionally retain personally identifiable information. The dataset is intended solely for research purposes. We acknowledge that the collected data may inherit biases from the original sources, including imbalances in garment categories, body appearance, pose, and scene distribution. We apply category balancing and quality filtering to mitigate these issues, although such biases may not be fully eliminated. Virtual try-on models may potentially be misused for deceptive or unauthorized image manipulation. Our work is intended for benign research applications, and we discourage any use that violates privacy, consent, or applicable laws. The authors take responsibility for the ethical conduct and presentation of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed MR-VTON architecture, including the adaptive main-reference selection branch and the target pose-guided feature aggregation branch, in the main paper. We also document the construction, filtering, split strategy, and statistics of the proposed MRG-4K benchmark to facilitate reproducible evaluation. Upon publication, we plan to release the source code, model configurations, and dataset-related resources necessary to reproduce the reported results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2019.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pp. 14131–14140, 2021.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, pp. 206–235. Springer, 2024.

Zheng Chong, Xiao Dong, Haoxiang Li, Wenqing Zhang, Hanqing Zhao, Dongmei Jiang, Xiaodan Liang, et al. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, volume 2025, pp. 66586–66601, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *WACV*, pp. 1414–1423, 2025.

Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pp. 8485–8493, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, pp. 7543–7552, 2018.

Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pp. 10471–10480, 2019.

Zijian He, Yuwei Ning, Yipeng Qin, Guangrun Wang, Sibei Yang, Liang Lin, and Guanbin Li. Vton 360: High-fidelity virtual try-on from any viewing direction. In *CVPR*, pp. 26388–26398, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, pp. 2366–2369. IEEE, 2010.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*, 2024.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, pp. 8176–8185, 2024.

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pp. 204–219. Springer, 2022.

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ToG*, 40(4):1–10, 2021.

Zhenyi Lin, Xiaofan Ming, Qilong Wang, Dongwei Ren, Wangmeng Zuo, and Qinghua Hu. Roomeditor: High-fidelity furniture synthesis with parameter-sharing u-net. *NeurIPS*, 38:32789–32816, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, pp. 5084–5093, 2020.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, pp. 2231–2235, 2022.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACMMM*, pp. 8580–8589, 2023.

OpenAI. Gpt-image-2 model card. https://platform.openai.com/docs/models/gpt-image-2, 2026. Accessed: 2026-07-14.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Naina Raisinghani. Introducing nano banana pro. *Google Blog*, 20, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16904–16915, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026a.

Tencent HY Team et al. Hy-wu (part i): An extensible functional neural memory framework and an instantiation in text-guided image editing. *arXiv preprint arXiv:2603.07236*, 2026b.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pp. 589–604, 2018.

Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. MV-VTON: Multi-view virtual try-on with diffusion models. In *AAAI*, pp. 7682–7690, 2025a.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025b.

Qilong Wang, Xiaofan Ming, Zhenyi Lin, Jinwen Li, Dongwei Ren, Wangmeng Zuo, and Qinghua Hu. Roomeditor++: A parameter-sharing diffusion architecture for high-fidelity furniture synthesis. *arXiv preprint arXiv:2512.17573*, 2025c.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *NeurIPS*, 34: 2598–2610, 2021.

Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, Xin Dong, Feida Zhu, and Xiaodan Liang. Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. *arXiv preprint arXiv:2207.13475*, 2022.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, pp. 23550–23559, 2023.

Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI*, pp. 8996–9004, 2025.

Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *CVPR*, pp. 7017–7026, 2024a.

Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D^4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *ECCV*, pp. 36–52. Springer, 2024b.

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *CVPR*, pp. 8372–8382, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and An-An Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *CVPR*, pp. 26399–26408, 2025.

Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *CVPR*, pp. 4606–4615, 2023.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

Our method is built upon the Qwen-Image-Edit-2511 version Wu et al. (2025). We adopt LoRA Hu et al. (2022) fine-tuning with rank 64 and train the model using the AdamW Loshchilov & Hutter (2019) optimizer with a learning rate of 1×10^{-4}. The batch size is set to 8, and the model is trained for 20,000 steps. All experiments are conducted on 8 NVIDIA H20 GPUs.

A.2 MULTI-REFERENCE GARMENT DATASET

A.2.1 DATA CONSTRUCTION.

The construction of MRG consists of three stages. **Stage 1: Data Collection.** We collect garment reference images from e-commerce platforms and extract person images from livestream videos through frame sampling. To improve diversity, the collected data cover a wide range of garment categories, human appearances, poses, viewpoints, and indoor scenes. In total, the initial data pool contains approximately 1.5M images. **Stage 2: Data Filtering, Refinement, and Anonymization.** The raw data are filtered using a combination of automated processing and human verification. A VLM-based Bai et al. (2025) pipeline removes low-quality samples, including images with severe blur, occlusion, duplicated content, incomplete bodies, or unclear garment details. Each retained image is further annotated with semantic attributes, which are first inferred automatically and then manually verified. To protect privacy, we blur the eye regions of all livestream subjects and anonymize visible livestream information, such as usernames, platform identifiers, and on-screen interface elements. **Stage 3: Triplet Construction and Quality Control.** The collected data do not naturally provide paired images before and after garment editing. We therefore employ large-scale image editing models, including GPT-Image-2 OpenAI (2026) and Qwen-Image-Edit Wu et al. (2025), to synthesize garment-changing results and construct triplets consisting of an input person image, an edited try-on image, and multiple garment references. During generation, the models are instructed to preserve the person identity, pose, body structure, and background while modifying only the garment appearance. Finally, we apply another round of VLM-based filtering and human inspection to remove samples with identity changes, pose deviations, garment inconsistencies, or obvious editing artifacts, resulting in the final MRG dataset.

Table 7: Reference image number distribution ratio in the training set

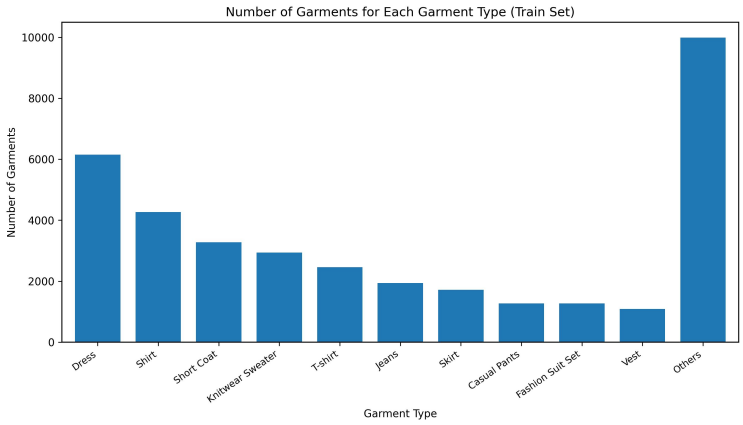
Product Images Count	Sample Count	Ratio
single reference image	12103	13.97%
two reference images	20627	23.81%
three referenece images	22276	25.71%
four reference images	19894	22.97%
five reference images	11727	13.54%

Table 8: Reference image number distribution ratio in the test set

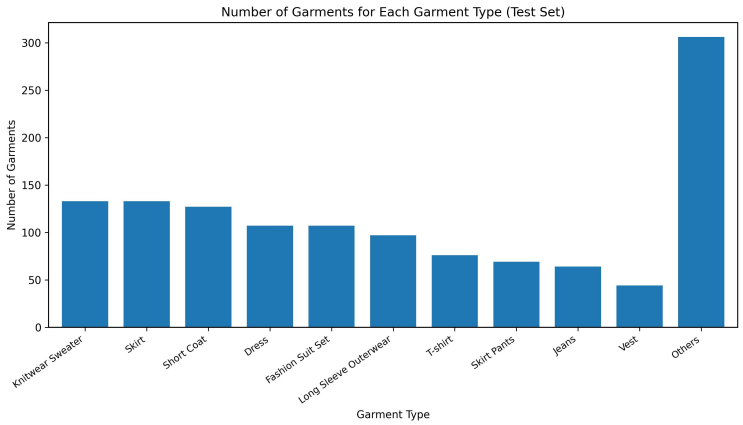
Product Images Count	Sample Count	Ratio
two reference images	263	20.29%
three referenece images	354	27.31%
four reference images	385	29.71%
five reference images	294	22.69%

A.2.2 DATA OVERVIEW.

Our dataset comprises a training set of 86,627 samples containing 431,650 images and a test set of 1,296 samples containing 4,505 images. The distributions of the number of reference images



(a) Training set.



(b) Test set.

Figure 7: The number of garments in each garment type for the training and test sets.

in the training and test sets are reported in Tab. 7 and Tab. 8, respectively. In addition to multi-reference samples, we retain a substantial number of single-reference samples during training. This design prevents the model from becoming overly dependent on multi-reference inputs, preserves its capability under the conventional single-reference setting, and improves its robustness to varying numbers of reference images. As illustrated in Fig. 7, the dataset covers a wide range of garment categories and styles. Such diversity exposes the model to more realistic and challenging scenarios, thereby improving its generalization ability and enabling it to effectively integrate complementary garment information from heterogeneous references.

A.3 Reference Image Selection Prompt

To construct the *Single-Reference** setting, we employ Gemini-2.5-Pro Comanici et al. (2025) to select the most informative reference garment image from multiple candidates. Specifically, the model prioritizes garment visibility, pose similarity to the target person, and the completeness of garment details. The complete prompt is shown below.

Listing 1: Prompt used for selecting the best reference garment image.

```
## ROLE & GOAL
You are an expert fashion image analyst. Your task is to
select the single best reference garment image for a
virtual try-on task.

## INPUTS
- Image 1: The target person image. The clothes worn by
  this person will be replaced.
```

```
- Images 2 to {ref_end}: Reference garment images
  ({num_refs} images in total), each showing the target
  garment from a different viewpoint or wearing pose.

## SELECTION CRITERIA
Select the best reference image according to the following
criteria, listed in descending order of priority:

1. Minimal occlusion:
   The garment should be fully visible with minimal
   occlusion. Avoid images in which the garment is blocked
   by hands or accessories, or is partially cropped.

2. Pose similarity:
   The pose and viewpoint of the person or mannequin in the
   reference image should be as similar as possible to those
   of the target person in Image 1.

3. Garment completeness:
   The image should clearly display the garment's overall
   shape, structure, texture, color, and fine-grained details.

## OUTPUT FORMAT
Output only a single integer indicating the index of the
selected reference image. The index is 1-based and counts
only the reference images.

For example, if the third reference image is selected, output:
3

Do not output any explanation, reasoning, punctuation, or
additional text. Output only the integer.
```

A.4 More Qualitative Results

In this section, we present additional qualitative results across multiple benchmarks. Fig. 8, Fig. 9, Fig. 10, and Fig. 11 show the results on the MRG-4K, VITON-HD, DressCode, and StreetTryOn datasets, respectively. As observed, our method consistently preserves fine-grained garment appearance and identity while maintaining the original pose and body structure of the target person. This capability remains robust under both single-reference and multi-reference settings, as well as across in-shop and in-the-wild scenarios. These results demonstrate the strong generalization and robustness of our method across diverse input configurations, human poses, garment appearances, and scene conditions.



Figure 8: Additional visualization results on the MRG-4K dataset.



Figure 9: Additional visualization results on the VITON-HD dataset.



Figure 10: Additional visualization results on the DressCode dataset.



Figure 11: Additional visualization results on the StreetTryOn dataset.