RefTon: Reference person shot assist virtual Try-on

Liuzhuozheng Li[1,2*] Yue Gong[2*] Shanyuan Liu[2*] Zanyi Wang[3] Dengyang Jiang[4]
Leibucha Wu[2] Bo Cheng[2] Yuhang Ma[2] Dawei Leng[2†] Yuhui Yin[2]

[1] The University of Tokyo [2] 360 AI Research [3] University of California, San Diego
[4] The Hong Kong University of Science and Technology

Abstract

We introduce RefTon, a flux-based person-to-person virtual try-on framework that enhances garment realism through unpaired visual references. Unlike conventional approaches that rely on complex auxiliary inputs such as body parsing and warped mask or require finely designed extract branches to process various input conditions, RefTon streamlines the process by directly generating try-on results from a source image and a target garment, without the need for structural guidance or auxiliary components to handle diverse inputs. Moreover, inspired by human clothing selection behavior, RefTon leverages additional reference images (the target garment worn on different individuals) to provide powerful guidance for refining texture alignment and maintaining the garment details. To enable this capability, we built a dataset containing unpaired reference images for training. Extensive experiments on public benchmarks demonstrate that RefTon achieves competitive or superior performance compared to state-of-the-art methods, while maintaining a simple and efficient person-to-person design. Code and dataset is available at: `https://github.com/360CVGroup/RefTon`.

1. Introduction

The **Virtual Try-On (ViTON)** model aims to generate photo-realistic images of a person wearing target clothing, a tool crucial for applications in online retail and personalized fashion systems. ViTON methods are broadly categorized into Generative Adversarial Networks (GANs) [20] and Diffusion Models [27, 49]. Early ViTON research relied on GANs [8, 24, 57], which typically employed warping modules to deform clothing for alignment with the human body, followed by fusion to achieve visual harmony. However, GAN-based approaches frequently generate unrealistic artifacts, particularly when dealing with complex clothing textures or challenging human poses. Recently, methods based on latent diffusion models (LDMs) [5, 60] have gained traction, significantly enhancing clothing warping and addressing structural arrangement and texture preservation during denoising [9, 30, 61, 68]. Despite these advances, current diffusion-based ViTON technologies still generally rely on extensive auxiliary conditions, such as clothing region masks, garment masks, human poses, key points, or multi-modal inputs like text prompts [10, 11, 31, 60, 62].

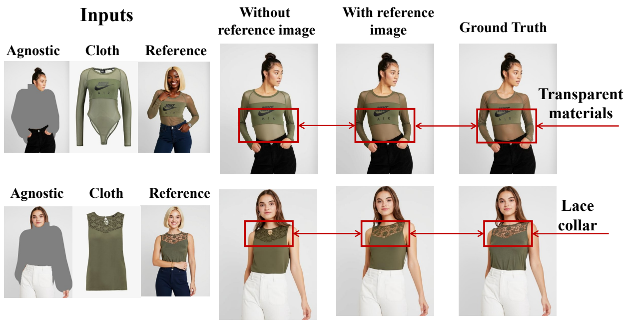


Figure 1. **Effect of reference images in virtual try-on.** From left to right: inputs (agnostic, cloth, reference), results without reference images, results with reference images, and ground truth. Using reference images in both training and inference improves visual fidelity and preserves fine details (e.g., transparent materials and lace collars). Zoom in for a better comparison.

Despite the remarkable progress of prior virtual try-on approaches, they are still constrained by two critical limitations that hinder the authenticity of the try-on results and broader applicability: **First**, these approaches rely on multiple external models and internal modules, such as pose estimators [3, 4, 22, 55, 58], human parse models [15, 37], segmentation models [34, 48], to process different conditions, which compromises the practicality. To process diverse inputs, additional modules are integrated into the model, which consequently increases the overall framework complexity. Moreover, in practical applications, the quality of conditional inputs—such as the cloth mask—has a substantial impact on the quality of the final try-on results; **Second**,

---

*Equal contribution.
†Corresponding author

many aspects of clothing, such as style, texture, and detailed design, cannot be fully perceived from the garment image alone; instead, it is more important to consider the overall appearance when a model wears the garment. Therefore, in real-world try-on scenarios, such as online shopping, users are typically more interested in model images rather than the garment itself. They tend to see how the target garment looks when worn on a real person, rather than relying solely on the isolated garment image as a reference. For example, as illustrated in Fig. 1, at the garment in the first row, it is difficult to tell whether it is a green translucent fabric or a light green opaque one, whereas the reference image clearly reveals its green translucent material. We cannot accurately identify transparent materials or intricate designs, such as lace collars, solely from cloth images. In contrast, the reference images of human models wearing the garments reveal such details. However, existing virtual try-on methods do not support such references due to the lack of corresponding reference data in public datasets [8, 17, 26, 43, 44, 51, 66].

Based on the above observation, we propose **RefTon**, a flux-based person-to-person virtual try-on framework that achieves strong performance *without relying on any external models or auxiliary components*, while being further enhanced by *additional reference images* that offer more accurate and context-aware guidance for the try-on model. First, to ensure the best performance, we adopt the powerful image editing model *Flux-kontext* as our base and apply adaptation on its position index for RoPE [54] to make it suitable for multi-condition/resolution virtual try-on inputs. Similar to [10], RefTon eliminates the need for auxiliary inputs such as segmentation masks using a two-stage training strategy, allowing for simple inference with only a source image and the target garment as inputs. Second, we introduce the use of images of a different person clothed in the target garment as the visual references, like the model images in online shopping, which better reflect users' real-world behavior when choosing clothes and enable the preservation of fine garment details that existing methods cannot achieve. To achieve these objectives, we propose a reference data generation pipeline, by which we construct a dataset with supplementary reference images and use unpaired person-cloth samples to train our own model to utilize reference images as additional visual guidance. These improvements empower RefTon to achieve both a simplified model structure and a streamlined inference process, while simultaneously delivering superior generation results.

In summary, the main contributions are as follows:

- We propose incorporating **additional reference images** into the virtual try-on pipeline. This significantly enhances the authenticity and visual quality of the try-on results, achieving **State-of-the-Art** performance in preserving fine garment design details.
- We designed a **reference data generation framework** to create the necessary reference images for both the clothing and target ground truth samples. Based on this pipeline, we built the VFR dataset upon existing benchmarks (e.g., VITON-HD, DressCode, ViViD), providing a robust new resource to improve the practicality and evaluation of virtual try-on models.
- We present an adaptation of the *Flux-Kontext* I2I model with a modified *Rescaled Position Indexing* mechanism to support **flexible multi-conditional and multi-resolution inputs**, along with a two-stage training strategy for virtual try-on. Our framework enables integration of varying numbers and types of reference images and effectively supports mask-based and person-to-person try-on within a single model. It achieves **state-of-the-art** performance and demonstrates strong generalization to in-the-wild person–clothing scenarios.

2. Related Works

2.1. Generative Model via Flow Matching

Generative modeling has rapidly progressed with diffusion models (DMs) [52], score-based generative models (SGMs) [53], and flow-based methods. Recent work has explored controllable generation, layout-conditioned synthesis, unified planning-generation frameworks [7, 25], as well as diffusion transformer variants for multilingual generation, efficient sampling, and fine-grained control [19, 39, 40, 42]. Flow-based methods [13, 14, 32] have been further advanced to address the inefficiency of continuous normalizing flows (CNFs), which require costly backpropagation through ODE solvers during training [6]. Flow Matching (FM) [38] mitigates this limitation by learning a time-dependent vector field that deterministically transports a simple before the data distribution, using a simulation-free objective that avoids numerical integration during training. By directly parameterizing the probability flow, FM achieves competitive or superior sample quality compared to diffusion models with significantly fewer sampling steps. Our method builds upon the *Flux-Kontext* architecture, where input images are encoded into latent representations, flattened into sequences, and concatenated with Gaussian noise ϵ. It is also closely related to recent FLUX-based appearance editing and transfer approaches [67].

2.2. Diffusion-based Virtual Try-on

Diffusion models [27, 49] have enabled significant progress in garment or makeup transfer [8, 30, 67, 68]. Leveraging the flexibility of Stable Diffusion [30, 45], prior works exploit text guidance and inpainting for garment synthesis. Extensions such as DiffusionCLIP [29] introduce semantic control via CLIP, while methods like DCI-VTON [21] and IDM-VTON [8] adopt two-stage pipelines to align and fuse garments, improving structural consistency.

Recent approaches, including CatVTON [10], OmniTry [18], Any2AnyTryon [23], and OmniVTON [62], explore person-to-person try-on without explicit masks. However, these methods typically rely on additional conditions (e.g., pose) or fail to unify mask-based and mask-free settings. Moreover, P2P pipelines such as TryOffDiff [56] and ViTON-GUN [64] adopt a "try-off–then–try-on" strategy, which introduces error accumulation and loss of garment details. In contrast, our method directly leverages the reference image, avoiding the try-off stage and better preserving garment structure and material fidelity. In summary, existing diffusion-based virtual try-on methods either rely on heavy auxiliary annotations or lack support for clothed reference images. To address these limitations, we propose RefTON, which adapts the virtual try-on task to the Flux-Kontext framework, enabling end-to-end, reference-guided generation while supporting both mask-based inpainting and mask-free editing.

3. Method

3.1. Preliminary

RefTon is built upon DiT [47], a scalable Transformer architecture for diffusion-based generation. Images are encoded into a latent space via an autoencoder [33] and then patched into tokens [16]. The diffusion process [27] operates on these tokens, with the Transformer consuming noisy tokens and predicting their denoised results.

We consider the problem of generating images under the condition $\boldsymbol{y}$, which may represent garment images, semantic maps, human pose, or other modality-specific control signals. Let $\boldsymbol{x}$ denote the latent image representation obtained from a VAE encoder. The goal of *Flux.1* [35, 36] is to approximate the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{y})$ by learning a time-dependent velocity field $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}, t)$ that transports a sample from a simple prior $\mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I})$ at $t = 0$ to the data distribution $p_{\text{data}}(\boldsymbol{x}|\boldsymbol{y})$ at $t = 1$. The dynamics of the conditional probability density $p(\boldsymbol{x}|\boldsymbol{y}, t)$ over time t are governed by the continuity equation:

$$\begin{aligned} \frac{\partial}{\partial t} p(\boldsymbol{x}|\boldsymbol{y}, t) &= -\nabla_{\boldsymbol{x}} \cdot \big(\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}, t) \cdot p(\boldsymbol{x}|\boldsymbol{y}, t)\big), \\ \boldsymbol{x}_{t=0} &\sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}),\ \boldsymbol{x}_{t=1} \sim p_{\text{data}}. \end{aligned} \tag{1}$$

To estimate $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}, t)$, we train a diffusion-transformer backbone to approximate the neural velocity field $\boldsymbol{v}_{\boldsymbol{\theta}}$ using the conditional flow matching objective [38, 41]:

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\theta}} &= \mathbb{E}_{t, \boldsymbol{x}_i, \boldsymbol{\epsilon}, \boldsymbol{y}_i} \Big[\big\| \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}_i, t) - (\boldsymbol{x}_i - \boldsymbol{\epsilon}) \big\|_2^2 \Big], \\ \boldsymbol{x} &= (1 - t)\, \boldsymbol{x}_i + t\, \boldsymbol{\epsilon}, \end{aligned} \tag{2}$$

where $t \sim \mathcal{U}(0, 1)$, $\boldsymbol{x}_i \sim \mathcal{X}_{\text{train}}$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. This training objective encourages the model to learn a velocity field $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, t)$ that consistently guides the noisy samples toward the data distribution conditioned on $\boldsymbol{y}$, following the probability flow ODE starting from the Gaussian prior:

$$d\boldsymbol{x} = \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}, t)dt, \tag{3}$$

enabling controllable image synthesis at inference time.

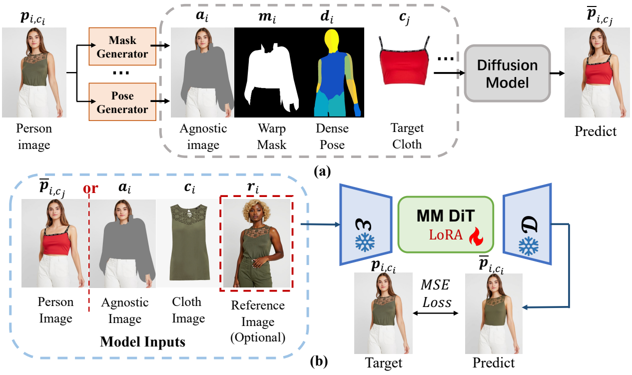


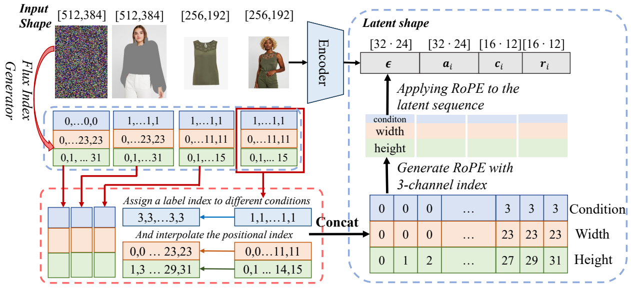
Figure 2. Overview of our two-stage training strategy. (a) Stage 1 follows mask-based try-on paradigms to generate person images wearing random garments from masked person inputs, providing training data for Stage 2. (b) Stage 2 uses these synthesized person images, along with the target garment and optional reference images, to train a person-to-person virtual try-on model that directly fits the target garment onto the person.

In the virtual try-on setting, let $\boldsymbol{x}_i$ denote the image of a person wearing the target cloth, and let $\boldsymbol{y}_i$ represent a collection of conditional inputs, including the cloth-agnostic image $\boldsymbol{a}_i$, the target cloth $\boldsymbol{c}_i$, the visual references $\boldsymbol{r}_i$, and others. Formally, we write $\boldsymbol{y}_i = [\boldsymbol{a}_i, \boldsymbol{c}_i, \dots]$. The objective is to progressively transform a Gaussian noise sample $\boldsymbol{\epsilon}$ into the target image $\boldsymbol{x}_i$ guided by conditions $\boldsymbol{y}_i$.

3.2. Person To Person Virtual Try-on Model with Two Stage Training

Our goal is to dress the person directly with the target garment, without relying on auxiliary conditions such as DensePose [22] or segmentation masks. To achieve this, we train a diffusion model on clothing-person pairs $(\boldsymbol{c}_i, \bar{\boldsymbol{p}}_{i,\boldsymbol{c}_j})$ to generate the target image $\boldsymbol{p}_{i,\boldsymbol{c}_i}$, where the person wears the target garment, as shown in Fig. 2(b). This setup requires unpaired triplets $[\bar{\boldsymbol{p}}_{i,\boldsymbol{c}_j}, \boldsymbol{c}_i, \boldsymbol{p}_{i,\boldsymbol{c}_i}]$, where $\boldsymbol{c}_j \neq \boldsymbol{c}_i$. However, existing open-source benchmarks typically provide only paired data $[\boldsymbol{c}_i, \boldsymbol{p}_{i,\boldsymbol{c}_i}]$. Following the two-stage training strategy of CATVTON [10], we adopt a similar pipeline for our try-on model and further exploit richer conditions, including agnostic masks and DensePose, to improve unpaired image generation. Specifically, in the first stage of RefTon training, we synthesize unpaired person images $\bar{\boldsymbol{p}}_{i,\boldsymbol{c}_j}$ using a mask-based try-on model.

As illustrated in Fig. 2(a), we train a virtual try-on model using agnostic person images $\boldsymbol{a}_i$, clothing images $\boldsymbol{c}_i$, dense-



(a) (b)

Figure 3. **Rescaled three-channel position index.** (a) The first channel encodes conditional inputs, while the second and third capture spatial positions for varying resolutions. (b) The concatenated indices are used to generate $RoPE$, which is applied to the latent sequence in the attention mechanism.

pose maps $\boldsymbol{d}_i$, warp masks $\boldsymbol{m}_i$ as inputs. During inference, a random unpaired garment $\boldsymbol{c}_j$ is selected to generate the corresponding synthesized person image $\bar{\boldsymbol{p}}_{i,c_j}$. To ensure the quality of the synthesized person image $\bar{\boldsymbol{p}}_{i,c_j}$, the agnostic–cloth pairs $[\boldsymbol{a}_i, \boldsymbol{c}_i]$ used for training must belong to the same garment category (e.g., if $\boldsymbol{a}_i$ is from the "dresses" subset, the selected garment $\boldsymbol{c}_j$ should also come from "dresses" rather than "upper body" or "lower body"). Otherwise, the generated person image may appear unrealistic due to mismatches between the clothing mask region and the target garment (e.g., fitting a skirt onto the upper body or a shirt onto the lower body).

After obtaining the unpaired person images, we train the person-to-person model with either the agnostic image $\boldsymbol{a}_i$ or the generated person image $\bar{\boldsymbol{p}}_{i,c_j}$ from the first stage, sampled with equal 50% probability. Instead of training a try-on model from scratch, we freeze the encoder and decoder of *Flux-kontext* [36] and fine-tune only the transformer blocks using Low-Rank Adaptation [28]. The model is optimized with the flow-matching objective in equation. 2. In addition, to help the model capture garment appearance from another person, we provide an extra reference image $\boldsymbol{r}_i$ with a probability of 25%, as illustrated in Fig. 2(b).

3.3. Multi-input Training and Adaptation

The latent embeddings are concatenated into the sequence $[\boldsymbol{\epsilon}_i, \boldsymbol{a}_i, \boldsymbol{c}_i]$ after image encoding in vanilla *Flux-kontext*, and an index generator produces a position index of shape $[L, 3]$, which is further converted into Rotary Position Embeddings (RoPE) for self-attention. Its first channel is a binary mask that distinguishes Gaussian noise $\boldsymbol{\epsilon}_i$ from conditional image/text inputs, while the second and third channels encode horizontal and vertical coordinates, respectively (Fig. 3). However, this binary design is insufficient for our setting, where the model must handle multiple heterogeneous image conditions. Therefore, we extend the first channel from a binary flag to discrete condition labels, allowing the transformer to distinguish different input types such as person, garment, and reference images.

Specifically, our model follows the *Flux-kontext* to encode each image condition independently into latent patches, and concatenates to form a sequence such as $[\boldsymbol{a}_i, \boldsymbol{c}_i, \boldsymbol{r}_i, \ldots]$, which is then concatenated with the noise latent $\boldsymbol{\epsilon}_i$ along the sequence dimension. For each condition, we generate an individual three-channel position index: the first channel indicates the condition identity, and the second and third channels store integer spatial coordinates rescaled by the resolution ratio between the target image and the corresponding condition image, which preserves spatial alignment across different resolutions. This ***Rescaled Position Index*** design enables flexible integration of multi-type and multi-resolution image conditions within a unified DiT framework. Similar indexing strategies have been explored in prior work [23], where positional indices are constructed on a concatenated image canvas, placing different conditions in a shared spatial layout. In contrast, our method assigns positional indices to each condition independently rather than generating them from a pixel-space concatenated canvas, resulting in a more flexible formulation for multi-resolution inputs. The modified DiT position indexing scheme is illustrated in Fig. 3.

3.4. Virtual-Tryon Generation with Extra Visual Reference

For both humans and generative virtual try-on models, garment images and conditions extracted by external models alone are insufficient to capture the realistic visual effect of wearing a target garment, as illustrated in Fig. 1. A garment's style, texture, and fine design details are more faithfully presented when worn by another person than when shown in isolation. Therefore, we introduce reference person images $\boldsymbol{r}_i$, in which another person wears the target garment, to provide more intuitive visual guidance during virtual try-on generation. To obtain such references, we construct pairs of the form $[\boldsymbol{c}_i, \boldsymbol{r}_i]$, representing "different persons wearing the target garment." Since these reference images $\boldsymbol{r}_i$ are unavailable in existing open-source virtual try-on datasets, we develop a reference data generation pipeline to synthesize them. This pipeline augments current datasets with supplementary references and enables model training with additional visual guidance.

Existing open-source datasets like VITON-HD [8], DressCode [44], and IGPairs [51], lack unpaired reference images $\boldsymbol{r}_i$ showing *different persons wearing the target garment*. To overcome this limitation, we employ editing models to synthesize such reference images. To serve as accurate and informative visual guidance, the generated reference data $\boldsymbol{r}_i$ should satisfy the following requirements:

- **Preserve the target garment faithfully.** The target gar-

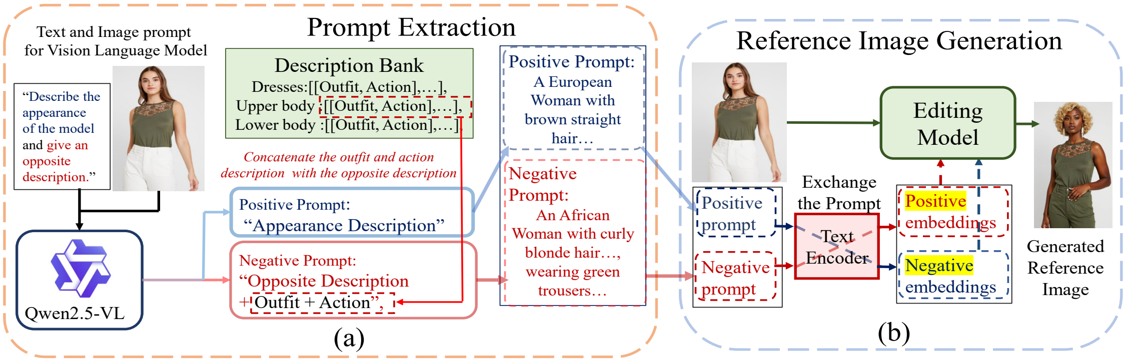


Figure 4. Reference image generation pipeline. Given a target image of a European woman wearing the target upper-body garment, *Qwen2.5-VL* [1] produces an appearance description (e.g., "a European woman with brown straight hair") and an opposite description (e.g., "an African woman with curly blonde hair"). In (a), the opposite description is combined with the action and non-target clothing attributes to form the positive prompt, while the appearance description is used as the negative prompt. In (b), the image and prompts are fed into *Flux-kontext* to generate reference images of different individuals wearing the target garment.

ment's color, texture, and design must remain unchanged to ensure an accurate reference.

- **Introduce diversity in the person's appearance.** The person wearing the target garment in the reference image should be different from the target person wearing the same garment. Otherwise, the model may learn shortcuts and overfit to the target image. This diversity can be achieved by altering hairstyle, hair color, skin tone, body pose, or facial expression.
- **Vary the non-target garments to provide outfit diversity.** While the target garment remains unchanged, other garments should be modified. For example, if the target garment is an upper-body item, the reference image should retain the same upper-body garment while altering the lower-body clothing, shoes, or accessories.

Since the reference image is edited from the target image $\boldsymbol{p}_{i,\boldsymbol{c}_i}$ in existing datasets, the first requirement ensures faithful garment preservation, while the second and third promote diversity in non-target regions to prevent overfitting and prevent the model from taking the shortcut of directly copying the target image in the generated try-on results.

As illustrated in the pipeline, we use *Flux-kontext* [36] to synthesize reference-person images. Given a target image and carefully designed text prompts, *Flux-kontext* generates images that faithfully preserve the target garment's color, texture, and design while varying the wearer's appearance and non-target clothing. Specifically, we first feed the target image and a textual instruction into *Qwen2.5-VL* [1] to obtain a detailed description of the person's appearance, together with an opposite description. We then collect a set of non-target garments and action descriptions to introduce diversity in clothing and pose; examples of these garments and action items are provided in Appendix A. Finally, we concatenate the opposite appearance description with the action and non-target garment descriptions to form the *positive prompt*, while the original appearance description is used as the *negative prompt*. These prompts are then fed into *Flux-kontext* to synthesize the reference images.

We supplement existing virtual try-on benchmarks including VITON-HD [24], DressCode [44], ViViD [17], FashionTryOn [66] and IGPairs [51] by generating corresponding reference pairs for each target garment image $\boldsymbol{c}_i$, forming data pairs $[\boldsymbol{c}_i, \boldsymbol{r}_i]$ that are used for both model training and evaluation.Some open-source datasets, such as FashionTryOn [66] and IGPairs [51], contain numerous duplicated or low-quality samples. We compare the CLIP features of images to filter out redundant samples and employ *Qwen2.5-VL* to identify distorted or unclear images, as well as images where the person faces away from the camera, ensuring overall data quality before generating the final reference set via our data generation pipeline. Finally, we enrich open-source virtual try-on datasets with additional visual references $\boldsymbol{r}_i$ and person images $\bar{\boldsymbol{p}}_{i,\boldsymbol{c}_j}$, and combine them to form our own dataset, named **Virtual Fitting with Reference (VFR)**, for training our RefTon model. The detailed data collection, filtering procedures, and visualization of the samples are provided in Appendix A.

4. Experiments

We mainly evaluate our method on two public benchmarks, DressCode [44] and VITON-HD [8], both containing images with a resolution of 1024 × 768. The VITON-HD comprises 13,670 upper-body image pairs of women, split into 11,647 pairs for training and 2,032 pairs for testing. The DressCode includes three subsets—upper body, lower body, and dresses—with 48,392 training and 5,400 testing pairs. Since DressCode does not provide wrapped cloth masks or agnostic images $\boldsymbol{a}_i$, we generate them using the mask generation tool from CatVTON [10]. We further conducted experiments on the StreetTryOn [12] and included comparisons with the methods mentioned to demonstrate the model's P2P and in-the-wild capabilities. In the P2P evaluation, we did not use the cloth-only image as input and only used the person image as a reference.

During training, we fine-tune our model on the *flux-kontext* backbone using the Low-Rank Adaptation (LoRA) technique, with a rank of 64 and $\alpha = 128$, optimized by *AdamW*. For quantitative evaluation, all generated images are resized to 512×384 to ensure a fair comparison with previous methods. In a single-dataset experiment, the model is trained independently for 20,000 steps on VITON-HD and 48,000 on DressCode with a batch size of 128, using 8 NVIDIA H100 GPUs. To further enhance generalization and robustness to in-the-wild inputs, we train an additional model on the mixed VFR dataset introduced in Sec. 3.4. We report cross-dataset evaluation results of our RefTon model at a resolution of 1024×768 on both VITON-HD and DressCode, and demonstrate its in-the-wild performance using images collected by ourselves.

4.1. Quantative result

We provide the numerical results of our model on the VITON-HD and DressCode datasets, distinguishing between paired and unpaired try-on settings. For paired try-on settings with ground truth in test datasets, we utilize four metrics to assess the similarity between the synthesized images and their corresponding authentic images: the Structural Similarity Index *(SSIM)* [46], Learned Perceptual Image Patch Similarity *(LPIPS)* [65], Fréchet Inception Distance *(FID)*[50], and Kernel Inception Distance *(KID)*[2]. For unpaired settings, where we measure the distributional similarity between the synthesized and real samples, we specifically rely on the *Fréchet Inception Distance (FID)* and *Kernel Inception Distance (KID)*. As shown in Table 1.

Our method (**RefTon**) consistently performs better than prior baselines, demonstrating higher try-on fidelity and strong alignment with the target person's pose. With the addition of reference images ("**+R**"), the quality and detail consistency of the try-on results are further improved compared with the results without reference images, establishing new state-of-the-art results across multiple metrics. Notably, even in the mask-free setting—without agnostic masks or auxiliary inputs—our method maintains garment style correctness and pose consistency, while reaching accuracy on par with or superior to baseline methods, highlighting its robustness and practicality.

Table 1 summarizes the quantitative evaluation on the DressCode dataset. Our method (RefTon) outperforms all baselines, delivering higher try-on quality and achieving strong consistency with the target person's pose and body structure. Integrating reference images ("+R") further enhances the results, establishing a new state-of-the-art. Importantly, even in the mask-free setting—without agnostic masks or additional inputs—our method correctly preserves garment styles (e.g., clothing length and design) and maintains high pose alignment, while achieving accuracy comparable to or surpassing prior baselines, demonstrating robustness and practicality. Additional quantitative results on the DressCode subset are provided in Appendix B in detail.

Figure 5. Qualitative comparison on the VITON-HD dataset. "reference" denotes using additional reference $\boldsymbol{r}_i$ for the inference.

Results in Table 2 show that our method achieves state-of-the-art performance on the StreetTryOn benchmark. Notably, our model is neither trained on the StreetTryOn domain nor designed to handle the absence of garment images during training. These results demonstrate its strong generalization ability and flexibility, and further confirm that it can directly extract garment information from the reference image for virtual try-on.

4.2. Qualitative comparison

Previous methods can fit garments onto a person, but often fail to preserve wearing-critical attributes such as cut, style, and fine details. By contrast, our method achieves improved visual fidelity among all baselines. As shown in Fig. 5, on VITON, RefTon realistically renders challenging materials such as hollow and semi-transparent fabrics even without reference images. For example, it faithfully preserves the perforated structures and transparency of lace garments, while baselines often produce solid textures or spurious dotted artifacts. Our method also better maintains garment patterns, keeping printed letters and logos clear and consistent with the input. Adding reference images further improves generation quality, validating the effectiveness of our reference assist framework. Even without agnostic masks, directly transferring clothing $\boldsymbol{c}_i$ onto person images $\bar{\boldsymbol{p}}_{i,\boldsymbol{c}_j}$ still allows our method to outperform most baselines. More results on VITON-HD, DressCode, and in-the-wild examples are provided in Appendix C.

Our model also performs well on a benchmark with a

Table 1. Quantitative comparison on VITON-HD [8] and DressCode [44]. Best and second-best results are shown in **bold** and underline, respectively. "+R" denotes the use of reference images, "MF" indicates mask-free inputs, and "–" denotes missing values. Subscripts p and u denote the *paired* and *unpaired* settings, respectively.

Method	**Input**		**VITON-HD**						**DressCode**					
	Mask	**Pose**	**LPIPS**$_p$ ↓	**SSIM**$_p$ ↑	**FID**$_p$ ↓	**KID**$_p$ ↓	**FID**$_u$ ↓	**KID**$_u$ ↓	**LPIPS**$_p$ ↓	**SSIM**$_p$ ↑	**FID**$_p$ ↓	**KID**$_p$ ↓	**FID**$_u$ ↓	**KID**$_u$ ↓
CAT-DM* [63]	✓	✓	0.080	0.877	5.60	0.83	8.93	1.37	–	–	–	–	–	–
IDM-VTON* [9]	✓	✓	0.102	0.870	6.29	–	–	–	0.062	0.920	8.64	0.90	–	–
OOTDiffusion* [60]	✓	–	0.071	0.878	8.81	0.82	–	–	0.045	**0.927**	4.20	**0.37**	–	–
CatVTON* [10]	✓	–	0.057	0.870	5.43	0.41	9.02	1.09	0.046	0.892	3.99	0.82	6.14	1.40
CatVT2ON* [11]	✓	✓	0.057	**0.890**	8.10	2.25	11.22	2.99	0.037	0.922	5.72	2.34	8.63	3.84
OmniVTON* [62]	✓	✓	0.145	0.832	7.76	–	9.62	–	0.119	0.865	5.34	–	6.45	–
PromptDresser$^*_{pose}$ [31]	✓	✓	0.097	0.878	9.07	1.16	–	–	–	–	–	–	–	–
PromptDresser* [31]	✓	–	0.112	0.869	8.54	**0.67**	–	–	–	–	–	–	–	–
RefTON (Ours)	✓	–	0.057	0.873	5.45	0.82	8.58	1.06	0.037	0.912	3.48	1.20	5.31	1.36
RefTON+R (Ours)	✓	–	**0.049**	0.879	**4.69**	0.68	**8.43**	**0.91**	**0.031**	0.918	**2.94**	0.95	**5.07**	**1.15**
							Mask-Free setting							
CatVTON(Mask-Free)* [10]	–	–	0.061	0.870	5.89	**0.51**	9.29	1.17	0.045	0.902	4.78	1.30	7.40	2.62
Any2AnyTryon* [23]	–	–	0.088	0.839	6.93	0.74	8.97	0.98	–	–	–	–	–	–
TryOffDiff* [56]	–	–	–	–	–	–	11.9	2.60	–	–	–	–	7.90	2.70
RefTON/MF (Ours)	–	–	0.061	0.866	5.98	1.04	8.40	0.81	0.041	0.901	3.84	1.33	**5.00**	**1.17**
RefTON+R/MF(Ours)	–	–	**0.053**	**0.872**	**5.11**	0.82	**8.32**	**0.78**	**0.035**	**0.906**	**3.34**	**1.15**	5.02	1.28

Table 2. Quantitative comparisons on the StreetTryOn. Sh, St, P, denotes the shop, street, and person(model) image, respectively.

Method	**Required Input**			**Sh-to-St**	**P-to-P**	**P-to-St**	**St-to-St**
	Mask	**Pose**	**Text**	**FID**↓	**FID**↓	**FID**↓	**FID**↓
StreetTryOn	✓	✓	✗	34.054	12.185	34.191	33.039
OmniVTON	✓	✓	✓	33.919	8.983	33.450	23.470
RefTON	✓	✗	✗	**28.991**	**8.870**	**25.429**	**16.452**

Table 3. Cross-dataset comparison with OOTDiffusion [60] on VITON-HD and DressCode under both paired and unpaired settings. For the mask-free setting, we report only the unpaired results, as the paired setting is not meaningful when no garment–person alignment is enforced.

Methods	**VITON-HD (Paired)**				**DressCode (Paired)**			
	SSIM↑	**FID**$_p$↓	**KID**$_p$↓	**LPIPS**↓	**SSIM**↑	**FID**$_p$↓	**KID**$_p$↓	**LPIPS**↓
OOTDiffusion* [60]	0.839	11.22	2.72	0.123	0.915	11.96	1.21	0.061
RefTON (Ours)	0.851	6.23	0.80	0.072	0.896	3.70	1.13	0.045
RefTON+R (Ours)	**0.859**	**5.13**	**0.62**	**0.060**	**0.903**	**3.14**	**0.97**	**0.038**

Methods	**VITON-HD (Unpaired)**		**DressCode (Unpaired)**	
	FID$_u$↓	**KID**$_u$↓	**FID**$_u$↓	**KID**$_u$↓
RefTON (Ours)	9.11	1.08	5.22	1.20
RefTON+R (Ours)	8.59	0.87	5.03	1.11
RefTON/MF (Ours)	8.88	0.82	5.03	1.23
RefTON+R/MF (Ours)	**8.39**	**0.65**	**4.87**	**1.10**

wider variety of clothing types. As shown in Fig. 6, we evaluate our method on the DressCode dataset, where garments are categorized into *upper body*, *lower body*, and *dresses*. Our approach produces more faithful and natural try-on results compared to previous methods. In particular, it renders reflective materials, such as leather and metallic fabrics, with superior realism, avoiding the over-smoothing or distortion artifacts commonly observed in other methods. Moreover, even without agnostic masks, our model can still perform consistent try-on guided by garment style, accurately preserving the length, structure, and overall design without introducing mismatched or inconsistent shapes.

4.3. Training on VRF Dataset and Evaluation

To evaluate the effectiveness of our person-to-person virtual try-on framework and data construction pipeline, we build a mixed dataset, *Mixed-Virtual-Ref*, by combining samples from DressCode, VITON-HD, FashionTryOn [59], ViViD [17], and IGPairs [9]. We use *Qwen2.5-VL* to filter low-quality images, generate agnostic masks via [10], and obtain reference and unpaired person images through our data generation process. In total, 103,936 image pairs are collected for training, using the same hyperparameters as in the DressCode and VITON-HD experiments.

We evaluate the model on the DressCode and VITON-HD test sets, with quantitative results shown in Table 3. Our method (RefTon) consistently achieves superior or comparable performance to OOTDiffusion [60] across both benchmarks and paired and unpaired settings. Notably, despite not being trained on DressCode or VITON-HD individually, the mixed-dataset model surpasses dataset-specific baselines on most metrics (e.g., FID and LPIPS), demonstrating strong cross-dataset generalization and the robustness of our data construction strategy.

4.4. Ablation Study

We conduct an ablation study to examine our model under four settings(w/&w/o mask,w/&w/o Ref.). As shown in Table 1, our model maintains consistently strong performance across all settings. Introducing a reference image yields clear improvements in both mask-based and mask-free modes, while moving from mask-based to mask-free inputs causes only mild metric fluctuations, confirming the model's stable robustness without masks.

Figure 6. **Qualitative comparison on the DressCode dataset.**, and the model is trained following the pipeline in Fig. 2 (b). “reference” denotes the reference $\boldsymbol{r}_i$ image is used during the inference.

Figure 7. **Qualitative results of the ablation study across different settings.** “Ref.” denotes that a reference image is provided, while “MF” indicates mask-free inputs using the original person image instead of a masked agnostic image.

Fig. 7 further provides qualitative comparisons. The first two rows show that reference images help the model correctly infer garment structures and materials that are ambiguous in the flat images (e.g., hollow textures, semi-transparency). The last two rows illustrate that mask quality heavily affects mask-dependent models: overly aggressive masks remove important items (e.g., handbags), while conservative masks retain unwanted regions (e.g., legs), leading to incorrect garment geometry. In contrast, our mask-free model consistently produces correct outputs regardless of mask or reference conditions, demonstrating that mask-free capability reduces reliance on mask quality and enables more flexible and stable try-on performance.

Table 4. **Rescaled position index (PI) ablation.**. We compare the original Flux-Kontext position index with our rescaled PI at different condition scales; MF denotes mask-free inputs. Conditional inputs are resized to $0.5\times$ and $0.25\times$ of the target resolution.

Method	**0.5 Scale**				**0.25 Scale**			
	FID↓	KID↓	FID_{MF}↓	KID_{MF}↓	FID↓	KID↓	FID_{MF}↓	KID_{MF}↓
w/ o Rescaled PI	5.29	0.84	4.75	0.72	6.08	0.93	**5.35**	0.76
w/ Rescaled PI	**5.09**	**0.79**	**4.71**	**0.69**	**6.01**	**0.77**	5.37	**0.71**

To validate the rescaled position index, we experiment on VITON-HD with cloth image $\boldsymbol{c}_i$, dense pose $\boldsymbol{d}_i$, warp mask $\boldsymbol{m}_i$, and reference image $\boldsymbol{r}_i$ as conditions. In the masked and mask-free settings, the target cloth is transferred to the person image and the agnostic image, respectively. We evaluate two condition scales, $0.25\times$ and $0.5\times$, where the person or agnostic image is resized to the target resolution and the conditional inputs are rescaled accordingly. As shown in Table 4, the rescaled position index outperforms the original *Flux-Kontext* position index on FID and KID under both settings, showing its effectiveness for multi-condition inputs with varying resolutions.

5. Conclusion

This paper introduces **RefTon**, a virtual try-on framework that supports both mask-based and mask-free inference and leverages reference images to guide the try-on process. We extend *Flux-Kontext* to handle multi-condition inputs of varying resolutions via a rescaled position index. To train RefTon, we propose a reference data generation pipeline integrating *Qwen2.5-VL* and *Flux-Kontext*. This design allows RefTon to faithfully preserve translucent fabrics, intricate designs, and fine details, consistently outperforming existing methods both quantitatively and qualitatively, achieving state-of-the-art performance in all settings.

6. Acknowledgments

This research was supported by 360 AI Research. We sincerely thank 360 AI Research for its support of this work.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[5] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In *European Conference on Computer Vision*, pages 124–142. Springer, 2024. 1

[6] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Kristjanson Duvenaud. Neural ordinary differential equations. In *Neural Information Processing Systems*, 2018. 2

[7] Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in neural information processing systems*, 37:128886–128910, 2024. 2

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 1, 2, 4, 5, 7

[9] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer, 2024. 1, 7

[10] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 1, 2, 3, 6, 7

[11] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025. 1, 7

[12] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. *arXiv preprint arXiv:2311.16094*, 2023. 6

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2

[15] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 843–850, 2014. 1

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. 2, 5, 7

[18] Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks. *arXiv preprint arXiv:2508.13632*, 2025. 3

[19] Yue Gong, Shanyuan Liu, Liuzhuozheng Li, Jian Zhu, Bo Cheng, Liebucha Wu, Xiaoyu Wu, Yuhang Ma, Dawei Leng, and Yuhui Yin. Cta-flux: Integrating chinese cultural semantics into high-quality english text-to-image communities. *arXiv preprint arXiv:2508.14405*, 2025. 2

[20] 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. 1

[21] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607, 2023. 2

[22] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 1, 3

[23] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19085–19096, 2025. 3, 4, 7

[24] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 1, 5

[25] Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18143–18154, 2025. 2

[26] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. *arXiv preprint arXiv:2407.10625*, 2024. 2

[27] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 1, 2, 3

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[29] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 2

[30] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8176–8185, 2024. 1, 2

[31] Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. *arXiv preprint arXiv:2412.16978*, 2024. 1, 7

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[35] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 3

[36] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 3, 4, 5

[37] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[38] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. 2, 3

[39] Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Bridge diffusion model: Bridge chinese text-to-image diffusion model with english communities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5541–5549, 2025. 2

[40] Shanyuan Liu, Jian Zhu, Junda Lu, Yue Gong, Liuzhuozheng Li, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Nanocontrol: A lightweight framework for precise and efficient control in diffusion transformer, 2025. 2

[41] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[42] Yuhang Ma, Bo Cheng, Shanyuan Liu, Hongyi Zhou, Liebucha Wu, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Nami: Efficient image generation via bridged progressive rectified flow transformers. *arXiv preprint arXiv:2503.09242*, 2025. 2

[43] Ming Meng, Qi Dong, Jiajie Li, Zhe Zhu, Xingyu Wang, Zhaoxin Fan, Wei Zhao, and Wenjun Wu. Hf-vton: High-fidelity virtual try-on via consistent geometric and semantic alignment. *arXiv preprint arXiv:2505.19638*, 2025. 2

[44] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022. 2, 4, 5, 7

[45] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM international conference on multimedia*, pages 8580–8589, 2023. 2

[46] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 6

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[49] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1, 2

[50] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, 2020. Version 0.3.0. 6

[51] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinghui Tang. Imagdressing-v1: Customizable virtual dressing. In *AAAI Conference on Artificial Intelligence*, 2024. 2, 4, 5

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[53] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[54] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2

[55] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[56] Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Enhancing person-to-person virtual try-on with multi-garment virtual try-off. *arXiv preprint arXiv:2504.13078*, 2025. 3, 7

[57] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 1

[58] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1

[59] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 7

[60] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI Conference on Artificial Intelligence*, 2024. 1, 7

[61] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1

[62] Zhaotong Yang, Yuhui Li, Shengfeng He, Xinzhe Li, Yangyang Xu, Junyu Dong, and Yong Du. Omnivton: Training-free universal virtual try-on. *arXiv preprint arXiv:2507.15037*, 2025. 1, 3, 7

[63] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8372–8382, 2024. 7

[64] Nannan Zhang, Zhenyu Xie, Zhengwentai Sun, Hairui Zhu, Zirong Jin, Nan Xiang, Xiaoguang Han, and Song Wu. Viton-gun: Person-to-person virtual try-on via garment unwrapping. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7740–7751, 2025. 3

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[66] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM international conference on multimedia*, pages 266–274, 2019. 2, 5

[67] Jian Zhu, Shanyuan Liu, Liuzhuozheng Li, Yue Gong, He Wang, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Dawei Leng, et al. Flux-makeup: High-fidelity, identity-consistent, and robust makeup transfer via diffusion transformer. *arXiv preprint arXiv:2508.05069*, 2025. 2

[68] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4606–4615, 2023. 1, 2