# STRUCTURE-DETAIL DECOUPLED AUTOREGRESSIVE GENERATION FOR FAST AND HIGH-FIDELITY VIRTUAL TRY-ON

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Virtual try-on (VTON) is a bi-conditional image generation problem that requires not only accurate person preservation but also faithful garment deformation and detail synthesis. Diffusion-based VTON methods can jointly model these factors in a compressed latent space, but suffer from high-frequency detail loss due to inherent latent compression, even with costly multi-step denoising. Recent visual autoregressive (VAR) models offer a promising alternative for high-quality generation with faster inference, yet remain unexplored for VTON due to the lack of effective bi-conditioning mechanisms. To bridge this gap, we first introduce VAR-VTON, a VAR-based VTON model that incorporates garment conditioning and structural guidance for efficient latent-space VTON. Despite its efficacy, latent-space generation still struggles to preserve fine-grained garment details. We argue that different VTON sub-tasks should be addressed in different representation spaces: structural synthesis such as garment warping and person layout is suited to the latent space, whereas fine-grained detail recovery should be tackled in the pixel space. Motivated by this insight, we further propose STAR-VTON, a Two-Stage AutoRegressive framework that builds upon VAR-VTON by decoupling latent-space structural synthesis from pixel-space detail recovery. Our idea is to resort to a matching-informed refiner to establish dense correspondences between the stage-one generation and the source garment to directly map fine-grained pixel-space details. Extensive experiments show that STAR-VTON achieves an impressive efficiency-fidelity trade-off: VAR-VTON runs at least $4\times$ faster than diffusion-based counterparts without degrading quality, and the pixel-space refiner effectively restores fine details and acts as a plug-and-play module that can benefit existing VTON approaches.

## 1 INTRODUCTION

Given a person image and a garment image, virtual try-on (VTON) aims to synthesize a photorealistic image in which the person is naturally and accurately dressed in the target garment (Han et al., 2018). It is a typical bi-conditional generative problem, because a faithful VTON generation requires not only preserving the identity and pose of the person but also the fidelity of the garment.

Achieving high-fidelity VTON, however, is non-trivial. It needs to i) correctly deform the garment to align with the target body pose and shape, ii) well preserve person attributes including appearance and pose, and iii) faithfully retain garment details such as logos and text. Early GAN-based methods (Han et al., 2018; Wang et al., 2018; Choi et al., 2021) typically

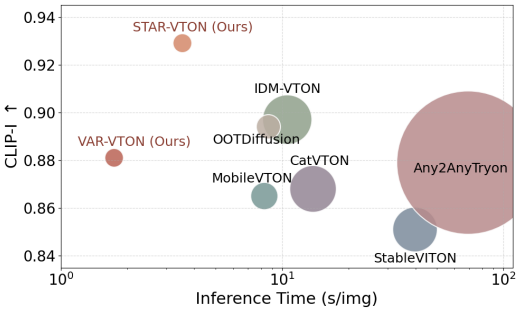


Figure 1: **Comparison with the state of the art in terms of detail fidelity and efficiency.** Bubble size denotes computational cost (TFLOPs). We evaluate detail preservation using CLIP-I (Hessel et al., 2021), with higher values indicating better fidelity. Our solutions achieve better fidelity-efficiency trade-off.



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.

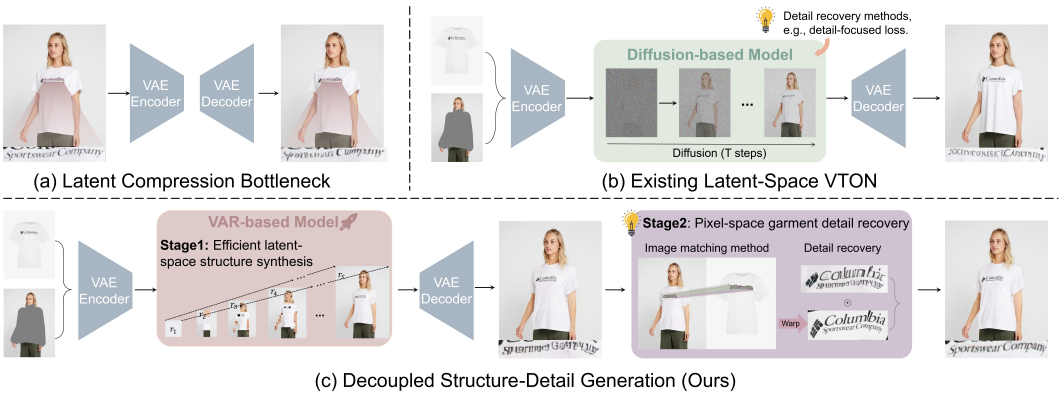


Figure 2: **Comparison between latent-space virtual try-on and our structure-detail decoupled generation framework.** (a) Existing VTON frameworks rely on compressed latent representations, where high-frequency garment details are inevitably lost during VAE encoding. Consequently, (b) latent-space generation struggles to faithfully preserve fine-grained garment textures and textual patterns. (c) Our framework leverages the VAR model for efficient holistic structure synthesis and dense image matching for pixel-space garment detail recovery, enabling both fast inference and faithful detail preservation.

follow a two-stage pipeline: a module first aligns garment deformation and human pose, followed by a generator (Goodfellow et al., 2020) that synthesizes the try-on image. However, the reliance on explicit warping renders them vulnerable to large pose changes and non-rigid deformations.

Recent diffusion-based methods (Kim et al., 2024; Choi et al., 2024; Chong et al., 2025) alleviate the limitations above by jointly modeling garment deformation, person preservation, and garment-detail synthesis as a single generative process in the compressed VAE latent space (Kingma & Welling, 2013), obviating explicit warping while yielding improved geometric alignment and photorealism. Despite their visual appeal, these methods suffer from i) high inference latency from multi-step denoising and ii) the inherent loss of high-frequency details due to VAE compression (Fig. 2(a)). Consequently, they struggle to maintain pixel-level detail fidelity, particularly for fine-grained patterns and text (Fig. 2(b)).

Unlike diffusion models, visual autoregressive (VAR) models (Tian et al., 2024; Han et al., 2025) have demonstrated high-quality image generation with substantially fewer decoding steps. Interestingly, this promising direction remains largely unexplored for efficient VTON. A possible reason is that VAR has not yet developed mature bi-conditional image generation toolkits such as ControlNet (Zhang et al., 2023) and ReferenceNet (Zhu et al., 2023; Hu, 2024). To fill this gap, we first introduce **VAR-VTON**, a VAR-based VTON model for fast latent-space VTON, focusing on garment-person interaction and structural integrity. It is featured by i) a decoupled garment-prefix conditioning strategy that enables sufficient person-garment interaction during VAR generation, and ii) a scale-wise embedding mechanism of garment-agnostic representation that encourages the structural consistency of the target person. However, we observe that our VAR-VTON, while more efficient, still fails to address the garment detail degradation (Fig. 2(c)).

The phenomenon above motivates us to rethink the current one-stage latent-space VTON. Indeed structural synthesis (e.g., pose alignment and deformation) is appropriately tackled in the compact latent space, but high-frequency details cannot be well recovered without accessing uncompressed pixel-level features. In this work, we return to the previous two-stage paradigm. Instead of burdening a single latent-space generator to resolve all challenges simultaneously, we posit that *different subtasks should be tackled in representation spaces suited to their intrinsic properties*. This motivates us to reformulate VTON as a structure-detail decoupled generation problem. To this end, we propose **STAR-VTON**, a novel two-stage VTON approach based on VAR-VTON, which decouples efficient structural synthesis from faithful garment-detail recovery, as illustrated in Fig. 2(c).

Given the stage-one STAR-VTON generation, we focus on pixel-space garment detail recovery in the second stage and introduce dense matching (Shen et al., 2024) to recover high-frequency garment

details from the original garment image. Specifically, we propose a matching-informed refiner that identifies texture-rich regions and dynamically injects warped garment details into the stage-one generation. By transferring garment details directly in pixel space, our method effectively preserves thin texts, small logos, and intricate patterns that are typically lost in latent-space generation.

Extensive experiments demonstrate that STAR-VTON achieves a remarkable trade-off between generation fidelity and efficiency. Specifically, VAR-VTON runs over $4\times$ faster than diffusion-based VTON counterparts while maintaining competitive generation quality, and the proposed pixel-space refiner consistently sets a new state of the art in garment-detail preservation on both VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). Notably, the refiner is model-agnostic and can be readily integrated into existing VTON methods, consistently improving detail preservation across different generative backbones. Overall, our results suggest that a two-stage framework that explicitly assigns structural synthesis and detail recovery to their respective representation spaces provides a new paradigm for designing efficient and high-fidelity virtual try-on.

## 2 Related Work

Our work is related to image-based VTON, controllable VAR generation, and dense image matching.

### 2.1 Image-based Virtual Try-On

Early image-based VTON approaches (Han et al., 2018; Wang et al., 2018; Yu et al., 2019; Choi et al., 2021; Lee et al., 2022) commonly adopt a two-stage pipeline by first warping the garment to the target pose and then synthesizing the try-on image (Goodfellow et al., 2020). But inaccurate warping often leads to structural distortion under large pose variations.

Recently diffusion-based models (Ho et al., 2020; Peebles & Xie, 2023) have renewed the field with a unified generative process. Within this paradigm, a primary research focus has been dedicated to resolving the challenge of detail preservation. Early diffusion-based methods such as LaDI-VTON (Morelli et al., 2023) and DCI-VTON (Gou et al., 2023) leverage CLIP (Radford et al., 2021) embeddings to encode garment semantics, but these embeddings prioritize high-level concepts and fail to retain fine-grained textures. To mitigate this, StableVITON (Kim et al., 2024) adopts a ControlNet-like Zhang et al. (2023) architecture, and dual-stream designs (Zhu et al., 2023; Choi et al., 2024; Xu et al., 2025; Zhou et al., 2025; Wan et al., 2025) process the source garment and the target person independently to improve garment fidelity. However, parallel streams introduce substantial computational overhead; CatVTON (Chong et al., 2025) and TPD (Yang et al., 2024) therefore consolidate feature extraction into a single-stream U-Net to improve efficiency. In parallel with detail enhancement, accelerating inference has also received much attention, e.g., MobileVTON (Wan et al., 2026) reduces computational cost through model distillation, and CAT-DM (Zeng et al., 2024) addresses the speed bottleneck by initializing the reverse diffusion process with a GAN-generated implicit distribution, thereby drastically reducing the number of sampling steps without sacrificing visual quality.

Despite these attempts to jointly optimize detail retention and inference efficiency, existing pipelines remain hindered by their fundamental architectural constraints. In particular, the iterative denoising procedure imposes a strict lower bound on generation speed, while the reliance on the bottlenecked VAE latent space inevitably discards high-frequency information during compression. To overcome these limitations, we explore a decoupled generation framework. We leverage next-scale VAR prediction (Tian et al., 2024) for rapid global structure synthesis, and establish direct correspondences in the uncompressed pixel domain to recover garment details.

### 2.2 Controllable Visual Autoregressive Learning

Recent VAR models (Tian et al., 2024; Han et al., 2025) have achieved competitive image generation quality with far fewer decoding steps than diffusion models, making them attractive for efficient generation. To enable conditional control, existing methods typically inject guidance either at the pre-filling stage or during decoding.

Pre-filling based methods (Li et al., 2024; Qu et al., 2025; Dong et al., 2026) prepend conditional tokens before VAR generation, enabling interaction between conditioning signals and generated to-

kens, which is suitable for appearance conditions that are spatially misaligned with the target image. In contrast, decoding-phase methods (Li et al., 2025; Yao et al., 2024; Xu et al., 2026) inject control signals throughout generation, providing stronger structural consistency when the conditioning signal is spatially aligned with the target image, such as pose maps or segmentation layouts.

VTON requires both strong garment-person interaction and faithful preservation of body structure. To satisfy both requirements, we adopt a hybrid conditioning strategy that injects garment appearance during pre-filling and garment-agnostic structural cues during decoding, enabling both human-garment interaction and person-consistent generation.

### 2.3 Dense Image Matching

Recent advances in dense image matching have significantly improved the robustness of pixel-level correspondence under large viewpoint changes and non-rigid deformations (Sun et al., 2021; Edstedt et al., 2023; Shen et al., 2024). In particular, generalizable matching frameworks trained on large-scale video datasets (Shen et al., 2024) have demonstrated a remarkable ability to establish reliable dense correspondences across diverse appearances and geometric transformations. These developments make it increasingly feasible to recover fine-grained visual details through correspondence-based transfer rather than generative synthesis.

Motivated by this opportunity, we revisit garment-detail preservation from a correspondence perspective. Rather than relying solely on generative models to synthesize fine-grained textures, we leverage dense image matching to directly recover garment details from the source image. This perspective bypasses the information bottleneck of latent-space generation and opens a new direction for preserving garment texture fidelity in VTON.

## 3 STAR-VTON: Two-Stage Autoregressive Virtual Try-On

We begin with the problem formulation and then present the technical details of STAR-VTON.

### 3.1 Problem Formulation

Given an image of a target person $\boldsymbol{x}_p$ and an image of a source garment $\boldsymbol{x}_g$, we extract a set of garment-agnostic representations $\mathcal{A} = \{\boldsymbol{m}, \boldsymbol{x}_m, \boldsymbol{x}_d\}$ from $\boldsymbol{x}_p$, where $\boldsymbol{m}$, $\boldsymbol{x}_m$, and $\boldsymbol{x}_d$ denote the garment mask, masked person image, and dense pose, respectively. In the first stage, structure synthesis is modeled via a VAR framework. A multi-scale vector-quantized (VQ)-encoder processes $\boldsymbol{x}_p$ to produce $S$ discrete token maps at progressively increased scales, denoted by $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_S)$. Conditioned on the representation set $\mathcal{A}$ and the garment $\boldsymbol{x}_g$, a generator predicts the token maps in a coarse-to-fine manner. This process commences with the generation of the $1 \times 1$ map $\boldsymbol{r}_1$ and sequentially produces the map of each subsequent scale, that is,

$$p(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_S \mid \mathcal{A}, \boldsymbol{x}_g) = \prod_{s=1}^{S} p(\boldsymbol{r}_s \mid \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{s-1}, \mathcal{A}, \boldsymbol{x}_g)\,. \tag{1}$$

A VQ-decoder then maps the generated sequence of tokens back to the pixel space, yielding an intermediate try-on result $\hat{\boldsymbol{x}}_{tr}$. While $\hat{\boldsymbol{x}}_{tr}$ generally captures the global topology and semantic coherence, it inherently lacks high-frequency details due to latent compression. Therefore, in the second stage, a pixel-level refinement process $f$ is introduced to recover these fine-grained details directly from the uncompressed garment $\boldsymbol{x}_g$, yielding the final high-fidelity try-on generation $\boldsymbol{x}_{tr}$ by

$$\boldsymbol{x}_{tr} = f(\hat{\boldsymbol{x}}_{tr}, \boldsymbol{x}_g)\,. \tag{2}$$

### 3.2 Technical Challenges and Overview

Applying VAR to virtual try-on is not a straightforward replacement of diffusion denoising with autoregressive token prediction. Unlike text- or class-conditioned generation, VTON requires jointly modeling two image-derived conditions with distinct roles: the source garment provides transferable appearance, while the garment-agnostic representation specifies the target body layout. This introduces two VAR-specific challenges: i) the garment should be globally accessible throughout

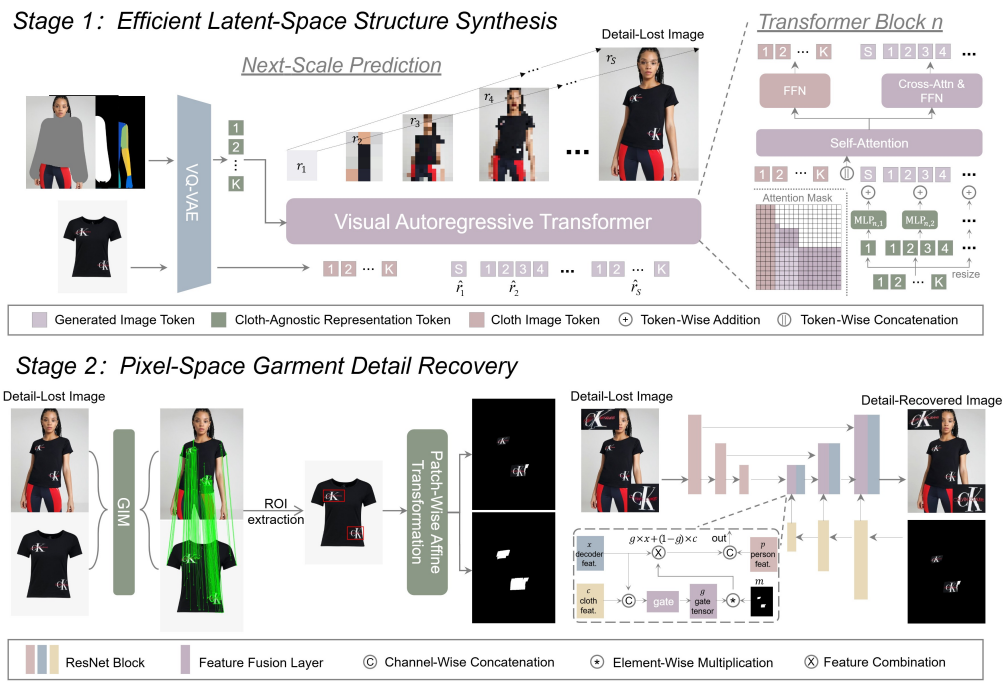


Figure 3: **Overview of STAR-VTON.** Our method adopts a two-stage framework. The first stage consists of two components: i) *decoupled garment prefix conditioning*, which fuses garment tokens and image tokens within self-attention while maintaining separate processing pathways thereafter; and ii) *scale-wise garment-agnostic representation embedding*, where multi-scale garment-agnostic features are progressively integrated into the generated tokens via dedicated MLPs at each scale and block. The second stage employs iii) a *matching-informed refiner*, which establishes dense correspondences between the source garment and the generated output to recover details.

autoregressive generation without being entangled with the target token sequence, and ii) structural errors introduced at coarse scales can propagate and accumulate during multi-scale decoding.

We propose STAR-VTON, a two-stage structure-detail decoupled framework for efficient and high-fidelity virtual try-on. To handle the VAR-specific issues discussed above, the first stage performs latent-space structure synthesis with two mechanisms: i) *decoupled garment prefix conditioning*, which uses garment tokens as condition prefixes to guide multi-scale token prediction; and ii) *scale-wise garment-agnostic representation embedding*, which injects target-person structural cues across autoregressive scales to maintain geometric consistency. The second stage further overcomes the latent compression bottleneck in fine-grained detail preservation through iii) a *matching-informed refiner*, which establishes dense correspondences between the original garment $\boldsymbol{x}_g$ and the intermediate VAR output $\hat{\boldsymbol{x}}_{tr}$, and selectively transfers high-frequency details to produce the final high-fidelity VTON result. The overall pipeline is illustrated in Fig. 3.

### 3.3 Latent-Space Autoregressive Structure Synthesis

**Decoupled Garment Prefix Conditioning.** The garment information is injected in a prefix-based manner. We find that directly concatenating the garment tokens and the multi-scale image tokens at the input end may entangle garment information with the text conditioning. To preserve garment-specific information, we adopt a decoupled prefix-based encoding. Specifically, we concatenate them only before the self-attention layer. After self-attention, the two sequences are separated: the image tokens proceed to the subsequent cross-attention and FFN, while the garment tokens are processed by another FFN. Within self-attention, we design a causal mask (Fig. 3). It allows the generated tokens to attend unrestrictedly to the garment tokens, thereby extracting visual cues as needed. At the same time, it prevents garment tokens from attending to the generated tokens, preserving the autoregressive sampling trajectory.

**Scale-Wise Garment-Agnostic Representation Embedding.** SCALAR (Xu et al., 2026) and ControlAR (Li et al., 2025) show that for spatially aligned generation tasks, simple additive conditioning is sufficient for effective control. Following this observation, we inject garment-agnostic representations into the VAR block through scale-specific embeddings. Specifically, the masked person image $\boldsymbol{x}_m$ and dense pose $\boldsymbol{x}_d$ are encoded into latent variables $\boldsymbol{z}_m$ and $\boldsymbol{z}_d$ via a VQ-VAE, while the mask $\boldsymbol{m}$ is downsampled to match the resolution of the latent codes. The masked image latent $\boldsymbol{z}_m$, the dense pose latent $\boldsymbol{z}_d$, and the resized mask $\boldsymbol{m}$ are concatenated along the channel dimension to form the conditioning representation $\mathcal{A}$. To match the length of the generated tokens, $\mathcal{A}$ is resized to multiple scales. For each VAR block and at each scale of the generated tokens, a dedicated MLP processes the correspondingly scaled garment-agnostic representation, and the output is then added to the generated tokens. Formally,

$$\mathcal{A} = \boldsymbol{z}_m \copyright \boldsymbol{z}_d \copyright \boldsymbol{m}, \quad \mathcal{A}_{r_1}, \mathcal{A}_{r_2}, \ldots, \mathcal{A}_{r_S} = \text{Resize}(\mathcal{A}), \tag{3}$$

$$\boldsymbol{x}_{i,j} = \boldsymbol{x}_{i,j} + \text{MLP}_{i,j}(\mathcal{A}_{r_j}), \tag{4}$$

where $\boldsymbol{x}_{i,j}$ denotes the token at the $j$-th scale output by the $i$-th block, $\mathcal{A}_{r_j}$ denotes the garment-agnostic representation at the $j$-th scale, and © denotes channel-wise concatenation.

### 3.4 Pixel-Space Matching-Informed Detail Recovery

We introduce a matching-informed refiner to recover lost details in VAE latents. The pixel correspondences between $\hat{\boldsymbol{x}}_{tr}$ and $\boldsymbol{x}_g$ are initially established using GIM (Shen et al., 2024). GIM produces a dense matching field $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$, which maps garment pixels to their corresponding locations in the try-on image and thus provides the basis for pixel-space detail transfer. Nevertheless, using $\Phi$ to warp the entire garment is prone to misalignment, especially in textureless, occluded, or heavily deformed regions. We therefore use the matching points not only for warping but also for selecting reliable regions to warp. Specifically, we construct a density map on the source garment from the matching points using Gaussian kernels and extract the high-density regions via thresholding, as the high-density regions generally indicate the areas that require textural refinement. To achieve precise alignment with the try-on image, each selected region is further partitioned into multiple small patches that are warped independently and then composited. This process is formalized as

$$\boldsymbol{c}_w, \boldsymbol{c}_m = \sum_{p=1}^{P} \text{Warp}(\boldsymbol{c}_p, \Phi), \tag{5}$$

where $P$ denotes the number of patches, $\boldsymbol{c}_p$ is the $p$-th garment patch, $\boldsymbol{c}_w$ is the warped garment, and $\boldsymbol{c}_m$ is the corresponding mask. An interesting phenomenon is that, even though textural details are lost in the try-on result, image matching remains accurate to produce precise garment warping.

Based on the warped garment regions, we employ a lightweight dual-encoder U-Net refiner to selectively inject the recovered pixel-space details into the stage-one generation. The refiner takes the intermediate try-on result $\hat{\boldsymbol{x}}_{tr}$ together with the warped garment $\boldsymbol{c}_w$ as input. Specifically, the person encoder extracts a set of multi-scale structural features $\{\boldsymbol{f}_p^l\}_{l=1}^L$ from $\hat{\boldsymbol{x}}_{tr}$, while the garment encoder extracts $\{\boldsymbol{f}_c^l\}_{l=1}^L$ from $\boldsymbol{c}_w$, where $L$ denotes the number of decoding stages. The mask $\boldsymbol{c}_m$ is resized into a multi-scale pyramid $\{\boldsymbol{c}_m^l\}_{l=1}^L$ to constrain texture injection at each scale. At the $l$-th decoding stage, let $\boldsymbol{f}_{dec}^l$ be the decoder feature. We update $\boldsymbol{f}_{dec}^l$ via a gating mechanism :

$$\boldsymbol{g}^l = \sigma\big(\text{Conv}\big(\text{Concat}[\boldsymbol{f}_c^l, \boldsymbol{f}_{dec}^l]\big)\big) \in [0,1]^{h \times w}, \tag{6}$$

$$\boldsymbol{g}^l \leftarrow \boldsymbol{g}^l \odot \boldsymbol{c}_m^l, \tag{7}$$

$$\tilde{\boldsymbol{f}}_{dec}^l = \boldsymbol{f}_c^l \odot \boldsymbol{g}^l + \boldsymbol{f}_{dec}^l \odot (1 - \boldsymbol{g}^l), \tag{8}$$

where $\sigma$ is the sigmoid activation and $\odot$ denotes element-wise multiplication. The gate provides a spatially adaptive reliability measure for texture injection. Instead of directly pasting the warped garment features, it allows the decoder to selectively borrow high-frequency details from $\boldsymbol{f}_c^l$ only where the warped texture is reliable, while preserving the synthesized structure in $\boldsymbol{f}_{dec}^l$ where local misalignment may occur. The mask $\boldsymbol{c}_m^l$ further confines the gating operation to the matched garment regions, preventing irrelevant background or unreliable pixels from affecting the generated person structure. The updated feature $\tilde{\boldsymbol{f}}_{dec}^l$ is then concatenated with the person skip feature $\boldsymbol{f}_p^l$ and propagated to the next decoding stage. Training details are provided in the appendix.

Table 1: Quantitative comparison on VITON-HD and DressCode. * denotes metrics computed on the detail-rich subset using garment-region crops, as described in Sec. 4.1. Inference latency is measured on an RTX 3090 GPU using bf16 precision.

| Method | VITON-HD | | | | | DressCode | | | | | Steps | TFLOPs | Latency (s/img) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paired | | | Unpaired | | Paired | | | Unpaired | | | | |
| | FID ↓ | KID ↓ | CLIP-I* ↑ | FID ↓ | KID ↓ | FID ↓ | KID ↓ | CLIP-I* ↑ | FID ↓ | KID ↓ | | | |
| StableVITON (Kim et al., 2024) | 6.39 | 1.11 | 0.851 | 9.30 | 1.42 | - | - | - | - | - | 50 | 370.5 | 39.81 |
| IDM-VTON (Choi et al., 2024) | 6.15 | 1.07 | 0.897 | 9.24 | 1.32 | 3.59 | 1.06 | 0.909 | 5.46 | 1.46 | 30 | 451.9 | 10.54 |
| OOTDiffusion Xu et al. (2025) | 6.22 | 1.79 | 0.894 | 9.25 | 1.99 | 4.03 | 0.83 | 0.909 | 7.56 | 2.59 | 20 | 108.7 | 8.66 |
| CatVTON (Chong et al., 2025) | 7.27 | 1.68 | 0.868 | 9.24 | 1.81 | 3.96 | 1.02 | 0.897 | **5.34** | **1.07** | 50 | 404.0 | 13.78 |
| Any2AnyTryon (Guo et al., 2025) | 8.08 | 1.74 | 0.879 | 10.17 | 2.20 | - | - | - | 7.61 | 2.56 | 30 | 3791.1 | 69.20 |
| MobileVTON (Wan et al., 2026) | 6.74 | 1.79 | 0.865 | 10.49 | 2.22 | - | - | - | - | - | 28 | 138.1 | 8.30 |
| **VAR-VTON (Ours)** | 5.86 | 0.85 | 0.881 | 8.80 | 0.86 | 3.35 | 0.74 | 0.913 | 5.62 | 1.61 | **13** | **64.4** | **1.74** |
| **STAR-VTON (Ours)** | **5.84** | **0.84** | **0.929** | **8.80** | **0.85** | **3.33** | **0.74** | **0.930** | 5.56 | 1.57 | 14 | 66.3 | 3.54 |

# 4 EXPERIMENTS

Here we present our experimental setup, results, and discussion.

## 4.1 EXPERIMENTAL SETUP

**Datasets.** We evaluate our method on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). VITON-HD contains $13,679$ upper-body image pairs, split into $11,647$ training pairs and $2,032$ testing pairs. DressCode contains $53,795$ pairs from three categories: $15,366$ upper-body garments, $8,951$ lower-body garments, and $29,478$ dresses. We follow the official split of DressCode, using $1,800$ pairs per category for testing and the remaining pairs for training.

**Implementation details.** The VAR model initializes weights from the Infinity (Han et al., 2025). Training proceeds in two stages: first at a resolution of $512 \times 384$ for 10k steps, then at $1024 \times 768$ for 15k steps. We train the refiner on $2,000$ detail-rich image pairs from VITON-HD. All training is conducted on 4 NVIDIA A6000 GPUs (48G memory each) with the AdamW optimizer and a batch size of 16. The learning rate is set to $1 \times 10^{-5}$.

**Evaluation.** We evaluate overall image quality in both paired and unpaired settings using Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018). To specifically assess garment-detail preservation, we construct detail-rich subsets from both VITON-HD and DressCode and measure the CLIP Image Similarity (CLIP-I) Hessel et al. (2021) between the generated and ground-truth garments. To focus exclusively on texture recovery, CLIP-I is computed only within the garment region. In addition, we report Structural Similarity Index (SSIM) (Wang et al., 2004) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) on paired samples in the appendix as supplementary evaluations.

**Baselines.** We select StableVITON (Kim et al., 2024), IDM-VTON (Choi et al., 2024), OOTDiffusion (Xu et al., 2025), CatVTON Chong et al. (2025), Any2AnyTryon (Guo et al., 2025), and MobileVTON (Wan et al., 2026) as baselines. For evaluation, each method generates try-on images at $1024 \times 768$ resolution when supported; otherwise, we generate results at $512 \times 384$, upscale them to $1024 \times 768$ using interpolation or super-resolution (Wang et al., 2021), and report the best performance. For fair latency comparison, all methods are evaluated with bf16 precision on a single RTX 3090 GPU, using the official sampling steps and averaging inference time over $10$ images.

## 4.2 MAIN RESULTS

Table 1 and Fig. 1 present quantitative comparisons on VITON-HD and DressCode. Overall, STAR-VTON achieves a superior trade-off between generation fidelity and efficiency. The results demonstrate that autoregressive generation serves as a strong alternative to diffusion models for VTON, achieving competitive or even superior image quality while significantly reducing computational cost and inference latency. We additionally report results at $512 \times 384$ resolution in the appendix for a more comprehensive comparison.




Figure 4: Qualitative comparison with baselines. Our method faithfully preserves fine-grained garment details, such as logos, text, and intricate patterns, that are often blurred or lost by existing methods. Zoom in for better viewing.

Beyond efficiency, the structure-detail decoupled design further improves fine-grained garment preservation. As shown in Table 1 and Fig. 4, the pixel-space refiner enhances texture fidelity without degrading the generation quality of the latent-space synthesis stage. The higher CLIP-I score of STAR-VTON further demonstrates more faithful preservation of garment appearance, especially fine-grained textures and visual details. These results support our hypothesis that global structure synthesis and detail recovery are better handled in latent and pixel spaces, respectively.

Furthermore, appendix experiments show that applying the proposed refiner to existing VTON methods consistently improves texture fidelity across existing VTON backbones, demonstrating the complementarity of pixel-space detail recovery and supporting our latent–pixel decoupling formulation.

### 4.3 Ablation Study

**Ablation on Garment Condition Embedding Strategy.** We compare different garment condition embedding strategies in VAR-VTON. As shown in Table 2, the decoding-phase conditioning strategy of SCALAR (Xu et al., 2026) performs worse, indicating that additive injection is insufficient for spatially misaligned garment appearance. Prefix conditioning improves all metrics by enabling direct interaction between garment tokens and generated tokens within self-attention. Our decoupled prefix design achieves the best performance, since it protects garment representations from subsequent conditional interference and allows more accurate appearance modeling.

**Ablation on Garment-Agnostic Condition Embedding Strategy.** We study different parameter-sharing strategies for garment-agnostic embedding in VAR-VTON. As shown in Table 3, using only block-specific or only scale-specific MLPs leads to inferior performance, while combining both performs best. This indicates that structural cues play different roles across VAR scales and depths, making scale- and block-specific projections more effective for preserving person structure and pose.

Table 2: Ablation for garment condition embedding strategy on VITON-HD ($512 \times 384$).

| embedding Strategy | FID↓ | KID↓ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|
| SCALAR | 6.59 | 0.42 | 0.854 | 0.070 |
| prefix | 5.87 | 0.35 | 0.852 | 0.065 |
| decoupled prefix | **5.22** | **0.28** | **0.864** | **0.054** |

Table 3: Ablation for garment-agnostic condition embedding strategy on VITON-HD ($512 \times 384$).

| Scale mlp | Block mlp | FID↓ | KID↓ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|---|
| ✗ | ✓ | 5.57 | 0.41 | 0.861 | 0.057 |
| ✓ | ✗ | 5.56 | 0.34 | 0.862 | 0.058 |
| ✓ | ✓ | **5.22** | **0.28** | **0.864** | **0.054** |

Table 4: Ablation for garment encoding module on VITON-HD ($512 \times 384$).

| Encoder | FID↓ | KID↓ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|
| Trainable conv | 10.21 | 1.57 | 0.845 | 0.106 |
| VQ-VAE | **5.22** | **0.28** | **0.864** | **0.054** |

Table 5: Ablation for classifier-free guidance on VITON-HD ($1024 \times 768$).

| CFG Strategy | FID↓ | KID↓ | SSIM↑ | LPIPS↓ | $t$ (s/img)↓ |
|---|---|---|---|---|---|
| w/ CFG | **5.75** | **0.80** | **0.878** | **0.073** | 2.48 |
| **w/o CFG** | 5.86 | 0.85 | 0.876 | 0.076 | **1.74** |




Figure 5: **Ablation on the gating strategy in the refiner.** Direct concatenation causes noticeable color shifts, while mask blending introduces obvious pasted artifacts. In contrast, the proposed gating mechanism produces more natural and coherent garment-detail recovery. Zoom in to compare.

**Ablation on Garment Encoding Module.** Table 4 compares garment encoding strategies in VAR-VTON. A trainable convolutional encoder performs notably worse than the pretrained VQ-VAE encoder, indicating that VQ-VAE latent features are better suited for autoregressive VTON generation.

**Ablation on Classifier-Free Guidance.** Table 5 evaluates the effect of classifier-free guidance (CFG). Although CFG brings moderate improvements in generation quality, it also increases inference latency. Since our primary goal is efficient virtual try-on generation, we adopt the setting without CFG as the default configuration.

**Ablation on Gating Strategy.** We design three feature fusion strategies for the proposed refiner: i) *Direct concatenation*, which directly concatenates the garment features $\boldsymbol{f}_c$ and decoder features $\boldsymbol{f}_{dec}$ along the channel dimension; ii) *Mask blending*, which directly combines the two features using the warped garment mask: $\boldsymbol{f}_{dec} \leftarrow \boldsymbol{f}_c \cdot \boldsymbol{c}_m + \boldsymbol{f}_{dec} \cdot (1 - \boldsymbol{c}_m)$ and iii) the proposed *Gated fusion*, which adaptively fuses the two features through the learned spatial gate. The results in Fig. 5 demonstrate the importance of adaptive spatial fusion for natural garment-detail recovery.

## 5 Conclusion

In this work, we present STAR-VTON, a novel virtual try-on framework that reformulates VTON as a decoupled structure-detail generation problem. Instead of relying on a unified latent-space generator to simultaneously handle garment deformation, person preservation, and fine-grained texture synthesis, our method separates efficient structure synthesis from faithful garment-detail recovery. Specifically, we explore Visual Autoregressive (VAR) modeling for fast latent-space try-on generation, and introduce a matching-informed pixel-space refiner to directly recover high-frequency garment textures through dense correspondence transfer. Extensive experiments demonstrate that the proposed framework significantly reduces inference cost while preserving fine-grained garment details with high fidelity. We hope this work highlights the importance of modeling different synthesis objectives in representation spaces that best match their intrinsic properties, and provides a new perspective for future research on efficient and high-fidelity virtual try-on generation.

### AI USE STATEMENT

In this work, we used generative AI tools solely for polishing the language of the manuscript, including improving grammar, clarity, and readability. No generative AI tools were used for research ideation, methodology development, experimental design or execution, code implementation, result analysis, or interpretation. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

## REFERENCES

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

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*, 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 *Proceedings of the European Conference on Computer Vision (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 *International Conference on Learning Representations*, 2025.

Ruixiao Dong, Zhendong Wang, Keli Liu, Li Li, Ying Chen, Kai Li, Daowen Li, and Houqiang Li. Echogen: Generating visual echoes in any scene via feed-forward subject-driven auto-regressive model. In *International Conference on Learning Representations*, 2026.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17765–17775, 2023.

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.

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*, pp. 7599–7607, 2023.

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*, pp. 19085–19096, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552, 2018.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 694–711. Springer, 2016.

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*, pp. 8176–8185, 2024.

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

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 204–219, 2022.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. In *International Conference on Learning Representations*, 2025.

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*, 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 *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8580–8589, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. *arXiv preprint arXiv:2501.18993*, 2025.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems*, 37:84839–84865, 2024.

Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on. In *International Conference on Learning Representations*, 2025.

Zhenchen Wan, Ce Chen, Runqi Lin, Jiaxin Huang, Tianxi Chen, Yanwu Xu, Tongliang Liu, and Mingming Gong. Mobile-vton: High-fidelity on-device virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38081–38090, 2026.

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)*, pp. 589–604, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11379–11387, 2026.

Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7017–7026, 2024.

Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024.

Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10511–10520, 2019.

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*, pp. 8372–8382, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2491–2501, 2025.

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*, pp. 4606–4615, 2023.

# A APPENDIX

## A.1 TRAINING DETAILS

We train the VAR model following the configuration of the official Infinity (Han et al., 2025) repository. For the refiner, we select 2,000 detail-rich image pairs from VITON-HD. The refiner is optimized using MSE loss $\mathcal{L}_{mse}$, perceptual loss (Johnson et al., 2016) $\mathcal{L}_p$, adversarial loss $\mathcal{L}_{adv}$, and Discrete Fourier Transform (DFT) loss $\mathcal{L}_{fft}$. The overall objective is formulated as:

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{mse} + \lambda_p\mathcal{L}_p + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{fft}\mathcal{L}_{fft}\,, \tag{9}$$

where $\lambda_{mse}$, $\lambda_p$, and $\lambda_{adv}$ are set to 1, and $\lambda_{fft}$ is set to 0.01.

To simulate the information loss introduced by latent compression, we first compress and reconstruct the images using VQ-VAE at a resolution of $512 \times 384$, and then upsample them to $1024 \times 768$ for refiner training. During inference, imperfect detail generation may cause mismatches between the warped garment patterns and the synthesized images. To improve robustness against such inconsistencies, we first train the refiner on aligned data and subsequently fine-tune it using training samples with slight distortions applied to the warped garment patterns.

## A.2 MORE EXPERIMENTAL RESULTS

**Additional paired evaluation with SSIM and LPIPS.** We report SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) under the paired setting in Table 6. MobileVTON achieves the best SSIM and LPIPS scores. However, despite these favorable scores, MobileVTON shows weaker preservation of fine-grained garment details in Fig. 4 and Fig. 8, particularly for text, logos, and local patterns. This discrepancy suggests that SSIM and LPIPS mainly capture overall structural or perceptual similarity and are less sensitive to localized high-frequency details, further motivating garment-region CLIP-I as our primary metric for garment-detail fidelity.

Table 6: Supplementary evaluation on paired VTON settings using structural similarity (SSIM) and perceptual similarity (LPIPS). These metrics serve as complementary evaluation to the main results reported in Table 1.

| Method | VITON-HD | | DressCode | |
|---|---|---|---|---|
| | SSIM ↑ | LPIPS ↓ | SSIM ↑ | LPIPS ↓ |
| StableVITON (Kim et al., 2024) | <u>0.890</u> | 0.077 | - | - |
| IDM-VTON (Choi et al., 2024) | 0.869 | <u>0.075</u> | 0.902 | **0.052** |
| OOTDiffusion Xu et al. (2025) | 0.839 | 0.087 | 0.884 | 0.069 |
| CatVTON (Chong et al., 2025) | 0.875 | 0.096 | 0.899 | 0.073 |
| Any2AnyTryon (Guo et al., 2025) | 0.870 | 0.112 | - | - |
| MobileVTON (Wan et al., 2026) | **0.896** | **0.068** | - | - |
| **VAR-VTON (Ours)** | 0.876 | 0.076 | <u>0.903</u> | <u>0.056</u> |
| **STAR-VTON (Ours)** | 0.876 | 0.076 | **0.903** | 0.057 |

**Comparison at $512 \times 384$ resolution.** For fair comparison with prior methods evaluated at lower resolutions, we additionally report results on VITON-HD at $512 \times 384$ resolution in Table 7. VAR-VTON is trained and evaluated directly at $512 \times 384$, while for baselines supporting $1024 \times 768$ generation, we first generate high-resolution results and then downsample them to $512 \times 384$. Even under this setting, VAR-VTON achieves the best overall performance while maintaining substantially lower computational cost and faster inference speed. These results further demonstrate the effectiveness and efficiency of the proposed VAR-based global synthesis framework.

**Additional comparison on DressCode.** We further provide additional qualitative comparisons on the DressCode dataset in Fig. 7. These results suggest that the proposed VAR-based global synthesis framework generalizes effectively beyond VITON-HD and can achieve competitive synthesis quality even for more challenging garments and poses.

**Refiner performance under challenging poses.** We further evaluate the robustness of the proposed pixel-space refiner under challenging poses and partial occlusions. As shown in Fig. 6, when garment regions are partially occluded or distorted due to complex human poses, the refiner is still able to recover high-frequency textures from visible regions by leveraging dense image correspondences between the synthesized result and the source garment. This demonstrates that the refiner does not simply perform local sharpening, but instead relies on matching-informed texture transfer, enabling adaptive restoration of unoccluded garment details even under significant geometric deformation. These results further highlight the generalization capability of the proposed pixel-space refinement mechanism beyond standard frontal or well-aligned cases.

**Compatibility of the refiner.** To further validate the generality of the proposed refiner, we apply it to the outputs of several existing VTON methods, as shown in Fig. 8. The refiner consistently enhances fine-grained garment details, including small logos and textual patterns, while preserving the

Table 7: Quantitative comparison on VITON-HD ($512 \times 384$ resolution)

| Method | Paired | | | | Unpaired | | Steps | TFLOPs | Latency (s/img) |
|---|---|---|---|---|---|---|---|---|---|
| | FID ↓ | KID ↓ | SSIM ↑ | LPIPS ↓ | FID ↓ | KID ↓ | | | |
| StableVITON (Kim et al., 2024) | 6.44 | 0.94 | 0.854 | 0.091 | 11.05 | 3.91 | 50 | 92.8 | 8.91 |
| IDM-VTON (Choi et al., 2024) | <u>5.76</u> | <u>0.73</u> | 0.850 | 0.060 | 9.84 | 1.12 | 30 | 117.4 | 5.85 |
| OOTDiffusion Xu et al. (2025) | 9.31 | 4.09 | 0.819 | 0.088 | 12.41 | 4.69 | <u>20</u> | <u>27.3</u> | <u>1.14</u> |
| CatVTON (Chong et al., 2025) | 5.43 | 0.41 | 0.870 | 0.057 | 9.02 | 1.09 | 50 | 101.0 | 2.59 |
| Any2AnyTryon (Guo et al., 2025) | 6.93 | 0.76 | 0.838 | 0.088 | <u>8.79</u> | <u>0.89</u> | 30 | 1091.4 | 17.91 |
| MobileVTON (Wan et al., 2026) | 5.95 | 1.07 | **0.892** | **0.049** | 9.75 | 1.66 | 28 | - | - |
| **VAR-VTON (Ours)** | **5.22** | **0.28** | <u>0.864</u> | <u>0.054</u> | **8.34** | **0.47** | **10** | **19.9** | **0.78** |

Table 8: Effectiveness of the proposed pixel-space refiner on existing VTON methods. We apply the refiner to multiple baseline models and report **CLIP-I** scores.

| Datasets & Methods | | StableVITON | IDM-VTON | OOTDiffusion | CatVTON | Any2AnyTryon | MobileVTON |
|---|---|---|---|---|---|---|---|
| VITON-HD | Latent-space synthesis | 0.851 | 0.897 | 0.894 | 0.868 | 0.879 | 0.865 |
| | +Pixel-space refinement | 0.923 | 0.920 | 0.922 | 0.922 | 0.916 | 0.919 |
| DressCode | Latent-space synthesis | - | 0.918 | 0.909 | 0.897 | - | - |
| | +Pixel-space refinement | - | 0.940 | 0.926 | 0.921 | - | - |




Figure 6: **Refiner results under challenging poses with partial occlusion.**

global structure generated by the original models. We further provide quantitative results in Table 8, where the proposed refiner is applied to multiple baselines across VITON-HD and DressCode. As shown, the refinement module consistently improves CLIP-I scores across all methods and datasets, demonstrating that the proposed pixel-space refinement is model-agnostic and can serve as a general plug-and-play enhancement for improving garment-detail fidelity.

### A.3 Limitations

Although the proposed STAR-VTON achieves strong efficiency and garment-detail preservation, several limitations remain. In this work we mainly focus on recovering fine-grained textures in pixel space, while the global synthesis stage primarily relies on a standard VAR formulation without dedicated detail-preservation mechanisms inside latent-space generation. As a result, although the proposed VAR-based framework achieves competitive overall synthesis quality, its intrinsic ability to preserve fine-grained textures within latent space is still weaker than diffusion-based models. Importantly, the proposed pixel-space refiner should not be viewed as a replacement for detail preservation during structure synthesis. The effectiveness of the refiner still depends on the quality of the generated try-on image. When the global synthesis stage fails to preserve sufficient garment details, the number and reliability of image matching correspondences decrease accordingly, which can negatively affect the subsequent pixel-space refinement process. Specifically, the refinement process becomes more sensitive to hyperparameters such as the number of warped patches. In some challenging cases, a single refinement stage may also be insufficient for fully recovering complex textures, potentially requiring iterative or multi-stage refinement strategies. We believe that jointly improving latent-space detail preservation and pixel-space refinement is an important direction for future work.

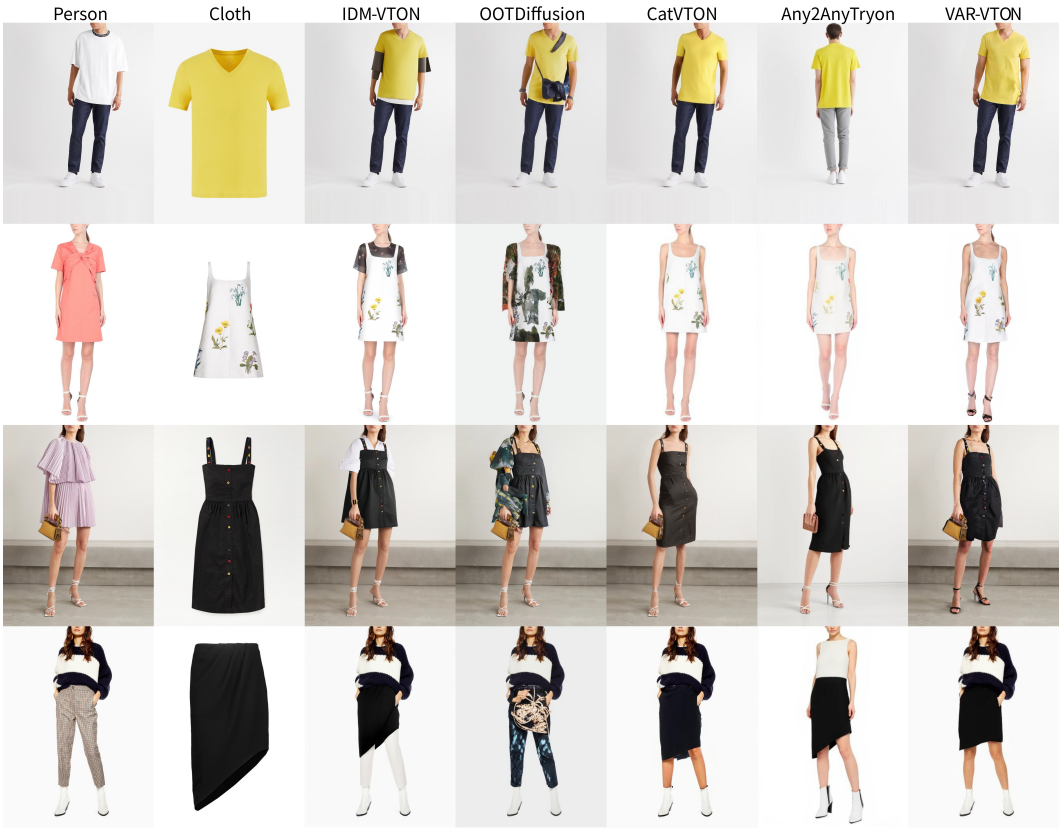


Figure 7: **Qualitative comparison on the DressCode dataset.** Our method produces realistic garment deformation across diverse garment categories.

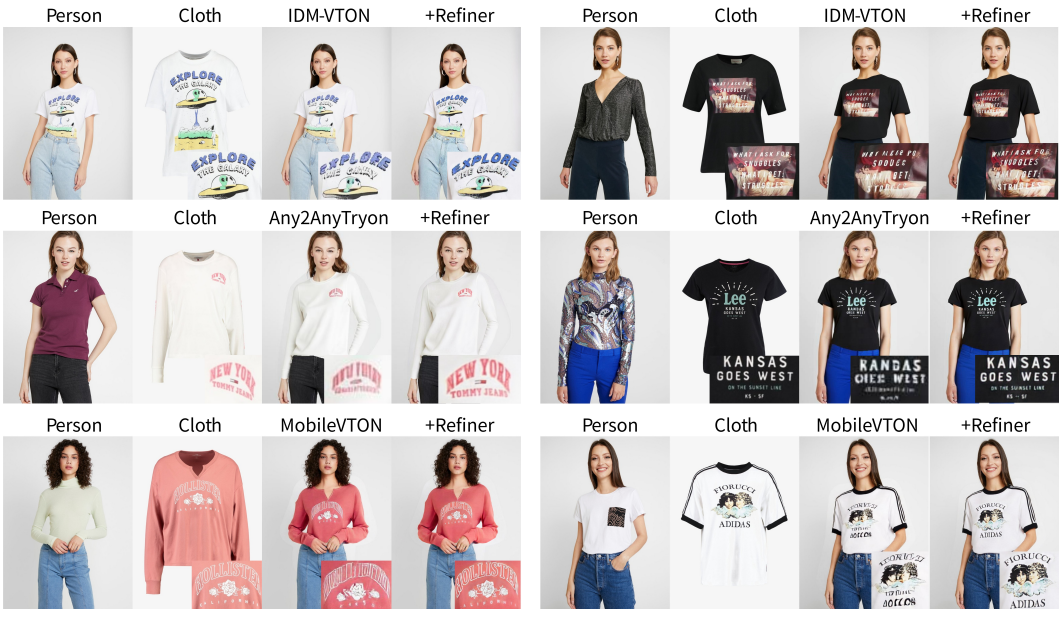


Figure 8: **Compatibility of the proposed refiner with different baseline methods.** The proposed refiner consistently improves the preservation of fine-grained garment textures, while maintaining the overall appearance generated by the original models.