LOOKING BELOW THE LATENT TOKEN: DETAIL RETRIEVAL VIA CORRESPONDENCE SUPERVISION FOR VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-space virtual try-on (VTON) models often distort fine-grained garment details, such as small text, logos, stitching, and repetitive patterns. While existing methods improve garment-person correspondence, such correspondence is defined over spatially coarse latent tokens and therefore provides only coarse localization of reference garment content. We introduce **DECOR-VTON**, a two-phase framework that uses latent correspondence as a spatial routing signal for selectively accessing finer garment information. Through layer-wise analysis and controlled attention interventions, we identify late dual-stream blocks as promising correspondence-aware representations. Phase I improves correspondence in these blocks using probabilistic warp-consistency supervision without requiring dense correspondence annotations or an external correspondence teacher during training. In Phase II, the refined correspondence guides the selection of relevant garment regions, from which higher-resolution convolutional features are retrieved through local cross-attention and injected into downstream transformer blocks via a timestep-conditioned residual gate. This coarse-to-fine design complements latent-space generation with selectively accessed fine-grained reference features without introducing global high-resolution attention. Experiments on VITON-HD and DressCode show consistent improvements in overall synthesis quality, while qualitative comparisons indicate better preservation of fine-grained garment structures. Extensive ablations further demonstrate the complementary roles of correspondence supervision and high-resolution detail retrieval.

1 INTRODUCTION

Image-based Virtual Try-On (VTON) aims to synthesize an image of a person wearing a given garment without physical interaction (Song et al., 2025). Modern diffusion-based VTON models (Morelli et al., 2023; Gou et al., 2023; Kim et al., 2024; Zeng et al., 2024; Choi et al., 2024; Hong et al., 2025; Chong et al., 2024; Kim et al., 2025) have substantially improved synthesis quality across diverse body shapes and poses. However, fine-grained garment details such as small text, logo contours, stitching, and repetitive patterns are still often blurred, distorted, or omitted.

Recent studies improve garment-person correspondence in generative attention to preserve reference appearance (Kim et al., 2024; Zhou et al., 2025; Kim et al., 2026). However, correspondence over coarse latent tokens localizes garment regions without resolving finer structures within them. Rather than adopting computationally expensive pixel-space generation (Yu et al., 2026; Li & He, 2026), we retain pretrained latent-space backbones (Rombach et al., 2022; Black Forest Labs, 2024) and selectively retrieve higher-resolution garment features. Inspired by coarse-to-fine matching and attention-guided recognition (Sun et al., 2021; Hu et al., 2019), latent correspondence determines where to retrieve, while higher-resolution features provide finer appearance information; see Fig. 1.

Reliable detail retrieval depends on selecting an appropriate correspondence source. Prior methods (Kim et al., 2024; Zhou et al., 2025) operate on multi-scale U-Nets (Ronneberger et al., 2015), where feature resolution helps identify correspondence-aware layers. In contrast, Diffusion Transformer (DiT) (Peebles & Xie, 2023) blocks typically operate at the same spatial resolution, making

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.

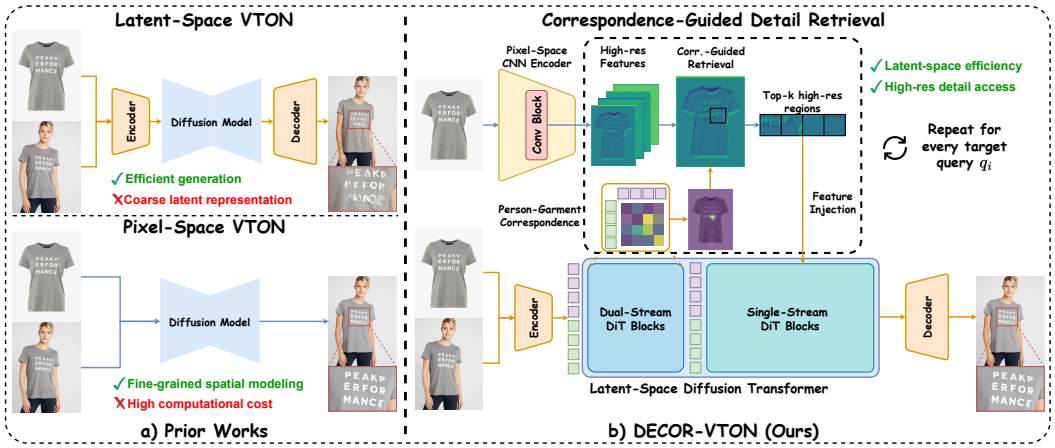

Figure 1: Overview of DECOR-VTON. Latent garment-person correspondence provides spatial routing, while retrieved higher-resolution garment features provide finer reference information.

layer selection less straightforward. We therefore analyze garment-person attention across the DiT backbone and identify late dual-stream blocks as promising correspondence sources.

Based on these observations, we introduce **DECOR-VTON** (**DE**tails retrieval via **COR**respondence supervision for **V**irtual **T**ry-**ON**), a two-phase coarse-to-fine framework. Phase I improves garment–person correspondence through probabilistic warp-consistency supervision. Phase II uses the refined correspondence to select relevant garment regions and retrieve higher-resolution features for downstream generation. Experiments on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) show consistent improvements in overall synthesis quality, while qualitative comparisons indicate better preservation of fine-grained garment structures. Our contributions are:

- We formulate **fine-detail preservation as a coarse-to-fine retrieval problem**, using latent correspondence for spatial routing and higher-resolution features for garment information.
- We analyze **correspondence across DiT blocks** and identify late dual-stream blocks as promising correspondence sources.
- We introduce **probabilistic warp-consistency supervision** to improve correspondence without dense correspondence annotations or an external correspondence teacher during training.
- We propose **correspondence-guided detail retrieval** to selectively retrieve higher-resolution garment features for downstream generation.

2 Related Work

2.1 Image-based Virtual Try-On

Early image-based VTON methods (Han et al., 2018; Raj et al., 2018; Wang et al., 2018) combined explicit garment warping with image synthesis, but warping errors could propagate to the generated results. Diffusion-based methods subsequently improved synthesis quality through various conditioning strategies, including dual-UNet architectures (Choi et al., 2024; Kim et al., 2025; Xu et al., 2025), learned textual representations (Morelli et al., 2023), appearance-flow guidance (Gou et al., 2023; Dam et al., 2024), and ControlNet-style conditioning (Kim et al., 2024; Zeng et al., 2024).

Recent methods further explore flexible conditioning and reference information. Guo et al. (2025) leverages adaptive position embeddings to perform versatile clothing generation tasks. Li et al. (2026) incorporates an additional reference person image to improve garment fidelity, while Yang et al. (2026) addresses inter-task interference in unified fashion generation. Despite these advances, faithfully preserving fine-grained reference appearance remains challenging. Our work approaches this problem by using latent garment-person correspondence as a spatial routing signal for selectively accessing finer reference information.

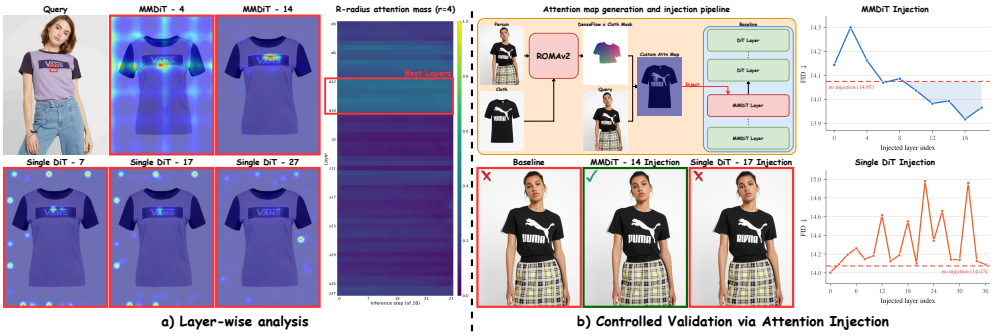

Figure 2: Layer-wise analysis of correspondence-aware attention.

2.2 Correspondence Learning and Attention Supervision

Prior VTON studies show that garment–person correspondence emerges in generative attention, but the lack of dense annotations makes supervision challenging (Kim et al., 2024; Zhou et al., 2025; Kim et al., 2026). Kim et al. (2024) introduce Attention Total Variation (ATV) to encourage smooth correspondence, while Zhou et al. (2025) supervise attention through garment warping. Both methods are developed for multi-scale U-Nets, where feature resolution provides a natural cue for layer selection. This cue is less obvious in transformer-based VTON, whose blocks typically operate at the same resolution. More recently, Kim et al. (2026) align DiT query-key attention with external DINOv3 (Siméoni et al., 2025) correspondences, but require a large pretrained teacher.

Probabilistic Warp Consistency (PWarpC) (Truong et al., 2022) learns dense correspondence by enforcing consistency between a known synthetic warp and composed probabilistic mappings. Inspired by this formulation, we supervise garment–person correspondence encoded in DiT attention layers through geometric consistency, without dense annotations or an external pretrained teacher.

2.3 Coarse-to-Fine Retrieval

Coarse-to-fine processing combines coarse localization with fine-grained feature refinement. Sun et al. (2021) establish coarse correspondences before refining matches using higher-resolution features, while Hu et al. (2019) employ coarse attention to identify informative regions for detailed feature extraction. These works suggest that fine-resolution processing can be restricted to a small set of regions once reliable coarse localization is available. Inspired by this principle, we use latent garment–person correspondence to guide selective retrieval of higher-resolution garment features.

3 Analysis

Unlike multi-scale U-Nets, DiT blocks typically operate at the same spatial resolution, making correspondence-aware layers difficult to identify. We therefore analyze garment–person attention across a VTON-finetuned DiT backbone to identify reliable correspondence sources.

3.1 Layer-wise Correspondence Analysis

We fine-tune FLUX.1 Fill (Black Forest Labs, 2024) for VTON and inspect garment–person attention across transformer blocks. Following Yang et al. (2025), we partition the backbone into early dual-stream (0–9), late dual-stream (10–18), early single-stream (19–36), and late single-stream (37–56) blocks. As shown in Fig. 2a, early dual-stream blocks exhibit diffuse attention, whereas late dual-stream blocks show stronger garment localization. Several single-stream blocks instead attend to background regions or image boundaries. These observations motivate a quantitative analysis of attention concentration across layers.

We measure attention concentration using the r-radius attention mass, inspired by Truong et al. (2021a). Given the head-averaged attention $A_{ij}^{(l,t)}$ from target query i to reference token j at layer l and timestep t, we compute

$$\boldsymbol{\mu}_i^{(l,t)} = \sum_{j=1}^{N_{\text{ref}}} A_{ij}^{(l,t)} \mathbf{c}_j, \qquad C_i^{(l,t)}(r) = \sum_{j:\, \|\mathbf{c}_j - \boldsymbol{\mu}_i^{(l,t)}\|_\infty \le r} A_{ij}^{(l,t)}, \tag{1}$$

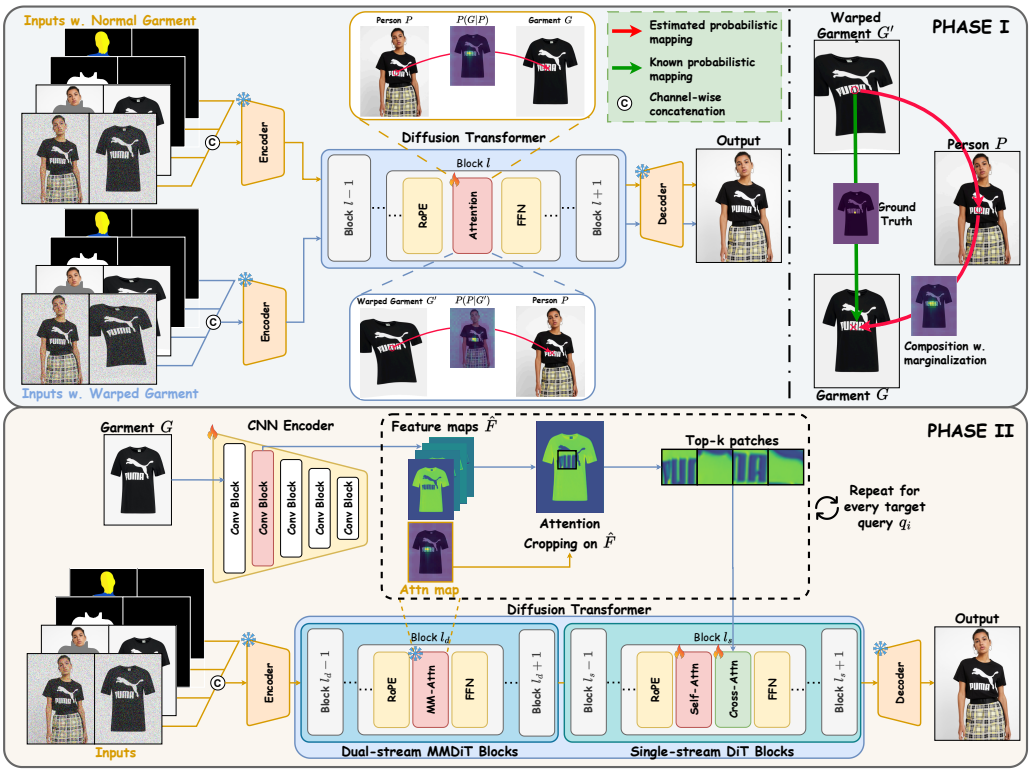

Figure 3: Overview of DECOR-VTON. Phase I enhances garment–person correspondence with probabilistic warp consistency. Phase II uses the learned correspondence to select the top-k garment regions, retrieve fine-grained features from higher-resolution CNN feature maps, and inject them into downstream target tokens.

where $\mathbf{c}_j$ is the reference-token coordinate and $\|\cdot\|_\infty$ denotes the Chebyshev distance. We average the score over target garment-region queries from 50 randomly sampled VITON-HD (Choi et al., 2021) test pairs. With $r = 4$ (a 9×9-token neighborhood), late dual-stream blocks exhibit the highest and most stable concentration across timesteps. Since concentration alone does not guarantee correct localization, we further validate their functional role through controlled attention interventions.

3.2 Controlled Validation via Pseudo-Oracle Attention Injection

The r-radius analysis measures attention concentration but cannot establish correspondence accuracy. We therefore perform controlled attention interventions to assess the functional role of individual layers, using RoMa v2 (Edstedt et al., 2025) to estimate dense correspondence between the reference garment and target garment region. The estimated flow is converted into a pseudo-oracle attention map and injected into individual transformer blocks during inference, with all other components unchanged. We evaluate the intervention effects on 500 VITON-HD test cases manually selected for accurate flow and garment warping. As shown in Fig. 2b, only late dual-stream blocks benefit from attention injection, supporting their functional role in garment–person correspondence.

4 Methodology

DECOR-VTON is motivated by the observation that accurate pixel-level retrieval relies on reliable garment-person correspondence. We therefore adopt a two-phase training strategy. In Phase I, we improve the correspondence encoded by the attention layers using an additional correspondence objective. In Phase II, the refined correspondence guides details retrieval from the reference garment, and we inject the retrieved features into the target representation to refine fine-grained details.

4.1 Rectified-flow Virtual Try-On Baseline

As shown in Fig. 3, we adopt a transformer-based rectified-flow backbone for VTON. Given a target person image P and a reference garment image G, both images are first encoded into latent

representations using a pretrained autoencoder. We spatially concatenate the person and garment latents to form a joint token sequence, allowing direct interaction between the two domains through transformer attention. Additional conditioning inputs, including the cloth-agnostic mask P_{mask}, cloth-agnostic image P_{agn}, and DensePose P_{pose}, are processed in the same manner. For conditions defined only on the person side, we replace the corresponding garment-side input with an all-zero image before concatenation.

4.2 Phase I: Correspondence Supervision via Probabilistic Warp Consistency

To train VTON attention as a probabilistic correspondence representation, we construct a warp-consistency graph (Truong et al., 2021b; 2022) over the target person, reference garment, and a synthetically warped garment. Unlike standard semantic matching, VTON inputs are asymmetric and expose correspondence only in the person-to-garment direction. We overcome these asymmetries by exploiting two properties specific to VTON.

Identity-Preserving Triplet Construction. Given a reference garment G and a randomly sampled mapping M_W, we create a warped copy $G' = G \circ M_W$. We assume that moderate geometric deformation preserves garment identity. Therefore, G and G' are expected to yield the same underlying try-on target on a shared person P. The shared target P thus serves as a common intermediate domain, forming the triplet (G, G', P) and the corresponding warp-consistency graph shown in Fig. 3.

Bidirectional Probabilistic Correspondence from Attention. We use uppercase letters to denote image domains and lowercase letters to denote their token locations. Following the probabilistic mapping convention of (Truong et al., 2022), $\mathcal{P}_{A\leftarrow B}(a \mid b)$ denotes the probability that token b in domain B corresponds to token a in domain A. At timestep t, we extract from the l-th transformer block the person-garment affinity matrix

$$S^{l,t}_{P,G} = \frac{Q^{l,t}_P \left(K^{l,t}_G\right)^\top}{\sqrt{d}}, \tag{2}$$

where $Q^{l,t}_P$ and $K^{l,t}_G$ denote the query features of person tokens and key features of garment tokens. The resulting matrix contains the pairwise affinity scores between the two domains. Normalizing the affinity matrix and its transpose yields the two conditional correspondence distributions:

$$\mathcal{P}^{l,t}_{G\leftarrow P} = \text{Softmax}\left(S^{l,t}_{P,G}\right), \qquad \mathcal{P}^{l,t}_{P\leftarrow G} = \text{Softmax}\left((S^{l,t}_{P,G})^\top\right). \tag{3}$$

where Softmax is applied along the last dimension. Thus, the same query-key affinity can be interpreted as a bidirectional probabilistic correspondence space.

Probabilistic Bipath Supervision. Given the triplet (G, G', P), we construct two input pairs, (P, G) and (P, G'), which share the same target person and differ only in the garment reference. During training, both input pairs are processed by the same model at the same timestep t, forming two parallel branches. The (P, G) branch provides the person-to-garment correspondence $\mathcal{P}^{l,t}_{G\leftarrow P}(g \mid p)$, while the (P, G') branch provides the reverse correspondence $\mathcal{P}^{l,t}_{P\leftarrow G'}(p \mid g')$. We compose these two distributions along the bi-path $G' \to P \to G$ by marginalizing over person tokens within the garment region $\mathcal{M}_P$:

$$\widehat{\mathcal{P}}^{l,t}_{G\leftarrow P\leftarrow G'}(g \mid g') = \sum_{p\in\mathcal{M}_P} \mathcal{P}^{l,t}_{G\leftarrow P}(g \mid p)\mathcal{P}^{l,t}_{P\leftarrow G'}(p \mid g'). \tag{4}$$

We supervise the composed distribution only at valid warped-garment tokens:

$$\mathcal{L}^{l,t}_{\text{bi}} = \sum_{g'\in\mathcal{V}} \mathcal{H}\left(\widehat{\mathcal{P}}^{l,t}_{G\leftarrow P\leftarrow G'}(\cdot \mid g'), \mathcal{P}_W(\cdot \mid g')\right), \tag{5}$$

where $\mathcal{V}$ contains valid tokens within the warped garment mask. Since VTON provides explicit garment masks, we directly use them to restrict the bipath composition and supervision. We aggregate the bipath objective over the selected correspondence-aware layers $\mathcal{L}_c$:

$$\mathcal{L}_{\text{corr}} = \frac{1}{|\mathcal{L}_c|} \sum_{l\in\mathcal{L}_c} \mathcal{L}^{l,t}_{\text{bi}}. \tag{6}$$

The final Phase-I training objective is defined as

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{RF}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}}, \tag{7}$$

where $\mathcal{L}_{\text{RF}}$ denotes the standard rectified-flow training objective.

4.3 Phase II: Detail Retrieval from High-Resolution Garment Features

Although Phase I improves garment-person correspondence, the resulting attention still operates at the latent-token resolution and therefore cannot resolve fine-grained structures within each garment token. In Phase II, we use the refined correspondence to first localize relevant garment regions and then retrieve finer details from higher-resolution convolutional features, as illustrated in Fig. 3.

Correspondence-Guided Region Selection. For each target person token p, we use the attention distribution from the correspondence-aware layers to select the k most relevant garment tokens:

$$\Omega_p = \mathrm{TopK}\left(\mathcal{P}_{G\leftarrow P}(\cdot \mid p),\, k\right), \tag{8}$$

where Ω_p denotes the set of selected garment-token indices. Rather than attending to the entire high-resolution garment representation, we use these tokens only as coarse spatial anchors.

Intra-Token Detail Retrieval. Let $\hat{F} \in \mathbb{R}^{H_f \times W_f \times d_f}$ denote the high-resolution garment feature map. For each target token p, the top-k garment indices Ω_p determine k local feature windows, which are gathered into a candidate set $\mathcal{F}_p$. Fine-grained garment information is then retrieved through local cross-attention:

$$R_p^l = W_o^l\ \mathrm{Attn}\left(Q^l(h_p^l), K^l(\mathcal{F}_p), V^l(\mathcal{F}_p)\right). \tag{9}$$

Detail Injection. The retrieved feature is injected into the corresponding target token through a timestep-conditioned residual gate:

$$\widetilde{h}_p^l = h_p^l + \alpha_l(t) R_p^l, \qquad \alpha_l(t) = \tanh\left(f_{\mathrm{gate}}^l(e_t)\right), \tag{10}$$

where e_t denotes the time embedding. The top-k correspondence is computed once from a selected dual-stream block and shared across the downstream single-stream injection blocks. Phase II is optimized using only $\mathcal{L}_{\mathrm{RF}}$, allowing the retrieval and injection modules to be learned end-to-end from the generation objective. Additional details are provided in section B of the Appendix.

5 Experiments

5.1 Experiment Settings

Implementation Details. We build DECOR-VTON on FLUX.1 Fill (Black Forest Labs, 2024), removing text conditioning and operating only on image tokens. Most experiments are conducted at 512×384 resolution, while an additional 1024×768 variant is trained for resolution-matched comparison with recent high-resolution VTON baselines. We first fine-tune the backbone for VTON for 50K steps. Phase I further trains the correspondence-aware attention modules for 20K steps using eq. (7) with $\lambda_{\mathrm{corr}} = 10^{-3}$. Correspondence supervision is applied only at normalized flow times $t < \theta_t$, with $\theta_t = 0.5$. In Phase II, the dual-stream blocks are frozen, while the single-stream attention and newly introduced retrieval modules are optimized for 50K steps. For detail retrieval, we use VGG-16 (Simonyan & Zisserman, 2014) features at twice the spatial resolution of the latent representation along each dimension and select the top-k garment tokens with $k = 4$ for each target token. Further training details are provided in section B of the Appendix.

Dataset. We evaluate our method on two public VTON benchmarks, VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). VITON-HD contains 13,679 frontal-view upper-body person-garment pairs, split into 11,647 training and 2,032 test pairs. DressCode contains 48,392 training pairs and 5,400 test pairs across three garment categories: upper-body, lower-body, and dresses.

Evaluation Metrics. Following prior work, we report FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020) for paired evaluation, and FID/KID for unpaired evaluation. We compute FID/KID using Clean-FID (Parmar et al., 2022), SSIM/LPIPS using TorchMetrics (Detlefsen et al., 2022), and DISTS using its official implementation. We additionally evaluate correspondence using PCK@16, the percentage of attention-derived matches within 16 pixels of DINOv3-derived (Siméoni et al., 2025) pseudo-ground-truth correspondences. Further details are provided in section B of the Appendix.

Compared Methods. We compare DECOR-VTON with eight recent VTON methods, including StableVITON (Kim et al., 2024), CAT-DM (Zeng et al., 2024), CatVTON (Chong et al., 2024), ITA-MDT (Hong et al., 2025), Leffa (Zhou et al., 2025), Any2Any (Guo et al., 2025), RefVTON (Li et al., 2026), and OrthoTry-On (Yang et al., 2026).

Table 1: Quantitative comparison on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). The best performance for each metric is indicated in **bold**, and the second-best is underlined.

Method	Resolution	VITON-HD					DressCode				
		FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
StableVITON (Kim et al., 2024)	512 × 384	7.06/10.36	2.34/2.94	0.881	0.060	0.093	-	-	-	-	-
CAT-DM (Zeng et al., 2024)	512 × 384	9.58/10.29	9.11/9.41	0.847	0.078	0.118	-	-	-	-	-
CatVTON (Chong et al., 2024)	512 × 384	5.43/8.91	0.41/1.04	0.870	0.069	0.094	5.12/7.64	1.02/2.28	0.875	0.083	0.118
ITA-MDT (Hong et al., 2025)	512 × 512	5.92/9.24	1.38/1.74	0.874	0.057	0.082	3.82/5.78	1.60/1.99	0.915	0.034	0.082
Leffa (Zhou et al., 2025)	1024 × 768	4.77/9.03	**0.11**/1.25	0.870	0.048	0.070	2.79/**4.68**	0.39/**0.69**	0.909	0.041	0.071
Any2Any (Guo et al., 2025)	1024 × 768	6.93/8.97	0.74/0.98	0.839	0.091	0.116	3.15/5.72	1.00/1.80	0.869	0.042	0.078
RefTon (Li et al., 2026)	1024 × 768	7.56/9.11	1.93/0.91	0.881	0.056	0.094	3.48/7.01	1.20/3.15	**0.947**	0.060	0.095
OrthoTryOn (Yang et al., 2026)	512 × 384	10.90/12.74	6.12/5.80	0.851	0.083	0.116	-	-	-	-	-
DECOR-VTON	512 × 384	4.61/8.73	0.21/0.78	0.895	**0.040**	0.065	2.34/5.11	0.21/1.38	0.923	**0.029**	0.065
DECOR-VTON	1024 × 768	**4.49/8.48**	0.13/**0.62**	**0.904**	0.049	**0.063**	**2.28**/4.85	**0.20**/1.12	0.926	0.034	**0.057**



Figure 4: Qualitative comparison on VITON-HD (Choi et al., 2021). Enlarged regions highlight fine-grained garment details, including text, local motifs, and repetitive patterns.

5.2 Quantitative Comparison

Tab. 1 reports the quantitative comparison with existing VTON methods. On VITON-HD, DECOR-VTON achieves the best paired/unpaired FID of 4.49/8.48 at 1024×768, together with the highest SSIM (0.904), lowest unpaired KID (0.62), and lowest DISTS (0.063). The 512×384 variant achieves the lowest LPIPS (0.040). Compared with Leffa, DECOR-VTON improves both paired and unpaired FID, SSIM, and DISTS at the same resolution. On DressCode, our high-resolution model achieves the best paired FID (2.28), paired KID (0.20), and DISTS (0.057), while ranking second in unpaired FID (4.85), unpaired KID (1.12), and SSIM (0.926). Our lower-resolution variant achieves the lowest LPIPS (0.029). Furthermore, Phase I alone reduces paired/unpaired FID from 4.89/8.96 to 4.81/8.87, demonstrating improvements even before high-resolution detail retrieval.

5.3 Qualitative Comparison

Fig. 4 compares DECOR-VTON with prior methods on challenging unpaired VITON-HD (Choi et al., 2021) examples featuring text, local motifs, and repetitive patterns. While competing methods generally recover the overall garment structure, they often blur or distort fine details. In the first row, several methods correctly locate the logo but deform its contours and accompanying text, whereas DECOR-VTON better preserves the reference typography and logo structure. In the second row, our method retains distinct floral motifs that competing methods tend to blur or merge. In the third row, DECOR-VTON better maintains the continuity and spacing of high-frequency stripes. These comparisons demonstrate more faithful preservation of local garment details. Additional qualitative results are provided in section D of the Appendix.



Figure 5: Effect of correspondence supervision on VITON-HD. (a) Generated outputs with correspondence and warped-garment visualizations. (b) Attention maps for the same selected query. (c) Dense correspondences for the same target query points.

Table 2: Ablation of correspondence supervision on VITON-HD. Best and second-best results are shown in **bold** and underlined, respectively.

Method	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓	PCK@16 ↑
Baseline	4.89/8.96	0.19/0.87	**0.890**	**0.044**	0.069	0.885
Baseline + ATV Loss	4.98/8.98	**0.12**/0.81	0.887	0.046	0.072	0.784
Baseline + Leffa Loss	4.96/8.94	0.23/**0.81**	0.877	0.051	0.073	0.879
DECOR-VTON (Phase I)	**4.81/8.87**	0.13/0.83	**0.890**	**0.044**	**0.068**	**0.980**

Table 3: Ablation of two-phase training on VITON-HD. Best and second-best results are shown in **bold** and underlined, respectively.

Phase I	Phase II	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
✗	✗	4.89/8.96	0.19/0.87	0.890	0.044	0.069
✓	✗	4.81/8.87	**0.13**/0.83	0.890	0.044	0.068
✗	✓	4.74/8.93	0.22/0.91	0.892	0.042	0.067
✓	✓	**4.61/8.73**	0.21/**0.78**	**0.895**	**0.040**	**0.065**



Figure 6: Effect of two-phase training. (a) Try-on results. (b) Retrieved regions for the same query tokens. Colors denote corresponding query–retrieval pairs.

5.4 Ablation Studies

We conduct ablation studies on VITON-HD to evaluate the contributions of correspondence supervision and detail retrieval, as well as the key design choices in both phases of DECOR-VTON. All experiments use the same backbone, training data, and evaluation protocol. Within each design ablation, we vary one factor while keeping the remaining settings fixed. Complete quantitative results and experiment-specific configurations are provided in section C of the Appendix.

Effect of Phase-I Correspondence Supervision. Prior VTON methods introduced correspondence-aware objectives such as ATV (Kim et al., 2024) and Leffa (Zhou et al., 2025). To isolate the effect of our Phase-I supervision, we reimplement both under the same backbone and training protocol. As shown in Tab. 2, our method achieves the best FID and DISTS, matches the best SSIM and LPIPS, and improves PCK@16 from 0.885 to 0.980 over the baseline. Fig. 5 further illustrates more coherent correspondence and sharper attention. Compared with alternative objectives, our supervision produces less distorted warps, better preserving garment shape and fine details with fewer mismatched and crossing correspondences.

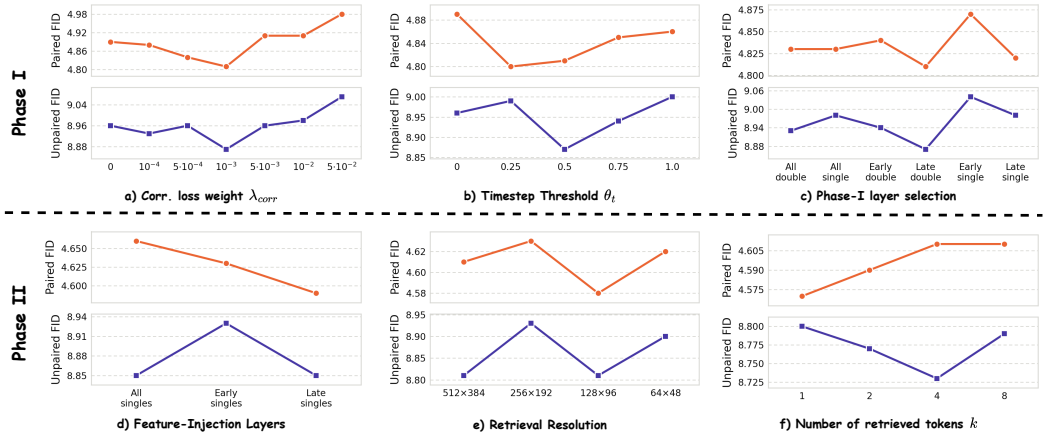

Figure 7: Ablation of key design choices on VITON-HD. Paired and unpaired FID are reported for (a) correspondence-loss weight λ_{corr}, (b) timestep threshold θ_t, (c) Phase-I layer selection, (d) feature-injection layers, (e) retrieval resolution, and (f) the number of retrieved tokens k.

Effect of two-phase training. Tab. 3 shows that Phase I and Phase II provide complementary gains, while the full model achieves the strongest overall performance. As illustrated in Fig. 6, training Phase II without Phase I leads to less consistent garment retrieval for selected query tokens. Such inaccurate retrieval can introduce irrelevant features and noise, which may overwrite or obscure fine garment details. In contrast, the full model provides more coherent retrieval and better preserves local garment structures in the final try-on results, resulting in improved quantitative performance.

Phase-I Design Choices. We examine three Phase-I design choices: the correspondence-loss weight, timestep threshold, and supervised-block selection, as shown in Fig. 7(a)–(c), respectively. Setting $\lambda_{\text{corr}} = 10^{-3}$ achieves the lowest paired and unpaired FID, while larger weights degrade performance. Applying correspondence supervision at timesteps $t < 0.5$ yields the lowest unpaired FID with near-optimal paired performance. Supervising the late dual-stream blocks achieves the lowest FID in both settings, consistent with our layer-wise correspondence analysis.

Phase-II Design Choices. We examine three Phase-II design choices: the feature-injection position, retrieval resolution, and number of selected garment tokens, as shown in Fig. 7(d)–(f), respectively. Injecting retrieved features into late single-stream blocks achieves the lowest paired FID and ties the lowest unpaired FID. Using VGG Block 3 features at 128×96 resolution yields the lowest paired FID and matches the best unpaired result. Selecting $k = 4$ garment tokens achieves the lowest unpaired FID and KID, whereas $k = 1$ performs best on the paired metrics. Accordingly, we adopt late single-stream injection, VGG Block 3, and $k = 4$ in our final model.

Additional ablations on mask-based marginalization, warp-target representation, joint branch optimization, and generalization to other VTON architectures are provided in the Appendix section C.

6 Conclusions

We introduced DECOR-VTON, a two-phase framework for fine-detail preservation in virtual try-on. Our layer-wise analysis identifies late dual-stream blocks as promising garment–person correspondence sources. Phase I improves correspondence through probabilistic warp-consistency supervision without dense annotations or an external teacher. Phase II uses this correspondence to selectively retrieve higher-resolution garment features through local cross-attention. This coarse-to-fine design accesses fine reference details without global high-resolution attention. Extensive experiments and ablations demonstrate the complementary benefits of both phases in preserving text, patterns, and other fine-grained garment details.

Limitations. Nevertheless, detail retrieval remains sensitive to correspondence errors, and the latent-space backbone is still constrained by the reconstruction capacity of its pretrained VAE. These limitations motivate future exploration of uncertainty-aware retrieval and pixel-space refinement. Further discussion of limitations and future directions is provided in section E of the Appendix.

Ethics Statement

This work uses established fashion and human-image datasets and does not collect new personal data or involve human-subject experimentation. We will document the licenses and permitted uses of all training and evaluation sources and will release only artifacts compatible with those terms. Because the task generates images of people, outputs may reproduce identity, body-shape, skin-tone, pose, or garment-category biases present in the data. The model is intended for research and virtual-fashion evaluation; it should not be used for identity impersonation, non-consensual image manipulation, or decisions about people.

AI Use Statement

In this work, we used generative AI tools only for language assistance and manuscript polishing. We reviewed all AI-assisted revisions and take responsibility for the final content of this work, including its text, claims, and artifacts.

References

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

Black Forest Labs. FLUX.1 Fill [dev]. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev, 2024. Hugging Face model repository. 1, 3, 6, 16

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 14

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. 2, 4, 6, 7, 22

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*, pp. 206–235. Springer, 2024. 1, 2

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, 6, 7, 20

Phuong Dam, Jihoon Jeong, Anh Tran, and Daeyoung Kim. Time-efficient and identity-consistent virtual try-on using a variant of altered diffusion models. In *European Conference on Computer Vision*, pp. 35–51. Springer, 2024. 2

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022. 6

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 6

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. Roma v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025. 4

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. 1, 2

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19085–19096. IEEE, 2025. 2, 6, 7

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*, pp. 7543–7552, 2018. 2

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. 6

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 14

Ji Woo Hong, Tri Ton, Trung X Pham, Gwanhyeong Koo, Sunjae Yoon, and Chang D Yoo. Ita-mdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28284–28294, 2025. 1, 6, 7

Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019. 1, 3

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. 1, 2, 3, 6, 7, 8, 20

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16026–16036, 2025. 1, 2

Jiyoung Kim, Youngjin Shin, Siyoon Jin, Dahyun Chung, Jisu Nam, Tongmin Kim, Jongjae Park, Hyeonwoo Kang, and Seungryong Kim. Coral: Correspondence alignment for improved virtual try-on, 2026. URL https://arxiv.org/abs/2602.17636. 1, 3, 16

Liuzhuozheng Li, Yue Gong, Shanyuan Liu, Zanyi Wang, Dengyang Jiang, Leibucha Wu, Bo Cheng, Yuhang Ma, Dawei Leng, and Yuhui Yin. Refton: Reference person shot assist virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14915–14925, 2026. 2, 6, 7

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026. 1

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 14

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. 2, 6, 7, 23

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. 1, 2

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 6

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023. 1

Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *European Conference on Computer Vision*, pp. 679–695. Springer, 2018. 2

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015. 1

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3, 6

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 16

Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, Mohan Kankanhalli, and An-An Liu. Image-based virtual try-on: A survey. *International Journal of Computer Vision*, 133(5): 2692–2720, 2025. 1

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 14

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021. 1, 3

Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5710–5720. IEEE, 2021a. 3

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10326–10336. IEEE, 2021b. 5, 15

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8698–8708. IEEE, 2022. 3, 5, 15

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. 2

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. 6

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*, pp. 8996–9004, 2025. 2

Yitong Yang, Yinglin Wang, Changshuo Wang, Yongjun Zhang, Ziyang Chen, and Shuting He. Splitflux: Learning to decouple content and style from a single image. *arXiv preprint arXiv:2511.15258*, 2025. 3, 18

Zhaotong Yang, Ying Tai, Jiahui Zhan, Yu Zheng, Jianjun Qian, and Jian Yang. Orthotryon: Geometric orthogonalization for conflict-free unified fashion generation. *arXiv preprint arXiv:2606.27880*, 2026. 2, 6, 7

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026. 1

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. 1, 2, 6, 7

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*, pp. 586–595, 2018. 6

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2491–2501. IEEE, 2025. 1, 3, 6, 7, 8

APPENDIX

The supplementary material is organized as follows:

- Section A reviews Flow Matching, Rectified Flow, Warp Consistency, and Probabilistic Warp Consistency.
- Section B details the implementation and training of correspondence supervision and high-resolution detail retrieval.
- Section C provides complete ablation results, additional analyses, and experiments on other VTON backbones.
- Section D presents additional qualitative comparisons on VITON-HD and DressCode.
- Section E discusses limitations and future research directions.

A PRELIMINARIES

A.1 FLOW MATCHING AND RECTIFIED FLOW

Flow Matching. Flow Matching (FM) (Lipman et al., 2022) is a family of generative models that combines the advantages of Continuous Normalizing Flows (CNFs) (Chen et al., 2018) and Diffusion Models (Ho et al., 2020; Song et al., 2020). It formulates generative modeling as transporting samples between a simple prior distribution, typically a standard Gaussian, and the data distribution through a continuous ordinary differential equation (ODE). Instead of solving the ODE during training, FM directly learns the velocity field associated with a predefined probability path, enabling simulation-free training. Given a data sample x_0, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and a normalized timestep $t \in [0, 1]$, an intermediate sample is constructed as

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \tag{11}$$

where α_t and σ_t are time-dependent coefficients that define the probability path and control the contributions of the data and noise, respectively. Following the convention that $t = 0$ corresponds to clean data and $t = 1$ to Gaussian noise, the boundary conditions are $\alpha_0 = 1$, $\sigma_0 = 0$, $\alpha_1 = 0$, and $\sigma_1 = 1$. The corresponding target velocity is obtained by differentiating the probability path with respect to the timestep:

$$u_t(x_t) = \frac{dx_t}{dt} = \frac{d\alpha_t}{dt} x_0 + \frac{d\sigma_t}{dt} \epsilon. \tag{12}$$

A neural velocity field $v_\theta(x_t, t)$ is then trained to approximate this target velocity by minimizing

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0, \epsilon, t} \left[\| v_\theta(x_t, t) - u_t(x_t) \|_2^2 \right]. \tag{13}$$

Rectified Flow. Rectified Flow (RF) further simplifies the FM formulation by adopting a linear probability path between the data sample and Gaussian noise. Given a clean sample x_0, noise $\epsilon \sim \mathcal{N}(0, I)$, and a normalized timestep $t \in [0, 1]$, the intermediate state is defined as

$$x_t = (1 - t)x_0 + t\epsilon. \tag{14}$$

This corresponds to setting $\alpha_t = 1 - t$ and $\sigma_t = t$ in the general Flow Matching formulation. Consequently, the target velocity becomes independent of t,

$$u_t = \frac{dx_t}{dt} = \epsilon - x_0. \tag{15}$$

The velocity network $v_\theta(x_t, t)$ is therefore trained with the objective

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{x_0, \epsilon, t} \left[\| v_\theta(x_t, t) - (\epsilon - x_0) \|_2^2 \right]. \tag{16}$$

By using a linear interpolation between the two endpoint distributions, RF provides a simple velocity-prediction objective and encourages straighter transport trajectories. At inference time, generation starts from Gaussian noise and follows the learned ODE toward the data distribution.

A.2 Warp Consistency and Probabilistic Warp Consistency

Warp Consistency. Warp Consistency (Truong et al., 2021b) is an unsupervised objective for learning dense correspondences without ground-truth annotations. Given a real image pair (I, J), a synthetic image I' is generated by applying a randomly sampled warp W to I, forming the triplet (I, I', J). The resulting consistency graph yields the W-bipath constraint, which enforces agreement between the known direct mapping $I' \to I$ and the composed path $I' \to J \to I$. This allows the known synthetic warp to supervise the unknown correspondence between the real image pair (I, J). Warp Consistency represents each correspondence as a deterministic 2D mapping.

Probabilistic Warp Consistency. Probabilistic Warp Consistency (PWarpC) (Truong et al., 2022) extends this formulation by representing each correspondence as a conditional probability distribution. Given a pixel j in J, its correspondence distribution over image I is

$$P_{I\leftarrow J}(i \mid j) = \frac{\exp\left(C_{I\leftarrow J}(i,j)\right)}{\sum_k \exp\left(C_{I\leftarrow J}(k,j)\right)}. \tag{17}$$

The estimated correspondence along the composed path $I' \to J \to I$ is obtained by marginalizing over intermediate locations j in J:

$$\widehat{P}_{I\leftarrow J\leftarrow I'}(i \mid i') = \sum_j P_{I\leftarrow J}(i \mid j) P_{J\leftarrow I'}(j \mid i'). \tag{18}$$

The composed distribution is then supervised to match the known probabilistic mapping $P_W(\cdot \mid i')$ induced by the synthetic mapping M_W. In our method, we adapt this formulation to attention-based correspondences by interpreting attention-derived affinities as a matching-cost space and composing the resulting conditional correspondence distributions along the bipath.

B Additional Implementation Details

B.1 Sampling of Warps M_W

Following the transformation-sampling strategy of PWarpC (Truong et al., 2022), we construct M_W by uniformly selecting among homography, Thin-Plate Spline (TPS), and affine-TPS transformations. We adopt milder transformation ranges for VTON. Specifically, $\sigma_H = 0.25$ controls the displacement range of the four image corners for homography and the 3×3 control points for TPS, while $\sigma_{\text{tps}} = 0.25$ controls the TPS deformation used in the affine-TPS transformation. For the affine component, $\tau = 0.25$ determines the scale range $[1 - \tau, 1 + \tau]$, the translation magnitude is sampled within $[-t_a, t_a]$ with $t_a = 0.15$, and $\alpha = \pi/18$ specifies the maximum rotation and shear angles, both sampled from $[-\alpha, \alpha]$. All spatial displacements are sampled in normalized coordinates. We additionally compensate the displacement magnitude for the non-square image aspect ratio. Unlike PWarpC, we disable horizontal flipping because it reverses orientation-sensitive garment details such as text and logos.

B.2 Creation of Ground-Truth Probabilistic Mapping $\mathcal{P}_W$

Following PWarpC (Truong et al., 2022), we adopt two representations to construct the ground-truth probabilistic mapping $\mathcal{P}_W$ from the known mapping M_W. We first resize M_W to the latent-token resolution and convert each warped-token location into its corresponding continuous coordinate on the original garment-token grid. For the one-hot representation, the full probability mass is assigned to the nearest garment token. For the smooth representation, the probability mass is first distributed among the four neighboring tokens according to their bilinear interpolation weights, assigning larger probability to spatially closer tokens. We then apply a 3×3 Gaussian kernel and renormalize the resulting map to obtain a valid conditional distribution. Tokens whose mapped coordinates fall outside the reference token grid are excluded from supervision. Unless otherwise specified, the smooth representation uses a Gaussian standard deviation of $\sigma = 0.7$, measured in token-grid cells. We use the one-hot representation as the default target for our probabilistic bipath objective and evaluate the smooth representation in section C.3.

B.3 Architecture and Training Hyperparameters

All experiments are implemented based on FLUX.1 Fill (Black Forest Labs, 2024) and trained in bf16 using AdamW with a learning rate of 1×10^{-5} and a constant learning rate schedule. The VTON baseline is first optimized for 50K steps using the rectified-flow objective. Phase I is then trained for an additional 20K steps with $\lambda_{\text{corr}} = 10^{-3}$. We express the timestep as a normalized variable $t \in [0, 1]$ and apply the correspondence objective only when $t < \theta_t$, with $\theta_t = 0.5$. During Phase I, only the attention modules are updated, while the correspondence objective is applied exclusively to the selected late dual-stream blocks identified by our layer-wise analysis.

In Phase II, all dual-stream blocks are frozen, and we optimize the single-stream self-attention modules together with the newly introduced retrieval cross-attention modules for 50K steps using a batch size of 8. Garment features are extracted with VGG-16 (Simonyan & Zisserman, 2014). The retrieval resolution is defined relative to the latent representation rather than fixed to an absolute feature-map size: for a latent grid of $H_L \times W_L$, retrieval is performed on a $2H_L \times 2W_L$ feature grid. Thus, the retrieval representation contains four times as many spatial locations as the latent representation while preserving the same relative configuration across input resolutions. For each target token, the top-k garment tokens are selected according to the learned correspondence distribution, with $k = 4$ by default. Their associated high-resolution feature regions are gathered and used as keys and values for local cross-attention. Unless otherwise specified, experiments are conducted at 512×384 resolution. We additionally train a 1024×768 variant using the same relative retrieval configuration for resolution-matched comparison with recent high-resolution VTON methods.

All experiments are conducted on a single NVIDIA B200 GPU. At 512×384 resolution, Phase I and Phase II require approximately 17 and 21 hours of training, respectively. Inference uses 28 timesteps. Unless otherwise specified, all experiments share the same hyperparameters and architectural settings. To improve reproducibility, we fix the random seed to 42 and enable deterministic algorithms in PyTorch to reduce run-to-run variability.

B.4 Extraction of Garment-Person Correspondence from Attention Maps

Equation (3) in the main paper presents a simplified formulation of garment–person correspondence, while its exact extraction from attention depends on the backbone architecture. Here, we detail the extraction procedure for our DiT backbone, which processes person and reference garment tokens jointly within the same attention sequence.. Similar to (Kim et al., 2026), the query and key representations at layer l and timestep t can therefore be written as:

$$Q^{l,t} = [Q_P^{l,t} \,\|\, Q_G^{l,t}], \qquad K^{l,t} = [K_P^{l,t} \,\|\, K_G^{l,t}], \tag{19}$$

where P and G denote the person and garment views, respectively. This joint attention naturally contains interactions both within and across the two views. However, rather than using a clothing mask to pre-select which person queries and garment keys participate in the attention computation like CORAL, we first preserve the original attention normalization over the complete key sequence. For each attention head h, the native attention distribution is

$$A_{p,u}^{l,t,h} = \text{Softmax}_u \left(\frac{(q_p^{l,t,h})^\top k_u^{l,t,h}}{\sqrt{d}} \right), \tag{20}$$

where u indexes all keys, and d is the head dimension. We then average the attention probabilities across heads,

$$\bar{A}_{p,u}^{l,t} = \frac{1}{H} \sum_{h=1}^{H} A_{p,u}^{l,t,h}. \tag{21}$$

To obtain the conditional person-to-garment correspondence, we retain the entries associated with the garment view and renormalize their probability mass:

$$\mathcal{P}_{G \leftarrow P}^{l,t}(g \mid p) = \frac{\bar{A}_{p,g}^{l,t}}{\sum_{g' \in \mathcal{G}} \bar{A}_{p,g'}^{l,t}}, \qquad g \in \mathcal{G}, \tag{22}$$

where $\mathcal{G}$ denotes the garment-view token set. The reverse distribution $\mathcal{P}_{P \leftarrow G}^{l,t}(p \mid g)$ is constructed analogously by retaining the person-view probability mass and renormalizing over person tokens.

This ordering preserves the original attention behavior before conditioning on the opposite view. In particular, attention heads that assign little probability mass to the garment contribute correspondingly less to the resulting garment distribution, rather than being forced to form an independently normalized correspondence over garment tokens. Clothing masks are therefore not used to define the attention correspondence itself, but are introduced only where required by the Phase-I probabilistic bipath objective.

B.5 Feature Retrieval and Injection

At each denoising timestep t, garment–person correspondence is computed once from a selected late dual-stream block ℓ_r and shared across all downstream injection layers. Following section 4, for each target token p, we select the top-k garment tokens:

$$\Omega_p = \mathrm{TopK}\left(\mathcal{P}^{\ell_r,t}_{G\leftarrow P}(\cdot \mid p), k\right). \tag{23}$$

These selected tokens serve as coarse spatial anchors for gathering local windows from the higher-resolution garment feature map $\hat{F}$. The retrieved feature set is constructed as

$$\mathcal{F}_p = \mathrm{Concat}_{g\in\Omega_p}\,\mathrm{Flatten}\left(\mathrm{Window}(\hat{F}, g)\right), \tag{24}$$

where each window contains $m \times m$ feature locations, yielding km^2 features per target token.

To preserve the spatial structure within each retrieved window, we apply window-relative rotary positional embeddings (RoPE) to the projected keys. For a feature location (u, v) within a window, its relative position is defined as

$$\mathbf{\Delta}_{u,v} = s_f\left(u - \frac{m-1}{2}, v - \frac{m-1}{2}\right), \tag{25}$$

where s_f denotes the feature-map stride in image pixels. Each window uses its own center as the positional origin, preserving local spatial relationships without introducing positional offsets between retrieved windows.

At injection layer ℓ, the target representation h^ℓ_p provides the query, while the retrieved features $\mathcal{F}_p$ provide the keys and values:

$$\begin{aligned} Q^\ell_p &= Q^\ell(h^\ell_p), \\ K^\ell_p &= \mathrm{RoPE}_{\mathbf{\Delta}}\left(K^\ell(\mathcal{F}_p)\right), \\ V^\ell_p &= V^\ell(\mathcal{F}_p), \\ R^\ell_p &= W^\ell_o\,\mathrm{Softmax}\left(\frac{Q^\ell_p (K^\ell_p)^\top}{\tau_\ell\sqrt{d}}\right) V^\ell_p, \end{aligned} \tag{26}$$

where τ_ℓ is a learnable attention temperature and d is the attention-head dimension. RoPE is applied only to the keys, while queries and values remain unrotated.

Following the timestep-conditioned residual gating in section 4.3, the retrieved feature is added only to the corresponding target attention output:

$$\widetilde{h}^\ell_p = h^\ell_p + \alpha^\ell(t) R^\ell_p, \qquad \alpha^\ell(t) = \tanh\left(f^\ell_{\mathrm{gate}}(e_t)\right). \tag{27}$$

Garment tokens are not directly modified by the injection branch. All downstream injection layers reuse the same selected garment indices Ω_p, while maintaining independent feature projections, attention temperatures, and gating modules.

C Additional Analyses

C.1 Full Quantitative Results for Design Ablations

For completeness, we report the full quantitative results corresponding to the design ablations presented in the main paper. In addition to paired and unpaired FID, the tables include KID, SSIM,

Table 4: Ablation of the correspondence-loss weight λ_{corr} on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

λ_{corr}	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
0 (no PWC)	4.89/8.96	0.18/0.86	0.890	0.044	0.069
1×10^{-4}	4.88/**8.93**	0.20/0.86	0.890	0.044	0.069
5×10^{-4}	4.84/8.96	0.17/0.85	0.890	0.044	0.069
1×10^{-3}	**4.81/8.87**	**0.13/0.82**	0.890	0.044	0.068
5×10^{-3}	4.91/8.96	0.19/0.91	0.890	0.044	0.069
1×10^{-2}	4.91/8.98	0.19/0.86	0.890	0.044	0.069
5×10^{-2}	4.98/9.07	0.23/0.99	0.890	0.044	0.070

Table 5: Ablation of the timestep threshold θ_t for Phase-I correspondence supervision on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

θ_t	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
0 (no PWC)	4.89/8.96	0.18/0.86	0.890	0.044	0.069
0.25	**4.80**/8.99	0.17/0.89	0.890	0.044	0.069
0.50	4.81/**8.87**	**0.13/0.82**	0.890	0.044	0.069
0.75	4.85/8.94	0.15/0.85	0.890	0.044	0.069
1.00 (all timesteps)	4.86/9.00	0.15/0.93	0.890	0.044	0.069

Table 6: Ablation of the feature-injection position on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Injection Position	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
All single-stream	4.66/**8.85**	0.24/0.91	0.895	0.039	0.065
Early single-stream	4.63/8.93	0.21/0.91	0.896	0.040	0.066
Late single-stream	**4.59/8.85**	**0.20/0.85**	0.895	0.039	0.065

Table 7: Ablation of Phase-I training layers on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Training Layers	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
All dual-stream	4.83/8.93	0.12/0.88	0.889	0.045	0.070
All single-stream	4.83/8.98	**0.10**/0.90	0.889	0.045	0.070
Early dual-stream	4.84/8.94	0.16/**0.82**	0.890	0.044	0.069
Late dual-stream	**4.81/8.87**	0.13/**0.82**	0.890	0.044	0.068
Early single-stream	4.87/9.04	0.17/0.85	0.890	0.044	0.069
Late single-stream	4.82/8.98	0.16/0.86	0.890	0.044	0.069

Table 8: Ablation of the retrieval resolution on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Retrieval Feature	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
VGG Block 1 (512×384)	4.61/**8.81**	0.21/0.83	0.896	0.039	0.065
VGG Block 2 (256×192)	4.63/8.93	0.21/0.91	0.896	0.040	0.066
VGG Block 3 (128×96)	**4.58/8.81**	**0.19/0.76**	0.896	0.040	0.065
VGG Block 4 (64×48)	4.62/8.90	0.23/0.87	0.896	0.040	0.066

Table 9: Ablation of the number of retrieved garment tokens k on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Top-k	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
1	**4.57**/8.80	**0.18**/0.79	0.896	0.039	0.065
2	4.59/8.77	0.19/0.79	0.895	0.040	0.065
4	4.61/**8.73**	0.21/**0.78**	0.895	0.040	0.065
8	4.61/8.79	0.21/0.81	0.896	0.040	0.065

LPIPS, and DISTS. Each ablation varies the factor of interest while keeping the remaining settings fixed within the corresponding experimental configuration. Since different ablations were evaluated under their respective controlled settings, absolute values across tables are not intended to be directly compared. The configuration adopted in the final model is highlighted in light green. The fixed settings used for each ablation are specified below.

C.1.1 Phase I Ablation Studies

Correspondence-loss weight. Tab. 4 reports the full results for the correspondence-loss weight ablation. We vary λ_{corr} from 0 to 5×10^{-2} while updating only the late dual-stream blocks and fixing the timestep threshold to $\theta_t = 0.5$.

Timestep threshold θ_t. Tab. 5 reports the full results for the timestep-threshold ablation. We update only the late dual-stream blocks, fix $\lambda_{\text{corr}} = 1 \times 10^{-3}$, and vary θ_t from 0 to 1, where the two endpoints correspond to disabling correspondence supervision and applying it across all timesteps, respectively.

Phase I training layers. Tab. 7 reports the full results for the Phase-I training-block ablation. Following the layer partitioning in (Yang et al., 2025), we divide the FLUX backbone into four groups: early dual-stream, late dual-stream, early single-stream, and late single-stream blocks, corresponding to block indices [0–9], [10–18], [19–36], and [37–56], respectively. We fix $\theta_t = 0.5$ and $\lambda_{\text{corr}} = 1 \times 10^{-3}$, and compare the application of correspondence supervision to different block groups, as well as to all dual-stream or single-stream blocks.

C.1.2 Phase II Ablation Studies

Feature-injection position. Tab. 6 reports the full results for the feature-injection position ablation. Following the block partitioning described above, we compare injecting the retrieved features into all, early, or late single-stream blocks. We fix the retrieval features to the second VGG Block and the number of retrieved garment areas to $k = 4$.

Retrieval feature resolution. Tab. 8 reports the full results for the retrieval-resolution ablation. We compare features extracted from four VGG blocks, with spatial resolutions ranging from 512×384 to 64×48. We fix the number of retrieved garment areas to $k = 4$ and apply feature injection to the early single-stream blocks.

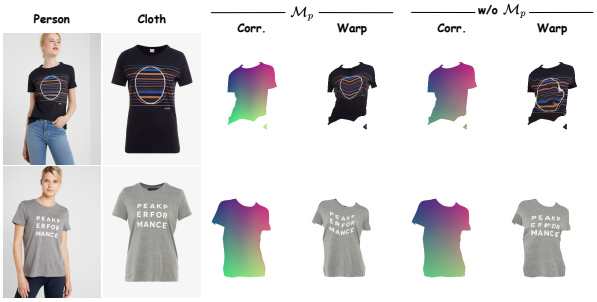

Figure 8: Effect of marginalization over $\mathcal{M}_P$



Figure 9: Comparison of one-hot and smoothed warp targets $\mathcal{P}_W$

Table 10: Ablation of marginalization over $\mathcal{M}_P$ for Phase-I correspondence supervision on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Marginalization	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
With mask	**4.81/8.87**	**0.13/0.82**	0.890	0.044	0.068
Without mask	4.87/8.99	0.18/0.83	0.890	0.045	0.069

Number of retrieved garment tokens k. Tab. 9 reports the full results for the number of retrieved garment areas. We vary $k \in \{1, 2, 4, 8\}$ while fixing the retrieval features to the third VGG Block and injecting the retrieved features into the late single-stream blocks.

C.2 Effect of Marginalization over the Person-Side Garment Mask $\mathcal{M}_P$

Our probabilistic bipath objective marginalizes over person tokens within the target garment region $\mathcal{M}_P$. We evaluate this design by comparing masked and full-image marginalization. The latter includes background tokens that lack meaningful correspondence with the reference garment, potentially introducing irrelevant paths into the composed probabilistic mapping and weakening correspondence supervision. Restricting marginalization to the target garment region improves paired/unpaired FID from 4.87/8.99 to 4.81/8.87 and KID from 0.18/0.83 to 0.13/0.82, as shown in Tab. 10. Qualitative comparisons in Fig. 8 further show that masked marginalization produces more coherent correspondence fields and better garment warps, particularly around printed details. In contrast, full-image marginalization can lead to distorted patterns and less consistent alignment.

C.3 Effect of the $\mathcal{P}_W$ Representation

As described in section B.2, we consider two representations of the ground-truth probabilistic mapping $\mathcal{P}_W$, including one-hot and smoothed targets. As illustrated in Fig. 9, one-hot supervision yields more concentrated attention around corresponding garment locations, whereas smoothed targets produce more diffuse responses that may extend into unrelated regions. This suggests that smoothing can reduce correspondence localization precision. Consistently, one-hot supervision achieves better FID (4.81/8.87 vs. 4.93/9.08) and KID (0.13/0.82 vs. 0.22/0.97), as shown in Tab. 11.

C.4 Effect of Joint Optimization of Both Bipath Branches

Phase I processes the original and synthetically warped garment pairs through two shared-weight branches. We empirically evaluate whether including both branch outputs in the flow-matching loss benefits Phase I training. As shown in Tab. 12, supervising both branches achieves better

Table 11: Ablation of $\mathcal{P}_W$ representations for Phase-I correspondence supervision on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Warp target	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
One-hot	**4.81/8.87**	**0.13/0.82**	0.890	0.044	0.068
Smooth	4.93/9.08	0.22/0.97	0.890	0.045	0.070

Table 12: Ablation of two-branch training for Phase-I correspondence supervision on VITON-HD. The best FID and KID are indicated in **bold**, and the second-best are underlined.

Training strategy	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
Both branches	**4.81/8.87**	**0.13/0.82**	0.890	0.044	0.068
Normal branch only	4.85/8.99	0.14/0.88	0.890	0.045	0.070

Table 13: Effect of our correspondence supervision framework across different VTON backbones on VITON-HD. The best results are indicated in **bold**, and the second-best are underlined.

Method	$\mathcal{L}_{\text{corr}}$	FID (pair/unp.) ↓	KID (pair/unp.) ↓	SSIM ↑	LPIPS ↓	DISTS ↓
CatVTON	✗	5.43/8.91	0.41/1.04	0.870	0.069	0.094
	✓	5.19/8.88	0.19/**0.75**	0.886	0.050	0.074
StableVITON	✗	7.06/10.36	2.34/2.94	0.881	0.060	0.093
	✓	6.66/9.63	1.70/1.82	**0.892**	0.049	0.080
Ours	✗	4.89/8.96	0.18/0.86	0.890	**0.044**	**0.069**
	✓	**4.81/8.87**	**0.13**/0.82	0.890	**0.044**	**0.069**

paired/unpaired FID (4.81/8.87 vs. 4.85/8.99) and KID (0.13/0.82 vs. 0.14/0.88) than supervising only the normal branch, while SSIM remains unchanged. Based on these results, we adopt two-branch training in our framework.

C.5 Generalization of Phase-I Correspondence Supervision to Other VTON Architectures

We further evaluate the applicability of our correspondence supervision by incorporating it into CatVTON (Chong et al., 2024) and StableVITON (Kim et al., 2024), in addition to our DiT backbone. As shown in Tab. 13, the proposed supervision consistently improves paired and unpaired FID and KID across all three models. Specifically, paired/unpaired FID decreases from 5.43/8.91 to 5.19/8.88 for CatVTON, from 7.06/10.36 to 6.66/9.63 for StableVITON, and from 4.89/8.96 to 4.81/8.87 for our backbone. KID exhibits similar improvements, while SSIM, LPIPS, and DISTS either improve or remain unchanged. These results empirically demonstrate the applicability of our correspondence supervision beyond the proposed DiT backbone.

D Additional Qualitative Results

Additional Qualitative Results on VITON-HD. Fig. 10 presents additional unpaired comparisons on VITON-HD, covering challenging typography, geometric patterns, and complex graphic prints. In the first two rows, several competing methods recover the overall placement of printed text but introduce distortions in letter shapes and spacing, whereas DECOR-VTON preserves more legible typography and cleaner contours. The third and fourth rows highlight geometric and repetitive patterns, where our method better maintains the continuity of circular motifs and the spacing of narrow stripes. In the last two rows, DECOR-VTON retains more recognizable graphic elements and accompanying text, while several baselines exhibit blurred or distorted details. These examples further illustrate the benefits of correspondence-guided high-resolution retrieval for preserving diverse garment details.

Additional Qualitative Results on DressCode. Fig. 11 presents additional unpaired comparisons across upper-body garments, lower-body garments, and dresses. In the first two rows, DECOR-VTON reproduces printed logos and text with clearer contours, while several competing methods introduce distortions in typography and graphic details. For lower-body garments, our method preserves the length and spacing of the striped shorts and the fitted silhouette of the buttoned skirt. In contrast, some baselines alter garment lengths or produce inaccurate silhouettes. In the last two rows, DECOR-VTON maintains recognizable floral motifs and the layered patterns of the dresses while preserving their overall structure. These results further illustrate the applicability of correspondence-guided detail retrieval across diverse garment categories.

E LIMITATIONS AND FUTURE WORK

Dependence on Human Priors. DECOR-VTON relies on auxiliary human priors, including DensePose, human parsing, and garment masks. While such dependencies are common among existing VTON methods, errors in these inputs may affect garment localization and the quality of generated results. Future work could explore reducing reliance on explicit human priors to improve robustness in unconstrained scenarios.

Dependence on Correspondence Accuracy. Our high-resolution detail retrieval relies on garment–person correspondence learned during Phase I. Incorrect or ambiguous correspondences, particularly under severe occlusion or repetitive patterns, may lead to the selection of irrelevant garment regions and inaccurate detail transfer. Incorporating correspondence uncertainty into the retrieval process could improve robustness to such errors.

Limitations of Latent-Space Generation. Although DECOR-VTON retrieves high-resolution garment features, it still operates within a latent-space generative framework, where target images are compressed by a pretrained VAE encoder during training. Consequently, fine-grained information lost during VAE compression may not be fully recovered, even with accurate correspondence and high-resolution retrieval. Future work could investigate lightweight pixel-space refinement to preserve high-frequency garment details beyond the reconstruction capacity of the pretrained VAE.



Figure 10: Additional qualitative results on the VITON-HD (Choi et al., 2021) dataset.



Figure 11: Additional qualitative results on the DressCode (Morelli et al., 2022) dataset.