CORAL: Correspondence Alignment for Improved Virtual Try-On

Jiyoung Kim[1] **Youngjin Shin**[1] **Siyoon Jin**[1] **Dahyun Chung**[1] **Jisu Nam**[1]
Tongmin Kim[1] **Jongjae Park**[2] **Hyeonwoo Kang**[2] **Seungryong Kim**[1]

[1]KAIST AI [2]NC AI

https://cvlab-kaist.github.io/CORAL/



(a) **CORAL** Framework (b) Virtual Try-On Results (c) Person-to-Person Results

Figure 1: **Teaser.** (a) CORAL is a Diffusion Transformer-based framework that explicitly enhances person→garment correspondence within the 3D attention. This leads to more accurate local details in the generated results, such as fewer artifacts like duplicated garment hems, as shown in the baseline. (b) Virtual Try-On results on VITON-HD and DressCode, comparing outputs without and with CORAL. (c) Person-to-Person garment transfer results on PPR10K in challenging in-the-wild images.

Abstract

Existing methods for Virtual Try-On (VTON) often struggle to accurately transfer garment appearance, especially in unpaired settings where accurate person–garment correspondence is required. These methods do not explicitly supervise person–garment alignment, leaving correspondence to be learned implicitly within the generation model. In this paper, we first analyze full self-attention in DiT-based architectures and show that person→garment query–key matching aligned with local geometric correspondence is closely related to try-on quality. Building on this insight, we introduce **COR**respondence **AL**ignment (**CORAL**), a DiT-based framework that explicitly aligns query–key matching with robust external correspondences. CORAL integrates two complementary components: a correspondence distillation loss that aligns reliable matches with person→garment attention, and an entropy minimization loss that sharpens the attention distribution. For evaluation, we further propose a VLM-based evaluation protocol to better reflect human preference. CORAL consistently improves over the baseline, enhancing both global shape transfer and local detail preservation. Extensive ablations validate our design choices.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

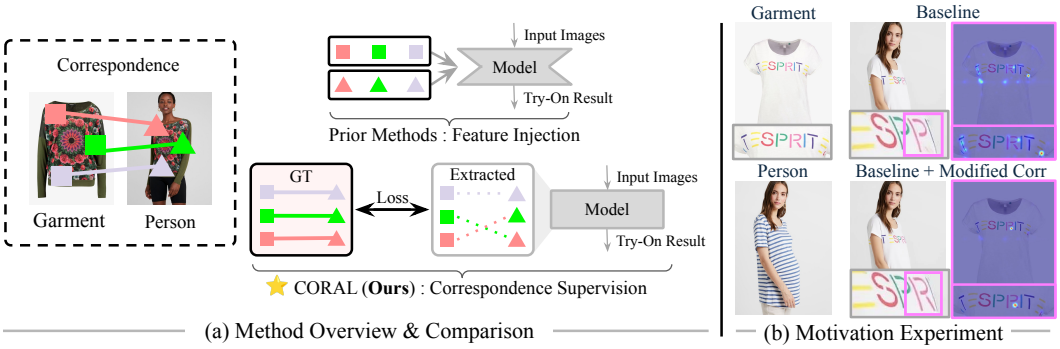


Figure 2: **Motivation and overview of our approach.** Pink markers denote the query points (the bottom-right corner of the letter "R"). (a) Prior methods [65, 8] inject separate shared features for the garment and person into U-Net-based generative models as auxiliary signals, where correspondence is not explicitly enforced. The matches must therefore be estimated at inference as well. In contrast, our method enforces explicit correspondence supervision in a DiT without modifying the feature representations, and requires no correspondence at inference. (b) We modify the baseline's query-key attention at inference time to focus on pseudo ground-truth correspondences from DINOv3. This intervention recovers the correct text and improves try-on quality. (Best viewed when zoomed in. Full images and experiment details are provided in Appendix H.5.)

1 Introduction

Given a pair of person and garment images, Virtual Try-On (VTON) aims to synthesize the same person wearing the given garment, while handling large geometric variations between them, including differences in pose, silhouette, and garment type or shape [13, 76]. Therefore, a key challenge in VTON is ***establishing accurate visual correspondence between the person and the garment***, as inaccurate matching can lead to suboptimal VTON outputs, including distorted garment shapes, misplaced textures, or incorrect fitting.

Visual correspondence has been a core component of VTON. Early approaches warped garment pixels through geometric transformations [66, 12] or learned flow fields [70]. As illustrated in Fig. 2a, diffusion-based methods built on U-Net dual-branch architectures incorporate correspondence as an auxiliary conditioning signal, typically by injecting paired point embeddings into the diffusion model [8, 65]. However, these methods mainly rely on a reconstruction objective during training, which only implicitly encourages person–garment matching and provides no direct supervision over the correspondence itself, often resulting in misaligned generation. Moreover, these methods take correspondence as an input and thus require it at inference, where accurate matching is far harder to obtain since try-on is unpaired.

Meanwhile, Diffusion Transformers (DiTs) [53, 17] have recently gained attention for their strong in-context interactions between input tokens via full 3D attention [59, 27]. Building on this, DiT-based VTON methods adopt a unified design [18], concatenating all tokens into a single sequence. This design removes the need for separate dual-branch architectures [13, 65, 8, 76] and enables person and garment tokens to directly interact within self-attention, providing a natural mechanism for person–garment correspondence to emerge.

However, such correspondence is not always reliably established. While the unified DiT design enables rich cross-token interactions and strong try-on synthesis, these interactions are not explicitly optimized for accurate person–garment matching. Since the diffusion loss is defined only on the output, attention tends to rely on appearance similarity, without establishing local geometric correspondence between body positions and garment parts. As a result, the model may produce visually plausible try-on results while relying on imperfect correspondences, often leading to misplaced textures, distorted garment structures, or incorrect fitting, especially when the person and garment differ in pose, scale, or illumination.

To verify the importance of correspondence in DiT-based VTON, we first conduct an inference-time intervention on the baseline model. Specifically, we use DINOv3, a pretrained vision foundation model that provides dense patch-level visual descriptors, to extract high-quality pseudo correspondences between person and garment regions. We then use these correspondences to locally guide the

baseline's person→garment query–key matching during inference, without updating model weights or modifying feature representations. As shown in Fig. 2, adding a correspondence bias to the person→garment query–key logits shifts the attention toward the matched garment regions, improving local garment appearance transfer and overall try-on quality. This suggests that strengthening the internal person–garment matching in DiT self-attention can directly improve VTON results. Motivated by this observation, we explicitly supervise query–key matching during training.

Based on this, we propose a novel DiT-based framework, **COR**respondence **AL**ignment (**CORAL**), which enhances person–garment correspondence by explicitly supervising query–key matching in the full self-attention of the DiT. Specifically, we introduce two complementary losses: (1) a **correspondence distillation loss** that aligns reliable local geometric matches obtained from the vision foundation model DINOv3 [56] to person–garment query–key matching in the full self-attention, and (2) an **entropy minimization loss** that sharpens the attention distribution by minimizing its entropy for localized and precise matching. Since the correspondence serves only as a training target, the model internalizes the matching and requires no correspondence at inference.

We demonstrate the effectiveness of CORAL on standard VTON benchmarks in both paired and unpaired settings. To further evaluate the robustness of our method, we introduce a new evaluation dataset that includes real-world, challenging person-to-person unpaired scenarios. CORAL achieves state-of-the-art performance on SSIM, LPIPS, and FID across all benchmarks, with the largest improvement on the challenging in-the-wild benchmark. We additionally evaluate our method using VLM-based assessment and human evaluation, achieving the best performance among all methods. Extensive ablation studies further validate our design choices.

In summary, our contributions are as follows:

- We identify correspondence-aware query–key matching as a key factor for improving person–garment alignment in DiT-based VTON.
- Based on these findings, we introduce CORAL, a DiT-based VTON framework that explicitly enhances person–garment correspondence through correspondence distillation and attention entropy minimization.
- We achieve consistent performance gains on standard metrics as well as human and VLM-based evaluations across VTON benchmarks, and validate our method on additional challenging settings that we introduce.

2 Preliminaries

Latent Diffusion Models (LDM). LDM [54] operates in latent space, consisting of VAE and denoiser $\epsilon_\theta(\cdot)$. The VAE encoder maps an image $I \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $\mathbf{z}_0 \in \mathbb{R}^{h\times w\times c}$, where $(H, W, 3)$ and (h, w, c) denote the height, width, and channel dimensions in RGB and latent spaces, respectively. At a timestep t, Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is added to $\mathbf{z}_0$, producing noisy latent $\mathbf{z}_t$. Given conditioning tokens $\mathbf{c}$, we follow rectified flow [17], where the model predicts the velocity transporting a clean latent toward Gaussian noise. Specifically, at a timestep t, a linearly interpolated latent is constructed as

$$\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

and the velocity network $\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$ is trained to approximate the target velocity $(\boldsymbol{\epsilon} - \mathbf{z}_0)$ conditioned on $\mathbf{c}$. The velocity training objective $\mathcal{L}_{\text{velocity}}$ is defined as

$$\mathcal{L}_{\text{velocity}} = \mathbb{E}_{z_0,\,\boldsymbol{\epsilon},\,t}\Big[\|\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\|_2^2\Big]. \tag{2}$$

Attention in DiTs. Diffusion Transformers (DiTs) [53, 17] consist of multiple layers of 3D full attention that jointly process visual and textual information. The noisy latent tokens $\mathbf{z}_t$ and the conditioning tokens $\mathbf{c}_t$ at time step t are concatenated to form the input sequence. The conditioning tokens are constructed by concatenating tokens from multiple conditions, $\mathbf{c}_t = [\, c_{t,1} \| ... \| c_{t,N} \,]$, where N denotes the number of conditions and $[\,\cdot\|\cdot\,]$ denotes token-wise concatenation. Therefore, the final input sequence is $[\,\mathbf{z}_t \| \mathbf{c}_t\,]$. At each timestep t and transformer layer l, the concatenated sequence is projected into queries $Q^{t,l}$ and keys $K^{t,l}$:

$$Q^{t,l} = [\, Q^{t,l}_{\mathbf{z}} \,\|\, Q^{t,l}_{\mathbf{c}_1} \,\|\, ... \,\|\, Q^{t,l}_{\mathbf{c}_N} \,], \quad K^{t,l} = [\, K^{t,l}_{\mathbf{z}} \,\|\, K^{t,l}_{\mathbf{c}_1} \,\|\, ... \,\|\, K^{t,l}_{\mathbf{c}_N} \,]. \tag{3}$$

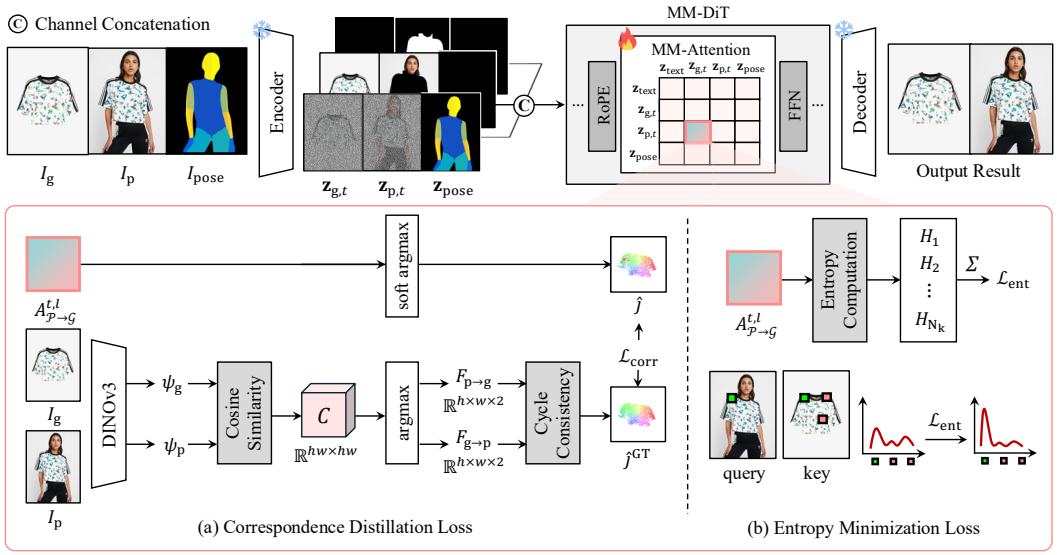


Figure 3: **Overall Architecture.** CORAL builds upon a baseline architecture that constructs the noisy latent $\mathbf{z}_t$ by horizontally concatenating the noisy garment latents $\mathbf{z}_{\mathrm{g},t}$ and person latents $\mathbf{z}_{\mathrm{p},t}$, and then channel-wise concatenates the conditioning canvas $\mathbf{z}_{\mathrm{diptych}}$ and mask canvas $\mathbf{m}_{\mathrm{diptych}}$ with $\mathbf{z}_t$ before the input projection layer. Pose is injected by adding $\mathbf{z}_{\mathrm{pose}}$ as tokens, with RoPE set to share spatial positions between person and pose tokens. $\mathcal{L}_{\mathrm{CORAL}}$ is applied to the person→garment matching cost $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ estimated from MM-Attention within DiT blocks: $\mathcal{L}_{\mathrm{corr}}$ aligns $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ to pseudo ground-truth correspondences extracted from DINOv3, while $\mathcal{L}_{\mathrm{ent}}$ is computed on $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ to encourage sharper, more localized matches.

Each token in the sequence is augmented with rotary position embeddings (RoPE) [58]. The attention map $A^{t,l}$ at timestep t and layer l is then computed as:

$$A^{t,l} = \mathrm{Softmax}\left(\frac{Q^{t,l}(K^{t,l})^\top}{\sqrt{d}}\right), \tag{4}$$

where $\mathrm{Softmax}(\cdot)$ is applied over the key dimension for each query, and d denotes a channel dimension. To sum up, this full attention determines how each latent or conditioning token attends to every other token in the sequence.

3 Method

3.1 Task Definition

Let $I_{\mathrm{p}} \in \mathbb{R}^{H\times W\times 3}$ denote a person image, $I_{\mathrm{g}} \in \mathbb{R}^{H\times W\times 3}$ a garment image, and $I_{\mathrm{pose}} \in \mathbb{R}^{H\times W\times 3}$ a pose image (estimated from I_{p}). Let $M_{\mathrm{p}}, M_{\mathrm{g}} \in \mathbb{R}^{H\times W}$ be binary masks indicating garment regions in I_{p} and I_{g}, respectively, and $M_{\mathrm{e}} \in \mathbb{R}^{H\times W\times 3}$ be a binary inpainting mask, specifying the region to be edited in I_{p}, where $M_{\mathrm{p}} \subseteq M_{\mathrm{e}}$. Given the inputs $\{I_{\mathrm{p}}, I_{\mathrm{g}}, I_{\mathrm{pose}}, M_{\mathrm{p}}, M_{\mathrm{g}}, M_{\mathrm{e}}\}$, our goal is to synthesize a try-on image $\hat{I}$ that (i) preserves the appearance of I_{p} within the region specified by M_{e}, (ii) follows the pose defined by I_{pose}, and (iii) accurately transfers the shape and local details of the garment I_{g} according to M_{g} into the person image, restricted to M_{e}.

3.2 Network Architecture

Diptych Formulation. We first build a DiT-based baseline with a simple conditioning design, leveraging DiT's in-context conditioning ability for reference-guided generation [55, 27]. We adopt a two-panel diptych layout that horizontally concatenates garment and person latents to enable direct token-level interaction through multimodal attention, instead of relying on auxiliary encoders to integrate garment. The overall architecture is illustrated in Fig. 3.

Specifically, we encode the garment and person images, I_{g} and I_{p}, with a VAE to obtain $\mathbf{z}_{\mathrm{g}}$ and $\mathbf{z}_{\mathrm{p}}$, and form a two-panel latent canvas. At timestep t, we construct the noisy diptych latent as:

$$\mathbf{z}_t = [\, \mathbf{z}_{\mathrm{g},t} \,\|\, \mathbf{z}_{\mathrm{p},t} \,], \tag{5}$$

where $\mathbf{z}_{\mathrm{g},t}$ and $\mathbf{z}_{\mathrm{p},t}$ are noisy garment and person latents at timestep t. For conditioning clean garment and masked person image, we use the clean garment latent and the masked person latent as:

$$\mathbf{z}_{\text{diptych}} = [\, \mathbf{z}_{\mathrm{g}} \,\|\, \mathbf{z}_{\mathrm{p}} \odot (1 - \mathbf{m}_{\mathrm{e}}) \,], \tag{6}$$

where $\mathbf{m}_{\mathrm{e}}$ is obtained by downsampling M_{e} into the latent space. We also define a binary mask canvas $\mathbf{m}_{\text{diptych}}$ following the same two-panel layout, assigning ones to the region to be replaced and zeros elsewhere:

$$\mathbf{m}_{\text{diptych}} = [\, \mathbf{0}_{h \times w} \,\|\, \mathbf{m}_{\mathrm{e}} \,]. \tag{7}$$

Pose Injection. Previous try-on methods often suffer from pose hallucination and hand inconsistencies. To mitigate this, a common design in prior methods is to concatenate a pose condition $\mathbf{z}_{\text{pose}}$, obtained by encoding I_{pose} with a VAE, with the noisy person latent $\mathbf{z}_{p,t}$ along the channel dimension [13, 35]. However, we find this approach suboptimal for strict pose alignment, as the person appearance and pose conditions become entangled in the latent space.

Following recent DiT-based work [59], our baseline instead concatenates clean pose condition $\mathbf{z}_{\text{pose}}$ along the token dimension, forming $[\mathbf{z}_{\mathrm{g},t}||\mathbf{z}_{\mathrm{p},t}||\mathbf{z}_{\text{pose}}]$ as input. We then modify RoPE [58] so that $\mathbf{z}_{\mathrm{p},t}$ and $\mathbf{z}_{\text{pose}}$ share the same positional indices, which provides strong positional correspondence and allows pose to be integrated more coherently through full attention. Finally, $\mathbf{z}_{\text{diptych}}$ and $\mathbf{m}_{\text{diptych}}$ are zero-padded to match the input channel dimension and channel-wise concatenated with $[\mathbf{z}_t||\mathbf{z}_{\text{pose}}]$ as the input for the embedding layer, and the resulting tokens are then passed to the DiT.

3.3 Person-Garment Correspondence Analysis

Motivation Correspondence has long been a core component of VTON, and prior methods have improved try-on quality by strengthening person–garment matching. In U-Net dual-branch architectures, this is typically achieved by injecting matched features into the generation network [65, 8]. However, when directly reproduced on our DiT-based baseline, these feature injection strategies yield only marginal improvements, as shown in Tab. 6.

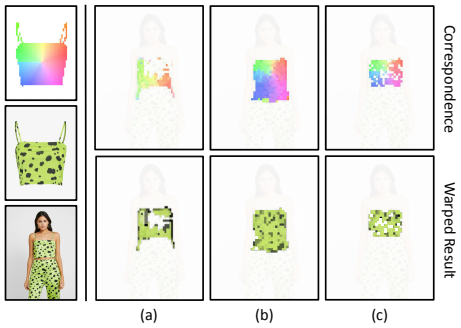


Figure 4: **Correspondence Visualization and Warped Results.** We visualize correspondence fields and the resulting warped garment, computed by mapping pixels from the garment reference image to the garment region of the person image. (a) Baseline attention-derived correspondence. (b) DINOv3 correspondence before reliability filtering. (c) Refined correspondence after cycle-consistency check. The baseline warp shows geometric distortion, while unfiltered DINOv3 can mistakenly match visually similar regions, such as lower-garment areas.

Simply injecting correspondence-aware features provides additional information, but does not explicitly optimize the person-garment matching relation itself. This limitation is particularly relevant in unified DiT architectures, where such features are jointly processed with person, garment, and text tokens within shared self-attention. As a result, the correspondence signal remains as conditioning signal, rather than a training objective that directly guides local person–garment matching. This motivates us to explore correspondence not only as an auxiliary input signal, but also as a direct supervision signal for matching inside DiT.

Another possible direction is to align DiT intermediate features with external matching descriptors(e.g. DINOv3 [56]), following recent representation alignment methods [72]. However, this can interefer with the model's learned generative representations and requires full retraining [72]. Moreover, intermediate features encode appearance information from value matrices that weakens spatial relationships, making them suboptimal for precise matching [1, 47].

To this end, we study query–key attention within the DiT's full self-attention, which implicitly models interactions between person and garment tokens and can be leveraged for explicit correspondence supervision without disrupting the underlying generative representations.

Correspondence Estimation. To examine the baseline's correspondence behavior, we extract query–key correspondences from the DiT's full 3D attention. Following Eq. 3, the query and key at timestep t and layer l consist of text, garment, person, and pose token components:

$$Q^{t,l} = [\, Q^{t,l}_{\mathbf{z}_{\text{text}}} \| Q^{t,l}_{\mathbf{z}_{\text{g}}} \| Q^{t,l}_{\mathbf{z}_{\text{p}}} \| Q^{t,l}_{\mathbf{z}_{\text{pose}}} \,], \quad K^{t,l} = [\, K^{t,l}_{\mathbf{z}_{\text{text}}} \| K^{t,l}_{\mathbf{z}_{\text{g}}} \| K^{t,l}_{\mathbf{z}_{\text{p}}} \| K^{t,l}_{\mathbf{z}_{\text{pose}}} \,]. \tag{8}$$

Let $m_{\text{p}}, m_{\text{g}}$ be the downsampled masks of $M_{\text{p}}, M_{\text{g}}$, and $\mathcal{P}, \mathcal{G}$ the corresponding person and garment token index sets within $Q^{t,l}_{\mathbf{z}_{\text{p}}}$ and $K^{t,l}_{\mathbf{z}_{\text{g}}}$. Using Eq. 4, we compute the person$\rightarrow$ garment attention with $Q^{t,l}_{\mathbf{z}_{\text{p}}}$ as queries and $K^{t,l}_{\mathbf{z}_{\text{g}}}$ as keys:

$$A^{t,l}_{\mathcal{P}\to\mathcal{G}} := A^{t,l}[\mathcal{P}, \mathcal{G}] \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{G}|}. \tag{9}$$

Following [47, 49], we treat $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ as the matching cost between person and garment latents. For simplicity, the variables t and l are omitted from the following equations. From these matching costs, we extract dense correspondences from person spatial location i to garment latents by taking $\arg\max$ over garment tokens:

$$\hat{j}_i = \arg\max_{j\in\mathcal{G}} A_{\mathcal{P}\to\mathcal{G}}(i, j). \tag{10}$$

Evaluation Benchmark. To enable correspondence supervision and evaluation, we require ground-truth person-garment correspondences. Since such annotations do not exist, we construct pseudo ground-truth matches using DINOv3 [56], a vision foundation model, shown to be highly effective for dense matching. CORAL only requires reliable local matches and can work with other dense correspondence estimators. Given I_{p} and I_{g}, we extract feature descriptors $\phi(\cdot)$ from DINOv3 and keep only tokens inside the downsampled masks:

$$\psi_{\text{p}} = \phi(I_{\text{p}}) \odot m_{\text{p}}, \quad \psi_{\text{g}} = \phi(I_{\text{g}}) \odot m_{\text{g}}. \tag{11}$$

Following standard dense matching protocols [24, 25, 11, 48], we compute the matching cost between descriptors using cosine similarity:

$$C(i, j) = \frac{\psi_{\text{p}}(i) \cdot \psi_{\text{g}}(j)}{\|\psi_{\text{p}}(i)\|_2 \, \|\psi_{\text{g}}(j)\|_2}, \tag{12}$$

where i and j index the spatial locations of the person and garment descriptors, respectively.

To obtain reliable pseudo ground-truth matches, we introduce a reliability mask $\mathbf{m}_{\text{rel}}(i)$ that filters out unstable correspondences using a cycle-consistency [32] constraint. We first estimate bidirectional flow, which are person to garment, $F_{\text{p}\to\text{g}} \in \mathbb{R}^{h\times w\times 2}$, and garment to person, $F_{\text{g}\to\text{p}} \in \mathbb{R}^{h\times w\times 2}$, by taking the $\arg\max$ over spatial locations of garment and person in the matching cost, respectively. The reliability mask identifies points whose matched locations in the garment features return to the original locations in the person features within a specified threshold γ:

$$\mathbf{m}_{\text{rel}}(i) = \begin{cases} 1, & \text{if } \|F_{\text{g}\to\text{p}}(F_{\text{p}\to\text{g}}(i)) - i\|_2 < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Finally, we compute the mean L2 loss between estimated correspondences and the pseudo ground-truth matches over the reliable query tokens:

$$\mathcal{L}_{\text{corr}} \;=\; \frac{1}{N_{\text{rel}}} \sum_{i=1}^{N} m_{\text{rel}}(i) \, \big\|\hat{j}_i - \hat{j}_i^{\text{GT}}\big\|_2^2, \qquad N_{\text{rel}} = \sum_{i=1}^{N} m_{\text{rel}}(i), \tag{14}$$

where N is the number of query tokens.

Effect of Correspondence Quality. We leverage constructed pseudo-matches at inference time to investigate how correspondence quality affects VTON performance. Using a pretrained DiT-based VTON model as the baseline, we modify its query-key attention to align with the pseudo-ground truth, sharpening the matching cost to concentrate around the target correspondence locations. As shown in Fig. 2(a), modifying the attention with target correspondences yields visibly improved try-on quality compared to the baseline, indicating that the attention pattern substantially influences generation quality, and directly correcting it without modifying the underlying features clearly improves results.

However, this inference-time intervention requires pseudo ground-truth correspondences for each input pair and remains applicable only to paired settings. To enable the model to internalize this correspondence prior and generalize to unpaired settings, we propose to directly supervise the query-key matching during training. This motivates CORAL, which we describe in the next section.

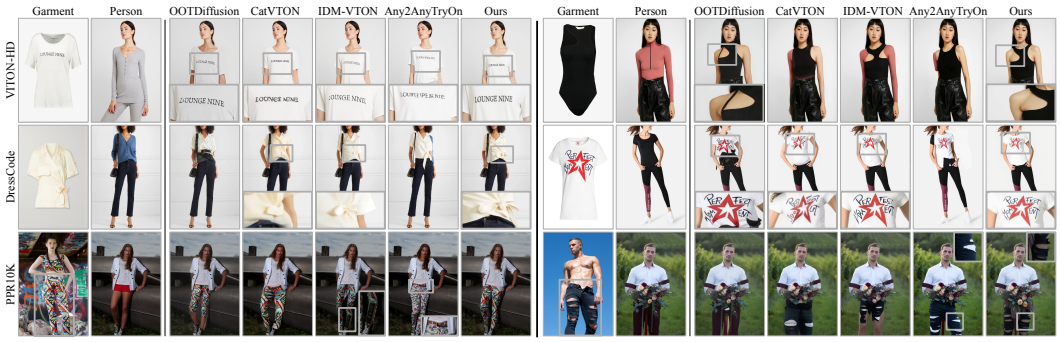


Figure 5: **Qualitative Comparison.** We show qualitative results on the standard benchmarks VITON-HD [12] and DressCode [45], as well as in-the-wild evaluation dataset built from PPR10K [41] (Best viewed when zoomed-in). Additional qualitative results are provided in Appendix N

Methods	Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
GPVTON [70]	0.878	0.067	8.938	4.257	11.993	4.570
StableVITON [35]	0.888	0.073	8.233	0.490	9.026	3.029
OOTDiffusion [71]	0.842	0.087	6.619	0.845	9.938	1.302
IDM-VTON [13]	0.866	0.062	6.009	0.838	9.198	1.203
CatVTON[14]	0.874	0.058	5.458	0.439	9.076	1.184
Any2AnyTryOn [18]	0.838	0.087	5.482	**0.384**	9.623	1.601
FitDiT [31]	0.834	0.105	8.502	1.563	10.310	1.270
CORAL (w/o $\mathcal{L}_{\text{CORAL}}$)	0.889	0.055	5.543	0.870	9.641	1.323
CORAL (w $\mathcal{L}_{\text{CORAL}}$)	**0.907**	**0.048**	**4.962**	0.565	**8.763**	**0.880**

Table 1: **Quantitative Comparison on VITON-HD [12].**

Methods	Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
GPVTON [70]	0.918	0.068	8.423	2.439	9.144	3.936
OOTDiffusion [71]	0.886	0.069	5.082	1.377	9.276	4.009
IDM-VTON [13]	0.904	0.052	3.472	0.882	5.343	1.321
CatVTON [14]	0.875	0.075	5.384	1.903	7.998	3.242
Any2AnyTryOn* [18]	-	-	-	-	5.573	1.458
CORAL (w/o $\mathcal{L}_{\text{CORAL}}$)	0.908	0.045	2.896	0.418	5.221	1.315
CORAL (w $\mathcal{L}_{\text{CORAL}}$)	**0.927**	**0.029**	**2.333**	**0.401**	**4.692**	**0.846**

Table 2: **Quantitative Comparison on DressCode [45].*** does not support paired setting on specific dataset.

Methods	Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
OOTDiffusion [35]	0.804	0.130	99.492	15.592	87.818	15.467
IDM-VTON [13]	0.844	0.111	64.250	1.911	63.638	3.600
CatVTON[14]	0.743	0.147	81.722	10.979	76.417	11.207
Any2AnyTryOn* [18]	-	-	-	-	**49.728**	**0.421**
CORAL (w/o $\mathcal{L}_{\text{CORAL}}$)	0.877	0.078	44.117	0.015	56.644	1.202
CORAL (w $\mathcal{L}_{\text{CORAL}}$)	**0.915**	**0.060**	**43.648**	**0.011**	53.164	1.101

Table 3: **Quantitative Comparison on PPR10K [41].** * does not support paired setting on specific dataset.

Train/Test	VITON-HD/VITON-HD			DressCode/DressCode			DressCode/PPR10K		
	GTC ↑	TAC ↑	FPC ↑	GTC ↑	TAC ↑	FPC ↑	GTC ↑	TAC ↑	FPC ↑
GPVTON* [70]	3.93	4.28	4.17	3.19	3.31	3.14	-	-	-
OOTDiffusion [71]	3.97	4.34	4.12	3.25	3.24	3.45	1.04	2.06	1.57
IDM-VTON [13]	3.96	4.40	4.25	3.25	3.28	3.61	1.72	2.21	2.37
CatVTON [14]	3.82	4.20	4.06	3.07	3.22	3.53	1.05	2.00	2.01
Any2AnyTryOn [18]	3.81	4.27	4.19	2.99	3.28	3.36	1.92	2.45	2.48
CORAL (Ours)	**3.99**	**4.40**	**4.26**	**3.47**	**3.31**	**3.83**	**2.07**	**2.56**	**2.89**

Table 4: VLM-based evaluation on unpaired settings. Scores are averages of per-sample GPT-4o ratings on a continuous 0–5 scale (Appendix J). * does not support inference on specific dataset.

3.4 CORAL: Correspondence Alignment

Based on our observation, we introduce CORAL, which integrates reliable pseudo ground-truth matches into the baseline's query–key matching for robust virtual try-on performance.

Correspondence Distillation Loss. We aim to align the estimated matches obtained from query–key attention with the pseudo ground-truth matches provided by DINOv3 [56]. Because the argmax operation is not differentiable (Eq. 10), we use a soft argmax to predict the estimated correspondences from the query–key attention, formulated as follows:

$$\hat{j}_i = \sum_{j=1}^{N} A_{\mathcal{P}\to\mathcal{G}}(i,j) \cdot j. \tag{15}$$

Finally, we compute the mean L2 loss between estimated correspondences and the pseudo ground-truth matches:

$$\mathcal{L}_{\text{corr}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{j}_i - \hat{j}_i^{\text{GT}}\|_2^2. \tag{16}$$

Entropy Minimization Loss. While correspondence distillation loss ($\mathcal{L}_{\text{corr}}$) aligns query-key matches with external pseudo-matches, this alignment is unreliable when the attention distribution becomes overly diffuse, as incorrect key positions can then dominate the weighted average. To

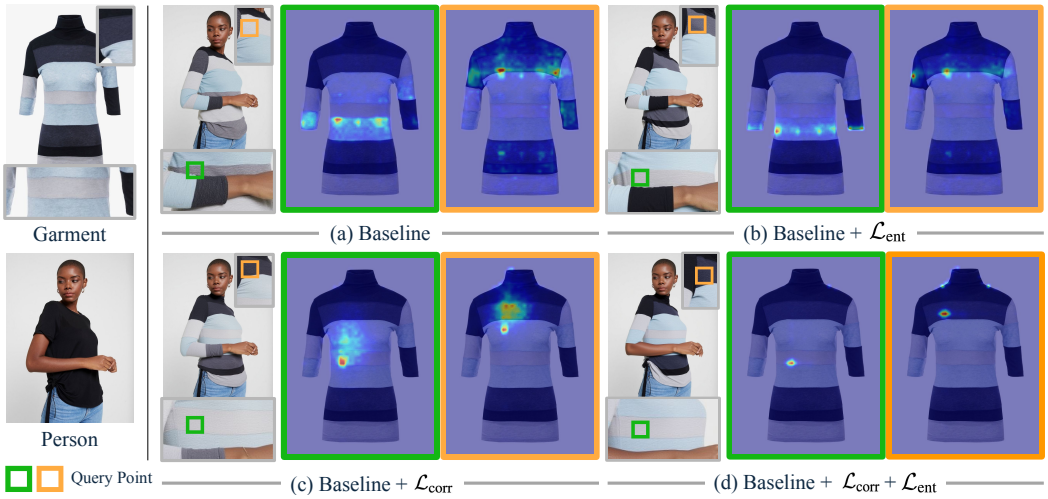


Figure 6: **Loss Component Ablations.** We show the effectiveness of $\mathcal{L}_{\text{ent}}$ and $\mathcal{L}_{\text{corr}}$. Orange and green markers denote query points, and attention maps outlined in the same colors indicate the matches for each variant. By combining $\mathcal{L}_{\text{ent}}$ with $\mathcal{L}_{\text{corr}}$, our model localizes the correct keys most accurately and exhibits the sharpest attention, yielding the best VTON performance.

ensure robust correspondence supervision, we introduce an entropy minimization loss that promotes confident, spatially localized query-key alignments. To implement this, we define the entropy based on the attention distribution. Following Shannon's definition of entropy [2], we compute the entropy H_i for a given query i by treating its attention weights $A(i, \cdot)$ over the keys as a probability distribution:

$$H_i \;=\; -\sum_{j=1}^{N_{\text{k}}} A(i,j)\,\log A(i,j), \tag{17}$$

where N_{k} is the total number of key tokens. Averaging H_i over the reliable query tokens yields the entropy minimization loss:

$$\mathcal{L}_{\text{ent}} \;=\; \frac{1}{N_{\text{rel}}} \sum_{i=1}^{N} m_{\text{rel}}(i)\, H_i. \tag{18}$$

Both losses are thus computed only over query tokens that pass the cycle-consistency check (Eq. 13). Tokens without a reliable match are left to the diffusion objective, as in the baseline, since there is no correct position to teach there.

Lower entropy corresponds to sharper and more confident attention, whereas higher entropy indicates diffuse or uncertain matching. Entropy minimization loss ($\mathcal{L}_{\text{ent}}$) therefore complements the correspondence distillation loss ($\mathcal{L}_{\text{corr}}$) by enforcing sharper, and more reliable alignments.

Total Loss. We define $\mathcal{L}_{\text{CORAL}} = \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}$ and optimize:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{velocity}} + \mathcal{L}_{\text{CORAL}}. \tag{19}$$

4 Experiments

4.1 Evaluation Setting

Evaluation Dataset. We train separate models on the standard VTON datasets, **VITON-HD** [20] and **DressCode** [45], both at a resolution of 1024×768, and evaluate them on their respective test sets. For more implementation details, see Appendix A. Additionally, we evaluate zero-shot generalization on the **PPR10K** [41], which provides less curated inputs than the in-shop product images. For this evaluation, we use a pre-trained checkpoint trained on DressCode. We also include the worn-garment setting, where the reference garment is shown on another person, which is offered in Appendix M. Additional details on PPR10K evaluation are provided in Appendix L.

Metrics. We report conventional metrics SSIM [15], LPIPS [74], FID [23] and KID [4] for paired settings, and measure FID and KID for unpaired settings. We additionally evaluate correspondence accuracy using Percentage of Correct Keypoints (PCK) on the VITON-HD test set. PCK measures whether predicted correspondences fall within a distance threshold α of the pseudo–matches.

VLM-based Evaluation. As prior studies note [69], FID and KID inadequately reflect garment transfer fidelity in unpaired settings. We therefore evaluate three aspects: Garment Transfer Consistency (GTC), Textual Attribute Consistency (TAC), and Fit Pose Coherence (FPC), capturing garment fidelity, attribute alignment, and natural fit, respectively. For more details, see Appendix J and K.

4.2 Quantitative Comparison

Main Results. Tab. 1 and Tab. 2 report quantitative comparisons on VITON-HD [12] and DressCode [45]. CORAL achieves the best overall performance on both benchmarks. Tab. 3 further shows that CORAL achieves the best results in the paired setting and remains competitive in the unpaired setting on the in-the-wild dataset, outperforming most baselines. Across all benchmarks, CORAL consistently improves every metric, demonstrating robust gains over strong baselines.

VLM Evaluation Results. Tab. 4 reports VLM-based evaluation results, where CORAL achieves the best performance across all metrics. By aligning query-key matches to correct person-garment regions, CORAL transfers appearance more accurately, improving GTC. More precise correspondences better preserve garment attributes, yielding higher TAC, and result in more natural wear and pose adherence, reflected in improved FPC.

4.3 Qualitative Comparison

Fig. 5 shows qualitative comparisons between prior methods and CORAL. CORAL preserves global garment structure while recovering fine details missed by baselines, and keeps garment text clear where others produce blur or distortion. It also maintains closer pose alignment to the input, avoiding subtle deviations. For more results, see Appendix N.

4.4 Ablation Study

Ablation of Loss Components. Fig. 6 and Tab. 5 highlight the effects of our loss design. (a) and (I) show the baseline, where attention is poorly localized and often dispersed away from query-relevant regions, yielding misaligned matches (e.g., misplaced patterns) and artifacts. (b) and (II) present the effect of $\mathcal{L}_{\text{ent}}$, which sharpens attention and improves fine detail transfer. However, since $\mathcal{L}_{\text{ent}}$ does not correct *where* the model matches, coarse properties like garment length often remain inaccurate, and sharpening on wrong matches can transfer inappropriate appearance. This explains why (II) improves SSIM and LPIPS over (I), while still exhibiting misalignment in challenging cases. (c) and (III) show the effect of $\mathcal{L}_{\text{corr}}$, which relocates correspondences to the correct garment regions and improves coarse properties (e.g., silhouette, pattern geometry). However, attention remains relatively diffuse, limiting fine detail transfer with low attention mass on the matched key. Compared to (b), (c) achieves more accurate geometry but lags in color fidelity. (d) and (IV), the combination of $\mathcal{L}_{\text{ent}}$ and $\mathcal{L}_{\text{corr}}$, integrate these complementary gains. This yields both accurate alignment and sharp detail transfer, leading to the best qualitative and quantitative results.

	$\mathcal{L}_{\text{corr}}$	$\mathcal{L}_{\text{ent}}$	Paired				Unpaired	
			SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
(I)	✗	✗	0.889	0.055	5.543	0.870	9.641	1.423
(II)	✗	✓	0.905	0.052	5.482	0.986	9.012	1.011
(III)	✓	✗	0.906	0.051	5.451	1.028	8.979	1.250
(IV)	✓	✓	0.907	0.048	4.962	0.565	8.763	0.880

Table 5: **Loss Components Ablation on VITON-HD.**

Comparison with Prior Approaches. We compare CORAL with two kinds of prior approaches, reproduced on our DiT baseline, to highlight the importance of direct supervision in DiT-based VTON. The first is feature injection. We reproduce SPM-Diff [65] by injecting garment features at matched positions using per-layer zero-initialized linear projections, following the original design (see Appendix D.2 for implementation details), and evaluate both the original sparse setting (25 matched points) and a dense correspondence variant. The second

Method	SSIM ↑	LPIPS ↓	FID ↓
Baseline	0.889	0.055	5.543
Baseline + DINOv3 conditioning (sparse, ∼25pts)	0.889	0.054	5.889
Baseline + DINOv3 conditioning (dense)	0.883	0.074	5.598
Baseline + ATV loss	0.892	0.054	5.652
CORAL (direct supervision)	**0.908**	**0.048**	**5.009**

Table 6: **Prior Approaches on VITON-HD**

is attention regularization without a correspondence target. We apply the attention total variation (ATV) loss of StableVITON [35], which acts on the same person-to-garment attention as ours but only smooths its attention center across neighbouring queries. As shown in Tab. 6, both feature injection variants yield negligible or negative gains over the baseline. ATV improves reconstruction slightly but worsens FID. It supervises neither of the two things our losses do: it names no position for attention to point at, and it never directly asks attention to sharpen onto one. CORAL improves all metrics. This confirms that auxiliary conditioning is insufficient for DiT-based VTON, and that direct supervision on attention-level correspondences is required, on both where attention points and how sharply.

Impact of CORAL on Attention Correspondence. Fig. 7 summarizes how $\mathcal{L}_{\text{CORAL}}$ affects $A^{t,l}[\mathcal{P}, \mathcal{G}]$ and generated results. Without $\mathcal{L}_{\text{CORAL}}$, the correspondence maps exhibit mismatches near local textures and garment boundaries, despite plausible vertical alignment seems to appear. These mismatches lead to unclear logo patterns in the generated result. Adding $\mathcal{L}_{\text{CORAL}}$ consistently corrects the correspondence maps, producing matches that more closely follow the DINOv3 pseudo ground-truth matches, leading to improved text details. Fig. 7 also shows this quantitatively: $\mathcal{L}_{\text{CORAL}}$ improves PCK ($\alpha = 16$) by $34\,\%$ while reducing attention entropy, indicating more accurate and sharper correspondences.

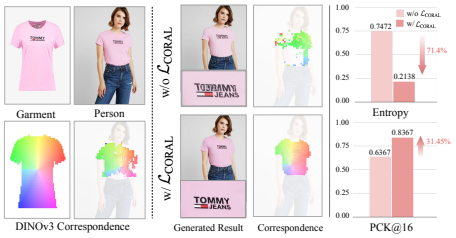


Figure 7: **Effect of $\mathcal{L}_{\text{CORAL}}$.** We visualize person→garment correspondences from attention and the resulting try-on images before and after applying $\mathcal{L}_{\text{CORAL}}$. The rightmost plot shows correspondence quality, averaged across timesteps and layers.

5 Conclusion

We introduce CORAL, a DiT-based framework that aligns person–garment query–key matches with robust DINOv3 matches. CORAL combines a correspondence distillation loss to guide attention and an entropy minimization loss to sharpen it. As a result, CORAL achieves outperforming performance across standard and in-the-wild benchmarks, supported by extensive analyses validating our design choices.

Acknowledgments and Disclosure of Funding

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents, Contribution: 50%) and by National IT Industry Promotion Agency (NIPA) grant funded by the Korea government (MSIT) (Development of Physical Intelligence Action Models and an Integrated Framework, No. RS-2026-25621677, Contribution: 50%).

References

[1] Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1103–1115, 2025.

[2] Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. Entropy-based attention regularization frees unintended bias mitigation from lists, 2022. URL `https://arxiv.org/abs/2203.09192`.

[3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model, 2023. URL `https://arxiv.org/abs/2211.12500`.

[4] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. URL `https://arxiv.org/abs/1801.01401`.

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023. URL https://arxiv.org/abs/2304.08465.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. doi: 10.1109/TPAMI.2019.2929257.

[7] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

[8] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment, 2024. URL https://arxiv.org/abs/2403.12965.

[9] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models, 2025. URL https://arxiv.org/abs/2503.09567.

[10] Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, Yanlin Li, Lei Ren, Wei Chen, Zhiyuan Huang, Mingjie Zhan, Xiaojie Wang, and Fangxiang Feng. Evaluating mllms with multimodal multi-image reasoning benchmark, 2025. URL https://arxiv.org/abs/2506.04280.

[11] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. In *European conference on computer vision*, pages 306–325. Springer, 2024.

[12] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021. URL https://arxiv.org/abs/2103.16874.

[13] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild, 2024. URL https://arxiv.org/abs/2403.05139.

[14] 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, 2025. URL https://arxiv.org/abs/2407.15886.

[15] Heng Cong, Lingzhi Fu, Rongyu Zhang, Yusheng Zhang, Hao Wang, Jiarong He, and Jin Gao. Image quality assessment with gradient siamese network, 2022. URL https://arxiv.org/abs/2208.04081.

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

[18] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks, 2025. URL https://arxiv.org/abs/2501.15891.

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018. URL https://arxiv.org/abs/1802.00434.

[20] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network, 2018. URL https://arxiv.org/abs/1711.08447.

[21] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion, 2023. URL https://arxiv.org/abs/2305.15581.

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

[23] 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, 2018. URL https://arxiv.org/abs/1706.08500.

[24] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022.

[25] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. *Advances in Neural Information Processing Systems*, 35:13512–13526, 2022.

[26] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[27] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers, 2024. URL https://arxiv.org/abs/2410.23775.

[28] Zaiyu Huang, Hanhui Li, Zhenyu Xie, Michael Kampffmeyer, Qingling Cai, and Xiaodan Liang. Towards hard-pose virtual try-on via 3d-aware global correspondence learning, 2022. URL https://arxiv.org/abs/2211.14052.

[29] Hyeonho Jeong, Chun-Hao Paul Huang, Jong Chul Ye, Niloy Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation, 2025. URL https://arxiv.org/abs/2412.06016.

[30] Boyu Jia, Junzhe Zhang, Huixuan Zhang, and Xiaojun Wan. Exploring and evaluating multimodal knowledge reasoning consistency of multimodal large language models, 2025. URL https://arxiv.org/abs/2503.04801.

[31] Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on, 2024. URL https://arxiv.org/abs/2411.10499.

[32] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images, 2021. URL https://arxiv.org/abs/2103.14167.

[33] Siyoon Jin, Seongchan Kim, Dahyun Chung, Jaeho Lee, Hyunwook Choi, Jisu Nam, Jiyoung Kim, and Seungryong Kim. Matrix: Mask track alignment for interaction-aware video generation, 2025. URL https://arxiv.org/abs/2510.07310.

[34] Siyoon Jin, Jisu Nam, Jiyoung Kim, Dahyun Chung, Yeong-Seok Kim, Joonhyung Park, Heonjeong Chu, and Seungryong Kim. Appearance matching adapter for exemplar-based semantic image synthesis in-the-wild, 2025. URL https://arxiv.org/abs/2412.03150.

[35] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on, 2023. URL https://arxiv.org/abs/2312.01725.

[36] 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, 2025. URL https://arxiv.org/abs/2412.16978.

[37] Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think, 2025. URL https://arxiv.org/abs/2503.08250.

[38] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. URL https://arxiv.org/abs/2407.07895.

[39] Nannan Li, Kevin J. Shih, and Bryan A. Plummer. Enhancing virtual try-on with synthetic pairs and error-aware noise scheduling, 2025. URL https://arxiv.org/abs/2501.04666.

[40] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing, 2019. URL https://arxiv.org/abs/1910.09777.

[41] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency, 2021. URL https://arxiv.org/abs/2105.09180.

[42] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. doi: 10.1109/CVPR.2016.124.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[44] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner, 2024. URL https://arxiv.org/abs/2306.06101.

[45] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on, 2022. URL https://arxiv.org/abs/2204.08532.

[46] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on, 2023. URL https://arxiv.org/abs/2305.13501.

[47] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024.

[48] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching, 2024. URL https://arxiv.org/abs/2305.19094.

[49] Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025.

[50] David Nordström, Johan Edstedt, Georg Bökman, Jonathan Astermark, Anders Heyden, Viktor Larsson, Mårten Wadenbäck, Michael Felsberg, and Fredrik Kahl. Loma: Local feature matching revisited, 2026. URL https://arxiv.org/abs/2604.04931.

[51] OpenAI and Josh Achiam et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[55] Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator, 2025. URL https://arxiv.org/abs/2411.15466.

[56] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

[57] Michael Soloveitchik, Tzvi Diskin, Efrat Morin, and Ami Wiesel. Conditional frechet inception distance, 2022. URL https://arxiv.org/abs/2103.11521.

[58] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

[59] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. URL https://arxiv.org/abs/2411.15098.

[60] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers, 2025. URL https://arxiv.org/abs/2503.08280.

[61] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. URL https://arxiv.org/abs/2306.03881.

[62] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences, 2021. URL https://arxiv.org/abs/2104.03308.

[63] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. URL https://arxiv.org/abs/2211.12572.

[64] Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. Improving virtual try-on with garment-focused diffusion models, 2024. URL https://arxiv.org/abs/2409.08258.

[65] Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on, 2025. URL https://arxiv.org/abs/2505.16977.

[66] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network, 2018. URL https://arxiv.org/abs/1807.07688.

[67] Chenhui Wang, Tao Chen, Zhihao Chen, Zhizhong Huang, Taoran Jiang, Qi Wang, and Hongming Shan. Fldm-vton: Faithful latent diffusion model for virtual try-on, 2024. URL https://arxiv.org/abs/2404.14162.

[68] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller, 2024. URL https://arxiv.org/abs/2405.14864.

[69] Hu Xiaobin, Liang Yujie, Luo Donghao, Peng Xu, Zhang Jiangning, Zhu Junwei, Wang Chengjie, and Fu Yanwei. Vtbench: Comprehensive benchmark suite towards real-world virtual try-on models, 2025. URL https://arxiv.org/abs/2505.19571.

[70] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning, 2023. URL https://arxiv.org/abs/2303.13756.

[71] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on, 2024. URL https://arxiv.org/abs/2403.01779.

[72] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025. URL https://arxiv.org/abs/2410.06940.

[73] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence, 2023. URL https://arxiv.org/abs/2305.15347.

[74] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

[75] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models, 2025. URL https://arxiv.org/abs/2505.23656.

[76] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen He. Learning flow fields in attention for controllable person image generation, 2024. URL https://arxiv.org/abs/2412.08486.

Appendix

This appendix provides supplementary material to support the main paper.

Contents

A Implementation Details **17**
A.1 Training Setup 17
A.2 Pseudo Ground-Truth Construction 17
B Training Dataset **17**
B.1 Dataset Overview 17
B.2 Dataset Preprocessing 17
C Related Works **18**
D Detailed Method Analysis **19**
D.1 Choice of Supervision Signal 19
D.2 Choice of Mechanism 20
D.3 Choice of Alignment Target 21
E Additional Analysis on Pseudo Ground-Truth **23**
E.1 Robustness to Pseudo-GT Errors 23
E.2 Generality across Pseudo-GT Sources 23
E.3 Behavior under Challenging Conditions 24
F Additional Ablation Studies **26**
F.1 Hyperparameter Sensitivity 26
F.2 Additional Visualizations of Loss Component Ablations 26
G Additional Details on Motivation Experiment **27**
H Runtime and Memory **27**
I CORAL on Other Baselines **29**
I.1 Baselines with Different Pose Injection Strategies 29
I.2 An Additional Efficient Baseline 30
I.3 Baseline with a Different Backbone 31
J VLM-based Evaluation **32**
K Human Evaluation **34**
L Evaluation on Additional Datasets **35**
L.1 Motivation 35
L.2 Evaluation Dataset Details 35
M Person-To-Person Garment Transfer **36**
N Additional Qualitative Results **39**
O Limitations and Discussion **44**

A Implementation Details

A.1 Training Setup

We initialize our model from FLUX.1-Fill-dev and finetune the attention layers. We train with the AdamW [43] optimizer using a learning rate of 2×10^{-5} and a batch size of 16 for a total of 8.4k steps on 8 NVIDIA B200 GPUs. We train and report results at a resolution of 1024×768. For training with $\mathcal{L}_{\text{CORAL}}$, same learning rate and batch size are used, with loss weights $\lambda_{\text{corr}} = 0.01$ and $\lambda_{\text{ent}} = 0.1$.

A.2 Pseudo Ground-Truth Construction

We construct pseudo ground-truth correspondences using DINOv3-B [56] and set the cycle-consistency threshold γ as 3. The DiT operates on latents from an $8\times$ VAE with a 2×2 patchify, so each attention token corresponds to a 16×16 pixel region, giving a 64×48 token grid at 1024×768. DINOv3-B has a patch size of 16 and receives the same 1024×768 input, so its token grid is also 64×48. The pseudo ground-truth is thus defined directly on the grid on which the loss operates, with no resampling. When a matcher has a different patch size, we adjust its input resolution so that it produces the same token grid, rather than interpolating its matches afterwards. We compute $\mathcal{L}_{\text{CORAL}}$ over all layers throughout training and average it across layers, with both terms restricted to query tokens that pass the cycle-consistency check (Eqs. 16 and 18).

B Training Dataset

B.1 Dataset Overview

VITON-HD [12] comprises 11,647 training images and 2,032 test images and focuses on upper-body garments captured in a studio setting. DressCode [45] consists of 48,392 training set and 5,400 test set, covering 13,563 upper, 7,151 lower and 27,678 dresses. The person images are generally full-body centered, with face region excluded. In both datasets, garment product images are typically captured on plain, white backgrounds and presented as flat product shots. Overall, person images in both benchmarks are captured under controlled backgrounds and lighting, and most samples present static, frontal poses with limited viewpoint variation.

B.2 Dataset Preprocessing

We follow prior works [14, 27] for input prompts, using a fixed input prompt for VITON-HD and adding the garment category for DressCode. For VITON-HD, we use both the masks from benchmarks and ones preprocessed by prior work [13] in a 1:1 ratio. For DressCode, we adopt masks preprocessed by prior work [13], and at each training step we apply a random directional dilation with a radius uniformly sampled from 1 to 20 pixels to augment the masks. Input masks can cover cloth-agnostic regions such as hair or bags, which makes it difficult to preserve their appearance from the input under occlusions. While one could refine the masks to exclude such regions, this comes with practical risks: if the masks become too tight, they can reveal the originally worn garment, while accurately segmenting fine structures such as hair across diverse poses and viewpoints is often costly and error-prone. Instead, we extract these regions using human-parser [40] or SAM3 [7], and provide them as additional guidance by stitching them into the DensePose [19] condition.

C Related Works

Virtual Try-On. VTON aims to precisely transfer the shape and details of the garment image onto a person image. Several works improve sampling and alignment at inference [3, 14, 67, 39]. Other efforts [13, 36, 70] add structural cues (e.g., parsing maps) to improve robustness against pose variations. Other methods inject global [35, 64, 46] or local features [13, 76, 71] from garment images to enhance detail preservation. Yet, they often lose garment details in unpaired settings that demand robust matching, leading to logo drift or boundary shifts.

Correspondence in Diffusion Models. The impressive performance of diffusion models comes from the rich visual representations they learn. Analyses of the internal attention mechanism in both UNet [47, 48, 29, 34, 21, 61, 68] and DiT-based [72, 49, 33, 37] models show that query-key matching encodes semantic correspondences. Editing and customization studies [47, 5, 34, 22, 63] consistently report that query-key matching captures structure and semantic alignment, while values carry an appearance that is routed to attended locations. These findings enable strong layout preservation and controllable appearance transfer, and also support downstream tasks such as tracking [29, 49].

Correspondence in VTON. Several VTON works [65, 8, 28] inject sparse, point-based correspondences into attention or warp features around a few matched points. Beyond sparse correspondences, some [76] derive dense flows from attention, but in unpaired settings, such photometric assumption [62] fails more often, degrading the performance. In these methods, correspondence is an input to the model, so it must also be estimated at inference. In training, the person wears the reference garment, so matches can be estimated from shared appearance. In actual use, the person has never worn the garment, so there is no shared appearance to match, and the estimated matches become unreliable. StableVITON [35] places a loss on person-garment attention, but it only encourages smooth attention across neighboring queries without specifying where attention should point. Our approach instead operates in DiTs and explicitly strengthens query-key matching by aligning attention to external, robust person-garment correspondences from DINOv3 [56], reducing reliance on photometric consistency. The correspondence is used only as a training target and verified by a cycle-consistency check, so inference requires no correspondence.

D Detailed Method Analysis

We provide a detailed analysis of CORAL's design from three perspectives: what kind of signal to supervise with (Sec. D.1), how correspondence should be used in the model (Sec. D.2), and where in the generation model to apply supervision (Sec. D.3). We discuss CORAL's design choices along these aspects and analyze why they are effective.

D.1 Choice of Supervision Signal: With What Kind of Signal?

We can supervise person-garment alignment with different signals. We compare 3 forms used in related approaches and motivate CORAL's choice.

Photometric Reconstruction. Photometric reconstruction is the most widely used form in VTON, computing reconstruction loss between the warped or generated output and the ground-truth person image. This is dominant in GAN-based methods [70, 12] that supervise on warped outputs from explicit flow estimators, and is also adopted in recent diffusion-based VTON methods [76]. However, photometric reconstruction relies on the assumption that pixel intensities are preserved between corresponding regions, which is often violated in practice due to lighting variations, material reflectance differences, and non-rigid deformations of garments on the body [66].

Global Semantic Alignment. A natural question is whether attention should disperse across all semantically similar regions, since they would share the same semantic identity. Recent works such as REPA [72] align the model's intermediate features with semantic descriptors from vision foundation models. However, multiple regions in a garment often share similar semantic features (e.g., repeated stripe patterns), and attending to all such regions introduces ambiguity in spatial correspondence. In contrast, geometric correspondence enforces a single, geometrically consistent match, providing a more reliable signal for generation.

Local Geometric Correspondence (Ours). CORAL adopts local geometric correspondence as its supervision signal. From features extracted by a vision foundation model, we compute pairwise similarity between every person token position and every garment token position, and extract for each person token the garment position of maximum similarity. This yields, for each person token, a target position in the garment that the model's attention should be directed toward. The supervision is then applied to the model's Q-K matching pattern. Unlike global semantic alignment, where the supervision target is the model's intermediate features, our supervision target is positional: it ties each person token to a specific garment location rather than to a semantic identity shared across multiple regions. This resolves the dispersion ambiguity faced by both photometric reconstruction and global semantic alignment. Since supervision operates on the matching pattern rather than the features themselves, it preserves the model's generative representations.

To empirically compare these supervision signals, we reproduce reconstruction-based supervision and global semantic alignment on our DiT baseline. As shown in Tab. D.1, local geometric correspondence supervision (CORAL) yields the best performance across all metrics, validating our choice.

Implementation Details for Reproduction. Both alternatives are reproduced by replacing the supervision target in our loss formulation (16) with different quantities, while keeping the attention map A and training configuration unchanged. For **photometric reconstruction**, we leverage reference RGB values from the person image. The input images are downsampled to match the attention map resolution:

Supervision Signal	SSIM ↑	LPIPS ↓	FID ↓
Baseline (no supervision)	0.889	0.055	5.543
Reconstruction-based	0.900	0.051	5.542
Global semantic alignment	0.887	0.059	5.724
Local geometric correspondence (Ours)	**0.908**	**0.048**	**5.009**

Table D.1: **Comparison of supervision signals on our DiT baseline.** Local geometric correspondence supervision yields the best performance compared to reconstruction-based and global semantic alternatives reproduced on the same DiT backbone.

$$\mathcal{L}_{\text{photo}} = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \left\| \sum_{i \in \mathcal{G}} A_{ij}\, \tilde{I}_g[i] - \tilde{I}_p[j] \right\|_2^2, \tag{21}$$

where $\tilde{I}_g, \tilde{I}_p$ denote the bilinear-downsampled garment and person images at the latent resolution. For **global semantic alignment**, we replace the position-based target with semantic features extracted by DINOv3 [56]:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \left\| \sum_{i \in \mathcal{G}} A_{ij}\, \phi(I_g)_i - \phi(I_p)_j \right\|_2^2, \tag{22}$$

where ϕ denotes the DINOv3 feature extractor and $\phi(\cdot)_i$ is the feature at token i.

D.2 Choice of Mechanism: How should Correspondence be Used?

The main paper empirically demonstrates that feature injection approaches yield limited gains on our DiT baseline. Here we analyze the architectural reasons: conditioning, which feature injection is one mechanism of, is inherently limited in DiT-based VTON, while direct supervision is more effective.

Conditioning Approaches in Prior VTON. Prior correspondence-based VTON methods [8, 65] integrate sparse correspondences into the model through conditioning: garment features extracted at matched positions are added to the model as additional input, and the model is trained only with the diffusion loss. The model is not explicitly guided to preserve the correspondence information carried by the added features, as illustrated in Fig. 2

Information Flow in UNet vs DiT Attention. Such conditioning approaches may operate on UNet-based dual-branch architectures, where garment information reaches the person branch through a single pathway: Garm-UNet features fed into augmented self-attention. Formally, $h_g^l = \text{GarmUNet}^l(I_g)$ depends solely on the garment image, so a correspondence feature added along this pathway has a consistent, predictable effect with no alternative route to bypass it.

In contrast, DiT-based VTON employs unified full attention where garment, person, pose, and text tokens become entangled at every layer: $h_g^l = f^l(I_g, I_p, \text{text}, \text{pose})$. Unlike UNet, unified attention offers many alternative routes for person tokens to access garment information, including indirect paths through tokens that have already incorporated it. A correspondence feature added through conditioning is only one route among many: if the model can obtain garment information through other routes, the correspondence signal is effectively bypassed without any explicit incentive to preserve it.

Empirical Verification. This is empirically verified in Section 4.4 of the main paper, where SPM-Diff [65] is reproduced on our DiT baseline. Both the original sparse formulation with $\sim$25 matched points and a dense extension that uses correspondences at all positions yield negligible or negative improvement over the baseline, consistent with our architectural analysis above.

Implementation Details for Reproduction. We implement a feature injection conditioning variant on our DiT baseline, following the design of SPM-Diff [65]. Projected DINOv3 features are injected into attention layers by adding flow-warped garment features to person-side query and key, and original garment features to garment-side key, with per-layer zero-initialized linear projections. As the inference code is not publicly available, we implement this conditioning from the paper description. We test two densities: a sparse variant using $\sim$25 matched points (as in the original formulation) and a dense variant using correspondences at all positions.

As shown in Tab. 6, both variants yield negligible or negative improvement compared to the baseline, consistent with our analysis. In contrast, CORAL's direct supervision explicitly enforces correspondence within attention, yielding substantial improvement across all metrics.

D.3 Choice of Alignment Target: Where to Apply Supervision?

Given that direct supervision is adopted, the final question is where within the generation model the supervision should be applied. We compare two alignment targets and motivate CORAL's choice of Q-K matching.

Intermediate Features vs Q-K Matching. We compare two alignment targets within DiT: aligning the model's intermediate features with feature descriptors from vision foundation models, as in prior approach [72], versus aligning the Q-K similarity between person and garment tokens, as in CORAL. The two targets differ in two key aspects.

First, in what they preserve in the model. Prior analyses [47] show that query-key similarities largely determine the spatial arrangement of generated image elements while value features mainly carry appearance details. Since intermediate features jointly encode both spatial and appearance information, aligning them directly can disrupt the model's learned appearance representation. CORAL supervises only the query-key matches, correcting spatial alignment without altering the value pathway, thus preserving the model's appearance representation.

Second, in their training assumptions. REPA alignment is designed to accelerate training from scratch rather than to fine-tune a pretrained model [75]. Hard alignment pulls intermediate features toward external descriptors, disrupting the inherent generative representations learned during pretraining and leading to slower convergence during fine-tuning. CORAL operates only on attention similarities, leaving the underlying features intact.

Implementation Details for Feature Alignment. We build the intermediate feature alignment baseline upon the same architectural modifications adopted in our main baseline described in Sec. 3.2. Following REPA [72], we attach a lightweight projection head to the output of each transformer block, where each head is a three-layer MLP with SiLU activations that maps DiT hidden states into the DINOv3 [56] embedding space. We use DINOv3-B as the external encoder, consistent with our default configuration.

Given a garment image I_{g} and a person image I_{p}, we concatenate them horizontally to form $\hat{I} \in \mathbb{R}^{H\times 2W\times 3}$, and extract DINOv3 features $\phi_{\mathrm{g}}, \phi_{\mathrm{p}} \in \mathbb{R}^{hw\times 768}$ from the frozen encoder. We concatenate them as:

$$\hat{\phi} = [\,\phi_{\mathrm{g}} \,\|\, \phi_{\mathrm{p}}\,] \in \mathbb{R}^{2hw\times 768}. \tag{23}$$

Let $\mathbf{h}$ be an intermediate hidden representation from the diffusion transformer $\mathbf{v}_\theta$, and let $f_\omega(\cdot)$ be a trainable projection head. We define a patch-wise alignment loss:

$$\mathcal{L}_{\mathrm{REPA}}(\theta,\omega) := -\mathbb{E}_{\hat{I},\epsilon,t}\left[\frac{1}{K}\sum_{k=1}^{K}\mathrm{sim}\left(\hat{\phi}_k, f_\omega(\mathbf{h}_k)\right)\right], \tag{24}$$

where k indexes patches and K is the number of patches. We compute this loss at the output of each transformer block using its corresponding projection head, and average it across layers.

The final objective adds the projection loss to the standard flow-matching loss:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{velocity}} + \lambda_{\mathrm{REPA}}\,\mathcal{L}_{\mathrm{REPA}}, \tag{25}$$

where we set $\lambda_{\mathrm{REPA}} = 0.1$ in all experiments.

	Methods	Paired				Unpaired	
		SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
(I)	Baseline	0.889	0.055	5.543	0.870	9.641	1.323
(II)	(I) + Feature Alignment (REPA) [72]	0.885 (−0.45%)	0.087 (−58.18%)	13.718 (−147.48%)	2.103 (−141.72%)	15.621 (−62.03%)	3.684 (−178.46%)
(III)	(I) + Query-Key Correspondence Alignment **(Ours)**	**0.907** (+2.02%)	**0.048** (+12.73%)	**4.962** (+10.48%)	**0.565** (+35.06%)	**8.763** (+9.11%)	**0.880** (+33.48%)

Table D.2: **Alignment Target Ablation on VITON-HD [12].**

Comparison and Analysis. Tab. D.2 compares the two alignment targets, showing that CORAL achieves better try-on quality with higher SSIM and lower LPIPS. Since intermediate features jointly encode both spatial and appearance information, directly aligning them affects the inherent representations learned during pretraining. By supervising query-key matches, CORAL strengthens

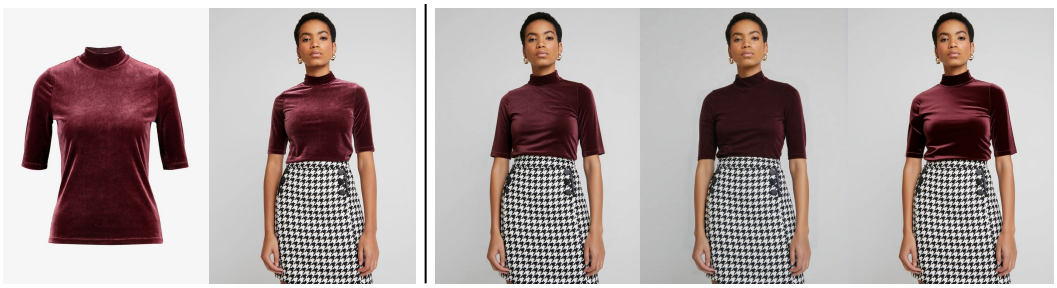


Figure D.1: **Intermediate Feature Alignment vs. Ours.**

person-garment alignment while largely preserving the model's learned appearance representation. In contrast, feature-level alignment shows a larger change in intermediate features (Fig. D.1), followed by a substantial performance drop across all metrics in Tab. D.2.

E Additional Analysis on Pseudo Ground-Truth

E.1 Robustness to Pseudo-GT Errors.

A natural concern is that the pseudo-GT correspondence may contain errors in challenging cases (e.g., complex poses, occlusions, ambiguous garment regions), which could propagate into the training signal. We address this concern through cycle consistency filtering: a correspondence between a person token and a garment position is retained only if the round-trip mapping (person $\rightarrow$ garment $\rightarrow$ person) returns to within a threshold of the original person position. Correspondences that fail this check are excluded from the supervision signal.

Tab. E.3 reports the effect of varying the cycle-consistency threshold on VITON-HD. Stricter filtering (threshold of 3) yields the best performance, confirming that cycle consistency effectively removes erroneous correspondences. Even without filtering (∞), CORAL still improves over the baseline, indicating that the supervision also tolerates a moderate amount of noise.

Threshold	Avg. #reliable	SSIM ↑	LPIPS ↓	FID ↓
∞ (no filtering)	6.2M	0.905	0.057	5.098
6	5.8M	0.907	0.050	5.021
3 (default)	4.2M	**0.908**	**0.048**	**5.009**

Table E.3: **Effect of cycle consistency threshold on VITON-HD.** Stricter filtering retains fewer but cleaner correspondences and yields the best performance.

E.2 Generality across Pseudo-GT Sources.

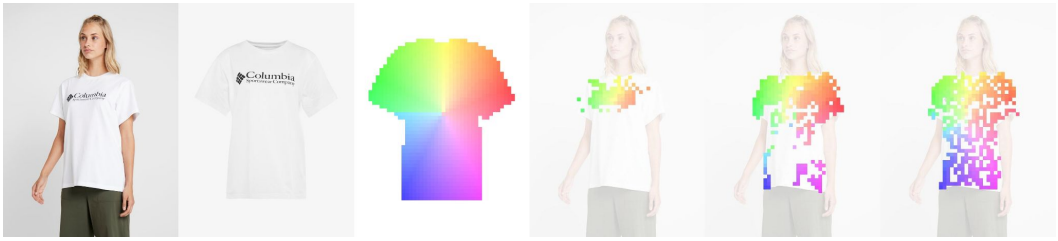


Figure E.2: **Visualization of different pseudo-correspondence sources.** For each spatial position on the garment, color-coded with a rainbow position map, we visualize the matched location on the person predicted by different correspondence sources: LoMa, DIFT, and DINOv3. DINOv3 produces the most spatially coherent mappings, while LoMa yields sparser and less consistent matches. Our method is compatible with any of these sources, with DINOv3 providing the strongest supervision signal.

CORAL relies on an external matching method to provide pseudo ground-truth correspondences. The choice of this matching method is independent of CORAL's design: any method that produces reliable correspondences between person and garment images can be **plugged in**. To verify this, we evaluate four alternatives spanning two categories: vision foundation models repurposed for correspondence (DINOv2 [52], DINOv3 (our default) [56], and DINOv2 combined with diffusion features [73]), and specialized dense matching methods such as LoMa [50] and DIFT [61].

Pseudo-GT Source	SSIM ↑	LPIPS ↓	FID ↓
Baseline (no supervision)	0.889	0.055	5.543
DINOv2 [52]	0.905	0.055	5.099
SD + DINOv2 [73]	0.907	0.049	5.087
LoMa [50]	0.891	0.054	5.540
DIFT [61]	0.902	0.055	5.321
DINOv3 (default) [56]	**0.908**	**0.048**	**5.009**

Table E.4: **Generality across pseudo-GT sources on VITON-HD.** CORAL improves over the baseline regardless of which correspondence source is used.

Fig. E.2 visualizes the correspondences extracted from each method on the same person-garment pair. While the matching quality varies across methods, all of them produce reasonable person-garment correspondences that can serve as the supervision signal for CORAL.

Tab. E.4 reports the quantitative results when each correspondence source is used. All four alternatives lead to substantial improvement over the baseline, confirming that CORAL is general across pseudo-GT sources. We also observe a slight performance increase with stronger correspondence sources, suggesting that future improvements in matching methods can directly benefit CORAL without any architectural modification.

E.3 Behavior under Challenging Conditions.

DINOv3 correspondence quality varies with input characteristics. Fig. E.3 shows three representative cases: **(a)** extreme viewpoint mismatch between person and garment (from VITON-HD), where DINOv3 distinguishes the two views and retains matches mainly along view-invariant boundary regions; **(b)** significant illumination change (from PPR10K), where matches concentrate in regions with consistent appearance; and **(c)** layered garments (from VITON-HD), where matches on regions without a counterpart in the reference are removed.

In all three cases, the removed regions are those without a reliable counterpart in the reference: the back of the body in (a), regions altered by strong illumination in (b), and the inner garment worn beneath the target garment in (c), which do not appear in a flat product photo. These regions are trained only with the diffusion loss, as in the baseline. Since they have no reliable position to learn, removing their matches does not lose useful supervision, and keeping them by loosening the filter hurts performance (Tab. E.3). The check does not catch every case (Appendix O), but the remaining regions, such as view-invariant boundaries, consistently lit regions, and outer garment parts, still receive supervision and anchor the garment structure.

To assess our method's behavior under in-the-wild conditions, we evaluate cycle-consistency filtering on PPR10K [41], which contains heavy occlusions, illumination changes, and challenging poses. We sweep the threshold γ across $\{3, 6, \infty\}$ (Tab. E.5). The number of reliable correspondences after filtering is significantly lower on PPR10K (1.1M vs. 4.2M at $\gamma = 3$ on VITON-HD), yet CORAL still improves over the no-supervision baseline across all threshold settings, and stricter filtering consistently yields better performance. This confirms that cycle-consistency filtering effectively prevents pseudo-label error propagation, and our supervision adapts gracefully to varying pseudo-GT density.

Threshold γ	Avg. # reliable points	SSIM ↑	LPIPS ↓	FID ↓
w/o pseudo-GT	—	0.897	0.068	52.201
∞	3.8M	0.899	0.064	51.995
6	2.5M	0.901	0.061	51.892
3 **(default)**	**1.1M**	**0.929**	**0.050**	**51.423**

Table E.5: **Robustness of cycle-consistency filtering on PPR10K.** Stricter thresholds yield better metrics, with $\gamma = 3$ providing the best balance between supervision density and pseudo-label quality.

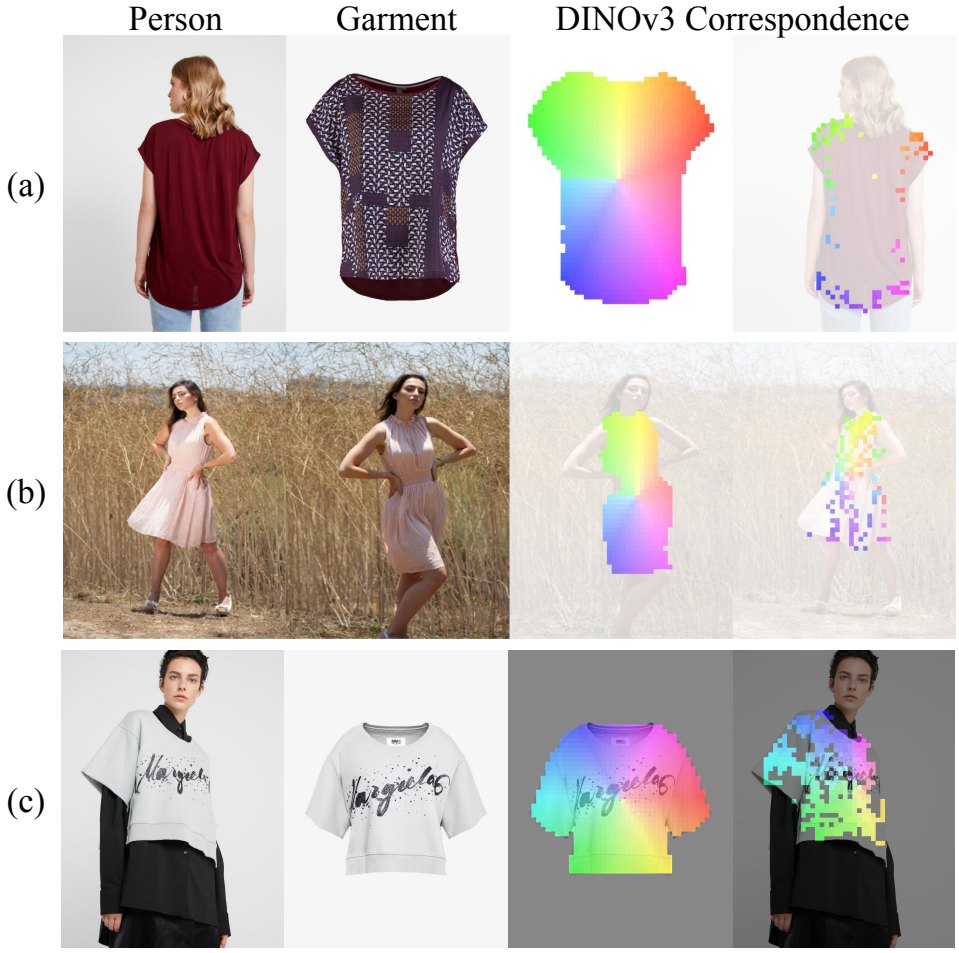


Figure E.3: **DINOv3 correspondence under extreme viewpoint, illumination variation, and layered garments.** For each garment position (color-coded with rainbow map), we visualize matched person locations predicted by DINOv3. **(a)** Opposite viewpoints from VITON-HD dataset (back-view person, front-view garment): DINOv3 distinguishes the two views, with correspondences remaining mainly along view-invariant boundary regions. **(b)** Significant illumination change: matches concentrate in regions with consistent appearance. **(c)** Layered garments from VITON-HD: matches on regions without a counterpart in the reference, such as the inner layer, are removed. Although coverage varies under such conditions, our supervision benefits from these partial yet reliable matches.

F Additional Ablation Studies

F.1 Hyperparameter Sensitivity

We analyze the sensitivity of CORAL to the choice of loss weights λ_{corr} and λ_{ent} in:

$$\mathcal{L}_{\text{CORAL}} = \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}. \tag{26}$$

Loss Weight Ablation. We sweep λ_{corr} and λ_{ent} across eight settings on VITON-HD, with all other hyperparameters fixed to the default training configuration. Tab. F.6 reports the results. CORAL maintains stable performance across a range of weight values, with the default setting ($\lambda_{\text{corr}} = 0.01$, $\lambda_{\text{ent}} = 0.1$) providing the best overall balance. Performance degrades when either weight is too small (insufficient supervision strength) or too large (loss imbalance disrupting the diffusion objective). These results indicate that CORAL is not overly sensitive to weight selection within reasonable ranges, and the chosen defaults generalize without per-dataset tuning.

λ_{corr}	λ_{ent}	SSIM ↑	LPIPS ↓	FID ↓
0	0	0.889	0.055	5.543
0	0.1	0.905	0.052	5.482
1.0	0.1	0.887	0.061	6.092
0.1	0.1	0.908	0.049	5.032
0.01	**0.1**	**0.908**	**0.048**	**5.009**
0.01	1.0	0.907	0.049	5.018
0.01	0.01	0.906	0.050	5.449
0.01	0	0.906	0.051	5.451

Table F.6: **Loss weight sensitivity on VITON-HD.** Default values are $\lambda_{\text{corr}} = 0.01$ and $\lambda_{\text{ent}} = 0.1$ (highlighted). Both losses are stable across a wide range of weights, with $\mathcal{L}_{\text{corr}}$ degrading at extreme values.

F.2 Additional Visualizations of Loss Component Ablations

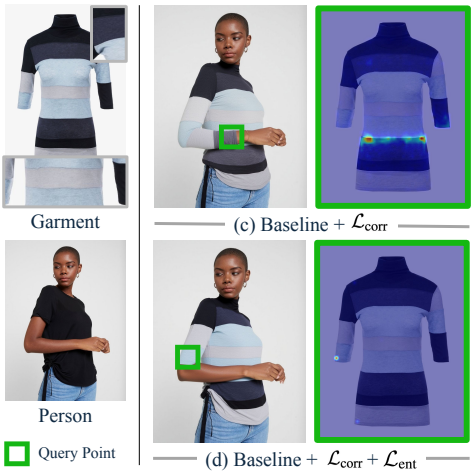


Figure F.4: **Additional attention visualizations for Fig. 6.** Queries are placed at the sleeve edge, where the outputs of (c) Baseline + $\mathcal{L}_{\text{corr}}$ and (d) CORAL differ.

Fig. F.4 provides additional attention visualizations for Fig. 6. Beyond the query points in Fig. 6, we place queries at the sleeve edge, where the outputs of (c) Baseline + $\mathcal{L}_{\text{corr}}$ and (d) CORAL differ. In (c), the attention spreads over the reference and reaches a dark band, whose fabric is drawn onto the arm. In (d), the attention narrows to the light-blue end of the reference sleeve. Since $\mathcal{L}_{\text{corr}}$ is satisfied once the attention averages out near its target, it leaves the attention spread, and content from other regions leaks into the output. Adding $\mathcal{L}_{\text{ent}}$ concentrates the attention on the corresponding position, which directly corrects the visible error.

G Additional Details on Motivation Experiment

We provide additional details and visual analysis for the motivation experiment in Fig. 2. The purpose of this experiment is not to introduce an inference-time guidance method, but to examine whether correspondence-aware query–key matching can affect the quality of DiT-based VTON. To this end, we apply a minimal intervention to the baseline model at inference time and visualize how the generated result changes.

Experimental setup. Given a selected query point on the person region, we obtain its corresponding location inside the garment region using the pseudo correspondence described in Sec. E. We then modify only its query–key matching to garment keys around the matched location. All other tokens, model weights, and intermediate feature representations remain unchanged. This allows us to observe the effect of locally guiding person-to-garment matching without changing the overall model.

Implementation details. For the matched garment location, we apply the correspondence bias to a small 5×5 neighborhood. This provides a localized guidance region while avoiding an overly restrictive single-token constraint. Our entropy-based objective shows a consistent effect in Fig. 7, reducing overly diffuse attention and bringing the attention entropy closer to the level observed in stronger baselines

Results. As shown in Fig. H.5, the baseline attention for the selected query point attends to an inaccurate or diffuse garment region. In the corresponding try-on result, this leads to imperfect transfer of local garment appearance, such as misplaced text or distorted patterns. After applying the local query–key bias, the selected person token attends more strongly to the corresponding garment region, and the generated result better preserves the local garment appearance around the queried area.

This experiment should be interpreted as a diagnostic analysis rather than a practical inference-time method. Since the intervention is applied only to selected query–key matching and does not update model parameters, the improved result suggests that query–key matching is a meaningful control point for person–garment correspondence in DiT-based VTON. Motivated by this observation, CORAL replaces manual inference-time intervention with train-time correspondence supervision, encouraging the model to learn correspondence-aware query–key matching directly.

H Runtime and Memory

$\mathcal{L}_{\text{CORAL}}$	GPU	GPU# (T/I)	Batch (T/I)	Speed (T/I)	VRAM (T/I)
✓	H100	4 / 1	1 / 1	0.42 / 0.063	56 / 40
✗	H100	4 / 1	1 / 1	0.45 / 0.063	52 / 40

Table H.7: **Speed and memory.** T denotes training and I denotes inference. Speed is in img/sec and VRAM is in GB.

In Table. H.7, we report the hardware setup, throughput, and peak GPU memory for each baseline setting.

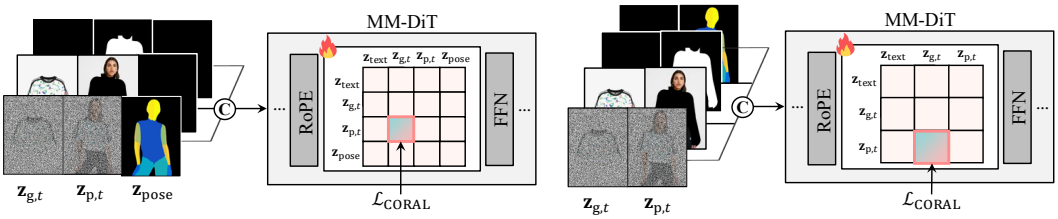


Figure H.5: **Zoomed-in visualization of the motivation experiment**, Pink markers denote the query points (the right bottom of the letter "R")

I CORAL on Other Baselines

CORAL requires no architectural change, no additional module, and no change at inference. It can be attached to any DiT architecture that exposes a person-garment attention map for supervision, and the model is then fine-tuned with the loss. To demonstrate this, we extend our application to additional DiT baselines.

I.1 Baselines with Different Pose Injection Strategies



(a) Pose Token Concat **(Ours)** (b) Pose Channel Concat

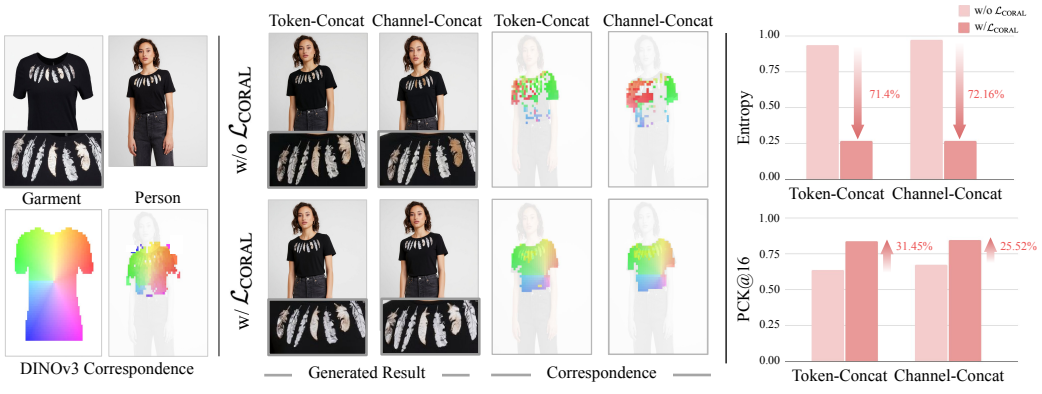
Figure I.6: **Architecture Comparison.**

Pose Token Concat. To leverage the DiT's token-based processing over the full sequence, we consider a baseline that injects pose as additional tokens. As shown in Fig. I.6 (a), we concatenate the pose condition $\mathbf{z}_{\text{pose}}$ along the token dimension. This preserves positional structure in the token space and provides explicit pose correspondences to the model. Our main experiments use this **Pose Token Concat.** baseline. For more details, see Sec. 3.2 of the main paper and Sec. A of Appendix.

Pose Channel Concat. As illustrated in Fig. I.6 (b), we additionally design a baseline that injects the pose condition $\mathbf{z}_{\text{pose}}$ along the channel dimension rather than the token dimension. Specifically, we first form a pose diptych $\mathbf{z}_{\text{pose-diptych}}$ by placing $\mathbf{z}_{\text{pose}}$ next to a zero-padded panel to match the diptych layout. We then concatenate $\mathbf{z}_t$, $\mathbf{z}_{\text{diptych}}$, $\mathbf{m}_{\text{diptych}}$, and the $\mathbf{z}_{\text{pose-diptych}}$ channel-wise, and feed the resulting tokens to the input projection layer within `FLUX.1-Fill-dev`. This increases the projection input dimension from 384 to 448, so we expand the projection layer and zero-initialize the newly added weights. Copying the original projection weights into the expanded channels led to degraded performance and unstable convergence.

Methods	$\mathcal{L}_{\text{CORAL}}$	Paired				Unpaired	
		SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
Pose Token Concat	✗	0.889	0.055	5.543	0.870	9.641	1.323
	✓	**0.907** (+2.02%)	**0.048** (+12.73%)	**4.962** (+10.48%)	**0.565** (+35.06%)	**8.763** (+9.11%)	**0.880** (+33.48%)
Pose Channel Concat	✗	0.886	0.060	5.554	0.540	9.841	1.221
	✓	0.888 (+0.23%)	0.059 (+1.67%)	5.372 (+3.28%)	0.744	9.323 (+5.26%)	1.077 (+11.79%)

Table I.8: **Effect of $\mathcal{L}_{\text{CORAL}}$ across Baselines with Different Pose Injection Strategies for VITON-HD [12].**



(a) Qualitative Analysis (b) Quantitative Analysis

Figure I.7: **Effect of $\mathcal{L}_{\text{CORAL}}$ across Baselines with Different Pose Injection Strategies.**

Effect of $\mathcal{L}_{\text{CORAL}}$. We evaluate CORAL on two pose-injection baselines, **Pose Token Concat.** and **Pose Channel Concat.**, with and without $\mathcal{L}_{\text{CORAL}}$. As reported in Tab. I.8, adding $\mathcal{L}_{\text{CORAL}}$ consistently improves VTON performance for both baselines, demonstrating that CORAL is model-agnostic across DiT variants that expose a person→garment attention. Between the two baselines, Pose Token Concat performs notably better than Pose Channel Concat. This is expected because token concatenation preserves spatially aligned pose tokens that the DiT can attend to directly, whereas channel concatenation injects pose more indirectly through the input projection. Fig. I.7 (a) further shows that $\mathcal{L}_{\text{CORAL}}$ refines person→garment mismatches in attention, leading to more accurate garment transfer: for instance, the fourth and fifth feather patterns no longer bleed into nearby brown tones and remain clean white as in the garment reference. Fig. I.7 (b) quantitatively supports that across both baselines, $\mathcal{L}_{\text{CORAL}}$ reduces attention entropy and increases PCK, indicating sharper and more correct correspondences.

I.2 An Additional Efficient Baseline

We introduce an additional efficient baseline trained under a different setup to test whether CORAL's improvements remain consistent beyond the settings in Sec. I .

Architectural Details. Additional baseline is designed to be computationally efficient. Following the compact token integration method in prior work [60], we first restrict denoising processes to the garment region in $\mathbf{z}_{\text{p}}$ using downsampled mask $\mathbf{m}_{\text{e}}$, since denoising the full person latent adds unnecessary computation and can also perturb garment-unrelated areas. At diffusion timestep t, the noisy person latent is defined as:

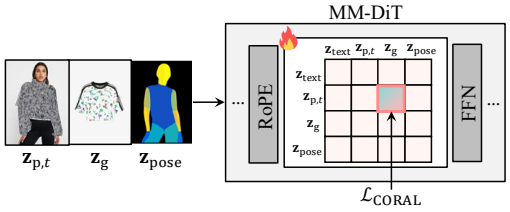


Figure I.8: **Efficient Baseline Architecture.**

$$\mathbf{z}_{\text{p},t} = \Big((1-t)\mathbf{z}_{\text{p}} + t\epsilon\Big) \odot \mathbf{m}_{\text{p}} + \mathbf{z}_{\text{p}} \odot (1-\mathbf{m}_{\text{p}}). \tag{27}$$

We also keep the garment latent clean to avoid unnecessary denoising process on the garment input. For pose conditioning, we concatenate the pose condition $\mathbf{z}_{\text{pose}}$ along the token dimension, as discussed in Sec. 3.2.

Implementation Details. We initialize our model from `FLUX.1-dev` and finetune with LoRA [26] with rank 128. We train with Prodigy [44] optimizer using a learning rate of 1.0 and batch size of 1 on 2 NVIDIA A100 GPUs. We train and report results at a resolution of 1024 x 1024. For training with $\mathcal{L}_{\text{CORAL}}$, same learning rate and batch size are used, with loss weights $\lambda_{\text{corr}} = 0.01$ and $\lambda_{\text{ent}} = 0.1$. All other implementation details follow Sec. A of Appendix.

$\mathcal{L}_{\text{CORAL}}$	VITON-HD						DressCode					
	Paired				Unpaired		Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
✗	0.864	0.084	9.521	1.353	11.251	0.456	0.905	0.065	5.752	0.473	8.723	0.736
✓	**0.871** (+0.81%)	0.084	**6.271** (+34.1%)	**0.563** (+58.5%)	**10.285** (+8.62%)	**0.315** (+30.9%)	0.905	**0.061** (+6.15%)	**5.221** (+9.22%)	0.581	**8.70** (+0.23%)	**0.193** (+73.78%)

Table I.9: **Effect of $\mathcal{L}_{\text{CORAL}}$ on Additional Efficient Baseline for VITON-HD [12] and DressCode [45].**

Effect of $\mathcal{L}_{\text{CORAL}}$. Applying CORAL improves the efficient baseline across overall metrics, with particularly large improvements compared to our main baseline. In this efficient training setup with LoRA finetuning, the baseline performance is relatively weaker, yet CORAL still yields substantial gains. The largest improvements appear in distribution-based metrics, which indicates that CORAL also enhances overall realism of generated outputs, not only the perceptual fidelity.
Overall, the improvements across baselines confirm the effectiveness of $\mathcal{L}_{\text{CORAL}}$ and highlight its suitability for DiTs, where attention provides an explicit controllable correspondence mechanism.

I.3 Baseline with a Different Backbone

To test whether the gain depends on the backbone → To examine whether the gain generalizes to other backbones, we further attach $\mathcal{L}_{\text{CORAL}}$ to a try-on model built on FLUX.2, a more recent diffusion transformer. As shown in Tab. I.10, CORAL improves every metric in both the paired and unpaired settings. To examine whether the gain transfers across backbones, we further attach $\mathcal{L}_{\text{CORAL}}$ to a try-on model built on FLUX.2, a more recent diffusion transformer. As shown in Tab. I.10, CORAL improves every metric in both the paired and unpaired settings.

$\mathcal{L}_{\text{CORAL}}$	Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
✗	0.894	0.053	5.433	0.529	9.682	0.860
✓	**0.907**	**0.047**	**4.941**	**0.382**	**9.018**	**0.674**
	+1.45%	+11.3%	+9.06%	+27.8%	+6.86%	+21.6%

Table I.10: **Effect of $\mathcal{L}_{\text{CORAL}}$ on a FLUX.2 Baseline for VITON-HD [12]**.

J VLM-based Evaluation

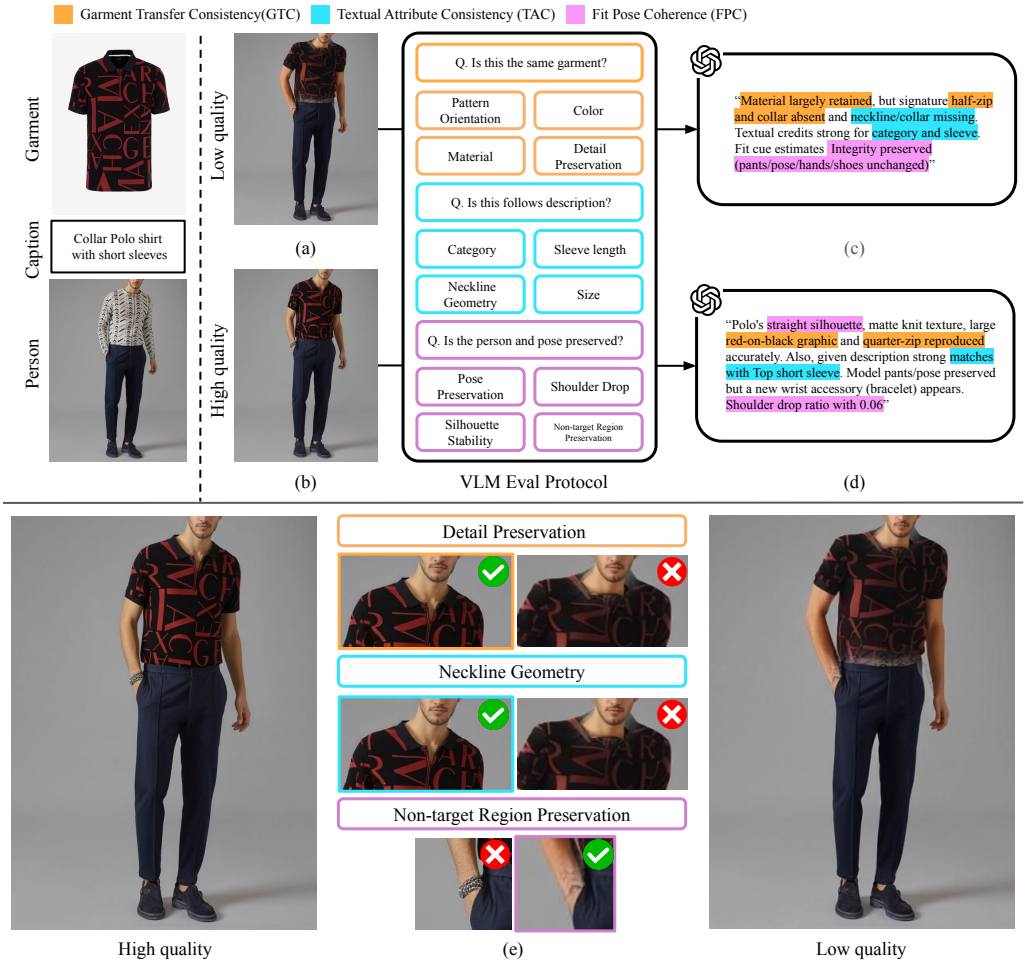


Figure J.9: **VLM-based Evaluation Results Comparison.**

Why VLM-based Evaluation is Needed? As discussed in Sec. 4.1, standard metrics such as FID [57], KID [4], SSIM [15] and LPIPS [74] are insufficient for evaluating VTON, as they mainly reflect overall realism or low-level similarity and do not assess whether the specific reference garment is faithfully transferred and plausibly worn. Following Sec. 4.1 of the main paper, we therefore propose a VLM-based evaluation protocol with three criteria, including Garment Transfer Consistency (GTC), Textual Attribute Consistency (TAC), and Fit Pose Coherence (FPC). For all VLM-based evaluations, we use GPT-4o as our evaluator. The full input prompt used for VLM evaluation is provided in the Fig. O.22.

Fig. J.9 motivates the need for this protocol. Although the two outputs in Fig. J.9 (a) and (b) are generated from the same input garment and person, they differ in whether garment details, attributes and fit-pose are preserved. Our evaluator produces GTC, TAC, and FPC scores together with natural language justifications that explain which visual evidence supports each score, as shown in Fig. J.9 (c) and (d). (e) provides the detailed view of the generated images aligned with each criterion. This design is motivated by recent findings [10, 30, 9] that VLMs [16, 38] excel at image and multi-image reasoning, making them well suited to judge garment-specific fidelity beyond distribution-level similarity.

VLM Evaluation Setup. For all VLM-based evaluations, we use GPT-4o as our evaluator. For each generated try-on result, the VLM receives 4 inputs: the input person image, the reference garment image, the generated output, and a caption describing the garment attributes (extracted

from the reference garment using GPT-4o [51]). The VLM is prompted to produce four sub-scores covering *textual alignment* (TAC), *visual match* (GTC), *fit quality*, and *non-target integrity*, the latter two combined into FPC. Each sub-score is reported on both a continuous (0.0-5.0) and discrete (1-5) Likert scale, accompanied by a brief JSON-formatted justification. To prevent overrating in challenging cases, the prompt enforces hard score caps when severe failures are detected (e.g., category mismatch, silhouette mismatch). The full prompt specification is provided in Fig. O.22.

Definition of Each Criterion. We evaluate three aspects that existing metrics inadequately reflect:

*(1) Garment Transfer Consistency (**GTC**)* measures whether the output preserves the garment's visual appearance, including its overall shape and silhouette as well as details like patterns, logos, and texture.

*(2) Textual Attribute Consistency (**TAC**)* captures whether the output satisfies garment attributes specified in text or metadata, covering requirements that may be ambiguous from the reference image alone (e.g., length, neckline, category, and fit).

*(3) Fit Pose Coherence (**FPC**)* reflects whether the garment is worn naturally on the target person, consistent with the input person pose and body geometry. It also considers whether non-target regions (e.g., hair, skin, hands, and other clothing) are preserved without unintended changes.

Reliability of VLM-based Evaluation. The VLM-based rankings are broadly consistent with user study results (Sec. K). To provide more rigorous validation, we have GPT-4o score the same 45 image-question pairs from our user study using the prompt described in Fig. O.22, with sub-scores aggregated into the same three criteria (GTC, TAC, FPC) used in the user study. From the 2,430 total human ratings (54 participants $\times$ 45 questions), we compute the average human score per question and measure Pearson correlation with GPT-4o scores across the 45 averaged data points, yielding $r = 0.71$ ($p < 0.001$, $n = 45$). This represents a large effect size with GPT-4o explaining approximately 50% of the variance in human scores ($r^2 \approx 0.50$), supporting the reliability of our VLM-based protocol as a proxy for human evaluation.

K Human Evaluation

Human Evaluation Details. We conducted a user study to measure virtual try-on quality under human perception and to validate that our VLM-based evaluation reflects human preference. Participants rated each generated result on the same three criteria used in Sec. 4.1, including Garment Transfer Consistency (GTC), Textual Attribute Consistency (TAC), and Fit-Pose Coherence (FPC), using a 5-point scale (from 1 to 5) for each criterion. Fig. O.21 presents example questions from the user study.

To make the criteria concrete and reduce ambiguity, we provided references for each question. For GTC, we showed the input garment image so participants could directly compare whether shape, patterns, and logos were preserved. For TAC, we provided a short textual description of the garment attributes, automatically extracted from the reference garment using a pretrained vision-language model [51]. For FPC, we showed the input person image so participants could judge whether the garment is worn naturally and follows the target pose and body geometry without altering non-target regions.

A total of 54 participants answers 45 questions. For a fair comparison, we sampled outputs from a large pool sharing the same input person image and garment image across 5 comparison models, including OOTDiffusion [71], IDM-VTON [13], CatVTON [14], Any2AnyTryOn [18] and CORAL (Ours) to ensure intra-rater reliability.

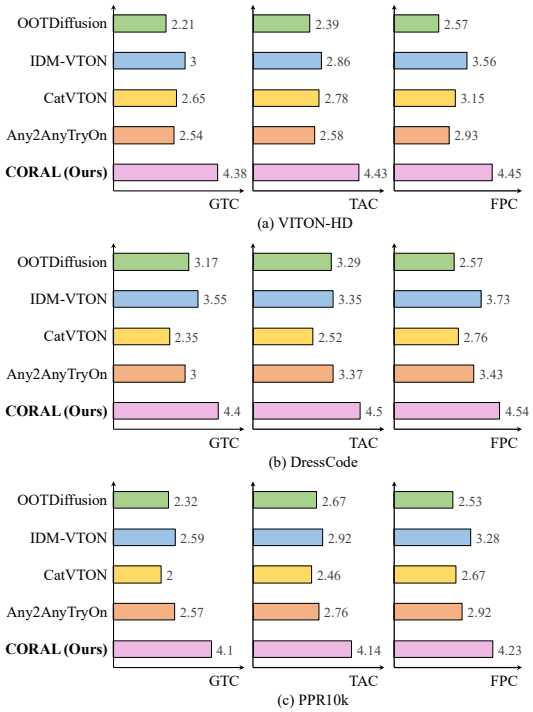


Figure K.10: **User Study Results.**

Human Evaluation Results. Fig. K.10 shows the results, demonstrating that our model outperforms others across all criteria.

Correlation with VLM-Evaluation. We find that our VLM-based scores align well with the user study outcomes, indicating that the proposed VLM evaluation reflects human preference more faithfully than standard metrics. In both Tab. 4 , methods that receive higher GTC/TAC/FPC scores also tend to be preferred by human participants, and CORAL performs strongly in both evaluations. This agreement suggests that GTC, TAC, and FPC capture perceptually important failure cases in try-on quality.

L Evaluation on Additional Datasets

L.1 Motivation

We evaluate whether the correspondences learned by CORAL generalize beyond standard VTON benchmarks and still leads to high-quality try-on results under input conditions that are less curated than the benchmark setting. For the person input, we consider photos with broader variation in poses, viewpoints, backgrounds, framing, and occlusions than typical benchmarks. For the garment input, we consider reference images that are not limited to clean product images.

Most existing VTON benchmarks follow the in-shop garment transfer setting, where the reference garment is a flat product image and the target person is photographed in a controlled studio environment. However, this setting does not fully reflect real-world inputs. User-provided person photos are often taken in everyday environments with diverse framing, non-frontal viewpoints, more dynamic poses, and frequent occlusions by accessories or surrounding objects. Garment reference images can also differ substantially in practice. It is often inconvenient for users to search for and provide a clean product photo for every garment. Instead, they typically rely on readily available photos, such as a casually taken garment photo with cluttered backgrounds or a worn-garment image from another person, which introduces additional noises compared to product images.

For this reason, we additionally build paired and unpaired test sets to evaluate CORAL under more challenging conditions while keeping the evaluation reliable. We evaluate with worn-garment references, where the reference garment is worn by another person, which is typically more challenging than a clean product image but still provides sufficient garment cues for try-on.

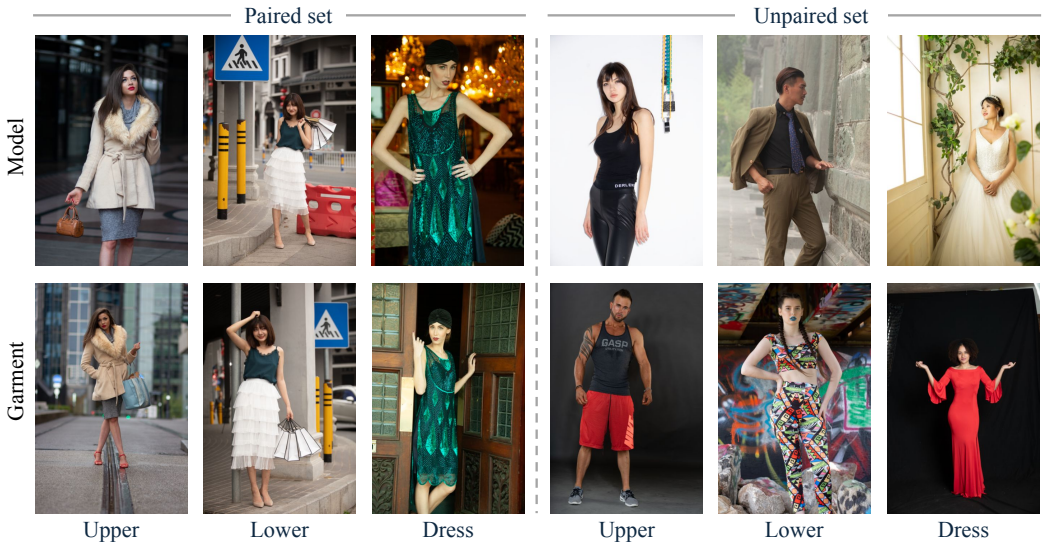


Figure L.11: **Additional Evaluation Dataset Samples.**

L.2 Evaluation Dataset Details

We additionally build paired and unpaired evaluation sets using PPR10K [41], since it offers high-resolution full-body portraits captured under diverse real-world conditions and provides grouped images of the same subject, which enables stable paired evaluation with consistent garment identity.

Dataset Overview. PPR10K is a large scale, high quality portrait photo retouching dataset designed to cover diverse real world conditions. It contains 11,161 high resolution raw portrait photos organized into 1,681 groups, where each group includes 3 to 18 photos of the same subject captured in the same scene within a short time span. The dataset was curated to cover diverse shooting purposes such as weddings, graduations, and advertisements. It also includes diverse subjects (e.g., children and people from diverse regions), varied indoor and outdoor backgrounds, and a wide range of lighting

conditions from day to night and across seasons. The portraits cover diverse framing from face-centric to full-body shots, and include various viewpoints excluding fully back-facing.

Dataset Preprocessing. We first filter person images to keep upper-body to full-body shots. We remove face-centric crops by extracting 2D keypoints with OpenPose [6] and retaining images where the body is visible at least down to the knees. We further heuristically filter samples by viewpoints using IUV maps extracted from DensePose [19] to exclude fully back and side views. For the reference garment setting, we exclude images where the garment region is too small to provide meaningful detail, as well as extreme cases such as overly dark or overexposed images and heavy occlusions. We use the same preprocessing pipeline as in the standard VTON benchmarks for extracting DensePose conditions. We obtain input masks following prior work [13]. After preprocessing, the final split contains 358 dresses, 320 upper-body samples, and 195 lower-body samples. In the paired setting there are 309 samples (138 dresses, 86 upper-body, 85 lower-body) and in the unpaired setting there are 564 samples (220 dresses, 234 upper-body, 110 lower-body).

M Person-To-Person Garment Transfer

Motivation. As mentioned in Sec. L.1, worn-garment transfer is a practically important setting where the garment is observed on another clothed person rather than as a clean product photo. We refer to this setup as *person-to-person* garment transfer. Compared to in-shop product photos, worn-garment references are more common in real world and also show how the garment looks when worn, but it is harder to determine what should be transferred from in-the-wild reference images. A standard in-shop setting often addresses this by estimating a garment mask from the reference image of a clothed person, but this mask can be unreliable under heavy occlusions or complex backgrounds. As illustrated in the first row of Fig. M.12, occlusion by hair can remove garment regions in the mask, and the try-on result can transfer this missing shape, making the garment look cut out in the same area. Garment masks can also remove global shape information such as overall length and silhouette, which is particularly critical for lower-garment transfer. As shown in Fig. M.12, for skirts and dresses, a garment-only reference cropped by a mask can hide where the garment ends on the body, making it difficult to preserve the correct length. For this reason, we train a separate model for *person-to-person* garment transfer using full reference images without estimating garment masks, and apply CORAL to improve person–garment correspondence under these in-the-wild references.



Figure M.12: **Person-To-Person Garment Transfer Motivation.**

Architectural Details. In the *person-to-person* garment transfer setting, the garment reference image is provided through another clothed person image rather than a clean product image. We therefore denote the target and reference person images as $I_{\mathrm{p_{tgt}}}$ and $I_{\mathrm{p_{ref}}}$, and encode them with a VAE to obtain the corresponding latents $\mathbf{z}_{\mathrm{p_{tgt}}}$ and $\mathbf{z}_{\mathrm{p_{ref}}}$. Following the same diptych layout as in Sec. 3.2, at diffusion timestep t, we construct the noisy diptych latent as:

$$\mathbf{z}_t = \left[\, \mathbf{z}_{\mathrm{p_{ref}},t} \,\|\, \mathbf{z}_{\mathrm{p_{tgt}},t} \,\right]. \tag{28}$$

For conditioning worn-garment reference image and the masked person image, we use the clean reference person latent and the masked target person latent as:

$$\mathbf{z}_{\mathrm{diptych}} = \left[\, \mathbf{z}_{\mathrm{p_{ref}}} \,\|\, \mathbf{z}_{\mathrm{p_{tgt}}} \odot (1 - \mathbf{m}_{\mathrm{e}}) \,\right], \tag{29}$$

where $\mathbf{m}_{\mathrm{e}}$ is obtained by downsampling M_e into the latent space. We use the same definition of the binary mask canvas $\mathbf{m}_{\mathrm{diptych}}$ as in Sec. 3.2.

Since DensePose [19] can be estimated from both $I_{\text{p}_{\text{tgt}}}$ and $I_{\text{p}_{\text{ref}}}$, we extract two pose conditions aligned to the target and reference persons, denoted as $I_{\text{pose}_{\text{tgt}}}$ and $I_{\text{pose}_{\text{ref}}}$, and then encode them with a VAE to obtain $\mathbf{z}_{\text{pose}_{\text{tgt}}}$ and $\mathbf{z}_{\text{pose}_{\text{ref}}}$. Due to additional pose inputs for the reference worn-garment image, pose injection through token concatenation would increase the number of tokens, leading to higher computation and memory cost. We therefore inject the pose condition along the channel dimension instead. As explained in Sec. I.1 of Appendix, we build the diptych layout for pose conditions, $\mathbf{z}_{\text{pose-diptych}} = [\,\mathbf{z}_{\text{pose}_{\text{ref}}} \,\|\, \mathbf{z}_{\text{pose}_{\text{tgt}}}\,]$, and then concatenate $\mathbf{z}_t$, $\mathbf{z}_{\text{diptych}}$, $\mathbf{m}_{\text{diptych}}$, and the $\mathbf{z}_{\text{pose-diptych}}$ channel-wise before the input projection layer in `FLUX.1-Fill-dev`.

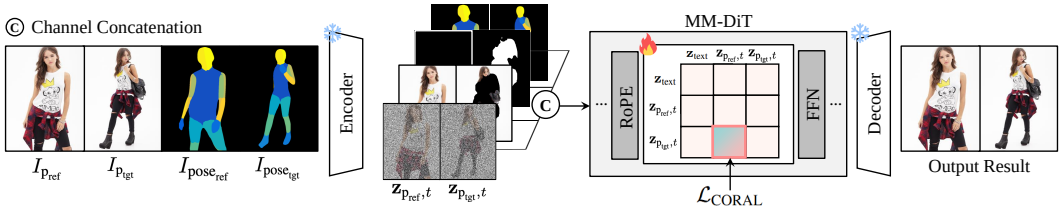


Figure M.13: **Person-To-Person Architecture.**

Training Dataset. Existing VTON benchmarks are built on curated paired sets where the reference image is a flat in-shop product image, typically centered with a clean background. The worn-garment setting, however, does not come with an established benchmark. Even though some public paired datasets are available where multiple subject images share the same garment identity, some images have incorrect garment identity annotations, and there is no annotation indicating which image in each group should be used as the reference worn-garment image. In the worn-garment setting, many images are in-the-wild with different viewpoints, framing, and occasional occlusions, so some images may include garments that are not visible enough to be used as a reliable reference input.

To build paired data for *person-to-person* garment transfer, we leverage DeepFashion [42] and an internal dataset. DeepFashion provides 14,120 pairs for the multi-view try-on task, and each pair typically includes multiple viewpoints across a range of garment categories, with reliable garment identity annotations. We filter out reference images with extreme viewpoints such as strong side or back views using DensePose [19] IUV maps, and overly tight crops using human parsing maps [40]. We also leverage an internal dataset covering tops, bottoms, and dresses, where images are paired by garment identity across different subjects, with garment identities annotated by human annotators. For in-the-wild images that are not full body, human parsing or OpenPose [6] can fail, so we additionally use SAM3 [7] to generate the input masks. In total, our final training set contains 23,678 paired samples.

Quantitative Results. As shown in Tab. M.11, training the model for *person-to-person* garment transfer with worn-garment references achieves better performance than using a model trained with in-shop product photos without retraining. In Tab. M.11, **CORAL-P2P** uses the unmasked reference image at inference time, while the other **CORAL** settings use a garment-only reference extracted with a garment mask. In this more challenging setting, $\mathcal{L}_{\text{CORAL}}$ helps the model focus on the intended garment and avoid transferring details from irrelevant parts of the reference image, which is reflected in the quantitative results of Tab. M.11 across both paired and unpaired settings.

Methods	Paired				Unpaired	
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓
CORAL (w/o $\mathcal{L}_{\text{CORAL}}$)	0.877	0.078	44.117	0.015	56.644	1.202
CORAL (w $\mathcal{L}_{\text{CORAL}}$)	0.915	0.060	43.648	0.011	53.164	1.101
CORAL-P2P (w $\mathcal{L}_{\text{CORAL}}$)	**0.923**	**0.050**	**38.590**	**0.008**	**52.047**	**1.1088**

Table M.11: **Quantitative Comparison on PPR10K.**

Qualitative Results. We provide qualitative results for the *person-to-person* garment transfer in Fig. M.12 and Fig. M.14. Models trained with in-shop product images often struggle when the garment reference is provided as a clothed person image rather than a clean product photo. As shown in Fig. M.12, the visible garment region in the reference image can be incomplete under occlusion, and the generated try-on result follow the edge of visible part, making the garment look cut off or visually deformed. Training on *person-to-person* pairs helps the model produce a more

natural garment shape even under such occlusions, where the shoulder part in shirts is generated seamlessly with **CORAL-P2P**. For lower garments and dresses, garment length is hard to infer from a garment-only reference cropped by a mask, since it removes global information about where the garment ends on the body and its overall silhouette, which can lead to shortened hems or incorrect length, as shown in Fig. M.12. Instead, Fig. M.14 further shows that **CORAL-P2P** preserves overall length and silhouette by conditioning on the full reference person image, keeping the skirt length just above the ankles, as in the reference.

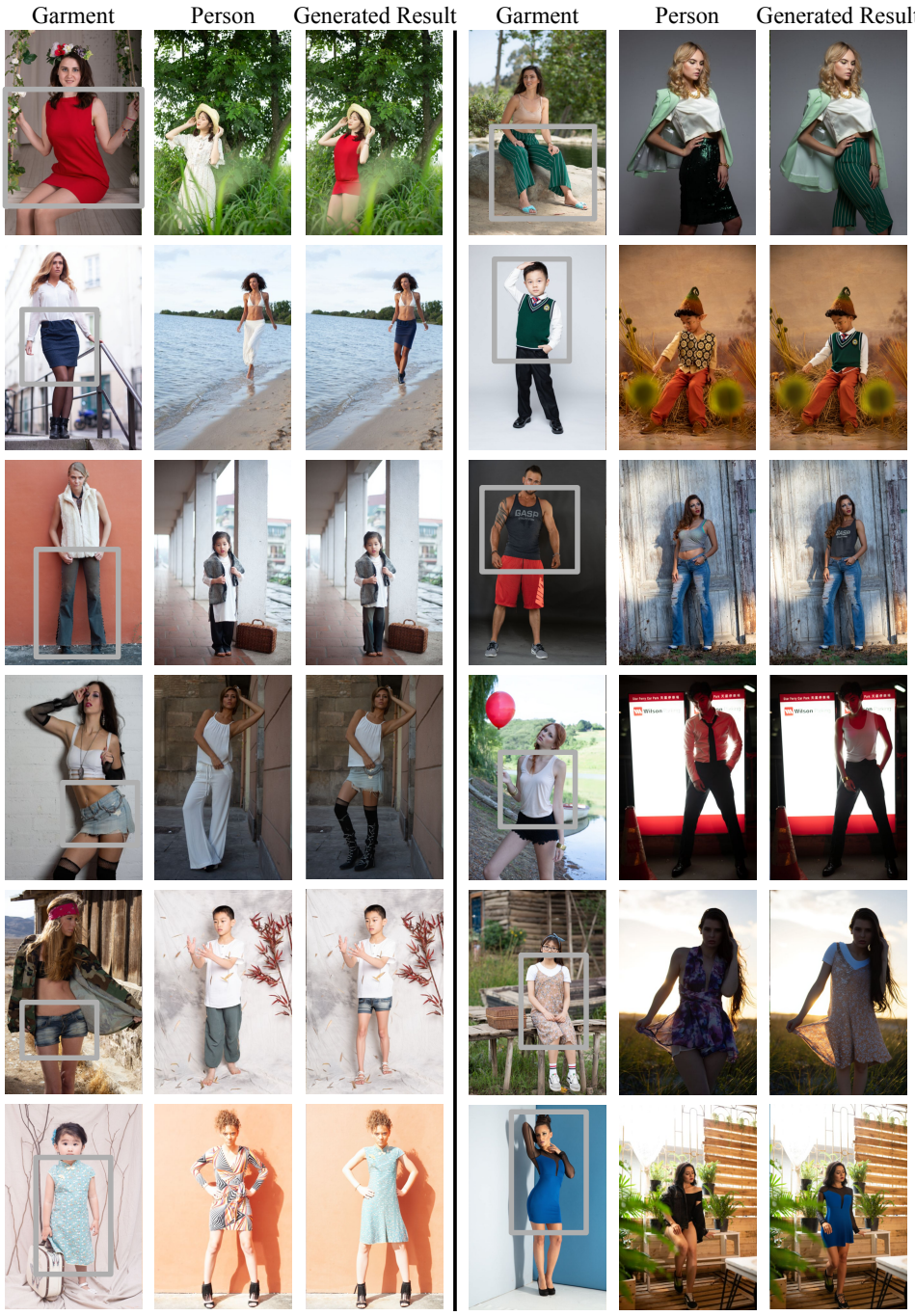


Figure M.14: **Qualitative Results of CORAL-P2P on PPR10K.**

N Additional Qualitative Results

We further compare CORAL with recent VTON methods on standard benchmarks and in-the-wild dataset in Fig. N.15, Fig. N.16, and Fig. N.17. We also report results for the baseline model trained without $\mathcal{L}_{\text{CORAL}}$ and with $\mathcal{L}_{\text{CORAL}}$, showing that CORAL consistently improves try-on quality by enhancing person–garment alignment and reducing misplacement artifacts in Fig. N.19 and Fig. N.20. We additionally evaluate CORAL on PPR10K to assess the generalization of our model trained only on standard virtual try on benchmarks in Fig. N.18.



Figure N.15: **Additional Qualitative Comparison of CORAL on VITON-HD [12]**.



Figure N.16: **Additional Qualitative Comparison of CORAL on DressCode [45].**



Figure N.17: **Additional Qualitative Comparison of CORAL on PPR10K [41].**

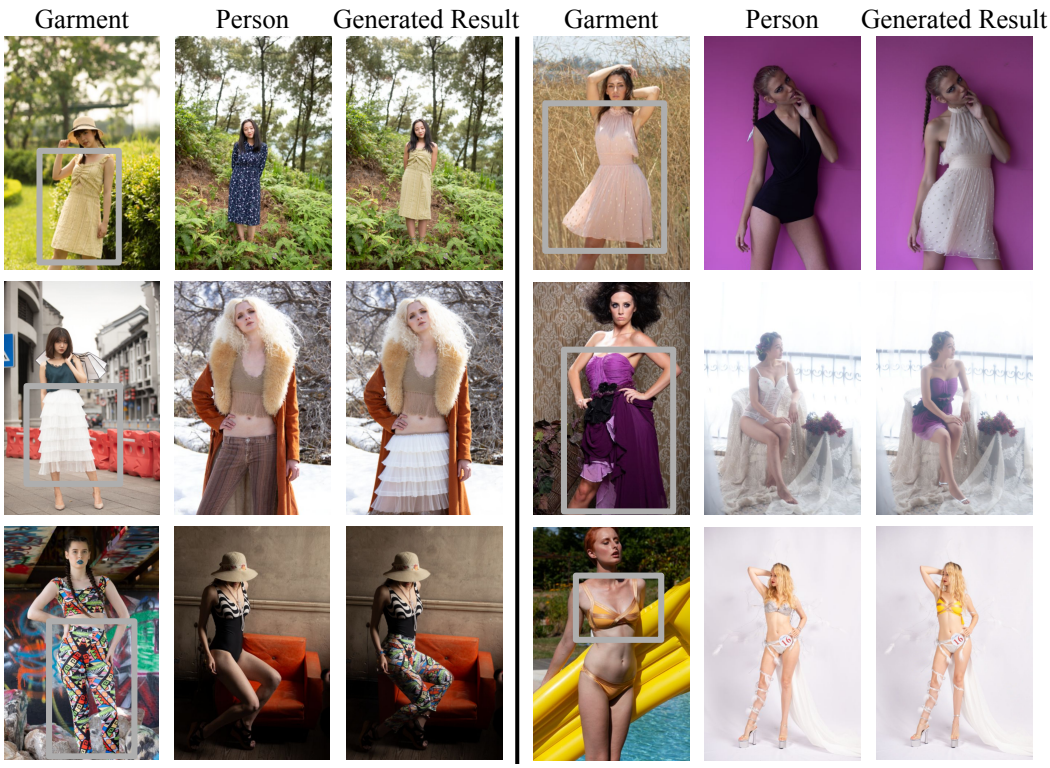


Figure N.18: **Additional Qualitative Results of CORAL on PPR10K.**

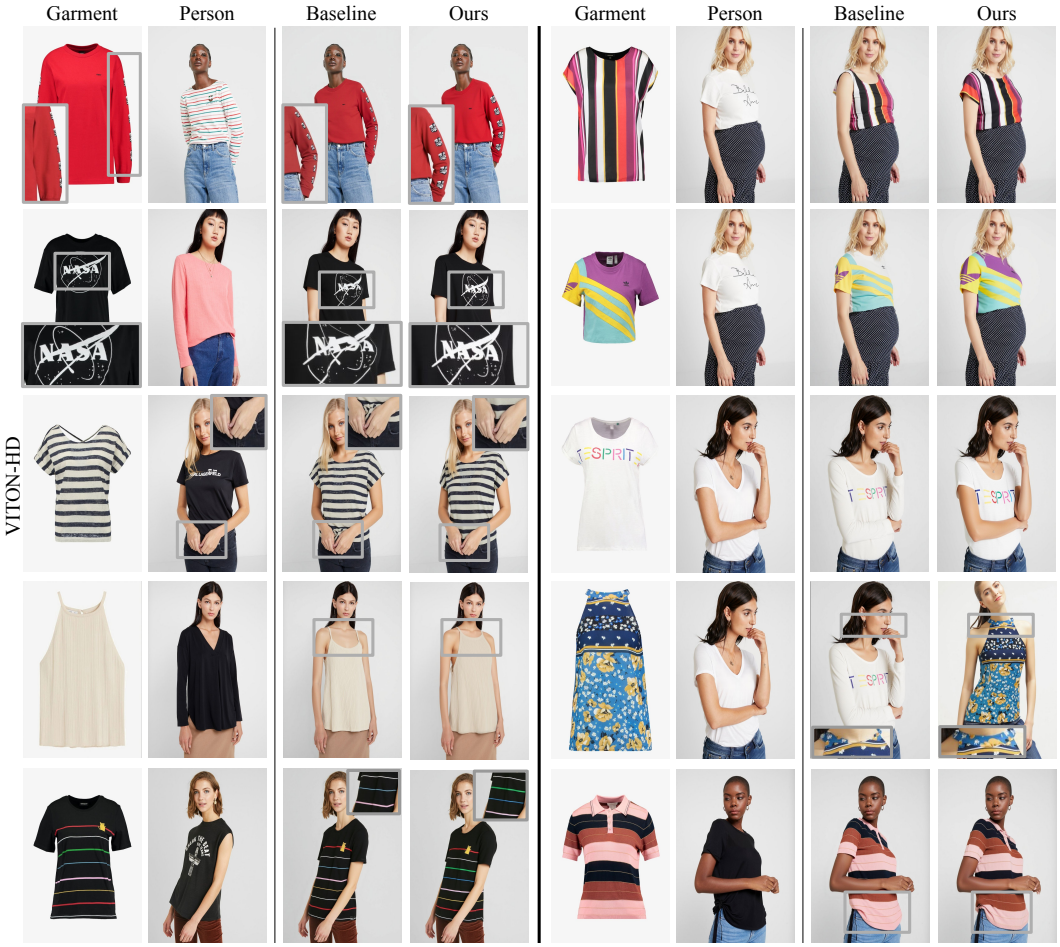


Figure N.19: **Qualitative Comparisons on VITON-HD [12] Before and After CORAL Loss.**

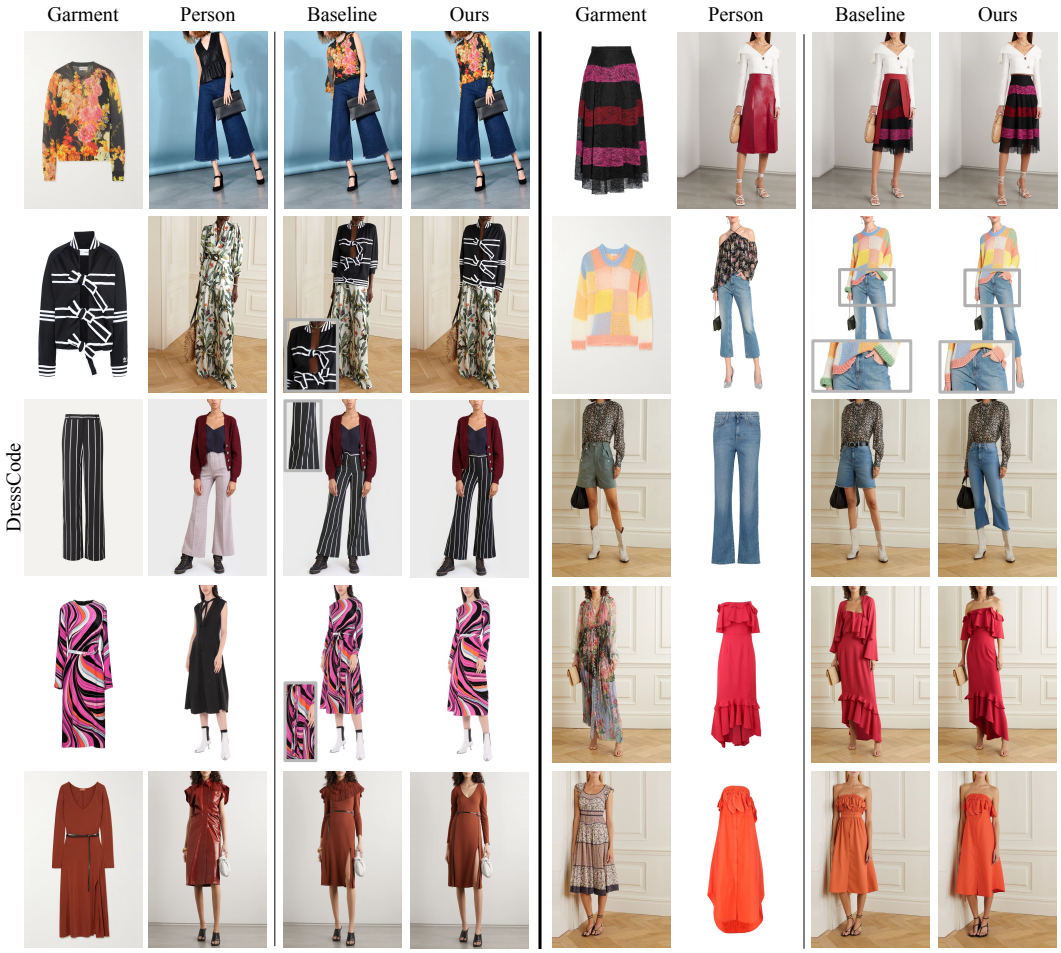


Figure N.20: **Qualitative Comparisons on DressCode [45] Before and After CORAL Loss.**

O Limitations and Discussion

Reliance on DINOv3 Pseudo Ground-Truth. CORAL supervises query–key matching using DINOv3-based pseudo ground-truth correspondences, so its correspondence quality can degrade when DINOv3 matching is unreliable, such as under strong appearance changes, heavy occlusions, or extreme viewpoints. While our cycle-consistency based reliability filtering removes many unstable matches, some noisy matches may remain and occasionally lead to incorrect supervision. A promising direction is to incorporate stronger confidence modeling, aggregate pseudo matches from multiple matching models, or introduce self-consistency checks to further reduce the impact of noisy pseudo labels. Our pseudo ground-truth assigns matches per token by argmax, without local or global consistency constraints. Consistency-aware matching, such as cost aggregation in CATs [24], could provide cleaner pseudo ground-truth and is a natural extension of our framework.

Analysis on Failure Cases. The failure cases we observe mainly come from heavily in-the-wild references, and in our experiments, the loss rarely introduced an artifact that the baseline did not already have. The first case is content missing from the reference. In the pants example of Fig. 5 (bottom-left), part of the reference garment is occluded by a hand. The garment transfers well overall, but a small spot on the right side does not match the reference, exactly where it is occluded. Since the reference provides little reliable correspondence for this occluded region, few matches there pass the cycle-consistency check, and the region is trained largely as in the baseline. The second case is content distorted in the reference, as in the last row of Fig. N.17, where the garment is visible but altered by extremely strong illumination. Here, few matches pass the check, so the affected regions are left to the diffusion objective as well. In both cases, the reference itself lacks reliable information for these regions, which supervision alone cannot recover.

Efficiency. Our baseline injects pose by concatenating pose latents along the token dimension, which increases the sequence length and can raise the training cost. In practice, this overhead can be reduced by using a compact pose representation.

Future Directions. Although we focus on VTON, an important future direction is to extend CORAL to broader reference-image-based customization settings, where a reference image specifies the target appearance while preserving the input content and structure. We also expect that leveraging stronger pseudo-correspondence estimators beyond DINOv3, or combining pseudo matches from multiple matchers, could further improve the quality and robustness of the correspondence supervision and generated results.

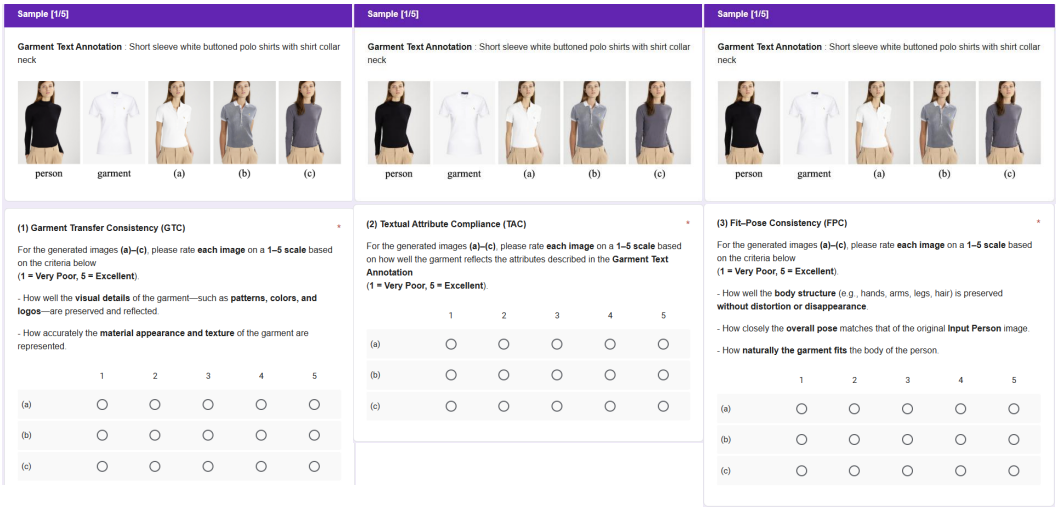


Figure O.21: **Example of User Study.**

You are a strict virtual try-on (VTON) evaluator.
Given a person image ***[IMG1]***, a garment product image ***[IMG2]***, a generated try-on image ***[IMG3]***, and a ***GARMENT CAPTION***, determine TEXTUAL ALIGNMENT, VISUAL MATCH, FIT QUALITY, NON-TARGET INTEGRITY, and OVERALL SCORE.

Scoring Rules (continuous 0.0–5.0, discrete 1–5):
- TEXTUAL ALIGNMENT: adherence to caption attributes (category, sleeve length, neckline, material/signature)
- VISUAL MATCH: agreement with ***[IMG2]*** in silhouette, material, sleeve length, neckline, hem band, and reproduction of signature details
- FIT QUALITY: geometric/ratio-based naturalness vs intended fit (slim/regular/relaxed) using torso width ratio, sleeve wrist ratio, shoulder drop ratio, bunching level, contour loss %
- NON-TARGET INTEGRITY: preservation of non-target regions (hair/skin/hands/other clothes) and valid pose consistency
- OVERALL: mean of all scores, then limited by Visual and Fit

Category-Scoped Ignore Rule: Decide garment type from ***[IMG2]+CAPTION*** ∈ {TOP, BOTTOM, DRESS}.
From ***[IMG1]*** → ***[IMG3]***, only that target garment category (TOP/BOTTOM/DRESS) may change .
All non-target regions must remain visually unchanged except minimal seam blending.

Hard Caps (applied after sub-scores):
- Category mismatch → TEXTUAL ≤2.0, VISUAL ≤2.0
- Length band mismatch (sleeve/hem) → TEXTUAL ≤2.0, VISUAL ≤2.0
- Silhouette mismatch → VISUAL ≤2.0
- Material/texture mismatch → VISUAL ≤2.0, FIT ≤3.0
- Missing essential signature detail → VISUAL ≤2.3, FIT ≤2.5
- If Visual silhouette component <0.6 or Visual_disc ≤2 → FIT ≤ VISUAL
- Severe artifacts → FIT ≤1.5
- Non-target changes → INTEGRITY ≤2.0, OVERALL ≤3.0

Caption Specificity Gate:
Caption explicitly containing ≥3 clothing attributes allows full (5.0) TEXTUAL score.
If only 1–2 attributes appear → cap TEXTUAL at 3.0–4.0.

Output Format (JSON only):

```
{
  "caption": "...",
  "garment_type": "TOP / BOTTOM / DRESS",
  "parsed_caption": ["category: ...", "sleeve: ...", "neckline: ...", "signature/material: ..."],
  "extracted_visual": ["silhouette: ...", "material/texture: ...", "signature: ...", "lower_length: ..."],
  "scores_continuous": { "textual": 0.0, "visual": 0.0, "fit": 0.0, "integrity": 0.0, "overall_continuous": 0.0 },
  "scores_discrete": { "textual": 1, "visual": 1, "fit": 1, "integrity": 1, "overall_discrete": 1 },
  "caps_applied": ["none"],
  "occlusions": ["none" | "..."],
  "reason": "<short numeric explanation>"
}
```

Figure O.22: **Details on VLM Evaluation Prompt.**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The main claims are supported by quantitative results on VITON-HD, DressCode, and PPR10K, VLM-based evaluation, human evaluation, qualitative comparisons, and ablation studies. The paper also analyzes correspondence quality and attention behavior to support the proposed design.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. We include limitations and discussion in the supplementary material, covering the reliance on DINOv3 pseudo correspondences, possible degradation under unreliable matching conditions such as occlusion or extreme viewpoints, efficiency overhead from pose-token concatenation, and future extensions beyond VTON.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results, formal guarantees, or proofs. The contributions are methodological and empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the model architecture, loss definitions, pseudo-correspondence construction, training setup, datasets, preprocessing, evaluation metrics, and ablation settings needed to reproduce or verify the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[Yes]

   Justification: We provide the datasets, preprocessing, evaluation metrics, and ablation settings needed to reproduce or verify the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. We provide training setup, optimizer, learning rate, batch size, training steps, hardware, resolution, loss weights, datasets, preprocessing, and evaluation metrics in the main paper and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or confidence intervals for the main benchmark results. For the VLM-based evaluation protocol, we additionally report its correlation with human scores, including Pearson correlation and statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. We report the compute setup used for training and evaluation, including GPU type, number of GPUs, throughput, and peak memory usage in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research conforms to it. The work uses public datasets and pretrained models following their terms of use and includes discussion of limitations and potential societal impacts.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts of improved virtual try-on, such as online fashion and accessibility, as well as possible negative impacts related to misuse, privacy, and unauthorized manipulation of person images.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Since the released model can be used for person-image editing, we include usage guidance and release the assets for research purposes. We do not redistribute third-party datasets or checkpoints, and users must follow the licenses and usage policies of the underlying models and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes. We cite the original papers for all pretrained models, datasets, baselines, and evaluation tools used in this work, including FLUX.1-Fill-dev, DINOv3, VITON-HD, DressCode, PPR10K, and VLM-based evaluation tools. We follow their respective licenses and terms of use and do not redistribute third-party datasets or pretrained checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released assets include code and model weights. We provide documentation, installation and inference instructions, evaluation scripts, license information, and usage notes alongside the released assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We conduct a human evaluation for virtual try-on quality. The supplementary material describes the evaluation protocol, criteria, number of participants, number of questions, comparison models, and the 5-point rating scale.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: We did not obtain formal IRB approval. The user study was limited to low-risk perceptual evaluation of generated images, collected no personally identifiable information, and participants were informed that their responses would be used for research evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use GPT-4o as part of the VLM-based evaluation protocol. The paper describes the evaluated criteria, scoring format, and prompt design, and uses the VLM only for evaluation rather than as part of the proposed training method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.