arXiv:2602.17636v1 [cs.CV] 19 Feb 2026

CORAL：用于改进虚拟试穿的对应关系对齐

金知英 [1] 申永镇 [1] 陈始允 [1] 郑多贤 [1] 南智秀 [1] 金东民 [1] 朴钟载 [2] 姜贤宇 [2] 金承龙 [1]

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



(a) **CORAL** 框架　(b) 虚拟试穿结果　(c) 人与人结果

图 *1.* **预览图。** (a) CORAL 是一种基于扩散Transformer的框架，可在 DiT 的三维注意力中显式增强人物与服装之间的对应关系。这使生成结果中的局部细节更加准确，例如减少伪影（如服装下摆重复），而基线中则存在此类问题。(b) VITON-HD 和 DressCode 上的虚拟试穿结果，对比未使用和使用 CORAL 时的输出。(c) 在具有挑战性的真实场景图像中，PPR10K 上的人物间服装迁移结果。

摘要

现有的虚拟试穿（VTON）方法往往难以保留精细的服装细节，尤其是在需要准确人物与服装对应关系的非配对设置中。这些方法并未显式强制人物与服装对齐，也未能解释对应关系如何在扩散 Transformer（扩散变换器）中形成。在本文中，我们首先分析基于 DiT 的架构中的完整三维注意力，并揭示人物→服装对应关系关键取决于完整三维注意力内精确的人物→服装查询–键匹配。基于这一见解，我们随后提出**COR**对应关系**AL**对齐（**CORAL**），这是一个基于 DiT 的框架，可将查询–键匹配与稳健的外部对应关系进行显式对齐。CORAL 整合了两个互补组件：一种将可靠匹配与人物→服装注意力对齐的对应关系蒸馏损失，以及一种锐化注意力分布的熵最小化损失。我们进一步提出一种基于视觉语言模型的评估协议，以更好地反映人类偏好。CORAL 相较基线持续取得提升，同时增强了全局形状迁移与局部细节保留。大量消融实验验证了我们的设计选择。

1. 引言

给定一张人物图像和一张服装图像，虚拟试穿（VTON）旨在合成同一个人穿着给定服装的效果，同时处理两者之间较大的几何差异

[1]KAIST AI [2]NC AI株式会社 对应关系：金承龙 < seungryong.kim@kaist.ac.kr>。

预印本。*2026*年*2*月*20*日。



(a) 注意力对应准确率和虚拟试穿质量

(b) 关键点正确率和图像质量指标

图 2. **查询-键匹配与虚拟试穿性能之间的相关性。** 粉色 标记表示查询点。(a) 展示了虚拟试穿性能与人物→服装注意力之间的定性相关性。(a) 中的所有输出均由基线生成。人类偏好的输出显示出定位准确、清晰的注意力，而低质量输出则显示出分散在错误位置的注意力。(b) 展示了定量相关性，其中 r 表示皮尔逊相关系数。人物→服装对应关系的关键点正确率与生成图像的结构相似性指数和学习感知图像块相似度相关，这表明更准确的定位可预测更高的视觉质量（放大查看效果最佳）。

，包括姿态、轮廓，以及服装类型或形状方面的差异（Choi 等人，2024年；Zhou 等人，2024年）。因此，虚拟试穿面临的一项关键挑战是在人与服装之间建立准确的对应关系，因为匹配不准确会导致虚拟试穿效果欠佳，包括服装形状扭曲、纹理错位或不合身。

尽管存在这一挑战，大多数现有工作仍着重于通过先进的推理技术（Bhunia 等，2023；Yang 等，2024；Chong 等，2025）、额外的条件信号（Choi 等，2024；Kim 等，2025；Xie 等，2023）或服装编码器（Kim 等，2023；Choi 等，2024；Xu 等，2024；Nam 等，2025b）来保留细粒度服装细节，却往往忽视了准确的人物与服装匹配。因此，它们仍无法忠实地迁移服装形状和局部细节（例如小型标志或重复图案）；尤其是在输入跨类别服装时，准确匹配更为必要。

此外，大多数先前工作（Zhou 等人，2024年）均基于扩散 U-Net 架构构建，对扩散 Transformer（DiT）（Peebles 和 Xie，2023年；Esser 等人，2024年）等更先进设计的探索有限。近期研究表明，DiT 通过完整的三维注意力，能够增强输入令牌之间的上下文内交互（Tan 等人，2025a）。这促使我们为虚拟试穿开发基于 DiT 的架构，以建立人物与服装输入令牌之间的强对应关系。

为此，我们首先分析了 DiT 中用于虚拟试穿的完整 3D 注意力机制的行为，并发现***RGB*** 空间中的人物-服装精确对齐，关键取决于 ***3D*** 注意力机制中精确的查询-键对应关系。如图1 和图2所示，我们证明，随着注意力导出的、人物查询与服装键之间的匹配越来越准确，试穿质量会线性提升。

基于这一分析，我们提出了一种新颖的基于 DiT 的帧级**COR**respondence**A**L**i**g**n**ment（**CORAL**），通过改善 DiT 完整三维注意力中的查询–键匹配，显式增强人物与服装对应关系。具体而言，我们引入两种互补损失：（1）对应关系蒸馏损失，将视觉基础模型 DINOv3 得出的可靠匹配（Sim´eoni 等，2025年）与完整三维注意力中的人物与服装查询–键匹配对齐；（2）熵最小化损失，通过最小化注意力分布的熵，使其更加尖锐，从而实现局部且精确的匹配。

我们证明了 CORAL 在配对和非配对设置下的标准虚拟试穿基准上均有效。为进一步评估我们方法的鲁棒性，我们引入了一个新的评估数据集，其中包含具有挑战性的真实世界人物间非配对场景。CORAL 在所有基准数据集上的 SSIM、学习感知图像块相似度和弗雷歇起始距离指标上均取得最先进的性能。此外，我们还采用基于视觉语言模型的评估和人工评估来评估我们的方法，其表现优于所有其他方法。大量消融研究进一步验证了我们的设计选择。

总而言之，我们的贡献如下：

- 我们揭示，虚拟试穿中的精确人物与服装对齐，关键取决于 DiT 完整三维注意力中准确的查询—键对应关系。

- 基于这些发现，我们提出 CORAL，这是一种基于 DiT 的虚拟试穿框架，通过对应关系蒸馏和注意力熵最小化，显式增强人物与服装之间的对应关系。

- 我们在各虚拟试穿基准上的标准指标以及人工评估和基于视觉语言模型的评估中均取得了稳定的性能提升，并在我们引入的其他具有挑战性的设置中验证了我们的方法。

2. 相关工作

虚拟试穿。虚拟试穿旨在将服装图像的形状和细节精确迁移到人物图像上。多项工作改进了推理阶段的采样与对齐（Bhunia 等，2023；Chong 等，2025；Wang 等人，2024年；Li 等人，2025年）。另一些研究（Choi 等，2024；Kim 等，2025；Xie 等，2023）加入结构线索（例如解析图），以提升对姿态变化的鲁棒性。其他方法则从服装图像中注入全局特征（Kim 等，2023；Wan 等人，2024年；Morelli 等人，2023年）或局部特征（Choi 等，2024；Zhou 等，2024；Xu 等，2024），以增强细节保留。然而，在需要稳健匹配的非配对设置中，它们往往会丢失服装细节，从而导致标志漂移或边界偏移。

扩散模型中的对应关系。扩散模型令人瞩目的性能源于它们所学到的丰富视觉表征。对 UNet（Nam 等，2024a；b；Jeong 等人，2025年；Jin 等人，2025b；Hedlin 等人，2023年；Tang 等人，2023年；Xiao 等人，2024年）和基于 DiT（Yu 等，2025；Nam 等人，2025a；Jin 等人，2025a；Lee 等人，2025年）的模型内部注意力机制的分析表明，查询-键匹配编码了语义对应关系。编辑与定制研究（Nam 等，2024a；Cao 等人，2023年；Jin 等人，2025b；Hertz 等人，2022年；Tumanyan 等人，2022年）一致表明，查询-键匹配捕捉结构与语义对齐，而值则承载外观，并将其导向被注意的位置。这些发现实现了强有力的布局保持和可控的外观迁移，同时也支持跟踪等下游任务（Jeong 等人，2025年；Nam 等人，2025a）。

虚拟试穿中的对应关系。若干虚拟试穿工作（Wan 等人，2025年；Chen 等人，2024年；Huang 等人，2022年）将稀疏的基于点的对应关系注入少数匹配点周围的注意力或扭曲特征中，这仅提供局部监督，并且在非配对设置下面对遮挡和大幅姿态变化时常常失效。除稀疏对应关系之外，一些方法（Zhou 等人，2024年）从注意力中推导稠密光流，但在非配对设置中，此类光度假设（Truong 等人，2021年）更常失效，从而降低性能。相反，我们的方法在扩散变换器中运行，并通过将注意力对齐到来自 DINOv3（Sim´eoni 等人，2025年）的外部稳健人物-服装对应关系，显式强化查询-键匹配，从而降低对光度一致性的依赖。

3. 预备知识

潜在扩散模型（LDM）。LDM（Rombach 等人，2022年）在潜空间中运行，由变分自编码器和去噪器组成 $\epsilon_\theta(\cdot)$。变分自编码器编码器将图像 $I \in \mathbb{R}^{H\times W\times 3}$ 映射为潜在表示 $\mathbf{z}_0 \in \mathbb{R}^{h\times w\times c}$，其中（$H$，$W$，3）和（$h$，$w$，$c$）分别表示 RGB 空间和潜空间中的高度、宽度和通道维度。在时间步 t，向 $\epsilon \sim \mathcal{N}(0,\ \mathbf{I})$ 中添加高斯噪声$\mathbf{z}_0$，生成噪声潜变量$\mathbf{z}_t$。给定条件标记$\mathbf{c}$，我们采用整流流（Esser 等人，2024年），模型预测将干净潜变量传输至高斯噪声的速度。具体而言，在时间步 t，构造线性插值后的潜变量如下：

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

并且速度网络 $\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c})$ 被训练以近似目标速度（$\boldsymbol{\epsilon} - \mathbf{z}_0$），并以 $\mathbf{c}$ 为条件。速度训练目标 $\mathcal{L}_{\text{velocity}}$ 定义为

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

DiT 中的注意力。扩散 Transformer（Peebles 与 Xie，2023年；Esser 等人，2024年）由多个三维全注意力层组成，这些层联合处理视觉信息和文本信息。噪声潜变量令牌 $\mathbf{z}_t$ 与时间步 t 的条件标记 $\mathbf{c}_t$ 拼接起来，形成输入序列。条件标记通过拼接多个条件的令牌构造而成，$\mathbf{c}_t = [\,c_{t,1}\|...\|c_{t,N}\,]$，其中 N 表示条件的数量，$[\,\cdot\|\cdot\,]$ 表示按令牌进行拼接。因此，最终的输入序列为 $[\mathbf{z}_t\|\mathbf{c}_t\,]$。在每个时间步 t 和变换器层 l中，拼接后的序列被投影为查询 $Q^{t,l}$ 和键 $K^{t,l}$：

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

$$K^{t,l} = [\,K^{t,l}_{\mathbf{z}} \,\|\, K^{t,l}_{\mathbf{c}_1} \,\|\, ... \,\|\, K^{t,l}_{\mathbf{c}_N}\,]. \tag{4}$$

序列中的每个令牌都通过旋转位置编码（RoPE）（Su 等人，2023年）进行增强。随后，时间步 t和层 l 处的注意力图 $A^{t,l}$ 计算如下：

$$A^{t,l} = \text{Softmax}\left(\frac{Q^{t,l}(K^{t,l})^\top}{\sqrt{d}}\right), \tag{5}$$

其中，Softmax(·) 针对每个查询沿键维度计算，d表示通道维度。总而言之，这种全注意力决定了序列中的每个潜变量或条件注入词元如何关注其他所有词元。

4. 方法

4.1. 任务定义

令 $I_{\text{p}} \in \mathbb{R}^{H\times W\times 3}$ 表示人物图像，$I_{\text{g}} \in \mathbb{R}^{H\times W\times 3}$ 表示服装图像，$I_{\text{pose}} \in \mathbb{R}^{H\times W\times 3}$ 表示姿态图像（根据 I_{p} 估计得出）。令 $M_{\text{p}}, M_{\text{g}} \in \mathbb{R}^{H\times W}$ 分别表示 I_{p} 和 I_{g} 中服装区域的二值掩码，$M_{\text{e}} \in \mathbb{R}^{H\times W\times 3}$ 表示二值图像修复掩码，用于指定 I_{p} 中需要编辑的区域，其中 $M_{\text{p}} \subseteq M_{\text{e}}$。给定输入 $\{I_{\text{p}}, I_{\text{g}}, I_{\text{pose}}, M_{\text{p}}, M_{\text{g}}, M_{\text{e}}\}$，我们的目标是合成一张 ˆ 试穿图像 I，使其 (i) 保留 I_{p} 在

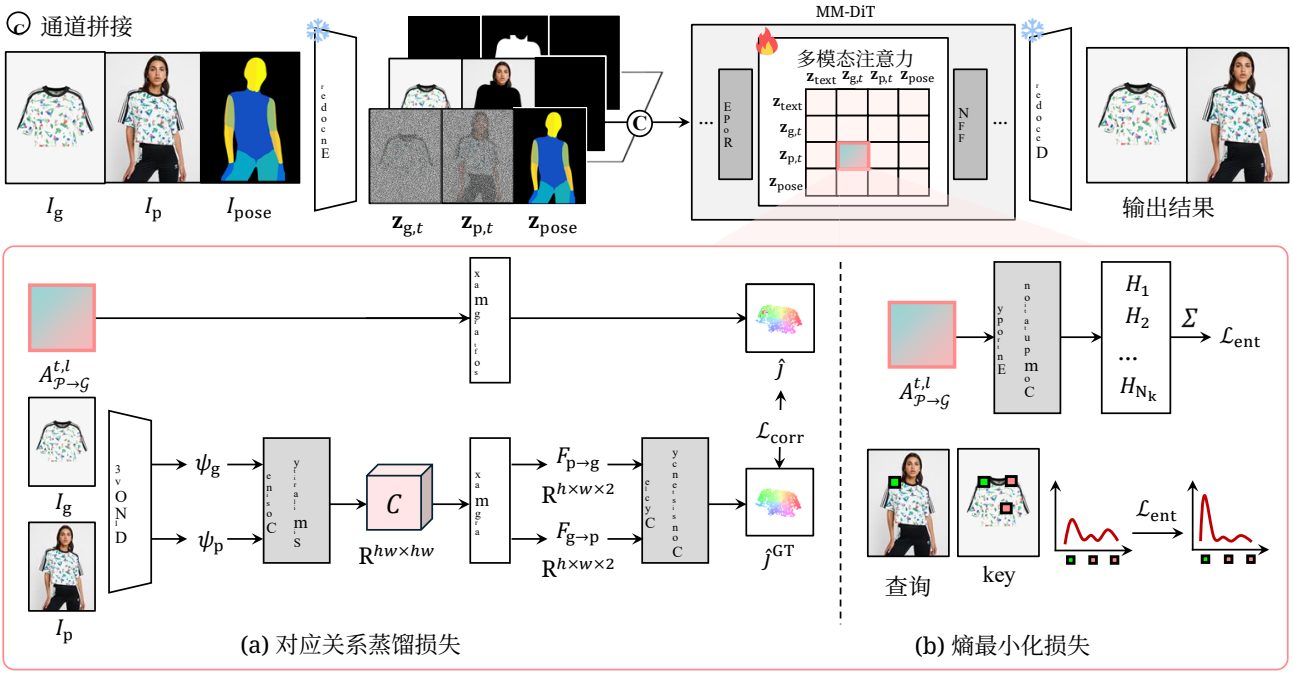


图3. **总体架构**。 CORAL 基于一种基线架构，该架构构造噪声潜变量 $\mathbf{z}_t$ ，即水平拼接噪声服装潜变量 $\mathbf{z}_{\mathrm{g},t}$ 与人物潜变量 $\mathbf{z}_{\mathrm{p},t}$，然后在输入投影层之前将条件画布$\mathbf{z}^{\mathrm{diptych}}$ 和掩码画布 $\mathbf{m}^{\mathrm{diptych}}$ 与 $\mathbf{z}_t$ 沿通道拼接。姿态通过将 $\mathbf{z}^{\mathrm{pose}}$ 作为令牌相加来注入，并将旋转位置编码设置为在人物令牌与姿态令牌之间共享空间位置。 $\mathcal{L}_{\mathrm{CORAL}}$ 应用于人物→服装匹配代价 $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ ，该代价由扩散变换器模块内的多模态注意力估计得到： $\mathcal{L}_{\mathrm{corr}}$ 将 $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ 与从 DINOv3 提取的伪真实对应关系对齐，而 $\mathcal{L}_{\mathrm{ent}}$ 在 $A^{t,l}_{\mathcal{P}\to\mathcal{G}}$ 上计算，以鼓励更锐利、更局部化的匹配。

由 M_{e} 指定的区域内的外观，(ii) 遵循 I_{pose} 定义的姿态，以及 (iii) 根据 M_{g} 将服装 I_{g} 的形状和局部细节准确地迁移到人物图像中，并将范围限制在 M_{e} 内。

4.2. 网络架构

双联画形式化。 我们首先构建一个基于 DiT 的基线，采用简单的条件注入设计，利用 DiT 的上下文条件注入能力进行参考引导生成（Shin 等人，2025年；Huang 等人，2024年）。我们采用双面板双联画布局，将服装潜变量与人物潜变量水平拼接，从而通过多模态注意力实现直接的令牌级交互，而不是依赖辅助编码器来整合服装。总体架构如图 3。

具体而言，我们使用变分自编码器对服装与人物图像 I_{g} 和 I_{p}进行编码，得到$\mathbf{z}^{\mathrm{g}}$ 和$\mathbf{z}^{\mathrm{p}}$，并构成一块双联潜变量画布。在时间步 t，我们将含噪双联画潜变量构造为：

$$\mathbf{z}_t = [\,\mathbf{z}_{\mathrm{g},t} \,\|\, \mathbf{z}_{\mathrm{p},t}\,], \tag{6}$$

其中$\mathbf{z}_{\mathrm{g},\ t}$ 和$\mathbf{z}_{\mathrm{p},\ t}$ 是在时间步 t 的含噪服装潜变量和人物潜变量。为了对干净服装图像和遮罩人物图像进行条件注入，我们使用干净服装潜变量和遮罩人物潜变量：

$$\mathbf{z}_{\mathrm{diptych}} = [\,\mathbf{z}_{\mathrm{g}} \,\|\, \mathbf{z}_{\mathrm{p}} \odot (1 - \mathbf{m}_{\mathrm{e}})\,], \tag{7}$$

其中$\mathbf{m}_{\mathrm{e}}$ 是将 M_{e} 下采样到潜空间得到的。我们还按照相同的双联布局定义二值掩码画布$\mathbf{m}^{\mathrm{diptych}}$，将待替换区域赋为1，其余位置赋为0：

$$\mathbf{m}_{\mathrm{diptych}} = [\,\mathbf{0}_{h\times w} \,\|\, \mathbf{m}_{\mathrm{e}}\,]. \tag{8}$$

姿态注入。以往的虚拟试穿方法常常存在姿态幻觉和手部不一致问题，即合成结果无法与输入姿势对齐，并扭曲手部形状。为缓解这一问题，先前方法中的一种常见设计是将姿态条件$\mathbf{z}^{\mathrm{pose}}$（通过对 I_{pose} 使用变分自编码器编码得到）与含噪人物潜变量$\mathbf{z}_{p,\ t}$ 沿通道维度拼接（Choi 等人，2024年；Kim 等人，2023年）。然而，我们发现该方法在严格姿态对齐方面并非最优，因为人物外观与姿态条件会在潜空间中相互纠缠。

受近期基于 DiT 的条件框架（Tan 等人，2025a）启发，我们的主要基线则沿令牌维度拼接干净的姿态条件$\mathbf{z}^{\mathrm{pose}}$，构成$[\mathbf{z}_{g,\ t}||\mathbf{z}_{p,\ t}||\mathbf{z}_{\mathrm{pose}}]$ 作为输入。随后，我们通过修改旋转位置编码（Su 等人，2023年）来确保空间对齐，使$\mathbf{z}_{p,\ t}$ 和$\mathbf{z}^{\mathrm{pose}}$共享相同的位置索引，从而提供强大的

位置对应关系，并允许姿态通过对联合序列进行全注意力处理而更加连贯地融入其中。为匹配输入通道维度，我们对$\mathbf{z}^{\text{diptych}}$和$\mathbf{m}^{\text{diptych}}$进行零填充。最后，将经过零填充的$\mathbf{z}^{\text{diptych}}$和$\mathbf{m}^{\text{diptych}}$与$[\mathbf{z}_t||\mathbf{z}_{\text{pose}}]$按通道维度拼接，作为嵌入层的输入，随后将生成的令牌传入 DiT。

4.3. 人像与服装对应关系分析

动机。虚拟试穿的一大挑战，是在人像与服装之间建立准确的对应关系，从而将服装局部细节和形状可靠地迁移到人物图像上。一种方法是将 DiT 中间特征与强匹配描述符（例如 DINOv3（Sim´eoni 等人，2025年））对齐，但这通常会扰乱模型学得的生成表征，并且需要完全重新训练（Yu 等人，2025年）。另一种方法是将可靠的对应关系蒸馏到 DiT 中间特征中。然而，正如先前研究（An 等人，2025年；Nam 等人，2024a）所讨论的，这些特征包含来自值矩阵的外观信息，会削弱空间关系，因此不适合精确匹配。更多探索见附录E。

为解决这些局限性，我们利用 DiT 完整三维注意力中的查询–键注意力机制，该机制能够隐式建模人物与服装词元之间的交互（Nam 等人，2025a），同时不扰乱外观。

对应关系估计。为考察基线扩散变换器如何建立服装特征与人物特征之间的空间对齐，我们从扩散变换器的注意力中提取查询–键对应关系。根据公式 4，时间步 t和层 l 上的查询和键由以下部分组成：

$$Q^{t,l} = [\,Q^{t,l}_{\mathbf{z}_{\text{text}}} \,\|\, Q^{t,l}_{\mathbf{z}_{\text{g}}} \,\|\, Q^{t,l}_{\mathbf{z}_{\text{p}}} \,\|\, Q^{t,l}_{\mathbf{z}_{\text{pose}}}\,], \tag{9}$$

$$K^{t,l} = [\,K^{t,l}_{\mathbf{z}_{\text{text}}} \,\|\, K^{t,l}_{\mathbf{z}_{\text{g}}} \,\|\, K^{t,l}_{\mathbf{z}_{\text{p}}} \,\|\, K^{t,l}_{\mathbf{z}_{\text{pose}}}\,]. \tag{10}$$

这里，$Q^{t,\,l}_{\mathbf{z}_{\text{text}}}$和$K^{t,\,l}_{\mathbf{z}_{\text{text}}}$对应于输入文本令牌。我们将$m_{\text{p}}$和$m_{\text{g}}$定义为$M_{\text{p}}$和$M_{\text{g}}$的下采样版本。令$\mathcal{P}$和$\mathcal{G}$分别为人物令牌和服装令牌的索引集，其空间位置分别落在m_{p}和m_{g}内，对应于$Q^{t,\,l}_{\mathbf{z}_{\text{p}}}$和$K^{t,\,l}_{\mathbf{z}_{\text{g}}}$的掩码子集。使用公式 5，我们以$Q^{t,\,l}_{\mathbf{z}_{\text{p}}}$为查询、以$K^{t,\,l}_{\mathbf{z}_{\text{g}}}$为键，计算人物与服装之间的注意力，得到：

$$A^{t,l}_{\mathcal{P}\to\mathcal{G}} := A^{t,l}[\mathcal{P},\mathcal{G}] \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{G}|}. \tag{11}$$

我们将$A^{t,\,l}_{\mathcal{P}\to\mathcal{G}}$视为人物潜变量与服装潜变量之间的匹配代价。这与（中的分析一致

Nam等人，2024a；2025a），但此前没有研究将虚拟试穿任务表述为 DiT 注意力中的人物与服装对应关系。为简便起见，以下方程中省略变量 t 和 l 。



(a) (b) (c)

图 *4*。**对应关系可视化与变形结果**。我们将对应场以及由此产生的变形服装可视化；该服装通过将像素从服装参考图像映射到人物图像的服装区域而生成。(a) 基线的基于注意力的对应关系。(b) 可靠性筛选前的 DINOv3 对应关系。(c) 经过循环一致性检查后的精炼对应关系。基线变形会产生几何失真，而未经筛选的 DINOv3 可能会错误匹配视觉上相似的区域，例如下装区域。

根据这些匹配代价，我们通过对服装令牌取 argmax，从人物空间位置 i 提取到服装潜变量的稠密对应关系：

$$\hat{j}_i = \arg\max_{j\in\mathcal{G}} A_{\mathcal{P}\to\mathcal{G}}(i,j). \tag{12}$$

评估基准。为验证基线架构中的查询–键对应关系，我们需要人物与服装真实对应关系。由于不存在此类标注，我们使用 DINOv3（Sim´eoni 等人，2025年）构建伪真实匹配。DINOv3 是一种视觉基础模型，已证明在密集匹配方面十分有效。给定 I_{p}和 I_{g}，我们从 DINOv3 中提取特征描述符 $\phi(\cdot)$，并仅保留下采样掩码内的令牌：

$$\psi_{\text{p}} = \phi(I_{\text{p}}) \odot m_{\text{p}}, \quad \psi_{\text{g}} = \phi(I_{\text{g}}) \odot m_{\text{g}}. \tag{13}$$

遵循标准的稠密匹配协议（Hong 等人，2022年；b；Cho 等人，2024年；Nam 等人，2024b），我们使用余弦相似度计算描述符之间的匹配代价：

$$C(i,j) = \frac{\psi_{\text{p}}(i)\cdot\psi_{\text{g}}(j)}{\|\psi_{\text{p}}(i)\|_2\,\|\psi_{\text{g}}(j)\|_2}, \tag{14}$$

其中，i 和 j 分别索引人物和服装描述符的空间位置。

为了获得可靠的伪真实匹配，我们引入可靠性掩码 $\mathbf{m}_{\text{rel}}(i)$，利用循环一致性（Jiang 等人，2021年）约束滤除不稳定的对应关系。我们首先估计双向光流，即从人物到服装的光流 $F_{\text{p}\to\text{g}} \in \mathbb{R}^{h\times w\times 2}$，以及从服装到人物的光流 $F_{\text{g}\to\text{p}} \in \mathbb{R}^{h\times w\times 2}$，方法是在空间位置上取最大值，得到服装和人物在匹配代价中的位置

，分别如此。可靠性掩码会识别出这样的点：其在服装特征中的匹配位置返回人物特征中的原始位置，且偏差在指定阈值γ以内：

$$\mathbf{m}_{\text{rel}}(i) = \begin{cases} 1, & \text{if } \|F_{\text{g}\to\text{p}}(F_{\text{p}\to\text{g}}(i)) - i\|_2 < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

最后，如图所示，我们将可靠性掩码应用于与各位置对应的服装点，从而获得伪真实对应关系。4(c)：

$$\hat{j}_i^{\text{GT}} = \mathbf{m}_{\text{rel}}(i) \cdot \arg\max_{j\in\Omega} C(i,j), \tag{16}$$

其中，Ω 表示服装空间域。

查询-键匹配与性能。通过将基于注意力的对应关系与来自 DINOv3 的伪真实对应关系进行比较，我们观察到匹配准确度与试穿结果质量之间存在相关性。如图所示。2，人物与服装的 RGB 对齐准确时，查询—键注意力清晰且定位精确；而服装上存在伪影的欠佳案例中，注意力则较为分散或位置偏移。

为量化这一观察结果，我们从 VITON-HD（Choi 等人，2021年）测试集中抽取 540 对样本，并使用正确关键点百分比（PCK）评估点准确度。该指标衡量预测点与真实对应点之间距离小于 α 的预测所占比例（公式17），其定义如下：

$$\text{PCK}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\|\hat{j}_i - \hat{j}_i^{\text{GT}}\|_2 < \alpha], \tag{17}$$

其中$\mathbf{1}$（·）为指示函数，N 为已评估对应关系的数量，由服装区域$\mathbf{m}_{\text{p}}$ 和可靠性掩码$\mathbf{m}_{\text{rel}}$决定，且 α 为距离阈值。虚拟试穿性能使用结构相似性指数（Cong 等人，2022年）和学习感知图像块相似度（Zhang 等人，2018年）。

如图2 （b）所示，关键点正确率与结构相似性指数呈正相关，与学习感知图像块相似度呈负相关，表明更高的查询–键匹配准确率会在虚拟试穿结果中带来更好的结构保真度和感知质量。这些发现凸显了准确的查询–键匹配对虚拟试穿的重要性，并促使我们显式引导注意力，以获得更强的对应关系。

4.4. CORAL：对应关系对齐

基于我们的分析，我们提出CORAL，将由DINOv3得到的可靠匹配整合到基线的查询–键匹配中，从而实现稳健的虚拟试穿性能。

对应关系蒸馏损失。 我们的目标是将查询—键注意力得到的估计匹配项与 DI-NOv3 提供的伪真实匹配对齐（Sim´eoni 等人，2025年）。由于 argmax 操作不可微（公式12），我们使用 soft argmax 根据查询—键注意力预测估计的对应关系，形式如下：

$$\hat{j}_i = \sum_{j=1}^{N} A_{\mathcal{P}\to\mathcal{G}}(i,j)\cdot j. \tag{18}$$

最后，我们计算估计的对应关系与伪真值匹配之间的平均 L2 损失：

$$\mathcal{L}_{\text{corr}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{j}_i - \hat{j}_i^{\text{GT}}\|_2^2. \tag{19}$$

熵最小化损失。对应关系蒸馏损失（$\mathcal{L}_{\text{corr}}$）将查询-键匹配与外部伪匹配对齐。然而，当注意力分布过于分散时，这种对齐并不可靠，因为错误的键位置可能会主导加权平均。为确保对应关系监督的稳健性，我们引入熵最小化损失，以促进置信度更高、空间定位更准确的查询-键对齐。

为实现这一点，我们根据注意力分布定义熵。遵循香农对熵的定义（Attanasio 等人，2022年），我们计算熵 H_i，针对给定查询 i，将其注意力权重 $A(i,\ \cdot)$ 在各个键上的分布视为概率分布，并将其表述为：

$$H_i \ = \ -\sum_{j=1}^{N_{\text{k}}} A(i,j)\ \log A(i,j), \tag{20}$$

其中，N_{k} 是键词元的总数。对所有 N 查询词元的 H_i 取平均，得到熵最小化损失：

$$\mathcal{L}_{\text{ent}} \ = \ \frac{1}{N}\sum_{i=1}^{N} H_i. \tag{21}$$

较低的熵对应更尖锐、置信度更高的注意力，而较高的熵则表示匹配分布分散或不确定。因此，熵最小化损失（$\mathcal{L}_{\text{ent}}$）通过强制实现更尖锐、更可靠的对齐，补充了对应关系蒸馏损失（$\mathcal{L}_{\text{corr}}$）。

总损失。 最后，我们将 CORAL 损失定义为对应项和熵项的加权组合：

$$\mathcal{L}_{\text{CORAL}} = \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}, \tag{22}$$

其中，λ_{corr} 和 λ_{ent} 根据经验设定。

然后，我们通过加入 CORAL 损失来优化最终目标：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{velocity}} + \mathcal{L}_{\text{CORAL}}. \tag{23}$$

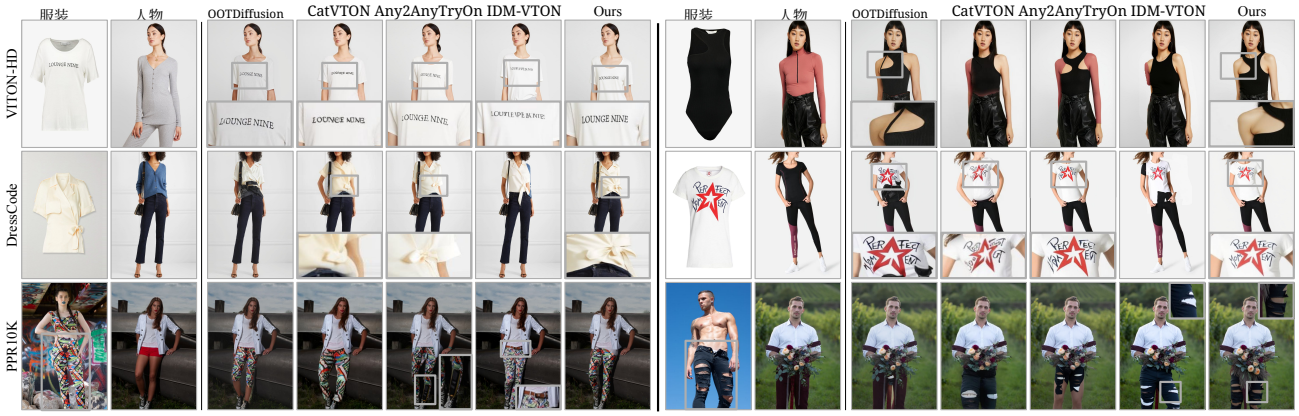


图 *5.* **定性比较**。 我们展示了标准基准数据集 VITON-HD（Choi 等人，2021年）和 DressCode（Morelli 等人，2022年）上的定性结果，以及基于 PPR10K（Liang 等人，2021年）构建的真实场景评估数据集上的结果（放大查看效果最佳）。更多定性结果见附录 J

表 *1.* **VITON-HD 上的定量比较（Choi 等人，2021年）**。

方法	成对				非成对	
	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
GPVTON(Xie 等人，2023年)	0.878	0.067	8.938	4.257	11.993	4.570
StableVTION (Kim 等人，2023年)	0.888	0.073	8.233	0.490	9.026	3.029
OOTDiffusion(Kim 等人，2023年)	0.842	0.087	6.619	0.845	9.938	1.302
IDM-VTON (Choi 等人，2024年)	0.866	0.062	6.009	0.838	9.198	1.203
CatVTON(Chong 等人，2025年)	0.874	0.058	5.458	0.439	9.076	1.184
Any2AnyTryOn（Guo 等人，2025年）	0.838	0.087	5.482	**0.384**	9.623	1.601
CORAL（不含 $\mathcal{L}_{\text{CORAL}}$）	0.889	0.055	5.543	0.870	9.641	1.323
CORAL（含 $\mathcal{L}_{\text{CORAL}}$）	**0.907**	**0.048**	**4.962**	0.565	**8.763**	**0.880**

表*2.* **DressCode 定量比较（Morelli 等人，2022年）**。* 不支持特定数据集上的配对设置。

方法	成对				非成对	
	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
GPVTON（Xie 等人，2023年）	0.918	0.068	8.423	2.439	9.144	3.936
OOTDiffusion（Xu 等人，2024年）	0.886	0.069	5.082	1.377	9.276	4.009
IDM-VTON（Choi 等人，2024年）	0.904	0.052	3.472	0.882	5.343	1.321
CatVTON（Chong 等人，2025年）	0.875	0.075	5.384	1.903	7.998	3.242
Any2AnyTryOn*（Guo 等人，2025年）	-	-	-	-	5.573	1.458
CORAL（不含 $\mathcal{L}_{\text{CORAL}}$）	0.908	0.045	2.896	0.418	5.221	1.315
CORAL（含 $\mathcal{L}_{\text{CORAL}}$）	**0.927**	**0.029**	**2.333**	**0.401**	**4.692**	**0.846**

5. 实验

5.1. 评估设置

评估数据集。 我们在标准虚拟试穿数据集上分别训练模型，**VITON-HD**（Han 等人，2018年）和**DressCode**(Morelli 等人，2022年），两者的分辨率均为1024 × 768，并在各自的测试集上进行评估。更多实现细节见附录A。

除了标准虚拟试穿基准之外，我们还在**PPR10K**（Liang 等人，2021年）上进一步评估零样本泛化能力。与店内商品图像场景相比，该数据集提供的输入未经精心筛选。此次评估中，我们使用在 DressCode 上训练的预训练检查点。我们还纳入了穿着服装场景，其中参考服装展示在另一位人物身上。关于 PPR10K 评估的更多细节见附录H。

指标。 我们报告常用指标结构相似性指数（Cong 等人，2022年）、学习感知图像块相似度（Zhang 等人，2018年）、弗雷歇起始距离（Heusel 等人，2018年）

表 *3.* **PPR10K 上的定量比较（Liang 等人，2021年）**。 * 不支持在特定数据集上进行配对设置。

方法	成对				非成对	
	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
OOTDiffusion（Choi 等人，2023年）	0.804	0.130	99.492	15.592	87.818	15.467
IDM-VTON（Choi 等人，2024年）	0.844	0.111	64.250	1.911	63.638	3.600
CatVTON(Chong 等人，2025年)	0.743	0.147	81.722	10.979	76.417	11.207
Any2AnyTryOn*（Guo 等人，2025年）	-	-	-	-	**49.728**	**0.421**
CORAL（不含 $\mathcal{L}_{\text{CORAL}}$）	0.877	0.078	44.117	0.015	56.644	1.202
CORAL（含 $\mathcal{L}_{\text{CORAL}}$）	**0.915**	**0.060**	**43.648**	**0.011**	53.164	1.101

表*4.* **非配对设置下基于视觉语言模型的评估**。* 不支持在特定数据集上进行推理。

训练/测试	VITON-HD/VITON-HD			DressCode/DressCode			DressCode/PPR10K		
	GTC ↑	TAC ↑	FPC ↑	GTC ↑	TAC ↑	FPC ↑	GTC ↑	TAC ↑	FPC ↑
GPVTON*（Xie 等人，2023年）	3.93	4.28	4.17	3.19	3.31	3.14	-	-	-
OOTDiffusion（Xu 等人，2024年）	3.97	4.34	4.12	3.25	3.24	3.45	1.04	2.06	1.57
IDM-VTON（Choi 等人，2024年）	3.96	4.40	4.25	3.25	3.28	3.61	1.72	2.21	2.37
CatVTON（Chong 等人，2025年）	3.82	4.20	4.06	3.07	3.22	3.53	1.05	2.00	2.01
Any2AnyTryOn（Guo 等人，2025年）	3.81	4.27	4.19	2.99	3.28	3.36	1.92	2.45	2.48
CORAL（我们的方法）	**3.99**	**4.40**	**4.26**	**3.47**	**3.31**	**3.83**	**2.07**	**2.56**	**2.89**

以及 KID（Bi´nkowski 等人，2021年），用于配对设置；对于非配对设置，则测量弗雷歇起始距离和 KID。

基于视觉语言模型的评估。 正如先前研究中所讨论的（Xi- aobin 等人，2025年），现有指标（如弗雷歇起始距离和 KID）不足以反映人类偏好，尤其是在非配对设置下。这类基于分布的指标主要衡量整体真实感，却无法验证参考图像中的特定服装是否得到忠实迁移并被自然地穿着。因此，我们评估现有指标未能充分反映的三个方面：**(1) 服装迁移一致性（GTC）**衡量输出是否保留了服装的视觉外观，包括整体形状和轮廓，以及图案、标志和纹理等细节。**(2) 文本属性一致性（TAC）**衡量输出是否符合文本或元数据中指定的服装属性，涵盖仅凭参考图像可能难以明确判断的要求（例如长度、领口、类别和合身度）。**(3) 穿着姿势协调性（FPC）**反映服装是否自然地穿在目标人物身上，并与输入人物的姿态和身体几何形态保持一致。同时还会考察非目标区域（例如头发、皮肤、手和其他衣物）是否得以保留且未发生意外变化。更多细节见

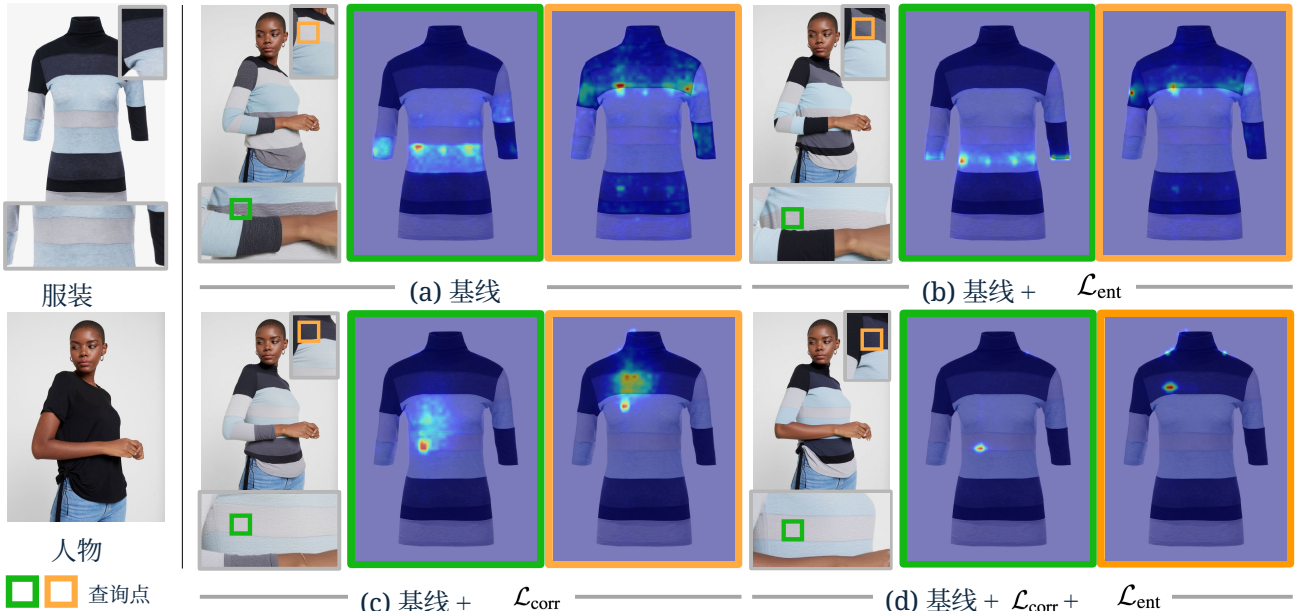


图6。 **损失组件消融实验。** 我们展示了这两种损失的有效性， $\mathcal{L}_{ent}$ 和 $\mathcal{L}_{corr}$。 橙色和绿色标记表示查询点，同色轮廓标出的注意力图则显示了各个变体对应的匹配结果。通过结合 $\mathcal{L}_{ent}$ 与 $\mathcal{L}_{corr}$，我们的模型能够以最高精度定位正确的键，并呈现最集中的注意力，从而取得最佳的虚拟试穿性能。



图7。 **.....的影响 $\mathcal{L}_{CORAL}$。** 我们将DiT注意力中的人物→服装对应关系，以及应用 $\mathcal{L}_{CORAL}$前后的虚拟试穿图像可视化。最右侧的图展示了对应关系质量，该质量通过关键点正确率（$\alpha = 16$）和注意力熵衡量，并在不同时间步和层上取平均。

附录 F和G 支持了我们所用评估协议的可靠性。

5.2 定量比较

主要结果。 表1和表2 报告了在 VITON-HD（Choi 等人，2021年）和 Dress-Code（Morelli 等人，2022年）上的定量比较。CORAL 在两个基准数据集上均取得了最佳整体性能。表 3 进一步表明，CORAL 在所有指标上均取得了配对设置下的最佳结果，并且在非配对设置下保持竞争力，同时在额外的真实场景数据集上显著优于其他基线。在这三张表中，我们的基线（不含 $\mathcal{L}_{CORAL}$）已经表现强劲，但加入 CORAL 后每项指标都得到持续提升，表明其在具有竞争力的基线之上仍能带来可靠增益，并且能够稳健地改善感知保真度，即使是在真实场景设置下也是如此。

表5。 **VITON-HD 上的损失组件消融实验。**

	$\mathcal{L}_{corr}$	$\mathcal{L}_{ent}$	成对				非成对	
			结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	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

基于视觉语言模型的评估结果。 表4 展示了基于视觉语言模型的评估定量结果。与以往方法相比，CORAL 在三项指标上均取得了最佳性能。通过在正确的人物-服装位置确定查询-键匹配，CORAL 能够从正确区域迁移外观，从而提升 GTC。更准确的空间对应关系还能更好地保留轮廓、长度和图案几何形状等服装属性，进而提高 TAC。这些增益带来了更自然的穿着效果和姿势跟随，并体现为最高的 FPC。

5.3. 定性比较

图5 展示了先前工作与CORAL之间的定性比较。CORAL保留了服装全局属性，同时恢复了基线经常遗漏的精细局部细节（例如，标志）。我们还观察到，CORAL使服装文字保持清晰可读，而其他方法往往会使其模糊、变形或去除。此外，一些基线会因细微的姿势偏移而偏离输入姿势，而CORAL则使姿态更接近输入。这一改进来自于校正DiT注意力中的人物-服装匹配，从而从正确位置迁移外观。与先前工作及基线的更多定性比较见附录J。

5.4. 消融研究

损失组件消融实验。图6 和表5 凸显了我们损失设计的效果。（a）和（I）展示了基线，其中注意力定位较差，常常分散并偏离查询相关区域，从而产生错位的对应关系和伪影。（a）中的图案放置经常不正确，而（I）的性能最弱。（b）和（II）展示了 $\mathcal{L}_{\text{ent}}$的效果，它锐化注意力并改进细节迁移（例如颜色、纹理）。然而， $\mathcal{L}_{\text{ent}}$ 并不能校正模型匹配的位置，因此长度等服装粗略属性往往仍然不准确。此外，当锐化作用于错误匹配时，模型往往会迁移不恰当的外观。这解释了为何（II）在感知指标和重建指标（结构相似性指数、学习感知图像块相似度）上优于（I），但在具有挑战性的情形中仍表现出错位。（c）和（III）展示了 $\mathcal{L}_{\text{corr}}$的效果，它将对应关系重新定位到正确的服装区域，并改进粗略属性（例如轮廓、图案几何形状）。尽管如此，注意力仍然相对弥散；当匹配到的键注意力质量较低时，会限制精细细节的准确迁移。与（b）相比，（c）的几何形状更准确，但在色彩保真度等外观丰富度方面较为落后。（d）和（IV）展示了 $\mathcal{L}_{\text{ent}}$ 与 $\mathcal{L}_{\text{corr}}$的组合，它整合了这些互补优势。由此同时实现准确对齐与清晰的细节迁移，从而在所有指标上取得最佳定性结果和最强定量性能。

CORAL 对注意力对应关系的分析。图7总结了 $\mathcal{L}_{\text{CORAL}}$ 对人物→服装注意力 A^{l}的影响$[\mathcal{P}, \mathcal{G}]$ 以及生成结果。若不加入 $\mathcal{L}_{\text{CORAL}}$，对应关系图中仍会在局部纹理和服装边界附近出现明显错配，即使整体垂直对齐看起来合理。这些错配会导致生成结果中的徽标图案不够清晰。加入 $\mathcal{L}_{\text{CORAL}}$ 后，对应关系图得到一致修正，生成的匹配结果更贴近图7中的 DINOv3伪真实匹配，从而改善文字细节，并正确生成字母。图7还从定量角度展示了这一点： $\mathcal{L}_{\text{CORAL}}$ 使关键点正确率提高了34%，同时降低了注意力熵，表明注意力中的对应关系更加准确、清晰。

6. 结论

我们分析了 DiT 的虚拟试穿完整三维注意力如何建立人物→服装对应关系，并表明 RGB空间人物与服装对齐依赖于准确的查询-键对应关系。我们提出了 CORAL，这是一种基于 DiT 的框架，利用稳健的 DINOv3 对应关系对齐人物与服装的查询-键匹配。CORAL 采用对应关系蒸馏损失，使稳健匹配与注意力对齐，并结合熵最小化损失来增强注意力。因此，CORAL 在所有标准基准数据集以及一个真实场景基准上均实现了最先进的性能。大量分析和消融研究进一步证明了我们设计选择的有效性。

致谢

本研究部分由韩国政府（MSIT）资助的韩国信息通信技术规划与评估院（IITP）项目提供支持（编号：RS-2025-25441313，多模态人工智能智能体专业人才培养计划，贡献比例：50%）。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能带来许多潜在的社会影响，但我们认为其中没有任何一项需要在此特别强调。

参考文献

An, H., Kim, J. H., Park, S., Jung, J., Han, J., Hong, S., and Kim, S. 跨视图补全模型是零样本对应关系估计器。载于 计算机视觉与模式识别会议论文集，第1103–1115 页，2025年。 5

Attanasio, G., Nozza, D., Hovy, D., and Baralis. 基于熵的注意力正则化使消除非预期偏差不再受列表限制，2022年。网址 https://arxiv.org/abs/2203.09192。 6

Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Laaksonen, J., Shah, M., and Khan, F. S. 通过去噪扩散模型进行人物图像合成，2023年。网址 https://arxiv.org/abs/2211.12500。 2，3

Bi´nkowski, M.、Sutherland, D.J.、Arbel, M.和Gretton, A. 揭秘最大均值差异生成对抗网络，2021年。网址 https://arxiv.org/abs/1801.01401。 7,19

Cao, M.、Wang, X.、Qi, Z.、Shan, Y.、Qie, X.和Zheng, Y. MasaCtrl：面向一致图像合成与编辑的免调参互自注意力控制，2023年。网址https://arxiv.org/abs/2304.08465。 3

Cao, Z.、Hidalgo, G.、Simon, T.、Wei, S.-E.和Sheikh, Y. OpenPose：基于部件亲和场的实时多人二维姿态估计。*IEEE*模式分析与机器智能汇刊，43(1):172–186，2021年。doi: 10.1109/TPAMI.2019.2929257。 22,23

Carion、N.、Gustafson、L.、Hu、Y.-T.、Debnath、S.、Hu、R.、Suris、D.、Ryali、C.、Alwala、K. V.、Khedr、H.、Huang、

A.、Lei、J.、Ma、T.、Guo、B.、Kalla、A.、Marks、M.、Greer、J.、Wang、M.、Sun、P.、R¨adle、R.、Afouras、T.、Mavroudi、E.、Xu、K.、Wu、T.-H.、Zhou、Y.、Momeni、L.、Hazra、R.、Ding、S.、Vaze、S.、Porcher、F.、Li、F.、Li、S.、Kamath、A.、Cheng、H. K.、Doll´ar、P.、Ravi、N.、Saenko、K.、Zhang、P.，以及 Feichtenhofer、C. Sa m 3：基于概念的万物分割，2025年。链接 https://arxiv.org/abs/2511.16719。14，23

Chen, M., Chen, X., Zhai, Z., Ju, C., Hong, X., Lan, J. 和 Xiao, S. Wear-any-way：通过稀疏对应对齐实现可操控的虚拟试穿，2024年。网址 https://arxiv.org/abs/2403.12965。3，18

Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T. 和 Che, W. 迈向推理时代：面向推理大型语言模型的长链式思维综述，2025年。网址 https://arxiv.org/abs/2503.09567。19

Cheng, Z., Xu, B., Gong, L., Song, Z., Zhou, T., Zhong, S., Ren, S., Chen, M., Meng, X., Zhang, Y., Li, Y., Ren, L., Chen, W., Huang, Z., Zhan, M., Wang, X. 和 Feng, F. 使用多模态多图像推理基准评估多模态大型语言模型，2025年。网址https://arxiv.org/abs/2506.04280。19

Cho, S., Huang, J., Nam, J., An, H., Kim, S., and Lee, J.- Y. 用于点跟踪的局部全配对对应关系。发表于欧洲计算机视觉会议，第 306–325 页。施普林格，2024年。5

Choi, S., Park, S., Lee, M., and Choo, J. Viton-hd：通过感知错位的归一化实现高分辨率虚拟试穿，2021年。网址：https://arxiv.org/abs/2103.16874。6，7，8，14，15，16，17，25，28

Choi, Y., Kwak, S., Lee, K., Choi, H., and Shin, J. 提升真实场景下真实感虚拟试穿的扩散模型，2024年。网址：https://arxiv.org/abs/2403.05139。2，3，4，7，14，20，22

Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., 和 Liang, X. CatVTON：对于基于扩散模型的虚拟试穿，拼接就是你所需要的一切，2025年。网址 https://arxiv.org/abs/2407.15886。2，3，7，14，20

Cong, H., Fu, L., Zhang, R., Zhang, Y., Wang, H., He, J., 和 Gao, J. 基于梯度孪生网络的图像质量评估，2022 年。网址 https://arxiv.org/abs/2208.04081。6，7，19

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip：通过指令微调实现通用视觉语言模型，2023年。网址 https://arxiv.org/abs/2305.06500。19

Esser,P.，Kulal,S.,Blattmann,A.,Entezari,R.,M¨uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. 扩展整流流变换器，实现高分辨率图像合成，2024年。网址 https://arxiv.org/abs/2403.03206。2，3

Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., 和 Liu, J. Any2AnyTryOn：利用自适应位置嵌入实现多样化的虚拟服装任务，2025年。网址 https://arxiv.org/abs/2501.15891。7，20

G¨uler, R. A., Neverova, N. 和 Kokkinos, I. Densepose：野外稠密人体姿态估计，2018年。网址 https://arxiv.org/abs/1802.00434。14，22，23

Han, X., Wu, Z., Wu, Z., Yu, R. 和 Davis, L. S. Viton：基于图像的虚拟试穿网络，2018年。网址 https://arxiv.org/abs/1711.08447。7

Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A. 和 Yi, K. M. 使用稳定扩散进行无监督语义对应关系，2023年。网址 https://arxiv.org/abs/2305.15581。3

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. 和 Cohen-Or, D. 使用交叉注意力控制进行提示到提示的图像编辑，2022年。网址 https://arxiv.org/abs/2208.01626。3

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. 和 Hochreiter, S. 采用双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡，2018年。网址 https://arxiv.org/abs/1706.08500。7

Hong, S., Cho, S., Nam, J., Lin, S. 和 Kim, S. 使用4D卷积 Swin Transformer进行代价聚合以实现少样本分割。载于 欧洲计算机视觉会议，第108–126页。施普林格，2022年。5

Hong, S., Nam, J., Cho, S., Hong, S., Jeon, S., Min, D. 和 Kim, S. 神经匹配场：用于视觉对应的匹配场隐式表示。 神经信息处理系统进展，35:13512–13526，2022年。5

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. 和 Chen, W. 低秩适配：大型语言模型的低秩适配，2021年。网址 https://arxiv.org/abs/2106.09685。16

Huang, L., Wang, W., Wu, Z.-F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., and Zhou, J. 面向扩散 Transformer 的上下文内低秩适配，2024年。网址 https://arxiv.org/abs/2410.23775。4,14

Huang, Z., Li, H., Xie, Z., Kampffmeyer, M., Cai, Q., and Liang, X. 通过三维感知全局对应关系学习实现困难姿态虚拟试穿，2022年。网址 https://arxiv.org/abs/2211.14052。 3,18

Jeong, H., Huang, C.-H. P., Ye, J. C., Mitra, N., and Ceylan, D. Track4gen：教会视频扩散模型追踪点，从而改进视频生成，2025年。网址 https://arxiv.org/abs/2412.06016。 3

Jia, B., Zhang, J., Zhang, H., and Wan, X. 探索并评估多模态大语言模型的多模态知识推理一致性，2025年。网址https://arxiv.org/abs/2503.04801。 19

Jiang, W.、Trulls, E.、Hosang, J.、Tagliasacchi, A. 和 Yi, K. M. COTR：用于跨图像匹配的对应关系变换器，2021年。URLhttps://arxiv.org/abs/2103.14167。 5Jin, S.、Kim, S.、Chung, D.、Lee, J.、Choi, H.、Nam, J.、Kim, J. 和 Kim, S. Matrix：用于交互感知视频生成的掩码轨迹对齐，2025a。URL https://arxiv.org/abs/2510.07310。 3Jin, S.、Nam, J.、Kim, J.、Chung, D.、Kim, Y.-S.、Park, J.、Chu, H. 和 Kim, S. 用于真实场景中基于示例的语义图像合成的外观匹配适配器，2025b。URLhttps://arxiv.org/abs/2412.03150。 3Kim, J.、Gu, G.、Park, M.、Park, S. 和 Choo, J. StableVITON：使用潜在扩散模型学习用于虚拟试穿的语义对应关系，2023年。URL https://arxiv.org/abs/2312.01725。 2， 3， 4， 7Kim, J.、Jin, H.、Park, S. 和 Choo, J. PromptDresser：通过生成式文本提示和提示感知掩码提升虚拟试穿的质量与可控性，2025年。URLhttps://arxiv.org/abs/2412.16978。 2， 3Lee, J.-Y.、Cha, B.、Kim, J. 和 Ye, J. C. 在扩散模型中将文本与图像对齐比你想象的更容易，2025年。URLhttps://arxiv.org/abs/2503.08250。 3Li, F.、Zhang, R.、Zhang, H.、Zhang, Y.、Li, B.、Li, W.、Ma, Z. 和 Li, C. LLaVA-NeXT-Interleave：解决大型多模态模型中的多图像、视频和3D问题，2024年。URLhttps://arxiv.org/abs/2407.07895。 19Li, N.、Shih, K. J. 和 Plummer, B. A. 使用合成对和误差感知噪声调度增强虚拟试穿，2025年。URL https://arxiv.org/abs/2501.04666。 3

Li, P.、Xu, Y.、Wei, Y. 和 Yang, Y. 用于人体解析的自我校正，2019年。URL https://arxiv.org/abs/1910.09777。 14， 23

Liang, J., Zeng, H., Cui, M., Xie, X., andZhang, L. PPR10K：一个包含人体区域掩码和组级一致性的大规模人像照片修饰数据集，2021年。网址 https://arxiv.org/abs/2105.09180。 7， 21， 26

Liu, Z., Luo, P., Qiu, S., Wang, X., 和 Tang, X. Deepfashion：利用丰富的标注增强服装识别与检索的鲁棒性。在 *2016 IEEE* 计算机视觉与模式识别会议（*CVPR*），第 1096–1104 页，2016 年。doi：10.1109/CVPR.2016.124。 23

Loshchilov, I.与Hutter, F. 解耦权重衰减正则化，2019年。URL https://arxiv.org/abs/ 1711.05101. 14 Mishchenko, K.与Defazio, A. Prodigy：一种快速自适应的无参数学习器，2024年。URL https: // arxiv.org/abs/2306.06101. 16 Morelli, D.、Fincato, M.、Cornia, M.、Landi, F.、Cesari, F.与Cucchiara, R. Dress Code：高分辨率多类别虚拟试穿，2022年。URL https://arxiv. org/abs/2204.08532. 7, 8, 14, 16, 25, 29 Morelli, D.、Baldrati, A.、Cartella, G.、Cornia, M.、Bertini, M.与Cucchiara, R. Ladi-vton：潜扩散文本反演增强的虚拟试穿，2023年。URL https: // arxiv.org/abs/2305.13501. 3 Nam, J.、Kim, H.、Lee, D.、Jin, S.、Kim, S.与Chang, S. Dreammatcher：用于语义一致的文本到图像个性化的外观匹配自注意力。载于IEEE/CVF计算机视觉与模式识别会议论文集，第8100–8110页，2024年a。3, 5, 17 Nam, J.、Lee, G.、Kim, S.、Kim, H.、Cho, H.、Kim, S.与Kim, S. 用于密集匹配的扩散模型，2024年b。URL https://arxiv.org/abs/2305.19094. 3, 5 Nam, J.、Son, S.、Chung, D.、Kim, J.、Jin, S.、Hur, J.与Kim, S. 视频扩散Transformer中涌现的时间对应关系。arXiv预印本 arXiv:2506.17220，2025年a。3, 5 Nam, J.、Son, S.、Xu, Z.、Shi, J.、Liu, D.、Liu, F.、Kim, S.与Zhou, Y. Visual persona：全身人体定制的基础模型。载于计算机视觉与模式识别会议论文集，第18630–18641页，2025年b。2 OpenAI与J. A.等人。GPT-4技术报告，2024年。URL https://arxiv.org/abs/2303.08774. 20

Peebles, W.与Xie, S. 基于Transformer的可扩展扩散模型, 2023年。URLhttps://arxiv.org/abs/2212.09748. 2,3

Rombach, R.、Blattmann, A.、Lorenz, D.、Esser, P. 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成，2022年。网址https://arxiv.org/abs/2112.10752。 3

Shin, C.、Choi, J.、Kim, H. 和 Yoon, S. 结合图像修复的大规模文本生成图像模型是一种零样本主体驱动图像生成器, 2025年。网址 https://arxiv.org/abs/2411.15466。 4

Sim´eoni, O.、Vo, H. V.、Seitzer, M.、Baldassarre, F.、Oquab, M.、Jose, C.、Khalidov, V.、Szafraniec, M.、Yi, S.、Rama- monjisoa, M.、Massa, F.、Haziza, D.、Wehrstedt, L.、Wang, J.、Darcet, T.、Moutakanni, T.、Sentana, L.、Roberts, C.、Vedaldi, A.、Tolan, J.、Brandt, J.、Couprie, C.、Mairal, J.、J´egou, H.、Labatut, P. 和 Bojanowski, P. DINOv3，2025年。网址https://arxiv.org/abs/2508.10104。 2，3，5， 6，14， 17

Soloveitchik, M., Diskin, T., Morin, E., andWiesel, A. 条件弗雷歇初始距离, 2022年. URLhttps://arxiv.org/abs/2103.11521. 19

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. RoFormer: 增强型变换器与旋转位置嵌入, 2023年. URLhttps://arxiv.org/abs/2104.09864. 3,4

Tan, Z., Liu, S., Yang, X., Xue, Q., and Wang, X. Ominicontrol: 面向扩散变换器的最小化通用控制, 2025a. URLhttps://arxiv.org/abs/2411.15098. 2,4

Tan, Z., Xue, Q., Yang, X., Liu, S., and Wang, X. Omnicontrol2: 面向扩散变换器的高效条件注入, 2025b. URLhttps://arxiv.org/abs/2503.08280. 16

Tang, L., Jia, M., Wang, Q., Phoo, C. P. 和 Hariharan, B. 基于图像扩散的涌现对应关系，2023年。链接https://arxiv.org/abs/2306.03881。 3Truong, P., Danelljan, M., Yu, F. 和 Gool, L. V. 用于稠密对应关系无监督学习的变形一致性，2021年。链接 https://arxiv.org/abs/2104.03308。3

Tumanyan, N., Geyer, M., Bagon, S. 和 Dekel, T. 用于文本驱动图像到图像转换的即插即用扩散特征，2022年。链接https://arxiv.org/abs/2211.12572。3

Wan, S., Li, Y., Chen, J., Pan, Y., Yao, T., Cao, Y., 和 Mei, T. 通过聚焦服装改进虚拟试穿

扩散模型，2024年。网址 https://arxiv.org/abs/2409.08258。 3Wan, S., Chen, J., Pan, Y., Yao, T., 和 Mei, T. 将视觉对应关系融入扩散模型以实现虚拟试穿，2025年。网址https://arxiv.org/abs/2505.16977。 3Wang, C., Chen, T., Chen, Z., Huang, Z., Jiang, T., Wang, Q., 和 Shan, H. FLDM-VTON：用于虚拟试穿的忠实潜在扩散模型，2024年。网址 https://arxiv.org/abs/2404.14162。3 Xiao, Z., Zhou, Y., Yang, S., 和 Pan, X. 视频扩散模型无需训练即可充当运动解释器和控制器，2024年。网址https://arxiv.org/abs/2405.14864。3

Xiaobin, H., Yujie, L., Donghao, L., Xu, P., Jiangning, Z., Junwei, Z., Chengjie, W. 和 Yanwei, F. VTBench：面向现实场景虚拟试穿模型的综合基准测试套件，2025年。链接 https://arxiv.org/abs/2505.19571。7

Xie, Z.、Huang, Z.、Dong, X.、Zhao, F.、Dong, H.、Zhang, X.、Zhu, F. 和 Liang, X. GP-VTON：通过协作式局部流与全局解析学习实现通用虚拟试穿，2023年。网址 https://arxiv.org/abs/2303.13756。 2，3、7

Xu, Y.、Gu, T.、Chen, W. 和 Chen, C. OOTDiffusion：基于潜扩散的服装外穿融合，用于可控虚拟试穿，2024年。网址https://arxiv.org/abs/ 2404.01089。 2、 3， 7、 20

Yang, X.、Ding, C.、Hong, Z.、Huang, J.、Tao, J. 和 Xu, X. 纹理保持扩散模型，用于高保真虚拟试穿，2024年。网址https://arxiv.org/abs/2404.01089。 2

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., 和 Xie, S. 用于生成的表征对齐：训练扩散 Transformer 比你想象的更容易，2025年。网址 https://arxiv.org/abs/2410.06940。 3， 5， 17

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., 和 Wang, O. 深度特征作为感知度量的惊人效用，2018年。网址 https://arxiv.org/abs/1801.03924。 6， 7， 19

Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., and Cheng, Y. Videorepa：通过与基础模型进行关系对齐来学习视频生成的物理规律, https://arxiv.org/abs/2505.23656。 17

Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong,Y.,Li, H., Xu, M., P´erez-R´ua, J.-M., Patel, A., Xiang, T., Shi, M., and He, S. 学习注意力中的流场，以实现可控人物图像生成，https://arxiv.org/abs/2412.08486。2，3

附录

本补充材料介绍我们提出的将方法推广至其他基于 DiT 的基线方法，并提供实验结果以及进一步方法 CORAL 的应用。 r

- 第 A 节介绍了我们全部实验的实现细节。
- 第B 节详细介绍了实验中使用的数据集及预处理流程。
- 第 C 节将我们的分析和应用扩展到其他 DiT 变体，并突出展示了 CORAL 损失的适用性。

- 第 D 节分析了计算复杂度。
- 第 E 节通过探索其他对齐方法，验证了我们的设计选择。
- 第 F节介绍了我们新颖的基于视觉语言模型的评估协议的细节。
- 第G 节报告了人工评估研究的细节和结果。
- 第H 节给出了人对人场景下的结果，以进一步展示我们的性能。
- 第I节介绍了CORAL在人人服装迁移中的另一项应用。
- 第J 节提供了更多定性结果。
- 第K 节讨论了我们工作的局限性和未来方向。

A. 实现细节

A.1. 训练设置

我们的模型初始化自`FLUX.1-Fill-dev`，并对注意力层进行微调。我们使用 AdamW（Loshchilov 与 Hutter，2019年）优化器进行训练，学习率为 2×10^{-5}，批量大小为 16，在 8 块 NVIDIA B200 GPU 上共训练 8.4k 步。训练和报告结果时使用的分辨率为 1024×768。使用 $\mathcal{L}_{\text{CORAL}}$进行训练时，学习率和批量大小保持不变，损失权重为 $\lambda_{\text{corr}} = 0.01$ 和 $\lambda_{\text{ent}} = 0.1$。我们在整个训练过程中计算 $\mathcal{L}_{\text{CORAL}}$ 所有层上的该指标，并对各层取平均；计算损失时排除循环一致性约束筛除的位置。

B. 训练数据集

B.1. 数据集概述

VITON-HD（Choi 等人，2021年）包含 11,647 张训练图像和 2,032 张测试图像，重点涵盖在影棚环境中拍摄的上身服装。DressCode（Morelli 等人，2022年）包含 48,392 张训练图像和 5,400 张测试图像，涵盖 13,563 件上装、7,151 件下装和 27,678 件连衣裙。人物图像通常为居中拍摄的全身照，并排除了面部区域。在这两个数据集中，服装商品图像通常是在纯白背景上拍摄的平铺商品图。总体而言，这两个基准数据集中的人物图像均在受控的背景和光照条件下拍摄，大多数样本呈现静态的正面姿势，视角变化有限。

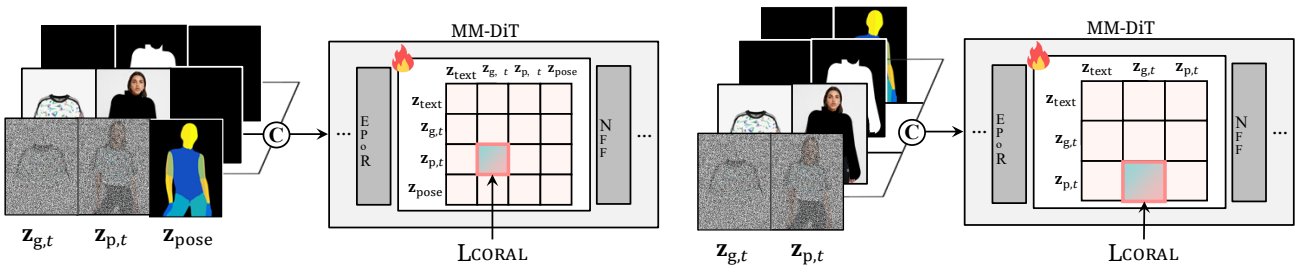
B.2. 数据集预处理

我们遵循先前工作（Chong 等人，2025年；Huang 等人，2024年）对输入提示词的设定：为 VITON-HD 使用固定的输入提示词，并为 DressCode 添加服装类别。对于 VITON-HD，我们以 1:1 的比例同时使用基准数据集中的掩码和由先前工作（Choi 等人，2024年）预处理的掩码。对于 DressCode，我们采用由先前工作（Choi 等人，2024年）预处理的掩码，并在每个训练步骤中应用随机方向性膨胀，膨胀半径从 1 到 20 像素的范围内均匀采样，以增强掩码。输入掩码可能覆盖头发或包等服装无关区域，这使得在存在遮挡时难以保留这些区域在输入中的外观。虽然可以细化掩码以排除这些区域，但这会带来实际风险：如果掩码过于紧凑，可能会暴露人物原本穿着的服装；而且，在不同姿态和视角下准确分割头发等精细结构，通常成本高昂且容易出错。为此，我们使用 human-parser（Li 等人，2019年）或 SAM3（Carion 等人，2025年）提取这些区域，并将其拼接到 DensePose（G¨uler 等人，2018年）条件中，作为额外引导提供。我们使用 DINOv3-B（Sim´eoni 等人，2025年）构造伪真实对应关系，并将循环一致性阈值 γ 设为 3。

C. CORAL 在其他基线上的表现

CORAL 旨在 与模型无关，可插入任何公开人物→ 注意力 ma 的 DiT 架构中 p
用于监督。为了演示 为此，我们将应用扩展到其他 DiT 基线。

C.1. 采用不同姿态注入策略的基线



（a）姿态令牌拼接 **（我们的方法）** (b) 姿态通道拼接

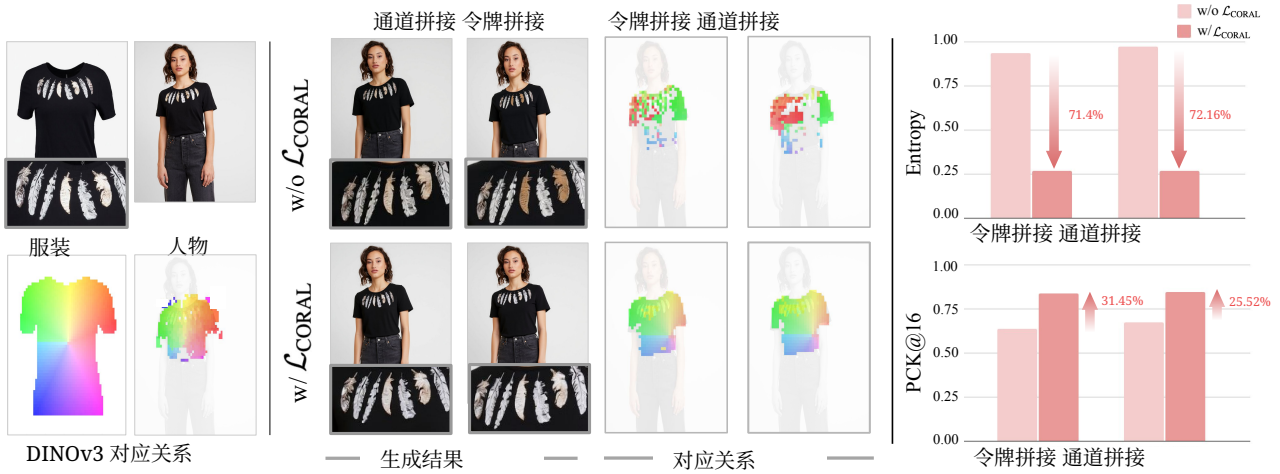
图 *C.1.* **架构对比。**

姿态令牌拼接。 为了利用 DiT 在完整序列上的基于令牌的处理方式，我们考虑了一种将姿态作为额外令牌注入的基线。如图 C.1（a）所示，我们沿令牌维度拼接姿态条件 $\mathbf{z}^{\text{pose}}$ 。这保留了令牌空间中的位置结构，并为模型提供了明确的姿态对应关系。我们的主要实验采用这一 **姿态令牌拼接。** 基线。更多细节请参见主论文第 4.2 节和附录第 A 节。

姿态通道拼接。如图C.1 (b) 所示，我们还设计了一种基线，将姿态条件$\mathbf{z}^{\text{pose}}$ 沿通道维度而非令牌维度注入。具体而言，我们首先通过将$\mathbf{z}_{\text{pose-diptych}}$放置在$\mathbf{z}^{\text{pose}}$ 旁边，并添加一个零填充面板以匹配双联画布局，从而构造姿态双联图。随后，我们沿通道维度拼接$\mathbf{z}_t$、$\mathbf{z}^{\text{diptych}}$、$\mathbf{m}^{\text{diptych}}$和$\mathbf{z}_{\text{pose-diptych}}$ ，并将生成的令牌输入`FLUX.1-Fill-dev`中的输入投影层。这会将投影输入维度从 384 增至 448，因此我们扩展投影层，并将新增权重初始化为零。将原始投影权重复制到扩展后的通道中会导致性能下降且收敛不稳定。

表 *C.1.* **不同姿态注入策略的基线中 $\mathcal{L}_{\text{CORAL}}$ 在采用不同姿态注入策略的基线中的效果，针对 VITON-HD（Choi 等人，2021年）。**

方法	$\mathcal{L}_{\text{CORAL}}$	成对				非成对	
		结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
姿态令牌拼接	✗	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%)
姿态通道拼接	✗	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%)



（a）定性分析 （b）定量分析

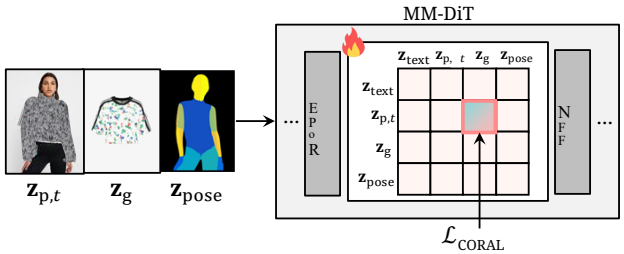
图 *C.2。* **不同姿态注入策略下各基线的效果。**

效果 $\mathcal{L}_{\text{CORAL}}$。我们在两个姿态注入基线上评估 CORAL，分别是**姿态令牌拼接**和**姿态通道拼接**，并分别测试是否使用 $\mathcal{L}_{\text{CORAL}}$。如表 C.1所示，添加 $\mathcal{L}_{\text{CORAL}}$ 后，两个基线的虚拟试穿性能均得到稳定提升，这表明 CORAL 不依赖于特定模型，适用于能够提供人物→服装注意力的各种 DiT 变体。在两个基线中，姿态令牌拼接的表现明显优于姿态通道拼接。这是意料之中的，因为令牌拼接保留了空间对齐的姿态令牌，DiT 可以直接对其进行注意力计算；而通道拼接则通过输入投影以更间接的方式注入姿态。图 C.2(a) 进一步表明，$\mathcal{L}_{\text{CORAL}}$ 能够改善人物→服装注意力中的不匹配，从而实现更准确的服装迁移：例如，第四和第五个羽毛图案不再渗入附近的棕色区域，而是像服装参考图中那样保持洁白。图C.2 (b) 从定量角度进一步证实，在两个基线中，$\mathcal{L}_{\text{CORAL}}$ 均能降低注意力熵并提高关键点正确率，表明对应关系更加清晰且准确。

C.2. 一个额外的高效基线

我们引入了一个在不同设置下训练的额外高效基线，以检验 CORAL 的改进在 C.1 节所述设置之外是否仍保持一致。

架构细节。 额外的基线旨在实现计算效率。遵循先前工作中的紧凑令牌集成方法（Tan 等人，2025b），我们首先使用下采样掩码 $\mathbf{z}^{\text{p}}$ 来将去噪过程限制在服装区域内 $\mathbf{m}_{\text{e}}$，因为对完整的人物潜变量进行去噪会带来不必要的计算，也可能扰动与服装无关的区域。在扩散时间步 t，带噪的人物潜变量定义如下：



图*C.3*。 **高效基线架构**。

$$\mathbf{z}_{\text{p},t} = \Big((1-t)\mathbf{z}_{\text{p}} + t\epsilon\Big) \odot \mathbf{m}_{\text{p}} + \mathbf{z}_{\text{p}} \odot (1-\mathbf{m}_{\text{p}}). \tag{21}$$

我们还保持服装潜变量干净，以避免对服装输入进行不必要的去噪。对于姿态条件注入，我们沿令牌维度拼接姿态条件 $\mathbf{z}^{\text{pose}}$，如 4.2节所述。

实现细节。 我们从 `FLUX.1-dev` 初始化模型，并使用低秩适配进行微调（Hu 等人，2021年），秩设为128。我们使用 Prodigy（Mishchenko 和 Defazio，2024年）优化器进行训练，学习率为 1.0，批量大小为 1，使用 2 块 NVIDIA A100 GPU。我们在 1024 x 1024 分辨率下训练并报告结果。对于使用 $\mathcal{L}_{\text{CORAL}}$ 进行训练，采用相同的学习率和批量大小，损失权重为 $\lambda_{\text{corr}} = 0.01$ 和 $\lambda_{\text{ent}} = 0.1$。其余实现细节遵循附录 A节。

表 C.2。 **对 $\mathcal{L}_{\text{CORAL}}$ 在 VITON-HD（Choi 等人，2021年）和 DressCode（Morelli 等人，2022年）的额外高效基线中的影响。**

$\mathcal{L}_{\text{CORAL}}$	VITON-HD						DressCode					
	成对				非成对		成对				非成对	
	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	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%)

效果： $\mathcal{L}_{\text{CORAL}}$。 应用 CORAL 可提升高效基线在各项整体指标上的表现，相较于我们的主基线，提升尤为显著。在这种采用低秩适配微调的高效训练设置中，基线表现相对较弱，但 CORAL 仍带来了显著提升。提升幅度最大的是基于分布的指标，这表明 CORAL 不仅增强了生成结果的感知保真度，也提升了其整体真实感。

总体而言，各基线上的改进证实了 $\mathcal{L}_{\text{CORAL}}$ 的有效性，并凸显了其对 DiT 的适用性，因为注意力提供了一种显式且可控的对应关系机制。

D. 运行时间和内存

在表C.3中，我们报告了每种基线设置的硬件配置、吞吐量和 GPU 显存峰值。

表 *C.3*。 **速度和内存**。 T 表示训练，I 表示推理。速度单位为张/秒，显存单位为 GB。

设置	$\mathcal{L}_{\text{CORAL}}$	GPU	GPU#（T/I）	批次（T/I）	速度（T/I）	显存（T/I）
令牌拼接	✓	B200	8 / 1	16 / 1	2.61 / 0.091	148 / 40
令牌拼接	✗	B200	8 / 1	16 / 1	3.01 / 0.105	148 / 40
通道拼接	✓	B200	8 / 1	16 / 1	3.41 / 0.119	148 / 40
通道拼接	✗	B200	8 / 1	16 / 1	3.92 / 0.137	148 / 40
高效	✓	A100	1 / 1	1 / 1	0.16 / 0.025	48 / 33
高效	✗	A100	1 / 1	1 / 1	0.18 / 0.025	48 / 33

E. 补充分析

E.1. 对齐目标

表 *E.4*。 **VITON-HD 上的对齐目标消融实验（Choi 等人，2021年）**。

	方法	成对				非成对	
		结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
(I)	基线	0.889	0.055	5.543	0.870	9.641	1.323
(II)	(I) + 特征对齐（REPA）（Yu 等人，2025年）	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) + 查询-键对应对齐**(我们的方法)**	**0.907** (+2.02%)	**0.048** (+12.73%)	**4.962** (+10.48%)	**0.565** (+35.06%)	**8.763** (+9.11%)	**0.880** (+33.48%)

我们比较了两种在训练过程中改进人物与服装对应关系的对齐方法。第一种方法将 DiT 中间特征与 $\phi(\cdot)$ 中的特征描述符对齐，并采用余弦相似度损失，遵循先前的工作（Yu 等人，2025年）。第二种是我们提出的方法 CORAL，通过监督查询-键匹配来对齐基于注意力的对应关系。先前的分析表明，查询-键相似度主要决定生成图像元素的空间布局，而值特征主要承载外观细节（Nam 等人，2024a），受此启发，我们对齐查询—键对应关系以纠正空间匹配，同时不改变模型学得的外观表示。先前的工作（Zhang 等人，2025年）强调，直接特征对齐方法（Yu 等人，2025年）旨在加速从头训练，而非通过微调迁移预训练模型。这种硬性对齐会将 DiT 中间特征拉向 DINOv3 特征，可能破坏预训练期间学得的固有生成表示，导致收敛速度变慢。我们在当前设定下评估了这种直接特征对齐基线，并报告了结果。



图 *E.4.* **中间特征对齐与我们的方法的对比**。

特征对齐的实现细节。 我们基于第4.2节所述主基线所采用的相同架构修改，构建中间特征对齐基线。为了将扩散变换器中的中间表示与外部语义特征对齐，我们遵循先前工作（Yu 等人，2025年），该方法将变换器特征与从预训练视觉编码器提取的表示进行对齐。我们使用 DINOv3-B（Sim´eoni 等人，2025年）作为外部编码器，与我们的主基线保持一致。

我们为每个变换器块的输出附加一个轻量级投影头。每个投影头是一个带有 SiLU 激活函数的三层多层感知机，将 DiT 隐藏状态映射到 DINOv3 嵌入空间。

给定服装图像 I_{g} 以及人物图像 I_{p}，我们将它们水平拼接，形成 $\hat{I} \in \mathbb{R}^{H\times 2W\times 3}$，并提取 t

来自冻结编码器的 DINOv3 特征 ϕ_{g}， $\phi_{\mathrm{p}} \in \mathbb{R}^{hw \times 768}$ 。我们将它们拼接为：

$$\hat{\phi} = [\, \phi_{\mathrm{g}} \,\|\, \phi_{\mathrm{p}} \,] \in \mathbb{R}^{2hw \times 768}. \tag{22}$$

令$\mathbf{h}$ 表示扩散Transformer$\mathbf{v}_\theta$的中间隐藏表示，并令 $f_\omega(\cdot)$为可训练的投影头。我们定义逐补丁对齐损失：

$$\mathcal{L}_{\mathrm{REPA}}(\theta, \omega) := -\mathbb{E}_{\hat{I}, \epsilon, t} \left[\frac{1}{K} \sum_{k=1}^{K} \mathrm{sim}\left(\hat{\phi}_k, f_\omega(\mathbf{h}_k)\right) \right], \tag{23}$$

其中， k 表示补丁索引， K 表示补丁数量。我们在每个变换器块的输出处，使用对应的投影头计算该损失，并对各层取平均。

最终目标是在标准流匹配损失上加上投影损失：

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{velocity}} + \lambda_{\mathrm{REPA}}\, \mathcal{L}_{\mathrm{REPA}}, \tag{24}$$

其中，我们在所有实验中均设定 $\lambda_{\mathrm{REPA}} = 0.1$。

比较与分析。 表 E.4 比较了这两种对齐方法，结果显示，CORAL 的试穿质量更高，结构相似性指数更高，学习感知图像块相似度更低。基于对应关系的虚拟试穿方法利用特征级对齐或基于特征的匹配来迁移服装（Chen 等人，2024年；Huang 等人，2022年）。既有分析表明，查询–键相似度与空间对应关系密切相关，而值通路则更多与外观相关。由于中间特征同时编码了这两类信息，直接对齐中间特征也会影响预训练期间学到的固有表示。通过监督查询键匹配，CORAL 加强了人物与服装对齐，同时在很大程度上保留了模型学到的外观表示。相比之下，特征级对齐会导致中间特征发生更大变化，随后在表 E.4 中所有指标均大幅下降，并在图 E.4中出现纹理错位。

F. 基于视觉语言模型的评估

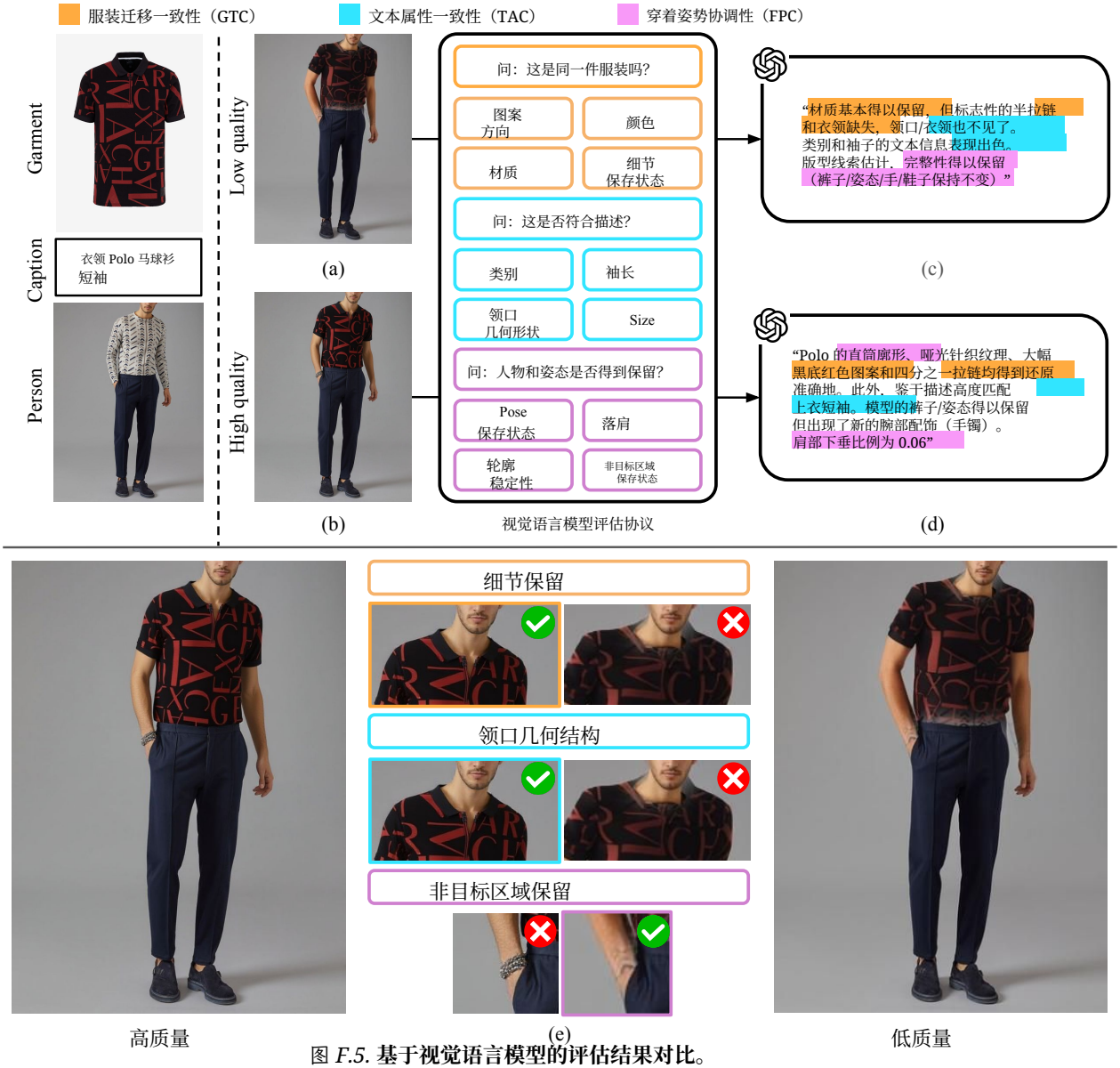


图 *F.5.* **基于视觉语言模型的评估结果对比。**

为何需要基于视觉语言模型的评估? 如第5.1节所述，弗雷歇起始距离（Soloveitchik 等人，2022年）、KID（Bi´nkowski 等人，2021年）、结构相似性指数（Cong 等人，2022年）和学习感知图像块相似度（Zhang 等人，2018年）等标准指标不足以评估虚拟试穿，因为它们主要反映整体真实感或低层次相似度，无法评估特定参考服装是否得到忠实迁移并且穿着效果是否合理。遵循主论文第5.1 节的做法，我们因此提出一种基于视觉语言模型的评估协议，包含三个标准：服装迁移一致性（GTC）、文本属性一致性（TAC）和穿着姿势协调性（FPC）。所有基于视觉语言模型的评估均使用 GPT-5 mini 作为评估器。用于视觉语言模型评估的完整输入提示见图K.18。

图 F.5 说明了采用该协议的必要性。尽管图 F.5 (a) 和 (b) 中的两个输出均由相同的输入服装和人物生成，但它们在是否保留服装细节、属性和服装贴合姿态方面存在差异。如图 F.5 (c) 和 (d) 所示，我们的评估器会给出 GTC、TAC 和 FPC 分数，并附上自然语言解释，说明哪些视觉证据支持各项分数。(e) 展示了与各项评判标准对应的生成图像细节视图。这一设计基于近期研究发现（Cheng 等人，2025年；Jia 等人，2025年；Chen 等人，2025年）指出，视觉语言模型（Dai 等人，2023年；Li 等人，2024年）擅长图像和多图像推理，因此非常适合评判超越分布级相似性的服装特定保真度。

G. 人工评估

人工评估详情。我们开展了一项用户研究，以衡量人类感知下的虚拟试穿质量，并验证基于视觉语言模型的评估是否能反映人类偏好。参与者按照第 5.1 节中使用的相同三项评判标准，对每项生成结果进行评分：服装迁移一致性（GTC）、文本属性一致性（TAC）和合身度与姿势协调性（FPC）。每项评判标准均采用五分制（1 至 5 分）。图 K.17 展示了用户研究中的示例问题。

为使评判标准具体明确并减少歧义，我们为每个问题提供了参考信息。对于 GTC，我们展示了输入服装图像，让参与者可以直接比较服装的形状、图案和标志是否得到保留。对于 TAC，我们提供了对服装属性的简短文本描述，该描述是使用预训练视觉语言模型从参考服装中自动提取的（OpenAI 等，2024年）。对于 FPC，我们展示了输入人物图像，以便参与者判断服装是否穿着自然、是否符合目标姿态和身体几何形态，同时不改变非目标区域。

共有 54 名参与者回答了 45 个问题。为确保公平比较，我们从大量输出中抽样，这些输出在 5 个对比模型中使用相同的输入人物图像和服装图像，包括 OOTDiffusion（Xu 等人，2024年）、IDM-VTON（Choi 等人，2024年）、CatVTON（Chong 等人，2025年）、Any2AnyTryOn（Guo 等人，2025年）和 CORAL（我们的方法），以确保评估者间的可靠性。

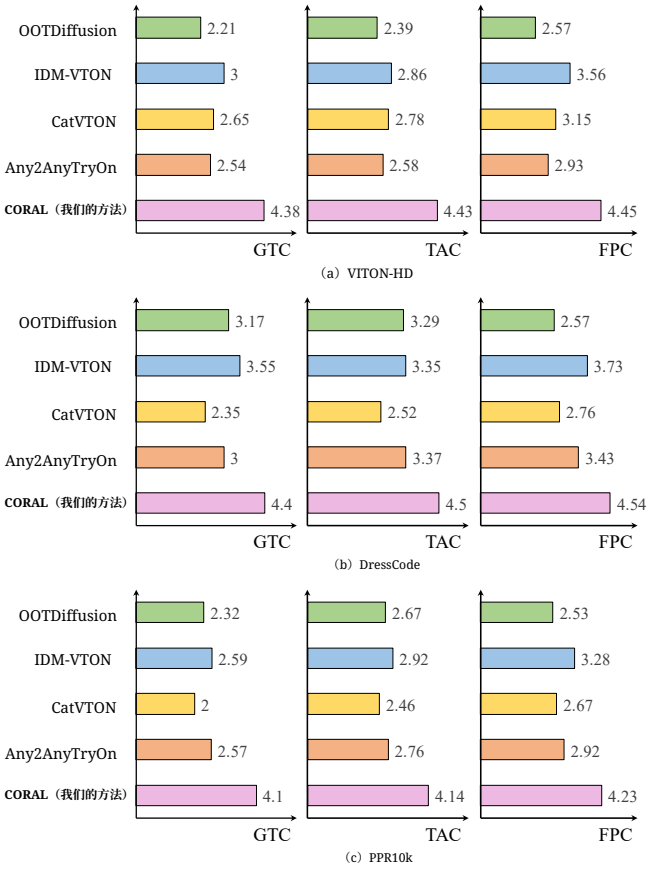


图 *G.6.* **用户研究结果**。

人工评估结果。图 G.6 展示了结果，表明我们的模型在所有评估标准上均优于其他模型。

与视觉语言模型评估的相关性。我们发现，基于视觉语言模型的分数与用户研究结果高度一致，这表明，与标准指标相比，所提出的视觉语言模型评估更忠实地反映了人类偏好。在表4 中，获得较高 GTC/TAC/FPC 分数的方法也往往更受人类参与者青睐，而 CORAL 在两种评估中均表现出色。这种一致性表明，GTC、TAC 和 FPC 能够捕捉对感知试穿质量至关重要的失败案例。

H. 在额外数据集上的评估

H.1. 动机

我们评估 CORAL 学到的对应关系能否泛化到标准虚拟试穿基准之外，并在不如基准设定那样经过精心筛选的输入条件下，依然生成高质量的试穿结果。对于人物输入，我们考虑姿态、视角、背景、取景和遮挡方面变化更大的照片，这些变化范围超出典型基准数据集。对于服装输入，我们考虑不限于干净商品图像的参考图像。

大多数现有的虚拟试穿基准遵循店内服装迁移设定，其中参考服装为平铺商品图，目标人物则在受控的影棚环境中拍摄。然而，这一设定并不能充分反映真实场景输入。用户提供的人物照片往往拍摄于日常环境，取景多样、视角非正面、姿态更为动态，并且经常被配饰或周围物体遮挡。服装参考图像在实际中也可能存在显著差异。用户往往不便为每件服装查找并提供一张干净的商品照片。相反，他们通常依赖随手可得的照片，例如背景杂乱的随手拍摄服装照片，或其他人穿着该服装的图像，这些相比商品图像会引入额外噪声。

因此，我们还构建了成对和非成对测试集，以便在更具挑战性的条件下评估 CORAL，同时确保评估结果可靠。我们使用穿着服装参考进行评估，即参考服装穿在另一个人身上；这通常比干净的商品图像更具挑战性，但仍能提供足够的服装线索用于虚拟试穿。



图 *H.7.* **其他评估数据集样本** **les.**

H.2. 评估数据集详情

我们还使用 PPR10K（Liang 等人，2021年）构建了成对和非成对评估集，因为它提供了在多种真实世界条件下拍摄的高分辨率全身肖像，并包含同一人物的分组图像，从而能够在服装身份一致的情况下进行稳定的配对评估。

数据集概述。 PPR10K 是一个大规模、高质量的人像照片修饰数据集，旨在涵盖多样化的真实世界条件。它包含 11,161 张高分辨率原始人像照片，分为 1,681 组，每组包含 3 至 18 张在同一场景中、短时间内拍摄的同一主体的照片。该数据集经过精心整理，涵盖婚礼、毕业典礼和广告等多种拍摄用途。它还包含多样化的拍摄对象（例如儿童和来自不同地区的人）、各类室内和室外背景，以及范围广泛的光照条件

从白天到夜晚、跨越不同季节。人物肖像涵盖多样的取景方式，从以面部为中心到全身照，并包含各种视角，但不包括完全背对镜头的视角。

数据集预处理。首先，我们筛选人物图像，保留上半身至全身的图像。我们使用 OpenPose 提取二维关键点（Cao 等人，2021年），并保留身体至少清晰可见至膝盖的图像，以剔除以面部为中心的裁剪图像。我们还利用从 DensePose 提取的 IUV 映射（G¨uler 等人，2018年）基于视角启发式地筛选样本，剔除完全背面和侧面视角。对于参考服装设置，我们剔除服装区域过小、无法提供有意义细节的图像，以及过暗、过曝和严重遮挡等极端情况。我们采用与标准虚拟试穿基准相同的预处理流程来提取 DensePose 条件。我们遵循先前工作的方法获取输入掩码（Choi 等人，2024年）。预处理后，最终划分包含 358 个连衣裙样本、320 个上半身样本和 195 个下半身样本。配对设置下有 309 个样本（138 个连衣裙、86 个上半身、85 个下半身），非配对设置下有 564 个样本（ 220 个连衣裙、 234 个上半身、 110 个下半身）。

I. 人人服装迁移

动机。 如 H.1 节所述，实际应用中一个重要的场景是服装穿在另一位着装人物身上，而非以干净的商品照片形式呈现。我们将这种设置称为人人服装迁移。与店内商品照片相比，穿着服装的参考图在现实世界中更为常见，也能展示服装穿上身后的效果；但从野外参考图像中判断哪些内容应当迁移则更加困难。标准的店内场景通常会通过着装人物的参考图像估计服装掩码来解决这一问题，但在存在严重遮挡或复杂背景时，该掩码可能不可靠。如图 I.8 第一行所示，头发遮挡可能导致掩码中缺失服装区域，而试穿结果会迁移这一缺失的形状，使服装看起来像是在相同区域被剪掉了一块。服装掩码也可能丢失整体形状信息



图 *I.8.* **人人服装迁移的动机**。

诸如整体长度和轮廓等信息，这对下装迁移尤为关键。如图I.8所示，对于裙子和连衣裙，经掩码裁剪的纯服装参考会掩盖服装在身体上的终止位置，从而难以保持正确的长度。因此，我们单独训练一个模型，用于人物到人物服装迁移，使用完整参考图像且不估计服装掩码，并应用 CORAL，以在这些真实场景参考下改进人物与服装对应关系。

架构细节。 在人物到人物服装迁移设定中，服装参考图像由另一张着装人物图像提供，而非干净的商品图像。因此，我们将目标人物图像与参考人物图像分别记为 $I_{\mathrm{p_{tgt}}}$ 和 $I_{\mathrm{p_{ref}}}$，并用变分自编码器对其进行编码，以获得相应的潜变量 $\mathbf{z}_{\mathrm{p_{tgt}}}$ 和 $\mathbf{z}_{\mathrm{p_{ref}}}$。沿用第 4.2 节所述的相同双联画布局，在扩散时间步 t，我们将含噪双联画潜变量构造为：

$$\mathbf{z}_t = \left[\, \mathbf{z}_{\mathrm{p_{ref}},t} \,\|\, \mathbf{z}_{\mathrm{p_{tgt}},t} \,\right]. \tag{25}$$

对于穿着服装参考图像和遮罩人物图像的条件注入，我们使用干净的参考人物潜变量和被遮罩的目标人物潜变量，定义如下：

$$\mathbf{z}_{\mathrm{diptych}} = \left[\, \mathbf{z}_{\mathrm{p_{ref}}} \,\|\, \mathbf{z}_{\mathrm{p_{tgt}}} \odot (1 - \mathbf{m}_{\mathrm{e}}) \,\right], \tag{26}$$

其中 $\mathbf{m}_{\mathrm{e}}$ 通过将 M_e 下采样至潜空间得到。二值掩码画布的定义与第 $\mathbf{m}^{\mathrm{diptych}}$ 节 4.2 中相同。

由于可以从 G¨uler 等人，2018年）以及 $I_{\mathrm{p_{tgt}}}$ 和 $I_{\mathrm{p_{ref}}}$中估计出 DensePose（ $I_{\mathrm{pose_{tgt}}}$ 和 $I_{\mathrm{pose_{ref}}}$，因此我们提取与目标人物和参考人物对齐的两种姿态条件，分别记为 $\mathbf{z}^{\mathrm{pose_{tgt}}}$ 和 $\mathbf{z}^{\mathrm{pose_{ref}}}$，然后使用变分自编码器对其进行编码，得到 。

由于参考穿着服装图像还需要额外的姿态输入，通过令牌拼接注入姿态会增加令牌数量，从而提高计算和内存开销。因此，我们改为沿通道维度注入姿态条件。如附录第 C.1 节所述，我们为姿态条件构建双联画布局，$\mathbf{z}_{\text{pose-diptych}} = [\mathbf{z}_{\text{pose}_{\text{ref}}} \| \mathbf{z}_{\text{pose}_{\text{tgt}}}]$，然后在 $\mathbf{z}_t$、$\mathbf{z}^{\text{diptych}}$、$\mathbf{m}^{\text{diptych}}$和 $\mathbf{z}_{\text{pose-diptych}}$ 输入投影层之前，按通道拼接 `FLUX.1-Fill-dev`中的输入。

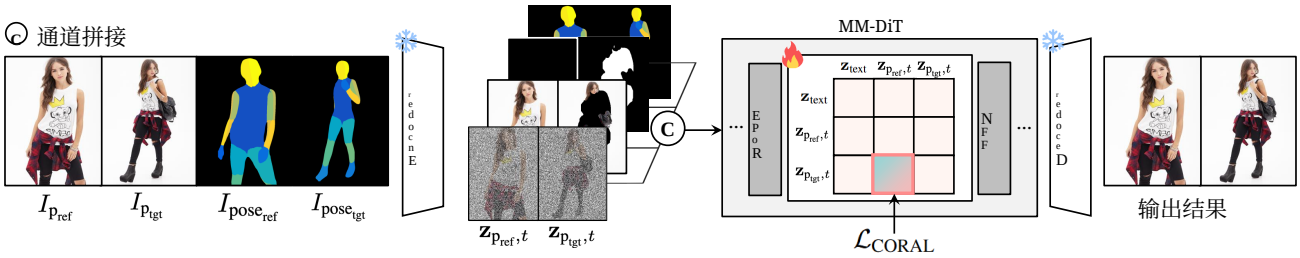


图 *I.9.* **人人架构**。

训练数据集。 现有虚拟试穿基准数据集基于经过筛选的配对集构建，其中参考图像是平铺的店内商品图像，通常居中摆放，背景干净。然而，穿着服装场景尚无成熟的基准数据集。尽管已有一些公开的配对数据集，其中包含多张穿着同一服装的主体图像，但部分图像的服装身份标注有误，而且没有标注指出每组中哪张图像应作为参考穿着服装图像。在穿着服装场景中，许多图像是在自然环境中拍摄的，视角和取景各异，偶尔还存在遮挡，因此有些图像中的服装可能不够清晰，无法作为可靠的参考输入。

为构建人物对人物服装迁移的成对数据，我们使用 DeepFashion（Liu 等人，2016）和一个内部数据集。DeepFashion 为多视角试穿任务提供了 14,120 对样本，每对通常包含涵盖多种服装类别的多个视角，并带有可靠的服装身份标注。我们使用 DensePose（G¨uler 等人，2018年）IUV 映射，筛除视角极端的参考图像，例如明显的侧面或背面视角；并使用人体解析图（Li 等人，2019）筛除裁剪过紧的图像。我们还使用一个涵盖上装、下装和连衣裙的内部数据集，其中不同人物所穿的服装按服装身份配对，服装身份由人工标注人员标注。对于非全身的野外图像，人体解析或 OpenPose（Cao 等人，2021年）可能会失效，因此我们还使用 SAM3（Carion 等人，2025年）生成输入掩码。最终，我们的训练集共包含 23,678 对样本。

定量结果。 如表I.5所示，使用穿着服装的参考图像训练人物到人物的服装迁移模型，其表现优于使用店内商品照片训练、且不经重新训练的模型。在表I.5中，**CORAL-P2P**在推理时使用未掩码参考图像，而其他**CORAL** 设置

表 *I.5.* **PPR10K 上的定量比较**。

方法	成对				非成对	
	结构相似性指数 ↑	学习感知图像块相似度 ↓	弗雷歇起始距离 ↓	KID ↓	弗雷歇起始距离 ↓	KID ↓
CORAL（不含 $\mathcal{L}_{\text{CORAL}}$）	0.877	0.078	44.117	0.015	56.644	1.202
CORAL（含 $\mathcal{L}_{\text{CORAL}}$）	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**

使用通过服装掩码提取的纯服装参考。在这一更具挑战性的设置中，$\mathcal{L}_{\text{CORAL}}$ 有助于模型聚焦于目标服装，避免迁移参考图像中无关部分的细节，这一点也体现在表I.5中涵盖配对和非配对设置的定量结果上。

定性结果。 我们在图*I.8*和图I.8 中展示了人物到人物服装迁移的定性结果。I.10。使用店内商品图像训练的模型，在服装参考以着装人物图像而非干净的商品照片形式提供时，往往难以处理。如图I.8所示，参考图像中的可见服装区域可能因遮挡而不完整，生成的试穿结果会沿着可见部分的边缘生成，使服装看起来像被截断或发生视觉变形。使用人物到人物 配对进行训练，有助于模型即使在此类遮挡情况下也能生成更自然的服装形状；使用**CORAL-P2P**时，衬衫的肩部能够无缝生成。对于下装和连衣裙，由掩码裁剪得到的纯服装参考难以推断衣长，因为它去除了服装在身体上的终止位置及整体轮廓等全局信息，可能导致下摆缩短或长度错误，如图I.8所示。相反，图I.10 进一步表明，**CORAL-P2P** 通过以完整参考人物图像作为条件注入，保留了整体长度和轮廓，使裙长保持在脚踝上方，与参考图像一致。

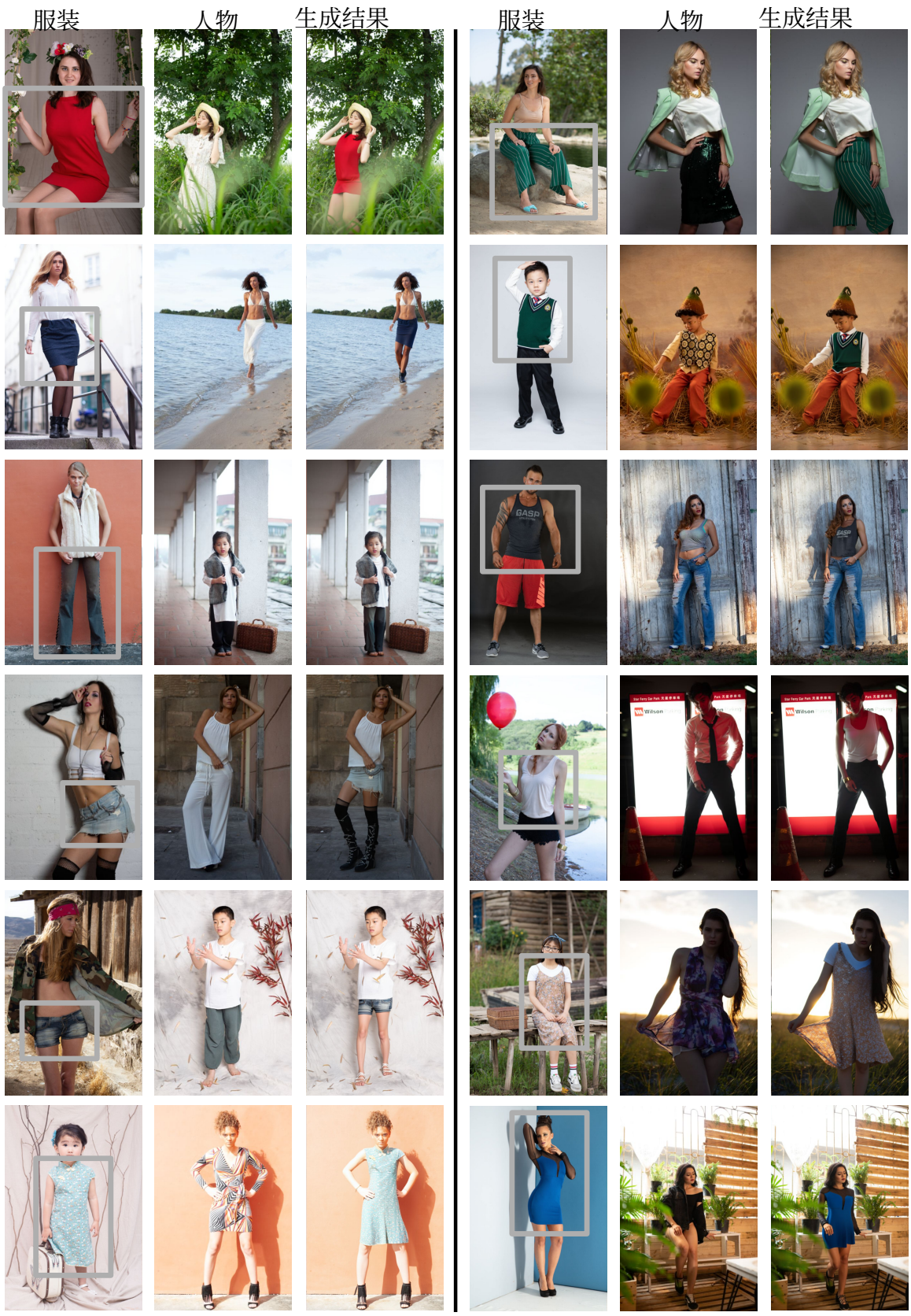


图 *I.10*。 **CORAL-P2P 在 PPR10K 上的定性结果。**

J. 附加定性结果

我们进一步在标准基准和真实场景数据集上，将 CORAL 与近期的虚拟试穿方法进行比较，结果见图 J.11、图 J.12和图 J.13。我们还报告了未使用 $\mathcal{L}_{\text{CORAL}}$ 以及使用 $\mathcal{L}_{\text{CORAL}}$训练的基线模型的结果，表明 CORAL 通过改善人物与服装对齐并减少错位伪影，始终能够提升试穿质量，结果见图 J.15 和图 J.16。此外，我们还在 PPR10K 上评估 CORAL，以检验仅在标准虚拟试穿基准上训练的模型的泛化能力，结果见图 J.14。



图 *J.11*。 **CORAL 在 VITON-HD 上的附加定性比较（Choi 等人，2021年）。**



图 *J.12*。 **CORAL 在 DressCode 上的附加定性比较（Morelli 等人，2022年）。**



图 *J.13.* 在 **PPR10K** 上对 **CORAL** 的额外定性比较（Liang 等人，2021年）。

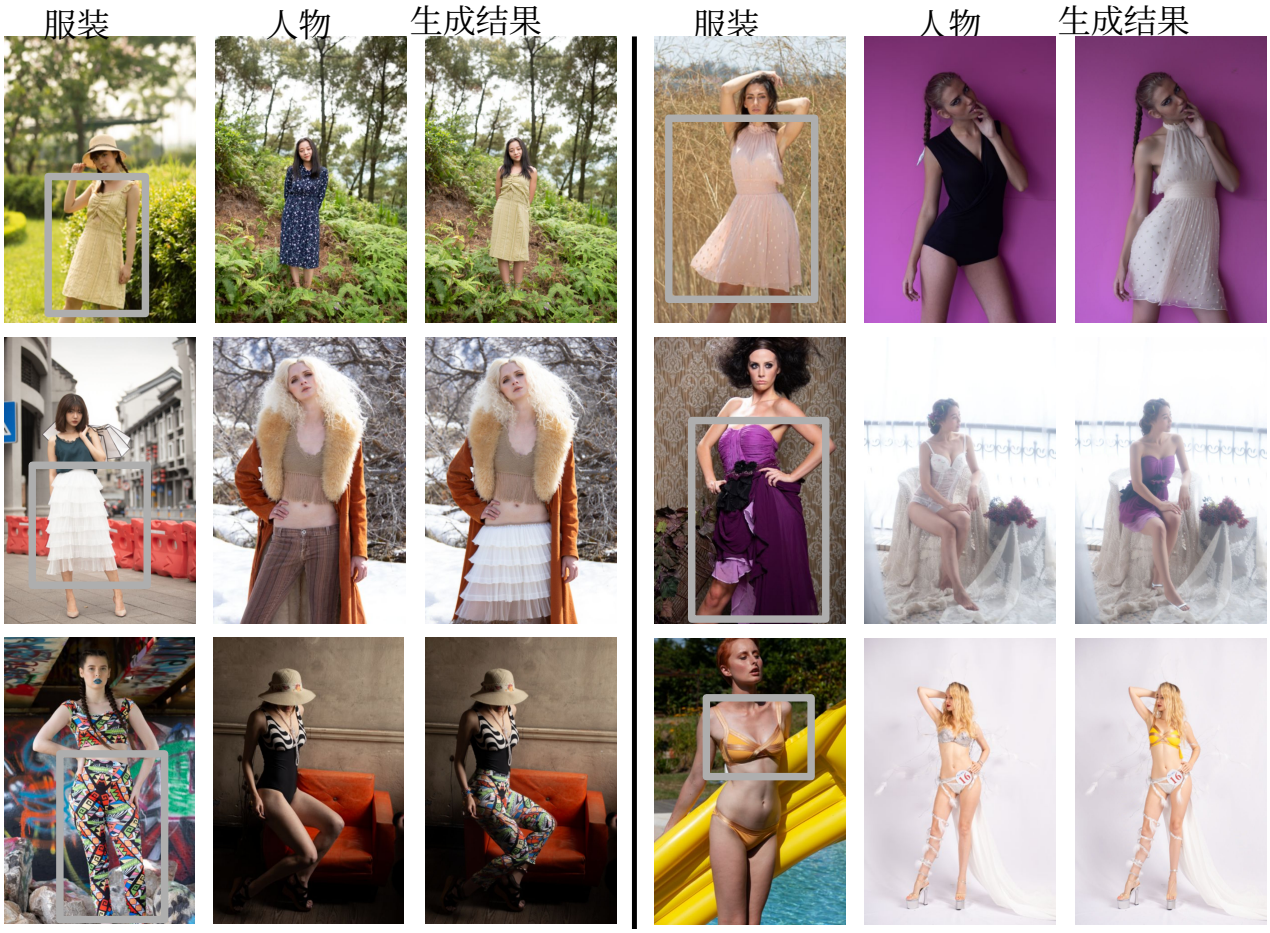


图 *J.14*。**CORAL 在 PPR10K 上的更多定性结果。**

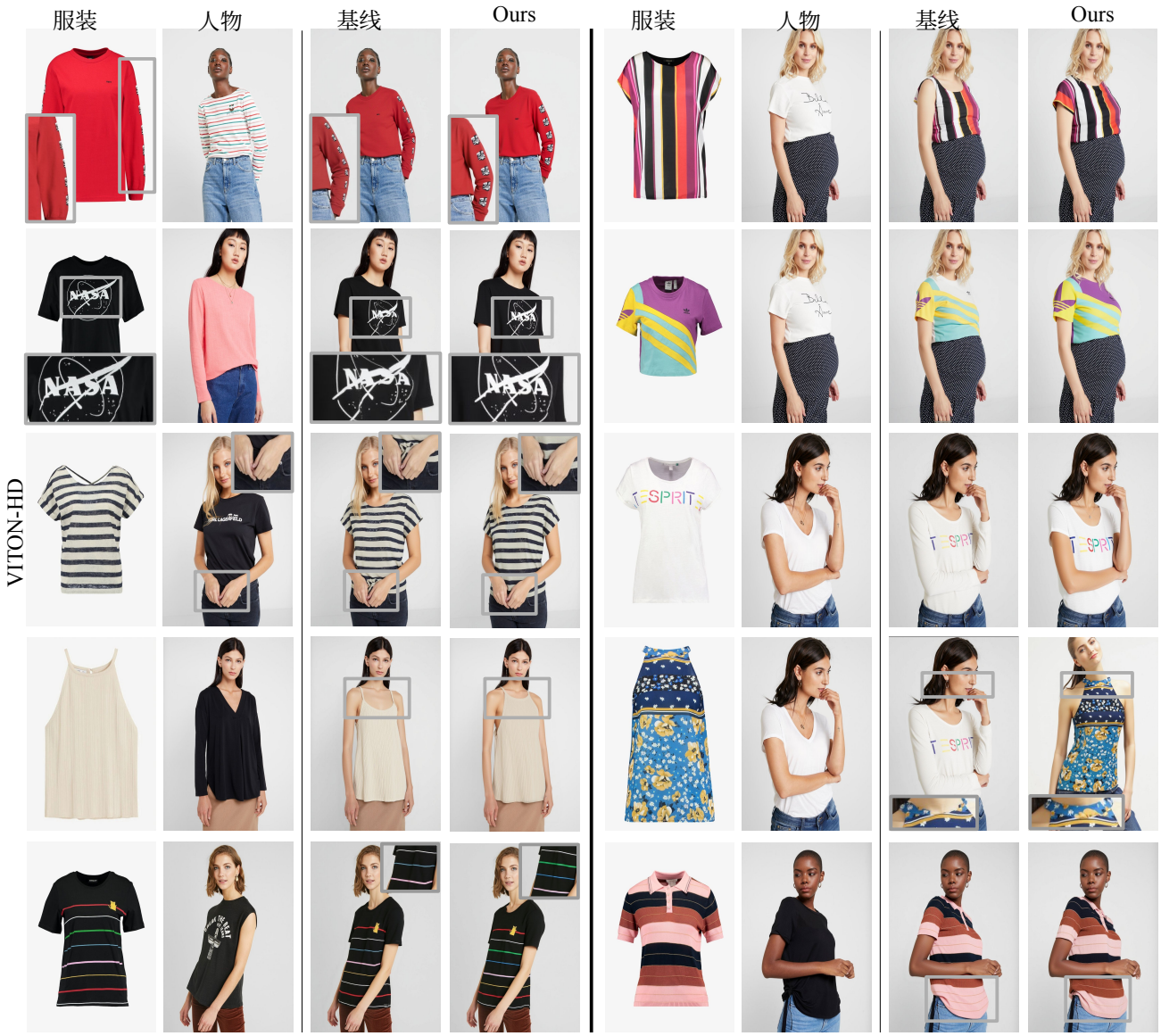


图 *J.15.* **VITON-HD 上的定性比较（Choi 等人，2021年）应用 CORAL 损失前后。**



图 *J.16.* 在 **DressCode** 上的定性比较（Morelli 等人，2022年）**CORAL** 损失前后。

K. 局限性与讨论

依赖 DINOv3 伪真实对应关系。 CORAL 使用基于 DINOv3 的伪真实对应关系来监督查询–键匹配，因此，当 DINOv3 匹配不可靠时，其对应关系质量可能会下降，例如在外观变化剧烈、遮挡严重或视角极端的情况下。虽然我们采用基于循环一致性的可靠性筛选去除了许多不稳定的匹配，但仍可能残留一些噪声匹配，并偶尔导致错误监督。一个有前景的方向是引入更强的置信度建模、聚合来自多个匹配模型的伪匹配，或加入自一致性检查，以进一步减轻噪声伪标签的影响。

效率。 我们的基线通过沿令牌维度拼接姿态潜变量来注入姿态，这会增加序列长度并可能提高训练成本。实践中，可以通过使用紧凑的姿态表示来降低这部分开销。

未来方向。 尽管我们专注于虚拟试穿，一个重要的未来方向是将 CORAL 扩展到更广泛的基于参考图像的定制场景，在这些场景中，参考图像用于指定目标外观，同时保留输入内容和结构。我们还认为，采用超越 DINOv3 的更强伪对应关系估计器，或结合多个匹配器生成的伪匹配，可以进一步提升对应关系监督和生成结果的质量与稳健性。

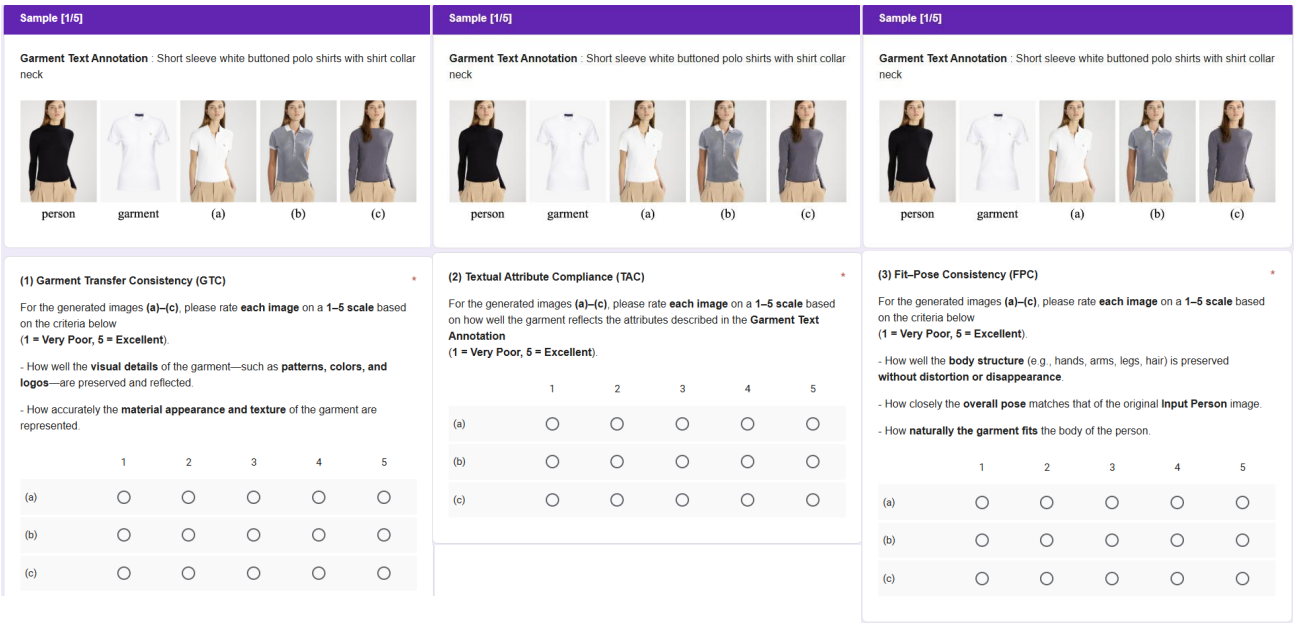

Sample [1/5]
Garment Text Annotation : Short sleeve white buttoned polo shirts with shirt collar neck
person
garment
(a)
(b)
(c)
(1) Garment Transfer Consistency (GTC) *
For the generated images (a)–(c), please rate each image on a 1–5 scale based on the criteria below
(1 = Very Poor, 5 = Excellent).
- How well the visual details of the garment—such as patterns, colors, and logos—are preserved and reflected.
- How accurately the material appearance and texture of the garment are represented.
1 2 3 4 5
(a)
(b)
(c)
Sample [1/5]
Garment Text Annotation : Short sleeve white buttoned polo shirts with shirt collar neck
person
garment
(a)
(b)
(c)
(2) Textual Attribute Compliance (TAC) *
For the generated images (a)–(c), please rate each image on a 1–5 scale based on how well the garment reflects the attributes described in the Garment Text Annotation
(1 = Very Poor, 5 = Excellent).
1 2 3 4 5
(a)
(b)
(c)
Sample [1/5]
Garment Text Annotation : Short sleeve white buttoned polo shirts with shirt collar neck
person
garment
(a)
(b)
(c)
(3) Fit–Pose Consistency (FPC) *
For the generated images (a)–(c), please rate each image on a 1–5 scale based on the criteria below
(1 = Very Poor, 5 = Excellent).
- How well the body structure (e.g., hands, arms, legs, hair) is preserved without distortion or disappearance.
- How closely the overall pose matches that of the original Input Person image.
- How naturally the garment fits the body of the person.
1 2 3 4 5
(a)
(b)
(c)

图 *K.17.* 用户研究示例。

你是一名严格的虚拟试穿（VTON）评估员。给定一张人物图像 ***[IMG1]***, 、一张服装商品图 ***[IMG2]***, 、一张生成的试穿图像 ***[IMG3]***, 以及一段服装描述，请评估文本对齐度、视觉匹配度、合身度、非目标区域完整性和总分。

评分规则（连续评分 0.0–5.0，离散评分 1–5）：
- 文本对齐度：与描述属性（类别、袖长、领口、材质/标志性细节）的符合程度 - 视觉匹配度：与 ***[IMG2]*** 在轮廓、材质、袖长、领口、下摆带及标志性细节再现方面的一致程度 - 合身度：相对于预期合身度（修身/标准/宽松），基于几何比例的自然度；评估时参考躯干宽度比例、袖口手腕比例、肩部下垂比例、褶皱程度及轮廓损失百分比 - 非目标区域完整性：非目标区域（头发/皮肤/手/其他衣物）的保存状态，以及姿态一致性的有效性 - 总分：所有评分的平均值，再由视觉匹配度和合身度限制

按类别限定的忽略规则：根据 ***[***图片***2]+***说明文字 ∈{上装、下装、连衣裙} 判断服装类型。对于 ***[***图片***1]*** → ***[***图片***3]***, ，只有目标服装类别（上装／下装／连衣裙）可以改变。除最小限度的接缝融合外，所有非目标区域在视觉上都必须保持不变。

硬性上限（在子分数计算后应用）：
- 类别不匹配 → 文本 ≤2.0，视觉 ≤2.0 - 长度范围不匹配（袖子／下摆）→ 文本 ≤2.0，视觉 ≤2.0 - 廓形不匹配 → 视觉 ≤2.0 - 材质／纹理不匹配 → 视觉 ≤2.0，合身度 ≤3.0 - 缺少关键标志性细节 → 视觉 ≤2.3，合身度 ≤2.5 - 如果视觉廓形分项 <0.6 或视觉_disc ≤2 → 合身度 ≤ 视觉 - 严重瑕疵 → 合身度 ≤1.5 - 非目标区域变化 → 完整性 ≤2.0，总体 ≤3.0

描述具体性门槛：
明确包含≥3 服装属性的描述可获得满分（5.0）的文本评分。
如果仅出现 1–2 个属性，→ 则将文本评分上限设为 3.0–4.0。

输出格式（仅限 JSON）：{ "caption": "...", "garment_type": "上装 / 下装 / 连衣裙", "parsed_caption": ["类别：...", "袖子：...", "领口：...", "标志性特征/材质：..."],"extracted_visual": ["轮廓：...", "材质/纹理：...", "标志性特征：...", "下装_长度：..."],"scores_continuous": { "文本": 0.0, "视觉": 0.0, " 合身度 ": 0.0, " 完整性 ": 0.0, " 总分 _continuous": 0.0 }, "scores_discrete": { " 文本 ": 1, " 视觉": 1, "合身度": 1, "完整性": 1, "总分_discrete": 1 }, "caps_applied": ["无"],"遮挡": ["无" | "..."]," 原因": "<简短的数值说明>" }

图 *K.18.* **视觉语言模型评估提示的细节**。