Garments2Look：兼具服装和配饰、面向高保真搭配级试穿的多参考数据集

Junyao Hu[1] Zhongwei Cheng[2] Waikeung Wong[1] Xingxing Zou[1 *]

[1]The Hong Kong Polytechnic University [2]Huhu AI Inc.

[*]Corresponding author

{junyao.hu,calvin.wong,xingxing.zou}@polyu.edu.hk, zcheng@huhu.ai

Abstract

虚拟试穿（*VTON*）在单件服装可视化方面已取得长足进展，但现实世界的时尚更关注完整搭配，包括多件服装、配饰、细粒度类别、层次叠穿以及多样化造型，这些仍然超出了当前 *VTON* 系统的能力范围。现有数据集在类别上受限，且缺乏造型多样性。我们提出 *Garments2Look*，这是首个用于搭配级 *VTON* 的大规模多模态数据集，包含覆盖 *40* 个主要类别和 *300+* 细粒度子类别的 *8* 万组多服饰到单造型配对数据。每一组数据包含一个由 *3–12* 张服饰参考图像（平均 *4.48* 张）组成的搭配、一张穿着该搭配的人体图像，以及对应的细粒度单品级与搭配级文本标注。为在真实感与多样性之间取得平衡，我们设计了一条合成数据管线：先通过启发式方法构造搭配列表，再生成对应的试穿结果，并对整个过程施加严格的自动过滤与人工质检，以保证数据质量。为探究任务难度，我们使用当前 *SOTA* 的 *VTON* 方法以及通用图像编辑模型构建基线。实验结果表明，现有方法在无缝试穿完整搭配以及推断正确叠穿关系和造型风格方面仍然存在困难，导致错位与伪影。我们的代码与数据已在 https://github.com/ArtmeScienceLab/Garments2Look 上开源。

1. 引言

虚拟试穿（*VTON*）在电商 *[13, 37]*、视效 *[1]*、时装设计 *[22, 39]*以及人机交互 *[31]*等领域展现出巨大应用潜力。目前，用户对 *VTON* 的期待已不再局限于单件服饰，而是进一步扩展到能够直观且精准地预览

图 1. 虚拟试穿数据集中不同数据格式的比较。我们的搭配级数据集由大规模真实图像构建与生成，每个样本都配有多样化的服饰与配饰，并显式包含搭配中叠穿层次与造型信息。

更加复杂的整体穿搭造型。已有工作开始探索多件单品 *[6, 24, 57, 58]*、叠穿 *[9, 38, 43, 54]*、细粒度类别 *[10, 44]*以及造型 *[20, 23, 58]*等方向，但尚未出现一种能对上述所有问题进行系统性全面建模的方法。

现有图像 *VTON* 数据集在结构设计上的不足，是上述局限性的直接原因之一。如图 *1* 所示，代表性数据集 *VITON-HD [5]* 和 *DressCode [27]* 虽然在图像质量与规模上都有所提升，但其最初仅针对单件服饰试穿任务进行设计。这类数据集忽视了配饰的作用，也缺乏关于穿搭技巧（如衬衫是否束入裤腰）以及服饰间协调关系（如搭配中的叠穿顺序）等文本标注。尽管 *OmniTry [12]* 在可穿戴类别的覆盖上更加丰富，其任务仍主要局限于单件单品；而 *M&M VTO [58]*、*BootComp [6]* 与 *DressCode-MR [57]* 虽然支持多参考输入，但在服饰类别多样性方面仍然受限。因此，亟需

表 1. 图像虚拟试穿数据集的比较。我们专注于搭配级虚拟试穿。我们的数据基于真实高分辨率的真实世界服饰图像和模特图像构建。Look 分辨率 = 目标图像的分辨率（高度×宽度）。≤/∼ = 像素数量小于/约等于产品数量。R/S = 真实/合成数据。叠穿顺序/造型技巧 = 数据集包含的服饰叠穿顺序/造型技巧标注。VLM = 由视觉语言模型生成的标注。公开性 = 数据集是否开源。

名称	时间	级别	#类别	#图像	#单品	#配对	#参考图像/配对	Look 分辨率	来源	叠穿顺序	造型技巧	公开性
VITON-HD [5]	CVPR 21	Item	1	27358	13679	13679	1	1024×768	R	✗	✗	Link
DressCode [27]	ECCV 22	Item	3	107584	53792	53792	1-2 (Avg 1.08)	1024×768	R	✗	✗	Link
M&M VTO [58]	CVPR 24	Outfit	2	-	-	18.8M	2	1024×512	R	✗	5 Types	✗
Street TryOn [10]	WACV 25	Item	13	14453	14453	0	1	512×320	R	✗	✗	Link
Shining Yourself [26]	CVPR 25	Item	4	-	-	64000	1	512×512	R	✗	✗	✗
BootComp [6]	CVPR 25	Outfit	7	-	-	54000	1-4	512×384	R+S	✗	✗	✗
Dresscode-MR [57]	arXiv 2508	Outfit	5	99260	71081	28179	1-4 (Avg 2.56)	1024×768	R+S	✗	✗	Link
Nano-Con OOTD [18]	arXiv 2508	Outfit	6	-	-	19000	2-6	≤1024×1024	R+S	✗	VLM	✗
OmniTry-Train [12]	arXiv 2508	Item	12	-	-	51195	1	≤1024×1024	R+S	✗	VLM	✗
OmniTry-Bench [12]	arXiv 2508	Item	12	825	360	6975	1	∼3000×3000	R	✗	VLM	Link
GO-MLVTON [54]	ICASSP 26	Outfit	2	10629	7091	3528	2	512×384	R	2 Layers	✗	Link
Garments2Look	CVPR 26	Outfit	40	429054	184367	80041	3-12 (Avg 4.48)	∼1024×1024	R+S	1-5 Layers	VLM	Link

一种新的虚拟试穿数据集，能够同时支持多样化的单品类别以及在整体搭配层面上的连贯组合。

与单品级 *VTON* 相比，搭配级 *VTON* 引入了新的技术挑战。服饰之间往往呈现复杂的叠穿与遮挡关系。例如，内外层的顺序可能发生变化（一件薄款针织开衫既可以作为最外层单品，也可以穿在大衣之内），而穿搭技巧也多种多样（可以常规穿着、披在肩上，或系在腰间）。对这类细节进行忠实刻画，对于结果的视觉质量与实际可用性都至关重要。

为此，我们提出 *Garments2Look*，为满足真实场景需求的更高级研究铺平道路。我们主要贡献如下：

- 我们构建了一个面向搭配级 *VTON* 的大规模多模态开源数据集，覆盖广泛的时尚单品类别，包含 *8* 万组高质量的单品–模特图像配对。
- 我们定义了一项新的 *VTON* 任务，利用丰富的结构化标注（文本描述、叠穿顺序、穿搭技巧），将多件参考单品及其匹配关系应用到给定模特上，从而生成灵活多样的整体穿搭试穿结果。
- 我们基于多种最新 *VTON* 方法开展了大规模实验，并进行了深入分析，以揭示其在搭配级试穿场景中的不足，并为未来方法改进提供有益参考。

2. 相关工作

2.1. 基于图像的虚拟试穿数据集

如表 *1* 所示，我们回顾了目前主流及近期发布的、面向图像虚拟试穿任务的数据集。*VITON-HD [5]* 将虚拟试穿图像的分辨率显著提升到较高水准，但其仅关注单一性别（女性）和单一服装类型（上衣）。*DressCode [27]* 和 *M&M VTO [58]* 认识到全身服饰的重要性，将服装类型扩展为三类（上装、下装与连体装）。在配饰试穿方面，*Shining Yourself [26]* 收集了覆盖四类配饰（手链、戒指、耳饰和项链）的成对图像。*BootComp [6]* 提出了一条基于换掉衣服（*Try-off*）的合成数据管线以及相应的数据筛选策略。*DressCode-MR [57]* 基于 *CatVTON [7]* 和 *FLUX [21]* 构建，引入了五类单品，新纳入了鞋履与包袋。*OmniTry [12]* 通过考虑更多可穿戴类型进一步拓展了 *VTON* 的应用场景，但其数据仍然是以单件单品成对图像为主。*Nano-Consistent-150K [18]* 中包含一个 *1.9* 万规模的 *VTON* 子集，但其忽略了姿态一致性问题。*GO-MLVTON [54]* 构建的数据集则主要针对双层上装这一特定难点场景。

为弥补现有数据集局限，我们提出的全新搭配级数据集具有多方面关键优势：它包含大量高质量的真实世界数据，支持搭配级参考输入，提供百万像素级分辨率的试穿结果，并附带关于单品与整体搭配描述、叠穿顺序和造型技巧的文本标注。我们的数据集在多个维度上全面超越以往的最先进方案，并将全部数据公开，以推动 *VTON* 领域的进一步研究。

2.2. 多参考图像数据合成

现有多参考图像生成工作，面向了不同类型的参考信息：主体、身份、风格、控制信号等。为构造成对数据，这类方法通常依赖开放词汇模型（如*Grounding DINO [25]* 和 *SAM2 [34]*）来获取实例的布局或分割结果作为参考输入 *[2, 28, 56]*。人们也提出了多种数据处

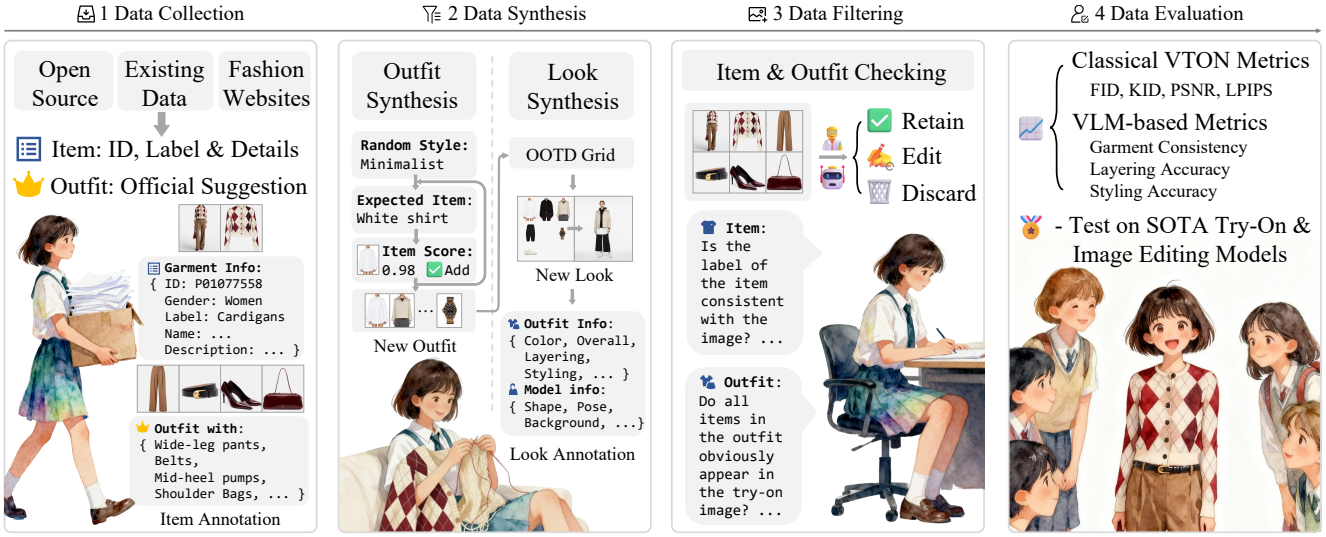


图 2. Garments2Look 构建流程概览。

理策略，以避免简单拷贝粘贴带来的伪影。*UNO* [49] 利用主体到图像模型合成参考图。*ComposeMe* [32] 构建了多图像身份数据集，使得身份、发型与服饰可以解耦控制，其中服饰被视为一个整体属性，而非若干独立单品。*USO* [48] 和 *DreamOmni2* [50] 则分别为目标图像生成风格参考图与内容参考图。*MultiRef* [4] 通过渲染引擎为同一物体生成不同类型的条件信号。近期的一系列工作，如 *Pico-Banana-400K* [33]、*MultiBanana* [29]、*MICo-150K* [46]、*UniRef-Image-Edit* [45] 与 *FireRed-Image-Edit* [41]，则采用 *Nano Banana* 系列 [8]、*Seedream* 系列 [36] 以及 *Qwen Image Edit* 系列 [47] 等先进模型作为核心合成引擎，通过严格筛选收集高质量的多参考图像。

与这些同期研究一致，我们同样采用“利用先进编辑模型进行数据合成与过滤”的通用范式，这一方法已逐渐成为生成高质量成对数据的业界共识。然而，与通用图像编辑研究不同，我们的工作聚焦于 *VTON* 场景，更加重视完整穿搭的一致性，以及诸如叠穿顺序与造型技巧等在通用合成框架中常被忽略的细节。

3. Garments2Look 数据集

如图 2 所示，*Garments2Look* 的构建大致分为四个步骤：*(1)* 数据采集：从不同来源获取真实世界的服饰单品以及对应的穿搭建议；*(2)* 数据合成：通过生成新的搭配列表和 *look* 图像，丰富数据集的内容与多样性；*(3)* 数据筛选：确保视觉一致性与数据质量，包括对服饰图像、搭配列表与 *look* 图像的标注；以及 *(4)* 数据评估：验证数据质量，为搭配级虚拟试穿任务设计新的评价指标，并对 *SOTA* 模型进行测试。

3.1. 数据采集

为了构建适用于搭配级虚拟试穿的数据集，我们需要成对的输入与输出图像：输入由若干单品图像构成（如多层上装、下装、配饰等），输出则是一张在人体模特上连贯展示完整穿搭效果的 *look* 图像。然而，完全匹配且高质量的成对数据往往十分稀缺、难以直接获取。因此，我们根据数据的完备性与可用性，将其分为如下几类：*(1)* **标准数据**：包含一组服饰单品图像及其对应的模特穿着图像，自然构成高质量的输入–输出配对。*(2)* **成对服饰图像**：已知搭配组合，但缺少对应的 *look* 图像。*(3)* **非成对服饰图像**：仅包含原始单品图像，缺乏已知搭配列表信息。*(4)* **仅 *look* 图像**：只有 *look* 图像，没有对应参考单品图像。

为了在数据质量与数量之间取得平衡，我们主要整合了第 *1* 类（*50.2%*）、第 *2* 类（*24.0%*）和第 *3* 类（*25.8%*）数据：一方面，我们利用高质量的标准数据来确保试穿效果的真实度（模型需要知道真实是什么样子）。另一方面，对于非成对图像，我们通过数据合成（见小节 *3.2*）提升数据量和多样性。

数据主要来自四个互补来源：*(1)* 穿搭兼容性学习的基础工作 [19, 59]。*(2)* 精选的开源时尚数据集，例如 *Maryland PolyVore* [15]，其提供了高质量且可靠的搭配数据。*(3)* 遵循严格授权与隐私合规要求采集的公开网络图像。*(4)* 由图像生成和理解模型合成的数据。

3.2. 数据合成

我们的数据合成主要聚焦于两个方面：*(1)* **搭配合成**：为利用非成对的服饰图像，我们采用类似检索增强生成的思路，以启发式方式构建搭配数据。*(2)* ***look* 图像合成**：为充分利用现有的非标准搭配数据以及新合成的搭配数据，我们使用图像生成模型合成试穿 *look* 结果，并结合图像理解模型生成细粒度标注。

3.2.1. 搭配合成

搭配合成流水线概览：我们首先从预先构建的时尚风格知识库中随机选择一个风格，作为生成的锚点。随后，大语言模型（*LLM*）基于选定风格生成潜在用户场景与偏好的详细描述。*LLM* 在此上下文及风格知识的共同约束下生成搭配列表。对于列表中的每件单品，我们在数据库中执行图像检索，找到最相关的候选项，并通过重加权采样策略最终选出合适单品。

步骤 *1* - 搭配风格知识库构建：为在确保不同风格边界清晰的同时覆盖尽可能广泛的时尚风格，我们采用“风格指导生成 + 时尚专家审阅”相结合的策略来构建搭配风格知识库。该知识库涵盖 *65* 种主流及次文化时尚风格（女性 *35* 种，男性 *30* 种），如 *Y2K* 风、清新风、学院风等。对于每一种风格，我们首先要求 *LLM* 严格遵循预设的大纲与 *Markdown* 结构生成技术化的风格指南。该指南精细地定义了该风格的偏好要点、禁忌事项、典型搭配示例与扩展造型规则。随后，由时尚专家对这些指南进行审阅与修订，最终形成精确的风格提示词与知识文件，用于约束后续生成。

步骤 *2* - 用户驱动的上下文生成：用户上下文是推动搭配合成的核心驱动力。为保证生成的搭配既具备实际应用价值又具有足够多样性，我们在随机选定风格和用户性别的基础上，引导 *LLM* 启发式构想生成多样化用户画像与具体着装场景。这些属性包括人口统计学特征（如年龄、职业、兴趣爱好）以及明确的场合信息（如晚宴、日常外出）。上下文描述涵盖四个关键维度：场合、配色方案、主题以及服饰类型，从而确保后续搭配列表生成在语境上合理、风格上契合。

步骤 *3* - 搭配列表生成：在获得详细的上下文与风格知识后，我们利用 *LLM* 生成搭配列表。模型被明确要求严格遵循用户需求与风格指南，输出由 *3–9* 件单品组成的完整搭配列表。为模拟真实生活中较为复杂的穿搭，我们特别要求模型关注叠穿效果，例如，允许组合中最多包含三层上装。生成的列表需遵循由上到下、由内到外、由服饰到配饰的顺序，以保证逻辑上的连贯性与层次感。

步骤 *4* - 单品检索：对于 *LLM* 生成的搭配列表中每一件单品的描述，我们在图像数据库中检索对应类别的前 *128* 个最相关候选，形成候选集。为缓解因平台数据偏差导致部分单品长期被忽略的问题，我们引入了一种重加权采样机制，以改进传统的相似度驱动选择策略。具体而言，我们根据候选项在历史搭配数据中的使用频次来调整其被采样概率：单品被选中的概率与其在既有搭配数据中出现的次数成反比。历史使用频率较低的单品因此会获得相对更高的被选中机会。该策略抑制对热门单品的反复选择，使整体语料中的单品分布更加均匀，提升原始数据利用效率。

关于指南与检索策略的更多细节见小节 *B.1*。

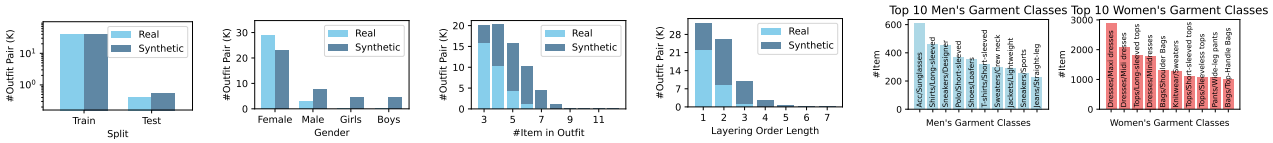
3.2.2. *look* 图像合成

为将非标准搭配数据以及合成的搭配数据转换为 *look* 图像，我们基于“每日穿搭”（*outfit-of-the-day, OOTD*）网格图来生成。我们将一个搭配列表中的全部单品图像排布成一个二维矩阵形式的网格图，将其作为图像生成模型的输入，主要使用 *Nano Banana*（*Gemini-2.5-Flash-Image*） *[8]*。与直接以多张分离图像作为输入相比，将 *OOTD* 网格图作为输入可以更好地保持各单品之间的一致性（见小节 *4.2*）。我们进一步考察了在 *OOTD* 网格图中改变单品位置、随机排布以及基于先验位置进行排布等设置对最终 *look* 图像质量的影响，结果表明差异并不显著。为增强 *look* 图像的创意性和视觉吸引力，我们通过提示工程显式注入叠穿顺序与造型技巧信息。在叠穿顺序方面，我们明确指定服饰的上下层顺序。在造型技巧方面，我们借鉴已有工作 *[3, 20, 23, 42, 58]* 中的五类典型穿搭手法，例如“将上衣束入下装”和“卷起袖口”。在部分设置中，我们显式指定期望的叠穿顺序与造型技巧；在另一些设置中，则允许模型自主选择合适的造型方式。

此外，当以 *look* 图像作为输入时，多模态大模型（*VLM*）能够给出更加丰富的文本描述，从而为数据集的文本模态提供更多信息（见小节 *B.2*）。

3.3. 数据筛选

为保证数据质量，我们在时尚领域专家的协助下，从单品图像、搭配列表以及服饰–*look* 成对图像三个方



(a) 按数据划分统计的搭配列表数量。 (b) 按性别类型统计的搭配列表数量。 (c) 按搭配长度统计的搭配列表数量。 (d) 按叠穿顺序长度统计的数量。 (e) 按性别划分的前 10 大服饰类别（按单品数量统计）。

(f) 前 8 种搭配类型的列表数量。 (g) 单品图像描述的词云。 (h) look 图像描述的词云。 (i) 造型描述的词云。 (j) 抽样 look 图像的美学评分分布。

Eval Type	Mean ± SD
Garments	4.74 ± 0.28
Layering	4.35 ± 0.62
Styling	4.58 ± 0.45

(k) 抽样搭配列表的人类主观评估结果。

图 3. Garments2Look 的数据分布统计。

面进行了数据筛选。对于单品图像，我们基于元数据以及现有工作 *[6, 10, 12, 26, 57]* 中广泛采用的常见类别，定义了 *40* 个主要服装与配饰类别，涵盖 *300* 余个细粒度子类别。对于搭配列表，尽管部分原始数据提供了预定义的搭配列表，我们的搭配合成流水线也能够生成新的列表，但这些列表中仍可能存在逻辑冗余（如一个人同时穿两条连衣裙在现实中并不常见）。为此，我们基于时尚领域知识设计了一套规则驱动的搭配合理性验证机制。当某个搭配违反约束时，我们通过剔除冗余单品来提取合理的子集。

对于服饰–*look* 图像配对，我们重点保留两种类型的图像：一是清晰展示整套搭配的完整单品图；二是从正面视角完整展示模特整套搭配的 *look* 图。我们使用 *Gemini-2.5-Flash [8]* 筛选合适图像，并结合 *DWPose [51]* 等工具对 *look* 图像进行姿态相关分类。

为进一步保障合成数据质量，我们在筛选流程中招募了 *10* 名时尚相关专业学生和 *3* 位专家参与审阅。如果某套搭配中任一单品与 *look* 图不一致，相关 *look* 图将被重新生成或剔除。最终，仅约 *∼40%* 的合成 *look* 图像被保留，每张都通过专家审阅。

关于服饰类别与专家评审的更多细节见小节 *B.3*。

3.4. 数据评估

统计分析： *Garments2Look* 共包含 *80K* 个搭配级配对样本，图 *3* 给出了数据集的基本统计信息。在最终数据集中，真实数据与合成数据的比例大致维持在 *∼1:1*（图 *3a*）。我们收集的数据覆盖了多种性别（图 *3b*）、每套搭配中不同数量的单品图像（图 *3c*）、不同长度的叠穿顺序（图 *3d*），以及广泛的服饰类别（图 *3e*）和搭配组合模式（图 *3f*）。我们同样重视文本标注，以便支持未来的多模态研究，标注内容包括单品图像描述、*look* 图像描述以及造型技巧描述。在图 *3g to 3i* 中，三幅词云分别展示了数据集中三类文本标注的核心维度。单品描述强调内在属性与材质纹理，高频词如 *“leather”“elegant”“sophisticated”* 表明这些标注主要用于刻画材料特性、风格取向与设计细节；*look* 描述则聚焦整体视觉效果与协调性，诸如 *“ensemble”“relaxed”“chic”* 等关键词突显了模特整体造型的气质；造型描述重点关照具体的穿着状态，高频出现的动词（如 *“tucked”“unbuttoned”“rolled”*）清晰反映出对服饰与人体之间物理交互的关注。整体来看，这些多维度文本线索为实现高保真、精准的虚拟试穿提供了全面指导。

我们还使用 *aesthetic-predictor-v2-5 [11]* 评估 *look* 图像的美学质量。以往工作 *[35]* 常采用 *5.0* 的绝对美学阈值，但对于以人物形象为中心的时尚图像，这一固定阈值可能并非最优选择。因此，我们针对每个数据子集，以经验均值作为阈值剔除美学分数较低的图像，从而移除明显低质量的样本。剩余候选随后再进行人工筛选。在图 *3j* 中，我们分别从 *Garments2Look* 的两个子集中各抽取 *1* 万样本进行最终评估。在一致性与准确性方面，如图 *3k* 所示，我们邀请 *13* 位时尚领域专家，基于 *Likert* 五点评分量表（*1–5* 分）对训练集中的 *100* 个随机样本进行评估。分数越高，表示一致性或准确性越好。

评估协议： 在对模型性能进行自动评估时，我们采用经典的 *VTON* 指标：*FID [30]*、*KID [40]*、*SSIM [40]* 和 *LPIPS [55]*。针对我们的搭配级 *VTON* 任务，我

表 2. 针对多参考虚拟试穿任务，在 DressCode-MR [57] 上对 Nano Banana (Gemini-2.5-Flash-Image) [8] 的性能评估。我们通过两种参考策略进行评估。

方法	成对				非成对	
	FID↓	KID↓	SSIM↑	LPIPS↓	FID↓	KID↓
CatVTON [7]	16.131	6.980	0.856	0.106	18.339	7.458
IP-Adapter [52]	14.459	4.144	0.861	0.089	24.139	10.783
FitDiT [17]	14.722	5.471	0.850	0.122	15.956	5.645
FastFit [57]	**9.311**	**1.512**	**0.859**	**0.079**	**12.059**	**2.123**
Nano-Banana [8] (N Ref)	15.220	6.703	0.801	0.199	15.369	6.916
Nano-Banana [8] (2 Ref)	15.980	7.494	0.797	0.231	16.393	7.535

们使用 *Gemini-3-Flash* 作为 *VLM* 评审者，从三项度量对结果进行评估，并报告二分类准确率。服饰一致性在单品维度上进行评估：遮挡导致的局部不可见是可以接受的，但若出现结构性不匹配（如口袋形状或位置错误）则被视为不一致。叠穿准确性通过仅验证相邻层之间的内外关系，将评估复杂度优化为线性；造型准确性同样在单品层面进行评估。所有评估结果需同时输出分类结论与相应理由，以保证可解释性。

关于数据集的更多细节见小节 *A*。

4. 实验

我们设计了一系列实验，从两个方面验证所提出 *Garments2Look* 数据集的价值：*(1)* **数据集难度**：现有方法在包含配饰、叠穿顺序和造型技巧的搭配级虚拟试穿任务上表现欠佳；*(2)* **可操作洞见**：超越纯视觉的结构化标注（叠穿顺序、造型技巧以及更丰富的文本描述）可以为生成过程提供有效指导。为此，我们首先在一个已有的多参考虚拟试穿数据集上，对强大的图像编辑模型进行基准评测，以在相对简单的设置下标定其性能上界；随后，我们在 *Garments2Look* 的测试集上，对多种虚拟试穿模型和通用图像编辑模型进行定量与定性评估，以暴露当前方法的瓶颈，并分析结构化标注如何提供帮助。

4.1. 编辑模型能胜任虚拟试穿吗？

随近期通用图像编辑模型的迅速发展，其性能受到广泛关注。虚拟试穿作为具有较高实际应用价值的任务，常被商业厂商用来展示编辑能力；在许多通用编辑产品的宣传材料中，虚拟试穿场景都占据了显著位置。这引出了一个看似矛盾的问题：如果商业编辑模型已经能够有效处理多单品试穿，那么我们提出的数据集与任务是否还有必要？反之，如果这些商业模型尚不能真正解决多单品试穿问题，我们又该如何方便、可扩展地构建大规模多单品数据集？为了解答这一问题，我们设计了第一组研究：先进图像编辑模型能否在现有数据集上合成可用的多参考试穿数据？其目标是在相对简单的多参考基准上建立可行基线，解释当前编辑模型为何仍不足以应对我们更具挑战性的设置，并凸显它们在可扩展数据合成方面的潜力。

具体而言，我们在最新的多参考虚拟试穿数据集 *DressCode-MR [57]* 上评估了当前 *SOTA* 图像编辑模型 *Nano Banana*（*Gemini-2.5-Flash-Image*）*[8]*。如表 *2* 所示，*Nano Banana* 仍明显落后于最优的专用虚拟试穿方法。乍看之下，这似乎与小节 *3.2.2* 中的描述相矛盾：我们正是采用 *Nano Banana* 进行 *look* 图像合成。我们的模型选择是通过专家交叉验证支持的：三位时尚专家在 *500* 个样本上，对两种虚拟试穿模型（*OmniTry [12]* 和 *FastFit [57]*）以及三种编辑模型（*Nano Banana [8]*、*Seedream 4.0 [36]* 和 *Qwen-Image-Edit-2509 [47]*）进行了比较；*Nano Banana* 在 *66%* 的案例中被评为最佳。进一步分析表明，存在两个关键限制：*Nano Banana* 并不原生支持图像修补（*inpainting*），且缺乏对人体骨架姿态的精确显式控制。因此，尽管其整体视觉质量令人信服，但由于细粒度可控性有限，其定量指标和细节保真度仍不及专用虚拟试穿模型。即便我们将骨架图像作为参考之一输入，图像编辑模型依然难以严格保持与原图相同的姿态（见小节 *C.2*），而标准虚拟试穿评估指标对姿态一致性高度敏感，姿态保持也是虚拟试穿长期以来的基本要求。

综上，由于缺乏图像修补能力和显式姿态控制，*Nano Banana* 在保真度和结构一致性方面仍落后于专用虚拟试穿系统。换言之，要让编辑模型稳健地解决虚拟试穿任务，必须在固定目标身份与条件设定下进一步提升生成一致性。另一方面，对于合成数据生成而言，在不依赖骨架或修补掩膜的前提下，直接通过文本对整幅图像进行编辑，仍有可能产生高质量多单品试穿结果。鉴于高质量多参考数据采集的成本，以及我们对合成结果进行的专家审核，上述因素共同凸显构建此类数据的挑战性，也进一步说明本工作价值。

4.2. *Garments2Look* 有多具挑战性？

作为一个新提出的数据集以及虚拟试穿家族中的进阶任务，*Garments2Look* 既具有明确的实际应用价

图 4. 在 Garments2Look 测试集 4 个具有代表性的真实样本上，3 个 SOTA 虚拟试穿模型与 4 个通用图像编辑模型的结果对比。QIE-2509 = Qwen-Image-Edit-2509，NB = Nano Banana，N Ref 表示使用模特图像和多张单品图像作为输入，2 Ref 表示使用模特图像和一张 OOTD 图像作为输入。黄色框标出与 look 图（GT）的差异，黑色箭头指示明显伪影边界。第 1 行：4 件单品，1 层，无配饰；第 2 行：5 件单品，2 层，无配饰；第 3 行：8 件单品，3 层，2 件配饰；第 4 行：9 件单品，3 层，3 件配饰。

表 3. 针对搭配级虚拟试穿任务，在 Garments2Look 测试集上对各种模型的性能评估。我们报告了经典 VTON 指标和由 VLM 评估的准确性指标。

模型	FID↓	KID↓	SSIM↑	LPIPS↓	服饰↑	叠穿↑	造型↑
虚拟试穿模型							
IP-Adapter [52]	6.5511	6.7624	0.7960	0.1472	0.4950	0.6036	0.5693
BootComp [6]	8.6301	8.9129	0.6110	0.3364	0.5369	0.3127	0.3547
OmniTry [12]	6.5559	10.0715	0.7244	0.2295	0.4612	0.1674	0.2609
FastFit [57]	3.5919	4.5827	0.8550	0.0956	0.6237	0.1305	0.3400
图像编辑模型							
GPT-4o [16] (N Ref)	4.0478	2.8334	0.6648	0.2565	0.8355	0.7901	0.6401
GPT-4o [16] (2 Ref)	2.1496	1.4154	0.7577	0.1558	0.8921	0.8487	0.6943
Seedream 4.0 [36] (N Ref)	4.1436	6.2291	0.7134	0.1831	0.7709	0.8312	0.6452
Seedream 4.0 [36] (2 Ref)	3.7117	6.0362	0.7358	0.1683	0.7712	0.8808	0.6808
QIE-2509 [47] (N Ref)	33.1751	83.7688	0.5093	0.5360	0.5144	0.4716	0.5120
QIE-2509 [47] (2 Ref)	1.7562	1.0117	0.8111	0.0910	0.7940	0.8168	0.6729
NB (N Ref) [8]	1.1619	**0.1628**	0.8523	0.0663	0.8045	0.9250	**0.7482**
NB (2 Ref) [8]	**1.0412**	0.2523	**0.8583**	**0.0619**	0.9253	0.8847	0.7386
NBP (N Ref)	1.3182	0.4002	0.8167	0.0903	**0.9836**	**0.9363**	0.7356
NBP (2 Ref)	1.3462	0.5041	0.8202	0.0863	0.9707	0.9317	0.7213

值，又为现有方法引入了非平凡的挑战。为判定该任务究竟是过于简单、不足以支撑深入研究，还是过于

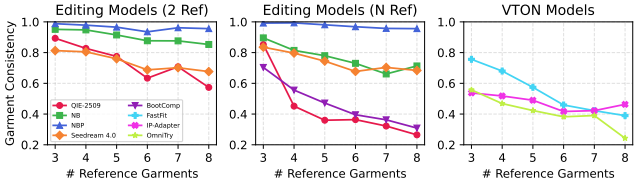


图 5. 不同模型类型下服饰一致性随参考图像数量变化。

困难、缺乏可行解，我们在 *Garments2Look* 上对最新方法进行了定量（表 *3 and* 图 *5*）和定性（图 *4*）评估。我们的分析聚焦于以下四个问题：

***Q1.* 同时可以穿多少件单品？** 受限于固定标签集合的虚拟试穿模型在处理大量单品时表现不佳。迭代式的 *OmniTry [12]* 范式更倾向于替换而非叠加更多服饰层，通常只保留一件上装。相比之下，通用编辑模型更加灵活：使用自由形式的多图像输入或将 *OOTD* 图作为统一参考，有助于叠加更多单品。诸如 *Nano Banana [8]* 等编辑模型，结合多图像输入与 *OOTD* 图，在搭配复杂度增加时仍能较好地提升叠穿深度。当单

品数量超过 4 件时，未在此类基数上充分训练过的虚拟试穿模型往往会漏掉部分单品，或仅渲染最外层而忽略内层服饰。相比之下，编辑模型在可变长度搭配列表上表现出更好的鲁棒性。在虚拟试穿范式内部，支持可变长度输入，或通过 *OOTD* 风格的组合方式构建输入，似乎是适应多样多单品搭配的一条可行且有前景的路径。这些发现表明，训练阶段所接触的输入模态，从根本上限制模型在搭配组合能力上的上限。

***Q2. look* 图像与参考的匹配程度如何？** 图 4 表明，随着参考数量的增加，退化现象在各模型中普遍存在：（1）形状扭曲：几乎所有示例中，配饰（如包袋、耳饰）的形状都出现了明显偏差；（2）纹理变化：在第 1 行和第 2 行中，衣物上的文字内容与风格发生了改变（“*PRADA*” 与“*LOWEWE*”），在第 3 行中，毛衣的斜条纹纹理无法被一致保留（*GPT-4o* 和 *NB (2 Ref)*）；（3）颜色偏差：在第 4 行中，*BootComp* 生成的大衣颜色以及 *NB (N Ref)* 生成的针织衫颜色都与参考图像存在差异；（4）单品融合：在第 3 行中，*BootComp* 将本应彼此独立的两件上装错误地融合在一起。此外，在同一样本上，2 参考策略的结果往往优于 *N* 参考策略（见图 4 *and* 5 *and* 表 3 中 *2 Ref* 和*N Ref* 的对比）。将整套搭配视为一个整体参考，能够携带更强的上下文共现与隐式关系，而当前模型对这种整体参考的利用效率明显高于对多个分散参考的利用。

***Q3.* 整体试穿效果有多好？** 这一问题关注的是整体视觉印象，评估全局协调性以及对特殊穿搭技巧的处理能力。我们观察到明显的修补伪影，例如 *FastFit [57]* 在图 4 中呈现的人体–背景过渡不自然，这很可能源于其训练中对白色纯背景试穿图的偏好。叠穿服饰仍是难点：虽然编辑模型可以解析包含叠穿信息的提示词，但当参考单品数量增多时，其控制能力显著下降。与单件服饰情形（第 1 行）相比，叠穿场景（第 2–3 行）的结果一致性更弱，包括第 2 行中白色内搭 *T* 恤文字缺失，第 3 行中西装纽扣数量错误，第 4 行中内衬条纹密度畸变（*QIE-2509* 与 *NB (N Ref)*），以及第 4 行中外套衣长不准确（*Seedream 4.0*）。造型控制同样不理想：在第 4 行中，目标 *look* 的中间层上衣未束入下装且仅扣上一两颗纽扣；即便给出明确提示，大多数模型生成的仍是整齐、束入下装的穿法，表明当前方法在非标准造型上的细粒度控制能力不足。

***Q4.* 为何表 3 中 *Garments2Look* 上的结果与表 2 中 *DressCode-MR [57]* 上的结果表现相反？** 我们发现，在 *Garments2Look* 上，大多数编辑模型优于虚拟试穿模型。这主要可以归因于两个方面：*(1) Garments2Look* 覆盖的类别更加多样。编辑模型不受单品类别的严格限制，因而能够容纳更多类型；相比之下，虚拟试穿模型通常只支持有限的类别，导致诸如围巾、手套、胸针等单品无法被测试——这些单品在 *DressCode-MR [57]* 中缺失，而在 *Garments2Look* 中被纳入。*(2) Garments2Look* 中单品数量更多、叠穿与造型关系更加复杂。编辑模型在此类设置下更具灵活性，而 *SOTA* 虚拟试穿模型则主要聚焦于单层试穿，难以同时合成多件服饰，或仅支持简单的指令控制。

总之，*Garments2Look* 提出了一个既具挑战性又可操作的任务与数据集：它足以对当前方法构成严苛压力，同时又为取得实质性研究进展提供了肥沃土壤。

4.3. 它带来了哪些新洞见？

现有 *SOTA* 虚拟试穿方法在 *Garments2Look* 上表现不佳，只有少数方法能够完整跑通整条试穿流水线。我们将其归结为三个主要原因：（*1*）大规模时尚库存带来的范式转变；（2）叠穿服饰之间的层次依赖；（3）对细粒度设计细节的高度敏感。这些问题揭示了纯视觉框架的根本局限。为缓解这些限制，*Garments2Look* 引入了丰富的结构化文本标注，这在传统虚拟试穿数据集中几乎是缺失的维度。不同于 *BootComp [6]* 和 *OmniTry [12]* 等方法仅将文本视为独立信号，我们的数据集在视觉输入与复杂搭配语义之间搭建起桥梁。如小节 3.2 所述，我们提供了与图像高度对齐的高保真标注，覆盖了单品级与搭配级描述，包括类别、单品描述、搭配描述、叠穿逻辑、造型技巧、模特属性等信息。这种结构化设计契合了多模态学习演进方向 *[14, 20, 53]*，为复杂搭配合成提供了必要语义引导。

我们的实验验证了文本引导在复杂虚拟试穿场景中的有效性。在表 4 中，我们进行了消融实验以评估文本粒度对结果的影响。当仅使用单品类别作为基础引导（第 1 行）时，性能提升有限；在此基础上加入搭配级信息（第 2 行），*FID* 与 *KID* 指标均有可观改善。随后，我们逐步叠加更细粒度的属性（第 3–5 行），模型在所有指标上都取得了显著跃升。这一趋势表明，*Garments2Look* 中的文本模态为解决搭配级虚拟试穿任务提供了至关重要且与视觉特征高度协同的指导信息。

表 4. 针对搭配级虚拟试穿任务，在 Garments2Look 测试集上对 Nano Banana 生成的样本，进行文本提示的细节级别研究。

文本输入类型	FID↓	KID↓	SSIM↑	LPIPS↓
单品类别	23.272	1.271	0.817	0.141
单品类别 + 整体搭配概述	22.135	0.752	0.810	0.155
单品类别 + 叠穿和造型	21.831	0.733	0.814	0.148
单品类别 + 身材和姿态	21.783	0.750	0.823	0.133
单品类别 + 身材和姿态 + 叠穿和造型	**21.545**	**0.641**	**0.825**	**0.131**

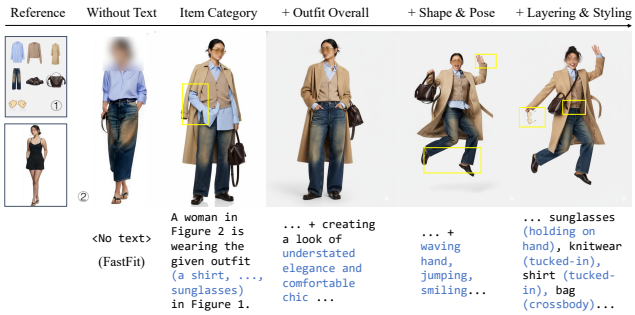


图 6. 文本模态贡献的可视化示例。

图 6 中的代表性示例进一步印证了这一观察。借助我们完整而独特的文本标注，模型能够生成更具时尚感且质量更高的服饰图像。图中最后两个示例还展示了在姿态与风格上的可控性提升，凸显了 *Garments2Look* 所支持的灵活生成潜力。

这些结果共同表明，超越纯视觉的特征维度对于将虚拟试穿扩展到现实多层次衣橱和对属性高度敏感的时尚应用场景至关重要。

5. 总结

本文针对现有数据集中搭配级虚拟试穿（*Outfit-level VTON*）的关键空白展开研究：这些数据集普遍缺乏对多件服饰叠穿、配饰融合以及细粒度叠穿与造型标注的支持。我们提出了 *Garments2Look*，这是首个面向搭配级虚拟试穿的大规模多模态数据集，包含 *80* 万组具有结构化标注的高保真配对数据。对多种 *SOTA* 方法的评估表明，它们在生成搭配级结果时仍存在显著不足，这一现象凸显了我们设定所带来的新挑战，也提供了指向该方向潜力的若干启示。未来工作将聚焦于设计更契合任务特点的评价指标，并探索融合视觉与文本线索的模型，实现具备精细细节控制的端到端搭配级虚拟试穿。

致谢

本工作得到了香港特别行政区研究资助局（项目编号：*PolyU/RGC Project 25211424*）和香港理工大学大学启动基金（项目编号：*P0047675*）的支持。

References

[1] *Peter Bug and Melina Bernd. The future of fashion films in augmented reality and virtual reality. In* Fashion and film: moving images and consumer behavior. *Springer, 2019. 1*

[2] *Bowen Chen, Brynn zhao, Haomiao Sun, Li Chen, Xu Wang, Daniel Kang Du, and Xinglong Wu. XVerse: Consistent multi-subject control of identity and semantic attributes via dit modulation. In* NeurIPS, *2025. 2*

[3] *Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In* ICCV, *2023. 4*

[4] *Ruoxi Chen, Dongping Chen, Siyuan Wu, Sinan Wang, Shiyun Lang, Petr Sushko, Gaoyang Jiang, Yao Wan, and Ranjay Krishna. Multiref: Controllable image generation with multiple visual references. In* ACM Multimedia 2025 Dataset Track, *2025. 3*

[5] *Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In* CVPR, *2021. 1, 2*

[6] *Yisol Choi, Sangkyung Kwak, Sihyun Yu, Hyungwon Choi, and Jinwoo Shin. Controllable human image generation with personalized multi-garments. In* CVPR, *2025. 1, 2, 5, 7, 8*

[7] *Zheng Chong, Xiao Dong, Haoxiang Li, shiyue Zhang, Wenqing Zhang, Hanqing Zhao, xujie zhang, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In* ICLR, *2025. 2, 6*

[8] *Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.* arXiv, *2025. 3, 4, 5, 6, 7, 15*

[9] *Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In* ICCV, *2021. 1*

[10] *Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In* WACV, *2025. 1, 2, 5*

[11] *discus0434. Aesthetic predictor v2.5.* https://github.com/discus0434/aesthetic-predictor-v2-5, *2024. 5*

[12] *Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks.* arXiv, *2025. 1, 2, 5, 6, 7, 8*

[13] *Adelya Gabriel, Alina Dhifan Ajriya, Cut Zahra Nabila Fahmi, and Putu Wuri Handayani. The influence of augmented reality on e-commerce: A case study on fashion and beauty products.* Cogent Business & Management, *2023. 1*

[14] *Jingyi Guo, Pengfei Duan, Chenghu Du, and Shengwu Xiong. Enhancing virtual try-on with text-image fusion guidance. In* ICIC, *2025. 8*

[15] *Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In* ACM Multimedia, *2017. 3*

[16] *Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card.* arXiv, *2024. 7*

[17] *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. 6*

[18] *Ye Junyan, Jiang Dongzhi, Wang Zihao, Zhu Leqi, Hu Zhenghao, Huang Zilong, He Jun, Yan Zhiyuan, Yu Jinghua, Li Hongsheng, He Conghui, and Li Weijia. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025. 2*

[19] *Pang Kaicheng, Zou Xingxing, and Wai Keung Wong. Modeling fashion compatibility with explanation by using bidirectional lstm. In* CVPRW, *2021. 3*

[20] *Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. In* CVPR, *2025. 1, 4, 8*

[21] *Black Forest Labs. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2*

[22] *Agnė Lagė and Kristina Ancutienė. Virtual try-on technologies in the clothing industry: basic block pattern modification.* International Journal of Clothing Science and Technology, *2019.* 1

[23] *Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In* WACV, *2024.* 1, 4

[24] *Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario.* NeurIPS, *2024.* 1

[25] *Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.* arXiv, *2023.* 2

[26] *Yingmao Miao, Zhanpeng Huang, Rui Han, Zibin Wang, Chenhao Lin, and Chao Shen. Shining yourself: High-fidelity ornaments virtual try-on with diffusion model. In* CVPR, *2025.* 2, 5

[27] *Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. In* ECCV, *2022.* 1, 2

[28] *Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In* SIGGRAPH Asia, *2024.* 2

[29] *Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In* CVPR, *2026.* 3

[30] *Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In* CVPR, *2022.* 5

[31] *M Prakash, Nithes Arunkumar, et al. Gesture-driven innovation: Exploring the intersection of human-computer interaction and virtual fashion try-on systems. In* ICNWC, *2024.* 1

[32] *Guocheng Gordon Qian, Daniil Ostashev, Egor Nemchinov, Avihay Assouline, Sergey Tulyakov, Kuan-Chieh Jackson Wang, and Kfir Aberman. Composeme: Attribute-specific image prompts for controllable human image generation.* arXiv, *2025.* 3

[33] *Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing, 2025.* 3

[34] *Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos.* arXiv, *2024.* 2

[35] *Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models.* NeurIPS, *2022.* 5

[36] *Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation.* arXiv, *2025.* 3, 6, 7, 15

[37] *Khawla Sekri, Olfa Bouzaabia, Haifa Rzem, and David Juárez-Varón. Effects of virtual try-on technology as an innovative e-commerce tool on consumers' online purchase intentions.* EJIM, *2025.* 1

[38] *Yidi Shao, Chen Change Loy, and Bo Dai. Towards multi-layered 3d garments animation. In* ICCV, *2023.* 1

[39] *Wenda Shi, Waikeung Wong, and Xingxing Zou. Generative ai in fashion: Overview.* ACM TIST, *2025.* 1

[40] *JD Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In* ICLR, *2018.* 5

[41] *Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 techinical report.* arXiv, *2026.* 3

[42] *Minh Tran, Johnmark Clements, Annie Prasanna Manoharan, Tri Nguyen, and Ngan Le. Dualfit: A two-stage virtual try-on via warping and synthesis. In* ICCV, *2025.* 4

[43] *Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Mgt: Extending virtual try-off to multi-garment scenarios. In* ICCV, *2025.* 1

[44] *Siqi Wan, Jingwen Chen, Qi Cai, Yingwei Pan, Ting Yao, and Tao Mei. VTON-VLLM: Aligning virtual try-on models with human preferences. In* NeurIPS, *2025.* 1

[45] *Hongyang Wei, Bin Wen, Yancheng Long, Yankai Yang, Yuhang Hu, Tianke Zhang, Wei Chen, Haonan Fan, Kaiyu Jiang, Jiankang Chen, et al. Uniref-image-edit: Towards scalable and consistent multi-reference image editing.* arXiv, *2026.* 3

[46] *Xinyu Wei, Kangrui Cen, Hongyang Wei, Zhen Guo, Bairui Li, Zeqing Wang, Jinrui Zhang, and Lei Zhang. Mico-150k: A comprehensive dataset advancing multi-image composition. In* CVPR, *2026.* 3

[47] *Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report.* arXiv, *2025.* 3, 6, 7, 15

[48] *Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning.* arXiv, *2025.* 3

[49] *Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In* ICCV, *2025.* 3

[50] *Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation.* arXiv, *2025.* 3

[51] *Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In* ICCV, *2023.* 5

[52] *Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.* arXiv, *2023.* 6, 7

[53] *Xiaomin Yu, Yi Xin, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Xiaoxing Hu, Xinlei Yu, Ziyue Qiao, Hao Tang, Xue Yang, et al. Modality gap-driven subspace alignment training paradigm for multimodal large language models.* arXiv, *2026.* 8

[54] *Yang Yu, Yunze Deng, Yige Zhang, Yanjie Xiao, Youkun Ou, Wenhao Hu, Mingchao Li, Bin Feng, Wenyu Liu, Dandan Zheng, et al. Go-mlvton: Garment occlusion-aware multi-layer virtual try-on with diffusion models. In* ICASSP, *2026.* 1, 2

[55] *Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In* CVPR, *2018.* 5

[56] *Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation.* arXiv, *2024.* 2

[57] *Chong Zheng, Lei Yanwei, Zhang Shiyue, He Zhuandi, Wang Zhen, Zhang Xujie, Dong Xiao, Wu Yiling, Jiang Dongmei, and Liang Xiaodan. Fastfit: Accelerating multi-reference virtual try-on via cacheable diffusion models, 2025.* 1, 2, 5, 6, 7, 8

[58] *Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In* CVPR, *2024.* 1, 2, 4

[59] *Xingxing Zou, Kaicheng Pang, Wen Zhang, and Waikeung Wong. How good is aesthetic ability of a fashion model? In* CVPR, *2022.* 3

Garments2Look：兼具服装和配饰、面向高保真搭配级试穿的多参考数据集

Supplementary Material

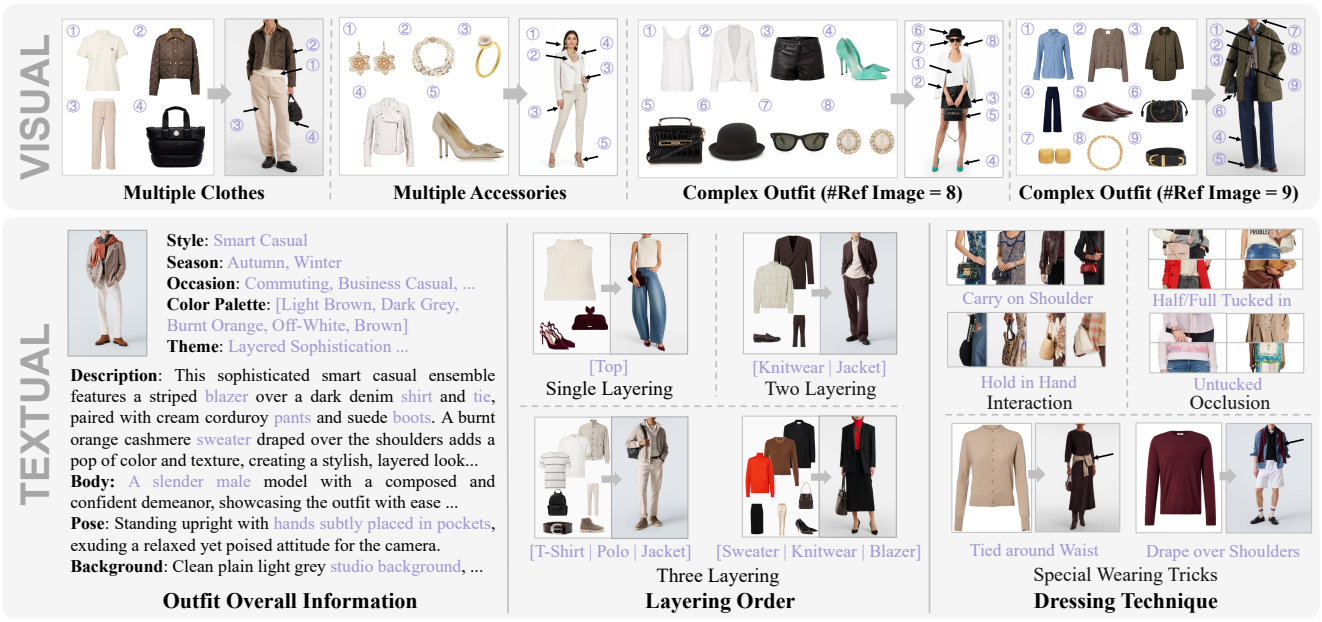


图 S1. Garments2Look 中更多示例数据。

A. 数据集详情

A.1. 更多示例数据

我们在图 *S1* 中展示了所提出数据集的更多示例数据。该数据集提供了高质量、高精度且高度多样化的复杂完整穿搭组合样本。

在视觉层面上，我们从真实应用场景出发，覆盖了比以往数据集更复杂的虚拟试穿场景。这包括从基础单品到丰富配饰、从单层穿搭到多层叠穿、从单一风格到多种造型方式等多种情形。以一套由 *T* 恤、短裤、鞋子和包袋构成的简单组合为例，在真实场景中往往还会搭配耳饰、手链、帽子或眼镜等配饰作为造型焦点；也可以在其外叠穿一件薄款针织开衫来增强层次感；还可以将围巾在胸前或腰间打结系绑。除了在维度上尽量贴近实际穿搭需求之外，我们同样十分重视数据质量：从单品图像与搭配数据的收集，到搭配组合的生成，再到试穿结果的合成，全流程都由时尚专家参与审阅把关。

此外，考虑到文本模态的重要性——以及某些信息（如穿搭技巧）更适合通过语言来表达——与以往试穿数据集不同，我们为所有样本提供了丰富而细致的文本描述。这些描述涵盖单品类别、细粒度属性（风格、季节、场合、颜色、主题等）、多单品整体搭配的整体说明、各单品的穿着顺序与方式，以及直接影响试穿效果的因素，如模特的身材体型与姿态等。

我们期待该数据集能够有力推动虚拟试穿领域的深入研究，并支撑更广泛的应用场景。

A.2. 数据划分

如表 *S1* 所示，我们构建了一个约 *∼1* 千样本规模的测试集用于评估。该测试集通过对完整的 *8* 万样本进行分层抽样得到，以确保具有较高的代表性以及数据来源的真实性。其余所有样本均被划分为训练集。

表 S1. Garments2Look 数据集划分。

类别	训练集	测试集	合计
真实	39819	402	40222
合成	39278	541	39819
合计	79097	944	80041

B. 数据处理

B.1. 搭配合成

风格指南：首先，我们基于网络检索得到的流行风格分类来确定风格类别。随后，由 *Gemini-2.5-Flash* 生成初稿文本。最后，三位资深时尚专家对内容进行修订，以确保其专业性与准确性。一个关于极简风格（*Minimalist Style*）的风格指南示例见表 *S4*。

主要时尚风格：美式复古（*American Vintage*），中性风（*Androgynous*），海滩风（*Beach*），波西米亚风（*Bohemian*），商务休闲（*Business Casual*），休闲风（*Casual*），经典风（*Classic*），舒适风（*Comfort*），新田园风（*Cottagecore*），乡村风（*Country*），牛仔风（*Cowboy*），混搭风（*Eclectic*），优雅风（*Elegant*），正式风（*Formal*），法式风（*French chic*），甜美风（*Girly*），华丽风（*Glamorous*），哥特风（*Gothic*），嘻哈风（*Hip-Hop*），港风（*Hong Kong Vintage*），可爱风（*Kawaii*），分层风格（*Lagenlook*），军事风（*Military*），极简风（*Minimalist*），森女风（*Mori Girl*），机车风（*Moto*），老钱风（*Old Money*），趣味风（*Playful*），学院风（*Preppy*），朋克风（*Punk*），古怪风（*Quirky*），浪漫风（*Romantic*），热辣风（*Spicy*），运动风（*Sporty*），街头风（*Street*），工装风（*Workwear*），*Y2K* 风（*Y2K*）。

检索采样：我们提出了一种基于逆频率重加权的检索采样机制。首先，从检索结果中选出相似度最高的 N 个候选单品 $\{x_1, \ldots, x_N\}$。随后，我们统计每个单品 x_i 在当前搭配数据集中出现的历史次数，记为 c_i，并据此定义其逆频率权重 $w_i = (c_{\max} - c_i) + 1$，其中 $c_{\max} = \max_{j=1}^{N}(c_j)$。在采样过程中，单品 x_i 被选中的概率 $f(c_i)$ 由其权重在总权重中的占比决定：$f(c_i) = \frac{w_i}{\sum_{j=1}^{N} w_j}$。这种设计通过给历史高频单品分配更低的权重、历史低频单品分配更高的权重，引导模型更多地选择以往较少出现的单品。

图 S2. Garments2Look 中更多叠穿与造型示例。

B.2. Look 合成

***Look* 图像的文本标注**：我们为每一张试穿图像生成了丰富的文本标注，主要包括三类信息：*(i)* 对整张试穿图像的整体描述；*(ii)* 每件单品的穿着方式，包括叠穿顺序与造型技巧；*(iii)* 与试穿效果直接相关的模特信息，如身材体型、姿态以及背景环境等。所有文本标注均由 *Gemini-2.5-Flash* 生成，随后由时尚专家进行审阅修订。

叠穿与造型 如图 *S2* 所示，我们通过为同一件单品提供多种搭配实例，充分刻画了其多样化的使用方式。这样的设计使得同一件服饰可以出现在不同的叠穿配置中（如作为单独的基础层、与外套叠穿、或与非传统款式连衣裙叠搭），也可以呈现出不同的造型方式（如携在手臂上、披在肩上等）。这类多样性对于推进真实场景虚拟试穿与服饰叠穿顺序编辑等研究非常有价值，因为它提供了丰富的例子来展示服饰之间如何相互作用，更好地还原现实世界时尚呈现的复杂性。

B.3. 数据过滤

主要服饰类别列表：运动装（*Activewear*），包类配饰（*Bag Accessories*），包类（*Bags*），海滩装（*Beachwear*），腰带（*Belts*），手链（*Bracelets*），胸针（*Brooches*），外套（*Coats*），袖扣（*Cuff Links*），领带夹（*Tie Clips*），连衣裙（*Dresses*），耳环（*Earrings*），眼镜（*Glasses*），手套（*Gloves*），发饰（*Hair Accessories*），帽子（*Hats*），夹克（*Jackets*），牛仔裤（*Jeans*），连体衣（*Jumpsuits*），针织衫（*Knitwear*），内衣（*Lingerie*），项链（*Necklaces*），裤子（*Pants*），戒指（*Rings*），围巾（*Scarves*），衬衫（*Shirts*），鞋子

（*Shoots*），短裤（*Shorts*），滑雪镜（*Ski Goggles*），裙子（*Skirts*），滑雪服（*Skiwear*），袜子（*Socks*），连裤袜（*Tights*），套装（*Suits*），太阳镜（*Sunglasses*），领带（*Ties*），领结（*Bow Ties*），上衣（*Tops*），*T* 恤（*T-shirts*），手表（*Watches*）。

类别校正：我们对原始元数据中的类别结构进行了人工校正。我们合并了语义上相近的子类别（如将“*Sportswear Jacket*”和“*Suit Jacket*”统一归入“*Jacket*”类别），并剔除了无关类别（如家居用品、手机壳等）。这一过程确保后续的搭配与 *look* 合成更聚焦于可穿戴服饰本身。

时尚专家评审流程：为保证合成数据的质量，我们设计并实施了一套严格的质控流程，邀请 *13* 位时尚领域专家参与评审：*(1)* 指南制定（*Guideline Development*），建立关于质量与造型的评估标准；*(2)* 试点研究（*Pilot Study*），对约 *2%* 的数据进行双盲标注以细化与修正评估准则；*(3)* 大规模标注（*Mass Annotation*），在标注阶段采用交叉验证机制，并由资深专家仲裁分歧，以确保评审结论的一致性。

C. 更多实验

C.1. 划分结果

如表 *S2* 所示，我们给出了按数据子集划分的实验结果；*Nano Banana* 系列在两个测试子集上均保持了 *SOTA* 表现，整体优于所有基线方法。

C.2. 显式姿态控制的影响

为研究显式姿态控制的影响，我们在测试集中抽取样本，并采用三种参考输入进行图像生成：带遮罩的模特图像、骨架图像以及 *OOTD* 图像。根据表 *S3* 所示的实验结果，我们观察到：加入骨架图像作为额外参考并不能显著提升模型指标，反而略有下降。这表明，通用图像编辑模型在同时接收不同类型的参考信息（姿态结构和物体参考）时，可能难以有效融合并消解二者的歧义。当我们将骨架图像作为第三路输入，并将其与服饰单品图像视为同等重要时，模型可能无法正确理解骨架仅作为姿态约束的角色；相反，这一额外输入会引入信息冗余或混淆，干扰模型的编辑能力，最终导致指标下降。这与专门为虚拟试穿设计的模型形成对比，后者能够更有效地利用骨架信息。

表 S2. 针对搭配级虚拟试穿任务，在 Garments2Look 测试集上对各种模型的性能评估。我们报告了经典 VTON 指标和由 VLM 评估的准确性指标。QIE-2509 = Qwen-Image-Edit-2509，NB = Nano Banana，NBP = Nano Banana Pro。

方法	KID↓	SSIM↑	LPIPS↓	服装一致性↑	叠穿准确性↑	造型准确性↑
IP-Adapter	8.3390 / 8.0255	0.8588 / 0.7492	0.1076 / 0.1768	0.5224 / 0.4747	0.6424 / 0.5948	0.6001 / 0.5506
BootComp	13.5787 / 14.3927	0.7702 / 0.4925	0.2399 / 0.4083	0.6354 / 0.4636	0.3059 / 0.3142	0.4374 / 0.3048
OmniTry	9.5100 / 14.8091	0.8036 / 0.6654	0.2011 / 0.2507	0.5292 / 0.4105	0.1371 / 0.1743	0.3576 / 0.2024
FastFit	4.6095 / 6.1260	0.8559 / 0.8544	0.0968 / 0.0948	0.7189 / 0.5527	0.1266 / 0.1313	0.4727 / 0.2599
GPT-4o (2 Ref)	3.7019 / 0.8719	0.8179 / 0.7128	0.1378 / 0.1693	0.9320 / 0.8624	0.8987 / 0.8374	0.7044 / 0.6881
Seedream 4.0 (2 Ref)	12.4870 / 13.7404	0.8126 / 0.6786	0.1740 / 0.1640	0.7805 / 0.7643	0.8513 / 0.8875	0.6286 / 0.7124
NB (N Ref)	0.6837 / 0.9384	0.8745 / 0.8357	0.0753 / 0.0596	0.8562 / 0.7659	0.9272 / 0.9245	0.7125 / 0.7697
NB (2 Ref)	0.4981 / 1.1012	0.8769 / 0.8445	0.0699 / 0.0559	0.9503 / 0.9066	0.8703 / 0.8880	0.6844 / 0.7714
NB Pro (N Ref)	0.6596 / 0.2797	0.8072 / 0.8238	0.1156 / 0.0714	0.9808 / 0.9783	0.9483 / 0.9336	0.6966 / 0.7592
NB Pro (2 Ref)	1.0233 / 0.1314	0.8112 / 0.8269	0.1108 / 0.0680	0.9796 / 0.9641	0.9082 / 0.9370	0.6712 / 0.7516
QIE-2509 (2 Ref)	0.4754 / 1.6250	0.8874 / 0.7543	0.0609 / 0.1135	0.8704 / 0.7371	0.8639 / 0.8061	0.7074 / 0.6520
QIE-2509 (2 Ref) + FT	0.1116 / 0.2199	0.8997 / 0.8674	0.0481 / 0.0563	0.9410 / 0.8944	0.9873 / 0.9539	0.8453 / 0.8679

表 S3. 针对搭配级虚拟试穿任务，在 Garments2Look 测试集上对各种模型的性能评估。我们报告了经典 VTON 指标和由 VLM 评估的准确性指标。QIE-2509 = Qwen-Image-Edit-2509，NB = Nano Banana，NBP = Nano Banana Pro。

模型	FID↓	KID↓	SSIM↑	LPIPS↓
Seedream 4.0 [36]	34.294 / 39.810 (+5.516)	10.446 / 13.310 (+2.864)	0.757 / 0.729 (-0.028)	0.335 / 0.349 (+0.014)
QIE-2509 [47]	29.030 / 32.773 (+3.743)	7.142 / 9.654 (+2.512)	**0.827** / 0.825 (-0.002)	0.116 / **0.114** (-0.002)
NB [8]	**21.545** / 21.484 (-0.061)	**0.641** / 0.867 (+0.226)	0.825 / 0.824 (-0.001)	0.131 / 0.131 (+0.000)
NBP	24.293 / 24.469 (+0.176)	2.160 / 2.217 (+0.057)	0.797 / 0.800 (+0.003)	0.174 / 0.179 (+0.005)

C.3. 微调结果

正文中的所有结果均未经过微调。我们在此补充了对 *Qwen-Image-Edit-2509* 在训练集上进行微调后的结果（由于时间限制，从训练集中随机采样 *10K* 作为微调数据）。如表 *S2*、图 *S3* 和图 *S4*（*QIE-2509+FT*）所示，定量指标、定性效果以及用户研究结果均表明性能获得了提升。

C.4. 叠穿准确性用户研究

我们开展了一项针对叠穿质量的用户研究，重点评估遮挡关系是否合理以及叠穿顺序是否准确等方面。如图 *S4* 所示，不同方法之间的性能存在显著差异，反映出它们利用标注信息能力的不同。受限于仅支持单层处理的范式，*FastFit* 获得的偏好评分最低。值得注意的是，在使用我们数据集的一小部分进行微调后，*QIE-2509* 的表现从 *0.41* 明显提升至 *0.627*。这一结果凸显了我们数据集本身的价值，也表明模型能够有效吸收并利用其中蕴含的丰富信息。

C.5. 更多示例和结果

在图 *S5 to S9* 中，我们展示了来自多个数据集的样本数据，以直观呈现本数据集所提供的丰富标注。在

图 S3. 在 Garments2Look 训练集中随机采样 1 万数据对 QIE-2509 进行微调后的定性结果示例。

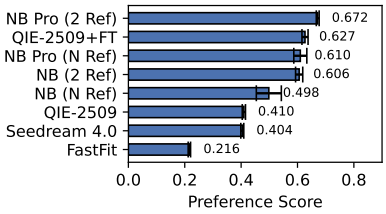


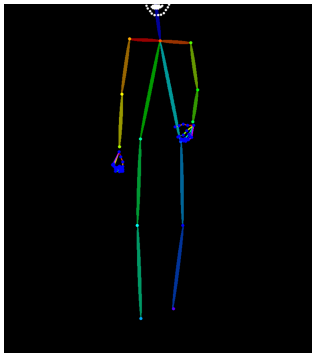
图 S4. 关于叠穿准确性的用户研究结果。

图 *S10 to S14* 中，我们给出了更多由 *SOTA* 虚拟试穿模型、通用图像编辑模型以及简单微调后的模型生成的结果。这些示例共同展示了不同模型在虚拟试穿任务上的表现、新任务（搭配级虚拟试穿）的内在难度，以及我们的数据集在该任务中的实用价值。

风格引导示例：极简风格	
模板	# 极简风穿搭指南 本指南总结了极简风穿搭的核心原则，强调通过利落线条、中性色调和高品质面料，在整体上实现有意图感、经久耐看与低调优雅。 ## I. 核心偏好与禁忌 | 类别 | 偏好 | 禁忌 | | **主色** | 白色、黑色、灰色、海军蓝、米色/驼色 | 明亮荧光色、过于响亮的基础色（如电光蓝）、高饱和宝石色调 | | **点缀色** | 橄榄绿、柔和藕粉、陶土色、浅卡其、深炭灰 | 荧光色、强烈对比的高饱和色、偏光/炫彩面料 | | **图案** | 纯色、低调条纹（如细针织条、细横条）、小格纹、低调人字纹、小型几何图案 | 大朵花卉、抢眼动物纹、夸张抽象图案、波点、繁复欧式花纹 | | **版型/剪裁** | 线条简洁、合身剪裁、有结构感、略宽松但有设计感、A 字型、直筒版型 | 紧身包臀、过分宽大/无版型、过度荷叶边与褶皱、杂乱无结构的不对称剪裁 | | **材质/面料** | 棉、亚麻、羊毛（羊绒、美利奴）、真丝、顺滑皮革、天丝/莫代尔 | 廉价感化纤、过度反光/亮面材质、粗糙花呢、严重做旧牛仔 | | **配饰** | 细致金/银首饰、有型皮质包、经典皮带、简洁墨镜、极简腕表、素色丝巾 | 夸张粗链条项链、重装饰包袋、卡通/新奇配饰、叠戴过多饰品 | | **整体氛围/情绪** | 轻松高级、平静克制、经典耐看、干练精致、有意图、从容冷静 | 浮夸追潮、夺目炫耀、视觉拥挤、过于随意/邋遢 | ## II. 搭配风格模式 本风格的穿搭注重有意识的简约、高质量基础单品，以及形式与功能的平衡，通过协调统一的轮廓与细节，营造出干净利落且经久耐看的视觉效果。 ### 1. 经典穿搭示例（4 个示例） | 风格 | 结构（单品组合） | 关键词 | | **轻松日常** | 白色略宽松纽扣衬衫 + 修身黑色阔腿裤 + 极简皮质乐福鞋 + 细圈金耳环 + 有型皮质托特包 | 干练、舒适、多场景 | | **精致通勤** | 驼色羊毛西装外套 + 米色真丝吊带背心 + 海军蓝直筒长裤 + 低跟皮鞋 + 简约极简腕表 | 精炼、适合商务、线条流畅 | | **高级休闲** | 灰色羊绒圆领针织衫 + 黑色 A 字中长裙 + 白色极简运动鞋 + 小巧斜挎包 + 简约吊坠项链 | 放松、优雅、低调 | | **现代都市** | 黑色假高领长袖上衣 + 浅色直筒牛仔裤 + 踝靴 + 黑色剪裁利落风衣外套 + 大框墨镜 | 现代、利落、都市感 | ### 2. 穿搭扩展规则总结 - **配色原则：**以中性色为主（黑、白、灰、米色、海军蓝），常见为单色系或低对比度搭配。点缀色应柔和且控制在小范围内，以保持平静感并避免视觉杂乱。 - **叠穿原则：**叠穿应具备明确目的，如增加保暖、增加层次或提升细微质感，而非单纯堆砌。各层之间需边缘整齐，保证整体轮廓干净利落并兼顾实穿性。 - **版型要求：**以干净、合身且舒适为主。衣物应轻贴身形但不过分紧绷，或在少数单品上有“刻意的宽松感”，避免任何过紧、过松或明显不合身的状态。 - **图案限制：**强烈偏好纯色。如需图案，只使用存在感极低的形式，如细针织条纹、小格纹或极淡纹理，确保不会破坏极简整体，而是作为细节兴趣点。 - **鞋包原则：**鞋与包的选择以经典设计、高品质材质与功能性为核心。造型干净、不做过多装饰，既能提升整体精致度，又能长久耐穿（如皮质乐福鞋、简洁高跟鞋、结构感托特包等）。 - **配饰要求：**配饰数量精简但质感良好，强调低调优雅与实用功能。首饰用于衬托整体而非成为视觉中心，更偏好细致金属首饰、经典腕表与简洁墨镜。 - **整体平衡：**通过对比例、材质与色彩的综合考量来达成视觉平衡。整体搭配避免任何单一元素过于抢眼或破坏干净线条，从而营造平静、克制又自然的高级感。 - **场景适应性：**穿搭需具备高度通用性，在办公室、日常外出、晚间聚会等场景间，只需少量调整即可切换，依托的是经典基础单品与不过时的极简风格。

表 S4. 极简风穿搭风格引导的提示词模板示例。

Model



Pose

Segmentation

beige cotton jersey t-shirt
Sydney cotton jersey T-shirt
clothing::tops::long-sleeved tops
Layer 1

burgundy wool and cashmere sweater
Wool and cashmere sweater
clothing::knitwear::sweaters
Layer 2

black mid-rise wide-leg jeans
Mid-rise wide-leg jeans
clothing::jeans::wide-leg jeans

beige suede-trimmed leather sneakers
Ball Star suede-trimmed leather sneakers
shoes::sneakers::low-top sneakers

black studded leather shoulder bag
Elena Small studded leather shoulder bag
bags::shoulder bags

earrings
Agoflus drop earrings
jewelry::fashion
jewelry::earrings

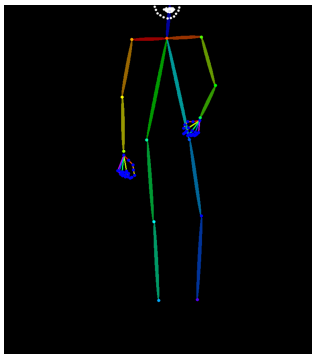
grey leather belt
Leather belt
accessories::belts

OUTFIT INFO
STYLE Casual Style
SEASON Autumn
OCCASION Daily Wear
TOPIC Cozy Comfort and Effortless Chic
OVERALL DESCRIPTION This casual ensemble features a rich deep red wool and cashmere sweater layered over a cream cotton t-shirt, paired with relaxed dark grey wide-leg jeans. A black studded shoulder bag and classic sneakers complete this comfortable yet stylish look, accented by delicate drop earrings.
PALETTTE Deep Red, Dark Grey, Cream, Black, Silver

MODEL INFO
BODY
Slender physique, exuding a relaxed and confident presence, showcasing the outfit's comfortable fit and modern silhouette.
POSE
Standing front-facing with one hand casually in a jeans pocket, conveying an approachable and natural demeanor.
BACKGROUND
Simple, clean off-white studio background provides a neutral canvas, effectively highlighting the outfit's details.

图 S5. Garments2Look 中更多示例数据（1/5）。

Model



Pose

Segmentation

white striped cotton poplin shirt
Striped cotton poplin shirt
clothing::shirts::casual::long-sleeved shirts
collar worn outside sweater neckline
Layer 1

blue cashmere polo sweater
Cashmere polo sweater
clothing::knitwear::polo sweaters
Layer 3

beige leather blouson
Leather blouson
clothing::jackets::leather
worn unbuttoned
Layer 4

jeans
Mid-rise slim jeans
clothing::jeans::straight-leg jeans
cuffed hem

black suede penny loafers
Suede penny loafers
shoes::loafers

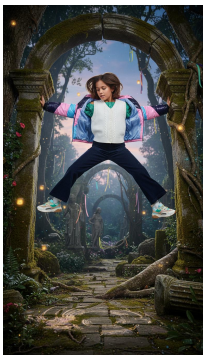
grey leather-trimmed canvas garment bag
Leather-trimmed canvas garment bag
bags::travel bags

red striped silk tie
Striped silk tie
accessories::ties
Layer 2

OUTFIT INFO
STYLE Smart Casual Style
SEASON Autumn, Spring
OCCASION Business Casual, Weekend Social
TOPIC Layered Sophistication, Refined Comfort
OVERALL DESCRIPTION A refined smart casual ensemble layers a beige blouson over a navy cashmere sweater, striped shirt, and burgundy tie. Dark wash slim jeans are paired with suede loafers, complemented by a sophisticated leather-trimmed canvas bag. The look offers a blend of classic menswear with contemporary ease and texture.
PALETTTE Navy Blue, Beige, White, Burgundy, Dark Brown, Grey

MODEL INFO
BODY
The model has a slender, well-proportioned physique, exuding a confident yet relaxed demeanor with a natural posture.
POSE
The model stands naturally with one hand casually in his jeans pocket, projecting a relaxed and approachable stance.
BACKGROUND
A minimalist, light grey studio background provides a clean and unobtrusive setting, focusing attention solely on the outfit.

图 S6. Garments2Look 中更多示例数据（2/5）。



Model

Pose

Segmentation

green printed ruffled top
Cycas printed ruffled cotton-blend top
clothing::girls' tops & shirts
Layer 1

white cable-knit sweater vest
Jobolene sweater vest
clothing::girls' knitwear::girls' cardigans & vests
Layer 2

multicoloured colorblocked puffer jacket
Super Mojo colorblocked down ski jacket
clothing::girls' skiwear
worn off-shoulder
Layer 3

dark blue corduroy pants
Junon cotton corduroy pants
clothing::girls' pants

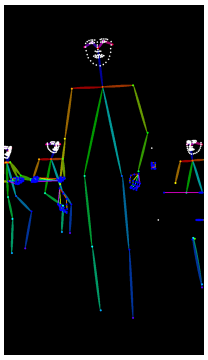
multicoloured colorblocked sneakers
Colorblocked sneakers
shoes::girls' sneakers

OUTFIT INFO
STYLE Youthful Sporty
SEASON Autumn / Winter
OCCASION Daily Wear / Outdoor Play
TOPIC Vibrant Layering and Playful Energy
OVERALL DESCRIPTION A child sports a dynamic, layered ensemble featuring a vibrant colorblocked jacket worn playfully, a classic white sweater vest, and dark flared pants. Completed with stylish colorblocked sneakers, this energetic look is perfect for adventurous outings.
PALETTTE Pink, Light Blue, Navy, White, Green, Beige

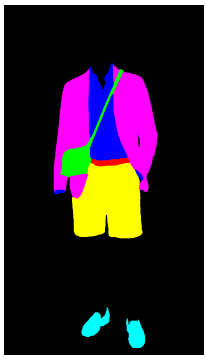
MODEL INFO
BODY
A youthful child with a slender build, demonstrating active and spirited physical engagement.
POSE
Captured mid-air during a jump, with arms and legs spread wide in an energetic, joyful expression.
BACKGROUND
An enchanting, moss-covered ancient ruin in a forest setting, illuminated by soft, magical glowing lights.

图 S7. Garments2Look 中更多示例数据（3/5）。

Model



Pose



Segmentation

blue striped cotton shirt
Striped cotton ooplin shirt
clothing::shirts::casual::short-sleeved shirts
partially unbuttoned, tucked-in
Layer 1

black single-breasted blazer
Icon single-breasted blazer
clothing::tailoring::suit jackets
worn open
Layer 2

blue cotton shorts
Cotton shorts
clothing::shorts::casual

brown suede loafers
Logo suede loafers
shoes::loafers

brown leather crossbody bag
Gear Small leather crossbody bag
bags::crossbody bags
crossbody worn

black woven leather belt
Intrecciato leather belt
accessories::belts
belted

OUTFIT INFO
STYLE Sophisticated Casual
SEASON Summer
OCCASION Business Casual Event, Resort Wear
TOPIC Smart Summer Professionalism
OVERALL DESCRIPTION A refined summer ensemble featuring tailored shorts paired with a striped shirt and blazer, exuding effortless elegance. The look balances professionalism with a relaxed, warm-weather appropriate aesthetic, complemented by classic accessories for a polished finish.
PALETTTE Navy Blue, Light Blue, Brown, Black

MODEL INFO
BODY
The model has a fit, athletic build with a neat beard, displaying a confident and composed demeanor.
POSE
Standing upright with one hand in a pocket, the model maintains a relaxed yet poised stance, engaging directly with the viewer.
BACKGROUND
The scene is set on a wooden boat deck with an ocean view under a clear, bright sky, suggesting an outdoor meeting.

图 S8. Garments2Look 中更多示例数据（4/5）。

Model

Pose

Segmentation

burgundy cable knit crew neck sweater
Women's Cable Knit Crew Neck Sweater
top
tucked-in front
Layer 1

black leather biker jacket
Acne Studios Women's Black Leather Biker Jacket
outwear
worn fully zipped
Layer 2

jeans
Women's Black High Waisted Ripped Skinny Jeans
pants

black patent platform lace-up ankle boots
Black Patent Platform Lace-Up Ankle Boots
shoes

black leather top handle satchel bag
Classic Black Leather Top Handle Satchel Bag
bag

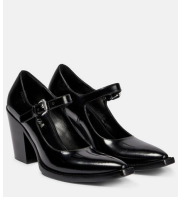
round half-rim black gold sunglasses
Round Half-Rim Matte Sunglasses
eyewear

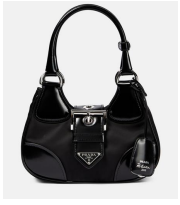
black leather chain fingerless gloves
Women's Black Leather Chain Fingerless Gloves
gloves

OUTFIT INFO
STYLE Edgy Street Style
SEASON Autumn
OCCASION Daily Wear
TOPIC Contrast and Texture, Bold and Effortless
OVERALL DESCRIPTION The outfit combines a rich burgundy cable knit sweater with a sleek black leather biker jacket layered over it. High-waisted ripped skinny jeans and glossy platform lace-up ankle boots emphasize a rebellious urban feel. Accented by fingerless leather gloves, round sunglasses, and a structured satchel, it exudes confident streetwear flair.
PALETTTE Black, Burgundy, Silver

MODEL INFO
BODY
Tall and slim physique with a poised, confident and composed demeanor.
POSE
Standing straight, slightly angled, left hand raised near face, right hand holding handbag relaxed.
BACKGROUND
Simple, clean white backdrop enhancing focus on the outfit and model.

图 S9. Garments2Look 中更多示例数据（5/5）。





white tank top
Logo cotton tank top
clothing::tops::sleeveless tops
tucked-in
Layer 1

brown asymmetric coat
Asymmetric coat
clothing::coats::knee-length coats
worn open
Layer 2

black leather midi skirt
Leather midi skirt
clothing::skirts::midi skirts

black leather Mary Jane pumps
Leather Mary Jane pumps
shoes::pumps::high-heel pumps

black Re-Nylon tote bag
Moon Re-Nylon tote bag
bags::shoulder bags

Reference

IP-Adapter

BootComp

OmniTry

FastFit

GPT-4o (N Ref)

GPT-4o (2 Ref)

Seedream 4.0 (N Ref)

Seedream 4.0 (2 Ref)

NB (N Ref)

NB (2 Ref)

QIE-2509 (N Ref)

QIE-2509 (2 Ref)

NBP (N Ref)

NBP (2 Ref)

QIE-2509 (2 Ref) + FT

GT

图 S10. Garments2Look 测试集上更多对比结果（1/5）。

multicoloured striped cotton shirt
Striped cotton shirt
clothing::tops::long-sleeved tops
partially unbuttoned, sleeves rolled-up, front-tucked

white denim shorts
Kate denim shorts
clothing::shorts::denim
worn open

beige raffia slide sandals
Ineni raffia slides
shoes::sandals::flat sandals

beige raffia pouch
Cassandre Large raffia pouch
bags::pouches

tan leather belt
Leather belt
accessories::belts

Reference

IP-Adapter

BootComp

OmniTry

FastFit

GPT-4o (N Ref)

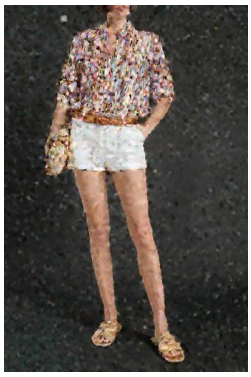
GPT-4o (2 Ref)

Seedream 4.0 (N Ref)

Seedream 4.0 (2 Ref)

NB (N Ref)

NB (2 Ref)



QIE-2509 (N Ref)

QIE-2509 (2 Ref)

NBP (N Ref)

NBP (2 Ref)

QIE-2509 (2 Ref) + FT

GT

图 S11. Garments2Look 测试集上更多对比结果（2/5）。

dress
Embroidered cotton jersey and tulle dress
clothing::girls' dresses
Layer 1

white down vest
Ghany down vest
clothing::girls' jackets
unbuttoned
Layer 2

blue wool cardigan
Giovanna wool cardigan
clothing::girls' knitwear::girls' cardigans & vests
tied around shoulders
Layer 3

beige appliqué cotton-blend sweatpants
Lou appliqué cotton-blend sweatpants
clothing::girls' sweatpants

green leather ballet flats
Dory leather ballet flats
shoes::girls' loafers & ballet flats

black bow headband
Jin Liberty headband
accessories::headbands

Reference

IP-Adapter

BootComp

OmniTry

FastFit

GPT-4o (N Ref)

GPT-4o (2 Ref)

Seedream 4.0 (N Ref)

Seedream 4.0 (2 Ref)

NB (N Ref)

NB (2 Ref)

QIE-2509 (N Ref)

QIE-2509 (2 Ref)

NBP (N Ref)

NBP (2 Ref)

QIE-2509 (2 Ref) + FT

GT

图 S12. Garments2Look 测试集上更多对比结果（3/5）。







blue argyle wool polo shirt
Argyle wool polo shirt
clothing::tops::short-sleeved tops
tucked-in

black pleated tennis skirt
Grand Slam pleated tennis skirt
clothing::activewear::skirts

black leather loafers
Penny leather loafers
shoes::loafers

burgundy leather crossbody bag
Cassandre leather crossbody bag
bags::crossbody bags
worn crossbody

gold drop earrings with freshwater pearls
Drop earrings with freshwater pearls
jewelry::fashion
jewelry::earrings

silver watch with diamonds
Serpenti Seduttori watch with diamonds
jewelry::watches::watches

brown leather belt
Downtown leather belt
accessories::belts

Reference


IP-Adapter


BootComp


OmniTry


FastFit


GPT-4o (N Ref)


GPT-4o (2 Ref)


Seedream 4.0 (N Ref)


Seedream 4.0 (2 Ref)


NB (N Ref)

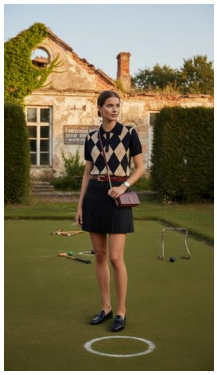

NB (2 Ref)


QIE-2509 (N Ref)


QIE-2509 (2 Ref)


NBP (N Ref)


NBP (2 Ref)


QIE-2509 (2 Ref) + FT


GT

图 S13. Garments2Look 测试集上更多对比结果（4/5）。

white shirt
Linen shirt
clothing::shirts::casual::long-sleeved shirts
tucked-in
Layer 1

neutral cashmere down vest
Cashmere down vest
clothing::jackets::vests
unbuttoned
Layer 3

black cotton and silk wide-leg pants
Xon cotton and silk wide-leg pants
clothing::pants::drawstring
drawstring tied

black croc-effect leather loafers
Marphy Loira croc-effect leather loafers
shoes::loafers

green leather tote bag
Marlo leather tote bag
bags::totes

black square-frame acetate sunglasses
Square-frame acetate sunglasses
accessories::sunglasses
worn on face

necklace
14kt white gold necklace with diamonds
accessories::jewelry::fine jewelry::necklaces

Reference

IP-Adapter | BootComp | OmniTry | FastFit

GPT-4o (N Ref) | GPT-4o (2 Ref) | Seedream 4.0 (N Ref) | Seedream 4.0 (2 Ref)

NB (N Ref) | NB (2 Ref) | QIE-2509 (N Ref) | QIE-2509 (2 Ref)

NBP (N Ref) | NBP (2 Ref) | QIE-2509 (2 Ref) + FT | GT

图 S14. Garments2Look 测试集上更多对比结果（5/5）。