# SMPL-GPTexture：使用文本到图像生成模型的双视角 3D 人体纹理估计

**Mingxiao Tu, Shuchang Ye, Hoijoon Jung, Jinman Kim***
School of Computer Science
University of Sydney
Sydney, Australia
jinman.kim@sydney.edu.au

## Abstract

生成高质量、逼真的三维人类头像纹理仍然是计算机视觉和多媒体领域中一个基础但具有挑战性的任务。然而，由于隐私、伦理以及获取成本等问题，真实的人体正反面成对图像很难获得，这限制了数据的可扩展性。此外，使用深度生成模型（如 GANs 或扩散模型）从图像输入学习先验知识来推断未见区域（如人体背面）通常会导致伪影、结构不一致或细节丢失。为了解决这些问题，我们提出了 SMPL-GPTexture（带皮肤的多个人线性模型 - 通用纹理），这是一种新的流程，它以自然语言提示作为输入，并利用最先进的文本到图像生成模型来生成人类主体的成对高分辨率正反面图像，作为纹理估计的起点。使用生成的成对双视图图像，我们首先采用人体网格恢复模型获得每个视角下图像像素与三维模型 UV 坐标之间的稳健 2D 至 3D SMPL 对齐。其次，我们使用了一种逆向光栅化技术，该技术将输入图像中的观察到的颜色显式地投影到 UV 空间中，从而生成准确且完整的纹理贴图。最后，我们应用基于扩散的修复模块来填充缺失区域，并融合机制将这些结果合并为统一的完整纹理映射。广泛的实验表明，我们的 SMPL-GPTexture 能够根据用户的提示生成与之对齐的高分辨率纹理。

## 1 介绍

高质量、照片级逼真的纹理生成对于 3D 人形模型来说是计算机视觉和多媒体领域中的一个关键问题，具有在虚拟试穿、电影动画、游戏人物以及远程呈现或根据单视图输入解决遮挡问题等直接应用。传统方法通常依赖于 360ř 多相机设置或全身三维扫描来生成准确的纹理，这需要投入大量时间和成本。因此，获取这样的数据集可能会非常昂贵，并且还会引发与实际人类相关的伦理、隐私和版权方面的担忧。

最近的进展探索了单视图方法，利用网上或开源数据库中广泛可用的二维图像来估算用于蒙皮多人线性模型（SMPL）或 SMPL-X 模型的完整 UV 纹理 [12, 15]。这些方法通常使用卷积神经网络（CNN）或基于注意力的架构直接回归或将“修复”颜色信息映射到 UV 空间。然

*corresponding author

Preprint. Under review.

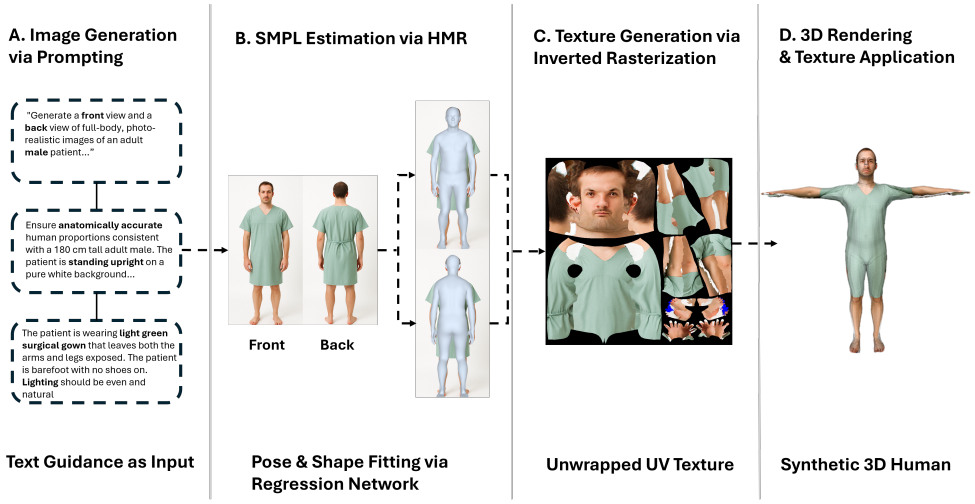


图 1: 提出的 SMPL-GPTexture 方法概览，该方法仅使用文本提示作为输入来生成前后一致的逼真图像。使用 HMR 估计 SMPL 身体模型，并使用逆栅格化生成未包裹的 UV 纹理贴图。

而，仅仅依赖单个正面视图不可避免地限制了全局上下文的捕捉，导致缺乏主体背面关键外观细节的不完整或错误纹理。为了解决这一局限性，引入额外视角，例如显式建模后视图，已被证明对于生成更准确且无伪影的纹理至关重要。然而，由于可扩展性的挑战，人类主体的真实前后图像很少可用。作为一种替代方案，使用学习先验、扩散或修复技术生成未见视角也可能导致视觉伪影，如模糊或不一致的纹理，特别是在被遮挡或未见区域 [9]。

为了解决这些问题，我们提出了 SMPL-GPTexture（SMPL-通用纹理），一种新颖的框架，利用最新的文本到图像生成模型来生成人体主体的前后配对图像作为纹理估计的起点，如图 1 所示。在文本到图像生成模型方面的最新进展，包括 GPT-4o[13] 已经展示了通过自然语言提示生成一致且高度逼真的图像的强大能力。通过生成两个互补的高质量视图，我们的方法克服了其他方法中固有的遮挡或推理歧义，并为纹理重建提供了更完整的依据。

通过由提议的文本提示生成的配对前后输入图像，SMPL-GPTexture 采用最先进的三维人体网格恢复（HMR）模型为每个视图获得稳健的 SMPL/SMPL-X 2D 拟合，确保图像像素与 3D 模型 UV 坐标的精准一一对应。然后它利用一种逆栅格化技术，明确地将从前和后图像中观察到的颜色投影到 UV 空间中，从而生成高分辨率（1024x1024）、准确、完整的纹理贴图。在后续阶段，基于扩散的修复模块填补缺失区域，并通过融合机制将这些结果整合成一个统一的高质量纹理贴图。

SMPLGPTexture 提供了几个显著的优势。首先，仅依赖自然语言提示进行数据合成而不使用真实的在线图像进行训练或推理，我们的框架避免了隐私、伦理和成本方面的担忧。其次，生成配对的正面和背面视图使模型能够同时显式捕捉正反面信息，这比依赖单视图输入、从学习到的先验知识推断背面视图或需要完整多视图设置的方法更能准确重建被遮挡区域。

## 2 相关工作

**单视图和多视图纹理补全** 现有的 3D 人体纹理估计方法以不同的方式解决了背面缺失的问题。Alldieck 等人 [1] 使用单目视频拟合一个 3D SMPL 模型，并将每一帧的像素投影到模型的 UV 图上，将这些部分纹理融合为完整的 360ř 表示。Zhao 等人 [20] 通过从一个视角预测纹理并然后从另一个视角渲染来强制执行跨视图一致性；然后使用人体部位分割线索将渲染

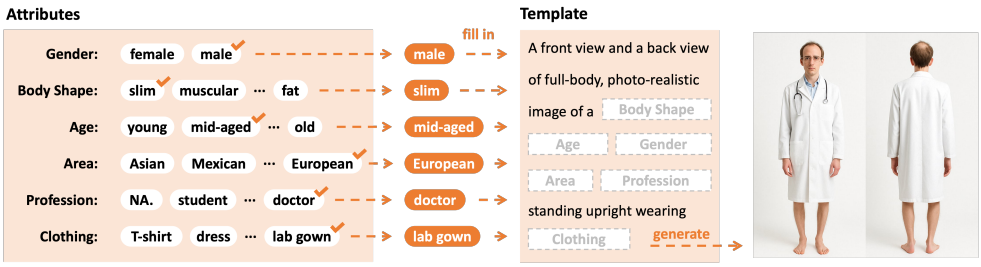


图 2: 结构化文本到图像提示管道用于全身人体图像生成。

图像与相应的视图进行比较。Xu 等人 [19] 采用变压器模型，其中仅使用单个图像推断未见的背面，从而产生统一的纹理。最近，SiTH[2] 采用了基于扩散的流水线来在观察到正面照片的条件下生成一个后视图图像，而 ConTex-Human[3] 则引入了一个由估计深度和文本提示引导的后视合成模块来生成一致的背面纹理。尽管这些方法主要旨在增强基于 SMPL 模型的 3D 纹理重建，但它们要么依赖于多视角设置，要么在填充未观察到区域时会出现伪影。

**文本到图像模型在纹理生成中的应用** 近期文本到图像生成模型的进步使得高质量纹理合成能够符合用户需求。AvataCLIP [8] 利用形状 VAE [5] 初始化 3D 几何结构，随后通过 CLIP [16] 进行基于文本引导的体积渲染优化。然而，生成的纹理通常仍然粗糙，且对 CLIP 特征的依赖可能导致指导不精确或语义纠缠。TexDreamer [10] 通过使用文本-纹理对微调一个文本到图像的潜在扩散模型 [17] 来生成高分辨率的人类 UV 纹理。然后它利用大型语言模型如 ChatGPT [14] 创建 50,000 个多样化的结构化描述，以引导由微调后的扩散模型进行的纹理合成过程。CPO [22] 利用 ChatGPT 生成正反对比提示，这些提示用于训练奖励模型，以增强长而复杂的用户提示与所产生的纹理之间的对齐。

**关于公开/开放的人像数据库/纹理包的总结。用于 3D 模型。**若干公开的数据集已被用于基于 SMPL 的纹理估计，这些数据集大致可以分为三类。第一类是在野外采集的图像数据集——例如行人重新识别数据集（如 [21]）和店内衣物检索数据集（[11]），以及多视角视频数据集——提供了具有不同视角的大量图像。然而，这些数据源仅提供弱监督且缺乏真实 UV 纹理。第二类是真实的扫描 3D 人体数据集（例如 THuman2.0[6]），包含了穿着各种衣物的人的高质量 3D 扫描结果以及精确纹理化的 SMPL 网格。尽管这些数据集提供了准确的纹理信息，但由于扫描过程的成本和资源需求较大，其规模通常有限。第三类是大规模多视角图像数据集，例如最近引入的 MVHumanNet[4]。以 MVHumanNet 为例，它收集了超过 6.45 亿帧覆盖 4,500 个身份（产生 9,000 种着装）的数据，并且还提供了每个序列服装的文字描述，这对验证文本到图像生成管道是有益的。然而，这些数据集稀缺且获取成本高昂，这促使需要高效的合成双视角方法来克服这些限制。

## 3 方法论

### 3.1 通过文本到图像模型生成双重视图图像

图 2 展示了一个结构化的文本到图像提示生成过程，用于根据预定义的属性和条件模板生成全身、照片级逼真的人类图像。左侧选择了一组语义属性，如性别、体型、年龄、地区、职业和服装等。然后将这些属性值（例如：男性、苗条、中年、欧洲人、医生、实验外套）填入预定义的文本模板中，描述所需的图像内容，包括外观和姿态。最后，我们使用最先进的文本到图像生成模型，包括 GPT-4o 和一个基于文本引导的自回归图像生成模型 Infinity[7]，根据这些结构化提示实现高质量、语义对齐的图像生成。

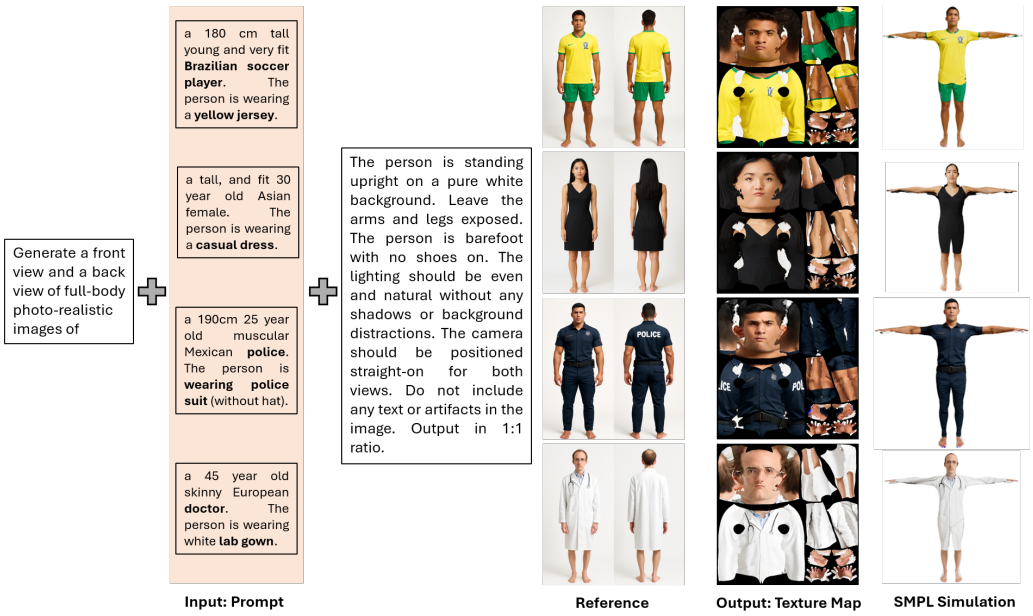


图 3: 选定纹理图由 SMPL-GPTexture（GPT-4o）生成的定性结果

### 3.2 SMPL/SMPL-X 拟合使用人体网格恢复

通过生成的正面和背面图像，我们应用最先进的 HMR 模型 [18] 来获得图像中合成人物的姿态和形状估计，这在图像像素与 SMPL/SMPLX 模型上的相应区域之间建立了精确的二维到三维对齐。初始的 SMPL 参数估计是通过回归网络得到的。然后将 SMPL 模型与其 2D 关键点对齐，方法是将其 3D 关节投影到图像中，并最小化与检测到的关键点之间的欧几里得距离，同时使用额外的先验知识确保姿态合理。使用 MPJPE 和 PA-MPJPE 等指标来评估对齐情况，从而得出稳健的三维人体模型估计。这一步建立了输入图像与 3D 身体部位之间可靠的对应关系，这是许多方法通过单纯依赖注意力机制框架而绕过的关键组成部分。

### 3.3 逆光栅化用于纹理投影

我们的方法然后采用逆栅格化技术，将合成的前后图像中的像素颜色信息反向投影到 SMPL 模型的展开 UV 空间中。与传统的正向渲染不同，后者将 3D 顶点映射到 2D 图像上，逆栅格化将 2D 颜色信息“抬升”回对应的 SMPL UV 坐标。每个合成视图独立投影，生成单独的 UV 纹理贴图。随后，融合模块整合这两个视角的投影：在两种纹理都可用的区域中，混合操作强制空间一致性，并通过基于扩散的修复模块填充间隙。这一过程产生了一个统一、高保真的纹理贴图，能够准确地符合底层 3D 几何形状。

## 4 实验

### 4.1 文本到图像生成模型

我们评估了两个最先进的文本到图像生成系统：（1）OpenAI 的 GPT-4o 图像生成模型和（2）Infinity 模型。使用相同的提示模板，我们为十个全身人体样本生成了正面和背面图像对。样本集包括五种日常服装（例如运动服、T 恤、商务正装）和五种职业制服（病人病号服、医护人员工作服、警察、士兵、消防员）。对于每张图像对，我们记录了三个关键指标：（i）输出

表 1: 图像生成模型的定量结果，注意这些是十个提示下的平均结果。

| Model | Resolution (px) | Time (s) | Size (MB) |
|---|---|---|---|
| GPT4o | 1024x1024 | 33.61 | 1.52 |
| Infinity | 1024x1024 | 3.60 | 0.16 |

表 2: 纹理估计的定量结果。

| Model | MPAE ↓ | OCE ↓ |
|---|---|---|
| GPT4o | **1.09** | **2.72** |
| Infinity | 1.32 | 2.97 |

分辨率（宽度 Œ 高度像素），(ii) 文件大小（以兆字节为单位），以及 (iii) 生成延迟（秒）。然后报告所有十个样本对的每个指标的平均值。

### 4.2 纹理估计

为了定量评估我们的纹理估计管道的保真度，我们采用了两个自洽性指标：平均投影角度误差（MPAE）和重叠一致性误差（OCE）。这两个指标都不需要外部的真实纹理，并且仅依赖于我们方法生成的几何形状和多视角投影。

**平均投影角度误差（MPAE）** MPAE 衡量了反投影的观察方向与每个可见体素处重建的表面法线之间的对齐程度。设 $\mathcal{P}$ 是所有接收到至少一个有效投影的体素集合。对于每个体素 $i \in \mathcal{P}$，设 $\boldsymbol{n}_i$ 为估计的表面法线，$\boldsymbol{v}_i$ 为从相机到该表面位置的归一化视图向量。$\theta_i$ 与 $\boldsymbol{n}_i$ 和 $\boldsymbol{v}_i$ 之间的角偏差是

$$\theta_i = \arccos\big(\boldsymbol{n}_i \cdot \boldsymbol{v}_i\big). \tag{1}$$

MPAE 然后被定义为这些偏差的平均值：

$$\mathrm{MPAE} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \theta_i. \tag{2}$$

**重叠一致性误差（OCE）** OCE 评估两个视图在两者都覆盖的纹理映射区域中的一致性。给定两个视图 $a$ 和 $b$，设 $\mathcal{O}$ 为在这两个视图中可见的纹素集合。对于每个 $i \in \mathcal{O}$，分别用 $t_i^a$ 和 $t_i^b$ 表示从视图 $a$ 和 $b$ 采样的相应纹理属性（例如，强度或阴影响应）。重叠一致性误差是 $\mathcal{O}$ 的平均绝对差异。

$$\mathrm{OCE} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \left| \, t_i^a - t_i^b \, \right|. \tag{3}$$

## 5 结果与讨论

### 5.1 文本到图像生成模型的分析

表 1 报告了每个模型在十个提示下的平均分辨率、生成时间和文件大小。GPT4o 和 Infinity 均以 1024Œ1024 px 的尺寸生成图像，确保了公平的比较基础。然而，在计算成本和存储占用方面，这两种模型出现了显著差异。GPT4o 平均需要 33.61 秒才能完成每个双视角配对，并输出大小为 1.52 MB 的文件。其较长的运行时间可能反映了内部采样更为复杂或为了视觉保真度分配了更大的内部网络层。Infinity 每个双视角生成只需要 3.60 秒，同时产生的文件体积较小（0.16 MB）。这表明其架构更为高效，并且输出图像具有更高的压缩率。

### 5.2 纹理估计分析

表 2 概述了两个自洽性指标——MPAE 和 OCE，它们量化了估计纹理的几何一致性，而无需参考真实值。GPT4o 的 MPAE 为 1.09，OCE 为 2.72。较低的 MPAE 表明其每个 texel 的反向投影角度与表面法线更加接近，且较低的 OCE 显示在重叠区域前视图和后视图之间的一致性更紧密。Infinity 展现了更高的 MPAE（1.32）和 OCE（2.97），暗示其每像素几何对齐和视角间一致性略为不准确。这些结果表明，GPT4o 更丰富的内部表示产生更为忠实的纹理放置，而 Infinity 的更快推理在投影保真度上有所妥协。定性结果显示在图 3 中，该图说明了从每个配置的输入提示中估计出的真实且高分辨率的人类纹理。

总结来说，当需要最大程度的纹理保真度并且计算资源或时间充足时，GPT4o 是更优选。对于高吞吐量或资源受限的设置，在可以接受几何精度略有降低的情况下，Infinity 是更好的选择。

## 6 结论

总之，提出的 SMPL-GPTexture 框架通过利用最先进的文本到图像生成模型有效解决了为 3D 人类角色生成高质量、照片级真实的纹理的挑战。通过仅使用自然语言提示来促进成对前后面图像的创建，我们的流程克服了依赖于图像输入的学习方法所面临的隐私和伦理问题。在输入图像上使用 HMR 和反向光栅化确保了产生的纹理贴图既准确又完整。实验结果表明，我们的方法实现了出色的几何一致性和纹理保真度。这一进展为虚拟试穿、电影动画等应用提供了重要的见解，开启了计算机视觉和图形领域的新可能性。

## 参考文献


[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video-based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018.

[2] A. Author and B. Author. Sith: A diffusion-based pipeline for hallucinating back-view textures from a front photo. In *arXiv preprint arXiv:230X.XXXXX*, 2023.

[3] C. Author and D. Author. Contex-human: Back-view synthesis for consistent 3d human texture estimation. In *CVPR*, pages 2345–2353, 2024.

[4] X. Author and Y. Author. Mvhumannet: Large-scale multi-view human dataset for 3d reconstruction. In *CVPR*, 2024.

[5] Simone Foti, Bongjin Koo, Danail Stoyanov, and Matthew J Clarkson. 3d shape variational autoencoder latent disentanglement via mini-batch feature swapping for bodies and faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18730–18739, 2022.

[6] Qianru Guo et al. Thuman2.0: A high-quality 3d human dataset. In *CVPR*, 2020.

[7] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

[8] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

[9] Donald Knuth. Knuth: Computers and typesetting. URL `http://www-cs-faculty.stanford.edu/~{}uno/abcde.html`.

[10] Yufei Liu, Junwei Zhu, Junshu Tang, Shijie Zhang, Jiangning Zhang, Weijian Cao, Chengjie Wang, Yunsheng Wu, and Dongjin Huang. Texdreamer: Towards zero-shot high-fidelity 3d human texture generation. In *European Conference on Computer Vision*, pages 184–202. Springer, 2024.

[11] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval. In *CVPR*, pages 1096–1104, 2016.

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[13] OpenAI. Addendum to GPT-4o system card: 4o image generation. `https://openai.com/index/gpt-4o-image-generation-system-card-addendum/`, 2025. Accessed: 2025-04-02.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

[18] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–915, 2024.

[19] Xiangyu Xu and Chen Change Loy. 3d human texture estimation from a single image with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13849–13858, 2021.

[20] Fang Zhao, Shengcai Liao, Kaihao Zhang, and Ling Shao. Human parsing based texture transfer from single image to 3d human via cross-view consistency. In *NeurIPS*, pages 11240–11250, 2020.

[21] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qiao Qi. Market-1501: A benchmark for large scale person re-identification. In *ICCV*, pages 1116–1124, 2015.

[22] Pengfei Zhou, Xukun Shen, and Yong Hu. Text-driven 3d human generation via contrastive preference optimization. *arXiv preprint arXiv:2502.08977*, 2025.