DressWild: Feed-Forward Pose-Agnostic Garment Sewing Pattern Generation from In-the-Wild Images

Zeng Tao[1,*] **Ying Jiang**[1,*] **Yunuo Chen**[1,*] **Tianyi Xie**[1],
Huamin Wang[2], **Ying Nian Wu**[1], **Yin Yang**[3], **Chenfanfu Jiang**[1]
[1]University of California, Los Angeles [2]Style3D [3]University of Utah

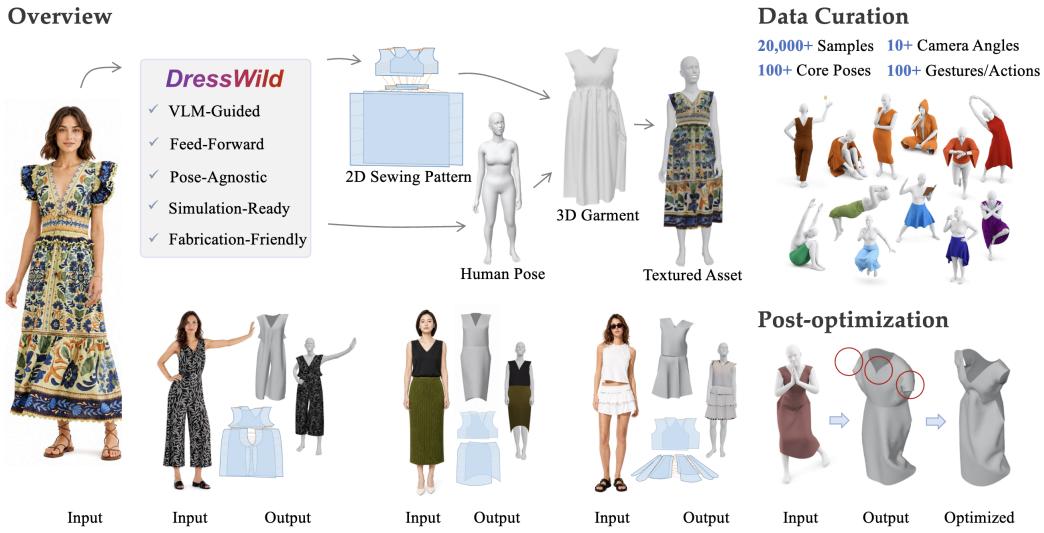

Figure 1: We propose **DressWild**, a feed-forward sewing-pattern generation pipeline that reconstructs simulation-ready, pose-agnostic 2D sewing patterns from a single in-the-wild image, while producing the corresponding pose-fitted 3D garment.

Abstract

Recent advances in garment pattern generation have shown promising progress. However, existing feed-forward methods struggle with diverse poses and viewpoints, while optimization-based approaches are computationally expensive and difficult to scale. This paper focuses on sewing pattern generation for garment modeling and fabrication applications that demand editable, separable, and simulation-ready garments. We propose **DressWild**, a novel feed-forward pipeline that reconstructs physics-consistent 2D sewing patterns and the corresponding 3D garments from a single in-the-wild image. Given an input image, our method leverages vision–language models (VLMs) to normalize pose variations at the image level, then extract pose-aware, 3D-informed garment features. These features are fused through a transformer-based encoder and subsequently used to predict sewing pattern parameters, which can be directly applied to physical simulation, texture synthesis, and multi-piece outfits virtual try-on. Extensive experiments demonstrate that our approach robustly recovers diverse sewing patterns and the corresponding 3D garments from in-the-wild images without requiring multi-view inputs or iterative optimization, offering an efficient and scalable solution for realistic garment simulation and animation.

*These authors contributed equally to this work.

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

1 Introduction

Traditional 3D garment creation follows a multi-stage pipeline that progresses from concept design to 2D sewing pattern drafting and finally to 3D garment assembly through virtual sewing and garment simulation. While this workflow enables precise control and fabrication-ready outputs, it is inherently time-consuming and requires substantial domain expertise, making it inaccessible to non-experts. Recent advances in visual computing and artificial intelligence have significantly lowered the barrier to 3D garment creation by enabling direct generation from images [1, 19], text prompts [16, 56], or point clouds [14], enabling efficient and scalable downstream virtual try-on and fabrication-aware garment design [49]. However, most existing approaches focus solely on producing visually plausible 3D garment geometry, without recovering the underlying 2D sewing patterns. The absence of pattern-level representations limits editability, parametric control, and physical manufacturability, posing a fundamental challenge for downstream design iteration and real-world fabrication.

Existing approaches for sewing-pattern-based garment generation can be broadly categorized into feed-forward methods [25, 26, 14] and optimization-based pipelines [23, 24]. Data-driven feed-forward methods leverage paired data to directly predict sewing patterns or 3D garments but their performance is tightly coupled to the training distribution, often restricting them to canonical body poses such as A- or T-poses and limiting their ability to generalize to diverse poses. Moreover, these approaches typically require multiple input images or controlled conditions, reducing their practicality in real-world scenarios. Optimization-based methods, on the other hand, explicitly enforce physical consistency between sewing patterns and 3D garments and can support multi-pose configurations; however, they rely on iterative simulation and gradient-based optimization, resulting in high computational cost and long runtimes. These limitations motivate the need for a more efficient framework that can recover sewing patterns and physically consistent 3D garments from an image.

To address these challenges, we introduce **DressWild**, a feed-forward sewing-pattern generation pipeline that reconstructs pose-agnostic 2D sewing patterns from a single in-the-wild image by leveraging VLM priors, while producing the corresponding pose-fitted 3D garment. Given an in-the-wild input image, we first leverage a VLM to generate an auxiliary standard T-pose 2D representation of the dressed person. We then perform garment segmentation on both the canonical T-pose image and the original input image, and extract reconstruction features from the segmented outputs to capture the garment's structural and appearance cues while suppressing background clutter. In parallel, we perform pose feature extraction on the input image to explicitly encode body articulation. These features are subsequently integrated through our feature fusion module, which injects VLM priors while selectively attending to complementary structure and pose cues. We predict the 2D sewing pattern representation based on the fused features. To summarize, our contributions are:

- We propose a feed-forward garment reconstruction pipeline that predicts diverse 2D sewing patterns and the corresponding physically consistent 3D garments, accurately fitted to people in arbitrary poses from a single in-the-wild image.
- We propose a data processing pipeline capable of handling diverse multi-view and multi-pose input. We further introduce a feature fusion design that effectively incorporates VLM priors, enabling robust sewing pattern recovery and garment reconstruction under challenging multi-pose configurations.
- We conduct extensive experiments demonstrating the effectiveness and versatility of our approach, successfully reconstructing diverse sewing patterns and multi-pose 3D garments from in-the-wild images.

2 Related Work

2.1 3D Reconstruction

Powerful generative models—including diffusions, transformers, and GANs—now synthesize 3D shapes from images [44, 35, 45, 50], videos [5], text [3, 34], and sketches [17]. Training strategies generally split into 3D supervision and 2D prior guidance. The former maps sparse inputs to 3D ground truth (e.g., meshes [46], voxels [51], or implicit fields [54]) but requires large-scale aligned data. The latter utilizes differentiable rendering and pretrained 2D priors, though ensuring multi-view consistency remains challenging [11, 37, 52]. Despite their performance, data-driven approaches

struggle with in-the-wild data due to distribution biases [23]. To bypass paired data requirements, optimization-based methods like SDS [34] and VSD [48] refine 3D representations via pretrained priors; however, their iterative nature hinders scalability. In contrast, we propose an efficient, scalable feed-forward framework for garment generation from in-the-wild images.

2.2 Garment Modeling and Generation

Garment geometry modeling explores diverse representations, including implicit 3D fields like UDF [55], SDF [30], and NeRF [29], as well as explicit meshes, point clouds, 3D Gaussian splats [47, 16, 36], and 2D sewing patterns [25, 10]. Early template-based pipelines [38, 7, 8, 32] predict deformations via per-vertex offsets or latents but require category-specific training. Other optimization-based methods [28, 16, 36, 2] refine meshes using multi-view or CLIP guidance yet still depend on predefined templates. To enhance diversity, recent generalized models synthesize garments from images or text without category-specific constraints. For instance, SMPLicit [6] and ClothWild [30] utilize implicit fields (UDF/SDF) for reconstruction, while Surf-D [55] employs diffusion for non-watertight geometry. However, these methods often lack explicit 2D sewing patterns, limiting downstream simulation. More recently, DressCode [10] and Sewformer [25] have utilized transformers to generate both 3D garments and 2D patterns. Our work similarly focuses on the joint generation of sewing patterns and 3D garments.

2.3 Sewing Pattern Generation

Large-scale datasets like GarmentCodeData [12] and SewFactory [25] have advanced data-driven sewing pattern generation. Prior works utilize diffusion, transformers, or autoregressive models [10, 4, 9, 26, 22] to infer vector representations from multimodal prompts. However, discrete latent spaces in these methods can hinder generalization. Conversely, rasterized implicit representations, as seen in Garment Image [39], ISP [21], GarmentRecovery [20], and DMap [19], provide more continuous spaces for efficient reconstruction. To enhance control, AIpparel [31], Design2GarmentCode [56], and ChatGarment [1] incorporate LLMs, LMMs, and VLMs, yet they often struggle with in-the-wild images due to training biases. While differentiable frameworks like Dress-1-to-3 [23] and DiffAvatar [24] achieve high-fidelity for out-of-distribution shapes, they remain computationally expensive and sensitive to extreme motions. Regarding texture, DeepIron [15] predicts unwarping for fixed patterns. In contrast, our work efficiently generates diverse, pose-agnostic, and simulation-ready textured patterns from in-the-wild images via fused multi-pose features.

3 Method

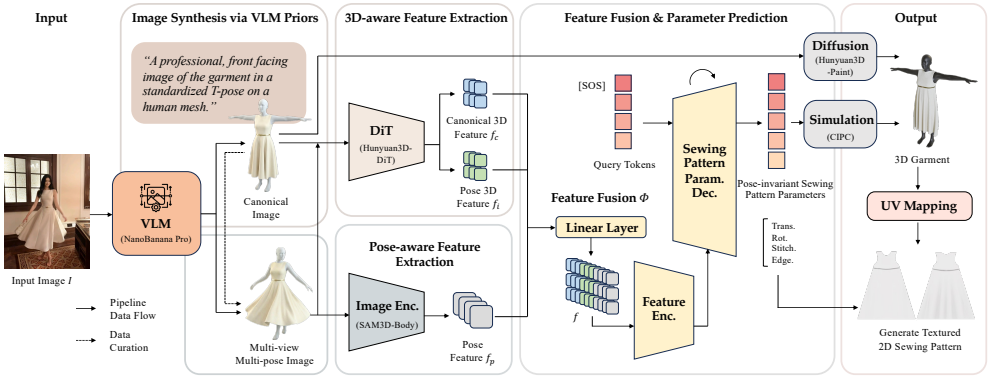

Figure 2: Overview of our **DressWild** pipeline. Given an in-the-wild image, our method reconstructs simulation-ready pi patterns and a corresponding 3D garment.

3.1 Overview

As shown in Fig. 2, we present **DressWild**, a novel framework for diverse and pose-agnostic sewing pattern generation that supports a wide range of garment categories, including T-shirts, jackets, pants, skirts, jumpsuits, dresses, and other apparel types. Given an input in-the-wild image I, which may exhibit arbitrary pose and viewpoint, **DressWild** first explores a VLM like NanoBanana Pro [40] to

transform the input image I into a canonical representation by synthesizing a garment image in a fixed T-pose and front-facing view I_c. This normalization is achieved using a pre-trained VLM, which provides strong priors for disentangling pose and viewpoint variations from garment appearance. By leveraging VLM priors, our framework aligns the normalized image I_c to the pose and view distribution of the training dataset while preserving garment-specific visual details. Given I_c and I, we extract image appearance features f_i, and canonical-space normalized image features f_c using Hunyuan3D [42]. In parallel, we employ SAM3D-Body [53], a state-of-the-art human reconstruction method, to extract pose-aware features f_p and recover a 3D human mesh M_h from the input image I. These embeddings are fused to form a joint feature

$$f = \Phi(f_p, f_i, f_c), \tag{1}$$

where $\Phi(\cdot)$ denotes a feature fusion module. f is further processed by a feed-forward garment modeling network to predict pose- and view-invariant sewing pattern parameters. This design enables **DressWild** to leverage generative models trained on curated datasets while robustly generalizing to diverse, in-the-wild images without requiring multi-view or pose annotations. In the following sections, we elaborate on each component of the proposed pipeline.

3.2 Data Processing

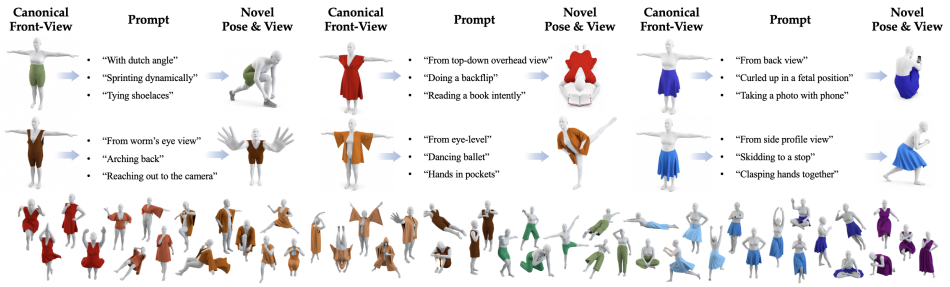

Figure 3: **Data Curation and Augmentation.** We use VLMs to generate novel-pose and novel-view images with consistent garments.

We curate a sewing pattern dataset featuring multi-pose and multi-view garment images to support learning-based modeling under diverse body configurations. Built upon [13], which contains over 20,000 variations across 12 categories, we extend its original dual-view T-pose limitation using a Vision-Language Model (VLM). Specifically, we synthesize images by randomly combining three configurable parameters: n_v viewpoints, n_p body poses, and n_g gestures. While our framework supports augmenting backgrounds, lighting, and materials, we exclude these in the final construction. Such variations increase synthesis instability and may cause inconsistencies between generated images and garment geometry. Since these factors primarily affect appearance rather than structural learning, we restrict our augmentation to pose, view, and gesture levels to ensure data fidelity.

More specifically, the viewpoint configuration includes short keyword prompts such as "low angle", "worm's-eye view", "Dutch angle", and "eye-level". The pose configuration is represented using concise descriptors including "standing", "arching back", "jumping", and "walking". The gesture configuration further augments motion diversity with prompt tokens such as "arms crossed", "hands on hips", and "wiping sweat from the forehead". This randomized combination of short, keyword-based prompts enables the generation of diverse and physically plausible variations while preserving garment structure and identity. In total, we define 12 distinct viewpoints, 105 body poses, and 112 gestures/actions, whose randomized combinations are used as prompt tokens for multi-view and multi-pose image generation.

3.3 Feature Extraction

Given an in-the-wild input image I, our goal is to extract pose-aware and pose-invariant garment features for robust sewing pattern prediction. To this end, we construct three complementary feature streams capturing image appearance, canonical garment structure, and human pose.

Canonicalization and Segmentation. We first leverage a VLM to normalize pose and viewpoint by synthesizing a canonical garment image I_c in a fixed front-facing T-pose. This canonicalization step

disentangles pose and view variations from garment appearance while preserving garment-specific visual details. We then apply HybridGL [27] to segment both the original input image I and the canonical image I_c, producing segmented garment images I_{si} and I_{sc}, respectively.

3D Reconstruction Features. We employ Hunyuan3D-2 [41] to extract 3D reconstruction features based on the segmented images. Specifically, we utilize the intermediate latent representations produced after the Hunyuan3D-DiT sampling stage and before the VAE decoder, which capture rich 3D-aware garment geometry. From the segmented original image I_{si}, we extract multi-view multi-pose reconstruction features f_i, while from the canonical T-pose image I_{sc}, we obtain pose-normalized reconstruction features f_c. Both feature types are derived using the same reconstruction pipeline to ensure consistency.

Pose Features. In parallel, we apply SAM3D-Body [53] to extract pose-aware embeddings f_p by taking I_{si} by the SAM3D-Body encoder. These features explicitly encode human pose information.

3.4 Model

Our model consists of two stages including feature fusion and parameter decoding. Multimodal features are fused into a pose and view invariant representation f, which is then decoded to predict structured sewing pattern parameters Θ including edge, stitching, rotation, and translation information.

Feature Fusion Module. Since the three feature types originate from different encoders and exhibit heterogeneous dimensions and semantics, we first project each feature into a shared embedding space of dimension d_w using separate linear transformations, $\tilde{f_i} = W_i f_i, \tilde{f_c} = W_c f_c, \tilde{f_p} = W_p f_p$ where W_i, W_c, and W_p denote learnable projection matrices. This operation ensures that all features are in the same feature space and can be jointly processed. We then concatenate the projected features along the sequence dimension, $\tilde{f} = [\tilde{f_i}; \tilde{f_c}; \tilde{f_p}]$, where the resulting sequence length equals $|f_i| + |f_c| + |f_p|$. The concatenated feature sequence is passed through a transformer encoder [43], which performs self-attention across all tokens to capture cross-modal interactions among image appearance, canonical garment structure, and pose information. This enables the model to adaptively aggregate complementary cues from different feature sources, producing a fused representation $f = \Phi(\tilde{f})$, where $\Phi(\cdot)$ denotes the feature fusion module.

Parameter Decoding. Given the fused feature representation f, we employ a transformer-based decoder [10] to autoregressively predict the sewing pattern parameters.

Each input token is formed by the sum of three embeddings: a positional embedding that identifies the panel associated with the token, a parameter type embedding that distinguishes edge geometry, rigid transformations, and stitching attributes, and a value embedding that encodes the quantized sewing pattern parameters. Each output token represents a specific element of the sewing pattern parameterization, including panel vertex coordinates, edge curvature control points, 3D translation, rotation, and stitching correspondences. Through iterative decoding, the decoder attends to the fused feature representation and generates a sequence of latent tokens, which are subsequently decoded into the structured sewing pattern parameters (see appendix A.3.1 for more details):

$$\Theta = \{\{\boldsymbol{v}_{i,k}\}, \{\boldsymbol{c}_{i,j}\}, \{(\boldsymbol{r}_i, \boldsymbol{T}_i)\}, \mathcal{S}\}.$$

By conditioning the decoder on pose- and view- invariant fused features, the predicted sewing patterns are robust to pose and viewpoint variations and are directly compatible with downstream physical simulation and animation.

Training Objective. At each decoding step t, the decoder predicts a categorical distribution over discrete pattern parameter tokens, corresponding to the conditional likelihood

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{L} \log p(f_t \mid f_{<t},\, f; \theta),$$

where f_t denotes the ground truth token at step t, $f_{<t}$ represents all previously generated tokens, and f is the fused conditioning feature. This formulation is equivalent to the standard token-wise cross-entropy loss used in autoregressive sequence modeling.

4 Post Process

4.1 Texture Generation

The final step of **DressWild** focuses on generating high-fidelity textures for both sewing patterns and their corresponding garment meshes, with the goal of supporting downstream garment fabrication. To this end, we introduce a dedicated texture generation module to reconstruct garment appearance. Unlike approaches that jointly generate geometry and texture, we explicitly decouple appearance synthesis from geometric reconstruction. Jointly predicting 2D sewing pattern textures and garment geometry is severely limited by the scarcity of paired supervision, making end-to-end learning difficult to scale. In addition, joint 3D texture and geometry generation from images is inherently ill-posed, as visual observations ambiguously entangle illumination, material appearance, and fine-scale geometric details, hindering reliable disentanglement of texture from shape.

Based on this design choice, we adopt a generative approach to first synthesize textures on the reconstructed 3D garment surface and subsequently transfer them to the sewing pattern domain. After obtaining sewing pattern parameters and generating a 3D garment mesh draped on the human body via physical simulation in Sec. 4.2, we extract multi-view consistent garment textures using Hunyuan3D-Paint [41], which goes through image de-lighting, multi-view consistent image generation, super-resolution and baking. The generated textures are then projected onto the UV parameterization induced by the sewing patterns, ensuring seam consistency and direct applicability to pattern-based fabrication workflows. This separation allows textures to be consistently aligned across seams, robust to pose-induced deformations, and directly applicable to physical garment production.

4.2 Garment Simulation

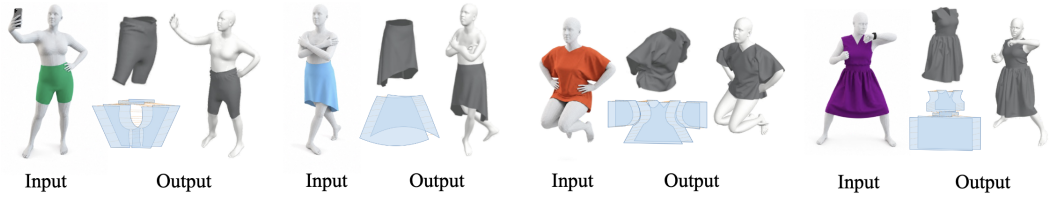

Figure 4: **Simulation Results with Human Poses.** Through pose estimation and simulation, the generated 2D sewing patterns can be fitted onto the target human body while preserving a shape consistent with the original image.

We adopt the SMPL-X model [33] as the articulated human body representation and follow Dress123 [23] to simulate garments draped on humans. We initialize garment patterns around a T-posed body, then utilize the Codimensional Incremental Potential Contact (CIPC) simulator [18] to stitch the panels around the body, while enforcing high stretching and bending stiffness to preserve the original shape of garments. The resulting configuration serves as an initial feasible state for CIPC-based dynamic simulation, after which the human body is interpolated from the T-pose to the reconstructed pose and simulated as a moving boundary condition, recovering the dressed human as its pose in the input image (see Fig. 4).

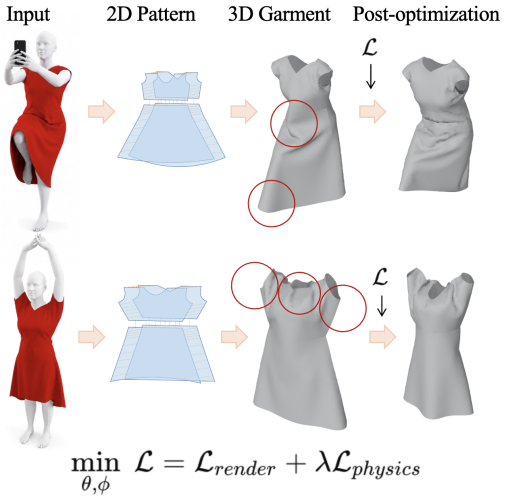

Figure 5: **Post-optimization.** We can further optimize garment pattern boundary with [23] to enhance localized features.

To ensure proper layering for multi-piece cases, stitched vertices are used to identify connected components, which are sorted vertically and sequentially fitted from bottom to top using CIPC. To prevent slippage of lower garments during pose interpolation, we locally shrink the rest shape

of triangles near the waist to induce sufficient friction; the garment rest pose otherwise remains unchanged.

4.3 Post-Optimization

After our feed-forward pipeline provides reliable prediction of sewing pattern type and shape, it can be seamlessly integrated with physical-simulation-based pattern optimization framework [23]. Through joint optimization of RGB-rendering loss and physical losses (such as tightness of clothes), we further improve localized details such as collar size and boundary shape of draped garments. As shown in Fig. 5, this optional post-optimization further improves reconstruction fidelity and enhances simulation-readiness of generated assets.

5 Experiments

5.1 Experiment Settings

Implementation Details We employ NanoBanana Pro [40] as the VLM. For image feature extraction, we utilize the pre-trained Hunyuan3D-2 [41] model. We employ the pre-trained HybridGL [27] model for referring segmentation, and SAM3D-Body Dino-v3 [53] is used for pose feature extraction. Regarding the model architecture, the decoder consists of 24 layers with 8 attention heads. We use the Adam optimizer with an initial learning rate set to 10^{-4}. The model is trained for 10 epochs on a single NVIDIA H100 GPU.

Dataset We utilize [13] and our generated data to train and evaluate the models. The sewing patterns are categorized into 12 classes based on the dataset guidance, which contains 25031 data points. We split the dataset into a training set, a validation set, and a testing set in a ratio of 7:2:1.

Baselines We conduct qualitative and quantitative comparisons with state-of-the-art sewing pattern generation methods, NeuralTailor and Sewformer, which generate sewing patterns from point clouds and single images, respectively.

5.2 Quantitative Comparison

We conduct quantitative comparisons with the baseline methods on the same test set across multiple evaluation metrics, including panel accuracy, edge accuracy, shape reconstruction error, Filtered shape (F-shape) error, and Chamfer Distance (CD). Table 1 presents our quantitative results. NeuralTailor relies on garment point cloud inputs, while Sewformer and our method predict sewing patterns directly from single in-the-wild images. For fair comparison, we use Hunyuan3D to generate the mesh from the input image and sample point clouds as inputs to NeuralTailor. To provide a more controlled comparison, we additionally retrained NeuralTailor and Sewformer on the DressWild dataset using the same dataset split and training setting. For each method, we selected the checkpoint with the best validation loss and evaluated it on the test set. As shown in Table 1, NeuralTailor achieves 88.16% panel accuracy and 79.45% edge accuracy, which reflects its limited robustness to pose and viewpoint variations. Sewformer improves structural prediction, reaching 92.70% panel accuracy and 83.17% edge accuracy, but still struggles to recover correct panel topology and geometry under arbitrary poses. In contrast, our method substantially outperforms both baselines, achieving 94.35% panel accuracy and 85.41% edge accuracy. Moreover, our approach yields significantly lower geometric errors, reducing the ℓ_2 shape error from 10.66 for NeuralTailor and 7.13 for Sewformer to 6.22, and achieving the lowest Chamfer Distance of 0.01899. These results demonstrate that our method produces both structurally correct and geometrically accurate sewing patterns from challenging in-the-wild images with diverse poses and viewpoints.

We additionally evaluated DressWild on the public SewFactory benchmark and compared it with existing approaches. We followed the official SewFactory protocol and created a unified train/validation/test split. Detailed results are provided in Table 2. Our method outperforms previous baselines on SewFactory, demonstrating that the performance improvement is not limited to our internally generated dataset split. These results demonstrate that DressWild maintains strong reconstruction quality beyond our generated dataset.

Table 1: **Quantitative Comparison.** Panel Acc. and Edge Acc. denote the prediction accuracy for the number of panels and edges, respectively, while Shape $\mathcal{L}^2$ Error represents the $\mathcal{L}^2$ distance of the parameters. Our method outperforms baselines across most of the metrics.

Method	Panel Acc. ↑	Edge Acc. ↑	Shape $\mathcal{L}^2$ ↓	F-Shape $\mathcal{L}^2$ ↓	CD ↓
NeuralTailor	88.16%	79.45%	10.66	11.38	0.02379
Sewformer	92.70%	83.17%	7.13	5.27	0.02164
DressWild	**94.35%**	**85.41%**	**6.22**	**5.07**	**0.01899**

Table 2: **Quantitative Comparison on SewFactory.**

Method	Panel Acc. ↑	Edge Acc. ↑	Shape $\mathcal{L}^2$ ↓	F-Shape $\mathcal{L}^2$ ↓
NeuralTailor	82.40%	89.61%	11.6826	9.3813
Sewformer	86.32%	91.09%	9.5862	**7.7829**
DressWild	**88.83%**	**92.68%**	**8.2836**	7.8053

5.3 Qualitative Comparison

Fig. 6 presents the results of qualitative comparisons with baseline methods on the same test set described in Sec. 5.1, focusing on sewing pattern prediction from in-the-wild images with arbitrary poses and viewpoints. NeuralTailor, which relies on point cloud inputs, struggles to recover complete and structurally consistent pattern pieces, especially for garments with complex topology such as dresses and jackets. Sewformer shows reasonable performance under limited viewpoint variation, but its predictions often degrade for non-frontal or rotated poses, resulting in fragmented or misaligned panels. In contrast, our method consistently reconstructs coherent and well-proportioned sewing patterns across diverse in-the-wild conditions. By internally leveraging canonicalized representations together with original-view information, our approach better preserves garment topology, symmetry, and semantic part correspondence. These results in the figure demonstrate the robustness of our method to pose and viewpoint variations and its ability to recover simulation-ready sewing patterns from challenging in-the-wild images.

5.4 Comparison With Reconstruction-based Method

Figure 7 illustrates the comparison between our feed-forward method and the reconstruction-based method GarmentRecovery [21, 20]. Our feed-forward approach generates 2D sewing pattern parameters, which can then be draped onto various human body models to obtain the final 3D garment. In contrast, the reconstruction-based method directly models the 3D garment mesh from images. The results demonstrate that our method achieves higher fidelity to the ground truth. Furthermore, the generated 3D garment meshes are smoother and more physically plausible. Our approach facilitates seamless transferability across different human body models and diverse motion sequences. Additionally, the explicit modeling of 2D sewing patterns provides practical utility for real-world fabric cutting and garment manufacturing.

5.5 Ablation Study

We conduct ablation studies of the key components of our framework, including additional visual features (canonical-space reconstruction features, pose features), and the feature fusion module, using the same garment images as in Sec. 5.1. The experiments assess how each proposed component contributes to the final sewing pattern reconstruction quality. See appendix A.2 for details.

6 Application

Sewing Pattern Generation from In-the-wild Images As illustrated in Fig. 8, our method is capable of directly processing in-the-wild images exhibiting diverse poses and viewpoints. Given an in-the-wild input image, we apply **DressWild** to robustly generate pose-agnostic and simulation-ready sewing patterns.

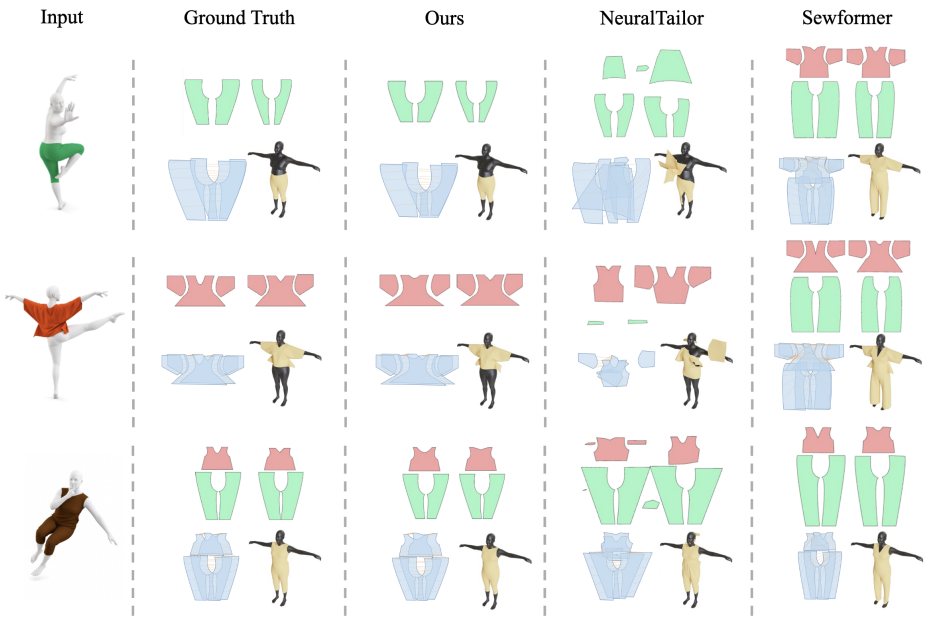

Figure 6: **Qualitative Comparison.** Our method produces high-quality 2D sewing patterns and corresponding 3D garments. Here we qualitatively compare our results with baseline approaches Sewformer [25] and NeuralTailor [14].

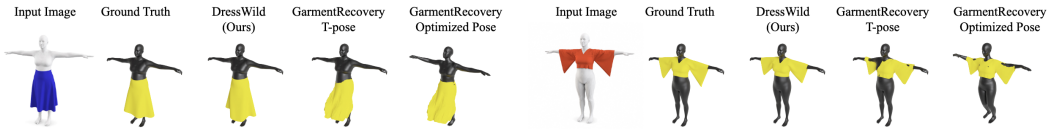

Figure 7: **Comparison with reconstruction-based method.** Our sewing-pattern-based method performs better than the reconstruction-based method.

Multi-piece Outfit Generation Given an in-the-wild image containing a multi-piece outfit, we utilize a VLM to identify and extract individual garments while preserving their relative ordering and semantic consistency. Each extracted garment layer is then processed independently using **DressWild** to predict its corresponding pose-agnostic and simulation-ready sewing pattern, as illustrated in Fig. 8. This design enables our framework to naturally generalize from single-piece garments to multi-piece outfits without requiring explicit layer annotations or modifications to the core model.



Figure 8: **Results on in-the-wild images with single- and multi-piece outfits.** Given an in-the-wild image as input, our method generates pose-agnostic and simulation-ready sewing patterns for multi-piece outfits.

7 Conclusion

In this paper, we present **DressWild**, a novel feed-forward framework for reconstructing pose-agnostic, simulation-ready 2D sewing patterns and corresponding 3D garments from a single in-the-wild image. By leveraging VLMs to generate canonical garment representations and integrating pose-aware,

canonical-space, and multi-view features via a hybrid mechanism, our approach effectively disentangles garment geometry from pose and viewpoint variations. Extensive experiments demonstrate that **DressWild** outperforms existing state-of-the-art methods in both quantitative metrics and qualitative comparison, particularly on challenging in-the-wild data with diverse poses and viewpoints. Furthermore, our pipeline supports multi-garment generation and produces physically plausible outputs compatible with downstream simulation and texture synthesis workflows, offering a scalable and efficient solution for realistic garment modeling and animation applications. **Limitations.** Our method relies on VLM-based canonicalization, and failures in image editing may affect the quality of sewing pattern reconstruction. We analyzed failure cases caused by VLM-based canonicalization. We found that this module is more likely to fail under the following conditions: Severe occlusions in the input image; Extreme poses or uncommon viewpoints; Garment categories significantly outside the training distribution. For example, when canonicalization incorrectly predicts garment attributes such as neckline or sleeve shape, this error may propagate to subsequent sewing pattern prediction. In addition, due to limitations in the training data distribution, the method may perform less effectively on detailed garments (e.g. pockets) or uncommon tailoring designs. These limitations could potentially be alleviated by expanding the diversity of garment types in the training dataset and leveraging more powerful VLMs.

Acknowledgements

We acknowledge support from TRI, Nvidia, Sony, Meshy, Nirvana, Cloudexe, and Envora.

References

[1] Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michael J Black, and Yao Feng. Chatgarment: Garment estimation, generation and editing via large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2924–2934, 2025.

[2] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, volume 41, pages 293–304. Wiley Online Library, 2022.

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023.

[4] Xipeng Chen, Guangrun Wang, Dizhong Zhu, Xiaodan Liang, Philip Torr, and Liang Lin. Structure-preserving 3d garment modeling with neural sewing machines. *Advances in Neural Information Processing Systems*, 35:15147–15159, 2022.

[5] Wenyan Cong, Hanqing Zhu, Kevin Wang, Jiahui Lei, Colton Stearns, Yuanhao Cai, Dilin Wang, Rakesh Ranjan, Matt Feiszli, Leonidas Guibas, et al. Videolifter: Lifting videos to 3d with fast hierarchical stereo alignment. *arXiv preprint arXiv:2501.01949*, 2025.

[6] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021.

[7] R Daněřek, Endri Dibra, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, volume 36, pages 269–280. Wiley Online Library, 2017.

[8] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8739–8748, 2019.

[9] Jingfeng Guo, Jinnan Chen, Weikai Chen, Zhenyu Sun, Lanjiong Li, Baozhu Zhao, Lingting Zhu, Xin Wang, and Qi Liu. Garmentx: Autoregressive parametric representations for high-fidelity 3d garment generation. *arXiv preprint arXiv:2504.20409*, 2025.

[10] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024.

[11] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16377–16387, 2025.

[12] Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung. Garmentcodedata: A dataset of 3d made-to-measure garments with sewing patterns. In *European Conference on Computer Vision*, pages 110–127. Springer, 2024.

[13] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3d garments with sewing patterns. *arXiv preprint arXiv:2109.05633*, 2021.

[14] Maria Korosteleva and Sung-Hee Lee. Neuraltailor: Reconstructing sewing pattern structures from 3d point clouds of garments. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

[15] Hyun-Song Kwon and Sung-Hee Lee. Deepiron: Predicting unwarped garment texture from a single image. *arXiv preprint arXiv:2310.15447*, 2023.

[16] Boqian Li, Xuan Li, Ying Jiang, Tianyi Xie, Feng Gao, Huamin Wang, Yin Yang, and Chenfanfu Jiang. Garmentdreamer: 3dgs guided garment synthesis with diverse geometry and texture details. In *2025 International Conference on 3D Vision (3DV)*, pages 1416–1426. IEEE, 2025.

[17] Haoxuan Li, Ziya Erkoç, Lei Li, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshpad: Interactive sketch-conditioned artist-reminiscent mesh generation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16227–16237, 2025.

[18] Minchen Li, Danny M Kaufman, and Chenfanfu Jiang. Codimensional incremental potential contact. *arXiv preprint arXiv:2012.04457*, 2020.

[19] Ren Li, Cong Cao, Corentin Dumery, Yingxuan You, Hao Li, and Pascal Fua. Single view garment reconstruction using diffusion mapping via pattern coordinates. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[20] Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. Garment recovery with shape and deformation priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1586–1595, 2024.

[21] Ren Li, Benoît Guillard, and Pascal Fua. Isp: Multi-layered garment draping with implicit sewing patterns. *Advances in Neural Information Processing Systems*, 36:40294–40319, 2023.

[22] Siran Li, Ruiyang Liu, Chen Liu, Zhendong Wang, Gaofeng He, Yong-Lu Li, Xiaogang Jin, and Huamin Wang. Garmagenet: A multimodal generative framework for sewing pattern design and generic garment modeling. *ACM Transactions on Graphics (TOG)*, 44(6):1–23, 2025.

[23] Xuan Li, Chang Yu, Wenxin Du, Ying Jiang, Tianyi Xie, Yunuo Chen, Yin Yang, and Chenfanfu Jiang. Dress-1-to-3: Single image to simulation-ready 3d outfit with diffusion prior and differentiable physics. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025.

[24] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. Diffavatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4368–4378, 2024.

[25] Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.

[26] Shengqi Liu, Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Lincheng Li, Mengxiao Bi, Xiaokang Yang, and Yichao Yan. Multimodal latent diffusion model for complex sewing pattern generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17640–17650, 2025.

[27] Ting Liu and Siyuan Li. Hybrid global-local representation with augmented spatial guidance for zero-shot referring image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29634–29643, 2025.

[28] Xinqi Liu, Jituo Li, and Guodong Lu. Modeling realistic clothing from a single image under normal guide. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3995–4007, 2023.

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[30] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *European conference on computer vision*, pages 184–200. Springer, 2022.

[31] Kiyohiro Nakayama, Jan Ackermann, Timur Levent Kesdogan, Yang Zheng, Maria Korosteleva, Olga Sorkine-Hornung, Leonidas J Guibas, Guandao Yang, and Gordon Wetzstein. Aipparel: A multimodal foundation model for digital garments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8138–8149, 2025.

[32] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7365–7375, 2020.

[33] 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.

[34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[35] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

[36] Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, and Rakesh Ranjan. Garment3dgen: 3d garment stylization and texture generation. In *2025 International Conference on 3D Vision (3DV)*, pages 1382–1393. IEEE, 2025.

[37] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[38] Xiu Su, Shan You, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Bcnet: Searching for network width with bilaterally coupled network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2175–2184, 2021.

[39] Yuki Tatsukawa, Anran Qi, I-Chao Shen, and Takeo Igarashi. Garmentimage: Raster encoding of garment sewing patterns with diverse topologies. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[40] Gemini Team, R Anil, S Borgeaud, Y Wu, JB Alayrac, J Yu, R Soricut, J Schalkwyk, AM Dai, A Hauth, et al. Gemini: A family of highly capable multimodal models, 2024. *arXiv preprint arXiv:2312.11805*, 10, 2024.

[41] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[42] Tencent Hunyuan3D Team. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details, 2025.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024.

[45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[46] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018.

[47] Ruiyan Wang, Zhengxue Cheng, Zonghao Lin, Jun Ling, Yuzhou Liu, Yanru An, Rong Xie, and Li Song. Semanticgarment: Semantic-controlled generation and editing of 3d gaussian garments. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9793–9802, 2025.

[48] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

[49] Zaiqiang Wu, I-Chao Shen, and Takeo Igarashi. Real-time per-garment virtual try-on with temporal consistency for loose-fitting garments. In *Computer Graphics Forum*, volume 44, page e70272. Wiley Online Library, 2025.

[50] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025.

[51] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019.

[52] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023.

[53] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint; identifier to be added*, 2025.

[54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.

[55] Zhengming Yu, Zhiyang Dou, Xiaoxiao Long, Cheng Lin, Zekun Li, Yuan Liu, Norman Müller, Taku Komura, Marc Habermann, Christian Theobalt, et al. Surf-d: Generating high-quality surfaces of arbitrary topologies using diffusion models. In *European Conference on Computer Vision*, pages 419–438. Springer, 2024.

[56] Feng Zhou, Ruiyang Liu, Chen Liu, Gaofeng He, Yong-Lu Li, Xiaogang Jin, and Huamin Wang. Design2garmentcode: Turning design concepts to tangible garments through program synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23712–23722, 2025.

A Technical appendices and supplementary material

A.1 Overview

In Sec A.2, we provide detailed ablation studies of different components of our model.

In Sec A.3, we provided supplementary explanations regarding the data curation paradigm for **DressWild**. We showcased various attempts in data curation along with their corresponding results, followed by detailed analysis and discussion.

In Section A.4, we provide additional experimental details and results, including more comparisons with baseline methods NeuralTailor [14] and Sewformer [25].

In Section A.5, we provide further discussion, including limitations and failure cases, as well as the broader impact of our work.

A.2 Ablation Study

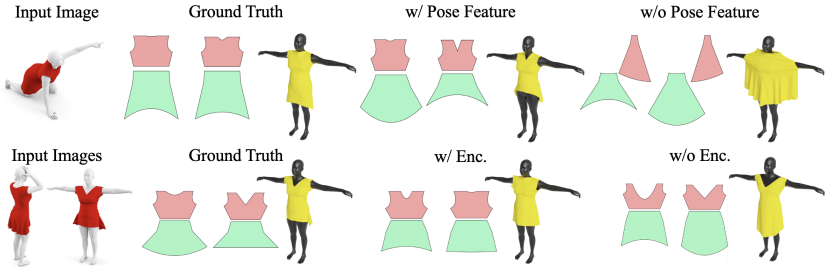

Figure 9: (Top) **Ablation of pose features.** The garment performs better with the pose feature when the human pose is challenging. (Bottom) **Ablation of feature fusion.** With feature fusion design, the results are more consistent with GT.

Feature Extraction. We first evaluate the effect of different feature choices in our framework. As shown in Tab. 3, incorporating canonical-space, front-view T-pose reconstruction features f_c consistently improves both panel and edge accuracy, while reducing the shape error, indicating that front features provide critical geometric cues for recovering more accurate panel boundaries and overall sewing pattern shapes. Regarding pose features f_p (Fig. 9 top row), predictions that incorporate pose information remain much closer to the ground truth in both panel shape and overall layout, producing more plausible boundaries. By providing explicit pose information, the model is able to disentangle pose-induced garment deformation from the underlying, pose-invariant pattern geometry. In contrast, removing pose features causes the network to misinterpret draped appearances under non-neutral poses as intrinsic pattern geometry, leading to distorted or warped panels and incorrect proportions.

Table 3: **Ablation of canonical-space reconstruction features.** The performance of sewing pattern reconstruction is significantly improved with the inclusion of this feature.

Method	Panel Acc. ↑	Edge Acc. ↑	Shape L2 ↓
w/o front features	91.18%	78.58%	12.11
w/ front features	**94.35%**	**85.41%**	**6.22**
w/o pose feature	91.71%	83.24%	8.35
w/ SMPLest-X pose feature	93.27%	85.36%	6.29
w/ SAM3D-Body pose feature	**94.35%**	**85.41%**	**6.22**
w/o encoder	92.70%	81.63%	9.88
w/ encoder	**94.35%**	**85.41%**	**6.22**
w/ SAM3D feature	94.24%	85.09%	6.87
w/ Hunyuan3D feature	**94.35%**	**85.41%**	**6.22**

Feature Fusion Encoder. Since the features f_i, f_c, and f_p originate from different encoders, we project them into a shared embedding space and fuse them with a transformer encoder using self-attention, producing a unified representation f. As shown in Fig. 9 bottom row, models equipped

with the feature fusion module produce sewing pattern predictions that closely match the ground truth in both panel shape and overall layout. By projecting f_i, f_c, and f_p into the shared embedding space, the model effectively captures cross-feature information. In contrast, removing the feature fusion module significantly degrades reconstruction quality, leading to incorrect proportions and distorted silhouettes, such as deformed collar shapes and inconsistent panel geometry. This degradation indicates that without explicit self-attention–based fusion, the model struggles to coherently aggregate complementary cues from different feature sources.

A.3 Data Curation

A.3.1 Sewing Pattern Representation

We follow the sewing pattern templates from [14] and [10], covering diverse garment categories. We represent a garment sewing pattern as a set of 2D panels together with their 3D placement and stitching topology. Specifically, a pattern consists of N_P panels $\{P_i\}_{i=1}^{N_P}$, where each panel P_i is defined as a closed planar polygon with an ordered set of vertices $\{\boldsymbol{v}_{i,k}\}_{k=1}^{N_i} \subset \mathbb{R}^2$. Consecutive vertices form edges $\{E_{i,j}\}_{j=1}^{N_i}$ that describe the panel boundary. Each edge is defined by two endpoints and an optional curvature term. Straight edges are represented as line segments, while curved edges are modeled using quadratic Bézier curves. A curved edge is parameterized by its start point $\boldsymbol{v}_s$, end point $\boldsymbol{v}_e$, and a single control point $\boldsymbol{c} \in \mathbb{R}^2$ specified in relative coordinates. To enable panel meshing and subsequent physical simulation, we discretize each curved sewing pattern edge by sampling points along its quadratic Bézier representation. Specifically, a point on a curved edge is computed as $\boldsymbol{B}(t) = (1-t)^2\boldsymbol{v}_s + 2(1-t)t\boldsymbol{c} + t^2\boldsymbol{v}_e, t \in [0,1]$, where uniformly sampling t yields a set of boundary points that approximate the continuous panel contour. To place panels in 3D space, each panel P_i is associated with a 6-DoF rigid transformation $(\boldsymbol{r}_i, \boldsymbol{T}_i)$, where $\boldsymbol{r}_i \in \mathbb{R}^3$ parameterizes the rotation and $\boldsymbol{T}_i \in \mathbb{R}^3$ denotes the translation, which defines its canonical 3D configuration. Stitching information is represented explicitly at the edge level using discrete stitching labels, where each label pairs two edges from different panels and encodes the sewing topology. Given an input image I, our network predicts the parameters of this structured representation, including panel vertex coordinates $\boldsymbol{v}_{i,k}$, edge curvature control points $\boldsymbol{c}_{i,j}$, panel-wise rigid transformations $(\boldsymbol{r}_i, \boldsymbol{T}_i)$, and stitching correspondences $\mathcal{S}$. By operating in this parametric sewing pattern space, the generated garments are pose- and view-invariant and directly compatible with physical simulation.

A.3.2 More Details on Data Curation

We leverage the robust capabilities of VLMs to bridge the pose discrepancy and data gap between the training and inference stages. While our training dataset consists primarily of canonical frontal T-pose data, in-the-wild images exhibit diverse and unconstrained poses. To address this, we design a novel data curation paradigm. Specifically, we craft a comprehensive set of tailored prompts for pose, view, and scene editing to synthesize a large volume of multi-pose and multi-view images. In the processing stage, both the canonical frontal T-pose images and the synthesized multi-pose/multi-view images undergo segmentation and feature extraction. The resulting features are then concatenated to form a robust representation, which serves as the input for the subsequent parameter prediction.

We utilize VLMs to generate multi-pose data for data augmentation. For each frontal T-pose image in the dataset, we process it using Nano Banana to generate images with diverse poses and viewpoints. Our prompt engineering follows a specific logic: we decompose the prompt into three core elements, like camera angle, core pose, and gesture/action, and use LLM to generate multiple sets of descriptions for each component. For every sample, we randomly select and combine these key elements to construct the final editing text prompt. These tailored text prompts facilitate stable and consistent batch image editing, as shown in Fig. 10.

Building upon this original prompt paradigm, we have designed more comprehensive prompt content that includes intricate backgrounds and textures, as well as a greater number of stability constraints. Specifically, the more comprehensive prompts consist of the following elements: Angle, Design, Pose, Gesture, Environment, and Light.

At the same time, the increased complexity of the elements can cause instability when using simple prompts. Consequently, we have integrated numerous robust constraints to maintain the stability of the editing process. The paradigm of text prompt is as follows: "CRITICAL REQUIREMENT: The silhouette, cut, and structural design of the clothing must be a perfect 1:1 replica of the garment in the

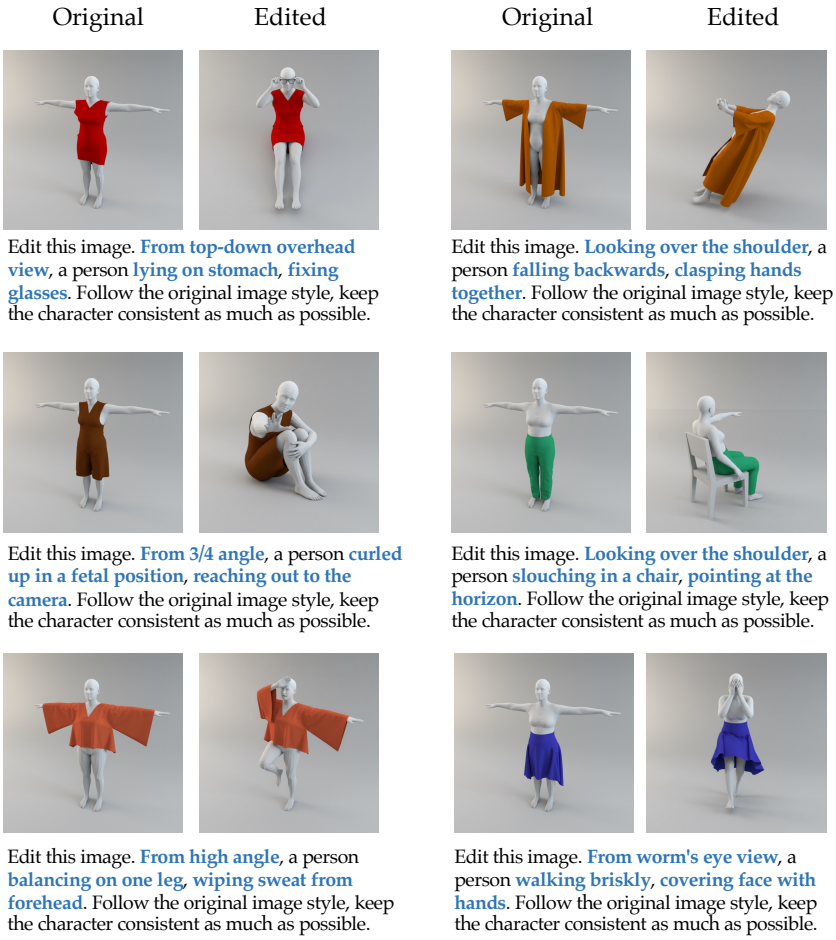

Figure 10: Data Curation Results. The paradigm of the text prompt is: "Edit this image. selected_angle, a person selected_pose, selected_gesture. Follow the original image style."

reference image. Angle: selected_angle. The clothing material reference is: selected_design. Subject: A real person wearing this exact outfit, selected_pose, selected_gesture. Environment: selected_env. Lighting: selected_light. Strict Requirements: 1. PHOTOREALISM & TEXTURE: The image must look like a raw DSLR photograph. The fabric must show hyper-realistic details. 2. INTEGRATION: Lighting on the person must match the background perfectly. 3. ANATOMICAL CONSISTENCY: Enforce strict human physiological constraints. If the requested gesture conflicts with the pose, prioritize a physically plausible stance. Ensure the subject has exactly two arms and natural joint articulation. Eliminate any hallucinations or morphing artifacts (e.g., extra limbs, floating hands) caused by comprehensive motion descriptions." The editing results are shown in Fig. 11.

Comprehensive prompt designs enable corresponding edits to the original image. However, due to the high complexity of the prompts, current models face the following issues in such settings, as shown in Fig. 12: (1) Alignment with Prompts: During the editing process, because a large amount of detailed information is provided, the model may fail to follow every instruction strictly. This results in cases where the editing output is inconsistent with the prompt. (2) Physical and Anatomical Irregularities: The editing results may occasionally violate physical laws or human anatomical structures, such as generating an incorrect number of limbs. Despite our use of constraints within the prompt templates to mitigate these issues, the performance bottlenecks of current pretrained models mean that results obtained from such comprehensive prompts still have a probability of exhibiting the aforementioned problems. Therefore, in this work, we adopt the original prompt design as shown in Fig. 10 to ensure

Original Edited




CRITICAL REQUIREMENT: The silhouette, cut, and structural design of the clothing must be a perfect 1:1 replica of the garment in the reference image. Angle: **From side profile view**. The clothing material reference is: **made of quilted nylon padding with a diamond stitching pattern**. Subject: A real person wearing this exact outfit, **hanging from a horizontal bar by hands, gesturing 'stop' with palm out**. Environment: **inside a grand abandoned industrial warehouse with dust motes dancing in light beams from high windows**. Lighting: **Rembrandt lighting (dramatic side lighting)**. Strict Requirements: 1. PHOTOREALISM & TEXTURE: The image must look like a raw DSLR photograph. The fabric must show hyper-realistic details. 2. INTEGRATION: Lighting on the person must match the background perfectly. 3. ANATOMICAL CONSISTENCY: Enforce strict human physiological constraints. If the requested gesture conflicts with the pose, prioritize a physically plausible stance. Ensure the subject has exactly two arms and natural joint articulation. Eliminate any hallucinations or morphing artifacts (e.g., extra limbs, floating hands) caused by complex motion descriptions.

Original Edited




CRITICAL REQUIREMENT: The silhouette, cut, and structural design of the clothing must be a perfect 1:1 replica of the garment in the reference image. Angle: **From high angle**. The clothing material reference is: **with a abstract marble texture print in black, white, and gold**. Subject: A real person wearing this exact outfit, **playing an air guitar enthusiastically, pulling hair in frustration**. Environment: **in a cozy, warm-lit vintage library filled with antique books and leather armchairs**. Lighting: **dramatic volumetric lighting with atmospheric fog**. Strict Requirements: 1. PHOTOREALISM & TEXTURE: The image must look like a raw DSLR photograph. The fabric must show hyper-realistic details. 2. INTEGRATION: Lighting on the person must match the background perfectly. 3. ANATOMICAL CONSISTENCY: Enforce strict human physiological constraints. If the requested gesture conflicts with the pose, prioritize a physically plausible stance. Ensure the subject has exactly two arms and natural joint articulation. Eliminate any hallucinations or morphing artifacts (e.g., extra limbs, floating hands) caused by complex motion descriptions.




CRITICAL REQUIREMENT: The silhouette, cut, and structural design of the clothing must be a perfect 1:1 replica of the garment in the reference image. Angle: **From eye-level**. The clothing material reference is: **made of delicate, sheer black lace with floral patterns layered over a solid base**. Subject: A real person wearing this exact outfit, **jumping for joy with limbs spread (star jump), clasping hands together in front**. Environment: on a gritty industrial loading dock with shipping containers. Lighting: dappled sunlight filtering through tree leaves (gobo effect). Strict Requirements: 1. PHOTOREALISM & TEXTURE: The image must look like a raw DSLR photograph. The fabric must show hyper-realistic details. 2. INTEGRATION: Lighting on the person must match the background perfectly. 3. ANATOMICAL CONSISTENCY: Enforce strict human physiological constraints. If the requested gesture conflicts with the pose, prioritize a physically plausible stance. Ensure the subject has exactly two arms and natural joint articulation. Eliminate any hallucinations or morphing artifacts (e.g., extra limbs, floating hands) caused by complex motion descriptions.




CRITICAL REQUIREMENT: The silhouette, cut, and structural design of the clothing must be a perfect 1:1 replica of the garment in the reference image. Angle: **From 3/4 angle**. The clothing material reference is: **featuring a repeating fleur-de-lis motif in gold thread on velvet**. Subject: A real person wearing this exact outfit, **sprinting at full speed (blur effect), beckoning someone to come closer**. Environment: **in a flower-filled alpine meadow with mountains in the background**. Lighting: **direct camera flash style with hard shadows (paparazzi style)**. Strict Requirements: 1. PHOTOREALISM & TEXTURE: The image must look like a raw DSLR photograph. The fabric must show hyper-realistic details. 2. INTEGRATION: Lighting on the person must match the background perfectly. 3. ANATOMICAL CONSISTENCY: Enforce strict human physiological constraints. If the requested gesture conflicts with the pose, prioritize a physically plausible stance. Ensure the subject has exactly two arms and natural joint articulation. Eliminate any hallucinations or morphing artifacts (e.g., extra limbs, floating hands) caused by complex motion descriptions.

Figure 11: Data Curation Results with Comprehensive Prompt.

editing stability and data quality. With the advancements in VLMs and their progressively enhanced image processing capabilities, we could design more comprehensive prompts and perform advanced pixel-level transformations in the future to acquire even more diverse data.



Figure 12: Failure Data Curation Case with Comprehensive Prompt.

A.4 Experiment

Fig 13, 14, 15, and 16 illustrate additional comparisons between our method and other 2D sewing pattern generation approaches, namely NeuralTailor [14] and Sewformer[25]. While NeuralTailor struggles with structural consistency in complex topologies and Sewformer fails to handle non-frontal poses, our method consistently reconstructs coherent, well-proportioned sewing patterns. By integrating canonicalized representations with original-view information, **DressWild** better preserves garment topology and symmetry. These results highlight our robustness to diverse in-the-wild conditions and the ability to generate simulation-ready patterns from challenging poses.

Our method demonstrates robustness and extensibility in processing complex in-the-wild images. As shown in Fig. 17 and 18, **DressWild** effectively handles diverse poses and viewpoints, generating pose-agnostic and simulation-ready sewing patterns from challenging real-world inputs. Beyond single garments, our framework naturally extends to multi-garment scenarios by leveraging VLM-based decomposition to extract individual layers while preserving their semantic consistency and relative ordering. This approach ensures topological fidelity across varying levels of garment complexity without requiring explicit layer annotations.

A.5 Further Discussion

Limitations and Failure Cases. Our method relies on VLM-based canonicalization, and inaccuracies in image editing may negatively affect the quality of sewing pattern reconstruction. In addition, due to limitations in the training data distribution, the model may struggle with garments containing fine-grained details or uncommon tailoring designs. For instance, because the dataset includes relatively few round-neck garment examples, the model may occasionally misclassify round-neck structures

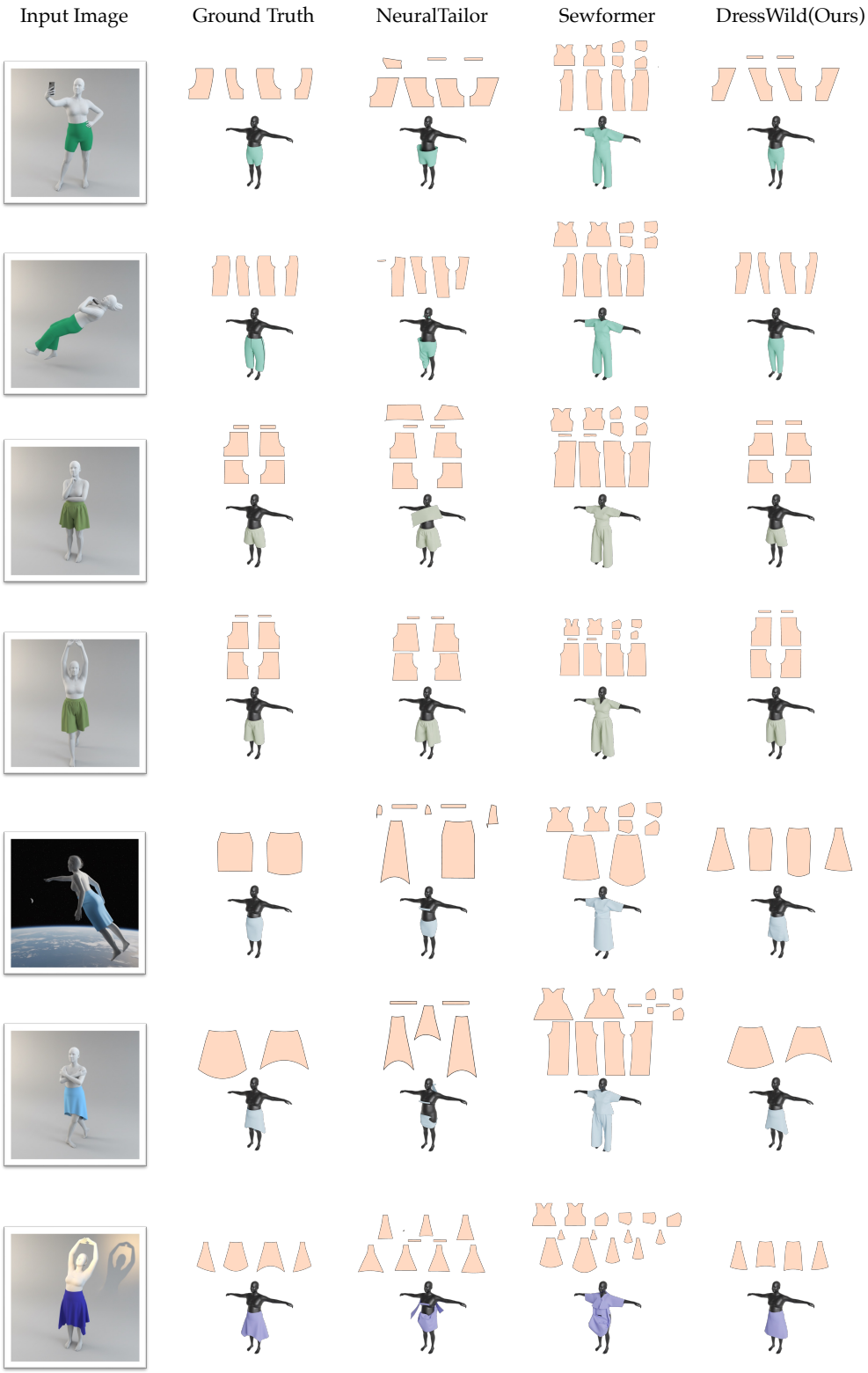

Figure 13: Comparison with baseline methods (1).

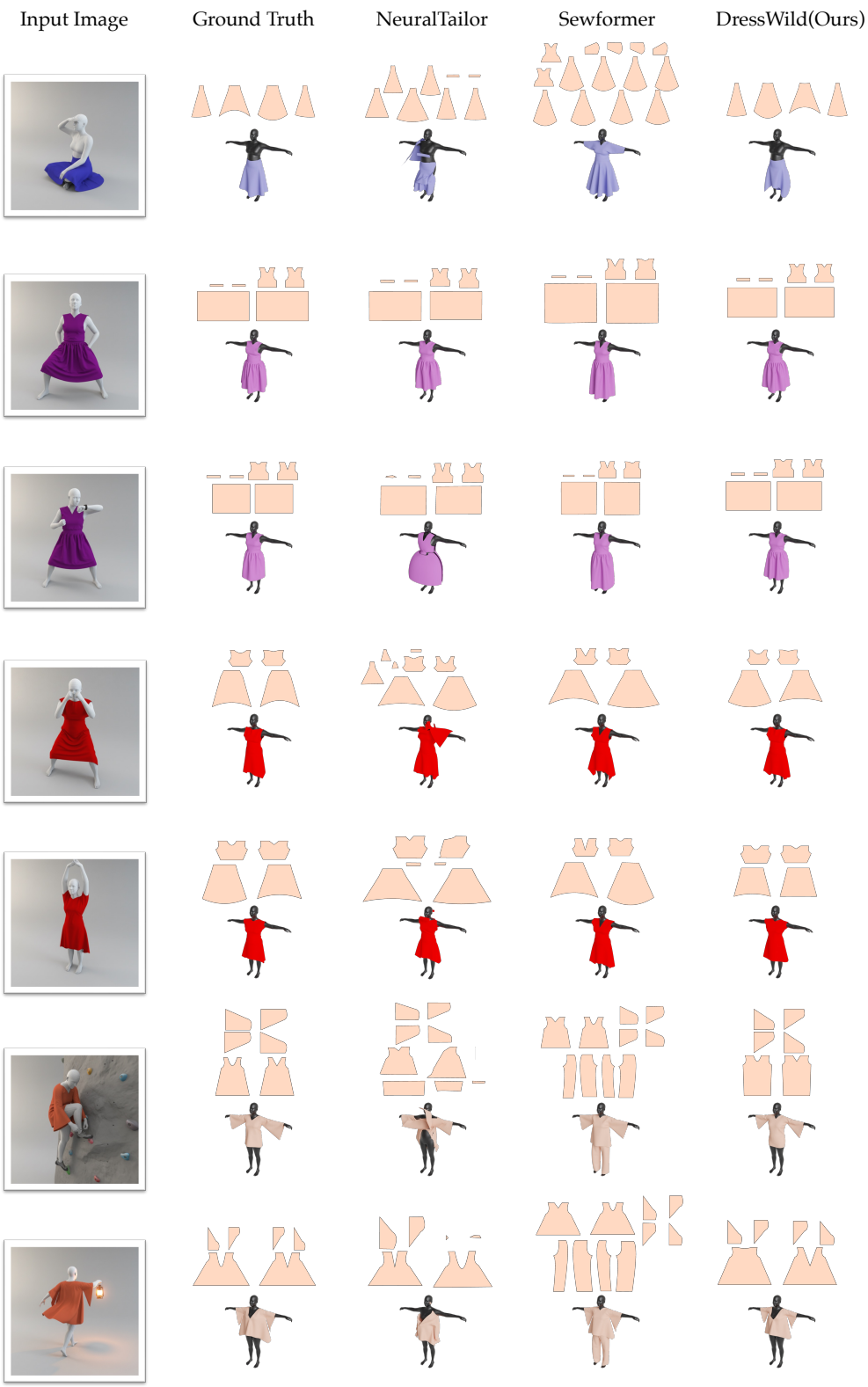

Figure 14: Comparison with baseline methods (2).



Figure 15: Comparison with baseline methods (3).

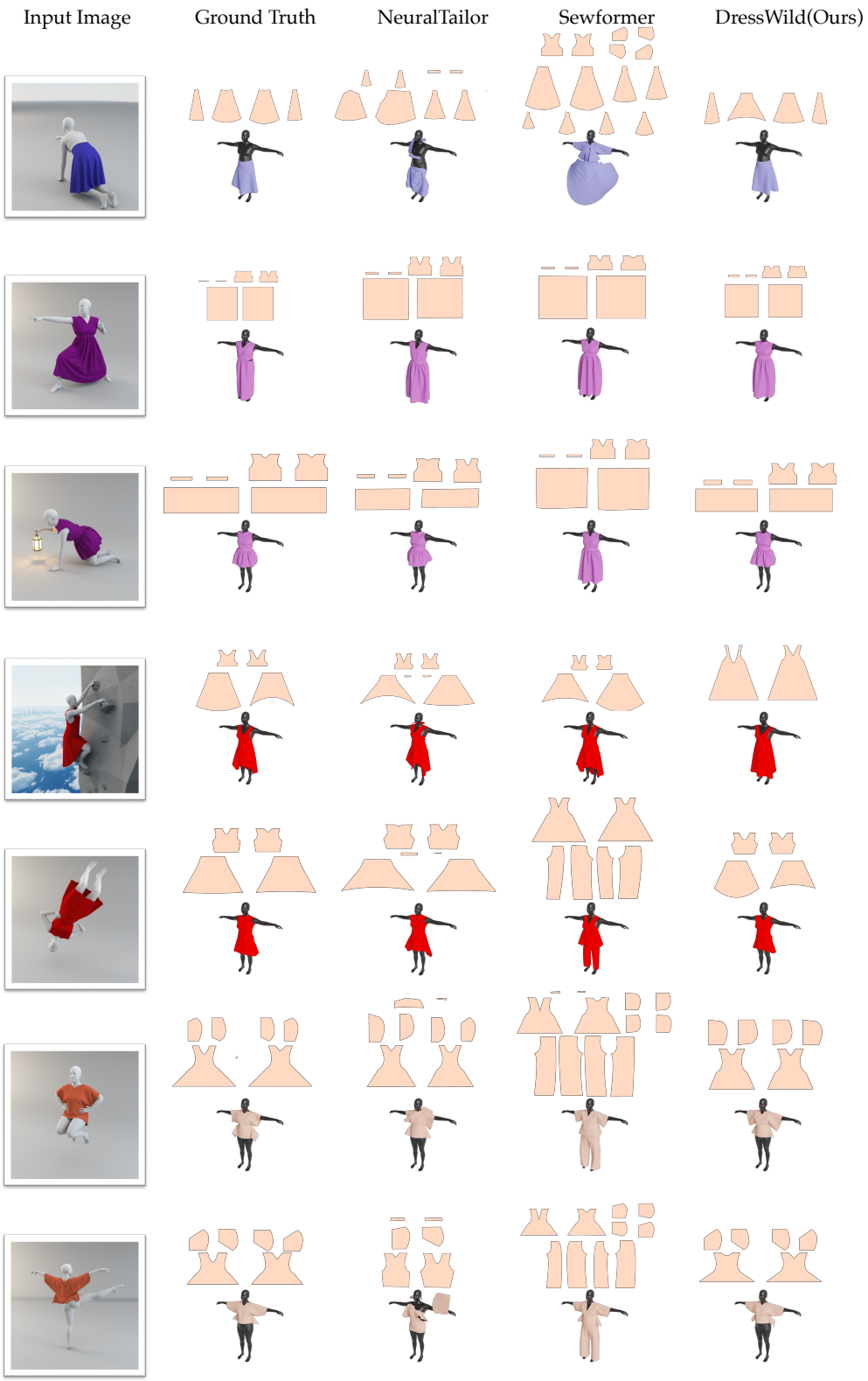

Figure 16: Comparison with baseline methods (4).



Figure 17: More garment generation results from in-the-wild images.



Figure 18: More garment generation results from in-the-wild images.

as V-neck designs, as shown in the black sleeveless top worn by the male subject in Figure 17. Furthermore, since the current sewing pattern representation does not explicitly model pockets, details such as the chest pocket on the yellow jumpsuit worn by the female subject in Figure 17 are absent in the reconstructed result. Moreover, DressWild currently does not support the modeling of accessories such as hats, shoes, or other wearable items. Joint modeling of human body shape and garment sizing also remains an important direction for future work. In addition, the textures shown in our results are generated through a separate post-processing baking stage, whereas jointly predicting garment patterns and textures in an end-to-end framework could substantially improve both efficiency and consistency. We believe these limitations can be progressively mitigated by expanding the diversity of garment categories in the training dataset and leveraging more advanced VLMs in future research.

Broader Impact. Our work has potential positive impacts on digital fashion design, virtual try-on, garment animation, and fabrication-aware content creation by reducing the technical barrier for creating simulation-ready garments from images. The proposed framework may also benefit applications in gaming, film production, e-commerce, and personalized apparel design. At the same time, the method could potentially be misused to generate unauthorized digital replicas of existing clothing designs or copyrighted fashion assets. In addition, inaccuracies in reconstruction under challenging poses or occlusions may produce unrealistic garment structures. We encourage responsible usage of generated assets and compliance with applicable copyright and data usage regulations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe DressWild as a feed-forward framework for pose-agnostic sewing pattern generation from a single image. The claimed contributions are consistent with the quantitative and qualitative results presented in the experiments.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in the conclusion section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper fully disclose all the information needed to reproduce the main experimental results of the paper, in experimental settings.

   Guidelines:

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

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

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

Answer: [No]

Justification: We will consider releasing code and model checkpoints upon acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The paper specify all the experimental setting and details.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: The paper doesn't report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper report the experiments compute resources.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work focuses on garment reconstruction and sewing pattern generation. The method is intended for virtual garment modeling and simulation.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The work has positive impacts for digital garment design, virtual try-on, animation, and fabrication-aware modeling.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the existing datasets, models, and tools used in the pipeline.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The assets in this paper are well introduced.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not conduct human-subject research or crowdsourcing studies.

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The paper explicitly describes the use of VLMs for canonical image generation, data curation, and garment-layer processing.

    Guidelines:

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