Multimodal Latent Diffusion Model for Complex Sewing Pattern Generation

Shengqi Liu[1], Yuhao Cheng[1], Zhuo Chen[1], Xingyu Ren[1], Wenhan Zhu[2],
Lincheng Li[3], Mengxiao Bi[3], Xiaokang Yang[1], Yichao Yan[1*]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
[2]Xueshen AI, [3]NetEase Fuxi AI Lab

Figure 1. SewingLDM can generate complex sewing pattern designs under the condition of texts, garment sketches, and body shapes, demonstrating detailed control ability. The generated garments can be seamlessly integrated into the CG pipeline for simulation and animation, achieving vivid and photo-realistic rendering results.

Abstract

Generating sewing patterns in garment design is receiving increasing attention due to its CG-friendly and flexible-editing nature. Previous sewing pattern generation methods have been able to produce exquisite clothing, but struggle to design complex garments with detailed control. To address these issues, we propose ***SewingLDM****, a multi-modal generative model that generates sewing patterns controlled by text prompts, body shapes, and garment sketches. Initially, we extend the original vector of sewing patterns into a more comprehensive representation to cover more intricate details and then compress them into a compact latent space. To learn the sewing pattern distribution in the latent space, we design a two-step training strategy to inject the multi-modal conditions,* i.e.*, body shapes, text prompts, and garment sketches, into a diffusion model, ensuring the generated garments are body-suited and detail-controlled. Comprehensive qualitative and quantitative experiments show the effectiveness of our proposed method, significantly surpassing previous approaches in terms of complex garment design and various body adaptability. Our project page: https://shengqiliu1.github.io/SewingLDM.*

1. Introduction

Clothes play a pivotal role in shaping human aesthetics and physique, where appropriate clothes can beautify their overall appearance and highlight human physical attributes. Therefore, garment design has always been a crucial component that significantly impacts both digital character creation and real-life human visual presentation. In recent years, many garment generation methods [4, 9, 20, 25–27, 29, 33, 37, 42, 43, 52, 53, 61, 71] have emerged to generate desired garments for users through various conditions.

*Corresponding author

Typically, garment generation can be classified into 2D and 3D methods. The 2D generation methods [4, 25, 42, 43] can produce visually appealing results but cannot maintain consistency across different views, failing to drape on human bodies. Therefore, many recent works are focusing on 3D cloth generation [29, 52]. Although these methods can generate high-quality mesh or neural field, they pose the challenge of clipping between clothes and bodies when draping clothes onto the human body, and they are incompatible with the digital garment production pipeline. Meanwhile, the sewing pattern is a more widely used representation for garments in the industry because it facilitates both physical simulation and animation in CG-friendly fashions [3, 7]. Current methods for sewing pattern generation [20, 26, 33] achieve fantastic garments generation but fall short in designing complex features, such as gathers and darts. Besides, these methods typically ignore human body shapes, preventing the creation of the made-to-measure garment. Apart from learning-based methods, parametric garment design tool [27] allows users to model complex garments and considers their relations with human body shapes. Nonetheless, this tool requires pre-defined templates and a delicate selection of control parameters, which require professional knowledge of garment designs, limiting its widespread promotion. In summary, there are two main challenges in suitable sewing pattern generation: 1) Designing a more general representation for complex designs of sewing patterns, and 2) Enabling control over garment details and ensuring the garments are body-suited.

To address these issues, we design a novel architecture named **SewingLDM** for generating complex sewing patterns under the control of texts, body shapes, and garment sketches. To represent complex designs of sewing patterns, we especially design an extended representation to encompass intricate types of edges and attachments of garments, enabling more general and complex garment learning. Subsequently, we train an auto-encoder model to compress the representation into a compact latent space for easier training while maintaining high reconstruction quality. To achieve multi-modal controlled and body-aware sewing pattern generation, we design a two-step training strategy to introduce the control signals into the latent diffusion model. In the first step, we train the diffusion model under the condition of texts, serving as a coarse fundamental model for additional control signal injection. In the second step, we further embed the knowledge of sketches and body shapes into the diffusion model by fusing the features after the first block and fine-tuning the output layers within the attention module, providing additional control of garment details and ensuring the generated garments fit various body shapes. Based on the proposed framework, SewingLDM can generate complex garments that fit various body shapes and align with user-provided text descriptions or garment sketches.

Our generated sewing patterns can be seamlessly integrated into subsequent CG pipelines, facilitating editing and animation processes. After simulation, the garment mesh can be combined with current texture generation methods [36, 63, 66, 69] or handcrafted texture to generate colored garments, as shown in Fig. 1, demonstrating our fantastic generation ability. Comprehensive qualitative and quantitative experiments show the superiority of our proposed method in terms of complex garment design and various body adaptability compared with previous methods. To summarize, our main contributions include:

- We design a novel architecture, dubbed SewingLDM, for sewing pattern generation conditioned by texts, body shapes, and garment sketches, enabling precisely controlled and body-suited garment generation.
- We design an extended representation to cover complex sewing patterns and compress it into a compact latent space for easier training of the generation model.
- We design a two-step training strategy to better inject the multi-modal control signals into a diffusion model, yielding superior generation performance and controllability.

2. Related Work

Multimodal-guided 3D Generation. The advancements in large models [1, 49, 51] have encouraged the emergence of recent 3D generation models [11, 21, 23, 34, 41, 57, 68], entering a new era of 3D generation. Some models [8, 37, 40, 47, 55, 65] focus on generating implicit neural radiance fields corresponding to the text description, whereas others [11, 56] extend their ability to generating 3D meshes with BRDF materials. Especially, some models [29, 37, 52, 53] dive into human clothes generation. Garment3DGen [52] can generate textured 3D mesh from images or text conditions by adjusting template meshes. GarmentDreamer [29] utilizes 3D Gaussian Splatting (GS) [24] as guidance to create 3D garment meshes from textual prompts. WordRobe [53] leverages unsigned distance field (UDF) [18] to represent 3D garments and generate 3D garment meshes under text guidance. GarVerseLOD [37] proposes a hierarchical framework to recover different levels of garment details through single images. Although these methods can generate visual-appealing garment meshes, their compatibility with CG pipelines remains a challenge, hindering seamless integration into modern industry workflows. In contrast, our method aims at sewing pattern generation, facilitating utilization in CG processes and daily clothing factories.

3D Sewing Pattern Modeling. Existing fashion CAD software tools, such as Clo3D [14] and Marvelous Designer [15], allow users to edit sewing patterns and simulate desired cloth outcomes. While these methods integrate the most advanced garment design, they heavily rely on artists to manually draw and adjust the shapes of sewing patterns,

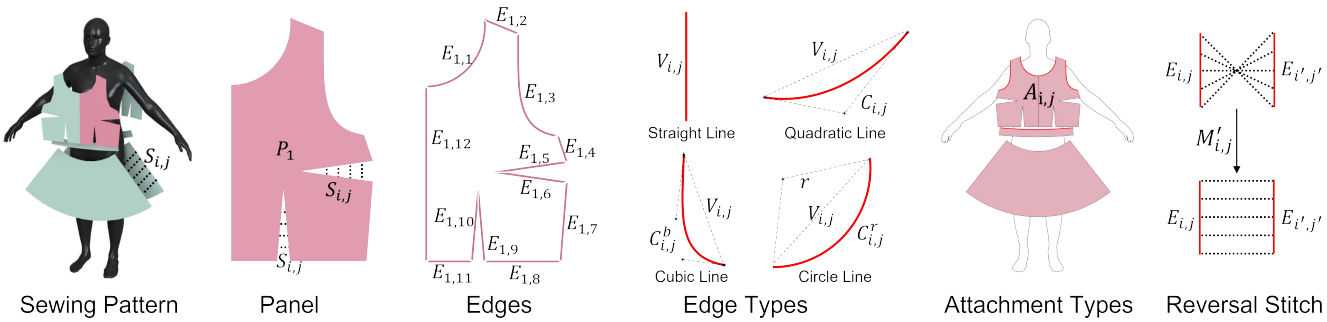

Figure 2. **Sewing pattern.** Sewing patterns are CAD representations of garments, containing 2D shapes and 3D placement of cloth. They consist of panels, and panels consist of edges joined from beginning to end. Between panels or inner panels, stitches are used to connect edges to form clothes. For each edge, different kinds of lines are utilized to conform to body contours. Additionally, complex sewing patterns need additional attachment constraints during simulation for certain edges, which is highlighted in red in the attachment types region. Besides, for the stitch between panels and inner panels, a reversal stitch flag is sometimes needed to reverse the stitch direction.

requiring a substantial of professional manual processing. Consequently, ongoing studies are focusing on automating the adjustment of sewing patterns [6, 16, 32, 38, 46, 58, 60], reconstructing sewing patterns [5, 12, 19, 22, 26, 33, 62], and assisting with complex garment design [13, 17, 30, 31, 59]. Recent studies [20, 26, 27, 33] on sewing patterns start to focus on autonomously generating diverse sewing patterns through different conditions rather than merely adjusting or producing a single garment. One of the recent SOTA methods, DressCode [20], first generates garments through natural language and yields visual-appealing appearances. However, the capability of DressCode is limited in modeling complex sewing patterns, and furthermore, it does not consider the relation between garments and body shapes, difficult to drape on various bodies. Another typical work, parametric sewing pattern [27], can control complex sewing patterns, while it requires predefined templates and a delicate selection of different control scale values, which is not user-friendly. Different from these methods, our SewingLDM not only has the ability to represent complex sewing patterns, but can also generate sewing patterns based on multi-modal intuitive conditions, *i.e.*, natural language, garment sketches, and body shape. These capabilities enable easily creating tailored garments that conform precisely to individual body shapes.

3. Method

To generate garments suited for various humans, we introduce **SewingLDM**, a latent-based diffusion model, to create complex 3D sewing patterns, conditioned by personalized body shapes, text prompts, and garment sketches. We first review the original sewing pattern representation [26] (Sec. 3.1) and then we improve it with special designs to cover complex sewing patterns (Sec. 3.2), as illustrated in Fig. 2. Subsequently, we compress the sewing pattern representation into a compact latent space for easier training of the generation model and reducing computational resources in Fig. 3 (Sec. 3.3). Finally, we train a latent diffusion model under multi-modal conditions through a two-stage training strategy, as illustrated in Fig. 4 (Sec. 3.4). Based on the proposed framework, SewingLDM can generate complex garments based on the body shape and align with the user-provided text description or garment sketches.

3.1. Preliminaries on Sewing Pattern

Sewing patterns are CAD representations of garments, representing 2D shapes and 3D placement of cloth, as shown in Fig. 2. Neuraltailor [26] first transfers sewing patterns into vector representations as inputs of the neural network. The sewing pattern contains N_p panels $\{P_i\}_{i=1}^{N_p}$, with each panel P_i including N_i edges $\{E_{i,j}\}_{j=1}^{N_i}$. For each edge $E_{i,j}$, a vector $\{V_{i,j}\}_{j=1}^{N_i} \in \mathbb{R}^2$ is utilized to represent the direction from its starting to ending point. In Neuraltailor [26], sewing patterns only have two kinds of edges, *i.e.*, straight lines and quadratic lines. Quadratic lines use two additional parameters $C_{i,j} = (c_x, c_y)$ representing the control point of the Bezier curve. Rotation $R_i \in SO(3)$ and translation $T_i \in \mathbb{R}^3$ are utilized to represent the 3D placement of each panel P_i. Moreover, to depict the stitching connecting each inner or outer panel edge, it incorporates per-edge stitch tags $\{S_{i,j} \in \mathbb{R}^3\}_{j=1}^{N_i}$ and stitch masks $\{M_{i,j} \in \{0, 1\}\}_{j=1}^{N_i}$. The stitch tag $S_{i,j}$ is determined by the 3D position of the corresponding edge, which utilizes the Euclidean distance between edges as a measure of stitch similarity. The stitch mask $M_{i,j}$ is a binary flag to indicate whether there are stitches on the edge.

3.2. Extended Representation for Sewing Pattern

For more complex garment designs, the modern industry will use special designs to make garments more fashionable, as illustrated in Fig. 2. Original sewing pattern representation in [26] can not cover complex garments with

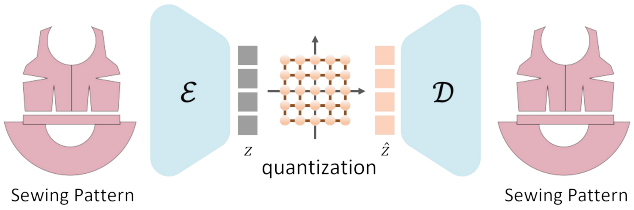

Figure 3. **Sewing pattern compression.** We compress the sewing pattern representations into a bound and compact latent space.

more kinds of curve lines, *i.e.*, cubic and circle lines, and additional attachment constraints in collars and waistbands to prevent cloth sliding from human bodies for certain garments, *e.g.*, strapless tops, and loose pants. Moreover, the stitches may intersect for some edge pairs, causing errors during simulation. To precisely represent complex clothes, we extend each edge feature to high dimensions to cover complex patterns. Then we preprocess them into a uniform tensor shape to feed into the neural network.

Representation. For each edge $E_{i,j}$, we append origin control parameters $C_{i,j}$ with cubic line control parameters $C^b_{i,j} \in \mathbb{R}^4$, representing two control points, and circle line control parameters $C^r_{i,j} \in \mathbb{R}^3$, which represents the radius r and the rotation angle, to cover more kinds of curve lines. Further, we use two binary flags $\{E^t_{i,j,k} \in \{0,1\}\}^2_{k=1}$ to denote these 4 different edge types. Moreover, 3 specific binary flags $\{A_{i,j,k} \in \{0,1\}\}^3_{k=1}$ are included to indicate the attachment type for certain edges, such as those associated with the collar and waistband, to prevent garments from sliding during simulation. We add one binary flag to $M_{i,j}$ as the new stitch mask $\{M^{'}_{i,j,k} \in \{0,1\}\}^2_{k=1}$ to additionally denote whether the stitch direction needs to be reversed to prevent stitch intersections.

Preprocessing. Before input to the neural networks, the vector representation needs to be the same size for all data during training. For each edge $E_{i,j}$, we concatenate all extended parameters and append with rotation R_i and translation T_i of panel P_i to form the high dimensional edge feature $E^f_{i,j}$. Furthermore, we design a binary flag $\{E^m_{i,j} \in \{0,1\}\}^{N_i}_{j=1}$ to denote the existence of each edge. All features are concatenated to form a 29-dimensional vector for each edge feature $E^f_{i,j}$, represented as follows:

$$E^f_{i,j} = V_{i,j} \oplus C_{i,j} \oplus C^b_{i,j} \oplus C^r_{i,j} \oplus S_{i,j} \oplus R_i \oplus T_i \oplus E^t_{i,j} \oplus E^m_{i,j} \oplus A_{i,j} \oplus M^{'}_{i,j}, \quad (1)$$

where i is in the range of $[1, max(N_p)]$, j is in the range of $[1, max(N_i)]$.

Then, all edge features $\{E^f_{i,j}\}^{N_i}_{j=1}$ are concatenated and padded with 0 to max edge number to get the representation of panels. Further, we concatenate all panels and pad with 0 to max panel number to get the representation of sewing pattern F, in the shape of $(max(N_p) \times max(N_i), 29)$. Before input to the neural network, all continuous values are standardized, and all binary flags are transformed into $\{-1, 1\}$.

3.3. Compact Latent for Sewing Pattern

The vector representation F inevitably incorporates redundant information as the panel and edge numbers increase, preventing generation models from learning the distribution of F. As indicated by the recent compression methods [39, 64, 70], it is necessary to compress F into a compact latent space and maintain the reconstruction quality. Following this idea, we train an auto-encoder to compress and quantize the F into a latent space where each dimension is bounded within the range $[-1, 1]$, as illustrated in Fig. 3. The sewing pattern representation F is encoded to z by the encoder $\mathcal{E}$ and quantized to $\hat{z}$ in the constrained latent space, subsequently reconstructed by the decoder $\mathcal{D}$. The process can be represented as:

$$z = \mathcal{E}(F_{\text{gt}}), \hat{z} = \frac{round(n \times tanh(z))}{n}, F_{\text{rec}} = \mathcal{D}(\hat{z}), \quad (2)$$

where n is an integer used to modify the spacing between each $\hat{z}$ in the latent space.

For training the encoder $\mathcal{E}$ and decoder $\mathcal{D}$, we combine the loss in previous works [26, 39] with additional binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ to constrain the newly incorporated binary flags:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{panel}} + \lambda_3 \mathcal{L}_{\text{stitch}} + \lambda_4 \mathcal{L}_{\text{BCE}}, \quad (3)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters to balance each loss term. $\mathcal{L}_{\text{rec}}$ is the MSE loss to keep reconstruction quality, while $\mathcal{L}_{\text{panel}}$ and $\mathcal{L}_{\text{stitch}}$ proposed in [26] to ensure the integrity of the garments.

After training, the sewing pattern representation F can be efficiently compressed into a bounded and compact latent space without compromising important information. Moreover, to facilitate the learning of generation models, each dimension of the latent is evenly distributed within the coordinates $\{-1, -0.5, 0, 0.5, 1\}$ by setting $n = 2$ in Eq. (2).

3.4. Multimodal Conditions of Diffusion Model

Inspired by the great power of controlled generation in the diffusion model, we employ latent diffusion [51] as our generation model. Our generation model is based on the DiT architecture [10, 45], which is scalable to different sizes of sewing patterns. To balance multi-modal conditions and facilitate future conditional scalability, we design a two-step training strategy: 1) In the first step, we train the latent diffusion model with IDDPM loss [44] only under the text guidance extracted by T5 tokenizer [50]; 2) In the second step, we embed the knowledge of body shapes and garment sketches into the diffusion model for detailed control and body-suited garment generation, as depicted in Fig. 4.

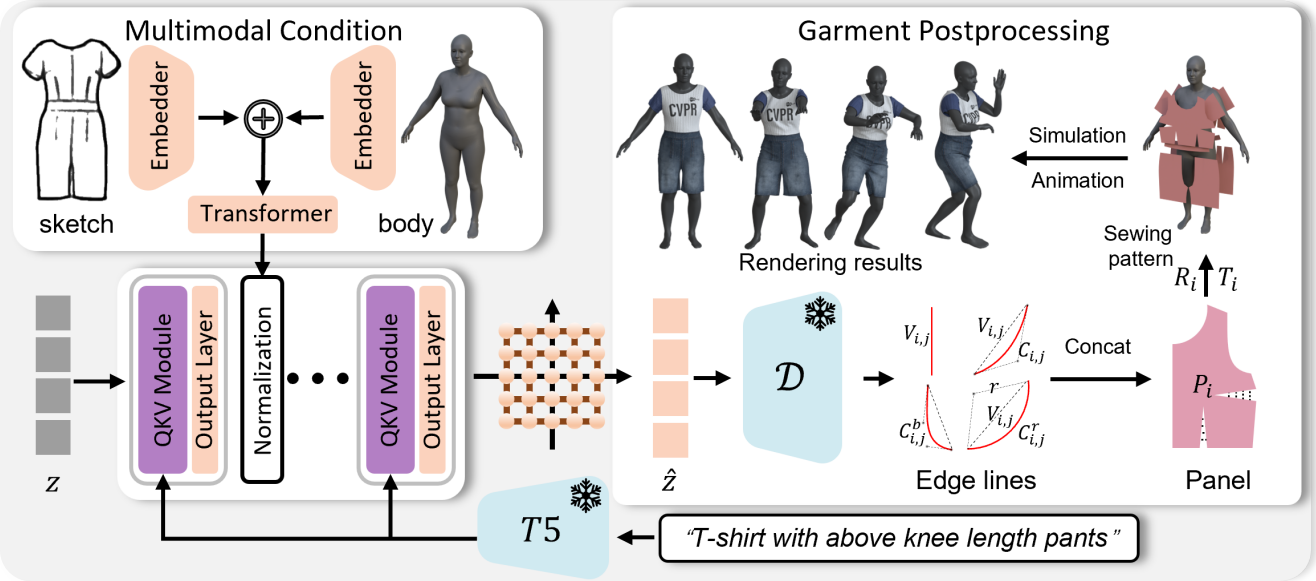

Figure 4. **Multimodal latent diffusion model.** After training the text-guided diffusion model, we fuse the features of sketches and body shapes and normalize them into the diffusion model, with fine-tuning minimal parameters of the diffusion model. The trained network parameters are depicted in orange, while the frozen parameters are shown in purple. The output latent is then quantized into a designed latent space and serves as the input of the decoder to yield all edge lines. Edge lines connect from beginning to end to form panels, placed on the corresponding body regions. Finally, we can get suited garments through the modern CG pipeline.

The text-guided latent diffusion model serves as a fundamental model for extensive multi-modal conditions injection. For the injection of body shapes and garment sketches, a naive idea is to inject them through two ControlNet [67] branches. While sketches will change along with body shape, *e.g.*, sketches will get wider when bodies grow fatter. We propose to first use embedders to extract the feature from sketches and body shapes, and then simply concatenate them together, and input it into a light transformer layer. During the light transformer layer, the features of sketches and body shapes can be thoroughly fused and get the relation between each other through self-attention modules and output as F_{bs}. Then we normalize mean $\boldsymbol{\mu}_{bs}$ and variance $\boldsymbol{\sigma}_{bs}$ of $\boldsymbol{F}_{bs}$ into the same mean $\boldsymbol{\mu}_z$ and variance $\boldsymbol{\sigma}_z$ with the latent features $\boldsymbol{F}_z$ and add them together during the middle block of the diffusion model.

$$\hat{\boldsymbol{F}}_z = \frac{(\boldsymbol{F}_{bs} - \boldsymbol{\mu}_{bs}) \times \boldsymbol{\sigma}_{bs}}{\boldsymbol{\sigma}_z + \epsilon} + \boldsymbol{\mu}_z + \boldsymbol{F}_z, \tag{4}$$

where ϵ is a small constant for numerical stability. After the normalization, we assume the new $\hat{\boldsymbol{F}}_z$ is similar to $\boldsymbol{F}_z$, which does not need to retrain the whole diffusion model in the second stage. We only fine-tune the output layer of the attention modules in each DiT block to transform the normalized features into the desired distribution. After two-stage training, our generation model can precisely follow the text guidance and sketches under various human shapes, enabling more body-suited and detailed controlled garment generation for individuals.

4. Experiments

4.1. Experiment Setup

Dataset. To train a generation model under the condition of texts or sketches, it is essential to acquire the corresponding paired data. We extend the current dataset [28] with additional textual annotations and garment sketches. The dataset [28] consists of 120,000 sewing patterns, covering a variety of clothing styles for different body types. Sewing patterns in [28] consider the relationship between various body shapes and garments, resulting in garments that are well-tailored to individual body types. Building on [28], we annotate each garment with text prompts according to its design parameters file, resulting in detailed text annotations. However, relying solely on textual descriptions may not precisely dictate garment shapes, potentially yielding undesirable outputs. To enhance control over the generation, we propose to generate more rich annotations like sketches. For each garment, we utilize PiDiNet [54], a pre-trained edge detection network, to extract garment sketches, thereby enriching the design details of the garment.

Implementation Details. We train our model on 4 RTX A6000 GPUs with 48G memory, where the auto-encoder requires 12 hours for training. The hyperparameters for training the auto-encoder, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are set as $5, 1, 1, 1$. Training the text-guided latent diffusion model takes 2 days in the first stage, and training the multi-modal conditions requires an additional 10 hours to reach convergence in the second stage. During the second stage, the sketch or text

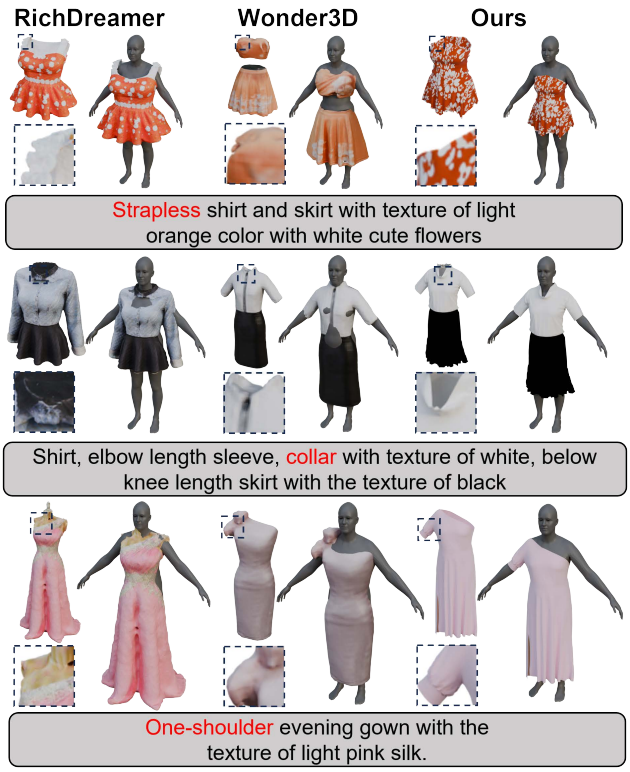

Figure 5. **Comparison with 3D mesh generation method.** We present the garments and draping results for each method. Our method successfully generates modern design garments with remarkable visual quality and close fitting to various body shapes. In contrast, Wonder3D [35] and RichDreamer [48] only generate close-surface meshes and contain obvious artifacts, resulting in human bodies clipping through the garments.

conditions are set to zero with a probability of 0.25 to ensure the model retains the capability to generate desired garments based on a single input condition.

4.2. Qualitative Comparison

We conduct qualitative comparisons with SOTA mesh generation methods and sewing pattern generation methods, respectively, to demonstrate our CG-friendly and superior generation results for various body shapes.

Comparison on 3D Mesh Generation Methods. We compare our SewingLDM with the current SOTA 3D mesh generation methods, *i.e.*, Wonder3D [35] and RichDreamer [48] with the same text prompts containing both the geometry and texture information. Note that, RichDreamer [48] is an image-guided generation method; thus, we feed the sketches and text prompts to ControlNet [67] with Stable Diffusion [51] to generate corresponding images. As the qualitative comparisons shown in Fig. 5, Wonder3D and RichDreamer can both generate garments aligned with text prompts. However, both of them are close-surface meshes and do not consider human body shapes, resulting in obvious clipping when draping on human bodies. In contrast, our method generates sewing patterns for various human bodies through two-stage training, which are easy to drape on human bodies and maintain the physical properties of clothing with fantastic clothes wrinkles. The results show that our garments are all well-fitted with complex geometry and aligned with the conditions.

	RichDreamer	Wonder3D	Sewformer	Dresscode	Ours
Runtime ↓	∼ 4 mins	∼ 4 hours	**∼ 3 mins**	**∼ 3 mins**	**∼ 3 mins**
Clothes-to-body Distance ↓	6.19 cm	6.54 cm	5.45 cm	3.69 cm	**2.20 cm**
Users Values ↑	1.89	1.88	2.10	3.56	**4.60**

Table 1. **Quantitative Comparison.** We compare the generation efficiency and the average clothes-to-body distance. Further, we conduct a comprehensive user study to judge the superiority of different methods. All metrics show our method generation superior results than other methods.

Comparison of Sewing Pattern Generation Methods. We also compare our method with current SOTA sewing pattern generation methods, *i.e.*, DressCode [20] and Sewformer [33]. DressCode and our SewingLDM are fed with the same text descriptions, and we use ControlNet to generate the corresponding images as the input of Sewformer [33]. Additionally, the sketches used to generate images are utilized in our SewingLDM. Furthermore, the generative sewing patterns are draped onto two different body shapes to validate the effectiveness of body-aware garment generation. As illustrated in Fig. 6, both Sewformer and DressCode fail to generate textual-aligned garments due to the complex garment descriptions. Moreover, they can not be worn to diverse body shapes, sliding from the body or just failing to simulate the results. In contrast, our method can generate complex sewing patterns, *e.g.*, mermaid skirt hem, one-shoulder gown, and circle neckline, and fit to various body shapes, which provides a more user-friendly approach to getting the desired made-to-measure garments. The results demonstrate the superiority of our method in complex garment design and various body adaptability.

4.3. Quantitative Comparison

Besides qualitative comparisons, we also perform quantitative comparisons with these SOTA methods, evaluating aspects including generation efficiency, clothes-to-body distance, and user study. The clothes-to-body distance is to indicate whether the clothes are close-fitted to body shapes calculated by averaging the minimum distance from each point in garments to human bodies. Besides, we also perform a user study to further assess the quality of garment generation. We take 10 text prompts to generate diverse garments and render the generated results draping on dif-

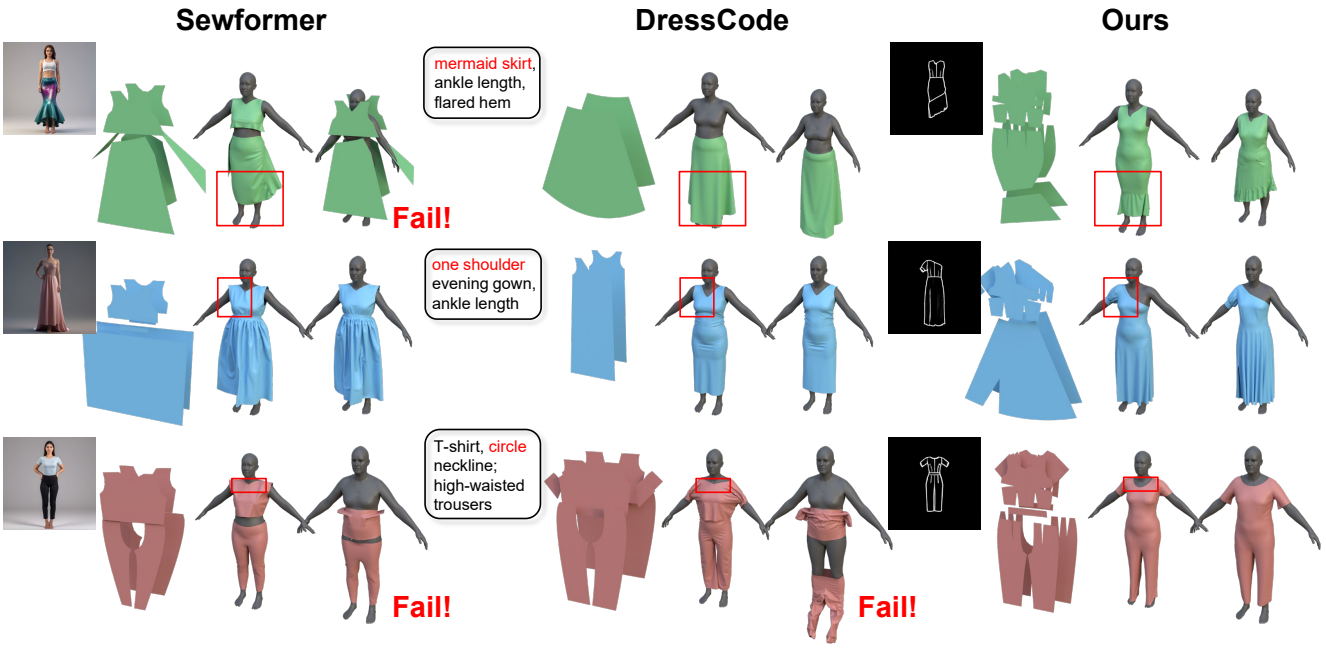

Figure 6. **Comparison of sewing pattern generation for various body shapes.** For each method, we present corresponding conditions and generated results, including sewing patterns, draping results on average body shape, and draping results on another body shape. Our method can generate complicated sewing patterns aligned with sketch conditions and text prompts, draping on various body shapes. In contrast, Sewformer [33] and DressCode [20] both fail to generate complex and body-suited garments.

ferent body shapes for each method. Then we ask 30 users to give a value of these rendering results with comprehensive consideration for two aspects: 1) consistency with text descriptions; and 2) well-fitting with human bodies. As illustrated in Tab. 1, the preference results demonstrate a notable superiority of our method over SOTA approaches in both aspects, highlighting in generating garments that are both well-suited to various bodies and exhibit high fidelity aligned with text descriptions.

4.4. Ablation Study

Sewing Pattern Compression. To maintain the reconstruction ability and dense compression of sewing patterns in the meantime, we have explored numerous parameters for the compression network, as shown in Tab. 2. We measure the reconstruction ability, generation ability, clothes-to-body distance, and codebook usage under different settings. The codebook usage is calculated by the number of used latent N_U dividing the latent number in latent space N_L as follows:

$$codebook\ usage = \frac{N_U}{N_L} = \frac{N_U}{(2n+1)^{n_f}}, \tag{5}$$

where n is an integer in pre-defined Eq. (2) for quantization, n_f is the last dimension length of latent. As illustrated in Tab. 2, without compression or lower compression, the latent space is inappropriate for the generation model to learn the distribution of latent. In contrast, with compact latent space, the latent is fully utilized, resulting in a well-generation ability and various body adaptability.

Compression shape	w/o compression	256*32 n=32	256*12 n=8	256*8 n=2	256*6 n=2	256*4 n=2
Reconstruction	✓	✓	✓	✓	✓	×
Generation	×	×	×	✓	✓	-
Clothes-to-body Distance ↓	-	-	-	2.87 cm	**2.20 cm**	-
Codebook usage	-	0%	0%	91%	100%	-

Table 2. **Different Compression.** We try different compression shapes for the latent and set the different values of n. ✓means it can well do this task, while × means it fails in this task.

Multi-modal Controllable Generation. In the context of injecting body shape and garment sketch conditions, we perform ablation studies on the optimized parameters of the output layers across different attention modules, *i.e.*, both self-attention and cross-attention, self-attention only, and cross-attention only. By default, we inject the additional condition after the first transformer block. Moreover, we investigate the impact of different injection positions, specifically after block 5, block 10, block 15, and block 20, as illustrated in Fig. 7. Notably, optimizing in both attention results in more desired circle necklines than only optimizing in cross-attention and is better aligned

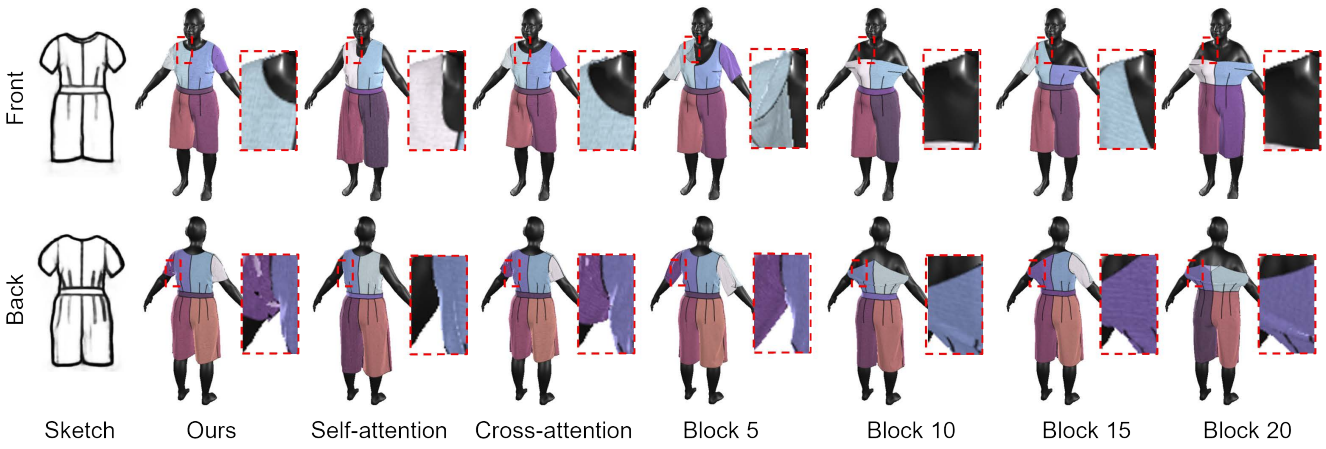

Figure 7. **Abalation of the multi-modal condition.** We have taken an ablation experiment on the training parameters of the output layer in different attention modules. We also explore the relationship between results and injection positions.

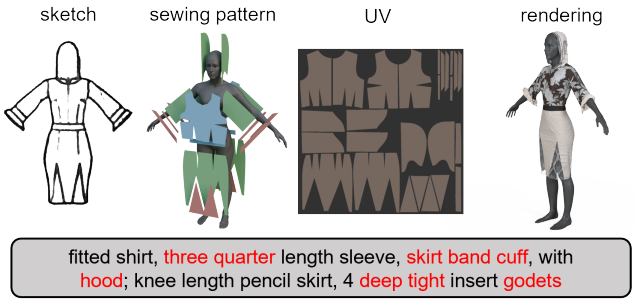

Figure 8. **Use case.** We present a example for an extremely complex sewing pattern generation. With the generated sewing pattern, we can easily paint the UV to produce visually appealing results.

Figure 9. **Limitations.** For intricate sketches, such as bridal gowns or additional accessories like pockets and zippers, our method may fail to generate the desired garments.

with sketches compared with only training output layers in self-attention, which fails to generate the desired neckline and sleeves. This shows that the additional conditional can closely resemble the latent features, which needs more learning across text prompts and the latent, rather than within the latent alone. Consequently, during the ablation study on injection positions, we fine-tune the output layers of both attention for better results. We observe that as the layer depth increases, the garment gradually loses key components, *e.g.*, sleeves or waistband, resulting in unable draping on the human body. In contrast, injecting the condition at shallower layers facilitates better fusion of the additional condition with the combined latent feature, leading to more accurate results. In summary, we train the output layers of both attention and inject the additional condition after block 0, which yields optimal results.

4.5. Use Case

Our SewingLDM is capable of generating intricately detailed clothing, meeting the current artistic demands for garment design across a wide range of styles, significantly advancing fashion garment design, and supporting everyday users in obtaining apparel tailored precisely to their needs. To demonstrate our superiority in garment generation, we present an extremely complex example of a sewing pattern in Fig. 8. With the detailed textual description and garment sketch, our method faithfully generates the complex sewing pattern, *e.g.*, skirt band cuff, hood, and godets, which significantly helps the artist in creating fantastic texture in UV space, *e.g.*, laces, leather pants, and hat brim.

5. Conclusion and Limitations

In conclusion, our SewingLDM can generate complex sewing patterns under the condition of text prompts, garment sketches, and body shapes. We propose an enhanced vector representation of sewing patterns and compress them into a bounded and compact latent space for more generalized garment designs and facilitating training of the diffusion model. To accommodate multi-modal conditioning and future conditions, we introduce a two-step training strategy. We first train a latent diffusion model only conditioned by text prompts. Subsequently, we incorporate the condition of garment sketches and body shapes by optimizing the output layers of the attention modules while maintaining the responsiveness to text-based guidance. Finally, our generation model can conditioned by multi-modal input, resulting

in body-suited generation and detailed control of garments.

Despite the promising results, our method still has several limitations that should be addressed in future work, as illustrated in Fig. 9. The major limitation is that our method encounters challenges with certain modern designs, *e.g.*, zippers, and pockets. Another limitation is that it occasionally struggles with aligning complex sketches of intricate garments. Our further work aims to explore comprehensive representations of daily garments and expand the range of conditions applicable during the generation process.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Adobe. Substance 3D Painter. https://creativecloud.adobe.com/apps/all/substance3d-painter, 2024. 2

[3] Autodesk, INC. Maya. https://autodesk.com/maya, 2019. 2

[4] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *ICCV*, pages 23393–23402, 2023. 1, 2

[5] Seungbae Bang, Maria Korosteleva, and Sung-Hee Lee. Estimating garment patterns from static scan data. *Computer Graphics Forum*, 40(6):273–287, 2021. 3

[6] Aric Bartle, Alla Sheffer, Vladimir G. Kim, Danny M. Kaufman, Nicholas Vining, and Floraine Berthouzoz. Physics-driven pattern adjustment for direct 3D garment editing. *TOG*, 35(4):50–1, 2016. 3

[7] Blender Foundation. Blender. https://www.blender.org/, 2022. 2

[8] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *CVPR*, pages 958–968, 2024. 2

[9] Beijia Chen, Yuefan Shen, Qing Shuai, Xiaowei Zhou, Kun Zhou, and Youyi Zheng. Anidress: Animatable loose-dressed avatar from sparse views using garment rigging model. *arXiv preprint arXiv:2401.15348*, 2024. 1

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 4

[11] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, pages 22246–22256, 2023. 2

[12] Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. Garment modeling with a depth camera. *TOG*, 34(6):1–12, 2015. 3

[13] Pinaki Nath Chowdhury, Tuanfeng Wang, Duygu Ceylan, Yi-Zhe Song, and Yulia Gryaditskaya. Garment Ideation: Iterative View-Aware Sketch-Based Garment Modeling. In *International Conference on 3D Vision*, 2022. 3

[14] CLO Virtual Fashion. Clo3d. https://clo3d.com/en/, 2022. 2

[15] CLO Virtual Fashion. Marvelous Designer. https://www.marvelousdesigner.com/, 2024. 2

[16] Xudong Feng, Huamin Wang, Yin Yang, and Weiwei Xu. Neural-assisted homogenization of yarn-level cloth. In *SIGGRAPH*, pages 1–10, 2024. 3

[17] Amelie Fondevilla, Damien Rohmer, Stefanie Hahmann, Adrien Bousseau, and Marie Paule Cani. Fashion Transfer: Dressing 3D Characters from Stylized Fashion Sketches. *Computer Graphics Forum*, 40(6):466–483, 2021. 3

[18] Benoit Guillard, Federico Stella, and Pascal Fua. Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *ECCV*, 2022. 2

[19] Nils Hasler, Bodo Rosenhahn, and Hans Peter Seidel. Reverse engineering garments. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 200–211, 2007. 3

[20] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *TOG*, 43(4):1–13, 2024. 1, 2, 3, 6, 7

[21] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 2

[22] Moon-Hwan Jeong, Dong-Hoon Han, and Hyeong-Seok Ko. Garment capture from a photograph. *Computer Animation and Virtual Worlds*, 26(3-4):291–300, 2015. 3

[23] Nikolay Jetchev. Clipmatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2109.12922*, 2021. 2

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):139–1, 2023. 2

[25] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, pages 8176–8185, 2024. 1, 2

[26] Maria Korosteleva and Sung-Hee Lee. Neuraltailor: Reconstructing sewing pattern structures from 3d point clouds of garments. *TOG*, 41(4):1–16, 2022. 2, 3, 4

[27] Maria Korosteleva and Olga Sorkine-Hornung. Garmentcode: Programming parametric sewing patterns. *TOG*, 42 (6):1–15, 2023. 1, 2, 3

[28] 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 *ECCV*, 2024. 5

[29] 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. *arXiv preprint arXiv:2405.12420*, 2024. 1, 2

[30] Minchen Li, Alla Sheffer, Eitan Grinspun, and Nicholas Vining. FoldSketch: Enriching garments with physically reproducible folds. *TOG*, 37(4):1–13, 2018. 3

[31] Chen Liu, Weiwei Xu, Yin Yang, and Huamin Wang. Automatic digital garment initialization from sewing patterns. *TOG*, 43(4):1–12, 2024. 3

[32] Kaixuan Liu, Xianyi Zeng, Pascal Bruniaux, Xuyuan Tao, Xiaofeng Yao, Victoria Li, and Jianping Wang. 3D interactive garment pattern-making technology. *CAD Computer Aided Design*, 104:113–124, 2018. 3

[33] Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *TOG*, 42(6):1–15, 2023. 1, 2, 3, 6, 7

[34] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2024. 2

[35] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, pages 9970–9980, 2024. 6

[36] Ivan Lopes, Fabio Pizzati, and Raoul de Charette. Material palette: Extraction of materials from a single image. In *CVPR*, 2024. 2

[37] Zhongjin Luo, Haolin Liu, Chenghong Li, Wanghao Du, Zirong Jin, Yinyu Nie, Weikai Chen, and Xiaoguang Han. Garverselod: High-fidelity 3d garment reconstruction from a single in-the-wild image using a dataset with levels of details. *TOG*, 2024. 1, 2

[38] Yuwei Meng, Charlie C.L. Wang, and Xiaogang Jin. Flexible shape control for automatic resizing of apparel products. *CAD Computer Aided Design*, 44(1):68–76, 2012. 3

[39] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *ICLR*, 2024. 4

[40] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, pages 12663–12673, 2023. 2

[41] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, pages 13492–13502, 2022. 2

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

[43] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *ACMMM*, 2023. 1, 2

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 4

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, pages 4195–4205, 2023. 4

[46] Nico Pietroni, Corentin Dumery, Raphael Guenot-Falque, Mark Liu, Teresa Vidal-Calleja, and Olga Sorkine-Hornung. Computational pattern making from 3D garment models. *TOG*, 41(4), 2022. 3

[47] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[48] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *CVPR*, pages 9914–9925, 2024. 6

[49] 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 *ICML*, pages 8748–8763, 2021. 2

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. 4

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4, 6

[52] Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, and Rakesh Ranjan. Garment3dgen: 3d garment stylization and texture generation. *arXiv preprint arXiv:2403.18816*, 2024. 1, 2

[53] Astitva Srivastava, Pranav Manu, Amit Raj, Varun Jampani, and Avinash Sharma. Wordrobe: Text-guided generation of textured 3d garments. In *ECCV*, pages 458–475, 2024. 1, 2

[54] Zhuo Su, Jiehua Zhang, Longguang Wang, Hua Zhang, Zhen Liu, Matti Pietikäinen, and Li Liu. Lightweight pixel difference networks for efficient visual representation learning. *TPAMI*, 45(12):14956–14974, 2023. 5

[55] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, pages 22819–22829, 2023. 2

[56] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 2

[57] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, pages 3835–3844, 2022. 2

[58] Jin Wang, Guodong Lu, Weilong Li, Long Chen, and Yoshiyuki Sakaguti. Interactive 3D garment design with constrained contour curves and style curves. *CAD Computer Aided Design*, 41(9):614–625, 2009. 3

[59] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popović, and Niloy J. Mitra. Learning a shared shape space for multimodal garment design. *TOG*, 37(6):1–13, 2018. 3

[60] Katja Wolff, Philipp Herholz, Verena Ziegler, Frauke Link, Nico Brügel, and Olga Sorkine-Hornung. Designing Person-

alized Garments with Body Movement. *Computer Graphics Forum*, 2023. 3

[61] Donglai Xiang, Fabian Prada, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica Hodgins, and Timur Bagautdinov. Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing driven by sparse rgb-d input. In *SIGGRAPH Asia*, pages 1–11, 2023. 1

[62] Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C. Lin. Physics-inspired garment recovery from a single-view image. *TOG*, 37(5):1–14, 2018. 3

[63] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *CVPR*, pages 4347–4356, 2024. 2

[64] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *ICLR*, 2024. 4

[65] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 2

[66] Cheng Zhang, Yuanhao Wang, Francisco Vicente Carrasco, Chenglei Wu, Jinlong Yang, Thabo Beeler, and Fernando De la Torre. FabricDiffusion: High-fidelity texture transfer for 3d garments generation from in-the-wild images. In *ACM SIGGRAPH Asia*, 2024. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 5, 6

[68] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *TOG*, 43(4):1–20, 2024. 2

[69] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *TOG*, 43(4):1–18, 2024. 2

[70] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024. 4

[71] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *ECCV*, 2024. 1

Multimodal Latent Diffusion Model for Complex Sewing Pattern Generation

Supplementary Material

In this supplementary document, we first provide comprehensive details of complex sewing patterns (Sec. 6). Afterward, we compare with the parametric method [27], which needs a delicate selection of values and professional knowledge of garment designs (Sec. 7). To substantiate the efficacy of our two-step training strategy, we perform an ablation study on the training strategy compared with the one-step training strategy (Sec. 8). Additionally, we provide examples of our user study, which ensure a fair and objective evaluation of our method compared to others (Sec. 9). We further include examples of generated garments that demonstrate the robustness and generative capabilities of our method across various body types (Sec. 10). We also provide a supplementary video demonstrating that our generated garments can be directly used in CG pipelines for animation production, showcasing high-fidelity simulation of cloth collisions and wrinkle formation. The garment visualization results are rendered using a camera that follows a circular trajectory, effectively emphasizing the superior fit of garments to body shape compared to other methods. The code of SewingLDM will be released publicly.

6. Representation Details

The binary concrete representations of different edge types, attachment types, and stitches are depicted in Fig. 10 alongside their corresponding annotations. For edges, in addition to the vector $V_{i,j}$ representing from the start point to the endpoint, the cubic line employs the control parameters $C^b_{i,j} \in \mathbb{R}^4$ to define two control points (x_1, y_1) and (x_2, y_2) in the 2D coordinate. The circle line uses additional control parameters $C^r_{i,j} \in \mathbb{R}^3$, which specify the radius r and four rotations with two binary flags, including the counterclockwise acute angle ($[0, 0]$), the clockwise acute angle ($[0, 1]$), the counterclockwise reflex angle ($[1, 0]$), and the clockwise reflex angle ($[1, 1]$). Furthermore, edge types are denoted as follows: $E^t_{i,j} = [0, 0]$ for the straight line, $E^t_{i,j} = [0, 1]$ for the quadratic line, $E^t_{i,j} = [0, 0]$ for the cubic line, and $E^t_{i,j} = [1, 1]$ for the circle line. The attachments are visually distinguished by highlighting the associated edges in red and annotating them with the name and value of $A_{i,j}$. Edges without attachment are not highlighted and use the default value of $A_{i,j} = [0, 0, 0]$. There are six kinds of attachment types, *i.e.*, lower interface ($[0, 0, 1]$), right collar ($[0, 1, 0]$), left collar ($[0, 1, 1]$), strapless top ($[1, 0, 0]$), right armhole ($[1, 1, 0]$), and left armhole ($[1, 0, 1]$). For reversal stitch $\{M'_{i,j,2} \in \{0, 1\}\}$, 0 means the stitch direction does not need reversal, while 1 means the stitch direction needs to be reversed. With the detailed representation of sewing patterns, users can convert the sewing patterns into vector representations as the input of neural networks.

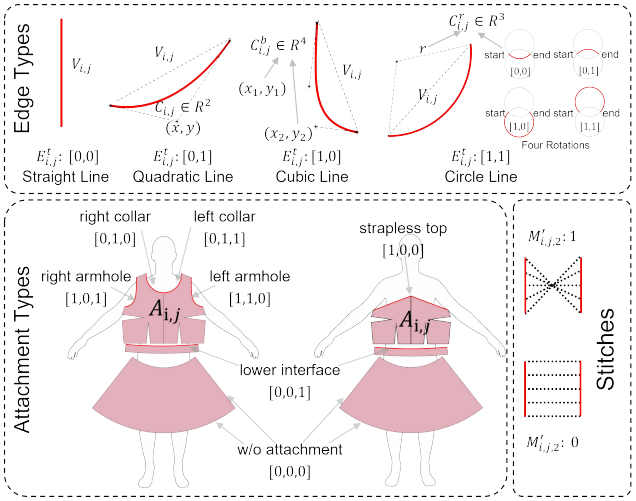

Figure 10. **Representation details.** We present various kinds of edges, attachments, and stitches with detailed annotations.

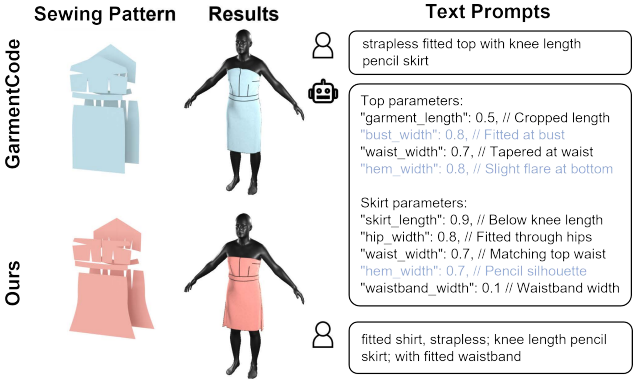

Figure 11. **Comparison with parametric method.** We present the garments and draping results for our SewingLDM and parametric method GarmentCode [27]. GarmentCode needs a delicate selection of values. In contrast, our method can generate garments under intuitive conditions like natural language or sketches, which provide an easier way for garment generation.

7. Comparison with Parametric Method

Except for generation methods, GarmentCode [27] allows users to model complex garments by selecting different parameters and producing desired sewing patterns. However, selecting various parameters is not intuitive, and needs professional knowledge of garment design, limiting its

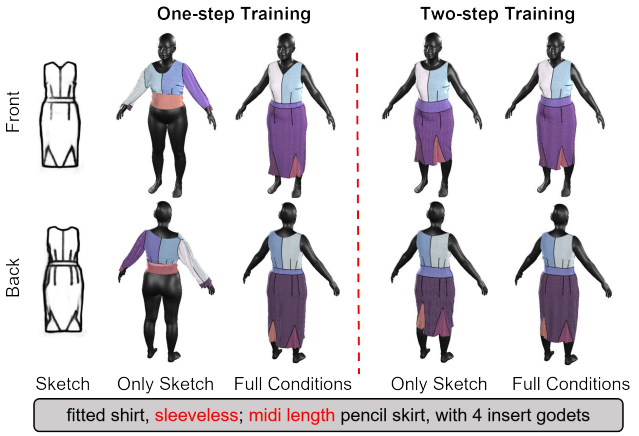

Figure 12. **Ablation on the training strategy.** One-step training shows an unbalance between the multi-modal conditions, failing under only the sketch. In contrast, two-step training helps to faithfully generate the ideal garments with only sketch conditions.

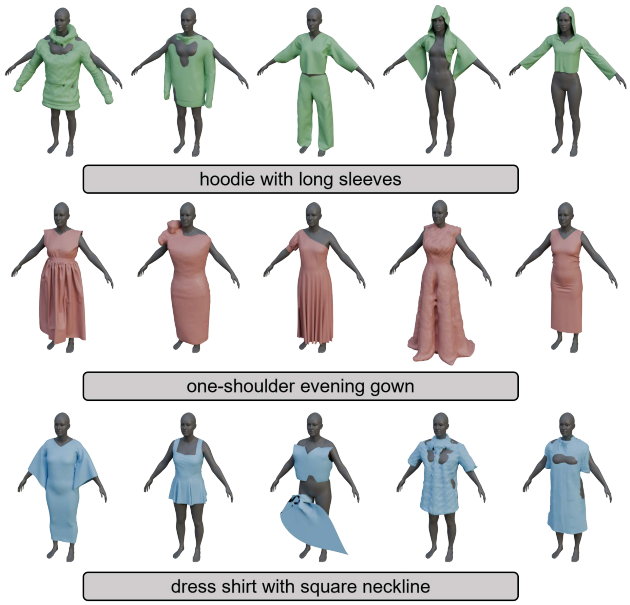

Figure 13. **User study examples.** We present 3 user study examples with random shuffled results.

widespread promotion. To enable the production of the desired sewing patterns through users' prompts, an easy way is to leverage the powerful ability of the large language model, like GPT4 [1]. We simply ask GPT4 to generate various values between 0 to 1 to satisfy the sewing pattern designs of [27], as illustrated in Fig. 11. With the designed prompt, GPT4 can truly provide instructions for garment design. However, most of the generated values are not concerned with garment shape in GarmentCode [27] and still require professional knowledge of garment designs and predefined templates. In summary, GarmentCode [27] needs indispensable manual processing to produce the desired garments. In contrast, our SewingLDM can generate the desired garment through more intuitive conditions, *i.e.*, text prompts and sketches, providing easier tools for garment designs and boosting daily garment production.

8. One-step Training v.s Two-step Training

We additionally take an ablation study on the training strategy. One-step training is unable to balance the multi-modal conditions, so that fails to generate the corresponding garment through only the sketch condition. As shown in Fig. 12, the generated garment loses its midi-length pencil skirt, failing to generate the desired garment. In contrast, the model under two-step training can faithfully generate the corresponding garment with the sketch only, which contains both the sleeveless fitted shirt and the corresponding midi-length pencil skirt. Therefore, two-step training can more effectively inject the sketch conditions into the diffusion model and provide additional control of garment designs, enabling wider usage of our SewingLDM. Moreover, combined with full conditions of text and sketch, SewingLDM can provide more precise control on desired garment generations, meeting users' requirements.

9. User Study Details

To ensure a fair and objective evaluation of our method compared to other methods, we randomly shuffle the results generated by different methods. Each result is paired with a corresponding textual description, and volunteers are asked to rate the results with a score of $1-5$ based on the consistency between the results and the texts, as well as the fitness between the clothes and the human bodies. Additionally, we provide 3 supplementary examples as shown in Fig. 13.

10. Qualitative Results

We additionally provide garments tailored to a wide range of body shapes, spanning variations such as short to tall and slim to broad. As illustrated in Fig. 14, our approach enables the creation of garments specifically adapted to different body types. Furthermore, the simulated garments are enriched with physically based rendering (PBR) textures, either generated by DressCode or designed using the Substance 3D Painter software [2], culminating in visually compelling garment representations as shown in Fig. 15.

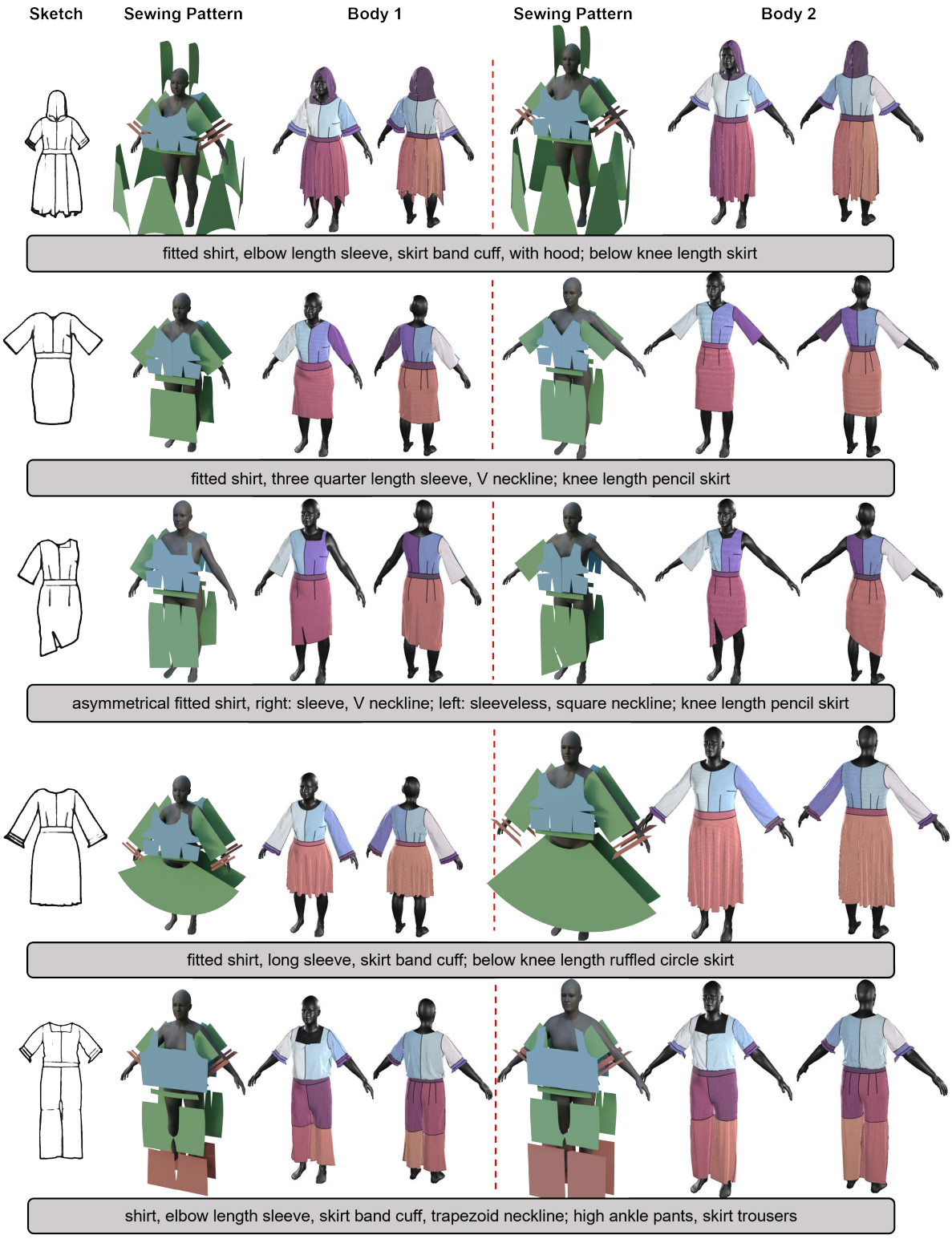

Figure 14. **Additional results for various body shapes.** We present identical garment designs tailored for two distinct body types, encompassing a spectrum of heights and body compositions, to demonstrate the effectiveness of our SewingLDM across diverse bodies.

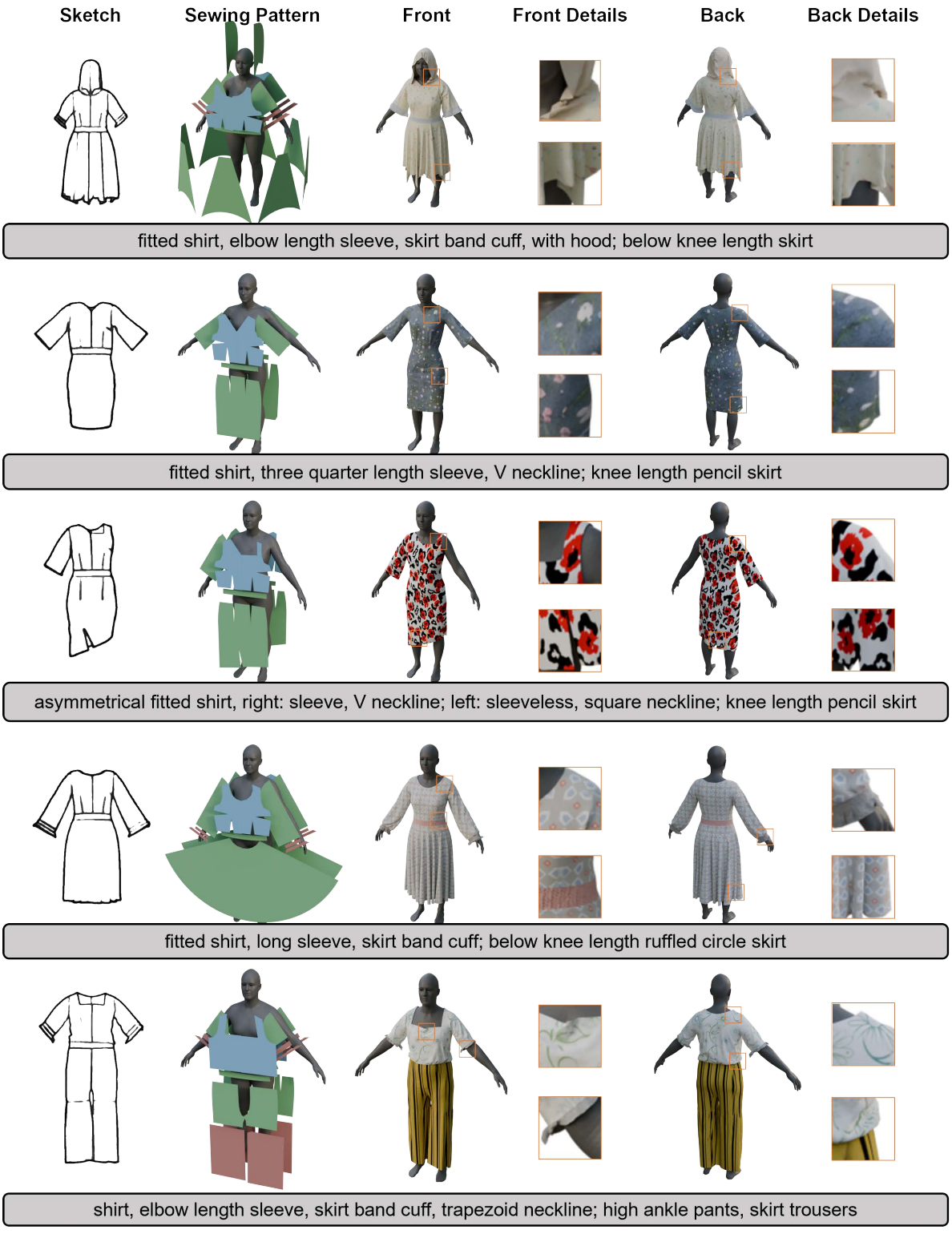

Figure 15. **Additional qualitative results.** By integrating Physically Based Rendering (PBR) textures, our generated outputs achieve visually compelling rendering effects, particularly for a wide range of intricate garment designs.