Image2Garment: Simulation-ready Garments from a Single Image

Anonymous CVPR submission

Paper ID

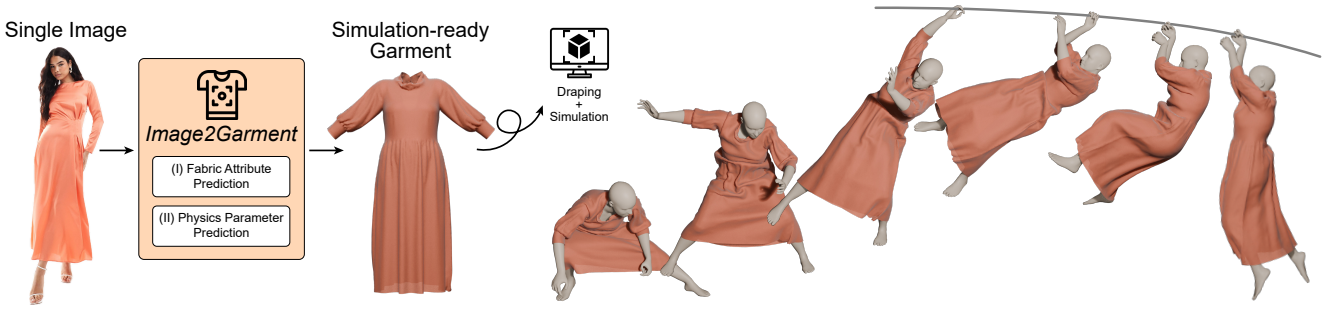

Figure 1. Image2Garment is a feedforward framework that predicts simulation-ready garments from a single image. For this purpose, garment geometry and physical fabric parameters are jointly predicted and used to simulate garment dynamics. Our framework is able to produce complex garments with delicate dynamics, as shown in this example.

Abstract

Estimating physically accurate, simulation-ready garments from a single image is challenging due to the absence of image-to-physics datasets, and the ill-posed nature of this problem. Prior methods either require multi-view capture and expensive differentiable simulation or predict only garment geometry without the material properties required for realistic simulation. We propose a feed-forward framework that sidesteps these limitations by first fine-tuning a vision–language model to infer material composition and fabric attributes from real images, and then training a lightweight predictor that maps these attributes to the corresponding physical fabric parameters using a small dataset of material–physics measurements. Our approach introduces two new datasets (FTAG and T2P) and delivers simulation-ready garments from a single image without iterative optimization. Experiments show that our estimator achieves superior accuracy in material composition estimation and fabric attribute prediction, and by passing them through our physics parameter estimator, we further achieve higher fidelity simulations compared to state-of-the-art image-to-garment methods.

1. Introduction

Generating simulation-ready garments from visual observations is increasingly important for applications in virtual reality, gaming, and fashion design. Although recent methods can recover physical fabric properties, they typically rely on multi-view capture setups or manual material measurement, both of which are labor-intensive and impractical outside controlled environments. While single-image inference would be an accessible alternative, estimating a garment's physical properties (e.g., stretch and bend stiffness, density, damping) from an in-the-wild image remains profoundly challenging due to limited viewpoints, ambiguous drape cues, and the lack of direct supervision.

Recent progress in single-image garment generation, driven by fine-tuning Vision-Language Models (VLMs), has enabled substantial improvements in predicting garment geometry and appearance [12, 45, 51, 56, 81]. However, these methods largely neglect physical parameters necessary for faithful simulation (see Fig. 2). Conversely, methods that optimize physical parameters via a differentiable simulation [27, 48, 49, 62, 80] require multi-view inputs, involve slow iterative optimization, and are typically restricted to simple garment categories, making them impractical for real-world deployment. As a result, obtaining simulation-ready garments from a single image remains a challenging problem. While optimization-based methods struggle with the limited supervision available from a single image, directly training a neural network for this task is equally infeasible due to the lack of appropriate data. No large-scale dataset that pairs real-world garment images

with the physical parameters required by cloth simulators exists, and creating such a resource would demand prohibitive manual effort.

Our key insight is to reformulate this inverse problem through a semantically grounded latent decomposition. Although direct image-to-physics supervision is unavailable, image-to-material information is abundant. Large online catalogs provide reliable material-composition labels (e.g., cotton, silk, polyester blends) as well as complementary garment attributes such as fabric family, weave structure, and thickness indicators. Importantly, these descriptors occupy a structured and relatively low-dimensional space that has a far more predictable relationship to the physical parameters used in cloth simulation. As a result, the material-to-physics mapping is substantially easier to learn and requires only a modest amount of material–physics data that can realistically be collected from industry-standard simulators.

Building on this observation, we introduce a two-stage factorization. First, we train a model to infer an interpretable set of material descriptors from the single input image. Second, using a compact set of independently gathered material–physics annotations, we train a mapper that converts these descriptors into the full set of simulator parameters. This latent-variable formulation regularizes the learning task, dramatically reduces data requirements, and resolves the ill-posedness that makes direct image-to-physics prediction impractical. The result is a fully feed-forward, optimization-free image-to-simulation pipeline.

In summary our contributions are:

- Two datasets: (1) The Fabric Attributes from Garment Tags (FTAG) dataset—containing images annotated with material composition, fabric family, and structure type—and (2) the Tag-to-Physics (T2P) Dataset which links fabric attributes with measurable physical parameters directly compatible with an industry standard simulator.
- A feed-forward framework for single-image to simulation-ready garment generation, jointly predicting garment geometry and interpretable fabric descriptors, and mapping them to physically realistic material parameters.
- Extensive experiments demonstrating superior accuracy and speed, showing that our approach predicts accurate garment physics parameters that can be directly plugged in to image-to-simulation workflows.

2. Related Work

Garment reconstruction and generation. Many works have explored how to reconstruct garments from multi-view images and videos. Works such as [2, 3, 11, 23, 34, 36, 78] take videos, multi-view images, or point clouds as input to recover the garment geometry. However, they do not model the material composition of the garments, and therefore cannot be directly used for physics simulation. Using inverse rendering and inverse physics, other approaches [27, 41, 58, 62, 80] optimize both the garment geometry and its physical parameters. However, they require studio-quality videos as input, making the algorithm impractical for casual users.

Figure 2. **Impact of garment fabric parameters on simulation.** We visualize the final frame of a jumping animation for four different fabrics (wool, cork–cotton, polyester, and a random material) each starting from the exact same initial condition. The choice of garment physics parameters changes the dynamics of the animation drastically. In turn, this makes it critical to estimate these parameters accurately in image-to-garment generation settings to faithfully predict shape, appearance, and dynamics of a garment from visual observations.

More recently, generative models have enabled single-image-to-garment generation. Works such as [4, 29, 31, 33, 39, 63, 71, 72] generate 3D meshes of garments fused with humans, requiring post-processing to separate the garment from the human mesh. Others [18, 43, 44, 55, 61] model the human mesh separately from the garment, but still lack the necessary physics parameters for simulation. Another line of works [12, 28, 32, 46, 50, 51, 56, 74, 81] directly generates the sewing pattern – a CAD representation of garments – from a single image by training on sewing pattern datasets [37, 38]. While this representation is more simulation ready as it automatically includes garment texture maps, users still need to specify physical parameters manually for each fabric panel.

The work closest to ours is Dress 1-to-3 [45], which estimates both sewing pattern and physical parameter from a single image, leveraging inverse optimization. However, their method requires hours of optimization per garment and only outputs specific physical parameters that are not generalizable for other types of simulators. Moreover, code is not available for this method.

Optimizing physics parameters from images and videos. Estimating physics from visual observations has become an increasingly important topic in computer vision [6, 7, 14, 20, 25, 26, 35, 40, 42, 67–69, 75, 77]. Several works have also attempted to obtain physics parameters for garments

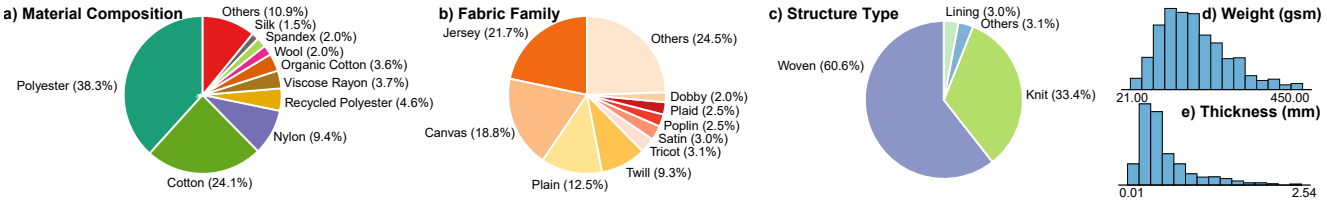

Figure 3. **Garment physics parameters dataset.** Our dataset contains 1,254 samples linking interpretable fabric attributes to measured fabric mechanical properties used for garment simulation. The fabric attributes include: material composition capturing fiber-level makeup that governs mechanical behavior; fabric family describing microstructure, drape, and surface appearance ; structure type identifying fabric construction that restricts the physically plausible range of simulation parameters; and density and thickness, which directly influence mass. Subfigures show distributions for (a) main fiber compositions, (b) fabric families, (c) structure types, (d) thickness, and (e) weight.

from visual data [8, 10, 27, 47–49, 54, 59, 62, 64, 70, 73, 80]. However, all of these methods require strong visual supervision with multi-view or video setups and most are not compatible with established cloth simulators. This clearly shows the difficulty of predicting physics parameters usable in general physics simulators from a single image.

High-quality estimation of garment physical parameters has been studied primarily in controlled settings. Wang et al. [65] fit elastic models from measurements of real fabric samples, and Bhat et al. [9] estimate parameters by matching simulated dynamics to video observations. The latter work refines these measurement-based approaches, including woven-fabric models [16], friction estimation [60], and differentiable-physics pipelines using multi-view or depth data [76, 79]. These methods achieve high fidelity but depend on specialized capture setups. In contrast, we target parameter estimation from a single RGB image.

3. Method

Overview. Our goal is to recover a *simulation-ready garment* from a single RGB image I. We define a simulation-ready garment as $G = (S, \boldsymbol{\theta})$, where S is the 3D garment shape and $\boldsymbol{\theta}$ are the physical parameters required by a cloth simulator (e.g., stretch, bend, damping, friction). We therefore seek to estimate

$$\hat{G} = \arg\max_{G} p(G \mid I) = \arg\max_{S,\boldsymbol{\theta}} p(S, \boldsymbol{\theta} \mid I). \quad (1)$$

While single-view shape estimation $p(S \mid I)$ has seen rapid progress, estimating $\boldsymbol{\theta}$ directly from I is underexplored: paired image–to–physics data do not exist and the mapping is ambiguous. Our key insight is to introduce a latent variable, *material* M (and auxiliary garment attributes Z; e.g., weave, thickness, finish), and decompose the posterior as[1]

$$p(S, \boldsymbol{\theta} \mid I) = p(S \mid I) \sum_{m,z} p(\boldsymbol{\theta} \mid m, z)\, p(m, z \mid I). \quad (2)$$

[1]We make the assumption that the garment geometry S and physical parameters $\boldsymbol{\theta}$ are conditionally independent given the input image I.

Therefore, to obtain a simulation-ready garment G from I, we use a three-stage pipeline: (i) predict a garment geometry G from I; (ii) infer $p(M, Z \mid I)$ using abundant web-scale image-to-material supervision; and (iii) learn a material-to-physics mapper $f_\phi : (M, Z) \mapsto \boldsymbol{\theta}$ from measured material datasets. In practice, we implement

$$\hat{S} = \arg\max_{S} p(S \mid I), \quad (3)$$

$$(\hat{M}, \hat{Z}) = \arg\max_{m,z} p(m, z \mid I), \quad (4)$$

$$\hat{\boldsymbol{\theta}} = f_\phi(\hat{M}, \hat{Z}), \quad (5)$$

so that $\hat{G} = (\hat{S}, \hat{\boldsymbol{\theta}})$. This split concentrates uncertainty into the first stage $p(M, Z \mid I)$ (where rich supervision exists) and turns physics estimation into a well-posed supervised mapping $p(\boldsymbol{\theta} \mid M, Z)$.

We use an existing image-to-garment geometry model to estimate G. To obtain (m, z) and $\boldsymbol{\theta}$, we develop a custom estimation framework from in-the-wild images, leveraging our novel datasets. In the following, we first describe how we obtain these datasets and then how we train the corresponding models.

3.1. Dataset Curation

Fabric attributes from garment tags (FTAG) dataset. We curate a garment material dataset for *fabric composition understanding* in apparel, containing **16,026** images of people wearing garments paired with the corresponding material composition, fabric family, and fabric structure type annotations. The dataset was curated from publicly available online clothing stores and subsequently processed to filter out multi-layered garments and invalid data. Each entry corresponds to a single retail garment with vendor-provided material composition including exact fiber percentages, and a text description of the garment, from which we classify the fabric family and the fabric structure type. Our dataset reflects global fiber production trends [21, 24], with a strong predominance of polyester and cotton, followed by viscose and nylon. We provide further details in the supplementary material.

Tag-to-Physics (T2P) Dataset. We curate a dataset to map intrinsic fabric properties to measurable physics parameters for simulation, containing **1,254** fabrics paired with their corresponding fabric attributes and physical measurements. The dataset was compiled from publicly available online sources. Each entry corresponds to a single digitized fabric characterized by (1) *fabric attributes*, including multi-material composition $C = \{(m_j, p_j)\}_{j=1}^{N}$, where N is the number of constituent materials, m_j denotes the j-th fiber type, and p_j its corresponding percentage, fabric family *f*, structure type *s*, areal density ρ (g/m^2), and thickness t (mm); and (2) *physics parameters*, which quantify the mechanical response of fabrics under deformation. We denote the full set of physics parameters as $\mathbf{P} \in \{\rho$, *friction, internal damping, buckling stiffness (g·mm^2/s^2), buckling ratio, bending stiffness (g·mm^2/s^2), shear stiffness (g/s^2), stretch stiffness (g/s^2)*$\}$. For parameters that are anisotropic, there are distinct values for the weft, warp, and bias directions. Each parameter describes a distinct aspect of a fabric's mechanical behavior—governing resistance to compression, bending, in-plane shear, and tension— and are used in standard commercial garment design software such as Blender [17], MarvelousDesigner [22], and Browzwear [53]. These values are derived from standardized mechanical testing of textiles, providing a supervised mapping from interpretable fabric attributes to their measurable physical response. Fig. 3 visualizes the resulting distributions of fabric families, weave structures, material compositions, thickness, and areal density.

3.2. Simulation-ready Garment Pipeline

Our single-image to simulation-ready garment generation pipeline consists of three stages. First, we reconstruct the garment geometry by estimating its sewing pattern using ChatGarment [12]. Because ChatGarment is trained on a diverse set of body poses to estimate sewing patterns draped on a standard A-pose human body, we can recover the garment geometry by draping the pattern with a standard cloth simulator [37]. Next, we predict the garment's material composition and fabric attributes, including fabric family, structure type, areal density, and thickness. Finally, we map these descriptors to the physical parameters required by the simulator using a learned material-to-physics model. An overview of the full pipeline is shown in Fig. 4. In the remainder of this section, we focus on our main technical contribution: predicting fabric attributes and simulator-compatible physical parameters from a single image.

3.3. Fabric Attribute Prediction

In this step, we estimate the material composition M and garment attributes Z from a single image, where M represents the fiber-level material mixture and Z includes attributes such as fabric family and structure type. A key observation is that these attributes differ significantly in how identifiable they are from visual cues alone. In particular, the conditional distributions $p(m \mid I)$ can be highly multimodal: many distinct material compositions produce nearly indistinguishable appearances in photographs due to similar textures, colors, and weave patterns. By contrast, higher-level descriptors such as fabric family or structure type tend to exhibit clearer visual signatures and therefore admit a more peaked distribution. Attributes like areal density or thickness, however, often lack direct visual indicators and are therefore the most ambiguous. We use this insight to motivate the model design in the following.

Material estimation model. These observations lead us to estimate the material composition with a generative model which not only already possesses general knowledge regarding fabric materials and garments but also supports multi-modal inputs that are essential for material prediction. In particular, we formulate the prediction of the fabric attributes as an image captioning task. For this, we prompt an existing VLM with the unaltered image and a custom text prompt and ask it to output a structured JSON string containing per-fiber percentages and other attributes. In the JSON string, we ask the VLM to include the following fabric properties for each input garment image that are essential for downstream physical parameter estimation:

- Material composition $\hat{C} = \{(\hat{m}_j, \hat{p}_j)\}_{j=1}^{\hat{N}}$ denotes that the garment contains $\hat{p}_j$ percent of fabric $\hat{m}_j \in \mathcal{M}$, a fixed set of apparel fibers [2]. This representation captures the fiber-level makeup of the fabric, which directly governs its mechanical behavior.
- Fabric family $\hat{f} \in \mathcal{F}$, where $\mathcal{F}$ is a set of common textile families[3] such as denim, chiffon, or jersey, describes the fabric's microstructure, drape, and surface appearance.
- Structure type $\hat{s} \in \{\text{knit}, \text{woven}, \text{others}\}$ identifies the fabric construction. This structural prior restricts the physically plausible range of density, thickness, and stiffness parameters used in downstream simulation.

We use Qwen-2.5VL [5] as our base VLM model and fine-tune it with LoRA [30]. Because there is an imbalance in the frequency of materials in the FTAG dataset, we counteract this by specifying weights w_{t_i} as the inverse of the class frequency of the corresponding material during cross-entropy loss computation. Specifically, the model is fine-tuned using a weighted token-level cross-entropy loss:

$$\mathcal{L}_{\text{VLM}} = -\sum_{i} w_{t_i} \log p_\theta(t_i \mid \mathcal{I}, \mathbf{t}_{<i}),$$

where $\mathbf{t}$ is the serialized target sequence.

[2] E.g., cotton, polyester, elastane; full list in the Supplementary Material.

[3] See the Supplementary Material for a full list.

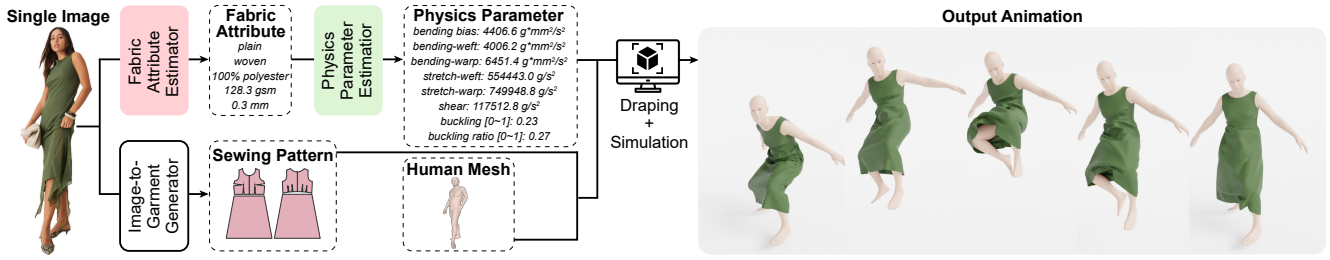

Figure 4. **Overview of the Image2Garment pipeline.** From a single image, we first generate the garment sewing pattern using ChatGarment [12]. Then we predict fabric attributes such as material composition, fabric family, structure type, weight and thickness aligned with standardized commercial garment tags. Finally, fabric physics parameters are estimated from the predicted attributes, following [19], yielding mechanically interpretable quantities that describe fabric deformation. The garment geometry and physical parameters are then used to produce simulation-ready garments and physically accurate draping animations for any given motion sequence or body poses such as SMPL [52].

Density–thickness estimator. Given predicted fabric attributes—composition $\hat{C} = \{(\hat{m}_j, \hat{p}_j)\}_{j=1}^{\hat{N}}$, fabric family $\hat{f}$, and structure type $\hat{s}$—we estimate areal density $\hat{\rho}$ and thickness $\hat{t}$ by sampling from our fabric dataset

$$\mathcal{D} = \{(C_i, f_i, s_i, \rho_i, t_i)\}_{i=1}^{M}, \quad C_i = \{(m_{ij}, p_{ij})\}_{j=1}^{N_i}.$$

We perform a hierarchical search among fabrics sharing the same (f_i, s_i): (1) exact composition match, (2) exact material type match, and (3) primary material match. If multiple candidates exist, $\hat{\rho}$ and $\hat{t}$ are jointly sampled from the same fabric entry within the most specific non-empty set using one of three modes—*mean*, *median*, or *random*. We train using a $70:15:15$ split (train / validation / test) of our fabric dataset and use 5-fold stratified cross-validation to select the sampling mode with the lowest validation MAE. This hierarchical retrieval preserves the empirical correlation between density and thickness while ensuring robustness when exact composition matches are unavailable.

3.4. Physics Parameter Prediction

We predict physics parameters $\mathbf{P}$ from fabric attributes (C, f, s, ρ, t) using independent Random Forest Regressors (RFRs) [13] for bending, shear, stretch, buckling stiffness, and buckling ratio. Each Random Forest regressor is trained on ground-truth fabric attributes and corresponding physics parameters, following [19]. We train using a $70:15:15$ split (train / validation / test) of our fabric dataset and tune hyper-parameters via a 50-iteration randomized search with 5-fold stratified cross-validation, selecting the configuration that minimizes validation MAE. During inference, we use predicted attributes $(\hat{C}, \hat{f}, \hat{s}, \hat{\rho}, \hat{t})$ as input to the same models. Friction μ and internal damping d are fixed constants, while ρ and t are directly taken from predicted fabric attributes. The final estimated physics parameter set is $\hat{\mathbf{P}} = \{\hat{B}, \hat{S}, \hat{E}, \hat{b}, \hat{r}, \mu, d, \hat{\rho}, \hat{t}\}$.

4. Experiments

4.1. Training Configuration

We fine-tune Qwen-2.5VL [5] on the FTAG dataset (Sec. 3.2) using its standard train/val/test split of 9,843/1,231/1,231 samples. To address class imbalance in material composition, we apply inverse-frequency token weights. The model is trained for 5,200 iterations with a global batch size of 9, using LoRA [30] (rank $R = 64$) applied to all modules except [`lm_head`, `embed_tokens`]. We optimize with AdamW (learning rate 1×10^{-5}, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.999)$) and a cosine schedule. Images are dynamically resized between 1500×1500 and 1700×1700 pixels using the Qwen2.5-VL image processor. Training runs on 9 Quadro RTX 6000/8000 GPUs and takes approximately 80 hours.

4.2. Simulation Setup

All garment simulations are performed in the Marvelous Designer (CLO3D) [15] engine, using its standard cloth solver and parameter definitions. We import the draped shapes together with our predicted physics parameters and simulate garments on the SMPL body for each sequence at 24 fps. Cloth resolution is set to 20 mm particle distance, with timestep 0.042 s, gravity $-9800\,\mathrm{cm/s^2}$, and default friction and damping settings (air damping 1.0). Self-collision and body–cloth collision are enabled using iteration-based collision detection (50 CG iterations). All methods are evaluated under identical simulator settings to ensure that differences in dynamic behavior arise solely from the predicted geometry and physics parameters.

4.3. Evaluation Datasets.

4D-Dress Dataset We evaluate garment reconstruction and dynamic draping quality on the 4D-Dress dataset [66]. The dataset contains 50 multi-view video sequences of clothed subjects from which we select 5 sequences for eval-

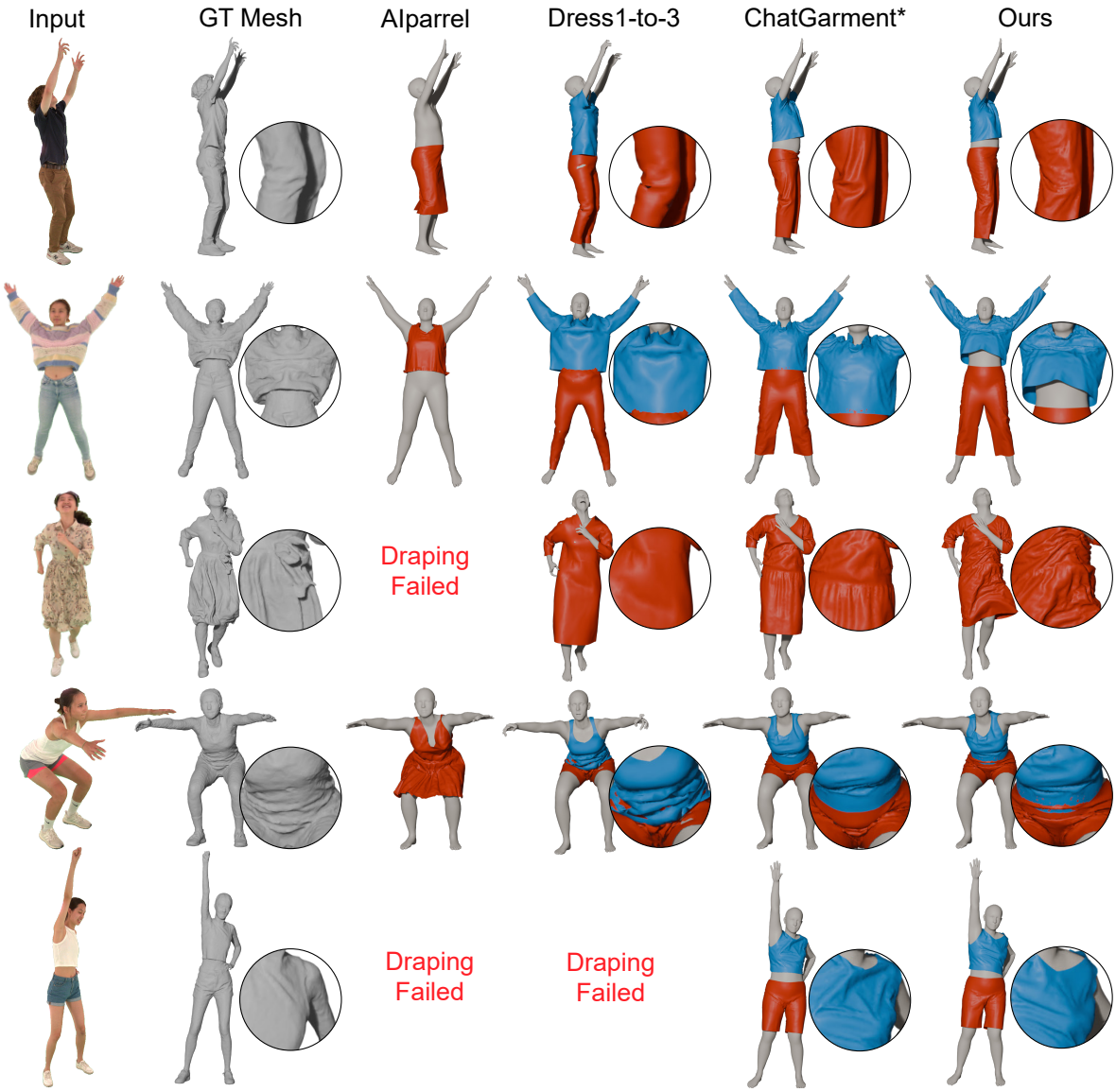

Figure 5. **Qualitative comparison of clothing reconstruction on 4D-Dress [66].** We visualize reconstructed garment meshes from different approaches. The zoomed-in image next to the garment meshes highlights the differences resulting from runningthe simulation. Compared to baselines, our model produces garments with more realistic shapes and wrinkle patterns that are better aligned with the ground truth geometry. *ChatGarment** uses the same garment geometry as our method but randomly sampled physical parameters, leading to incorrect garment deformation and dynamics.

uation. For each sequence, we use the first camera view as the single-image input and use the provided cloth meshes for evaluation.

FTAG Dataset For quantitative evaluation of material and fabric attribute prediction, we use the test split of our FTAG dataset, which includes 1,231 garments spanning 70 distinct material compositions (669 single-fiber and 562 multi-fiber blends), 28 fabric families, and 3 structure types, featuring challenging textures and diverse shapes.

4.4. Baselines

Single-image garment reconstruction. We compare against several state-of-the-art single-image garment reconstruction methods: ChatGarment [12], AIpparel [56], and Dress-1-to-3 [48]. We use the released checkpoints for ChatGarment and AIpparel. For Dress 1-to-3, we obtained the evaluating samples from the authors.

Material and fabric attribute prediction. We use OpenAI's GPT-5 [57], as a baseline as it is often credited with

Method	Lower		Upper	
	CD ↓	IoU ↑	CD ↓	IoU ↑
Dress 1-to-3 [48]	34.1	44.8	40.5	40.0
ChatGarment [12]	28.9	48.4	28.1	46.6
AIpparel [56]	380	28.0	380	28.0
Ours	**27.8**	**48.7**	**27.4**	**47.4**

Table 1. **Clothing reconstruction benchmark.** We report the average Chamfer Distance (CD ↓) and Intersection over Union (IoU ↑) between the ground-truth garment meshes and the reconstructed clothing averaged over time for the lower and upper garment categories in 4D-Dress. Our method achieves the best score in both metrics, showcasing the importance of accurate physical parameters for garment dynamics. Reported CD is multiplied by 1e4.

having the strongest visual recognition and reasoning capabilities. We test it in both a zero-shot and few-shot manner, providing it with the same information about plausible materials that we make available to our own model.

4.5. Evaluation Metrics

Garment prediction. For the simulation-ready garment prediction, we focus on the garment geometry across the entire simulation. We compare it to the ground truth using Chamfer Distance between the point clouds sampled from the predicted and ground-truth mesh, as well as the Intersection over Union (IoU) of the garment geometry. See the supplementary for details.

Fabric attribute prediction. For structure type and fabric family, we report macro-averaged accuracy and F1-score, computed by averaging the per-class metrics across all categories. For material composition, we report per-example accuracy and F1-score averaged across all test examples. For material composition accuracy, we exclude true negatives from the calculation, as the large vocabulary of valid materials combined with the sparsity of material usage (garments typically contain 1-3 materials) would inflate accuracy scores artificially. For numerical values (material percentages), we compare predictions with ground truth using both the Mean Absolute Error (MAE) and Normalized Mean Absolute Error (NMAE).

4.6. Results

Simulation-ready garment estimation. Tab. 1 reports quantitative results on single-image garment reconstruction and dynamic draping. Across all metrics, our approach outperforms existing state-of-the-art methods. In particular, our simulated garments achieve the lowest Chamfer Distance and highest IoU relative to ground-truth cloth geometry, demonstrating that both our reconstructed sewing patterns and our predicted physics parameters are functionally accurate across the entire simulation.

In addition to accuracy, our method is substantially more efficient than optimization-based approaches such as Dress-1-to-3, which require iterative fitting. In contrast, our feed-forward pipeline produces reconstruction and physics estimates in a single pass, enabling consistent performance on in-the-wild inputs without any human intervention.

Fig. 5 provides a visual comparison on the 4D-Dress dataset. While all methods recover plausible static garment shapes, differences become pronounced under motion. Competing approaches frequently produce garments that over-stiffen, collapse, or drift away from the ground-truth dynamics. Our method, by contrast, generates garment motion and silhouettes that closely follow the ground-truth cloth over entire sequences. This highlights the importance of accurate fabric attribute and physics prediction: our inferred bending, shear, and stretch parameters yield dynamic deformations that better capture the true material behavior.

Fabric attribute estimation. Table 2 reports the performance of our fabric attribute estimation model compared to strong vision-language baselines. Across all fields, our fine-tuned Qwen2.5-VL model substantially outperforms both zero-shot and few-shot ChatGPT. For categorical attributes such as fabric family, structure type, and material type, our method yields markedly higher accuracy and F1-scores, with gains of up to 30% absolute accuracy over zero-shot prompting. These improvements reflect the benefits of domain-specific finetuning and structured JSON supervision, which enable the model to resolve subtle appearance cues that general-purpose VLMs fail to capture. For continuous-valued fields, including material percentage, areal density, and thickness, our approach achieves the lowest MAE and NMAE across all metrics. Notably, thickness prediction error is reduced by nearly half compared to zero-shot ChatGPT, and weight estimation shows similar improvements. These results demonstrate that accurate fabric attribute prediction requires specialized training rather than generic prompting, and they establish our model as a reliable foundation for downstream physics parameter estimation and garment simulation.

Ablations. We conduct an ablation study to assess the importance of major components in our framework, as summarized in Table 3. Fine-tuning proves essential across all three tasks. Without fine-tuning, the zero-shot model's performance drops by 23, 23, and 35 percentage points on material composition, structure type, and fabric family estimation, respectively, compared to our fully fine-tuned model. This substantial performance gap demonstrates that task-specific adaptation is critical for reliable fabric property estimation from visual data alone. For material composi-

Attribute Field	ChatGPT (zero-shot)	ChatGPT (few-shot)	Ours
Categorical Fields (Accuracy % / F1-score) ↑			
Fabric Family	0.58 / 0.42	0.61 / 0.43	**0.75 / 0.72**
Structure Type	0.74 / 0.68	0.75 / 0.69	**0.86 / 0.85**
Material Type	0.65 / 0.70	0.66 / 0.70	**0.71 / 0.75**
Continuous Fields (MAE % / NMAE) ↓			
Material Percentage	23.3 / 0.45	22.4 / 0.43	**19.3 / 0.40**
Density (g/m^2)	64.38 / 0.121	75.62 / 0.143	**51.28 / 0.097**
Thickness (mm)	0.376 / 0.053	0.378 / 0.054	**0.227 / 0.032**

Table 2. **Performance of fabric attribute estimation baselines.** ChatGPT (zero-shot) and ChatGPT (few-shot) results are compared with our finetuned Qwen2.5-VL model. Categorical fields are evaluated using Accuracy and F1 Score, while continuous-valued fields are evaluated using MAE and NMAE (%). Weight and Thickness are estimated on the fabric test set using the thickness–density sampler (5-fold Cross Validation), and the remaining fields correspond to the stratified test split of our material composition estimation benchmark.

Table 3. **Ablation study.** We evaluate the impact of fine-tuning, image cropping, majority-only material prediction, and joint prediction with fabric attributes on material composition, structure type, and fabric family estimation tasks.

Method Variant	Accuracy ↑	F1-score ↑
Material Composition		
Full Model	**71**%	**0.75**
w/ Primary Fiber	62%	0.68
w/o Fabric Attributes	69%	0.73
w/o Fine-tuning	48%	0.56
w Cropped Images	63%	0.67
Structure Type		
Full Model	**86**%	**0.85**
w/o Fine-tuning	63%	0.33
w Cropped Images	75%	0.73
Fabric Family		
Full Model	**75**%	**0.72**
w/o Fine-tuning	40%	0.31
w Cropped Images	58%	0.48

tion estimation, predicting only the primary material type (w/ Primary Fiber) leads to a noticeable drop in both accuracy (71% → 62%) and F1-score (0.75 → 0.68), confirming that capturing the full material composition is critical. Removing jointly predicted fabric attributes causes additional degradation (71% → 69% accuracy), indicating that these complementary cues provide useful supporting information for material composition estimation. Following [19], we also evaluate using images of the fabrics only. To isolate the fabric region, we apply a 512×512 center crop to each image. Center-cropping consistently harms performance across all tasks, with severe drops in material composition (71% → 63%), structure type (86% → 75%), and especially fabric family estimation (75% → 58% accuracy). This demonstrates that high-resolution detail and complete garment context—including silhouette, draping patterns, and peripheral regions—provide essential visual information for reliably inferring fabric properties. Overall, each component contributes meaningfully to final estimation quality, with fine-tuning and full-image context being the most critical factors.

References

[1] Adobe Creative Cloud. Mixamo, 2021. 5

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4, 5

[6] Anton Bakhtin, Laurens van der Maaten Wu, Justin Johnson, Quoc Le, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *arXiv preprint arXiv:1908.05656*, 2019. 2

[7] Peter W Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for

learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[8] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Neural cloth simulation. *Computer Graphics Forum (Eurographics)*, 2022. 3

[9] Kiran S. Bhat, Christopher D. Twigg, and Jessica K. Hodgins. Estimating cloth simulation parameters from video. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 37–51, 2003. 3

[10] Kiran S Bhat, Christopher D Twigg, and Jessica K Hodgins. Estimating cloth simulation parameters from video. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2003. 3

[11] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[12] 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, 2025. 1, 2, 4, 5, 6, 7

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001. 5

[14] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016. 2

[15] CLO Virtual Fashion. Marvelous designer, 2024. 3D garment design cloth simulation software. 5, 2

[16] Donald Clyde, Rasmus Tamstorf, and Joseph Teran. Modeling and data-driven parameter estimation for woven fabrics. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 1–10, 2017. 3

[17] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4, 5

[18] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2

[19] Henar Dominguez-Elvira, Alicia Nicás, Gabriel Cirio, Alejandro Rodríguez, and Elena Garces. Practical Method to Estimate Fabric Mechanics from Metadata. *Computer Graphics Forum*, 2024. 5, 8, 2, 4

[20] Sebastian Ehrhardt, Aron Monszpart, Niloy Mitra, and Andrea Vedaldi. Unsupervised intuitive physics from visual observations. In *Asian Conference on Computer Vision (ACCV)*, 2018. 2

[21] Textile Exchange. Preferred fiber & materials market report 2021. *Textile Exchange: Lamesa, TX, USA*, 4, 2021. 3

[22] CLO Virtual Fashion. Clo 3d – 3d fashion design software, 2009. Accessed: 2025-10-07. 4, 5

[23] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[24] L Fernandez. Distribution of textile fibers production worldwide in 2020, by type, 2021. 3

[25] Pablo Garcia, Yue Sun, Taesung Kim, and Animesh Garg. Learning physics from video: Unsupervised physical parameter estimation for continuous dynamical systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[26] Oliver Groth, Kerstin Wimmer, Andrea Vedaldi, and Christian Rupprecht. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[27] Michelle Guo, Matt Jen-Yuan Chiang, Igor Santesteban, Nikolaos Sarafianos, Hsiao-yu Chen, Oshri Halimi, Aljaž Božič, Shunsuke Saito, Jiajun Wu, C Karen Liu, et al. Pgc: Physics-based gaussian cloth from a single pose. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21215–21225, 2025. 1, 2, 3

[28] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance, 2024. 2

[29] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, 2021. 2

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4, 5

[31] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3099, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 2

[32] Moon-Hwan Jeong, Dong-Hoon Han, and Hyeong-Seok Ko. Garment capture from a photograph. *Comput. Animat. Virtual Worlds*, 26(3–4):291–300, 2015. 2

[33] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, page 18–35, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[34] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[35] Venkatesh Kandukuri, Korbinian Riedhammer, and Vladlen Koltun. Physical representation learning and parameter identification from video via differentiable physics. *International Journal of Computer Vision (IJCV)*, 2021. 2

[36] Maria Korosteleva and Sung-Hee Lee. Neuraltailor: Reconstructing sewing pattern structures from 3d point clouds of garments. *ACM Trans. Graph.*, 41(4), 2022. 2

[37] Maria Korosteleva and Olga Sorkine-Hornung. GarmentCode: Programming parametric sewing patterns. *ACM Transaction on Graphics*, 42(6), 2023. SIGGRAPH ASIA 2023 issue. 2, 4

[38] 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 *Computer Vision – ECCV 2024*, 2024. 2, 5

[39] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-Degree Textures of People in Clothing from a Single Image . In *2019 International Conference on 3D Vision (3DV)*, pages 643–653, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2

[40] Long Le, Ryan Lucas, Chen Wang, Chuhao Chen, Dinesh Jayaraman, Eric Eaton, and Lingjie Liu. Pixie: Fast and generalizable supervised learning of 3d physics from pixels. *arXiv preprint arXiv:2508.17437*, 2025. 2

[41] Changmin Lee, Jihyun Lee, and Tae-Kyun Kim. Mpmavatar: Learning 3d gaussian avatars with accurate and robust physics-based dynamics, 2025. 2

[42] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016. 2

[43] Ren Li, Corentin Dumery, Benoit Guillard, and Pascal Fua. Garment Recovery with Shape and Deformation Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[44] 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. 2

[45] Ren Li, Cong Cao, Corentin Dumery, Yingxuan You, Hao Li, and Pascal Fua. Single view garment reconstruction using diffusion mapping via pattern coordinates, 2025. 1, 2

[46] Siran Li, Chen Liu, Ruiyang 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. *arXiv preprint arXiv:2504.01483*, 2025. 2

[47] Xiaoyu Li, Xinyu Chen, Joshua B. Tenenbaum, and Jiajun Wu. Neural garment dynamics via manifold-aware transformers. In *Eurographics*, 2024. 3

[48] 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. 1, 6, 7

[49] Yunzhu Li, Xinyang Liu, Guandao Yang, Gordon Wetzstein, Joshua B Tenenbaum, and Jiajun Wu. Diffavatar: Simulation-ready 3d avatars with differentiable physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 3

[50] Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023. 2

[51] 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. *International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5

[53] Browzwear Solutions Pte Ltd. Browzwear — digital apparel design and development platform, 2025. Accessed: 2025-11-12. 4, 2

[54] Weijia Mao, Chengcheng Tang, Robert Tong, Zhaoqi Wang, and Ming C Li. Video classification of cloth simulations using deep networks. In *Pacific Graphics (PG)*, 2020. 3

[55] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[56] 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. 1, 2, 6, 7

[57] OpenAI. Gpt-5, 2025. Large language model accessed via the OpenAI API. 6

[58] Bo Peng, Yunfan Tao, Haoyu Zhan, Yudong Guo, and Juyong Zhang. Pica: Physics-integrated clothed avatar. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–15, 2025. 2

[59] Gerard Pons-Moll, Sergi Pujades, Shihao Hu, and Michael J. Black. Clothcap: Seamless 4d clothing capture and retargeting. In *ACM Transactions on Graphics (SIGGRAPH)*, 2017. 3

[60] Ahmed Rasheed, Omar AlSibai, Luca Bergamini, Stelian Coros, and Bernhard Thomaszewski. Learning to measure the static friction coefficient in cloth contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4880–4889, 2020. 3

[61] Li Ren, Benoit Guillard, Edoardo Remelli, and Pascal Fua. DIG: Draping Implicit Garment over the Human Body. In *Asian Conference on Computer Vision*, 2022. 2

[62] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J. Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian Garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. In *International Conference on 3D Vision 2025*, 2025. 1, 2, 3

[63] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2

[64] Igor Santesteban, Elena Garces, and Miguel A. Otaduy. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[65] Huamin Wang, Ravi Ramamoorthi, and James F. O'Brien. Data-driven elastic models for cloth: Modeling and measurement. *ACM Transactions on Graphics (TOG)*, 30(4):109, 2011. 3

[66] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5, 6

[67] Nicholas Watters, Andrea Tacchetti, Theophane Weber, Razvan Pascanu, Peter Battaglia, and Daniel Zoran. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[68] Jiajun Wu, Ilker Yildirim, Joseph J Lim, William T Freeman, and Joshua B Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[69] Jiajun Wu, Erika Lu, Pushmeet Kohli, William T Freeman, and Joshua B Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[70] Donglai Xiang, Guandao Yang, Yang Zheng, Yunzhu Li, and Gordon Wetzstein. Inverse garment and pattern modeling with a differentiable simulator. In *CVPR Workshops*, 2024. 3

[71] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 2

[72] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[73] Yonghao Xu, Min Tang, and Robert Tong. Learning-based bending stiffness parameter estimation for cloth simulation. In *Computer Graphics Forum (Eurographics)*, 2021. 3

[74] 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. *ACM Trans. Graph.*, 37(5), 2018. 2

[75] Kexin Yi, Chuang Gan, Yunzhu Li, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[76] Dohyeong Yoon et al. Bayesian differentiable physics for cloth digitalization. *Journal of Computational Design and Engineering*, 12(8):29–44, 2025. 3

[77] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. *arXiv:2404.13026*, 2024. 2

[78] Yiqun Zhao, Chenming Wu, Binbin Huang, Yihao Zhi, Chen Zhao, Jingdong Wang, and Shenghua Gao. Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025. 2

[79] Qiuyu Zheng, Robert Smith, and Sanjiban Choudhury. Differentiable cloth parameter identification and state estimation in manipulation. *arXiv preprint arXiv:2311.05141*, 2023. 3

[80] Yang Zheng, Qingqing Zhao, Guandao Yang, Yifan Wang, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, and Gordon Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3

[81] 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. 1, 2

Image2Garment: Simulation-ready Garments from a Single Image

Supplementary Material

5. Supplementary Website

Please open the attached supplementary website to view our results in video, including dynamic draping simulations and qualitative comparisons.

6. Overview

This supplementary material provides additional details on our datasets, model hyperparameters, and evaluation protocols. We also include further quantitative and qualitative results to complement the main paper. Specifically, we cover:

- Dataset curation and normalization for the Fabric Attributes from Garment Tags (FTAG) and Tag-to-Physics (T2P) datasets;
- Implementation details of the density–thickness estimator, and the physics parameter regressor;
- Detailed evaluation metrics and protocols for both fabric attribute prediction and dynamic draping;
- Additional quantitative comparisons and qualitative examples, including typical failure cases.

7. Additional Dataset Details

7.1. FTAG: Fabric Attributes from Garment Tags

Material Composition Estimation Dataset

We curate a garment material dataset for *fabric attribute prediction*, containing **16,026** images of models wearing retail garments paired with three key labels: *material composition*, *fabric structure type*, and *fabric family*. Images were collected from publicly available online sources and processed through extensive cleaning, de-duplication, and normalization.

Each entry corresponds to a single retail garment with vendor-provided **material composition**

$$C = \{(m_j, p_j)\}_{j=1}^{N},$$

where each fiber type m_j belongs to a controlled vocabulary, including:

We further classify the garment's **fabric structure type**:

woven, knit, others

where woven fabrics correspond to interlaced yarns (e.g., denim, poplin, twill), knit fabrics correspond to interlooped yarns (e.g., jersey, rib, fleece), and others include non-woven categories such as mesh, lace, and sequin fabrics.

Each garment also receives a **fabric family** label:

chiffon, crepe, denim, satin, taffeta, poplin, twill, canvas, velvet, corduroy, flannel, organza, voile, lace, tulle, mesh, crochet, broderie, jacquard, sequin, metallic, leather, fleece, jersey, rib knit, french terry, interlock, pique, milano knit, ponte, scuba, tricot, plissé, gingham, chambray, plain, cheesecloth, terry, unknown

selected from a controlled taxonomy based on visual texture, weave pattern, drape, and surface appearance.

Processing and Normalization Details

All garment entries come with vendor-provided material composition and fabric descriptions that accompany each product image, allowing us to access standardized fiber information and free-form fabric terminology. The **material composition** is already reported in a consistent structured format (e.g., "95% Polyester, 5% Elastane"), allowing us to extract fiber types directly and retain only materials belonging to our validated vocabulary. During this step, we also map synonymous fiber names to a canonical form (e.g., *Spandex* and *Lycra Elastane* $\rightarrow$ *Elastane*), ensuring that all materials follow a unified naming scheme. Samples containing unrecognized fibers or malformed percentage formats are discarded.

For the **fabric family** and **fabric structure type**, the vendor descriptions include a wide variety of naming conventions (e.g., "satin-style", "ribbed jersey", "denim-like weave"). To ensure consistency, we perform an exhaustive search over these free-form descriptors and map each variant to its canonical category in our controlled taxonomy. Examples include normalizing "satin-style" and "sateen" to *satin*, "ribbed knit" to *rib knit*, and "mesh lace" to *mesh*.

The **structure type** is also provided in the vendor metadata and we standardize it into the categories *woven*, *knit*, or *others*. We remove entries where the structure type contradicts the fabric family (e.g., "woven jersey") or where the information is incomplete or ambiguous.

Our dataset also contains garments for which vendors report material composition under multiple headers (e.g., *Main*, *Shell*, *Lining*, *Upper*), each corresponding to a different visible or internal layer of the garment. Because our goal is to predict a single material composition from a single image, we retain only entries whose composition is provided under a primary header (*Main* or other non-secondary headers) and discard garments where the composition is split across multiple layers. This removes cases such as jackets worn over shirts or dresses with separate lining materials, where a single image cannot be associated with a unique fiber composition. The resulting FTAG dataset statistics and distributions are shown in Figure 6.

We additionally discard any sample for which one or more of the three attributes —material composition, struc-

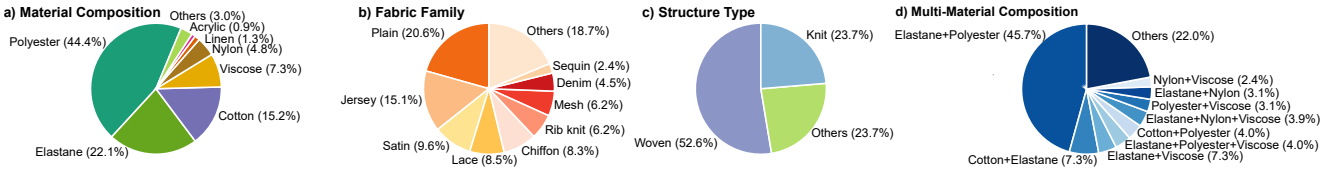

Figure 6. **Fabric attribute dataset.** Our Fabric Attributes from Garment Tags (FTAG) dataset contains 16,026 garment images paired with interpretable, vendor-provided fabric metadata. Each sample includes (i) *material composition*, specifying the fiber-level makeup that governs mechanical behavior; (ii) *fabric family*, characterizing microstructure, drape, and surface appearance; and (iii) *structure type*, indicating how yarns are interlaced or knitted to form the fabric. Subfigures show the distributions of (a) main fiber compositions, (b) fabric families, (c) structure types, and (d) multi-material fiber combinations.

ture type, or fabric family—was missing, inconsistent, or ambiguous. After applying these filtering steps, we obtain a final cleaned dataset consisting of **12,305** samples. We split this dataset in a stratified manner based on the joint label of material composition, structure type, and fabric family, resulting in **9,843** training, **1,231** validation, and **1,231** test samples for our fabric attribute estimation benchmark.

7.2. Tag-to-Physics (T2P) Dataset

We curate a dataset to convert intrinsic fabric properties to simulation physics parameters, containing **1,277** fabrics paired with corresponding fabric attributes and CLO 3D physics parameters. The dataset is collected from publicly available online sources and reflects real-world materials.

Each entry corresponds to a single digitized fabric with (1) fabric attributes, which consist of:

- *Material composition* including exact fiber percentages, $C \in \{$*Acrylic, Alpaca, Angora, Cashmere, Cotton, Cupro, Hemp, Jute, Linen, Lyocell, Metallic, Modal, Nylon, Organic Cotton, PE (Polyethylene), PP (Polypropylene), PU (Polyurethane), PVC (Polyvinyl chloride), Polyester, Raccoon, Recycled Nylon, Recycled Polyester, Silk, Spandex, Supima Cotton, TENCEL™, TPE (Thermoplastic Elastomer), TPU (Thermoplastic Poly Urethane), Triacetate, Viscose Rayon, Viscose from Bamboo, Wool, vinyl acetate copolymer*$\}$;
- *Fabric family* $f \in \{$*Canvas, Challis, Chambray / Oxford, Chiffon, Circular Knit Spacer, Clip Jacquard, Corduroy, Crepe / CDC, Denim, Dobby, Dobby / Jacquard, Double Knit / Interlock, Double Weave, Flannel, Fleece, French Terry, Gauze / Double Gauze, Georgette, Interlining, Jacquard / Brocade, Jersey, Lace, Loop Terry, Memory, Mesh / Tulle, Organza, PVC, Pique, Plaid, Plain, Pointelle, Polar Fleece, Ponte, Poplin, Raschel, Rib, Ripstop, Satin, Seersucker, Sequin, Stretch Lining, Taffeta, Taffeta Lining, Tricot, Tweed, Twill, Vegan Fur, Vegan Leather, Vegan Suede, Velour, Velvet, Voile, Waffle*$\}$;
- *Structure type* $s \in \{$*Knit, Woven, Lining, Others*$\}$;
- Areal density $\rho \in \mathbb{R}$ (g/m^2) and thickness $t \in \mathbb{R}$ (mm).

(2) CLO 3D physics parameters quantify the mechanical response of fabrics under deformation: *P* $\in \{\rho$, *friction, internal damping, buckling stiffness* $\{b_{bias\text{-}left}, b_{bias\text{-}right}, b_{warp}, b_{weft}\}$, *buckling ratio* $\{r_{bias\text{-}left}, r_{bias\text{-}right}, r_{warp}, r_{weft}\}$, *bending stiffness* $\{B_{bias\text{-}left}, B_{bias\text{-}right}, B_{warp}, B_{weft}\}$ *(g·mm^2/s^2), shear stiffness* $\{S_{left}, S_{right}\}$ *(g/s^2), and stretch stiffness* $\{E_{warp}, E_{weft}\}$ *(g/s^2)*$\}$.

Each parameter describes a distinct aspect of the fabric's mechanical behavior, governing resistance to compression, bending, in-plane shear, and tension, and is directly compatible with commercial garment design software such as Marvelous Designer [15], Browzwear [53], CLO 3D, and proprietary ones such as Seddi Author [19]. Each fabric sample is measured using the CLO Fabric Kit; we use these measurements without additional scale factors, enabling a supervised mapping from interpretable fabric attributes to measurable mechanical response. Figure 3 in the main paper visualizes the distributions of fabric families, structure types, material compositions, thickness, and areal density.

7.3. Material Prediction Metrics

We report two primary metrics for evaluating material prediction performance:

Average Accuracy:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{6}$$

Average accuracy across all examples, where for each example i, we compute the ratio of correctly predicted materials (true positives) to the total number of unique materials in either the ground truth or predictions. N is the total number of examples.

Average F1 Score:

$$\text{F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{7}$$

where $\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i}$ and $\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$.

Average F1 score across all examples, computed as the harmonic mean of precision and recall for each example.

For each example i, TP_i represents correctly identified materials, FP_i represents incorrectly predicted materials, and FN_i represents ground truth materials that were missed.

7.4. Material Percentage Prediction Metrics

We report two metrics for evaluating the accuracy of predicted material percentages:

Average Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|M_i|}\sum_{m\in M_i}|p_{i,m}^{\text{gt}} - p_{i,m}^{\text{pred}}| \tag{8}$$

Average mean absolute error across all examples, where for each example i, we compute the absolute difference between ground truth percentage $p_{i,m}^{\text{gt}}$ and predicted percentage $p_{i,m}^{\text{pred}}$ for each material m in the set M_i (all unique materials in either ground truth or predictions for example i). $|M_i|$ is the total number of materials for example i, and N is the total number of examples.

Average Normalized Mean Absolute Error (NMAE):

$$\text{NMAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|M_i|}\sum_{m\in M_i}\frac{|p_{i,m}^{\text{gt}} - p_{i,m}^{\text{pred}}|}{\max(p_{i,m}^{\text{gt}}, p_{i,m}^{\text{pred}})} \tag{9}$$

Average normalized mean absolute error across all examples, where the absolute error for each material is normalized by the maximum of the ground truth and predicted percentages to account for varying material proportions. This provides a scale-invariant measure of percentage prediction accuracy.

7.5. Structure Type Prediction Metrics

We report two metrics for evaluating fabric structure type classification:

Overall Accuracy:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(s_i^{\text{gt}} = s_i^{\text{pred}}) \tag{10}$$

Overall accuracy across all examples, where s_i^{gt} is the ground truth structure type for example i, s_i^{pred} is the predicted structure type, and $\mathbb{1}(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise. N is the total number of examples.

Macro-Averaged F1 Score:

$$\text{F1}_{\text{macro}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\text{F1}_s \tag{11}$$

where

$$\text{F1}_s = \frac{2\times\text{Precision}_s\times\text{Recall}_s}{\text{Precision}_s + \text{Recall}_s} \tag{12}$$

and $\text{Precision}_s = \frac{\text{TP}_s}{\text{TP}_s+\text{FP}_s}$, $\text{Recall}_s = \frac{\text{TP}_s}{\text{TP}_s+\text{FN}_s}$.

Macro-averaged F1 score computed by first calculating the F1 score for each structure type class $s \in \mathcal{S}$ (where $\mathcal{S}$ is the set of all structure types: {knit, woven, others}), then averaging across all classes. TP_s, FP_s, and FN_s represent true positives, false positives, and false negatives for class s, respectively.

7.6. Fabric Family Prediction Metrics

We report two metrics for evaluating fabric family classification:

Overall Accuracy:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(f_i^{\text{gt}} = f_i^{\text{pred}}) \tag{13}$$

Overall accuracy across all examples, where f_i^{gt} is the ground truth fabric family for example i, f_i^{pred} is the predicted fabric family, and $\mathbb{1}(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise. N is the total number of examples.

Macro-Averaged F1 Score:

$$\text{F1}_{\text{macro}} = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\text{F1}_f \tag{14}$$

where

$$\text{F1}_f = \frac{2\times\text{Precision}_f\times\text{Recall}_f}{\text{Precision}_f + \text{Recall}_f} \tag{15}$$

and $\text{Precision}_f = \frac{\text{TP}_f}{\text{TP}_f+\text{FP}_f}$, $\text{Recall}_f = \frac{\text{TP}_f}{\text{TP}_f+\text{FN}_f}$.

Macro-averaged F1 score computed by first calculating the F1 score for each fabric family class $f \in \mathcal{F}$ (where $\mathcal{F}$ is the set of all fabric families, including Poplin, Jersey, Lace, Satin, Chiffon, etc.), then averaging across all classes. TP_f, FP_f, and FN_f represent true positives, false positives, and false negatives for class f, respectively.

7.7. Density and Thickness Estimation Metrics

We use our T2P dataset to estimate density and thickness values from ground truth fabric family, structure type, and material composition. We also prompt GPT to perform the same estimation for comparison (see Table 2 in the main paper).

We report two metrics for evaluating continuous parameter estimation:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i^{\text{gt}} - y_i^{\text{pred}}| \tag{16}$$

Mean absolute error between ground truth values y_i^{gt} and predicted values y_i^{pred} across all N examples.

Normalized Mean Absolute Error (NMAE):

$$\text{NMAE} = \frac{\text{MAE}}{y_{\max} - y_{\min}} \tag{17}$$

Normalized mean absolute error, where MAE is divided by the range of ground truth values ($y_{\max} - y_{\min}$) to provide a scale-invariant measure of prediction accuracy.

7.8. Density–Thickness Estimator

Given predicted discrete attributes $(\hat{C}, \hat{f}, \hat{s})$ from the VLM, we estimate areal density $\hat{\rho}$ and thickness $\hat{t}$ using the T2P dataset $\mathcal{D} = \{(C_i, f_i, s_i, \rho_i, t_i)\}_{i=1}^{M}$. We perform hierarchical retrieval:

1. **Exact attribute match**: we first search for fabrics with $(f_i, s_i) = (\hat{f}, \hat{s})$ and a composition C_i matching $\hat{C}$ exactly up to a small tolerance in percentages.
2. **Material-set match**: if no exact match exists, we relax to fabrics whose *set* of fibers matches that of $\hat{C}$, ignoring small percentage differences.
3. **Primary fiber match**: if the previous sets are empty, we match fabrics whose primary fiber (highest percentage in C_i) equals the primary fiber in $\hat{C}$.

From the most specific non-empty candidate set, we estimate $(\hat{\rho}, \hat{t})$ either by taking the mean, median, or by randomly sampling a single fabric entry. We select the aggregation mode by performing 5-fold stratified cross-validation on the T2P dataset (train:val:test = 70:15:15) and choosing the mode with lowest validation MAE. The best performing mode was **mean aggregation**. This retrieval-based strategy preserves the empirical coupling between areal density and thickness while remaining robust when exact composition matches are not available.

7.9. Physics Parameter Estimator

We predict physics parameters P from fabric attributes (C, f, s, ρ, t) using a collection of independent Random Forest Regressors (RFRs), following the practical recipe of Dominguez-Elvira *et al.* [19]. We form a feature vector by concatenating:

- normalized fiber percentages over the vocabulary,
- one-hot encodings of fabric family f and structure type s,
- scalar features ρ and t (optionally log-transformed).

We train five separate Random Forest Regressors, one for each physics parameter group. We use a 70/15/15 train/val/test split and perform a 50-iteration randomized hyperparameter search with 5-fold stratified cross-validation to select the number of trees and features, the max depth, and the minimum samples per split and leaf.

For training all models, we use Mean Absolute Error as the loss function. The best hyperparameters selected via cross-validation are: for bending, stretch, shear, and buckling stiffness, $n_{\text{estimators}} = 100$, max_depth $= 20$, min_samples_split $= 0.0420$, min_samples_leaf $= 0.0094$, max_features $= 0.6911$; for buckling ratio, $n_{\text{estimators}} = 200$, max_depth $= 30$, min_samples_split $= 0.0703$, min_samples_leaf $= 0.0016$, max_features $= 0.9595$.

During inference, we feed predicted attributes $(\hat{C}, \hat{f}, \hat{s}, \hat{\rho}, \hat{t})$ into the same regressors to obtain $\hat{P}$. Friction μ and internal damping d are kept fixed across fabrics, while $\hat{\rho}$ and $\hat{t}$ are directly taken from the density–thickness estimator.

8. Garment Geometry and Simulation Details

8.1. ChatGarment-based Sewing Pattern Estimation

For garment geometry, we adopt ChatGarment as our image-to-sewing-pattern backbone. Given a single image, ChatGarment predicts a sewing pattern representation draped on a canonical SMPL body in A-pose. We export the predicted pattern and re-simulate it using our cloth simulator, ensuring that all methods (ours and baselines) are evaluated under identical numerical simulation settings.

We use the default ChatGarment resolution and panel parameterization and do not fine-tune ChatGarment on 4D-Dress to keep the comparison fair and emphasize the benefits of improved physics prediction.

9. Evaluation Protocols and Metrics

9.1. Dynamic Draping Metrics

To evaluate dynamic draping on 4D-Dress, we compare our simulated garment meshes against ground-truth cloth meshes over the entire sequence. We use two metrics:

- **Chamfer Distance (CD) ↓:** We sample points from both the predicted and ground-truth garment meshes at each frame and compute the symmetric Chamfer distance. The reported CD is averaged over frames and scaled by 10^4 for readability.
- **Intersection-over-Union (IoU) ↑:** We voxelize each mesh using 5cm voxels and compute the intersection-over-union of the occupied voxels between predicted and ground-truth garments, then average over frames.

We report separate metrics for upper- and lower-body garments, as in Table 1 of the main paper. Garments spanning upper and lower body, count towards both categories results.

9.2. Fabric Attribute Prediction Metrics

For fabric attribute estimation on FTAG, we evaluate:

- **Structure type** and **fabric family**: macro-averaged accuracy and F1-score, computed by averaging per-class metrics across all categories.
- **Material type**: multi-label accuracy and F1-score over the material vocabulary, ignoring true negatives to avoid inflated scores due to sparsity.

- **Material percentages, density, and thickness**: mean absolute error (MAE) and normalized MAE (NMAE) over all test samples.

These metrics are reported for our fine-tuned Qwen2.5-VL model and for ChatGPT baselines (zero- and few-shot) in the main paper.

10. Additional Results

10.1. Synthetic Dataset

Dataset creation. To compare our method further with exiting state-of-the-art methods we create the first synthetic evaluation dataset designed with full garment prediction in mind. In particular our dataset contains three outfits sampled from [38] that each have been assigned a material per garment. We drape this garment over SMPL [52] bodies and create distinct animations from Mixamo [1]. We then use CLO 3D's [22] simulator to simulate the garment under the animation. Finally, we render the garment at each step in the animation using Blender [17]. From this data, we use the first rendered frame as an input and the remaining frames and meshes for evaluation. Compared to existing evaluation benchmarks, this data has the advantage that it includes more complex garments and animations, and it does not suffer from measurement inaccuracies.

Quantitative Evaluation. In this section, we evaluate our method on the dataset we created and compare it against state-of-the-art garment shape estimators paired with random material samples. Tab. 4 reports the quantitative results. Our method achieves the best average performance across all 3D metrics and most 2D metrics. The only exception is SSIM, where our score is slightly lower; however, this can be attributed to the uniformly high SSIM values achieved by all methods, which reduces the discriminative power of this metric.

Qualitative Evaluation. Here, we present qualitative comparisons between our method and state-of-the-art garment generation approaches using random physics parameters. Fig. 7 shows results for the first sequence in our dataset. While AIpparel and DMap struggle to recover accurate garment shapes, our method produces a plausible geometry and estimates physics parameters that remain consistent with the ground truth throughout the animation. Compared to ChatGarment, our predictions avoid the excessive deformation observed in its outputs.

Fig. 8 presents results for the second sequence. As before, our predicted shape and physics parameters yield the closest visual match to the ground truth across the animation. In contrast, ChatGarment and AIpparel consistently generate dresses that appear unnaturally folded.

Finally, Fig. 9 shows the third sequence. While AIpparel again fails to estimate the correct shape, our method remains close to the ground truth. During the rapid, high-energy motion in this sequence, the ground-truth garment slides off the human body due to its loose fit, a behavior that our method successfully reproduces but none of the baselines capture.

Ablations. In addition to the parameter ablations presented in the main paper, we conduct further ablations at the 4D shape level using our curated dataset. To isolate the effect of physics parameter prediction, we use the ground-truth meshes and vary only the physics parameters. The quantitative results are reported in Tab. 5. Our predicted physics parameters yield substantial improvements across all metrics and sequences. Although both our method and the baseline exhibit some variance, our approach consistently outperforms the baseline overall, highlighting the importance of accurate physics parameter estimation.

10.2. Additional Image-to-Simulation Examples

Qualitative Evaluation. In the supplementary figures, we provide additional in-the-wild examples produced by our Image2Garment pipeline. For each example, we show the original sequence (first image as input), the reconstructed sewing pattern, the final simulation frame, and close-ups of characteristic wrinkles. Figures 10, 11, and 12 compare our results against simulations generated using random fabric materials. Across diverse garment categories, our method produces more realistic silhouettes, time-consistent wrinkle patterns, and garment dynamics that closely match those observed in the source video clips. Please refer to our project website for the full video results.

10.3. Typical Failure Modes

We observe several characteristic failure cases:

- **Ambiguous material appearance:** for visually ambiguous fabrics (e.g., cotton vs. viscose plain-weave), the model occasionally mispredicts the dominant fiber, leading to slightly over- or under-damped dynamics.
- **Highly layered garments:** garments with multiple visible layers (e.g., jackets worn over dresses) are filtered out in FTAG, but in-the-wild images may still contain layering; in such cases, our single-layer physics may not fully capture the observed dynamics.
- **Extreme accessories and trims:** heavy buttons, zippers, or appliqués are not explicitly modeled and can lead to local discrepancies between simulated and real wrinkles near those regions.

These cases suggest interesting future directions such as explicit layer reasoning, accessory-aware physics modeling, and leveraging multi-view or temporal cues when available.

Table 4. **Quantitative comparison of image-to-garment prediction.** We show the results we get for Geometry (CD, IoU), image-space-reconstruction (PSNR, SSIM, LPIPS), and physics parameter accuracy (NMAE) with respect to the ground truth of our created data. Best results in **bold**. Arrows indicate optimization direction.

Sequence	Method	#Frames	CD↓	IoU↑	PSNR↑	SSIM↑	LPIPS↓
Jumping Jack	DMap*	133	2030	1.87	16.86	0.910	0.129
	AIparrel*		98.1	14.6	20.62	0.942	0.064
	ChatGarment*		88.9	20.0	21.66	0.951	0.056
	Ours		**64.4**	**21.6**	**22.10**	**0.954**	**0.053**
Joyful Jump	DMap*	91	–	–	–	–	–
	AIparrel*		46.8	17.4	24.64	0.965	0.034
	ChatGarment*		7.73	37.5	27.88	**0.970**	**0.021**
	Ours		**7.49**	**38.6**	**28.05**	**0.970**	**0.021**
Northern Spin	DMap*	125	–	–	–	–	–
	AIparrel*		257	**14.0**	17.86	**0.881**	0.156
	ChatGarment*		525	13.9	18.32	0.879	0.150
	Ours		**163**	12.4	**18.91**	0.866	**0.145**
Average	DMap*		2030	1.87	16.86	0.910	0.129
	AIparrel*		134.0	15.3	21.04	0.929	0.085
	ChatGarment*		207.2	23.8	22.62	**0.933**	0.076
	Ours		**78.3**	**24.2**	**23.02**	0.930	**0.073**

Table 5. **Ablation of the Physics Parameter Prediction.** Geometry (CD, IoU), image reconstruction (PSNR, SSIM, LPIPS), and physics parameter accuracy (NMAE) with respect to ground truth. Our method estimates physical parameters through a feedforward method, while Random Parameters are sampled within the bounds of the simulator. Best results in **bold**. Arrows indicate optimization direction.

Sequence	Method	#Frames	CD↓	IoU↑	PSNR↑	SSIM↑	LPIPS↓
Jab Cross	Random Parameters	105	5.42	46.1	29.04	0.968	0.020
	Ours		**2.50**	**58.4**	**30.99**	**0.974**	**0.014**
Jumping Jack	Random Parameters	133	79.4	22.2	22.14	0.951	0.053
	Ours		**31.2**	**28.3**	**24.63**	**0.960**	**0.035**
Northern Spin	Random Parameters	150	68.6	25.1	23.92	0.932	0.063
	Ours		**1.35**	**76.4**	**33.60**	**0.970**	**0.009**
Average	Random Parameters		51.1	31.1	25.03	0.950	0.045
	Ours		**11.7**	**54.4**	**29.74**	**0.968**	**0.019**

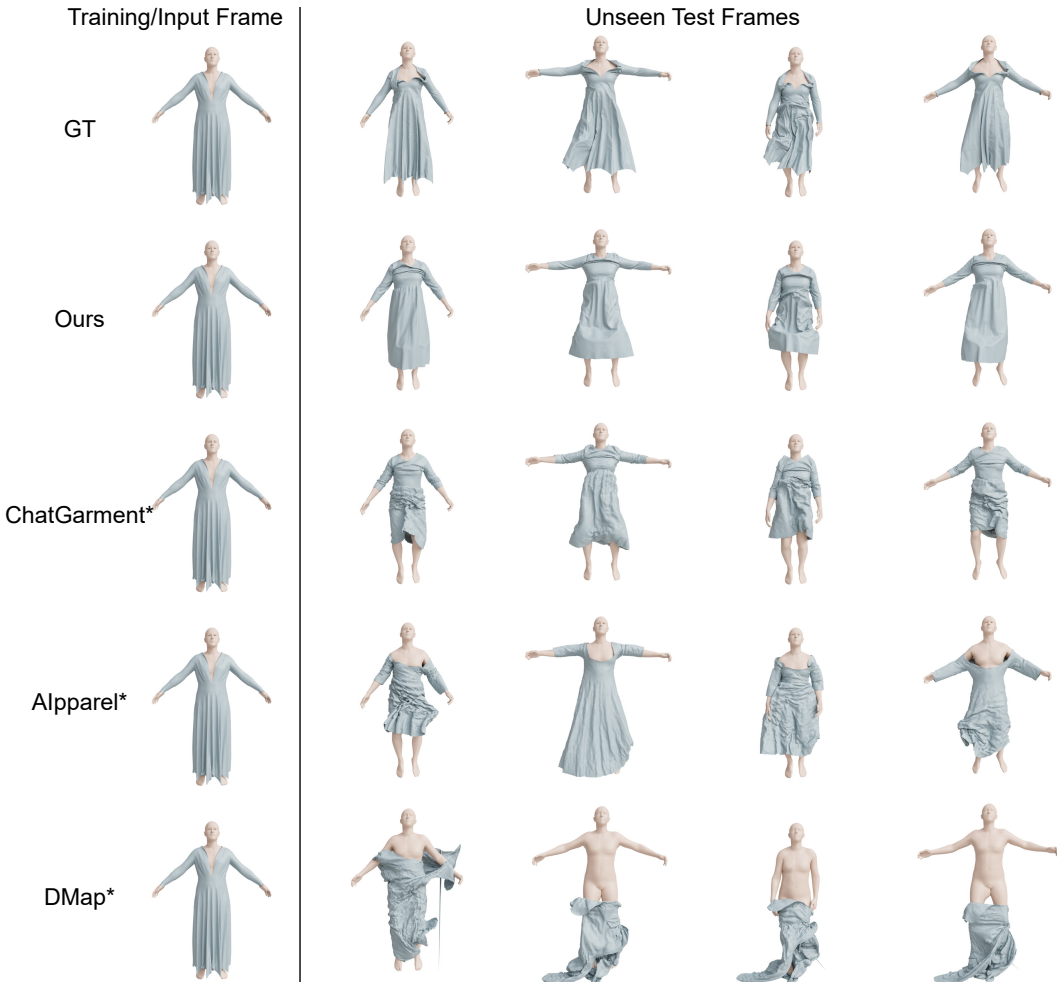

Figure 7. **Qualitative Comparison on Jumping Jack.** The left most column shows the input frames presented to each method and the columns on the right show the animated and rendered garments at different points in time of the animation. Our method's result most closely resemble the ground truth (top row).

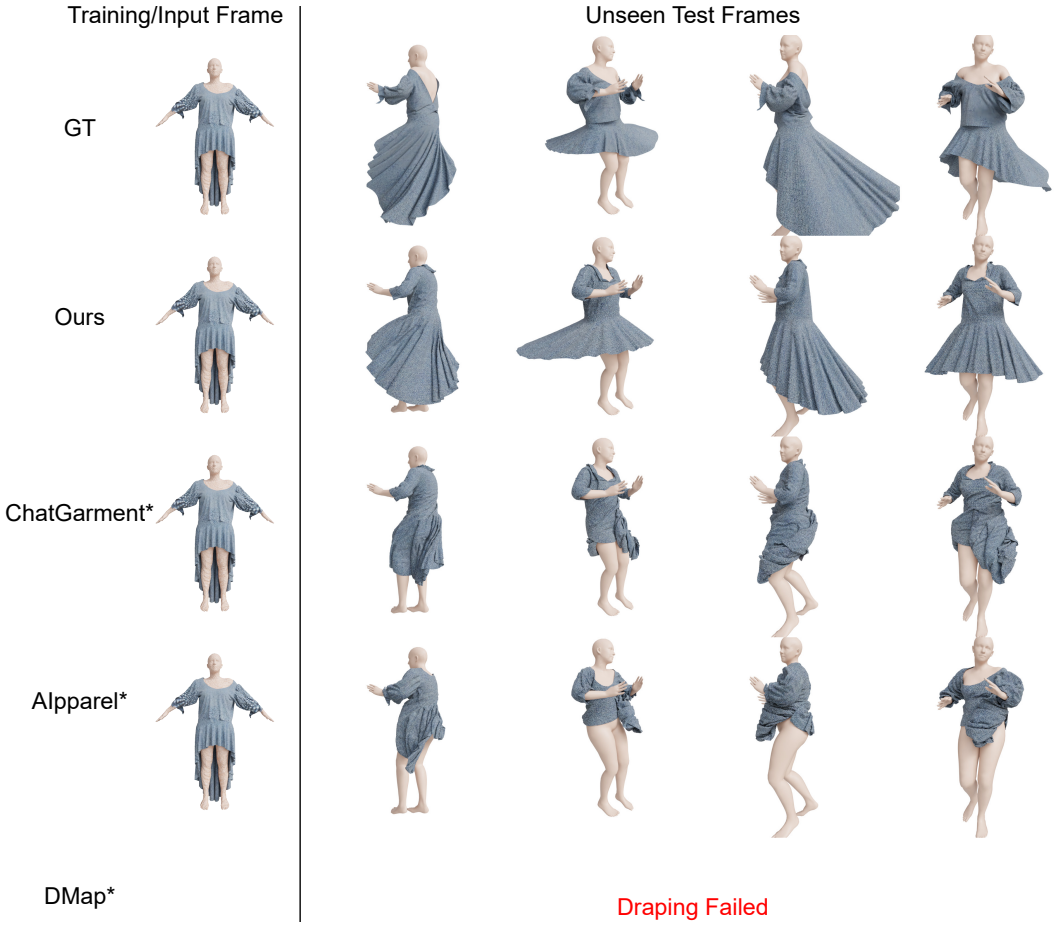

Figure 8. **Qualitative Comparison on Northern Spin.** The left most column shows the input frames presented to each method and the columns on the right show the animated and rendered garments at different points in time of the animation. Our method's result most closely resemble the ground truth (top row).

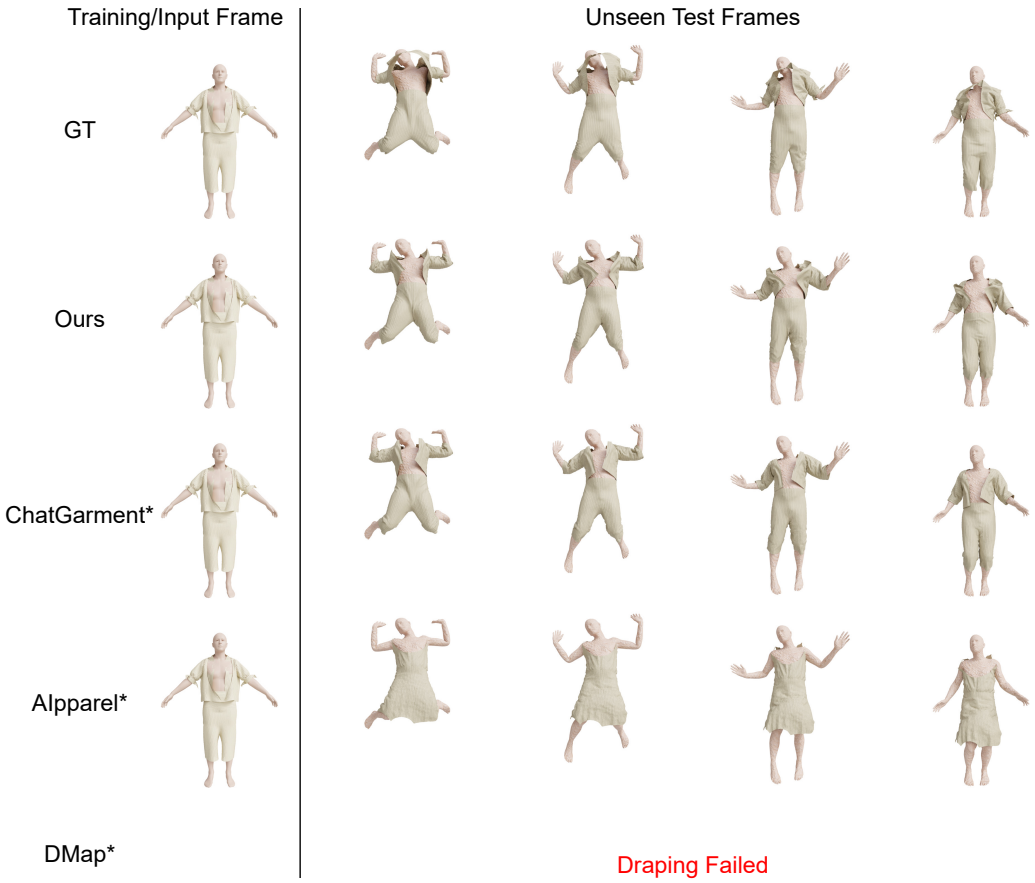

Figure 9. **Qualitative Comparison on Joyful Jumping.** The left most column shows the input frames presented to each method and the columns on the right show the animated and rendered garments at different points in time of the animation. Our method's result most closely resemble the ground truth (top row).

Figure 10. **Qualitative Comparison in-the-wild video.** The top row shows the original sequence. We only use the left-most frame as input. The rows below show renderings of the garments after simulation.

Figure 11. **Qualitative Comparison on in-the-wild video.** The top row shows the original sequence. We only use the left-most frame as input. The rows below show renderings of the garments after simulation.

Figure 12. **Qualitative Comparison on in-the-wild video.** The top row shows the original sequence. We only use the left-most frame as input. The rows below show renderings of the garments after simulation.