Hybrid Retrieval–Regression for Motion-Driven Loose-Fitting Garment Animation

Myeonjin Lee, Emmanuel Ian Libao and Sung-Hee Lee

Korea Advanced Institute of Science and Technology, Republic of Korea

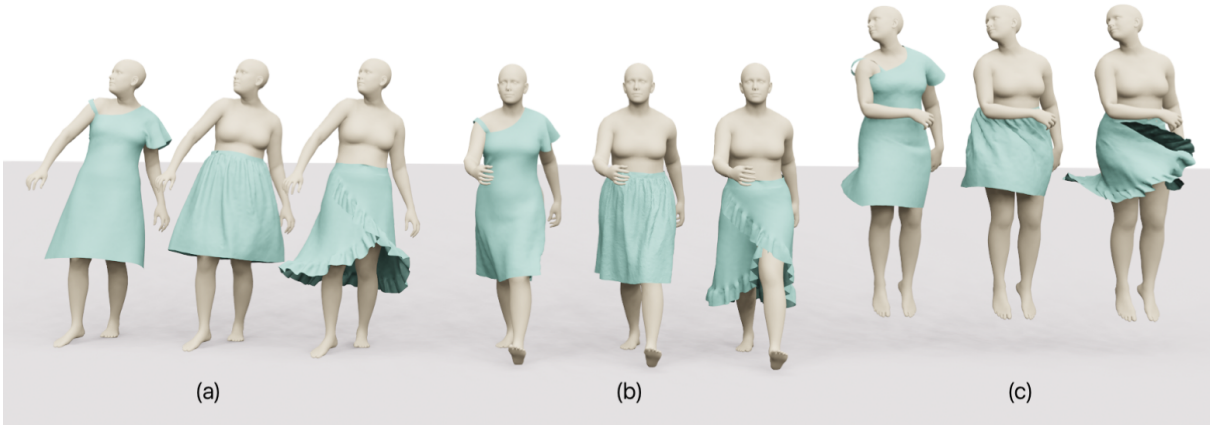

Figure 1: *Our hybrid retrieval-regression framework enables motion-driven animation of loose-fitting garments by leveraging a shared categorical latent space. The method addresses the challenge of capturing highly variable garment dynamics that are loosely correlated with body motion.*

Abstract

We present a hybrid retrieval-regression framework for motion-driven garment animation leveraging a shared discrete codebook. Our method targets the challenge of animating loose-fitting garments, whose dynamic behaviors exhibit high variability and less direct correlation with body motion—making them difficult to handle with conventional example-based approaches that assume tightly coupled motion–garment relationships. To address this, we project both motion and garment animation clips into a shared discrete codebook via Gumbel-Softmax-based quantization, allowing them to be aligned in a semantically consistent space where cross-retrieval can be performed using simple distance metrics. During inference, we adaptively switch between retrieval and regression based on the confidence derived from the codebook probability distribution, allowing the system to remain robust in the presence of ambiguous or unseen motions. We leverage a pre-trained mesh autoencoder to obtain garment latents that preserve local geometric structure, enabling smoother transitions and more geometrically consistent interpolation between retrieved and regressed animation segments efficiently. Experimental results demonstrate that our approach improves the accuracy and plausibility of garment animation for complex garments under diverse motion inputs, while maintaining robustness to unseen scenarios and achieving low simulation error for high-quality garment animation.

CCS Concepts

• ***Computing methodologies*** → ***Animation;*** *Learning latent representations;* ***Discrete space search;***

1. Introduction

In immersive applications such as games, virtual reality (VR), and interactive content, generating high-quality, real-time garment animation is critical to delivering realistic and engaging user experiences. Among various garment types, loose-fitting clothing poses unique challenges due to its rich and complex dynamics—including folding, flapping, and fluttering—driven not only by

DOI: 10.2312/pg.20251259

body motion but also by subtle physical forces such as acceleration and inertia. These behaviors are highly nonlinear and sensitive to motion variations, making them notoriously difficult to model accurately in real time. Even small prediction errors or motion noise can lead to visually implausible or unstable garment responses.

While physics-based simulation faithfully captures such intricate garment behaviors, its computational cost makes it impractical for real-time applications. Data-driven methods, particularly regression-based models, offer a promising alternative but suffer from two key limitations. First, they tend to oversmooth high-frequency garment motion, suppressing critical local details such as wrinkles and folds. Second, they are vulnerable to prediction instability, especially in autoregressive settings, where distributional shift accumulates over time, degrading the realism of long-term animations.

To address these issues, we shift the focus from numerical accuracy toward perceptual realism and dynamic plausibility. Rather than aiming to precisely replicate simulation outputs, our goal is to synthesize garment animations that are visually faithful and physically plausible, driven by motion.

We introduce a novel hybrid framework that blends the strengths of retrieval- and regression-based approaches within a shared, quantized latent space. At the core of our method is a Gumbel-Softmax-based vector quantization [JGP17], which discretizes continuous representations into learnable codebook entries that encodes motion and garment sequences into a discrete codebook. By restricting outputs to this discrete latent space, our model avoids the distribution drift and interpolation artifacts common in continuous regression, ensuring stable and semantically consistent garment behavior.

Furthermore, to address the intrinsic ambiguity in loose-fitting garments—where the same body motion can yield multiple plausible garment responses—we model the output as a probability distribution over discrete latent codes, enabling the system to express multiple plausible garment configurations for the same motion input via a learned codebook and Gumbel-Softmax-based vector quantization. This allows the system to express diverse garment dynamics such as asymmetric fluttering or localized folding, without compromising stability or coherence.

At inference time, we estimate the entropy of the predicted motion code distribution to assess confidence. For confident predictions, we retrieve a corresponding garment sequence from a latent codebook of precomputed examples. When confidence is low, we instead regress garment features from motion codes using a learned decoder. We further enhance temporal continuity by applying velocity-aware interpolation within a geometry-aware latent space, enabling smooth transitions between animation clips.

In summary, we present a hybrid garment animation framework that combines the visual sharpness of retrieval-based methods, the generalization capacity of regression-based models, and the stability and realism enabled by a quantized latent representation. Our approach enables expressive, robust, and physically plausible real-time garment animation driven solely by motion—without requiring expensive simulation. Our key contributions are as follows.

- **Hybrid Retrieval-Regression Framework** We propose a hybrid garment animation framework that combines retrieval and regression. Using a learned codebook, the system can either retrieve high-fidelity garment sequences from a database or generalize to unseen motions via regression.
- **Shared Discrete Latent Space** By encoding motion and garment sequences into a shared Gumbel-Softmax-based discrete codebook, we achieve semantically consistent alignment, enabling effective motion-to-garment cross-modal retrieval using simple distance metrics.
- **Confidence-Aware Hybrid Inference** We introduce an entropy-based confidence estimation mechanism that dynamically selects between the retrieval and regression mode, ensuring stable performance under diverse input conditions.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents experimental results that validate the effectiveness of our approach. Section 5 presents our conclusion, and Section 6 discusses the limitation of our work and future research directions.

2. Related Works

2.1. Physics-based Cloth Simulation

Cloth simulation has been traditionally done using physics-based simulations. Physics-based simulation involves modeling the forces acting on the cloth along with constraints and performing numerical methods of time integration. After the pioneering work of [BW98] that first introduced cloth simulation in computer graphics, various improvements have been introduced for accuracy and performance [VMTF09, WRK*10, WOR11, NSO12, TwL*18]. Recent advances in physics-based simulation have focused on differential simulators [LLK19, LDW*22, YZL*23] which enable a range of downstream applications in cloth simulation like system identification, trajectory optimization, closed-loop control and robotic manipulation. However, despite their lifelike wrinkle accuracy, these physics-based approaches generally remains more computationally expensive than learning-based methods.

2.2. Learning-based Cloth Simulation

Learning-based methods have been introduced to enable interactive and real-time cloth simulation, typically trained using ground-truth data from physics-based simulators or energy-based loss functions. A common strategy involves representing garments in relation to the human body using skinning weights [ZCM22, LTY*23a, LTY*23b] or offset displacements [SOC19, PLPM20, BME21]. While effective, these approaches are inherently limited to tight-fitting garments that should conform closely to the body.

To overcome this limitation, recent works have explored graph-based representations by using Graph Neural Networks (GNNs) models. This formulation allows for a more generalized representation that supports garments or fabrics of varying sizes and topologies [PFSB21]. Some methods further integrate physics-based information into the graph structure to better generalize on different motions [GBH23, LLKL23, TB23].

Other physics-informed approaches introduce loss functions

inspired by physical principles—modeling internal forces like stretching and bending, as well as external forces such as gravity and collisions [GCS*19, BME21, SOC22, BME22]. Notably, several self-supervised methods have emerged that eliminate the need for explicit ground-truth data by designing loss functions that approximate the underlying energy behavior of cloth, enabling the network to learn dynamics directly from the simulation process [SOC22, BME22, GBH23] .

Further research has focused on modeling loose-fitting garments, which exhibit complex deformations that do not tightly follow body movement [WSFM19, ZWCM21, PMJ*22, ZCM22]. While these methods show promise, they could easily struggle under highly dynamic conditions, sometimes producing unstable or temporally incoherent results. In general, learning-based methods lose fine-grained details in the results due to the approximation of the learned dynamics compared with the exact modeling of dynamics in traditional physics-based simulation.

2.3. Hybrid and Retrieval-based Cloth Simulation

Example-based garment simulation was introduced by [WHRO10] to animate tight-fitting garments by combining synthesized high-frequency wrinkles from a precomputed dataset with realtime coarse cloth simulation. This approach achieves visually detailed results at interactive rates for video game applications. However, it is constrained to tight-fitted garments where wrinkles are directly caused by body motion and would struggle with loose garments like skirts and dresses where a precomputed database cannot express the widely varying dynamics. It is also limited to poses represented in the database with interpolation providing only limited generalization.

This database-driven synthesis technique is also used in character animation through motion matching [Cla16], which selects appropriate motion clips from a database by matching features such as trajectory and velocity. This suffers from similar drawbacks of heavy dependence on large datasets and limited adaptability to unseen actions. Improvements have explored latent representations for motion models using LSTM [LLL18], CNN [HSK16, HKS17], and transformer [PBV21] architectures. While these latent representations capture smooth transitions and generalizations, they suffer from smoothed-out artifacts common to learning-based methods.

Hybrid frameworks utilizing vector quantization have emerged as a promising solution, representing continuous vectors through finite discrete codebook entries [VDOV*17] to address limitations of both database-driven synthesis and purely learning-based models. These approaches have been successfully applied to body motion synthesis [CKP*21, JCL*23, ZZLH23, YSZ*24], with techniques like Gumbel-Softmax [JGP17] enabling stable training through categorical reparameterization. Cross-modal representation learning has also been explored in garment design applications, where shared latent spaces enable unified modeling across different garment types [L*18].

Building on these advances, our work explores a hybrid framework that leverages vector quantization and shared representation learning specifically for loose-fitting garment animation, addressing the unique challenges of highly variable cloth dynamics that are loosely correlated with body motion.

3. Method

3.1. Overview

We propose a hybrid retrieval–regression framework for generating the motion-driven dynamics of loose-fitting garments in a reliable and high-fidelity manner. Unlike tight garments, loose garments exhibit nonlinear and indirect correlations with body motion, making example-based retrieval methods ineffective. The inherent variability and complex dynamics of loose-fitting garments mean that motion alone is insufficient to construct a meaningful latent space, as similar motions can produce different garment motions depending on previous states. In particular, defining a meaningful similarity metric to retrieve the most aligned garment sequence remains a significant challenge due to the inherently ambiguous and dynamic nature of loose garment behavior.

To address this, we observe that individual frame-based alignment is insufficient for loose garments, as their deformation depends on temporal context. We therefore process both motion and garment sequences as temporal windows and embed them into a shared discrete latent space using vector quantization [VDOV*17], allowing semantically similar motion-garment pairs to be represented by the same discrete code using a learned codebook via Gumbel-Softmax-based vector quantization. This enables cross-modal retrieval through simple similarity search in discrete latent space.

3.2. Garment Representation

Garment meshes inherently exhibit a high degree of freedom, making compression an essential step in downstream processing. To this end, we adopt a mesh-based autoencoder, specifically the Fully Convolutional Mesh Autoencoder (FCMAE) [ZWL*20]. This model learns a structured latent space from 3D garment meshes and is capable of accurately reconstructing both global shape and fine-grained local geometry. The learned latent space is fully localized, allowing for meaningful interpretation of different garment parts, and supports smooth interpolation with minimal visual artifacts. Using this model, we encode garments in each frame into compact, fixed-dimensional latent vectors, effectively replacing the high-dimensional mesh representation with a low-dimensional latent encoding for all subsequent processing. This transformation enables us to work with garment-motion pairs in the latent space, where garment dynamics are represented as sequences of latent vectors rather than raw mesh coordinates. For convenience, hereafter we refer to this latent representation simply as the garment, omitting the term "latent" unless otherwise specified.

3.3. Motion Feature Representation

We extract motion features from body movements to build representations that facilitate matching with dynamic garment sequence representations over time. Each motion frame is represented as a

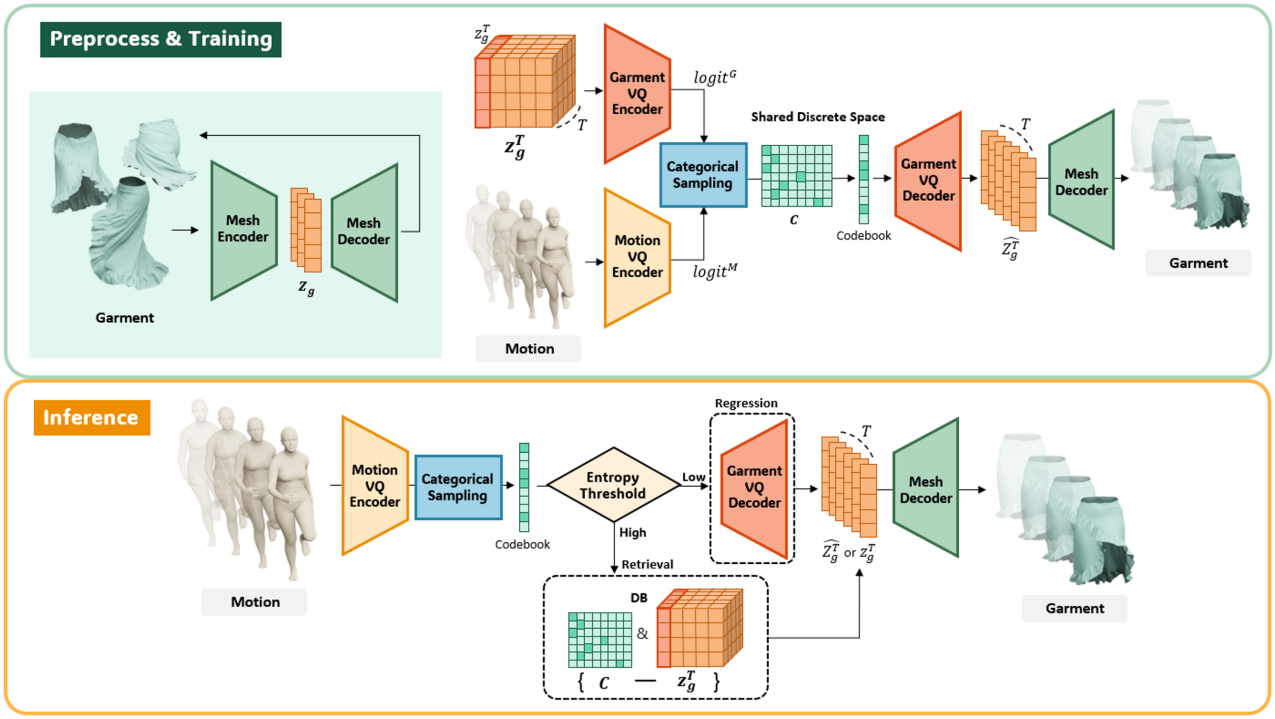

Figure 2: *Overview of our framework showing preprocessing & training (top) and inference (bottom) stages. During training, motion and garment sequences are encoded into a shared discrete codebook. At inference, entropy-based confidence switching determines between retrieval and regression modes.*

feature vector $\mathbf{f} \in \mathbb{R}^{124}$ that primarily encodes joint posture and dynamic motion characteristics. Specifically, for 20 major joints (excluding hands and feet), we convert the rotation matrices defined in the **root-relative** coordinate system into axis-angle representations $\mathbf{r} \in \mathbb{R}^{60}$ to quantify joint posture. Joint velocity is approximated using finite differences of root-relative joint positions between consecutive frames, computed as $v_t = (p_t - p_{t-1})/\Delta t$ where Δt is the frame interval, yielding $\mathbf{v} \in \mathbb{R}^{60}$. Additionally, to capture the global movement and rotational dynamics of the body, we include the global translation velocity of the root joint $\mathbf{t} \in \mathbb{R}^3$, as well as the yaw angular velocity around the vertical axis $\omega \in \mathbb{R}^1$, capturing global body rotation dynamics. The complete feature vector for each frame is formed as $\mathbf{f_t} = [\mathbf{r}; \mathbf{v}; \mathbf{t}; \omega] \in \mathbb{R}^{124}$. For temporal window processing, we concatenate W consecutive frames to form a window feature $\mathbf{F_t} = [\mathbf{f_t}, \mathbf{f_{t+1}}, \ldots, \mathbf{f_{t+W-1}}] \in \mathbb{R}^{124 \times W}$, which is then fed into a GRU-based encoder to capture temporal context across the window. Here we use $W = 10$ frames (0.33 seconds at 30 FPS) as the window size.

3.4. Shared Discrete Latent Space

The core of our framework lies in encoding garment and motion sequences separately, while enforcing semantically aligned pairs to share the same discrete latent code. To achieve this, we propose the following structure.

- **Garment Encoder.** Given a latent garment sequence extracted through a pre-trained mesh autoencoder, this module produces logits that map into a shared discrete codebook space, allowing alignment with motion representations.
- **Motion Encoder.** For a window of motion frames, this encoder outputs codebook logits using a structure analogous to the garment encoder. To better capture the diverse temporal dynamics inherent in loose garments, we adopt a GRU-based architecture refined with temporal attention.
- **Gumbel-Softmax Quantization.** We apply Gumbel-Softmax to the logits produced by both encoders, yielding soft probability distributions and corresponding hard one-hot vectors via the Straight-Through Estimator (STE). The soft vector represents the categorical distribution over codebook entries for differentiable training, while the hard vector provides discrete selections for stable inference, following the standard Gumbel-Softmax approach [JGP17]. This quantization process allows both encoders to operate within a shared codebook space, encouraging semantically aligned motion-garment pairs to share the same discrete code during training.
- **Loss functions** The model is optimized using both reconstruction loss and a code matching loss.

Reconstruction Loss Let $\mathbf{z}_g$ be the garment latent sequence encoded from the mesh autoencoder and $\hat{\mathbf{z}}_g$ be its reconstruction from the VQ decoder. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{z}_g - \hat{\mathbf{z}}_g\|_2^2 \tag{1}$$

Code Matching Loss Let $\mathbf{c}_g$ and $\mathbf{c}_m$ be the one-hot hard codes (after Gumbel-Softmax + STE) selected by the garment and motion encoders respectively. The code matching loss minimizes the Euclidean distance between their corresponding embeddings in the codebook $\mathcal{E}$:

$$\mathcal{L}_{\text{match}} = \|\mathcal{E}(\mathbf{c}_g) - \mathcal{E}(\mathbf{c}_m)\|_2^2 \tag{2}$$

To balance reconstruction quality and cross-modal alignment, we employ a two-stage training strategy following the curriculum learning approach [BVJS15]. Initially, we assign higher weight to the reconstruction loss ($\lambda_{rec} = 1.0$, $\lambda_{match} = 0.1$) to stabilize the garment latent space learning. When the reconstruction loss plateaus, indicating sufficient latent space stability, we increase the alignment weight ($\lambda_{match} = 0.5$) while maintaining the reconstruction weight, emphasizing cross-modal alignment to achieve semantic correspondence between motion and garment representations.

3.5. Confidence-aware Hybrid Inference

During inference, the input motion sequence is first mapped to a discrete latent code via the motion encoder. The entropy of the resulting categorical distribution is then analyzed to determine one of the following two inference paths:

Retrieval If the entropy is low—indicating a confident selection of a specific code—we retrieve the most similar garment code from a precomputed codebook database using cosine similarity, which is then mapped to the corresponding garment latent sequence.

Regression If the entropy is high - implying uncertainty in code selection - we bypass the retrieval and directly regress the garment latent sequence by feeding the motion code into the VQ decoder.

Let $\mathbf{p} \in \mathbb{R}^K$ denote the softmax probability distribution output by the motion encoder, where K represents the size of the discrete codebook. We compute the entropy $H(\mathbf{p})$ as:

$$H(\mathbf{p}) = -\sum_{k=1}^{K} p_k \log p_k \tag{3}$$

Hybrid Inference Strategy Based on the entropy, we choose the inference path:

$$\hat{\mathbf{z}}_g = \begin{cases} \text{Retrieve}(\mathbf{p}) & \text{if } H(\mathbf{p}) < \tau \\ \text{Decoder}(\mathbf{p}) & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{\mathbf{z}}_g$ denotes the inferred garment latent sequence and τ is a confidence threshold determined empirically.

The garment latent sequences are inferred on a window-by-window basis and concatenated to form the full animation with stride 1. To preserve temporal consistency across windows, overlapping frames are used to enable smooth blending at window boundaries. Importantly, our garment latent representation is built on a geometry-aware structure, which allows for interpolation across latent vectors without introducing visual artifacts or losing high-frequency geometric detail. The final garment mesh sequence is reconstructed via the decoder of a pretrained mesh autoencoder. This hybrid approach combines the sharpness and accuracy of retrieval-based methods with the generalization ability and robustness of regression, enabling stable and realistic garment animation under a wide variety of motion inputs.

Cloth	Vertices	Reduced Vertices	Sequences	Total Frames
Dress	12,146	114	56	7,117
Pleated Skirt	17,678	160	48	13,676
Asymmetric Skirt	10,018	96	89	40,342
VB Skirt02	10,018	96	89	40,342
VB Dress03	8,744	78	89	40,342

Table 1: *Dataset configuration showing five garment types with their mesh complexity, dimensionality reduction results, and sequence coverage.*

3.6. Model Implementation Details

All models were implemented in PyTorch and trained using the Adam optimizer with a learning rate of 1e-4 and a weight decay of 1e-4 for 1500 epochs. The garment encoder and decoder are composed of two-layer fully connected networks with 1024 hidden units and ELU activation. The motion encoder consists of a GRU backbone followed by a temporal attention refinement module. For each frame, the garment latent is defined as an 8-dimensional embedding per node, reduced by approximately 95% from the original mesh representation, and derived from a pre-trained mesh autoencoder. We use a temporal window size of 10 frames and a shared quantized codebook of size 4096 (comprising 512 channels with 8 dimensions per channel). All experiments were conducted on a single NVIDIA RTX 4090 GPU, with total training time of approximately 23 hours. Inference runs at 130 FPS or more, benefiting from optimized libraries such as FAISS, although performance may vary depending on garment type and retrieval conditions. Please refer to the supplementary material for full implementation details of the encoder, estimator, decoder, and the spatio-temporal attention architecture.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed hybrid garment animation framework. We begin by describing the dataset configuration, then compare our model with the existing state-of-the-art regression-based approach, dubbed as VirtualBones [PMJ*22], demonstrating superior accuracy and visual realism. Next, we conduct ablation studies to investigate the role of each component, specifically the contributions of retrieval and regression modes on animation accuracy and stability. To highlight the importance of learning a shared discrete latent space, we compare our approach against naive motion-to-garment retrieval baselines that operate using nearest-neighbor search directly on raw motion features. This comparison demonstrates substantial improvements in retrieval accuracy enabled by semantic alignment in the latent space. These experiments validate the effectiveness, expressiveness, and robustness of our method for generating realistic loose-fitting garment animations guided by motion input.

4.1. Dataset Configuration

We evaluate our method on five types of loose-fitting garments, each simulated on the default SMPL body model [LMR*15], ex-

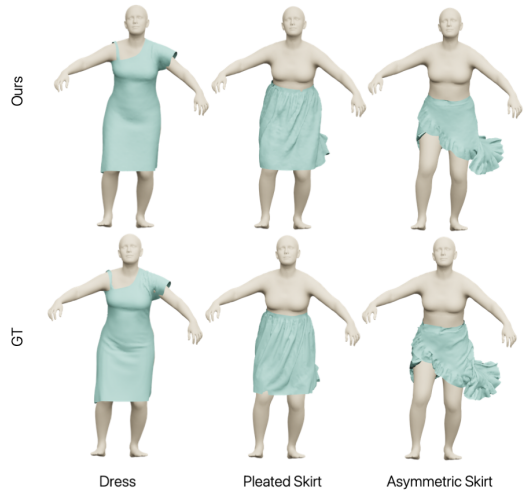

Figure 3: *Our method captures realistic garment dynamics including wrinkles and fluttering, demonstrating superior accuracy and visual realism.*

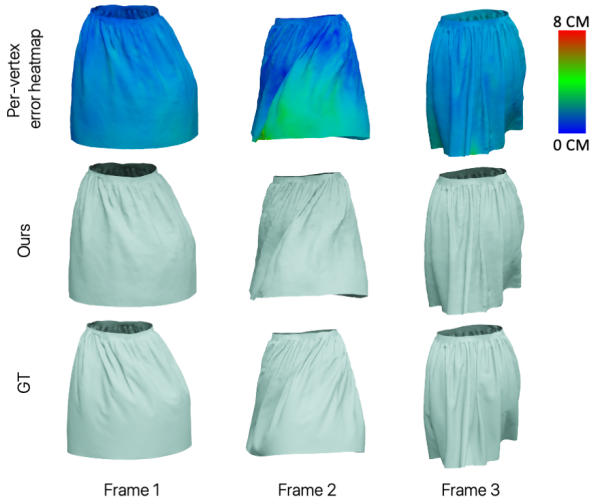

Figure 4: *Heatmap image showing the deviation of our method from the ground truth simulation.*

cept for VB Skirt02 and VB Dress03, which use the in-house body model from VirtualBones [PMJ*22]. The Asymmetric Skirt represents a resized version of VB Skirt02, adapted to fit the SMPL body.

The garments simulated on SMPL body were generated using CLO3D at 30 FPS with simulation parameters: collision distance of 4mm, iteration count of 30, and time step of 0.003333s. VB Skirt02 and VB Dress03 are retained in their original configurations to enable direct performance comparison with VirtualBones [PMJ*22]. Each garment type was trained separately due to topological differences between garment meshes.

For evaluation, we reserved 3-5 motion sequences per garment type as completely separate test sets, with the remaining data split 9:1 for training and validation to ensure reliable assessment. Our dataset configuration is summarized in Table 1.

4.2. Results and Comparison

Figure 3 and 4 demonstrate that our method generates sharp and realistic dynamics such as fluttering and folding, capturing the complex nonlinear behaviors characteristic of loose-fitting garments. This was possible thanks to our retrieval-based designs that leverages the superior reconstruction capabilities of the mesh reconstruction model, enabling highly accurate simulation of dynamic regions such as garment hems with severe wrinkles.

Quantitative comparison with VirtualBones in Table 2 shows that our model achieves significantly lower RMSE and per-vertex errors. As in the Figure 5, our method consistently preserves high-frequency geometric features essential for realistic animation, while the regression-based approach produces smoothed or unrealistic surface artifacts that fail to capture fine details.

Table 2: *Comparison results across different garment types.*

Garment Type	Method	RMSE (mm) (↓)	Per Vertex Error (mm) (↓)	STED (↓)
VB Skirt02	VirtualBones	20.27	15.58	0.371
	Ours	**5.81**	**4.66**	**0.316**
VB Dress03	VirtualBones	13.32	10.06	0.881
	Ours	**3.342**	**2.65**	**0.155**

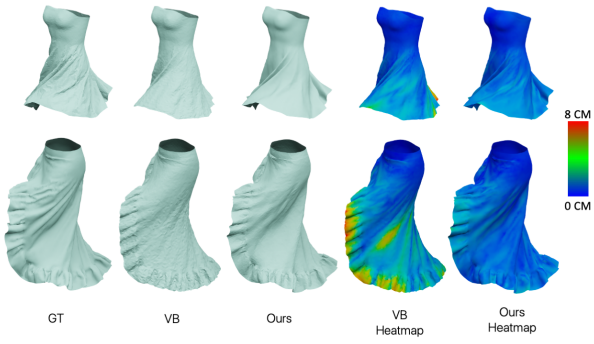

Figure 5: *Qualitative comparison showing our method's superior preservation of fine-grained details such as wrinkles and dynamic folding patterns compared to other baseline, particularly visible in high-motion areas like garment hems.*

4.3. Ablation Studies

4.3.1. Effect of Hybrid Design

To validate our hybrid design, we evaluate two ablated versions and a naive baseline. **Ours (w/o retrieval)** disables retrieval and uses only the VQ decoder for direct garment latent regression, while **Ours (w/o regression)** performs database lookup without regression fallback using cosine similarity in the quantized latent space. We also compare against **Naive NN Retrieval**, which uses direct

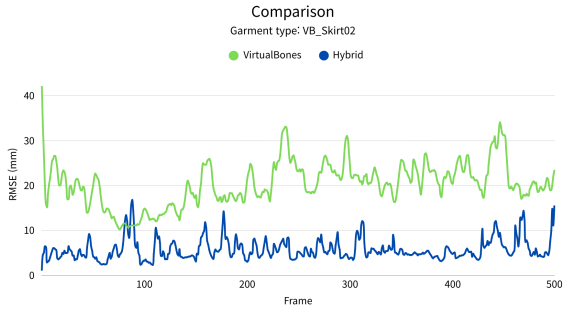

Figure 6: *Our model achieves significantly lower RMSE and Per Vertex Errors compared to regression-based model like VirtualBones.*

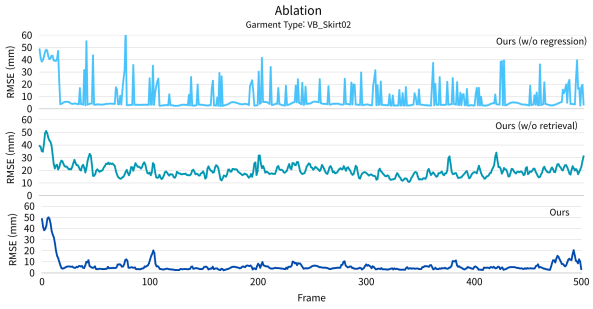

Figure 7: *Ablation study results showing the complementary nature of retrieval and regression components. Pure retrieval exhibits high variance due to occasional mismatches, while pure regression shows over-smoothing artifacts.*

motion features for nearest-neighbor search without learned codebook alignment. These comparisons demonstrate the necessity of both retrieval and regression components in our hybrid framework.

Ours vs. Ours (w/o retrieval). Disabling the retrieval path forces the VQ decoder to regress garment latents directly from codebook vectors. As shown in Table 3, this leads to degraded fine-grained details and noticeable error increase. The absence of retrieval results in smoothing the dynamic garment behavior, demonstrating the critical role of retrieval in preserving dynamic garment characteristics.

Ours vs. Ours (w/o regression). Relying solely on nearest-neighbor retrieval based on quantized motion codes fails under uncertain or ambiguous motions, often retrieving incorrect garment states and resulting in temporally inconsistent animations, shown as high temporal fluctuations in Figure 7. This confirms that regression fallback is essential for handling uncertain cases where retrieval would produce artifacts.

		Ours (w/o ret)	Ours (w/o reg)	Hybrid (Ours)
Pleated Skirt	RMSE (mm) ↓	10.86	5.17	4.16
	Vertex Error (mm) ↓	8.98	4.13	**3.29**
Asymmetric Skirt	RMSE (mm) ↓	16.85	10.18	**8.28**
	Vertex Error (mm) ↓	13.89	8.39	**6.78**

Table 3: *Ablation study comparing Ours (w/o retrieval), Ours (w/o regression), and hybrid approaches on two garment types, demonstrating the effectiveness of selective retrieval-regression switching.*

4.3.2. Confidence-Based Switching Mechanism

Our hybrid design requires a switching mechanism to decide when to use retrieval versus regression. We employ entropy from the retrieval distribution as a confidence measure: low entropy indicates confident retrieval, while high entropy suggests uncertainty requiring regression fallback.

We validate this mechanism by evaluating RMSE across different entropy thresholds ($\tau \in [0, 1)$) using 10-frame sliding windows with stride 1. When the entropy calculated from a given motion clip falls below τ, retrieval is used; otherwise, regression is applied. Retrieval usage is measured as the percentage of windows where the entropy falls below the threshold. As shown in Table 4, $\tau = 0.5$ achieves optimal RMSE performance, with even modest retrieval usage leading to substantial error reduction compared to regression-only approaches. This confirms that confidence-based selective retrieval is crucial for balances the complementary strengths of both approaches.

Table 4: *Effect of confidence threshold τ on retrieval usage and reconstruction error, showing optimal performance at $\tau = 0.5$. This approves that confidence-based switching mecahnism is important for balancing the strengths of both approaches.*

Threshold (τ)	Retrieval Usage (%)	Average RMSE (mm)
0	0	20.93
0.2	28.64	10.87
0.4	92.26	13.82
0.5	**98.64**	**12.75**
0.6	99.6	14.27
0.8	100	14.21

4.3.3. Effect of Learned Latent Alignment

To validate the effectiveness of our shared discrete latent space, we evaluate cross-modal retrieval performance under challenging conditions where ground-truth motion-garment pairs are deliberately excluded from the retrieval database. This setup ensures fair evaluation of generalization capability rather than memorization.

We compare three approaches: **Naive NN** performs traditional nearest-neighbor search in raw motion space without learned alignment; **Ours (w/o regression)** uses our learned shared codebook for pure retrieval without regression fallback; and **Hybrid (Full)** combines learned alignment with confidence-based switching between retrieval and regression.

As shown in Table 5, the naive approach performs significantly worse, demonstrating that raw motion similarity does not guarantee semantically appropriate garment retrieval. Our learned alignment achieves substantial improvement by mapping motion-garment pairs into a shared discrete space where semantically similar pairs share the same codes. Notably, as illustrated in Fig-

ure 8, even with ground-truth pairs excluded from the database, our retrieval-only method successfully identifies garment sequences remarkably similar to ground truth, significantly outperforming naive matching. This demonstrates the effectiveness of our semantically aligned latent space in capturing meaningful motion-garment relationships beyond exact matches.

Garment Type		Naive NN	Ours (w/o reg)	Hybrid (Ours)
VB Dress03	RMSE (mm) ↓	37.32	32.66	**13.97**
	Per Vertex Error (mm) ↓	45.73	41.23	**17.63**
VB Skirt02	RMSE (mm) ↓	58.54	46.64	**16.96**
	Per Vertex Error (mm) ↓	74.14	61.86	**20.25**

Table 5: *Quantitative evaluation of motion-to-garment retrieval with ground-truth exclusion. The learned shared latent space significantly outperforms naive nearest-neighbor matching, while the hybrid approach achieves optimal performance through confidence-based retrieval-regression switching.*

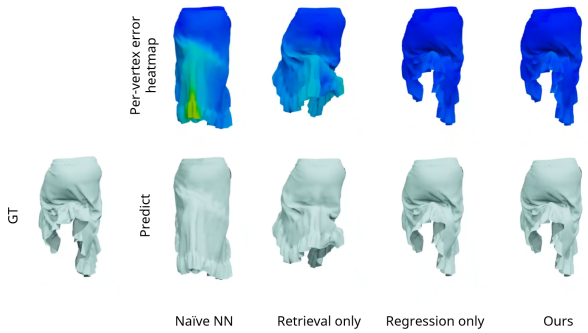

Figure 8: *Qualitative comparison of motion-to-garment retrieval with ground-truth exclusion. Our learned codebook significantly outperforms naive nearest-neighbor matching using shared motion-garment latent alignment, while the hybrid method achieves optimal results through confidence-aware switching.*

5. Conclusion

We presented a hybrid framework for real-time animation of loose-fitting garments that captures complex, nonlinear garment dynamics with high realism and stability. Our method unifies motion and garment sequences in a shared quantized latent space via Gumbel-Softmax-based vector quantization, enabling semantically aligned representations that support both high-fidelity retrieval and robust regression.

The key innovation is a confidence-aware hybrid inference mechanism that dynamically switches between retrieval and regression based on entropy-based confidence estimation. This design delivers sharp, realistic animation with preserved high-frequency geometric features for confident predictions while ensuring stable handling of unseen or ambiguous motions through regression fallback.

Comprehensive experiments demonstrate that our framework significantly outperforms state-of-the-art regression-based approaches like VirtualBones across multiple metrics including RMSE and per-vertex errors, while achieving fast inference time. Ablation studies confirm that both retrieval and regression components are essential, with our shared discrete latent space substantially improving retrieval accuracy compared to naive motion feature matching. The entropy-based switching mechanism with τ= 0.5 achieves optimal performance, demonstrating that even modest retrieval usage leads to substantial error reduction.

Our results establish that combining discrete latent modeling with hybrid inference provides an effective solution for data-driven garment animation, generating realistic loose-fitting garment dynamics without the computational cost of physics-based simulation. This approach opens new possibilities for real-time applications requiring high-quality garment animation with complex nonlinear behaviors.

6. Limitation and Future Work

While our hybrid framework effectively generates accurate and generalizable animation of loose-fitting garments, several limitations present promising directions for future research.

Generalization to Body-Garment Variations. Our model is currently trained on fixed body shapes and garment combinations, limiting adaptability to variations in body morphology or garment fitting. Future work could explore body-invariant garment representations or conditional models that enable animation across diverse body-garment configurations without retraining.

Memory and Inference Efficiency. The hybrid framework incorporates multiple components—codebook, garment database, and multiple encoders—introducing memory and computational overhead. Future improvements could involve codebook compression, latent space distillation, or architectural simplification to achieve a more lightweight system while maintaining performance.

Explicit Physical Collision Handling. Our approach relies on pre-simulated data where collisions are implicitly resolved, but lacks explicit collision-handling mechanisms during inference. This may result in garment-body penetrations or unrealistic intersections in novel scenarios. Future directions include incorporating collision-aware loss functions or post-processing refinement modules to ensure physically plausible behavior across diverse motion inputs.

Acknowledgments. This work was supported by Graduate School of Metaverse Convergence support program, IITP, Korea (IITP-2022(2025)-RS-2022-00156435) and CT R&D program, KOCCA, Korea (RS-2025-02307327).

References

[BME21] BERTICHE H., MADADI M., ESCALERA S.: Pbns: physically based neural simulation for unsupervised garment pose space deformation. *ACM Trans. Graph. 40*, 6 (Dec. 2021). URL: https://doi.org/10.1145/3478513.3480479, doi:10.1145/3478513.3480479. 2, 3

[BME22] BERTICHE H., MADADI M., ESCALERA S.: Neural cloth simulation. *ACM Trans. Graph. 41*, 6 (Nov. 2022). URL: https://doi.org/10.1145/3550454.3555491, doi:10.1145/3550454.3555491. 3

[BVJS15] BENGIO S., VINYALS O., JAITLY N., SHAZEER N.: Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems 28* (2015). 5

[BW98] BARAFF D., WITKIN A.: Large steps in cloth simulation. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1998), SIGGRAPH '98, Association for Computing Machinery, p. 43–54. URL: https://doi.org/10.1145/280814.280821, doi:10.1145/280814.280821. 2

[CKP*21] CHO K., KIM C., PARK J., PARK J., NOH J.: Motion recommendation for online character control. *ACM Trans. Graph. 40*, 6 (Dec. 2021). URL: https://doi.org/10.1145/3478513.3480512, doi:10.1145/3478513.3480512. 3

[Cla16] CLAVET S.: Motion matching and the road to next-gen animation. In *GDC* (2016). 3

[GBH23] GRIGOREV A., BLACK M. J., HILLIGES O.: Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16965–16974. doi:10.1109/CVPR52729.2023.01627. 2, 3

[GCS*19] GUNDOGDU E., CONSTANTIN V., SEIFODDINI A., DANG M., SALZMANN M., FUA P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 8738–8747. doi:10.1109/ICCV.2019.00883. 3

[HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Trans. Graph. 36*, 4 (July 2017). URL: https://doi.org/10.1145/3072959.3073663, doi:10.1145/3072959.3073663. 3

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph. 35*, 4 (July 2016). URL: https://doi.org/10.1145/2897824.2925975, doi:10.1145/2897824.2925975. 3

[JCL*23] JIANG B., CHEN X., LIU W., YU J., YU G., CHEN T.: Motiongpt: human motion as a foreign language. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2023), NIPS '23, Curran Associates Inc. 3

[JGP17] JANG E., GU S., POOLE B.: Categorical reparameterization with gumbel-softmax. In *ICLR* (2017). 2, 3, 4

[LDW*22] LI Y., DU T., WU K., XU J., MATUSIK W.: Diffcloth: Differentiable cloth simulation with dry frictional contact. *ACM Trans. Graph. 42*, 1 (Oct. 2022). URL: https://doi.org/10.1145/3527660, doi:10.1145/3527660. 2

[LLK19] LIANG J., LIN M., KOLTUN V.: Differentiable cloth simulation for inverse problems. In *Advances in Neural Information Processing Systems* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.), vol. 32, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/28f0b864598a1291557bed248a998d4e-Paper.pdf. 2

[LLKL23] LIBAO E. I., LEE M., KIM S., LEE S.-H.: Meshgraphnetrp: Improving generalization of gnn-based cloth simulation. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games* (New York, NY, USA, 2023), MIG '23, Association for Computing Machinery. URL: https://doi.org/10.1145/3623264.3624441, doi:10.1145/3623264.3624441. 2

[LLL18] LEE K., LEE S., LEE J.: Interactive character animation by learning multi-objective control. *ACM Trans. Graph. 37*, 6 (Dec. 2018). URL: https://doi.org/10.1145/3272127.3275071, doi:10.1145/3272127.3275071. 3

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG) 34*, 6 (2015), 1–16. 5

[LTY*23a] LI Y., TANG M., YANG Y., TONG R., AN B., YANG S., LI Y., KOU Q.: D-Cloth: Skinning-based cloth dynamic prediction with a three-stage network. *Computer Graphics Forum 42*, 7 (2023), e14937. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14937. 2

[LTY*23b] LI Y., TANG M., YANG Y., TONG R., YANG S., LI Y., AN B., KOU Q.: CTSN: Predicting cloth deformation for skeleton-based characters with a two-stream skinning network. *Computational Visual Media (Proceedings of CVM 2023)* (2023). 2

[L*18] LI M., ÇETINASLAN O., GUERRERO P., ET AL.: Learning a shared shape space for multimodal garment design. *ACM Transactions on Graphics* (2018). 3

[NSO12] NARAIN R., SAMII A., O'BRIEN J. F.: Adaptive anisotropic remeshing for cloth simulation. *ACM Trans. Graph. 31*, 6 (Nov. 2012). URL: https://doi.org/10.1145/2366145.2366171, doi:10.1145/2366145.2366171. 2

[PBV21] PETROVICH M., BLACK M. J., VAROL G.: Action-Conditioned 3D Human Motion Synthesis with Transformer VAE . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA, USA, Oct. 2021), IEEE Computer Society, pp. 10965–10975. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01080, doi:10.1109/ICCV48922.2021.01080. 3

[PFSB21] PFAFF T., FORTUNATO M., SANCHEZ-GONZALEZ A., BATTAGLIA P. W.: Learning mesh-based simulation with graph networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (2021), OpenReview.net. URL: https://openreview.net/forum?id=roNqYL0_XP. 2

[PLPM20] PATEL C., LIAO Z., PONS-MOLL G.: TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, June 2020), IEEE Computer Society, pp. 7363–7373. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00739, doi:10.1109/CVPR42600.2020.00739. 2

[PMJ*22] PAN X., MAI J., JIANG X., TANG D., LI J., SHAO T., ZHOU K., JIN X., MANOCHA D.: Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 Conference Proceedings* (New York, NY, USA, 2022), SIGGRAPH '22, Association for Computing Machinery. URL: https://doi.org/10.1145/3528233.3530709, doi:10.1145/3528233.3530709. 3, 5, 6

[SOC19] SANTESTEBAN I., OTADUY M. A., CASAS D.: Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum 38*, 2 (2019), 355–366. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13643, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13643, doi:https://doi.org/10.1111/cgf.13643. 2

[SOC22] SANTESTEBAN I., OTADUY M. A., CASAS D.: Snug: Self-supervised neural dynamic garments. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 8130–8140. doi:10.1109/CVPR52688.2022.00797. 3

[TB23] TIWARI L., BHOWMICK B.: Garsim: Particle based neural garment simulator. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2023), pp. 4472–4481. 2

[TwL*18] TANG M., WANG T., LIU Z., TONG R., MANOCHA D.: I-cloth: incremental collision handling for gpu-based interactive cloth simulation. *ACM Trans. Graph. 37*, 6 (Dec. 2018). URL: https://doi.org/10.1145/3272127.3275005, doi:10.1145/3272127.3275005. 2

[VDOV*17] VAN DEN OORD A., VINYALS O., ET AL.: Neural discrete representation learning. *Advances in neural information processing systems 30* (2017). 3

[VMTF09] VOLINO P., MAGNENAT-THALMANN N., FAURE F.: A simple approach to nonlinear tensile stiffness for accurate cloth simulation. *ACM Trans. Graph. 28*, 4 (Sept. 2009). URL: https://doi.org/10.1145/1559755.1559762, doi:10.1145/1559755.1559762. 2

[WHRO10] WANG H., HECHT F., RAMAMOORTHI R., O'BRIEN J. F.: Example-based wrinkle synthesis for clothing animation. *ACM Trans. Graph. 29*, 4 (July 2010). URL: https://doi.org/10.1145/1778765.1778844, doi:10.1145/1778765.1778844. 3

[WOR11] WANG H., O'BRIEN J. F., RAMAMOORTHI R.: Data-driven elastic models for cloth: modeling and measurement. In *ACM SIGGRAPH 2011 Papers* (New York, NY, USA, 2011), SIGGRAPH '11, Association for Computing Machinery. URL: https://doi.org/10.1145/1964921.1964966, doi:10.1145/1964921.1964966. 2

[WRK*10] WICKE M., RITCHIE D., KLINGNER B. M., BURKE S., SHEWCHUK J. R., O'BRIEN J. F.: Dynamic local remeshing for elastoplastic simulation. *ACM Trans. Graph. 29*, 4 (July 2010). URL: https://doi.org/10.1145/1778765.1778786, doi:10.1145/1778765.1778786. 2

[WSFM19] WANG T. Y., SHAO T., FU K., MITRA N. J.: Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Trans. Graph. 38*, 6 (Nov. 2019). URL: https://doi.org/10.1145/3355089.3356512, doi:10.1145/3355089.3356512. 3

[YSZ*24] YAO H., SONG Z., ZHOU Y., AO T., CHEN B., LIU L.: Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Trans. Graph. 43*, 4 (July 2024). URL: https://doi.org/10.1145/3658137, doi:10.1145/3658137. 3

[YZL*23] YU X., ZHAO S., LUO S., YANG G., SHAO L.: Diffclothai: Differentiable cloth simulation with intersection-free frictional contact and differentiable two-way coupling with articulated rigid bodies. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2023), pp. 400–407. doi:10.1109/IROS55552.2023.10341573. 2

[ZCM22] ZHANG M., CEYLAN D., MITRA N. J.: Motion guided deep dynamic 3d garments. *ACM Trans. Graph. 41*, 6 (Nov. 2022). URL: https://doi.org/10.1145/3550454.3555485, doi:10.1145/3550454.3555485. 2, 3

[ZWCM21] ZHANG M., WANG T., CEYLAN D., MITRA N. J.: Deep detail enhancement for any garment. *Computer Graphics Forum 40*, 2 (2021), 399–411. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.142642, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.142642, doi:https://doi.org/10.1111/cgf.142642. 3

[ZWL*20] ZHOU Y., WU C., LI Z., CAO C., YE Y., SARAGIH J., LI H., SHEIKH Y.: Fully convolutional mesh autoencoder using efficient spatially varying kernels. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2020), NIPS '20, Curran Associates Inc. 3

[ZZLH23] ZHU Q., ZHANG H., LAN M., HAN L.: Neural categorical priors for physics-based character control. *ACM Trans. Graph. 42*, 6 (Dec. 2023). URL: https://doi.org/10.1145/3618397, doi:10.1145/3618397. 3