HOOD: Hierarchical Graphs for Generalized Modelling of Clothing Dynamics

Artur Grigorev[1,2] Bernhard Thomaszewski[1] Michael J. Black[2]

Otmar Hilliges[1]

[1] ETH Zurich, Department of Computer Science
[2] Max Planck Institute for Intelligent Systems, Tubingen

1. Implementation details

1.1. Training details

Garment Initialization. During training, we want to step garment meshes forward in time from any point in the pose sequences. To evaluate the loss function at an arbitrary starting point, we must provide garment geometry for the two previous time steps. We approximate these geometries using linear blend skinning combined with the diffused body model formulation from [9]. We then remove any intersections between skinned garment meshes and the body. The initial garment meshes computed in this way are generally not in energetically optimal states. To reduce this internal energy, we find it useful to scale down the contribution of the inertia term using a coefficient $\alpha \in (0, 1]$. Using a smaller coefficient α for the first step during training iterations allows the garment to quickly relax into a state of lower potential energy. At inference time, when initialized from the resting pose, we always use $\alpha = 1$. We also pass α as an input to the network (see Sec. **??**) so it can adapt its prediction to different values. Finally, it should be noted that skinning is only used for initialization during training, not at inference time.

Normalization. Following [6], we find it crucial for convergence to normalize feature vectors at the beginning of each step using exponentially weighted averages of their mean and standard deviations. We also perform denormalization on the outputs to obtain accelerations. Since we do not have access to ground-truth data, we use statistics collected from linearly-skinned garments as described above.

Autoregressive Training. We start by predicting accelerations for only one next step, then gradually increase the number of predicted steps every k iterations up to 5. We set k to 5000 and find that predicting a maximum of 5 steps during training is enough for the model to autoregressively run for thousands of steps at inference time. In the supplementary material, we demonstrate the stability of our model for very long sequences.

1.2. Garments preprocessing

Hierarchical graph construction We tackle the problem of limited propagation radius in the message-passing networks by constructing several levels of coarsened garment graphs.

To generate coarse edges from the initial garment graph, we first find a centre node of a graph v_c, that is a node, whose eccentricity is equal to the graph's radius. Usually, garment graphs have more than one node with this property, so we randomly choose one of them. Then, for coarsened graph, we keep only those nodes, whose distance (i.e., the number of edges in the shortest path) to the centre is an even number and replace other nodes with edges between pairs of the kept nodes. We apply this process recursively to build several levels of coarse edges. See Algorithm 1 for details.

input : fine graph $G_f(V_f, E_f)$
output: coarse graph $G_c(V_c, E_c)$

$v_{center} \leftarrow$ center of G_f ;
for v_i *in* V_f **do**
  $d_i \leftarrow distance(v_{center}, v_i)$;
$V_c \leftarrow \{v_i \in V_f |\ d_i \mod 2 = 0\}$;
$E_c \leftarrow \{\}$;
$V^{interm} \leftarrow \{v_i \in V_f |\ d_i \mod 2 = 1\}$;
for v_i *in* V^{interm} **do**
  $V^{from} \leftarrow$
   $\{v_j \in V_f |\ e_{ij} \in E_f \ \ \& \ \ d_j = d_i - 1\}$;
  $V^{to} \leftarrow \{v_j \in V_f |\ e_{ij} \in E_f \ \ \& \ \ d_j = d_i + 1\}$;
  for v_j *in* V^{from} **do**
    for v_k *in* V^{to} **do**
      $push(e_{jk}, E_c)$

Algorithm 1: Building a coarse graph G_c from a given input graph G_f.

Resting pose geometry As we use resting pose ge-

ometries to compute the physical objective function (see Sec. 1.4), for some garments, we need to make a preprocessing step to generate these geometries.

Specifically, we find that canonical 3D geometries of the garments used in [8] are overly smooth and tight. At the same time [8] uses 2D triangle geometries in the UV space as canonical for computing the $\mathcal{L}_{stretching}$ term. However. we compute $\mathcal{L}_{stretching}$ using 3D resting pose triangles projected into 2D.

To better match the garment sizes used in [8] we generate the resting pose geometries with a preprocessing step. To do that, we run an LBFGS optimization using the exact same physical objectives used in [8] with the same 2D triangle geometries to get a relaxed version of the garments for the canonical SMPL pose.

For other garments, for example, those generated with a FoldSketch method [3] or the ones created manually in Blender, we don't run this preprocessing step and use their original 3D geometries as resting geometries.

Skinning weights Our model autoregressively predicts nodal accelerations for the garment mesh given its geometries from two previous time steps (this gives us nodal positions and velocities for the previous step). As we want to initialize the garment geometry from any point in the training sequences, during training we need a way to approximate the garment geometries for any given body pose.

To do that we follow [8] and, for each garment node, borrow skinning weights and blend shapes from the closest SMPL vertex in the canonical pose.

Although it works fine for tight-fitting garments, we find that when applied to loose garments (e.g. dresses), this results in overly stretched triangles. Hence, it is difficult for the model to start from a garment with severely sub-optimal potential energy. To overcome this issue, we employ diffused body model formulation from [9] to compute diffused skinning weights $\tilde{\mathcal{W}}$ and pose and shape blend-shapes $\tilde{\mathcal{B}}_{p,s}$.

$$\tilde{\mathcal{W}}(x) = \frac{1}{M} \sum_{q_n \sim N(x,d)} \mathcal{W}(\phi(q)) \tag{1}$$

$$\tilde{\mathcal{B}}_{p,s}(x) = \frac{1}{M} \sum_{q_n \sim N(x,d)} \mathcal{B}_{p,s}(\phi(q)), \tag{2}$$

where x is a position of a garment node in resting pose, d is the distance from this node to the closest SMPL vertex and $\phi(\cdot)$ is a function that for the given 3D point returns the closest SMPL vertex. However, instead of training an MLP, as in [9], for each garment node in the canonical pose, we directly sample M 3D points from the normal distribution $N(x_i, d)$. We find $M = 10000$ enough to get adequate initialization.

Pinned vertices For some of the garments, we need to specify a set of "pinned" vertices, i.e. vertices whose positions are rigidly connected to the body mesh (e.d. pants belt). We generate positions of these pinned vertices using our linear blend-skinning formulation and give them as input to the network while also giving these vertices a separate type label. Since there are only a few such vertices in the garment, using linear blend skinning (or any other rigid transformation w.r.t. the body mesh) does not affect the realism of the predicted dynamics.

1.3. Forward pass

Input feature vectors For each time step, we first endow each node and edge in the input graph with a designated feature vector that describes the state of this element in the previous time step along with the local material parameters. Here we list the contents of the input vectors for the nodes of the graph, garment edges and body edges. Also, see Fig. 1 for visualization.

The input vectors for the **nodes** of the graph consist of:

- velocity $\vec{v}$ of the not in the previous time step
- normal vector $\vec{n}$ of the node
- nodal mass m
- local material parameters: μ_{Lame}, λ_{Lame} and $k_{bending}$
- the weight of the inertial loss α (see Section 3.4 of the paper)
- one-hot encoding of the node type C (*"garment node"*, *"pinned garment node"* or *"body node"*)
- one-hot encoding of the deepest coarse level L the node is present in (separate code for body nodes)

for body nodes we set m, μ_{Lame}, λ_{Lame} and $k_{bending}$ to -1.

For each **garment edge** (including fine and coarse ones) connecting nodes i and j, each input vector is the concatenation of:

- a vector describing the relative position of the connected nodes in the current time step $\vec{\Delta x}_t^{(ij)} = (x_t^{(i)} - x_t^{(j)})$ and its' norm $\|\vec{\Delta x}_t^{(ij)}\|$
- a vector describing the relative position of the connected nodes in resting pose $\vec{\Delta x}_{rest}^{(ij)} = (x_{rest}^{(i)} - x_{rest}^{(j)})$ and its' norm $\|\vec{\Delta x}_{rest}^{(ij)}\|$
- local material parameters: μ_{Lame}, λ_{Lame} and $k_{bending}$ averaged across the connected nodes
- the weight of the inertial loss α (see Section 3.4 of the paper)

For each **body node** connecting garment node i and body node j, the input vector is the concatenation of:

- a vector describing the relative position of the garment node and the body node in the current time step $\vec{\Delta x}_t^{(ij)} = (x_t^{(i)} - x_t^{(j)})$ and its' norm $\|\vec{\Delta x}_t^{(ij)}\|$
- a vector describing the relative position of the garment node in the current time step and the body node in the next time step $\vec{\Delta x}_{t+1}^{(ij)} = (x_t^{(i)} - x_{t+1}^{(j)})$ and its' norm $\|\vec{\Delta x}_{t+1}^{(ij)}\|$
- the weight of the inertial loss α (see Section 3.4 of the paper)

Network architecture Our model consists of 3 parts: the encoder, N message-passing steps and the decoder. The encoder converts the input feature vectors into latent vectors with h dimensions. The message-passing steps update the latent vectors for each node and edge of the graph (see Fig. 2 for visualization). The decoder decodes the nodal latent vectors into scalar accelerations. We use the latent dimensionality $h = 128$ in our experiments.

Here, we describe the architectures for each of these parts.

The **encoder** comprises $M + 2$ multi-layer perceptrons (MLPs), where M is the total number of levels in the network. Each of the M MLPs encodes feature vectors of the garment edges on the specific level. The other two MLPs encode the nodal feature vectors and the feature vectors for the body edges.

Each of the **message-passing steps** consists of $L + 2$ MLPs. Here L is the number of levels processed by this specific step. The other two MLPs, again, process the nodal features and the body edge features.

The **decoder** is a single MLP that decodes the features of each garment node into the predicted accelerations.

Each MLP in the architecture has 2 hidden layers with ReLU activations and layer normalizations.

1.4. Physical supervision

We directly borrow the physical objective terms introduced in SNUG [8] to train our model, with the exceptions of the collision term $\mathcal{L}_{collision}$ and stretching term $\mathcal{L}_{stretching}$, which we slightly modify for our needs, and the novel term $\mathcal{L}_{friction}$.

Here we describe these terms in more detail.

Collision term We modify $\mathcal{L}_{collision}$ to compute the penetration of each garment node with respect to the closest *face* of the body mesh rather than to the closest *vertex*.

Apart from that, while we compute penetrations happening in the current time step, we use the body faces that were closest to the garment vertices in the previous time step. So the process for the computation of $\mathcal{L}_{collision}$ in time step t is as follows:

1. for each garment node v_i find the body face f_j, whose center f_j^{center} is closest to v_i in the time step t,
2. compute the size of the collision of vertex v_i with respect to the face f_j in the time step $t + 1$

The collision penalty is computed as a cubic energy term similarly to SNUG:

$$\mathcal{L}_{collision}(v_i) = max(\epsilon - d(v_i, f_j), 0)^3 \tag{3}$$
$$d(v_i, f_j) = (v_i - f_j^{center}) \cdot \vec{n_j} \tag{4}$$

where $\vec{n_j}$ is the normal vector of the face f_j.

In that way, the body-garment correspondences align with those used to build world edges. We find this little modification of the collision term crucial for the convergence of the model. Note, that for the metrics reported in Table 1 we use the regular collision term, with correspondences for the current time step.

Friction We introduce a novel physical objective $\mathcal{L}_{friction}$, which penalizes the sliding friction between the garment and the body. The friction energy is computed separately for every vertex. Following [1, 2], we penalize the movement of a node relative to the body for those vertices that fall into the r of the closest body face in both current and previous time frames.

Here we describe the process of computing the fiction term for a specific vertex v step-by-step.

First, we find indices u and w of two body faces, closest to v in the current and the previous time steps (see Figure 3).

$$u = \arg\min_p \|f_p^t - v^t\| \tag{5}$$
$$w = \arg\min_p \|f_p^{t+1} - v^{t+1}\| \tag{6}$$

Then we compute the node's projected position $\hat{v}^{t+1}$ in time step $t + 1$, assuming it moves the same way as f_u

$$\hat{v}^{t+1} = v^t + (f_u^{t+1} - f_u^t), \tag{7}$$

and find a vector from $\hat{v}^{t+1}$ to v^{t+1} and project it onto a plane with slope θ which is an average between slopes of f_m^t and f_k^{t+1}. Using the norm d of this projected vector, we compute the friction term with the following formula.

$$\mathcal{L}_{friction} = \mu_k \times g \times m \times \cos\theta \times d \tag{8}$$
$$\theta = (\theta_{f_u}^t + \theta_{f_w}^{t+1})/2 \tag{9}$$
$$d = \|proj(v^{t+1} - \hat{v}^{t+1}, \theta)\|_2, \tag{10}$$

where μ_k is friction ratio, g is gravitational acceleration, m is a mass of the node, and $proj(\cdot, \theta)$ is an operation of projecting a vector onto a plane with the slope θ_i

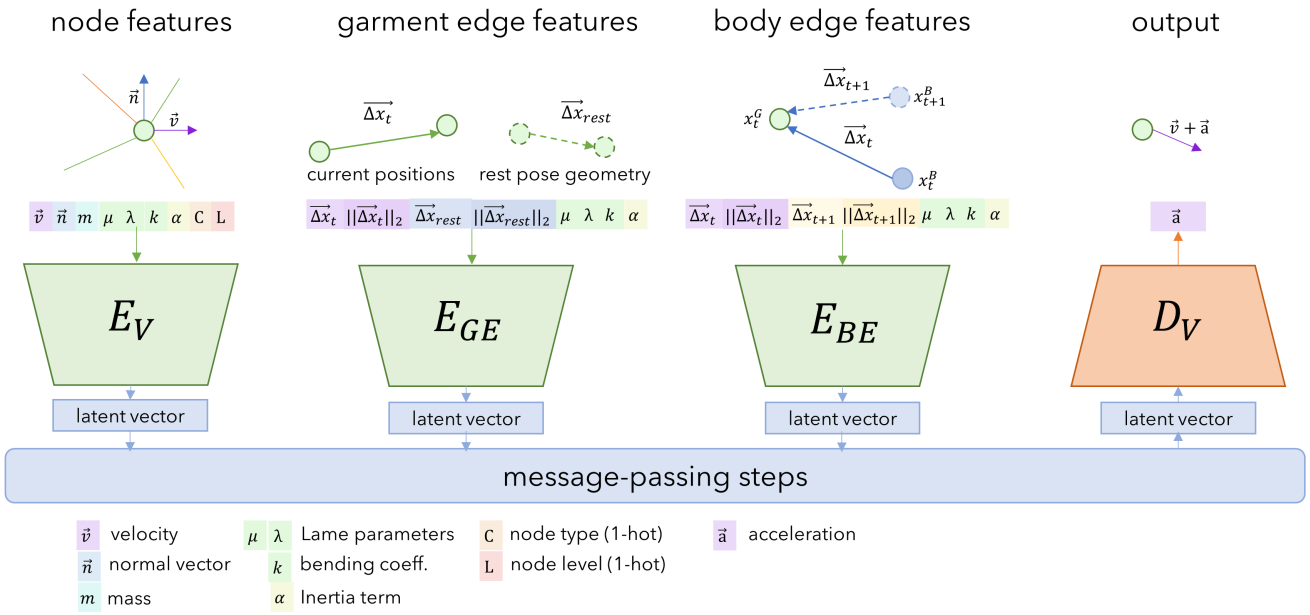

Figure 1. Visualisation of encoding and decoding steps of the HOOD model. E_V, E_{GE} and E_{BE} encode feature vectors of the graph nodes, garment edges and body edges respectively into the same latent space. After several message-passing steps, the latent vectors corresponding to garment nodes are decoded into acceleration vectors by D_V. To avoid clutter, we only show the encoder for the garment edges of the highest level. In reality, each level of edges is encoded with a separate encoder, while the set of input features for them is the same.

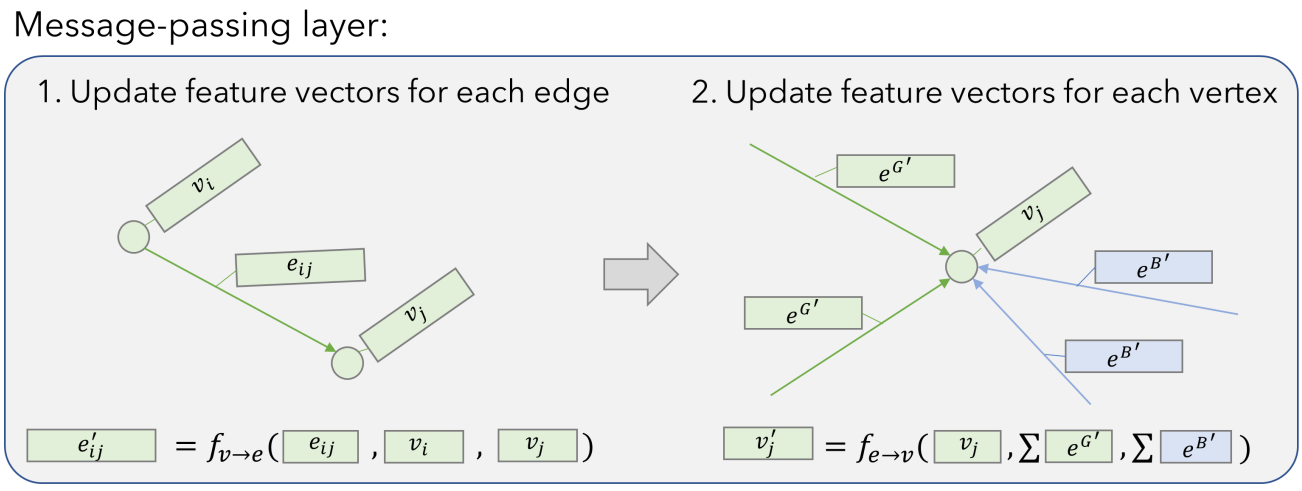

Figure 2. Illustration of one message-passing step. First, it updates the latent vectors of the edges of the graph with a dedicated MLP $f_{v\to e}$. Then, using the updated edge vectors, it updates the node latent vectors with another MLP $f_{e\to v}$. For simplicity, we only show edges of the highest level here, while in reality, edges of each level are processed by a separate processing MLP $f^l_{v\to e}$, and their aggregations are concatenated to the input vector of $f_{e\to v}$. Also note, that we use a bidirectional graph where there are always two edges between each pair of nodes. These two edges (a direct one and an inverse one) have independent latent vectors.

While it is a very rough approximation of the friction energy, we find that it allows for more realistic modelling of body-garment interactions. We demonstrate the effects of the friction term in the supplementary video on our project page (dolorousrtur.github.io/hood).

Vertex mass and canonical geometries Another slight difference to the physical objective formulation from SNUG is how we compute the mass of each garment vertex. As it

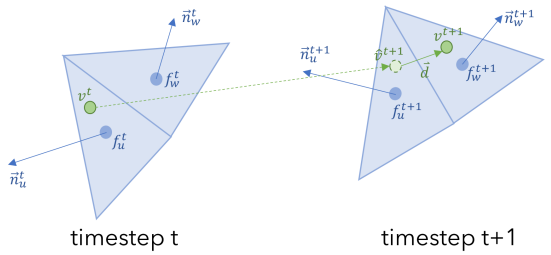

Figure 3. We model friction by penalizing the motion of garment nodes relative to the body.

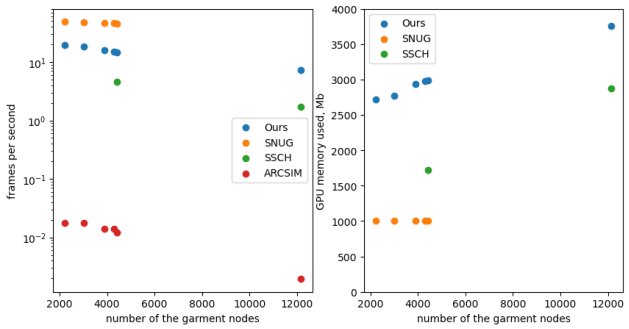

Figure 4. Inference speed and GPU memory footprint for our final model, two baseline architectures (SSCH [7] and SNUG [8]). Note the logarithmic scale for the speed plot as we also plot the performance of ARCSIM physical simulator [5]. Each point corresponds to one of the training garments. The speed was measured on NVIDIA GeForce RTX 3060 GPU.

SNUG, to compute the vertex mass, we need to know the material density and the areas of the adjacent triangle faces. The difference is that, while SNUG uses original 3D garment geometries to compute triangle areas, we use the relaxed ones, acquired using the preprocessing step described in Section 1.2.

Apart from that, SNUG uses original 2D triangles from UV space to compute $\mathcal{L}_{stretching}$, while we compute $\mathcal{L}_{stretching}$ using relaxed 3D resting pose triangles projected into 2D.

2. Experiment details

We trained our final model for 150000 training iterations which took around 26 hours on NVIDIA Quadro RTX 6000 GPU.

Below we provide some additional information concerning the experiments from the main paper.

2.1. Comparison to state-of-the-art

Since we do not learn from the physically simulated data as [9] does and use a slightly different objective function from SNUG [8] (See Sec. 1.4), it is difficult to quantitatively compare our method to them. However, we can compute the size and the number of body-garment penetrations by our method and the baselines. We provide these values in Table 1

	garments	$\mathcal{L}_{collision} \downarrow$	% penetrating vertices $\downarrow$
SNUG [8]	t-shirt, long sleeve top,	1.54e-5	6.83e-1
Ours	tank top, pants, shorts	**1.21e-7**	**1.15e-1**
SSCH [9]	t-shirt, dress	2.73e-5	6.54e-1
Ours		**1.27e-6**	**1.45e-2**

Table 1. Our method generates fewer garment-body penetrations compared to state-of-the-art methods in terms of both average penetration size and percentage of garment vertices penetrating the body. The numbers were averaged over the whole validation set.

2.2. Architecture analysis

In addition to Table 1 in the main paper, we provide detailed metrics averaged across the whole validation set for our final model *Ours* and two baseline architectures $Fine15$ and $Fine48$ in Table 2.

Different architectures with different sets of fine and coarse message-passing steps may result in models with different inference speeds, propagation radii and levels of fine details. We demonstrate the qualitative differences between different architectures in the supplementary video on our project page (dolorousrtur.github.io/hood).

2.3. Stability

Despite the autoregressive inference strategy, we empirically find that our model shows no signs of instabilities even for extremely long sequences. As a demonstration, Fig. 5 plots the total loss for a 100-second sequence composed of several cycles of a periodic animation clip.

References

[1] George E. Brown, Matthew Overby, Zahra Forootaninia, and Rahul Narain. Accurate dissipative forces in optimization integrators. *ACM Trans. Graph.*, 37(6), dec 2018. 3

[2] Moritz Geilinger, David Hahn, Jonas Zehnder, Moritz Bächer, Bernhard Thomaszewski, and Stelian Coros. Add: Analytically differentiable dynamics for multi-body systems with frictional contact. *ACM Trans. Graph.*, 39(6), nov 2020. 3

[3] Minchen Li, Alla Sheffer, Eitan Grinspun, and Nicholas Vining. Foldsketch: Enriching garments with physically reproducible folds. *ACM Transaction on Graphics*, 37(4), 2018. 2

[4] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6

[5] Rahul Narain, Armin Samii, and James F O'brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 5

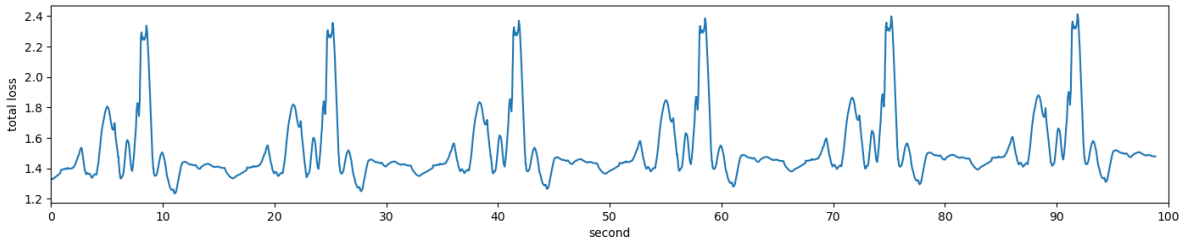

Figure 5. Total loss value plotted for a 100-second (3000 frames) sequence. We repeat the same sequence of a jumping person (*01_01* from CMU split of AMASS [4] dataset with a dress from vto dataset [7]) 6 times. The plot for the last part almost precisely matches the first part, demonstrating the stability of our method. Peaks correspond to rapid body motions.

	garments	average speed, fps ↑	$\mathcal{L}_{stretching}$ ↓	$\mathcal{L}_{bending}$ ↓	$\mathcal{L}_{inertia}$ ↓	$\mathcal{L}_{gravity}$ ↓	$\mathcal{L}_{collision}$ ↓	$\mathcal{L}_{friction}$ ↓
Fine15	only dress	6.65	1.68	1.56e-2	4.53e-3	1.22	2.54e-6	2.37e-3
Fine48		2.45	**1.66e-1**	**1.31e-2**	3.19e-3	**1.24**	1.85e-6	2.36e-3
Ours_full		**7.27**	2.08e-1	1.64e-2	**3.06e-3**	1.25	**1.77e-6**	**2.09e-3**
Fine15	all	13.1	7.71e-1	8.06e-3	2.79e-3	**8.95e-1**	7.25e-7	1.44e-3
Fine48		4.99	**1.25e-1**	**7.41e-3**	2.46e-3	8.99e-1	4.52e-7	1.32e-3
Ours_full		**13.6**	1.52e-1	8.46e-3	**2.37e-3**	9.04e-1	**3.28e-7**	**1.22e-3**

Table 2. Comparison of our final model to two ablations in terms of physical objectives and inference speed. We provide the metrics for the largest garment in the training set (*dress*, 12K vertices) separately.

[6] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020. 1

[7] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum*, 38(2):355–366, 2019. 5, 6

[8] Igor Santesteban, Miguel A Otaduy, and Dan Casas. SNUG: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8140–8150, 2022. 2, 3, 5

[9] Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11763–11773, June 2021. 1, 2, 5