Generating High-Fidelity Clothed Human Dynamics with Temporal Diffusion

SHIHAO ZOU, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
YUANLU XU, Meta Platforms, Inc., United States
NIKOLAOS SARAFIANOS, Meta Platforms, Inc., United States
FEDERICA BOGO, Meta Platforms, Inc., United States
TONY TUNG, Meta Platforms, Inc., United States
WEIXIN SI*, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
LI CHENG, University of Alberta, Canada

Clothed human modeling plays a crucial role in multimedia research, with applications spanning virtual reality, gaming, and fashion design. The goal is to learn clothed human dynamics from observations and then generate humans with high-fidelity clothing details for motion animation. Despite tremendous advancements in clothing shape analysis by existing approaches, the community still faces challenges in generating convincing visual effects of cloth dynamics, maintaining temporally smooth clothing details, and handling diverse clothing patterns. To address these challenges, we introduce ClothDiffuse, a temporal diffusion model that seamlessly integrates three key components into this task: temporal dynamics modeling, iterative refinement, and diversified generation. Our approach begins by using an encoder to extract high-level temporal features from input human body motions. These features are combined with a learnable pixel-aligned garment feature, serving as prior conditions for the shape decoder. The decoder then iteratively denoise Gaussian noise to produce clothing deformations over time on the input unclothed human bodies. To ensure that the results align with observations and adhere to physical plausibility for clothing shape inference, we propose two physics-inspired loss functions that preserve the intra-frame distances and inter-frame forces of clothing points. Additionally, the stochastic nature of the denoising process allows for the generation of diverse and plausible clothing shapes. Experiments show that our approach outperforms state-of-the-art methods in chamfer distance and visual effects, particularly for loose clothing such as dresses and skirts. Furthermore, our approach effectively adapts to out-of-domain clothing types and generate realistic clothes dynamics.

CCS Concepts: • **Computing methodologies** → **Shape representations**; **Shape inference**; **Shape analysis**.

Additional Key Words and Phrases: Diffusion Models, Clothed Human Modeling, Cloth Dynamics Generation

1 INTRODUCTION

Clothed human modeling is a vital research area in multimedia, with wide-ranging applications spanning gaming, virtual try-ons, digital fashion design, and more. The ability to produce high-fidelity clothed humans is essential for

*Corresponding Author.

Authors' Contact Information: Shihao Zou, sh.zou@siat.ac.cn, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Guangdong, China; Yuanlu Xu, Meta Platforms, Inc., Menlo Park, United States; Nikolaos Sarafianos, Meta Platforms, Inc., Menlo Park, United States; Federica Bogo, Meta Platforms, Inc., Menlo Park, United States; Tony Tung, Meta Platforms, Inc., Menlo Park, United States; Weixin Si, wx.si@siat.ac.cn, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Guangdong, China; Li Cheng, University of Alberta, Edmonton, Canada, lcheng5@ualberta.ca.

ACM 1551-6865/2025/1-ART
https://doi.org/10.1145/3712011

creating immersive digital experiences and enhancing user interaction. Specifically, this field focuses on learning clothed human dynamics from 3D point clouds or meshes of human bodies, enabling the animation of clothed humans with high-fidelity clothing details based on reference body motions. This task is inherently challenging due to the wide variety of clothes and human motions. Traditional methods often reply on rigging-and-skinning techniques [22] or physics-based simulations [32], which require intensive computations and domain expertise to create simulation-ready 3D assets for clothing shape inference.

In response to these issues, data-driven approaches have emerged as a promising alternative. These approaches utilize either implicit or explicit shape representations for clothed humans, effectively overcoming the limitations of traditional methods in shape analysis. Among them, point-based representations [21, 24, 25, 27, 49] have demonstrated their efficacy in modeling clothed humans, attributed to the efficiency, compactness and topological flexibility of point clouds. Surface Codec of Articulated Local Elements (SCALE) [24] is the first method applying the point-based surface representation on human clothes modeling and the following effort; while Power-of-Point (POP) [27] further demonstrates the ability of a single model for arbitrary clothes types. To address the issue of varying topology of clothes, First-Implicit-Then-Explicit framework (FITE) [21] is proposed to learn an implicit model to reconstruct a coarse template of clothes and then add explicit pose-dependent deformation. A similar idea is also employed in Skinned Refined Template-free approach (SkiRT) [25] that introduces a coarse-to-fine process. A following approach, Clothed humans on a continuous Surface with Explicit Template (CloSET) [49], learns pose features on a body surface to tackle the discontinuity of the UV map used in [24, 27]. Previous work, Dynamic Point Field (DPF) [35], models dynamic surfaces with a point-based representation of clothed humans.

Although existing approaches have made promising advancements, there are still unsolved challenges in this research field. *The first challenge* resides in the dynamics modeling, where the clothing deformations are supposed to be natural and smooth, both spatially and temporally, when people perform various actions. However, most existing learning-based approaches [8, 9, 21, 24, 25, 27, 39, 49] focus on the clothing deformations associated with static body poses only, ignoring the underlying correlation and continuity of clothing dynamics in a motion sequence. Although DPF [35] considers dynamics in the canonical surface reconstruction, it treats each frame separately without considering the explicit dynamics during inference, *i.e.*, during the reference motion animation. *The second challenge* lies in the lack of high-fidelity details, *i.e.*, clothes wrinkles. Earlier work adopts either a one-step end-to-end strategy [27, 35, 39, 49], or a two-step coarse-to-fine pipeline [21, 25] to solve this challenging task at different scales and levels of granularity. As such, they fail to fully exploit the benefits of iterative refinement in their model design. *The third challenge* is the lack of diversity in the results, which contradicts real-world observations. In reality, similar outfits in similar poses can result in different clothing patterns due to variations in motion context. For example, when an individual is walking forward or backward, they may present the same pose, but the clothing patterns can differ significantly. Existing methods [21, 25, 27, 35, 39, 49] are mostly deterministic, predicting clothing deformations based on a single pose, which limits their ability to account for the full range of variations that occur with specific outfits and motions.

To address these challenges, we propose *ClothDiffuse*, a temporal diffusion model that learns clothing dynamics to generate high-fidelity clothed humans in a given reference motion. *Our key insight is to develop a framework that integrates three critical factors: dynamics modeling, iterative refinement, and diversified generation.* Specifically, we use a sequence of unclothed human bodies, such as SMPL body shapes, as the input reference motion. The explicit dynamic features of vertices on these bodies, including 3D positions, velocities, and accelerations, are mapped to UV positional maps and then processed through a 3D CNN to encode high-level dynamic features of the motion sequence. In addition, we use a learnable tensor to encode pixel-aligned garment features. With these dynamic and garment features as prior conditions, a shape decoder iteratively denoises Gaussian noise to predict clothing wrinkle deformations and normal directions on the input unclothed human bodies over time. Finally, we obtain point-based clothed humans aligned with the input reference motion. The stochasticity in the sampling process allows our model to produce diverse yet plausible outcomes during inference. Furthermore, we

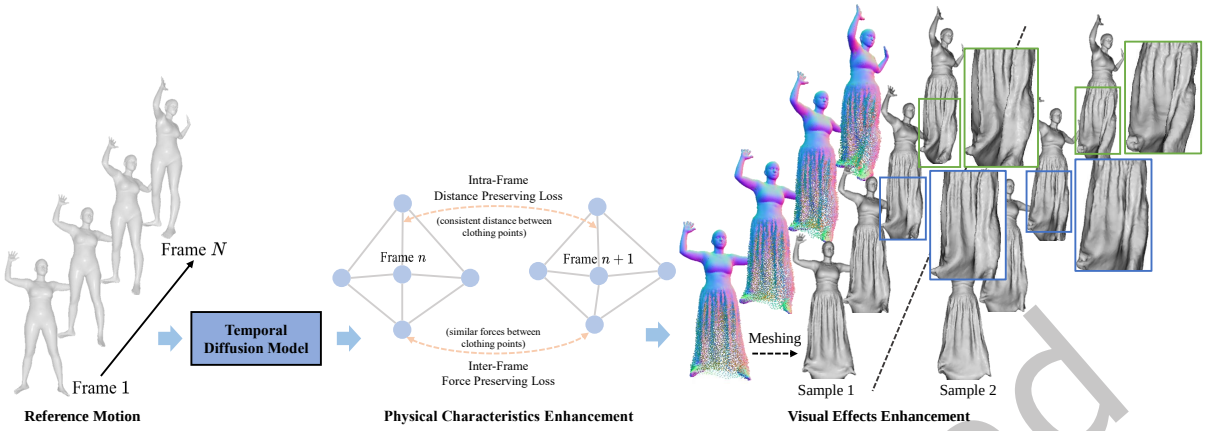

Fig. 1. **Scheme Overview of Our Method.** Given a reference body motion for animation, our temporal diffusion model generates point-based high-fidelity clothed humans with enhanced visual effects of clothing details over time through an iterative reverse denoising process. The stochasticity in this process also allows for diversified generations of clothed humans. During training, our model learns clothed human dynamics from 3D point clouds or meshes of clothed human bodies, with physical characteristics enhanced by dedicated loss functions.

propose two physics-inspired losses to enhance the physical characteristics of predicted results, which preserve the intra-frame distances and inter-frame forces of clothing points, as illustrated in Fig. 1. We conduct extensive experiments against state-of-the-art methods on three public benchmarks. Our approach outperforms existing methods and demonstrates a clear advantage in modeling loose clothes.

Our contributions are summarized as follows: (i) We propose ClothDiffuse, a temporal diffusion approach that learns clothed human dynamics and generates high-fidelity clothing details along with diverse patterns, showing a clear advantage in modeling loose clothes, *e.g.*, dresses and skirts. (ii) Our framework enhances the visual effects of clothed human dynamics through physics-inspired losses functions, *i.e.*, modeling cloth dynamics based on temporal motions and regularizing the intra-frame distances and inter-frame forces of clothing points. (iii) Comprehensive experiments demonstrate the superior performance of our framework and its applicability to real-world, unseen outfits and clothes.

2 RELATED WORKS

This section reviews relevant literature in three key areas: clothed human modeling, focusing on dynamic clothing simulation and animation; diffusion-based generative models, highlighting recent advancements in Artificial Intelligence Generative Content (AIGC); and physics-inspired losses, which contribute to enhancing the physical plausibility of generated content.

Clothed Human Modeling. This research field can be divided into two groups based on implicit and explicit human representations. Implicit representation defines surfaces as the zero level set of an implicit function, often powered by multi-layer perceptrons (MLPs) to predict occupancy values for any 3D position in continuous camera space. This approach, free from pre-defined templates, can model diverse, complex topologies and is widely used in human reconstruction [7, 12, 18, 30, 31, 38, 41, 47] and clothes modeling [6, 8, 28, 39, 44, 45, 50]. Most existing methods [6, 8, 12, 28, 39, 44] reconstruct clothed humans in a canonical space and animate them with predicted skinning weights and pose-aware deformations, while some others [18, 30] use part-based modeling. However, implicit representations [21, 27, 49] face cubic time computation costs when reconstructing explicit surfaces from

high-resolution volumes. In contrast, we use a point-based explicit representation, which is both compact and topologically flexible, avoiding the need for intensive computation. *(ii) Explicit representation* in clothes modeling typically uses a mesh-based template. Although robust and efficient, this approach is limited by its ability to generalize across diverse topologies and requires registration or canonization of raw scans. [4, 5, 17, 20, 43]. In contrast, point-based representations offer both compactness and support for arbitrary topologies. Early methods [13] generated sparse point sets for 3D object reconstruction, while later approaches [1, 11] structured point clouds into patches representing 2D UV maps, enabling dense surface geometry modeling. POP [27] demonstrated the ability to model pose-dependent clothing deformations using point clouds, handling various clothing types with a universal model. Subsequent works like FITE [21] and SkiRT [25] used a coarse-to-fine strategy, refining a coarse template with pose-dependent deformations. CloSET [49] addressed UV map discontinuities by directly learning pose-dependent features from the continuous body surface. Additionally, DPF [35] introduced a dynamic point field model combining explicit point-based representation with implicit deformation networks to model non-rigid 3D surfaces of clothed humans. Our proposed method differs from previous works by introducing a temporal diffusion framework with three key operations for this task: dynamics modeling, iterative refinement, and diversified generation. Temporal sequence offers richer context for robust dynamics modeling, iterative denoising refinement enhances detail representation, and the diffusion process enables diverse while natural outputs.

Diffusion-based Generative Model. Our work focuses on diffusion-based generative model, which is a paradigm based on the stochastic diffusion process and has been successfully applied in class and text-conditioned image generation [15], super-resolution, inpainting [16, 36, 37], and 3D object generation [34, 48]. Diffusion models work by progressively transforming a simple distribution, like Gaussian noise, into complex data distributions through a series of learned denoising steps. This iterative approach allows for the generation of highly realistic images, even in challenging tasks such as super-resolution and inpainting, where fine details must be reconstructed from partial or low-resolution inputs. In the context of generative clothed human modeling, diffusion models offer a powerful framework for capturing the intricate details of clothing and human body interactions. For instance, SMPLicit [10] proposes to learn a latent representation of body shapes and garments, enabling the generation of clothed humans in 3D space with high fidelity. Similarly, gDNA [6] leverages latent codes to generate detailed 3D canonical shapes of people in a variety of garments, along with corresponding skinning weights to ensure realistic deformations during motion. A concurrent work, DiffuStereo [42], introduces diffusion models into an iterative stereo matching network, demonstrating their capability in high-quality human reconstruction by refining depth estimates through a series of denoising steps. Unlike previous approaches that apply diffusion-based generative models to static 2D or 3D tasks, our proposed temporal diffusion model is specifically designed to capture the dynamics of clothing by denoising displacement changes over an entire temporal sequence, which can be considered as a 4D generation task. This framework offers a promising solution for complex temporal modeling tasks, enabling more realistic and consistent clothes modeling across time.

Physics-Inspired Losses. Physics-inspired losses have been pivotal in advancing clothed human modeling by improving the realism and efficiency of garment simulations. PBNS [2] introduces an unsupervised deep learning method to learn Pose Space Deformations (PSD) for rigged garments, integrating implicit Physically Based Simulations (PBS) to generate realistic cloth deformations without requiring extensive PBS data. Additionally, the approach [3] develops a general framework for neural cloth simulation that disentangles static and dynamic cloth subspaces, leveraging physics-inspired optimization schemes to enhance model performance and control motion predictions. However, these methods primarily focus on optimization schemes that minimize energy constraints to produce final outputs. In contrast, our approach incorporates differentiable physics constraints directly as loss functions, enabling the training of our end-to-end diffusion models. Additionally, SNUG [40] extends this by proposing a self-supervised approach that uses physics-based loss terms to train neural networks for dynamic 3D garment deformations, eliminating the need for ground-truth data and significantly speeding up

the training process. HOOD [14] employs graph neural networks and hierarchical message-passing to model clothing dynamics in real-time, effectively handling varying garment topologies and material properties. These approaches collectively highlight the role of physics-inspired losses in achieving realistic and efficient garment modeling across different scenarios. In contrast to previous approaches, we embed physics-inspired constraints directly into our temporal diffusion framework during training, enabling the model to incorporate these physical principles throughout the generation process. Furthermore, we have designed tailored formulations for our point-based clothed human representation. This setup enhances the model's capability to produce physically consistent, realistic motion for clothing dynamics generation.

3 PRELIMINARY

Temporal Diffusion Models. To be general, we represent $\boldsymbol{x}_0^{1:N}$ as an N-frame data sample drawn from the true data distribution. In the forward diffusion process, we add small amount of Gaussian noise to the sample over T steps, generating a sequence of noisy samples denoted as $\{\boldsymbol{x}_t^{1:N}\}_{t=1}^T$. This process is described by

$$q(\boldsymbol{x}_t^{1:N}|\boldsymbol{x}_{t-1}^{1:N}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}^{1:N}, \beta_t \boldsymbol{I}), \tag{1}$$

where the step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=0}^T$ and $\boldsymbol{I}$ represents the identity matrix. When the number of steps T is sufficiently large, $\boldsymbol{x}_T^{1:N}$ approaches a normal distribution, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. As noted in [15], for a specific diffusion step t, instead of repeatedly adding noises to $\boldsymbol{x}_0^{1:N}$, we can directly derive $\boldsymbol{x}_t^{1:N}$ through

$$\boldsymbol{x}_t^{1:N} = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{1:N} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{k=0}^t \alpha_k$.

The reverse process aims to reconstruct a true data sample from Gaussian noise $\boldsymbol{x}_T^{1:N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, utilizing a neural model $f_{\boldsymbol{\theta}}$ that progressively denoises $\boldsymbol{x}_T^{1:N}$ over T steps. This allows us to sample from the learned data distribution by gradually denoising a standard Gaussian noise. Formally, the process is defined as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^{1:N}|\boldsymbol{x}_t^{1:N}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y), \tilde{\beta}_t \boldsymbol{I}), \tag{3}$$

where t is the diffusion step, y includes prior conditions, such as text, category label and image, and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. According to previous work [15, 36], $\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y)$ can be estimated in three different ways by defining different outputs of the neural model $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y)$ and corresponding loss functions for training:

(1) The output is directly the mean values in Eq. (3), *i.e.*, $\boldsymbol{\mu}_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y)$. Then we have

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y), \tag{4}$$

$$\mathcal{L} = \mathbb{E}_{t\in[1,T], \boldsymbol{x}_{t-1}^{1:N} \sim q(\boldsymbol{x}_{t-1}^{1:N})}[\|\boldsymbol{x}_{t-1}^{1:N} - \mu_{\boldsymbol{\theta}}\|_2]. \tag{5}$$

(2) The output is the sampled Gaussian noise at diffusion step T, *i.e.*, $\hat{\boldsymbol{\epsilon}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y)$. Then we have

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t^{1:N} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}), \tag{6}$$

$$\mathcal{L} = \mathbb{E}_{t\in[1,T], \boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}[\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|_2]. \tag{7}$$

(3) The output is the denoised sample at diffusion step 0, *i.e.*, $\hat{\boldsymbol{x}}_0^{1:N} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{1:N}, t, y)$. Then we have

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t^{1:N} + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\hat{\boldsymbol{x}}_0^{1:N}, \tag{8}$$

$$\mathcal{L} = \mathbb{E}_{t\in[1,T], \boldsymbol{x}_0^{1:N} \sim q(\boldsymbol{x}_0^{1:N})}[\|\boldsymbol{x}_0^{1:N} - \hat{\boldsymbol{x}}_0^{1:N}\|_2]. \tag{9}$$

Using the estimated $\boldsymbol{\mu}_\theta$, we can draw the denoised sample $\hat{\boldsymbol{x}}_{t-1}^{1:N}$ at next diffusion step $t-1$ from the Gaussian distribution, $p_\theta(\boldsymbol{x}_{t-1}^{1:N}|\boldsymbol{x}_t^{1:N})$ defined in Eq. (3).

4 METHOD

An overview of our approach is illustrated in Fig. 2. We train a model to learn clothed human dynamics from a set of sequential 3D point clouds or meshes of human bodies, denoted as $\boldsymbol{V}^{1:N} = \{\boldsymbol{v}_i^{1:N}\}_{i=1}^{M_s}$, where N is the number of frames and M_s is the number of points in each frame. Similar to prior work [21, 25, 27, 49], we assume the corresponding fitted or registered unclothed bodies $\boldsymbol{P}^{1:N}$ are provided, usually represented by SMPL [23] or SMPL-X [33]. Guided by the learned clothing dynamics, our model takes N consecutive frames of these unclothed bodies in a reference motion as input and generate the corresponding clothed humans as output.

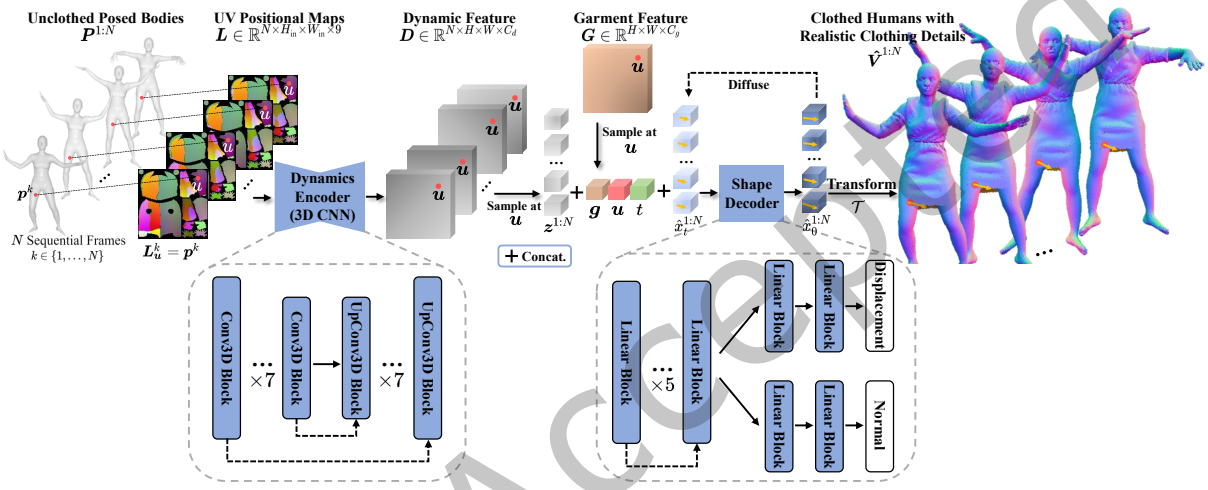

Fig. 2. **Pipeline of ClothDiffuse**: We take N sequential frames of unclothed posed bodies $\boldsymbol{P}^{1:N}$, such as SMPL bodies, as input. **(i) Dynamic Feature Encoding**: The explicit dynamic features of vertex $\boldsymbol{p}^k$ on the k-th frame's body surface are mapped to the corresponding 2D position $\boldsymbol{u}$ on a UV positional map $\boldsymbol{L}_{\boldsymbol{u}}^k = \boldsymbol{p}^k$, generating the input $\boldsymbol{L}$. A 3D CNN is then used to extract high-level dynamic features $\boldsymbol{D}$. **(ii) Diffusion-Based Cloth Generation**: We use a learnable tensor $\boldsymbol{G}$ to encode pixel-aligned garment features. $\boldsymbol{G}$ is specific to a particular outfit and is shared across all input motions for that outfit. Next we sample the respective dynamic features $\boldsymbol{z}^{1:N}$ and garment features $\boldsymbol{g}$ at position $\boldsymbol{u}$ on the UV maps via bilinear interpolation. Using these features along as prior conditions, a shape decoder progressively denoises Gaussian noise $\hat{\boldsymbol{x}}_T^{1:N}$ over T steps to yield the final prediction $\hat{\boldsymbol{x}}_0^{1:N}$, which includes cloth wrinkle displacements $\hat{\boldsymbol{r}}^{1:N}$ and normal directions $\hat{\boldsymbol{n}}^{1:N}$. Please refer to Sec. 4.3 for further details on training and inference using diffusion-based models. After applying the local transformations $\mathcal{T}$, we obtain the clothing points $\hat{\boldsymbol{v}}^{1:N}$ for the query points $\boldsymbol{p}^{1:N}$ over time. Finally, by densely querying the input unclothed bodies, we generate point-based clothed humans with realistic clothing details, represented as $\hat{\boldsymbol{V}}^{1:N} = \{\hat{\boldsymbol{v}}_i^{1:N}\}_{i=1}^{M_p}$.

4.1 Dynamic Feature Encoding

Dynamics Encoder is designed to encode explicit dynamic features of the input body motion sequence. Given the unclothed bodies $\boldsymbol{P}^{1:N}$ as input, we first generate the UV positional maps $\boldsymbol{L} \in \mathbb{R}^{N\times H_{\text{in}}\times W_{\text{in}}\times 9}$ by mapping the explicit dynamic features from the query points $\boldsymbol{p}^{1:N} \in \mathbb{R}^{N\times 9}$ on the bodies to the corresponding 2D position $\boldsymbol{u} \in \mathbb{R}^2$ on the UV maps. This mapping is represented by $\boldsymbol{L}_{\boldsymbol{u}}^k = \boldsymbol{p}^k$, where k is the frame index, ranging from 1 to

N. The explicit dynamic features include 3D position, velocity, and acceleration. We then feed the UV maps $\boldsymbol{L}$ to a 3D CNN to extract high-level dynamic features, denoted as $\boldsymbol{D} \in \mathbb{R}^{N\times H\times W\times C_d}$. These features encapsulate the spatiotemporal variations of the body motion at different levels of granularity.

Existing approaches [21, 25, 27, 49] mainly focus on pose-dependent modeling of clothed humans, often neglecting the correlation and continuity of dynamic features throughout a body motion sequence. To address this gap, we propose directly encoding explicit dynamic features. Combined with the physics-inspired losses described in Sec. 4.3, our approach enhances the modeling of clothed humans with motion dependency. Additionally, since the UV map of the template body has a fixed topology, the 3D CNN encoder can consistently extract vertex-aligned motion features, with each vertex $\boldsymbol{p}^k$ in every frame mapped to the same position $\boldsymbol{u}$ on the UV maps across time.

4.2 Diffusion-based Cloth Generation

Shape Decoder is designed to generate point-based clothed humans aligned with the input reference motion. We use a learnable tensor $\boldsymbol{G} \in \mathbb{R}^{H\times W\times C_g}$ to encode pixel-aligned garment features. $\boldsymbol{G}$ is specific to a particular outfit and is shared across all input motions for that outfit. During training, it is randomly initialized and optimized in an auto-encoding manner, similar to [27, 49]. Afterwards, given a query point $\boldsymbol{p}^{1:N}$ on the input unclothed bodies, we have its corresponding 2D position $\boldsymbol{u}$ on the UV map. Then we sample the corresponding motion-dependent dynamic features $\boldsymbol{z}^{1:N} \in \mathbb{R}^{N\times C_d}$ from $\boldsymbol{D}$ and the garment feature $\boldsymbol{g} \in \mathbb{R}^{C_m}$ from $\boldsymbol{G}$ at $\boldsymbol{u}$ on each frame. These features, along with the sampling position $\boldsymbol{u}$, serve as prior conditions for the diffusion reverse process, denoted as $\boldsymbol{y} = [\boldsymbol{z}^{1:N}, \boldsymbol{g}, \boldsymbol{u}]$. Note that $\boldsymbol{u}$ is mapped to an embedding vector via a linear layer, following the approach in [15]. Given these conditions, the shape decoder f_θ progressively denoises random Gaussian noise $\hat{\boldsymbol{x}}_T^{1:N}$ to generate the cloth wrinkle displacements $\boldsymbol{r}^{1:N} \in \mathbb{R}^{N\times 3}$ and corresponding normal directions $\boldsymbol{n}^{1:N} \in \mathbb{R}^{N\times 3}$ for the query points $\boldsymbol{p}^{1:N}$ across N frames, summarized as $\hat{\boldsymbol{x}}_0^{1:N} = [\hat{\boldsymbol{r}}^{1:N}, \hat{\boldsymbol{n}}^{1:N}]$.

Diffusion denoising process is described as follows. We represent $\boldsymbol{x}_0^{1:N}$ as the N-frame cloth wrinkle displacements $\hat{\boldsymbol{r}}^{1:N}$ and normals $\hat{\boldsymbol{n}}^{1:N}$ on the query points $\boldsymbol{p}^{1:N}$, drawn from the unknown true data distribution. Following the process in Sec. 3, the forward diffusion process incrementally add a small amount of Gaussian noise to the sample over T steps, resulting in a sequence of noisy samples denoted as $\{\boldsymbol{x}_t^{1:N}\}_{t=1}^T$.

The reverse process aims to generate cloth wrinkle displacements and normals $\hat{\boldsymbol{x}}_0^{1:N}$ on the query points $\boldsymbol{p}^{1:N}$ from Gaussian noise $\boldsymbol{x}_T^{1:N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. This is achieved through a neural model f_θ that progressively denoises $\boldsymbol{x}_T^{1:N}$ for T steps. As analyzed in Sec. 3, we follow [36], which directly predicts the denoised sample $\hat{\boldsymbol{x}}_0^{1:N} = f_\theta(\boldsymbol{x}_t^{1:N}, t; \boldsymbol{y})$ at each step, as this technique has shown to yield superior performance. Hence $\boldsymbol{\mu}_\theta$ can be obtained by Eq. (8). Finally, we sample $\boldsymbol{x}_{t-1}^{1:N}$ at denoising step $t-1$ from the Gaussian distribution $p_\theta(\boldsymbol{x}_{t-1}^{1:N}|\boldsymbol{x}_t^{1:N})$ defined in Eq. (3).

Local Transformations are applied to $\hat{\boldsymbol{r}}^{1:N}$ and $\hat{\boldsymbol{n}}^{1:N}$ on the query point $\boldsymbol{p}^{1:N}$. Following the approach in [27, 49], we model the predicted clothing deformation $\hat{\boldsymbol{r}}^{1:N}$ on the unclothed body in its canonical pose. This allows body articulation to account for a significant portion of shape variation, enabling the network to focus on modeling the remaining shape residuals. For each query point on the unclothed SMPL or SMPL-X body, we use Linear Blend Skinning (LBS) to calculate its local transformation via barycentric interpolation. Specifically, if the query point $\boldsymbol{p}^k$ on the k-th frame lies on a triangular face with three vertices $[\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3]$ and corresponding local transformations $[\boldsymbol{\mathcal{T}}_1, \boldsymbol{\mathcal{T}}_2, \boldsymbol{\mathcal{T}}_3]$ predefined by SMPL or SMPL-X, the local transformation of $\boldsymbol{p}^k$ is then obtained by barycentric interpolation, defined as

$$\boldsymbol{\mathcal{T}}^k = \sum_{j=1}^{3} \|\boldsymbol{p}^k - \boldsymbol{p}_j\|_2 \cdot \boldsymbol{\mathcal{T}}_j. \tag{10}$$

The final positions of the clothing points $\boldsymbol{v}^{1:N}$ over time can be obtained by

$$\hat{\boldsymbol{v}}^k = \boldsymbol{\mathcal{T}}^k \cdot \hat{\boldsymbol{r}}^k + \boldsymbol{p}^k, \tag{11}$$

where k is the frame index.

Finally, by performing dense querying on the input unclothed bodies, we obtain point-based clothed humans with realistic clothing deformations, represented as $\hat{\boldsymbol{V}}^{1:N} = \{\hat{\boldsymbol{v}}_i^{1:N}\}_{i=1}^{M_p}$, where M_p is the number of query points.

4.3 Training and Inference

Training. We follow the prior work [15, 36] to train our model. The first step is to uniformly sample a diffusion step t and derive the corresponding noisy sample $\boldsymbol{x}_t^{1:N}$ for each of M_p query points, along with the prior conditions $\boldsymbol{y}$. Next, we input them into the shape decoder, apply local transformations, and obtain the final outputs, denoted as $(\hat{\boldsymbol{r}}^{1:N}, \hat{\boldsymbol{n}}^{1:N}, \hat{\boldsymbol{v}}^{1:N})$. The total training loss $\mathcal{L}$ is defined as

$$\mathcal{L} = \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{n}}\mathcal{L}_{\mathrm{n}} + \lambda_{\mathrm{r}}\mathcal{L}_{\mathrm{r}} + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{g}} + \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{d}} + \lambda_{\mathrm{iso}}\mathcal{L}_{\mathrm{iso}}, \tag{12}$$

where λs are the weights to balance each loss.

Specifically, $\mathcal{L}_{\mathrm{c}}$ is the normalized Chamfer Distance that measures the average bi-directional squared distances between the predicted point clouds $\hat{\boldsymbol{V}}^{1:N} = \{\hat{\boldsymbol{v}}_i^{1:N}\}_{i=1}^{M_p}$ and ground-truth $\boldsymbol{V}^{1:N} = \{\boldsymbol{v}_i^{1:N}\}_{i=1}^{M_s}$, which is defined as:

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{L}_{\mathrm{c}}^k, \tag{13}$$

$$\mathcal{L}_{\mathrm{c}}^k = \frac{1}{M_s}\sum_{i=1}^{M_s}\min_j \|\boldsymbol{v}_i^k - \hat{\boldsymbol{v}}_j^k\|_2^2 + \frac{1}{M_p}\sum_{j=1}^{M_p}\min_i \|\boldsymbol{v}_i^k - \hat{\boldsymbol{v}}_j^k\|_2^2.$$

$\mathcal{L}_{\mathrm{n}}$ is the average L_1 distance between the normals of each predicted point and its nearest neighbor in the ground-truth point clouds:

$$\mathcal{L}_{\mathrm{n}} = \frac{1}{NM_p}\sum_{k=1}^{N}\sum_{i=1}^{M_p}\|\hat{\boldsymbol{n}}_i^k - \boldsymbol{n}_j^k\|_1, \tag{14}$$

where $j = \arg\min_{\boldsymbol{v}_j^k \in \boldsymbol{V}^k}\|\hat{\boldsymbol{v}}_i^k - \boldsymbol{v}_j^k\|_2$. $\mathcal{L}_{\mathrm{r}}$ and $\mathcal{L}_{\mathrm{g}}$ are the regularization of the norm for the predicted displacements and garment feature respectively, which is described by

$$\mathcal{L}_{\mathrm{r}} = \frac{1}{NM_p}\sum_{k=1}^{N}\sum_{i=1}^{M_p}\|\hat{\boldsymbol{r}}_i^k\|_2^2, \quad \mathcal{L}_{\mathrm{g}} = \frac{1}{HWC_g}\|G\|_2^2. \tag{15}$$

Additionally, inspired by physics-based cloth simulation, we propose two losses, $\mathcal{L}_{\mathrm{d}}$ and $\mathcal{L}_{\mathrm{iso}}$, to regularize the physical plausibility of the outputs, which have not been used in prior work on this task. $\mathcal{L}_{\mathrm{d}}$ is the inter-frame force preserving loss, which assumes points on the clothing in consecutive frames should undergo similar forces, defined as

$$\mathcal{L}_{\mathrm{d}} = \frac{1}{NM_p}\sum_{i=1}^{M_p}\sum_{k=2}^{N-1}\|(\hat{\boldsymbol{v}}_i^k - \hat{\boldsymbol{v}}_i^{k-1}) - (\hat{\boldsymbol{v}}_i^{k+1} - \hat{\boldsymbol{v}}_i^k)\|_2^2, \tag{16}$$

where the differences in acceleration of predicted points $\hat{\boldsymbol{V}}^{1:N} = \{\hat{\boldsymbol{v}}_i^{1:N}\}_{i=1}^{M_p}$ across consecutive frames are regularized. $\mathcal{L}_{\mathrm{iso}}$ is the intra-frame distance preserving loss, originating from [19], which enforces consistent distances between neighboring points across two consecutive frames, defined as

$$\mathcal{L}_{\mathrm{iso}} = \frac{1}{NM_p}\sum_{i=1}^{M_p}\sum_{k=1}^{N-1}\sum_{j=1}^{5}\|(\hat{\boldsymbol{v}}_i^k - \hat{\boldsymbol{v}}_j^k) - (\hat{\boldsymbol{v}}_i^{k+1} - \hat{\boldsymbol{v}}_j^{k+1})\|_2^2, \tag{17}$$

where j is the index of top-5 nearest neighboring points to point $\boldsymbol{v}_i$.

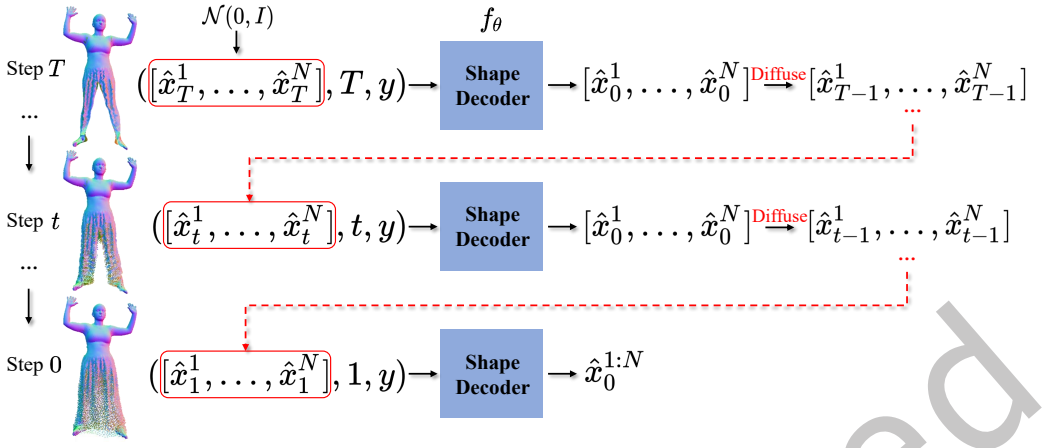

Fig. 3. **Inference process.** Sample from the learned distribution by gradually denoising Gaussian noise $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ for T steps using the shape decoder f_θ and finally yield the denoised sample $\hat{\boldsymbol{x}}_0^{1:N}$.

Inference. We sample from the learned distribution by iteratively denoising Gaussian noise $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. As illustrated in Fig. 3, at denoising step t, given the noisy sample $\boldsymbol{x}_t^{1:N}$ from the previous step as input, the shape decoder f_θ predicts the denoised sample $\hat{\boldsymbol{x}}_0^{1:N} = f_\theta(\boldsymbol{x}_t^{1:N}, t, \boldsymbol{y})$. This denoised sample is then diffused to $\hat{\boldsymbol{x}}_{t-1}^{1:N}$ using Eq. (8), which serves as the input for f_θ at the next denoising step $t-1$. After T steps of iterative denoising process, we obtain the final output $\hat{\boldsymbol{x}}_0^{1:N}$.

Our approach offers two main benefits over prior work [21, 25, 27, 35, 49]: *(i) Iterative refinement*: The sampling process involves T steps to achieve the final prediction, imitating the process of iterative refinement usually seen in artifact creation. As shown in Fig. 4 and 5, our method yields results with more refined clothing details compared to prior studies that model clothed humans in a single step only. *(ii) Diversified cloth generation*: Our approach generates diverse clothed humans, whereas previous techniques are deterministic during inference. The inherent stochasticity of our sampling process means that each inference produces a varied yet realistic outcome. This aligns well with the real-world observation that similar outfits in similar poses can present varied clothes patterns due to different motion context.

4.4 Generalization to Unseen Outfits

Since the garment feature $\boldsymbol{G}$ is unique to a particular outfit and individual, we initialize a set of garment features for all possible outfits in the training dataset, denoted as $\boldsymbol{G}_{\text{all}}$. During training, for a sample in a mini-batch, the corresponding garment feature is selected from $\boldsymbol{G}_{\text{all}}$ in the feedforward process and then updated accordingly. Each training sample is assigned an index to indicate its associated garment feature.

For a new outfit or individual with limited samples—specifically, a sequence of registered clothed human bodies represented as point clouds—our approach can still handle this unseen outfit following a scheme similar to [27]. We initialize a new garment feature $\boldsymbol{G}_{\text{new}}$ for this outfit, and optimize it by minimizing the loss defined in Eq. (12), keeping the dynamics encoder and shape decoder parameters fixed. When the sequence length is shorter than N, loss masks can be applied. The optimized garment feature then captures the characteristics of the unseen outfit, enabling our model to generate clothed humans in this outfit based on new input reference motion. As shown in Fig. 6, our approach effectively adapts to out-of-domain garment types and generate realistic dynamics.

5 EXPERIMENTS

Datasets. We evaluate our method and compare with baselines on three commonly used datasets. **ReSynth** [27] is a synthetic dataset with a larger variation in outfit shapes and motions in 13 different outfits. We follow [27] for training and test sets split. **CAPE** [26] is a dataset containing clothed human point clouds captured under a variety of motions in 14 different outfits. We follow [26] to split training and test sets, where test results of 3 subjects (00096, 00215, 03375) are reported due to their most extensive outfit variations. **THuman-CloSET** [49] provides more than 2,000 real-world scans of 15 outfits with a large variation in clothing style and poses imitating those in CAPE with fitted SMPL-X models. Following [49], we use the first 100 poses for training and the rest for testing. As the dataset comprises only single-frame point clouds rather than sequential ones, our model is constrained to utilizing single-frame inputs with $N = 1$. This means that our model on this dataset comes with diffusion-based iterative refinement, lacking dynamic modeling capability. To avoid data leakage and overfitting, we split the dataset such that, for the same outfit, some motion groups are included in the training set, while other distinct motion groups are reserved for the test set. Additionally, all presented results and evaluation metrics are derived from the test set.

Implementation details. are described as follows. The length of input sequence N is 8 and the input positional maps are of 128×128 resolution. The pose encoder in our model is a 7-layer 3D UNet while the shape decoder comprises 8-layer MLPs. The input dimension of UV positional maps is $N \times 128 \times 128 \times 9$, where N is the number of temporal frames. The dimension of dynamic feature, as the output of dynamics encoder, is $N \times 256 \times 256 \times 64$. Similarly, the garment feature dimension is also $256 \times 256 \times 64$. This means the dynamic and garment feature sizes, C_d and C_g, are both 64. The size of 256×256 gives around 43K query points as the final predicted point clouds of clothed human bodies. We follow [35] to use Lepard keypoint matcher to get the matched points in the ground truth point clouds for all the query points on the unclothed SMPL or SMPL-X body. Thus we can calculate $\boldsymbol{x}_0^{1:N}$ for the training with diffusion models. Note that when only pose-dependent features are considered (*i.e.* $N = 1$), our model has comparable parameters with prior work [27, 49]. We follow [29] to encode displacements $\boldsymbol{r}_i^{1:N}$, normals $\boldsymbol{n}_i^{1:N}$ and sampling position $\boldsymbol{u}$ with frequency being 6, as well as diffusion step t with frequency being 16. For all datasets, 100 diffusion steps and linear variance schedule of β are used. Our model is trained with uniformly sampled diffusion step for 90 epochs for all datasets on a single A100 GPU. The learning rate is 0.0001 and decays by 0.1 after training for 60 epochs. The batch size for training is 4 for $N = 8$ and 8 for $N = 1$. The weights of the losses are $1e4, 10, 30, 1, 3, 1$ for $\lambda_{\mathrm{c}}, \lambda_{\mathrm{n}}, \lambda_{\mathrm{r}}, \lambda_{\mathrm{g}}, \lambda_{\mathrm{d}}, \lambda_{\mathrm{iso}}$ respectively.

Metrics. Following prior work [25, 27, 49], we report the averaged Chamfer Distance (**CD**) from $\mathcal{L}_{\mathrm{c}}$ and averaged L_1 normal distance (**NML**) from $\mathcal{L}_{\mathrm{n}}$ across all test samples, in units of $\times 10^{-4} m^2$ and $\times 10^{-1}$, respectively. To capture model performance across outfits, especially challenging ones, we list results for each subject and outfit. In our ablation study, we report Dynamic Errors (**DE**) from Eq. (16), measuring average acceleration differences between points across frames, in units of $\times 10^{-4}$. We also evaluate the diversity of generated outcomes by the standard deviation of point clouds (**STD**) in $\times 10^{-2} m$, sampling our model 10 times.

5.1 Comparison with State-of-the-Art Methods

We compare our method, ClothDiffuse, with two implicit methods, SCANimate [39] and SNARF [8], and five closely top-performing point-based methods: SCALE [26], POP [27], SkiRT [25], CloSET [49] and DPF [35].

ReSynth dataset. We present quantitative results in Tab. 1, categorized by different outfits. The results reveal that all five recent point-based methods outperform the implicit method, SCANimate [39], by a significant margin. This underscores the advantages of using point cloud representations for clothed human modeling. In comparison with POP [27], our technique demonstrates consistent improvements in both CD and NML across most of the evaluated outfits. These enhancements are likely due to our use of dynamics feature encoding and iterative refinement through a diffusion process. Moreover, our method slightly improves upon the CD and

Methods	Outfits																	
	anna-001		beatrice-025		christine-027		janett-025		felice-004		carla-004		alexandra-006		eric-035		all	
	knee dress short sleeve		knee dress long sleeve		knee dress short sleeve		short skirt long sleeve		long dress tank top		puffy jacket long pants		loose blouse long pants		blazer jacket long pants		-	
	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓
SCALE [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.49	1.04
SCANimate [39]	1.34	1.35	0.74	1.33	3.21	1.66	2.81	1.59	20.79	2.94	0.90	1.52	2.28	1.84	2.54	1.94	4.30	1.77
POP [27]	0.62	0.82	0.34	**0.75**	1.72	**0.97**	1.24	0.89	7.34	1.24	0.51	**1.02**	1.71	**1.29**	1.34	1.16	1.36	1.02
SkiRT [25]	0.58	**0.81**	**0.31**	0.77	1.54	0.99	1.10	0.82	6.45	1.25	0.48	1.06	1.51	**1.29**	1.30	1.17	-	-
CloSET [49]	-	-	-	-	1.49	**0.97**	-	-	6.01	1.16	0.49	1.04	-	-	-	-	-	-
DPF [35]	0.96	0.96	0.46	0.99	2.88	1.24	2.51	1.19	16.07	2.50	-	-	-	-	-	-	-	-
Ours	**0.54**	**0.81**	0.33	0.76	**1.30**	**0.97**	**1.03**	**0.81**	**5.58**	**1.14**	**0.47**	1.05	**1.50**	**1.29**	**1.29**	**1.15**	**1.13**	**1.01**

Table 1. **Quantitative Comparison on ReSynth Dataset Across Diverse Outfits.** Our approach surpasses six prior methods in the majority of outfits, particularly excelling in the challenging long dresses ("felice-004"). The best results are in **bold**.

Fig. 4. **Qualitative Comparison on ReSynth Dataset.** Our approach delivers smooth and realistic details for loose clothes.

NML scores reported by SkiRT [25] for most outfits. Notably, we achieve a 13.4% improvement in CD for the challenging long dress outfit ("felice-004"), with scores of 5.58 compared to SkiRT's 6.45. Although SkiRT utilizes a coarse-to-fine strategy, our results affirm the value of multi-step progressive refinement for challenging outfits.

Methods	Outfits					
	blazerlong		shortlong		all	
	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓
SCALE [26]	1.07	1.22	0.89	1.12	-	-
POP [27]	0.78	1.29	0.57	1.24	0.59	1.11
CloSET [49]	0.71	1.15	0.54	1.09	-	-
Ours	**0.68**	**1.12**	**0.49**	**1.08**	**0.54**	**1.09**

Table 2. **Quantitative Comparison on CAPE Dataset.** Our approach obtains state-of-the-art performance for the "blazerlong" and "shortlong" outfits, as well as the average performance across all outfits. The best results are in **bold**.

Methods	Outfits					
	sweater-000		longshirt-001		skirt-005	
	CD↓	NML↓	CD↓	NML↓	CD↓	NML↓
SCANimate [39]	1.06	1.64	1.42	1.85	1.93	1.74
SNARF [8]	7.11	2.09	6.66	2.21	9.39	2.31
POP [27]	0.76	1.55	1.54	1.83	1.66	1.43
CloSET [49]	0.68	**1.48**	1.39	1.71	1.49	1.36
Ours	**0.66**	**1.48**	**1.33**	**1.70**	**1.42**	**1.35**

Table 3. **Quantitative Comparison on THuman-CloSET Dataset.** Our results are based on the model that uses single-frame input, as the dataset contains only single-frame point clouds. The best results are in **bold**.

Similar observations can be made when contrasting our method with CloSET [49]. While CloSET encodes features on the continuous body surface instead of UV map, it overlooks the importance of dynamics learning and iterative refinement in clothed human modeling. The qualitative results depicted in Fig. 4 further demonstrate that our model can generate clothed humans with smooth and natural clothes wrinkles over time. These results are notably superior to those from the four pose-dependent baselines [21, 27, 39, 49], highlighting the importance of dynamics modeling for this task. Moreover, our method produces higher point cloud density in regions where the clothing is distant from the human body, an improvement attributable to the iterative refinement facilitated by our diffusion-based framework.

CAPE dataset. The quantitative results are summarized in Tab. 2, which includes the results for the "blazerlong" and "shortlong" outfits, as well as the average performance across all outfits. Additionally, qualitative results are illustrated in Fig. 5. When compared to SCALE [26], our method demonstrates substantial improvements for both listed outfits. The inferiority of SCALE is primarily due to its reliance on low-resolution point clouds. Compared to POP [27], our technique yields improvements of 8.5% and 1.8% in CD and NML, respectively, when averaged across all outfits. These gains attest to the efficacy of our dynamic modeling and iterative refinement techniques for clothed human generation. These factors help our method obtain superior performance over CloSET [49].

THuman-CloSET dataset. Similar trend can be observed from the quantitative results presented in Tab. 3. While the dataset only provides single-frame point clouds, hence not involving dynamic modeling, our model still benefits from the iterative refinement via the diffusion generation process when compared with top-performing methods like POP [27] and CloSET [49].

5.2 Apply to Unseen Outfits

We present the qualitative comparison on unseen real-world outfits in Fig. 6, where the results of our approach, POP [27] and CloSET [49] are provided. We optimized garment features for each scan and generated clothed humans in reference motion animation, achieving natural and smooth wrinkles even for unseen loose clothing. It can be observed that POP shows quite noisy results and fails to model unseen loose dress, while our method gives smoother and more detailed clothing dynamics. This might attribute to the iterative refinement via denoising process, as well as the utilization of physics-based cloth simulation losses in training.

In ReSynth dataset, we showcase two illustrative outcomes in Fig. 7. These results demonstrate our framework's proficiency in adapting to a wide range of outfits or clothes with minimal effort, focusing instead on refining a

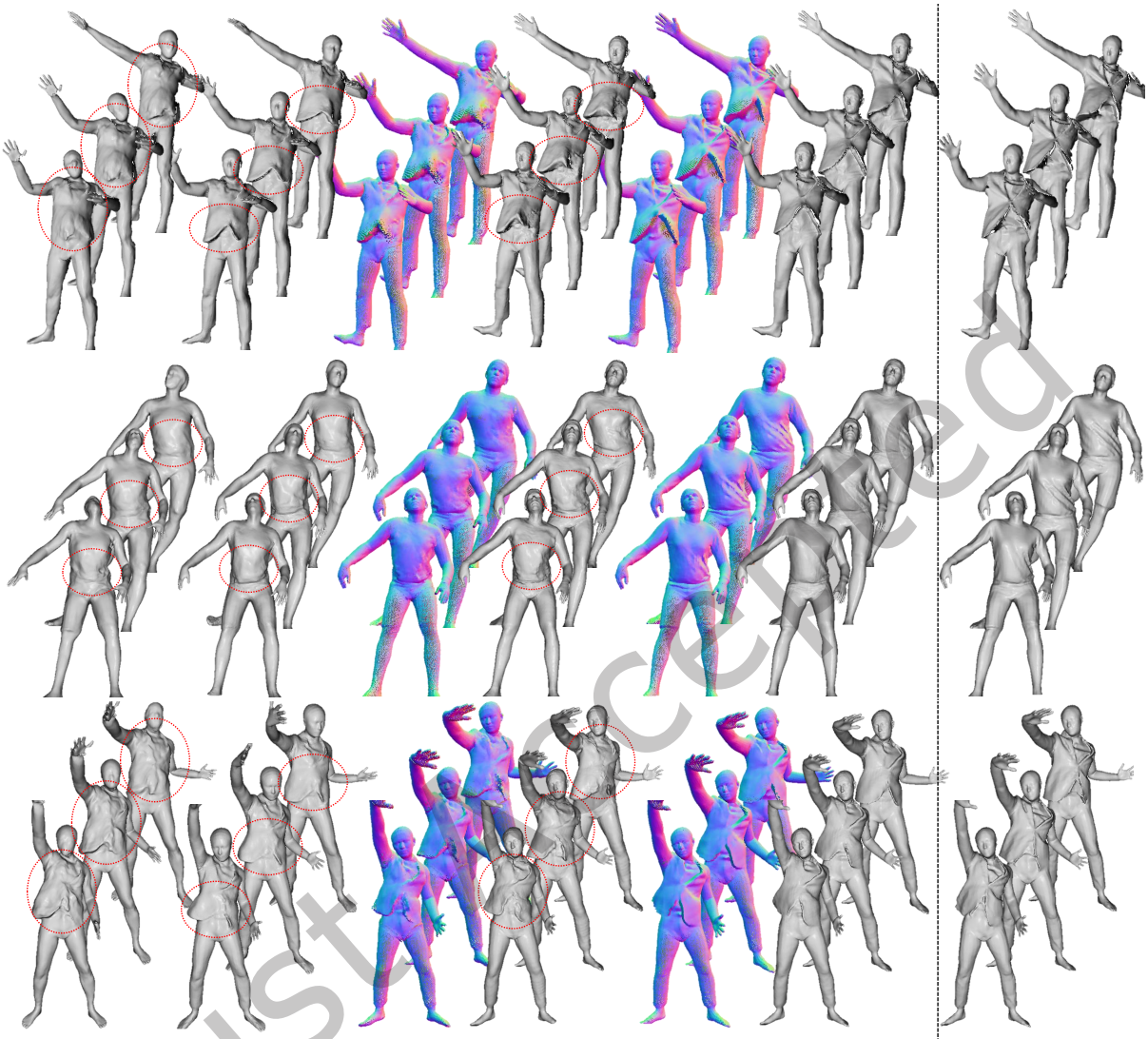

Fig. 5. **Qualitative Comparison on CAPE Dataset.**

garment feature uniquely associated with the presented outfit. Detailed results are included in the supplementary video.

5.3 Ablation Study

We conduct an ablation study on the ReSynth dataset to assess the contributions of two core components of our approach: dynamics feature encoding and diffusion-based iterative refinement. Quantitative results are presented in Tab. 4, with additional qualitative results provided in Fig. 8. Our study includes five variations:

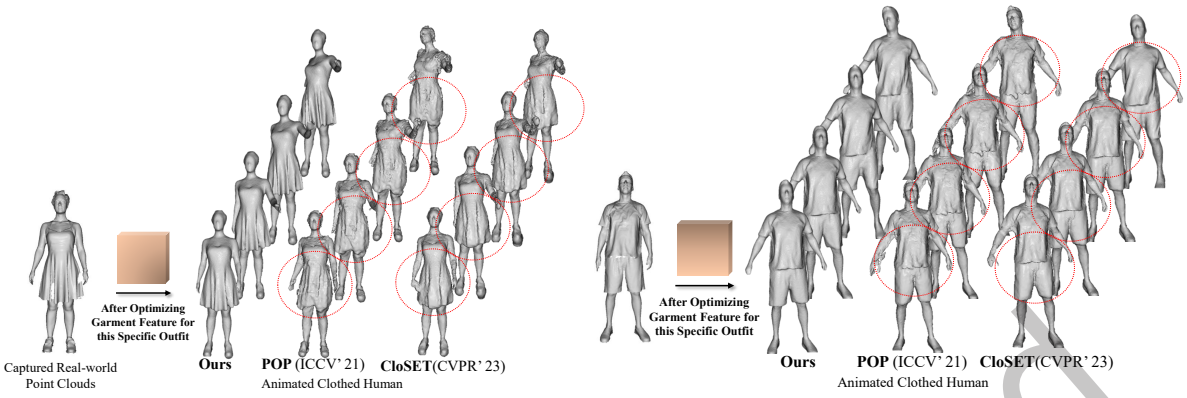

Fig. 6. **Unseen Real-World Outfits.** Qualitative comparison among Ours, POP [27] and CloSET [49].

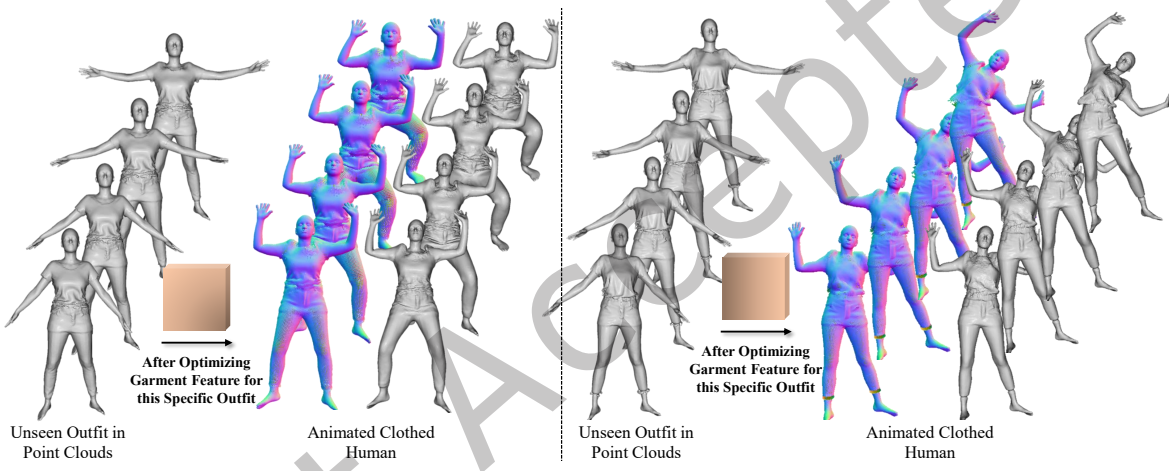

Fig. 7. **Unseen Outfits in ReSynth Dataset.**

- *Ours* (T=25): This variation utilizes fewer diffusion steps, specifically $T = 25$, as opposed to 100 steps typically used in our model. A noticeable increase in CD and decrease in STD are observed after reducing the number of diffusion steps, especially in the case of loose outfits like "felice-004", where CD rises from 5.58 to 6.33, and STD drops from 6.54 to 3.89. These outcomes emphasize the importance of diffusion models for progressive and diversified modeling of clothed humans.
- *w/o diffusion*: This variation omits diffusion-based progressive refinement, resulting in single-step deterministic predictions. Across the three outfits in Tab. 4, we observe a notable decrease in CD and also a lack of diverse outcomes, further reinforcing the importance of diffusion models in clothed human modeling.
- *w/o physics*: This variation omits two physics-based losses defined in Eq. (16) and (17). The rapid increase in CD and DE, compared with Ours (T=100), demonstrates the utility of physics-based constraints.
- *w/o dynamics*: This variation uses a single frame as input, where only pose-dependent features are considered, without dynamic modeling and two physics-based losses mentioned above. In Tab. 4, we observe larger errors in both CD and DE across three outfits where dynamic features are omitted, showing

the effectiveness of explicit dynamics modeling in generating clothed humans with smooth and natural clothing details.

- *w/o all*: This variation removes physics-based losses, dynamics modeling, and diffusion-based refinement from our model, resulting in the poorest performance.

Furthermore, we showcase two examples in Fig. 8 where we compare our model with the scenerios Ours(*w/o dynamics*) and Ours(*w/o diffusion*). It can be observed that *w/o dynamics* normally presents results lack of correct and smooth clothing patterns due to the ignorance of explicit dynamic features and physics-inspired loss functions. *w/o diffusion* typically gives results lack of clothing details especially for the loose skirt presented in Fig. 8. This illustrate the importance of iterative refinement introduced in our proposed diffusion-based model for the modeling of complicated clothing dynamics.

Methods	Outfits																	
	christine-027 knee dress short sleeve			janett-025 short skirt long sleeve			felice-004 long dress tank top			alexandra-006 loose blouse long pants			eric-035 blazer jacket long pants			all -		
	CD↓	DE↓	STD↑	CD↓	DE↓	STD↑	CD↓	DE↓	STD↑	CD↓	DE↓	STD↑	CD↓	DE↓	STD↑	CD↓	DE↓	STD↑
Ours (T=100)	**1.35**	**3.20**	**6.31**	**1.03**	**2.79**	**6.20**	**5.58**	**3.08**	**6.54**	**1.50**	**3.10**	**7.80**	**1.30**	**3.71**	**6.92**	**1.14**	**3.12**	**6.63**
Ours (T=25)	1.50	3.21	3.78	1.12	2.81	3.56	6.33	3.12	3.89	1.54	**3.10**	3.94	1.32	3.72	3.90	1.18	3.13	3.87
w/o diffusion	1.66	3.23	-	1.20	2.82	-	6.93	3.16	-	1.67	3.12	-	1.33	3.78	-	1.30	3.18	-
w/o physics	1.38	3.29	5.85	1.10	2.85	6.02	5.81	3.26	5.96	1.54	3.16	7.25	1.30	3.87	6.12	1.19	3.19	6.21
w/o dynamics	1.39	3.32	5.05	1.10	2.85	5.01	5.87	3.30	5.19	1.56	3.19	5.68	1.31	3.92	5.37	1.20	3.23	5.28
w/o all	1.72	3.36	-	1.23	2.88	-	7.30	3.37	-	1.70	3.25	-	1.35	3.99	-	1.36	3.27	-

Table 4. **Ablation Study.** Dynamic Error (DE) highlights the importance of dynamics modeling in our model, and the standard deviation (STD) of generations presents the capability of diversified modeling. We *underline* key comparisons **per column** for enhanced clarity of ablation results.

5.4 Discussions

Diverse Generation. Our diffusion-based approach can generate diverse patterns of clothes. In Fig. 9, we showcase two examples of this capability, each presented in three variations. These variations illustrate how our method produces distinct clothing details that vary from one another while maintaining a high level of naturalness and smoothness. For a more thorough exploration of the range and quality of these generated patterns, we invite you to view the supplementary video, which provides an in-depth depiction of the diverse garment generations.

Interpolate Different Garment Features. In Fig. 10, we present the qualitative results of applying linear interpolation to combine two distinct garment features. By blending these features, we explore how intermediate values transition between different styles, textures, and patterns, resulting in seamless transformations that showcase diverse aesthetic attributes. Through gradual shifts of garment features, the results highlight the flexibility of our model to synthesize realistic variations that retain elements from both original garment types.

Effect of Generation Time Duration. To examine the effect of varying temporal durations in the generated sequences, we downsample the FPS of an 8-frame input motion sequence to different rates. The original FPS of motions in the ReSynth dataset is 15. When downsampled to 15, 10, 5, 3, 1, 0.5 FPS, the output durations are 0.53, 0.8, 1.6, 2.7, 8, 16 seconds, respectively. Longer durations result in greater variation in the generated outputs. We evaluate the CD and NML on the typical outfit "felice-004" in the ReSynth dataset. The results are reported in Tab. 5. We observe a significant performance drop when the duration of the generated sequence exceeds 2.7

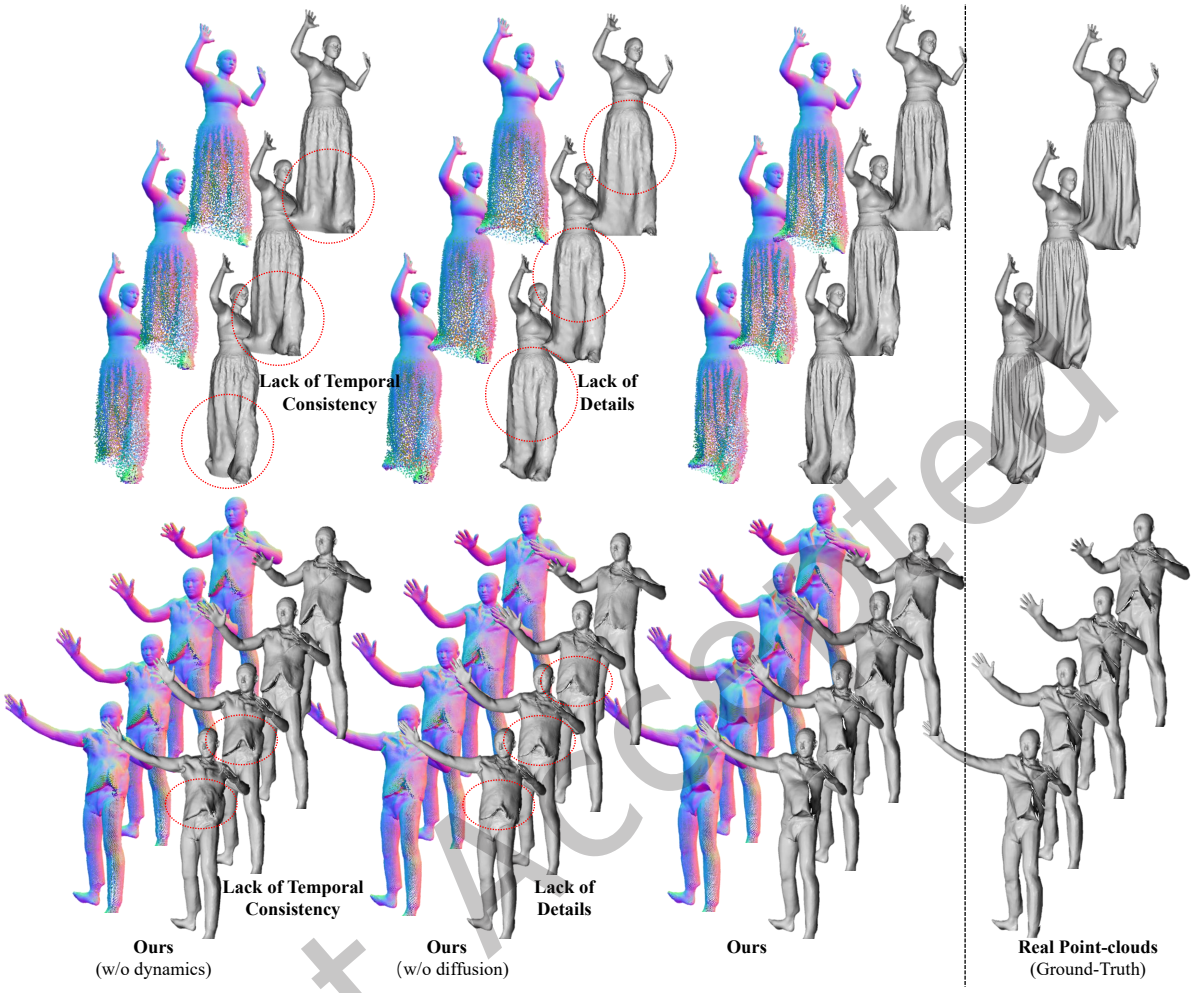

Fig. 8. **Qualitative Results on Ablation Study.** *w/o dynamics* presents results lack of temporally consistent clothing patterns, while *w/o diffusion* gives results lack of clothing details.

Input Motion FPS	15	10	5	3	1	0.5
Generation Time Duration	0.53s	0.8s	1.6s	2.7s	8s	16s
CD↓	5.58	5.61	6.52	7.63	11.25	12.3
NML↓	1.14	1.14	1.20	1.28	1.37	1.42

Table 5. **Evaluation of Generation Time Duration.** We report the CD and NML on the typical outfit "felice-004" in the ReSynth dataset.

seconds (3 FPS). This may be due to the greater variation in the input motion sequence over longer durations, which disrupts the smoothness of the generated clothing dynamics.

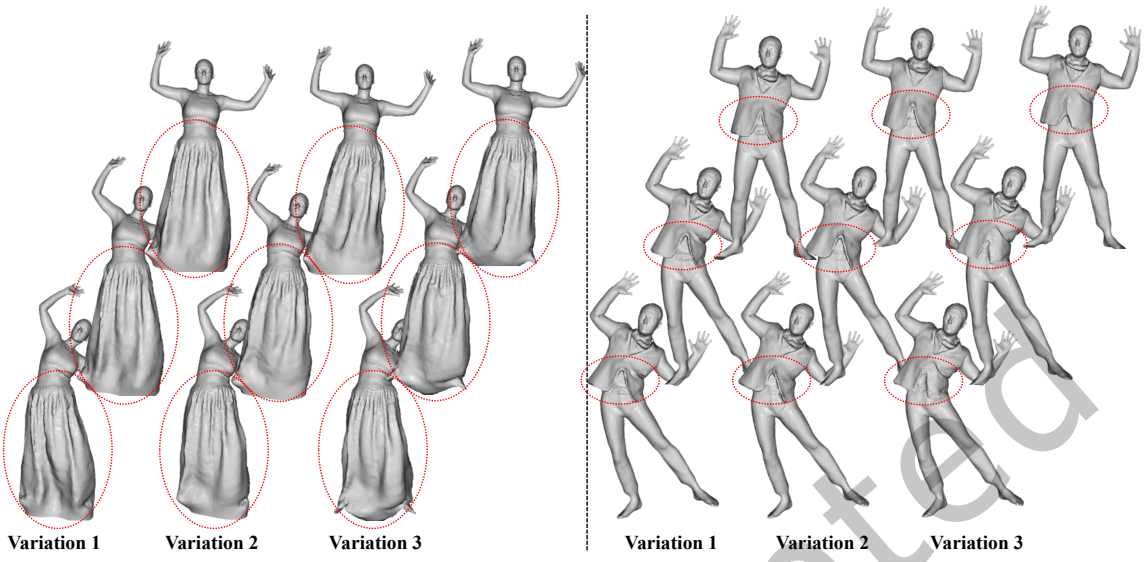

Fig. 9. **Diversified Generation of Clothes Pattern.**

	DPF [35]	POP [27]	Ours (T=100)	Ours (T=25)	w/o diffusion
Time	30†	0.020	0.514	0.141	0.022

Table 6. **Inference Time Per Frame in Seconds.** Results are obtained on a single A100 GPU, while †DPF [35] uses a more powerful RTX6000 GPU.

Limitations. Our method utilizes a diffusion-based iterative refinement process, which requires 0.514 seconds per frame for inference, as detailed in Tab. 6. While this is significantly more efficient compared to the most recent method, DPF (over 30 seconds), it remains slower than single-step inference approaches, such as our ablation case *w/o diffusion* (0.022 seconds) or POP (0.020 seconds). To address this, we could explore learning-based efficient sampling strategies, as utilized in some diffusion models [46]. Additionally, replacing linear blend skinning weights with learnable skinning methods [35] could further enhance performance.

Failure Cases. In Fig. 11, we present two failure cases from the CAPE dataset, highlighting specific challenges our method faces. It can be observed that our approach struggles with accurately modeling tight garments in the scenarios where certain parts of the clothing are significantly distant from the body during specific reference motions. This issue may stem from the regularization of clothing displacement lengths during model training, which might overly constrain the model, limiting its ability to capture large deformations. To address these challenges, introducing adaptive regularization techniques that account for varying garment types and their interaction with the body could improve the model's robustness. Additionally, expanding the dataset to include more diverse examples, particularly with tight garments in complex poses, could provide the model with a richer set of observations to learn from. Enhancing data augmentation techniques to simulate these challenging conditions could also help in overcoming the limitations.

In Fig. 7 of unseen outfit animation, some minor artifacts are observed around the neck area, specifically at the boundary between the clothes and neck. These artifacts likely result from the model encountering limited instances of loose clothing around the neck in the provided unseen point cloud frames. Despite this, our method

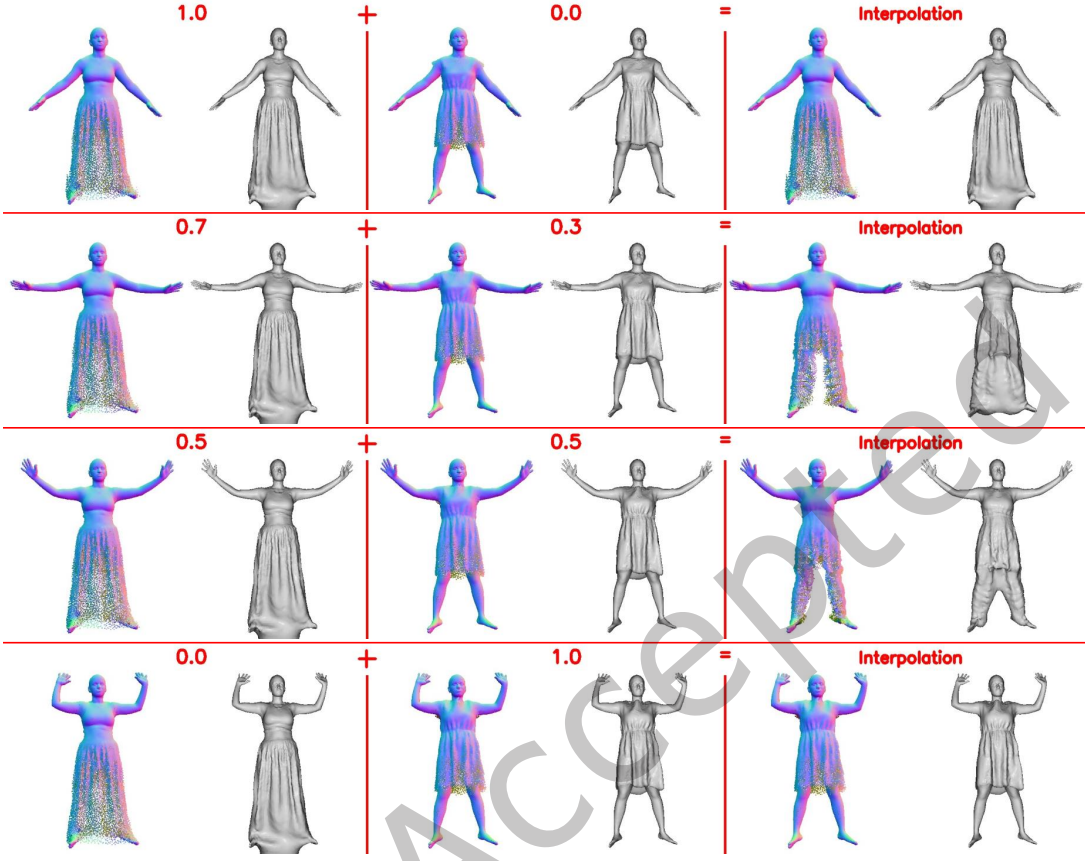

Fig. 10. **Interpolation Through Different Garment Features.**

remains effective overall, with these artifacts being exceptions rather than the norm. To mitigate these issues, future work could focus on enhancing the dataset with more diverse examples of loose clothing and refining the model's handling of complex boundaries through targeted training and post-processing techniques.

6 CONCLUSION

In this work, we introduced ClothDiffuse, a diffusion-based method for learning clothed human dynamics, enabling animations with natural clothing details from a reference body motion. ClothDiffuse is the first data-driven approach to integrate three significant elements into this task: dynamics modeling, iterative refinement, and diversified generation. Our method not only advances the generation of high-fidelity clothed humans but also addresses persistent challenges in the field, such as presenting convincing visual effects for cloth dynamics, maintaining temporal smoothness, and generating diverse clothing patterns. Empirical results demonstrate that incorporating these elements significantly enhances the generation of clothed humans with smooth and natural clothing details over time, especially for loose clothes, reducing chamfer distance and effectively adapting to out-of-domain clothing types.

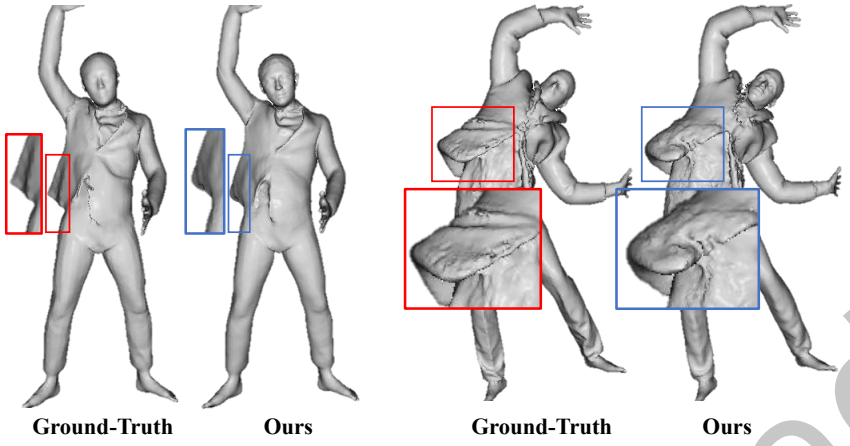

Fig. 11. **Failure Cases.**

Future work could focus on adaptive regularization techniques to better accommodate varying garment types and their interaction with the body, expanding the dataset to include more diverse examples, particularly with tight garments in complex poses, and refining data augmentation and post-processing methods to address the discussed limitations and further improve the robustness of ClothDiffuse.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work was partially supported by grants from National Natural Science Foundation of China (62372441, U22A2034), in part by Guangdong Basic and Applied Basic Research Foundation (2023A1515030268), in part by Shenzhen Science and Technology Program (Grant No. RCYX20231211090127030, JSGG20220831105002004).

REFERENCES

[1] Jan Bednarik, Shaifali Parashar, Erhan Gundogdu, Mathieu Salzmann, and Pascal Fua. 2020. Shape reconstruction by learning differentiable surface representations. In *CVPR*.
[2] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020. Pbns: Physically based neural simulator for unsupervised garment pose space deformation. *ACM TOG* (2020).
[3] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2022. Neural cloth simulation. *ACM TOG* (2022).
[4] Andrei Burov, Matthias Nießner, and Justus Thies. 2021. Dynamic surface function networks for clothed human bodies. In *ICCV*.
[5] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. 2021. Semi-supervised Synthesis of High-Resolution Editable Textures for 3D Humans. In *CVPR*.
[6] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. 2022. gdna: Towards generative detailed neural avatars. In *CVPR*.
[7] Xiaowei Chen, Xiao Jiang, Lishuang Zhan, Shihui Guo, Qunsheng Ruan, Guoliang Luo, Minghong Liao, and Yipeng Qin. 2023. Full-body Human Motion Reconstruction with Sparse Joint Tracking Using Flexible Sensors. *ACM Trans. Multimedia Comput. Commun. Appl.* (2023).
[8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. 2021. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*.
[9] Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *CVPR*.
[10] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. Smplicit: Topology-aware generative model for clothed people. In *CVPR*.
[11] Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. 2019. Learning elementary structures for 3d shape generation and matching. In *NeurIPS*.

[12] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. 2022. PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence. In *CVPR*.
[13] Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *CVPR*.
[14] Artur Grigorev, Michael J Black, and Otmar Hilliges. 2023. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *CVPR*.
[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.
[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research* (2022).
[17] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. 2022. BodyMap: Learning Full-Body Dense Correspondence Map. In *CVPR*.
[18] Boyan Jiang, Xinlin Ren, Mingsong Dou, Xiangyang Xue, Yanwei Fu, and Yinda Zhang. 2022. LoRD: Local 4d implicit representation for high-fidelity dynamic human modeling. In *ECCV*.
[19] Martin Kilian, Niloy J Mitra, and Helmut Pottmann. 2007. Geometric modeling in shape space. In *ACM SIGGRAPH*.
[20] Hyomin Kim, Hyeonseo Nam, Jungeon Kim, Jaesik Park, and Seungyong Lee. 2022. LaplacianFusion: Detailed 3D Clothed-Human Body Reconstruction. *ACM TOG* (2022).
[21] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. 2022. Learning implicit templates for point-based clothed human modeling. In *ECCV*.
[22] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. 2019. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM TOG* (2019).
[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM TOG* (2015).
[24] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. 2021. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *CVPR*.
[25] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. 2022. Neural Point-based Shape Modeling of Humans in Challenging Clothing. In *Int. Conf. 3D Vision*.
[26] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. 2020. Learning to dress 3d people in generative clothing. In *CVPR*.
[27] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. 2021. The power of points for modeling humans in clothing. In *ICCV*.
[28] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. 2021. LEAP: Learning articulated occupancy of people. In *CVPR*.
[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.
[30] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. 2022. SPAMs: Structured implicit parametric models. In *CVPR*.
[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*.
[32] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR*.
[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*.
[34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. In *CVPR*.
[35] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. 2023. Dynamic Point Fields. In *ICCV*.
[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
[38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*.
[39] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. 2021. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*.
[40] Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2022. Snug: Self-supervised neural dynamic garments. In *CVPR*.
[41] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. 2022. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*.
[42] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. 2022. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*.

[43] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for an animatable human avatar. *ACM TOG* (2021).
[44] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. 2022. Icon: Implicit clothed humans obtained from normals. In *CVPR*.
[45] Han Yan, Haijun Zhang, Jianyang Shi, Jianghong Ma, and Xiaofei Xu. 2023. Toward Intelligent Fashion Design: A Texture and Shape Disentangled Generative Adversarial Network. *ACM Trans. Multimedia Comput. Commun. Appl.* (2023).
[46] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* (2022).
[47] Hang Yu, Chilam Cheang, Yanwei Fu, and Xiangyang Xue. 2023. Multi-view Shape Generation for a 3D Human-like Body. *ACM Trans. Multimedia Comput. Commun. Appl.* (2023).
[48] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In *NeurIPS*.
[49] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. 2023. CloSET: Modeling Clothed Humans on Continuous Surface with Explicit Template Decomposition. In *CVPR*.
[50] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. 2022. Structured local radiance fields for human avatar modeling. In *CVPR*.

Received 29 August 2024; revised 7 November 2024; accepted 5 January 2025