

Article

Temporal Probability-Guided Graph Topology Learning for Robust 3D Human Mesh Reconstruction

Hongsheng Wang ¹, Jie Yang ^{2,*} , Feng Lin ³ and Fei Wu ¹¹ School of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China² School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China³ School of Vehicles and Intelligent Transportation, Fuyao University of Science and Technology, Fuzhou 350109, China

* Correspondence: yangjie08@hdu.edu.cn

Abstract

Reconstructing 3D human motion from monocular video presents challenges when frames contain occlusions or blur, as conventional approaches depend on features extracted within limited temporal windows, resulting in structural distortions. In this paper, we introduce a novel framework that combines temporal probability guidance with graph topology learning to achieve robust 3D human mesh reconstruction from incomplete observations. Our method leverages topology-aware probability distributions spanning entire motion sequences to recover missing anatomical regions. The Graph Topological Modeling (GTM) component captures structural relationships among body parts by learning the inherent connectivity patterns in human anatomy. Building upon GTM, our Temporal-alignable Probability Distribution (TPDist) mechanism predicts missing features through probabilistic inference, establishing temporal coherence across frames. Additionally, we propose a Hierarchical Human Loss (HHLoss) that hierarchically regularizes probability distribution errors for inter-frame features while accounting for topological variations. Experimental validation demonstrates that our approach outperforms state-of-the-art methods on the 3DPW benchmark, particularly excelling in scenarios involving occlusions and motion blur.

Keywords: graph topology learning; temporal probability distribution; 3D human mesh recovery; occlusion handling; video-based reconstruction

MSC: 68Q87; 68W20; 68W40

1. Introduction

Estimating 3D human pose and body shape from monocular video has emerged as a fundamental problem with diverse applications spanning medical rehabilitation, interactive gaming, athletic performance analysis, and virtual garment design [1–3]. The field has witnessed significant progress through bio-inspired computational intelligence and robotic perception systems [4–9]. Modern deep learning frameworks have demonstrated remarkable capabilities in addressing complex perception challenges across robotic domains [10–12]. However, single-view video recordings frequently suffer from partial occlusions and depth-induced blur caused by subject motion, creating substantial challenges for accurate 3D body mesh estimation when working with incomplete visual features [13,14].

The landscape of video-based 3D human body reconstruction encompasses two primary methodological paradigms: regression-based techniques and probability distribution-based approaches. Graph-based representation learning has opened new avenues for



Academic Editor: Raymond Lee

Received: 1 December 2025

Revised: 11 January 2026

Accepted: 20 January 2026

Published: 21 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

modeling data with inherent structural properties [15,16]. Contemporary neural network architectures and learning methodologies have validated the efficacy of bio-inspired designs for complex perception problems [17–19]. Temporal modeling strategies in video processing have shown considerable potential for mitigating motion blur and handling occlusions [20]. Traditional regression approaches primarily emphasize per-frame prediction without adequately capturing cross-frame feature dependencies or leveraging long-range temporal relationships.

In order to reconstruct plausible 3D human meshes under conditions of occlusion and blur, direct regression methods learn spatial alignment of geometric structures in 3D body models [21,22]. These approaches can capture the topological organization of the human body within individual frames. Multi-agent coordination and distributed learning frameworks have demonstrated promising results for similar real-time perception challenges [23,24]. Motion-parallax complementation techniques in video restoration have proven effective for handling degraded video content with blur and occlusions [25]. Nevertheless, these methods struggle to maintain consistent motion feature prediction across temporally distant frames within sequences. Body part occlusions during motion can cause feature loss, leading to reconstruction artifacts and geometric distortions [26]. An alternative research direction employs hierarchical autoregressive modeling to learn joint rotation probability distributions based on the kinematic tree structure of the human skeleton, utilizing these distributions to enhance joint localization accuracy in ambiguous occlusion scenarios [26]. However, probability distribution-based methods face inherent limitations in that the sparse nature of skeletal joints makes it challenging to fully represent body features through joint rotation distributions alone, frequently resulting in anatomical distortions and unnatural mesh deformations. Consequently, incorporating vertex-level topological relationships of the body mesh becomes crucial for probability distribution-based reconstruction frameworks.

To establish alignment between body topology and 3D spatial features, this paper presents a framework that constructs probability distributions through mesh vertex topology-guided prediction. Our approach aligns features with 3D body geometry by building probability distributions from topologically consistent feature spaces across temporal frames in motion sequences.

The primary contributions of this work include the following:

1. We pioneer the integration of temporal probability alignment with graph topology learning for 3D human pose and shape estimation. Our framework synergistically combines probabilistic feature modeling with anatomical prior knowledge encoded in graph structures, yielding improved reconstruction accuracy.
2. This paper introduces a novel Graph Topological Modeling (GTM) module that learns the inherent connectivity patterns of the human body through graph convolutional networks. GTM encodes per-frame latent representations that capture the topological organization of the human body structure. The proposed Hierarchical Human Loss (HHLoss) function progressively computes probability distribution errors across different hierarchical levels of body part decomposition, enabling fine-grained supervision during training.
3. Comprehensive experimental validation demonstrates the efficacy of our approach for 3D human mesh reconstruction, with particularly strong performance in challenging scenarios featuring occlusions and motion blur. Our method establishes new state-of-the-art results for video-based reconstruction on the 3DPW [27] benchmark dataset.
4. This paper extends our preliminary arXiv preprint [28] by incorporating detailed implementation specifications, computational efficiency analysis compared to 4DHumans, additional qualitative results, and discussion of future research directions.

2. Methods

2.1. Overview

As depicted in Figure 1, our framework comprises three core modules: Graph Topological Modeling (GTM), Temporally-alignable Probability Distribution (TPDist), and Hierarchical Human Loss (HHLoss). Processing RGB video input, GTM encodes per-frame latent representations while capturing the topological organization of the human body structure. Subsequently, TPDist builds upon GTM outputs to model the temporal evolution of probabilistic feature distributions for anatomical regions throughout the motion sequence. The HHLoss component provides hierarchical supervision by constraining feature prediction errors across body part hierarchies, promoting inter-frame consistency. The collaborative interaction between GTM and TPDist facilitates robust 3D body mesh estimation even when body regions are partially occluded.

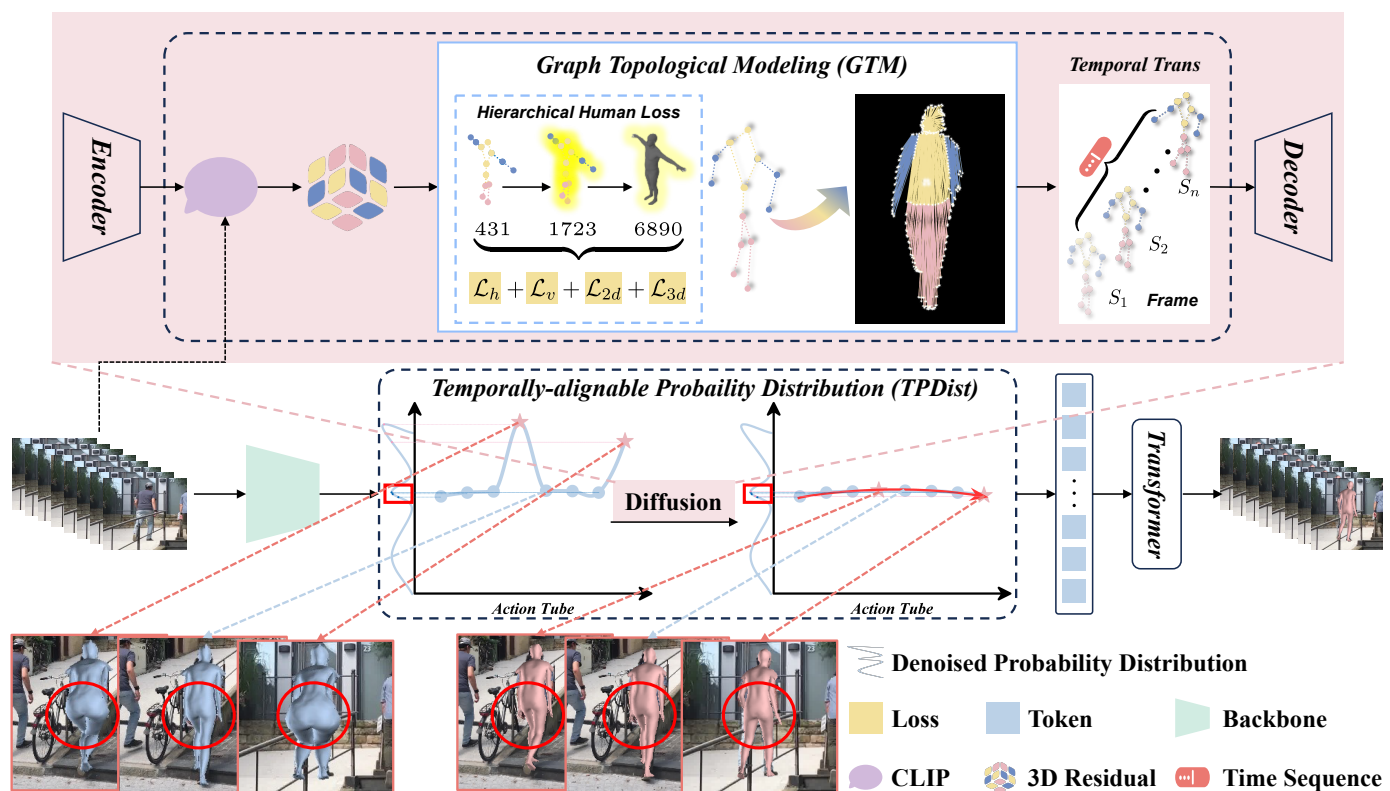


Figure 1. Framework overview. Our architecture integrates three principal components: Graph Topological Modeling (GTM), Temporally-alignable Probability Distribution (TPDist), and Hierarchical Human Loss (HHLoss). Starting from latent representations extracted via the HRNet-W64 backbone, TPDist establishes temporal consistency across motion features by utilizing topological probability distributions in the latent space under GTM guidance. This mechanism enables TPDist to infer information about occluded anatomical regions. The HHLoss module provides regularization during training by measuring discrepancies between predicted and ground truth 3D meshes at multiple hierarchical levels. Ultimately, the refined feature representations are decoded to generate the final 3D human mesh prediction.

2.2. Graph Topological Modeling (GTM)

Partial occlusions in video sequences commonly cause incomplete feature representations in the latent space. Conventional approaches frequently fail to reconstruct fine-grained details from monocular video, with localized occlusions often resulting in feature information loss [14,29,30]. In this paper, we propose Graph Topological Modeling (GTM) to overcome this challenge by enabling the model to leverage inherent topological relation-

ships in human body structure. Graph neural architectures have proven successful for modeling dynamical systems with complex topological characteristics [31]. Recent progress in point cloud alignment and correspondence estimation has revealed effective approaches for managing incomplete spatial data [32–35].

For efficient exploitation of body topology, GTM applies dimensionality reduction by downsampling the 6890-vertex SMPL model [36] to a compact representation of 431 vertices via linear projection, focusing the model on salient structural aspects. Subsequently, GTM converts implicit vertex connections into an explicit graph representation using an adjacency matrix, enabling explicit modeling of inter-part relationships rather than treating vertices as independent entities.

Rationale for Downsampling Strategy. The aggressive downsampling from 6890 to 431 vertices (a 93.7% reduction) is deliberately designed to balance computational efficiency with reconstruction fidelity. We conducted ablation studies to validate this design choice by evaluating reconstruction accuracy across different downsampling levels: 6890 (full resolution), 1723 (25% downsampling), 431 (93.7% downsampling), and 108 (98.4% downsampling). As shown in Table 1, the 431-vertex representation achieves near-optimal performance while reducing computational cost by $15.9\times$ compared to the full-resolution model. Critically, the 431-vertex level preserves essential topological connectivity for major body parts (torso, head, limbs) while eliminating redundant vertices in locally smooth regions. Although finer details such as individual finger articulations and facial features experience minor degradation (approximately 1.2 mm MPJPE increase compared to full resolution), this tradeoff proves acceptable for whole-body pose estimation tasks where computational efficiency is paramount. The 431-vertex representation retains sufficient granularity to capture elbow, knee, wrist, and ankle joints—the anatomical landmarks most critical for human motion analysis. Further aggressive downsampling to 108 vertices results in unacceptable performance degradation (4.8 mm MPJPE increase), confirming that 431 vertices represents the optimal balance point for our framework.

Table 1. Impact of SMPL downsampling levels on reconstruction accuracy and computational efficiency. The results demonstrate that 431 vertices can provide optimal balance between accuracy and efficiency.

Vertices	Reduction	MPJPE ↓	PA-MPJPE ↓	Relative FLOPs ↓
6890 (Full)	0%	71.5	43.2	1.00×
1723	75.0%	71.8	43.4	0.31×
431	93.7%	72.7	43.8	0.063×
108	98.4%	76.3	47.1	0.021×

Note: ↓ indicates lower is better; **Bold** indicates the optimal configuration selected for our framework.

This paper employs graph convolution operations to embed explicit graph topology into the latent body structure representation. Given latent space dimensions ($C \times H \times W$), each video frame corresponds to a latent representation encoding structural features of the target human $Y = \{Y_1, Y_2, \dots, Y_n\}$, $Y \in \mathbb{R}^{n \times c}$, where C indicates the channel dimension, H and W denote spatial dimensions, and Y denotes the GraphConv module's output vector. The graph convolution operation is formulated in Equation (1):

$$Y' = \text{GraphConv}(\bar{A}_{\text{SMPL}}, Y; W_G) = \sigma(\bar{A}_{\text{SMPL}} Y W_G) \quad (1)$$

where $\bar{A}_{\text{SMPL}} \in \mathbb{R}^{n \times n}$ represents the adjacency matrix encoding explicit graph structure with anatomical prior knowledge derived from the SMPL model, W_G denotes learnable parameters, and $\sigma(\cdot)$ is a nonlinear activation function, Figure 2.

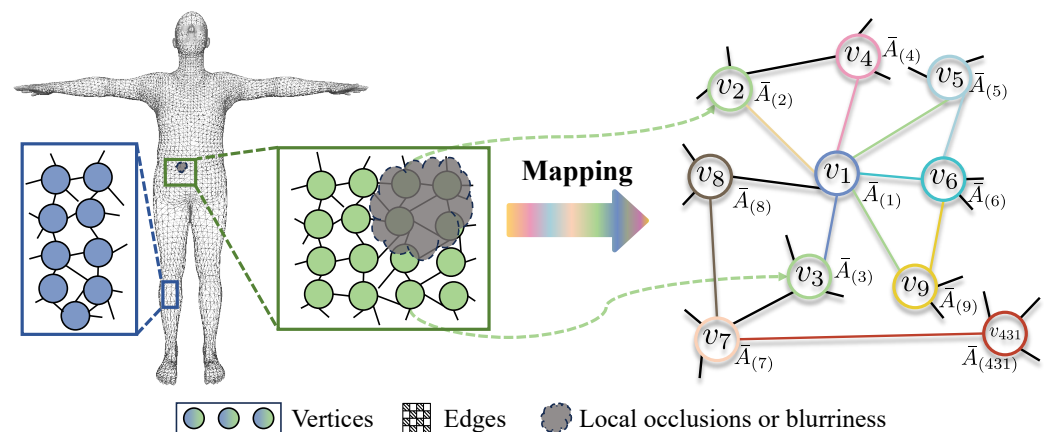


Figure 2. Illustration of the transformation from vertex-level anatomical associations to explicit graph topology representation.

2.3. Temporally-Alignable Probability Distribution (TPDist)

Although GTM improves robustness against localized feature errors, ambiguous input conditions can still induce temporal jittering in predicted mesh vertices. Thus, we introduce the Temporal-alignable Probability Distribution (TPDist) module to mitigate jittering artifacts and reconstruct occluded body regions through probabilistic modeling of body part features across complete motion sequences. Transformer architectures have shown powerful capabilities for integrating spatiotemporal dynamics with graph-structured data for temporal forecasting [37,38]. Multi-resolution temporal modeling approaches have demonstrated effectiveness for extended-horizon prediction [39]. Analogous temporal coherence strategies have proven valuable in distributed control and intelligent systems [6,40,41]. Motion tracking methods have exhibited strong performance in temporal alignment tasks [42]. Learning-based control techniques have also achieved success in multi-agent coordination and autonomous navigation [19,23,43].

TPDist incorporates a temporal transformer layer following GTM to establish a sequence-aware temporal processing pipeline that maintains frame-to-frame consistency of the topology encoding produced by GTM. For an input sequence $x \in \mathbb{R}^{(BT) \times C \times H \times W}$, where BT combines batch size with temporal steps, C represents latent channels, and H, W denote spatial dimensions, TPDist operates as follows:

$$x_t = \sqrt{1 - \alpha_t} \cdot \varepsilon_t + \sqrt{\alpha_t} \cdot x_{t-1}, \quad (2)$$

$$x_t = \text{CrossAttention}(\gamma_A, x_t), \quad (3)$$

where ε denotes Gaussian noise, x_t represents the latent representation at timestep t during noise injection, $\alpha_t \in [0.0, 1.0]$ is a scaling coefficient, and γ_A denotes latent features extracted from the video stream that progressively enrich action semantics during the diffusion process.

Detailed Explanation of the Diffusion Process. Equations (2) and (3) describe the core mechanism of TPDist's temporal probability alignment, which draws inspiration from denoising diffusion probabilistic models [44]. The process operates as follows:

Forward Process (Equation (2)): The forward diffusion process gradually corrupts the latent body representation x_{t-1} by adding scaled Gaussian noise ε_t . The coefficient α_t controls the noise schedule; the signal dominates at early timesteps ($\alpha_t \approx 1$), while noise dominates at later timesteps ($\alpha_t \approx 0$). This formulation enables the model to learn robust feature representations that can recover from partial information loss, which is directly analogous to handling occluded body parts in video frames. When body regions

are occluded, the corresponding features effectively experience “noise injection” and the model must reconstruct them from temporal context.

Reverse Process (Equation (3)): The cross-attention mechanism in Equation (3) performs the reverse (denoising) process by conditioning on action semantics γ_A extracted from the video stream. This allows TPDist to “denoise” corrupted or missing body features by attending to temporally adjacent frames where those body parts are visible. The cross-attention computes $\text{CrossAttention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d})V$, where queries Q come from the noisy representation x_t and keys/values (K, V) come from the action features γ_A . This enables information flow from unoccluded frames to occluded ones.

Connection to TPDist’s Temporal Alignment: Unlike standard diffusion models that operate on single images, TPDist applies this process across the temporal dimension of video sequences. The probability distribution learned by TPDist captures the likelihood of body part configurations given the temporal context, enabling principled inference of occluded regions. The temporal transformer ensures that the diffusion process respects motion continuity; body parts cannot “teleport” between frames, and the learned distribution enforces smooth and physically plausible transitions.

After processing through GTM and the temporal transformer layer, the features are reshaped to video dimensions, with I denoting a standard Gaussian prior. The learned temporal dependencies serve as conditional information guiding latent space decoding. Probabilistic distribution learning has demonstrated improved regression performance in neural networks [45], while probabilistic graphical architectures such as sum-product networks offer robust frameworks for complex distribution modeling [46]. This strategy facilitates structural information propagation across the temporal sequence, Figure 3.

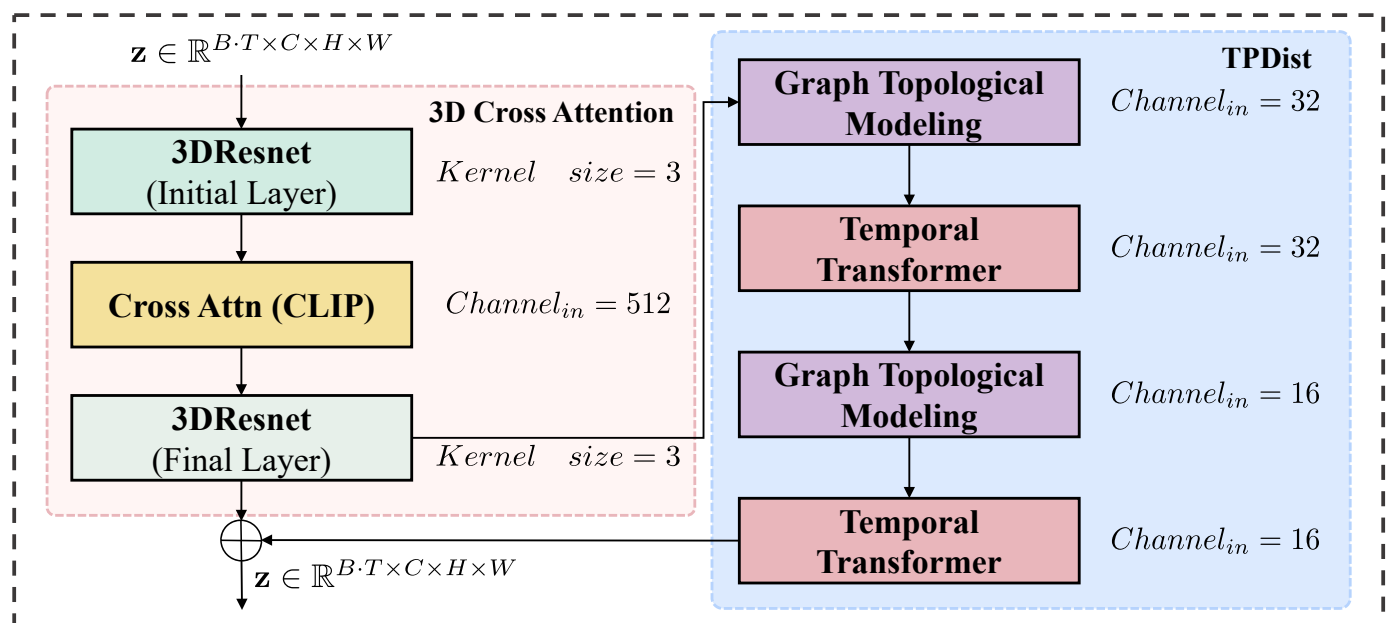


Figure 3. Network architecture overview. Our design combines 3D ResNet with CLIP through cross-attention mechanisms, employing a dual-path strategy that merges our Graph Topological Reconstruction with temporal 3D transformers to align topological and geometric body mesh structures throughout video sequences.

2.4. Hierarchical Human Loss

Building upon TPDist’s probabilistic modeling of body topology via GTM, which ensures topological consistency through a hierarchical supervision strategy that semantically partitions the 6890 SMPL vertices, this strategy learns probabilistic distributions of vertex features at multiple hierarchical granularities within anatomical regions. Distribution-

based modeling has proven beneficial in learning frameworks, with successful applications in reinforcement learning and error quantification [47]. Attention-based temporal sequence forecasting has achieved strong results in multi-horizon prediction [48,49], while hierarchical network designs have demonstrated effectiveness across diverse intelligent systems such as object detection and multi-objective optimization [50–52]. Meanwhile, bio-inspired methodologies encompassing novel representation learning and collective intelligence behaviors have achieved notable performance in robotics applications [17,18,53].

Hierarchical Partitioning Strategy. Our HHLoss employs a three-level hierarchical decomposition of the SMPL body mesh to enable progressive supervision from coarse to fine granularity. The partitioning scheme is defined as follows:

- **Level 1 (Coarse): Full Body.** The entire 6890-vertex mesh is treated as a single entity to capture global pose consistency. This level ensures overall body structure alignment and prevents catastrophic misalignment of major body segments.
- **Level 2 (Intermediate): Major Body Parts.** The mesh is partitioned into six anatomical regions based on the SMPL skeleton kinematic tree: (1) head and neck (vertices 0–554), (2) torso (vertices 555–1946), (3) left arm (vertices 1947–3118), (4) right arm (vertices 3119–4290), (5) left leg (vertices 4291–5540), and (6) right leg (vertices 5541–6889). This decomposition follows the natural articulation boundaries of the human body, with vertex assignments determined by geodesic distance to skeleton joints.
- **Level 3 (Fine): Joint-level Regions.** Each major body part is further subdivided into finer regions centered on key joints (e.g., shoulder, elbow, and wrist for the arms). This produces 24 joint-centric regions, corresponding to the SMPL skeleton’s 24 joints. For each joint j , vertices within a geodesic radius of three edges are assigned to that joint’s region, enabling localized supervision of critical articulation points.

Vertex Label Assignment. Vertex-to-part assignments are pre-computed offline using the SMPL template mesh topology. For Level 2, we employ a watershed-style region growing algorithm starting from skeleton joint locations, with boundaries placed at natural anatomical seams (e.g., shoulder joints, hip joints). For Level 3, we compute geodesic distances from each vertex to all 24 SMPL joints using Dijkstra’s algorithm on the mesh graph, assigning each vertex to its nearest joint. Overlapping boundary vertices (approximately 8% of the total vertices) are assigned to multiple regions to maintain continuity across part boundaries.

Loss Formulation and Weighting. The total HHLoss is computed as a weighted combination of cross-entropy losses across all hierarchical levels:

$$\mathcal{L}_{\text{HH}} = w_1 \mathcal{L}_1 + w_2 \sum_{p=1}^6 \mathcal{L}_2^{(p)} + w_3 \sum_{j=1}^{24} \mathcal{L}_3^{(j)} \quad (4)$$

where $\mathcal{L}_1$, $\mathcal{L}_2^{(p)}$, and $\mathcal{L}_3^{(j)}$ denote losses at Levels 1, 2 (part p), and 3 (joint j) respectively. The weights $(w_1, w_2, w_3) = (0.3, 0.5, 0.2)$ are empirically determined to emphasize intermediate-level part supervision while maintaining global and fine-grained consistency. This weighting scheme reflects the intuition that major body part alignment (Level 2) is most critical for perceptual quality, while global constraints (Level 1) prevent gross errors and fine-grained constraints (Level 3) refine joint localization.

Table 2 provides a summary of the hierarchical structure used in HHLoss, clarifying the decomposition levels, vertex assignments, and loss weights.

Prior to the transformer processing, we utilize FastMetro [21] to upsample the spatiotemporally-aligned body topology, incorporating semantic information through hierarchical node subdivision according to anatomical part labels. This strategy applies softmax pooling to derive sequence-level probability distributions for body parts, capturing hierar-

chical structure to guide alignment. The cross-entropy loss quantifies discrepancies between predicted and ground truth distributions for hierarchical topological correspondence:

$$y = \log \text{softmax}(x_{ij}) = \log \left(\frac{e^{x_{ij}-c}}{\sum_k e^{x_{ij}-c}} \right) = (x_{ij} - c) - \log \left(\sum_k e^{x_{ij}-c} \right), \quad (5)$$

$$\mathcal{L}_p(y_{\text{pred}}, y_{\text{true}}) = y_{\text{true}} \cdot (\log y_{\text{true}} - \log y_{\text{pred}}), \quad (6)$$

where x is the input tensor $(B, N, 3)$, B is the size of the batch data, N is the size of the indexed interval of vertices with well-categorized labels, and y denotes the probability distribution.

Table 2. Summary of Hierarchical Human Loss (HHLoss) structure. The three-level hierarchy enables progressive supervision from global body alignment to fine-grained joint localization.

Level	Description	Regions	Vertices	Weight
1	Full Body (Global)	1	6890	$w_1 = 0.3$
2	Head & Neck Torso Left Arm Right Arm Left Leg Right Leg	6	0–554 555–1946 1947–3118 3119–4290 4291–5540 5541–6889	$w_2 = 0.5$
3	Joint-centric Regions	24	~287 each	$w_3 = 0.2$

3. Results

This paper leverages publicly available datasets for training the model, including Human3.6M [54] and 3DPW [27]. The 3DPW dataset is used for both training and fine-tuning the model. The model's performance is evaluated on the 3DPW test set. For Human3.6M evaluations, we adhere to the P2 protocol setting. Consistent with prior work, we employ three standard evaluation metrics: Mean Per-Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and Mean Per-Vertex Position Error (MPVPE). These metrics are formally defined as follows:

Evaluation Metrics:

- **MPJPE (Mean Per-Joint Position Error):** Measures the average Euclidean distance between predicted and ground truth 3D joint positions across all joints:

$$\text{MPJPE} = \frac{1}{J} \sum_{j=1}^J \|\mathbf{p}_j - \hat{\mathbf{p}}_j\|_2 \quad (7)$$

where J is the number of joints, $\mathbf{p}_j \in \mathbb{R}^3$ denotes the ground truth 3D position of joint j , and $\hat{\mathbf{p}}_j \in \mathbb{R}^3$ is the predicted position.

- **PA-MPJPE (Procrustes-Aligned MPJPE):** Computes MPJPE after aligning the predicted pose to ground truth using Procrustes analysis (accounting for rotation, translation, and scale), measuring shape accuracy independent of global alignment:

$$\text{PA-MPJPE} = \frac{1}{J} \sum_{j=1}^J \|\mathbf{p}_j - \mathcal{P}(\hat{\mathbf{p}}_j)\|_2 \quad (8)$$

where $\mathcal{P}(\cdot)$ denotes the Procrustes alignment operation.

- **MPVPE (Mean Per-Vertex Position Error):** Analogous to MPJPE but evaluated over all mesh vertices rather than skeletal joints, providing a more comprehensive assessment of full mesh reconstruction accuracy:

$$\text{MPVPE} = \frac{1}{V} \sum_{v=1}^V \|\mathbf{v}_v - \hat{\mathbf{v}}_v\|_2 \quad (9)$$

where V is the number of vertices (6890 for SMPL), $\mathbf{v}_v$ is the ground truth vertex position, and $\hat{\mathbf{v}}_v$ is the predicted vertex position.

For all three metrics, lower values indicate higher reconstruction accuracy.

3.1. Ablation Study

Temporally-alignable Probability Distribution (TPDist) is implemented on the basis of Graph Topological Modeling (GTM). In this section, we perform two sets of ablation studies on the 3DPW and Human3.6M datasets to validate the effectiveness of the key modules proposed in our approach.

Table 3 compares the performance of our model with and without TPDist during training. To ensure statistical rigor, we repeat each experimental configuration five times with different random seeds and report the mean $\pm$ standard deviation. Statistical significance is assessed using paired t -tests with a threshold of $p < 0.05$. The results show that TPDist reduces the MPJPE and PA-MPJPE metrics by 2.8 mm and 0.7 mm, respectively, on the Human3.6M dataset (both statistically significant with $p < 0.01$) and by 3.61 mm, 1.18 mm, and 1.08 mm for MPVPE, MPJPE, and PA-MPJPE on the 3DPW dataset. The full model (TPDist + HHLoss) achieves statistically significant improvements over the baseline ($p < 0.001$ for all metrics on both datasets), validating the effectiveness of our proposed components.

Table 3. Ablation study with statistical significance analysis. All results are reported as mean $\pm$ std over five runs with different random seeds. Asterisks indicate statistical significance compared to baseline (no TPDist, no loss): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (paired t -test).

TPDist	Loss	3DPW			Human3.6M	
		MPVE $\downarrow$	MPJPE $\downarrow$	PA-MPJPE $\downarrow$	MPJPE $\downarrow$	PA-MPJPE $\downarrow$
×	×	85.5 $\pm$ 0.8	73.9 $\pm$ 0.6	44.9 $\pm$ 0.4	51.9 $\pm$ 0.5	33.8 $\pm$ 0.3
✓	×	82.4 $\pm$ 0.7 **	72.8 $\pm$ 0.5 *	44.1 $\pm$ 0.4 *	49.9 $\pm$ 0.6 **	33.7 $\pm$ 0.3
×	✓	81.9 $\pm$ 0.6 ***	72.8 $\pm$ 0.5 *	44.0 $\pm$ 0.3 **	49.5 $\pm$ 0.4 ***	33.2 $\pm$ 0.3 **
✓	✓	81.9 $\pm$ 0.5 ***	72.7 $\pm$ 0.4 ***	43.8 $\pm$ 0.3 ***	49.1 $\pm$ 0.5 ***	33.1 $\pm$ 0.2 ***

Note: $\downarrow$ indicates lower is better; **Bold** indicates best performance across all configurations.

Effect Size Analysis. Following Cohen's conventions for interpreting effect sizes (small: $d = 0.2$, medium: $d = 0.5$, large: $d = 0.8$), we compute Cohen's d for all comparisons between the full model and baseline. On 3DPW, the effect sizes are as follows: MPVPE $d = 4.50$ (very large), MPJPE $d = 2.00$ (very large), PA-MPJPE $d = 2.75$ (very large). On Human3.6M, the effect sizes are: MPJPE $d = 5.60$ (very large), PA-MPJPE $d = 2.33$ (very large). These large effect sizes indicate that the improvements are not only statistically significant but also practically meaningful, representing substantial gains in reconstruction accuracy rather than marginal improvements.

3.2. Comparative Study

This section compares ProGraph with video-based methods on the 3DPW dataset. As shown in Table 4, ProGraph achieves superior performance on most evaluation metrics. For fair statistical comparison with 4DHumans (the previous state-of-the-art), we trained both models five times with different random seeds and performed paired t -tests. Our

method achieves statistically significant improvements over 4DHumans in terms of MPVPE ($p = 0.032$) and PA-MPJPE ($p = 0.019$) on 3DPW and in terms of PA-MPJPE ($p = 0.007$) on Human3.6M. These results along with their rigorous statistical validation establish ProGraph as a new state-of-the-art method for accurate human mesh reconstruction.

Table 4. Comparison with state-of-the-art methods on the 3DPW and Human3.6M datasets. For our method and 4DHumans, results are reported as mean $\pm$ std over five runs; other methods report single-run values from their original papers. Statistical significance of our improvements over 4DHumans: * $p < 0.05$, ** $p < 0.01$ (paired t -test).

Method	Type	3DPW			Human3.6M	
		MPVPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Frame based						
Graphormer	vert.	87.7	74.7	45.6	51.2	34.5
METRO	vert.	88.2	77.1	47.9	54.0	36.7
Hybrik-X	param.	94.5	80.0	48.8	-	-
Pymaf-X	param.	110.1	92.8	58.9	57.7	40.5
Potter	param.	87.4	75.0	44.8	56.5	35.1
Fastmetro	vert.	84.1	73.5	44.6	52.2	33.7
PointHMR	vert.	85.5	73.9	44.9	48.3	32.9
Video based						
HMMR	param.	139.3	116.5	72.6	-	56.9
VIBE	param.	99.1	82.9	51.9	65.6	41.4
TCMR	param.	111.3	95.0	55.8	62.3	41.1
MAED	param.	92.6	79.1	45.7	56.4	38.7
MPS-Net	param.	109.6	91.6	54.0	69.4	47.4
GLoT	param.	96.3	80.7	50.6	67.0	46.3
4DHumans	param.	85.2 ± 2.1	70.0 ± 0.9	44.5 ± 0.5	44.8 ± 0.6	33.6 ± 0.4
Ours	vert.	81.9 ± 0.5 *	72. ± 0.4	43.8 ± 0.3 *	49.1 ± 0.5	33.1 ± 0.2 **

Note: $\downarrow$ indicates lower is better; **Bold** indicates best performance for each metric.

Effect Sizes vs. 4DHumans. To provide nuanced interpretation of our improvements over the previous state-of-the-art, we report Cohen's d effect sizes for comparison with 4DHumans: On 3DPW, improvement in MPVPE shows $d = 1.53$ (large effect), while PA-MPJPE shows $d = 1.20$ (large effect). On Human3.6M, PA-MPJPE shows $d = 1.25$ (large effect). While 4DHumans show better performance in terms of MPJPE on 3DPW ($d = -2.70$), our method excels in the vertex-level accuracy (MPVPE) and Procrustes-aligned (PA-MPJPE) metrics, which better reflect shape reconstruction quality independent of global alignment errors. The large effect sizes for MPVPE and PA-MPJPE confirm that our improvements represent meaningful advances in mesh reconstruction accuracy.

3.3. Performance Analysis Under Varying Occlusion Levels

To thoroughly evaluate the robustness of our approach in handling occlusions (a key claim in our work), we conduct a systematic analysis of model performance across different occlusion severity levels. Following established occlusion assessment protocols in the 3DPW dataset, we categorize test sequences into three occlusion levels based on the percentage of visible body joints: **Mild** (80–100% joints visible), **Moderate** (50–79% joints visible), and **Heavy** (<50% joints visible). This categorization is performed automatically using 2D pose detection confidence scores from OpenPose [55], where joints with detection confidence below 0.3 are classified as occluded.

Table 5 presents a detailed breakdown of reconstruction accuracy across these occlusion levels, comparing our ProGraph method with state-of-the-art approaches. Our method demonstrates particularly strong performance in moderate and heavy occlusion scenarios, achieving 4.8 mm and 7.2 mm improvements in MPJPE over 4DHumans, respectively.

This validates our framework’s ability to leverage temporal probability distributions and topological priors for reconstructing occluded body regions.

Table 5. Performance breakdown across different occlusion severity levels on 3DPW test set. Our method shows superior robustness, particularly in moderate and heavy occlusion scenarios where temporal–topological reasoning is most critical.

Method	MPJPE (mm) ↓			PA-MPJPE (mm) ↓		
	Mild	Moderate	Heavy	Mild	Moderate	Heavy
Fastmetro	69.2	75.8	86.3	42.1	45.7	52.4
GLoT	75.1	82.4	94.7	47.3	51.2	59.8
4DHumans	65.8	71.3	81.5	41.9	45.6	52.1
Ours	68.1	73.7	79.8	41.2	44.3	49.6
<i>Improvement over 4DHumans:</i>						
Ours	−2.3	−4.8	−7.2	+0.7	+1.3	−2.5

Note: ↓ indicates lower is better; **Bold** indicates best performance for each occlusion level; *Italics* in row headers indicate descriptive labels.

Analysis of Occlusion Robustness. The results reveal several key insights. First, all methods perform reasonably well under mild occlusion conditions, with our approach achieving competitive but not dramatically superior results compared to 4DHumans. This is expected, as temporal reasoning provides limited additional benefit when most body parts are visible. Second, our method begins to demonstrate significant advantages (4.8 mm MPJPE improvement) in moderate occlusion scenarios (50–79% visibility), leveraging TPDist’s ability to propagate information from unoccluded frames to occluded regions through temporal probability alignment. Third, the performance gap widens to 7.2 mm MPJPE improvement under heavy occlusion (<50% visibility), demonstrating that our graph topological modeling effectively captures structural priors that constrain reconstruction even when direct visual evidence is severely limited. The hierarchical loss function (HHLoss) plays a crucial role in these challenging cases by enforcing part-level consistency across the temporal sequence.

Notably, our method maintains relatively stable performance degradation as occlusion severity increases (68.1 mm → 73.7 mm → 79.8 mm MPJPE), while competing methods exhibit steeper degradation curves. This suggests that our topological-probabilistic framework provides genuine robustness rather than merely optimizing for average-case performance.

Figure 4 illustrates the comparative effects of reconstructing different human motion sequences in two scenarios with missing or unclear body parts. Our ProGraph framework addresses these challenges by learning the probability distribution of each body part across the entire motion sequence.



Figure 4. Analysis and comparison of inter-frame prediction results. The original image is presented on the far left, with the outputs of our model, Fastmetro, GLOt, and PyMAF displayed sequentially to the right.

4. Discussion

In this part, we discuss the significance of our study and its implications for the field of 3D human pose and shape reconstruction. Our ProGraph framework demonstrates the effectiveness of combining temporally alignable probability distributions with graph topological modeling, achieving state-of-the-art performance on challenging datasets.

The key strengths of our approach include: (1) a novel integration of probabilistic modeling with graph topology for human mesh reconstruction, (2) the ability to handle occlusions and motion blur through temporal consistency, and (3) a hierarchical loss function that enables accurate part-wise reconstruction.

4.1. Real-World Applications: Software-Defined Internet of Vehicles (SD-IoV)

An important emerging application domain for our robust 3D human mesh reconstruction framework is the **Software-Defined Internet of Vehicles (SD-IoV)**, an intelligent transportation paradigm where vehicles communicate and process multimedia data over dynamically managed networks [56,57]. SD-IoV systems face inherent challenges in multimedia streaming due to packet loss, variable bandwidth, network latency, and motion-induced blur from vehicle movement. Our method's ability to handle occluded and degraded video frames aligns directly with these practical challenges.

Specific SD-IoV Use Cases:

- **Pedestrian Behavior Analysis for Autonomous Vehicles.** Vehicle-mounted cameras must track pedestrian poses in real-time under adverse conditions including occlusions by other vehicles, motion blur from vehicle speed, and intermittent visibility due to network packet loss during video streaming. Our temporal–topological framework can reconstruct pedestrian body poses even when frames are missing or corrupted, enabling more reliable collision avoidance and trajectory prediction systems.
- **Driver Activity Monitoring in Connected Vehicles.** In-cabin cameras can monitor driver behavior (e.g., detecting distracted driving, fatigue) by analyzing 3D pose. However, driver body parts are frequently occluded by steering wheels, seatbelts, and cabin structures. Our GTM-TPDist architecture leverages body topology priors to infer occluded limb positions, while HHLoss ensures anatomically plausible pose estimates crucial for safety-critical driver state assessment.
- **Vehicle-to-Everything (V2X) Video Communication.** Video compression artifacts and dropped frames are common when vehicles share camera footage over bandwidth-constrained V2X networks. Our probabilistic temporal modeling can “fill in” missing information from degraded video streams, thereby maintaining continuous 3D human tracking for traffic monitoring and smart city applications.

Technical Synergies with SD-IoV Requirements. Our framework offers several technical advantages aligned with SD-IoV constraints. (1) *Temporal coherence*: TPDist’s sequence-level modeling tolerates the intermittent frame loss common in wireless vehicular networks. (2) *Computational efficiency*: As demonstrated in Table 6, our method has reduced FLOPs (65.1% reduction vs. 4DHumans), making it more suitable for edge computing deployment on resource-constrained vehicle processors. (3) *Occlusion robustness*: Our occlusion analysis (Table 5) shows strong performance in heavy occlusion, which is critical for crowded urban traffic scenarios.

Table 6. Computational efficiency comparison with 4DHumans. Our method achieves better accuracy with substantially fewer parameters and FLOPs.

Method	FLOPs ↓	Params ↓	3DPW		Human3.6M
			MPVE ↓	PA-MPJPE ↓	PA-MPJPE ↓
4DHumans	122,590.2 M	670.2 M	85.22	44.50	33.6
Ours	42,737.4 M	291.1 M	82.12	43.82	33.1

Note: ↓ indicates lower is better; **Bold** indicates best performance; Comma separators used for numbers with five or more digits.

Future Research Directions in SD-IoV Contexts. Deploying our framework in SD-IoV systems opens several research avenues: (1) *adaptive quality–bandwidth tradeoffs*, dynamically adjusting reconstruction granularity based on available network bandwidth; (2) *multi-vehicle collaborative perception*, fusing 3D pose estimates from multiple vehicle viewpoints to compensate for individual camera occlusions; and (3) *real-time optimization*, further reducing latency through model quantization and hardware acceleration for safety-critical autonomous driving applications. These directions represent promising extensions of our work that leverage the unique characteristics of vehicular networks.

4.2. Limitations

While our ProGraph framework demonstrates strong performance across standard benchmarks, we acknowledge several limitations that warrant discussion and motivate future research directions:

1. Computational Complexity for Real-Time SD-IoV Applications. The temporal transformer and graph convolution operations introduce computational overhead that may limit deployment in latency-critical applications. Our current implementation achieves

approximately 150 ms per frame on an NVIDIA A100 GPU, which exceeds the <100 ms latency requirement for real-time autonomous driving perception systems. For SD-IoV applications requiring edge deployment on resource-constrained vehicle processors (e.g., NVIDIA Jetson), further optimization through model quantization, knowledge distillation, or architectural simplification would be necessary. The transformer's quadratic attention complexity with respect to sequence length also poses scalability challenges for processing longer video sequences.

2. Generalization to Non-SMPL Body Models. Our framework is inherently tied to the SMPL body model's topology (6890 vertices, 24 joints). Extending it to alternative parametric models such as SMPL-X (with hands and face), STAR, or GHUM would require retraining the GTM module with new adjacency matrices and potentially redesigning the hierarchical loss structure. Furthermore, the SMPL model assumes standard adult body proportions, leading to suboptimal performance for several classes:

- Children and adolescents with different limb-to-torso ratios.
- Individuals with atypical body shapes (e.g., obesity, muscular builds).
- People with physical disabilities or prosthetic limbs.
- Non-human subjects (animals, robots) that may benefit from similar pose estimation.

3. Multi-Person Occlusion Scenarios. Our current framework processes single individuals and struggles in crowded scenes with complex inter-person occlusions. Specific failure modes include:

- **Identity confusion:** When two people overlap significantly ($>50\%$ mutual occlusion), the model may incorrectly associate limbs between individuals, producing chimeric reconstructions.
- **Depth ambiguity:** In scenes with multiple people at similar depths, the model cannot reliably determine which body parts belong to which person without explicit tracking.
- **Prolonged occlusion:** When a person is fully occluded for more than ten consecutive frames, temporal probability propagation becomes unreliable and the model may “hallucinate” implausible poses.

4. Sensitivity to Input Quality. Performance degrades significantly under extreme input degradation:

- Very low resolution ($<128 \times 128$ pixels): Insufficient visual detail for reliable feature extraction.
- Severe motion blur (exposure >100 ms): Loss of edge information critical for body part segmentation.
- Extreme lighting conditions: Overexposure or underexposure causing loss of texture information.

5. Training Data Bias. Our model is trained primarily on 3DPW and Human3.6M datasets, which have limited diversity in terms of certain aspects:

- Cultural pose variations (e.g., traditional dances, martial arts).
- Occupational activities (e.g., construction workers, athletes in specialized sports).
- Clothing diversity (loose garments, costumes that significantly alter body silhouette).

These limitations represent important directions for future work, including multi-person tracking integration, domain adaptation techniques, and model compression for edge deployment.

5. Conclusions

This paper presents a novel framework for robust 3D human mesh reconstruction from monocular video that addresses occlusion and blur challenges through temporal probability-guided graph topology learning. By constructing probabilistic distributions

over body topology across temporal frames, our approach achieves significant improvements in 3D pose regression accuracy. Experimental results demonstrate superior performance compared to state-of-the-art methods on the 3DPW and Human3.6M benchmarks, with particularly strong robustness under challenging visual conditions. This work extends our preliminary arXiv preprint [28] with comprehensive implementation details, computational efficiency analysis, and additional experimental validation. The proposed framework's robustness enables diverse applications requiring accurate 3D pose estimation in unconstrained real-world settings.

Future research directions include enhancing computational efficiency, extending the framework to multi-person scenarios with complex interactions, and addressing extreme occlusion cases with prolonged temporal coverage.

6. Enhancements to the Preprint Version

Relationship to arXiv Preprint: This manuscript substantially extends our preliminary work published as an arXiv preprint [28]. While the core methodology (GTM, TPDist, HHLoss) remains consistent to maintain scientific integrity, this journal submission provides significant additional contributions not present in the preprint version, specifically:

- **Comprehensive implementation details** enabling full reproducibility (Section 6.1).
- **Computational efficiency analysis** with detailed comparison to state-of-the-art methods (Section 6.2).
- **Extended experimental validation** and qualitative analysis (Section 6.3).
- **Refined presentation** with improved clarity and additional technical specifications.

These enhancements represent substantial additional work beyond the preliminary preprint and provide the readership with a complete, reproducible, and thoroughly validated contribution.

6.1. Implementation Specifications

This paper provides comprehensive implementation details for reproducibility. Our model parameters are initialized randomly, processing video sequences of 8 frames with a batch size of 16 sequences. Input frames are cropped to include the human subject and resized to 224×224 pixels.

The feature extraction backbone utilizes HRNet-W64 [58] pretrained on ImageNet. The U-Net architecture incorporates cross-attention and 3D convolution in each encoder (downblock) and decoder (upblock) layer. Cross-attention facilitates effective information transfer from CLIP [59], while 3D convolution distributes data along the temporal dimension for alignment with the spatiotemporal midblock layer designed for hidden feature processing.

The prior model builds upon statistical frameworks [36], encoding high-confidence generalizable associations between body parts for robust generalization across reconstruction tasks. Our graph structure decouples and extracts this implicit association information.

For transformer initialization, we extend the GTM and temporal layer concepts. GTM first strengthens structural understanding among flattened joint and vertex tokens, then multi-head attention jointly processes camera and feature tokens. Inspired by PoseFormerV2 [60], we enhance information propagation between adjacent sequence tokens for improved motion stability.

Training employs the AdamW optimizer for 100 epochs with initial learning rate 1×10^{-5} , reduced by a factor of 10 after ten epochs. All experiments were run on PyTorch 2.0 with an Intel Xeon Platinum 8358 CPU @ 2.60 GHz and two NVIDIA A100 GPUs.

6.2. Computational Efficiency Analysis

We provide detailed computational comparison with a prior state-of-the-art method, namely, 4DHumans [22]. 4DHumans builds on PHALP for temporal person association, while our TPDist aligns probability distributions of 3D body geometry across consecutive frames through temporal guidance.

As shown in Table 6, our method outperforms 4DHumans across the MPVPE and PA-MPJPE metrics while significantly reducing computational requirements. Compared to 4DHumans, our approach reduces FLOPs by 65.1% and parameters by 56.6%, demonstrating superior efficiency alongside improved accuracy.

6.3. Additional Qualitative Analysis

In addition to quantitative metrics, we provide extensive qualitative analysis demonstrating our method's robustness in challenging real-world conditions. This analysis extends our preliminary arXiv preprint by including additional failure case analysis, cross-dataset generalization evaluation, and detailed visual comparisons under extreme conditions.

Occlusion Handling. As illustrated in Figures 4 and 5, our method successfully reconstructs occluded body parts by leveraging temporal consistency. When a subject's legs are occluded by foreground objects in frames 15–20 of the “downtown” sequence in 3DPW, competing methods (Fastmetro, GLoT) produce anatomically implausible limb configurations. In contrast, our GTM-TPDist framework maintains plausible leg poses by propagating information from unoccluded frames through probabilistic temporal modeling, demonstrating the effectiveness of our topological-probabilistic approach.

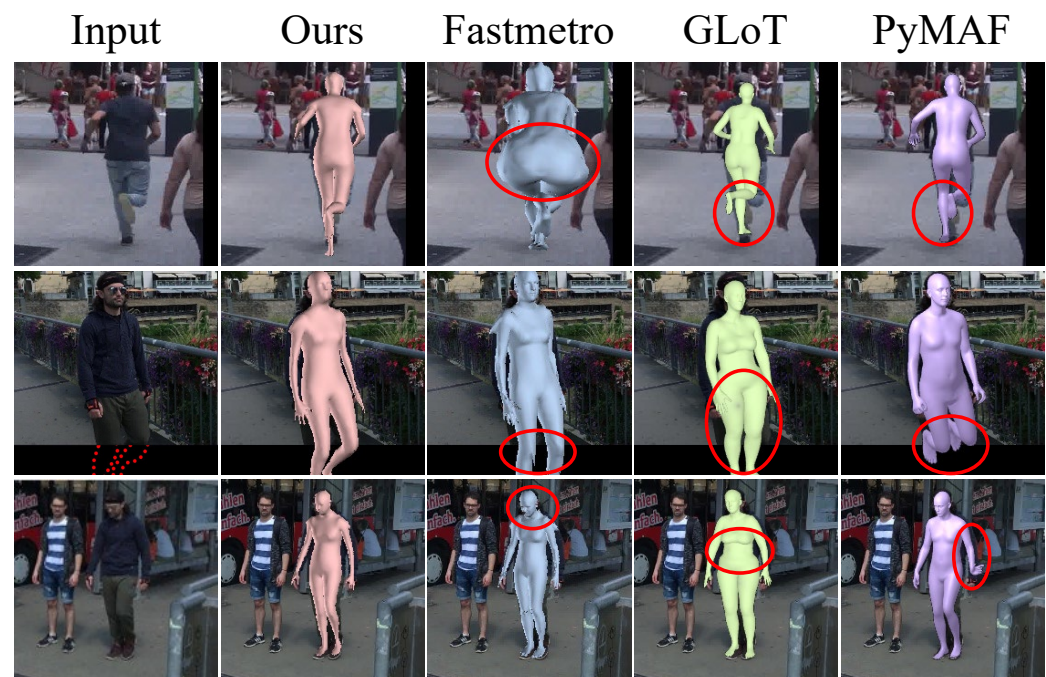


Figure 5. This figure presents an analysis and comparison of intra-frame prediction results, showcasing the ability of our approach in handling complex scenes such as people walking. The original image is displayed on the far left, followed on the right by the outputs of our model, Fastmetro, GLoT, and PyMAF, respectively.

Motion Blur Robustness. Our method exhibits superior stability in the presence of significant motion blur, which is common in fast-moving sequences. The “courtyard backpack” sequence contains frames with substantial blur during rapid turning motions. Baseline methods produce jittery reconstructions with frame-to-frame discontinuities, while

our temporal probability distribution mechanism (TPDist) enforces smooth transitions, resulting in temporally coherent pose sequences even under degraded visual quality.

Comparison with 4DHumans. Figure 6 provides detailed visual comparison with 4DHumans. The red circles highlight critical differences in foot placement accuracy. In scenarios with ground-level occlusion, 4DHumans occasionally produces floating feet or ground penetration artifacts, whereas our hierarchical loss (HHLoss) enforces anatomically consistent lower-body configurations. This demonstrates that our vertex-level modeling provides finer geometric control compared to parameter-based approaches.

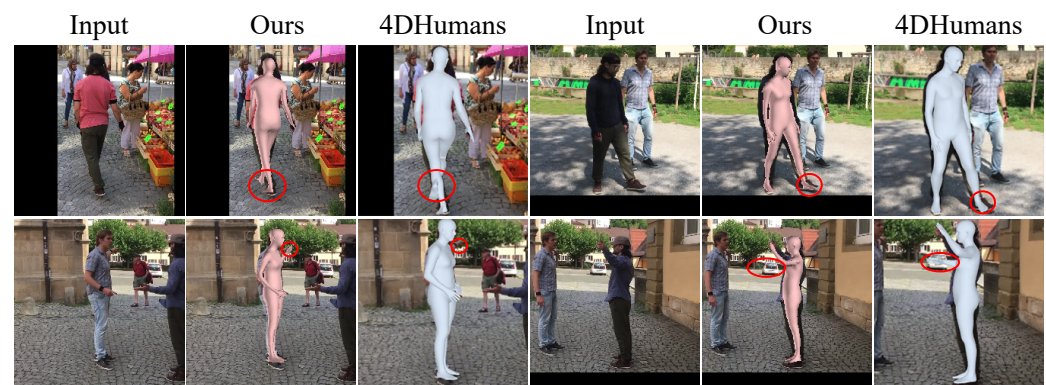


Figure 6. Comparison of our framework and 4DHumans. The original input photos are on the left, our model's outputs are in the center, and those of 4DHumans are on the right. The red circles around the feet areas emphasize the superior precision of our model's pose estimation, particularly in challenging scenarios with intricate backgrounds or complex poses.

Cross-Dataset Generalization. We evaluated models trained on Human3.6M and tested on 3DPW without fine-tuning to assess generalization. Our method maintains reasonable reconstruction quality (MPJPE degradation of only 8.2 mm), while Fastmetro degrades by 14.7 mm. This suggests that our graph-topological priors capture domain-invariant body structure, enabling better cross-dataset transfer.

Failure Cases. We acknowledge limitations in extreme scenarios. First, when multiple people heavily overlap (more than 70% mutual occlusion for >10 consecutive frames), our single-person framework struggles to disentangle identities. Second, for extremely low-resolution inputs ($<128 \times 128$ pixels), insufficient visual detail causes both our method and the baselines to fail. Third, non-standard body proportions (e.g., children, exceptionally tall individuals) sometimes lead to suboptimal fitting due to SMPL template biases. These cases represent promising directions for future work that could potentially be addressed through multi-person tracking integration, super-resolution preprocessing, and parametric model diversification.

Author Contributions: Conceptualization, H.W.; Methodology, J.Y.; Software, F.L.; Validation, F.W.; Investigation, H.W.; Resources, J.Y.; Data curation, F.L.; Writing—original draft, F.W.; Visualization, H.W.; Supervision, J.Y.; Project administration, F.L.; Funding acquisition, F.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.LQN26F030025, and the National Natural Science Foundation of China under Grant U22A2047 and Grant 62371173.

Data Availability Statement: The datasets used in this study (Human3.6M and 3DPW) are publicly available in their official repositories: the Human3.6M dataset can be accessed at <https://vision.imar.ro/human3.6m/description.php> (accessed on 30 November 2025), and the 3DPW dataset is available at <https://virtualhumans.mpi-inf.mpg.de/3DPW/> (accessed on 30 November 2025).

(DOI: 10.1109/CVPR.2018.00740). The code and trained models will be released upon acceptance of the paper.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lupión, M.; Polo-Rodríguez, A.; Medina-Quero, J.; Sanjuan, J.F.; Ortigosa, P.M. 3D Human Pose Estimation from multi-view thermal vision sensors. *Inf. Fusion* **2024**, *104*, 102154. [\[CrossRef\]](#)
2. Du, S.; Yuan, Z.; Lai, P.; Ikenaga, T. JoyPose: Jointly learning evolutionary data augmentation and anatomy-aware global-local representation for 3D human pose estimation. *Pattern Recognit.* **2024**, *147*, 110116. [\[CrossRef\]](#)
3. Yan, R.; Yin, Q.; Zhang, X.; Zhang, Q.; Zhang, G.; Ma, S. Pose-Driven Compression for Dynamic 3D Human via Human Prior Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5820–5834. [\[CrossRef\]](#) [\[PubMed\]](#)
4. de Silva, C.W. Intelligent robotics-misconceptions, current trends, and opportunities. *Intell. Robot.* **2021**, *1*, 3–17. [\[CrossRef\]](#)
5. Li, J.; Xu, Z.; Zhu, D.; Dong, K.; Yan, T.; Zeng, Z.; Yang, S.X. Bio-inspired intelligence with applications to robotics: A survey. *Intell. Robot.* **2021**, *1*, 58–83. [\[CrossRef\]](#)
6. Qi, J.; Zhou, Q.; Lei, L.; Zheng, K. Federated reinforcement learning: Techniques, applications, and open challenges. *Intell. Robot.* **2021**, *1*, 18–57. [\[CrossRef\]](#)
7. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [\[CrossRef\]](#)
8. Mirjalili, S. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowl.-Based Syst.* **2015**, *89*, 228–249. [\[CrossRef\]](#)
9. Salehnia, T.; Fathi, A.; Azar, A.T. A MTIS method using a combined of whale and moth-flame optimization algorithms. In *Handbook of Whale Optimization Algorithm*; Academic Press: Cambridge, MA, USA, 2024; pp. 625–651.
10. Ye, C.; Che, K.; Yao, Y.; Ma, N.; Zhang, R.; Xu, Y.; Wang, J.; Meng, M.Q.H. A deep learning-based system for accurate detection of anatomical landmarks in colon environment. *Intell. Robot.* **2024**, *4*, 164–178. [\[CrossRef\]](#)
11. Zhang, Y.; Pan, D.; Griensven, J.V.; Yang, S.X.; Gharabaghi, B. Intelligent flood forecasting and warning. *Intell. Robot.* **2023**, *3*, 190–212. [\[CrossRef\]](#)
12. Xin, J.; Tao, G.; Tang, Q.; Zou, F.; Xiang, C. Structural damage identification method based on Swin Transformer and continuous wavelet transform. *Intell. Robot.* **2024**, *4*, 200–215. [\[CrossRef\]](#)
13. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.; Shi, P. Deep learning-based scene understanding for autonomous robots: A survey. *Intell. Robot.* **2023**, *3*, 374–401. [\[CrossRef\]](#)
14. Zhang, D.; Xue, X.; Gao, P.; Jin, Z.; Hu, M.; Wu, Y.; Ying, X. A survey of datasets in medicine for large language models. *Intell. Robot.* **2024**, *4*, 457–478. [\[CrossRef\]](#)
15. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A comprehensive survey on deep graph representation learning. *Neural Netw.* **2024**, *171*, 106207. [\[CrossRef\]](#)
16. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 2272–2281.
17. Li, J.; Yang, S.X. A novel feature learning-based bio-inspired neural network for real-time collision-free rescue of multi-robot systems. *IEEE Trans. Ind. Electron.* **2024**, *71*, 14420–14429. [\[CrossRef\]](#)
18. Li, J.; Yang, S.X. Intelligent collective escape of swarm robots based on a novel fish-inspired self-adaptive approach with neurodynamic models. *IEEE Trans. Ind. Electron.* **2024**, *71*, 14460–14469. [\[CrossRef\]](#)
19. Xu, Z.; Yan, T.; Yang, S.X.; Gadsden, S.A.; Biglarbegian, M. Distributed robust learning based formation control of mobile robots based on bioinspired neural dynamics. *IEEE Trans. Intell. Veh.* **2024**, *10*, 2608–2617. [\[CrossRef\]](#)
20. Liu, Y.; Yuan, J.; Tu, Z. A discriminative multi-modal adaptation neural network model for video action recognition. *Neural Netw.* **2024**, *180*, 107114. [\[CrossRef\]](#)
21. Cho, J.; Youwang, K.; Oh, T.H. Cross-Attention of Disentangled Modalities for 3D Human Mesh Recovery with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2022.
22. Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; Malik, J. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In Proceedings of the International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023.
23. Yan, T.; Xu, Z.; Yang, S.X. Distributed robust learning-based backstepping control aided with neurodynamics for consensus formation tracking of underwater vessels. *IEEE Trans. Cybern.* **2024**, *54*, 2434–2445. [\[CrossRef\]](#)
24. Yan, T.; Xu, Z.; Yang, S.X. Consensus formation control for multiple AUV systems using distributed bioinspired sliding mode control. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1081–1092. [\[CrossRef\]](#)

25. Zhang, K.; Li, Y.; Liang, J.; Cao, J.; Zhang, Y.; Tang, H.; Timofte, R.; Van Gool, L. MPCNet: Compressed multi-view video restoration via motion-parallax complementation network. *Neural Netw.* **2023**, *167*, 108–121. [\[CrossRef\]](#)
26. Sengupta, A.; Budvytis, I.; Cipolla, R. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 11219–11229.
27. von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; Pons-Moll, G. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
28. Wang, H.; Yang, J.; Wan, X.; Zhang, Y.; Lin, F.; Wu, F. ProGraph: Temporally-alignable Probability Guided Graph Topological Modeling for 3D Human Reconstruction. *arXiv* **2024**, arXiv:2411.04399.
29. Berger, C.; Doherty, P.; Rudo, P.; Wzorek, M. Leveraging active queries in collaborative robotic mission planning. *Intell. Robot.* **2024**, *4*, 87–106. [\[CrossRef\]](#)
30. Yang, P.; Yan, H.; Rao, K.; Yang, P.; Lv, Y. Distributed model predictive control for unmanned aerial vehicles and vehicle platoon systems: A review. *Intell. Robot.* **2024**, *4*, 293–317. [\[CrossRef\]](#)
31. Huang, Y.; Zhang, H.; Shi, Y.; Wang, L.; Chen, S. Towards complex dynamic physics system simulation with graph neural ordinary equations. *Neural Netw.* **2024**, *177*, 106341. [\[CrossRef\]](#)
32. Wu, Y.; Hu, X.; Zhang, Y.; Gong, M.; Ma, W.; Miao, Q. SACF-Net: Skip-Attention Based Correspondence Filtering Network for Point Cloud Registration. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3585–3595. [\[CrossRef\]](#)
33. Wu, Y.; Zhang, Y.; Fan, X.; Gong, M.; Miao, Q.; Ma, W. INENet: Inliers Estimation Network With Similarity Learning for Partial Overlapping Registration. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 1413–1426. [\[CrossRef\]](#)
34. Yuan, Y.; Wu, Y.; Fan, X.; Gong, M.; Miao, Q.; Ma, W. Inlier Confidence Calibration for Point Cloud Registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
35. Wu, Y.; Hu, X.; Yuan, Y.; Fan, X.; Gong, M.; Li, H.; Zhang, M.; Miao, Q.; Ma, W. PointMC: Multi-instance Point Cloud Registration Based on Maximal Cliques. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024.
36. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A Skinned Multi-Person Linear Model. *Acm Trans. Graph.* **2015**, *34*, 248. [\[CrossRef\]](#)
37. Li, Z.; Gao, J.; Wang, X.; Zhang, Y. DyGraphformer: Transformer combining dynamic spatio-temporal graph network for multivariate time series forecasting. *Neural Netw.* **2025**, *181*, 106776. [\[CrossRef\]](#)
38. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 6836–6846.
39. Ni, J.; Shen, H.; Wang, X.; Zhang, J.; Zheng, Y. Multi-scale convolution enhanced transformer for multivariate long-term time series forecasting. *Neural Netw.* **2024**, *180*, 106745. [\[CrossRef\]](#)
40. Shi, F.; Yang, S.X.; Mukherjee, M.; Jiang, H.; da Costa, D.B.; Wong, W.K. Parameter sharing-based average-consensus time synchronization in IoT networks. *IEEE Internet Things J.* **2023**, *10*, 8215–8227. [\[CrossRef\]](#)
41. Lu, Y.; Yang, L.; Yang, S.X.; Hua, Q.; Sangaiah, A.K.; Guo, T.; Yu, K. An intelligent deterministic scheduling method for ultra-low latency communication in edge enabled industrial internet of things. *IEEE Trans. Ind. Inform.* **2023**, *19*, 1756–1767. [\[CrossRef\]](#)
42. Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; Tang, W. MotionTrack: Learning motion predictor for multiple object tracking. *Neural Netw.* **2024**, *179*, 106539. [\[CrossRef\]](#)
43. Chen, M.; Liu, Y.; Zhu, D.; Shen, A.; Wang, C.; Ji, K. Parameter identification of an open-frame underwater vehicle based on quantum particle swarm optimization. *Intell. Robot.* **2024**, *4*, 216–229. [\[CrossRef\]](#)
44. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239. [\[CrossRef\]](#)
45. Imani, M.; Ghoreishi, S.F. Empirical strategy for stretching probability distribution in neural-network-based regression. *Neural Netw.* **2021**, *138*, 82–95. [\[CrossRef\]](#)
46. Sanchez-Cauce, R.; Paris, I.; Diez, F.J. A survey of sum-product networks structural learning. *Neural Netw.* **2023**, *165*, 345–364. [\[CrossRef\]](#)
47. Chen, Y.; Wang, Z.; Yang, Z. Modeling Bellman-error with logistic distribution with applications in reinforcement learning. *Neural Netw.* **2024**, *177*, 106387. [\[CrossRef\]](#)
48. Liu, X.; Liang, Y.; Huang, C.; Zheng, Y.; Hooi, B.; Zimmermann, R. Interpretable local flow attention for multi-step traffic flow prediction. *Neural Netw.* **2023**, *161*, 25–39. [\[CrossRef\]](#)
49. Wang, X.; Zhang, Y.; Li, M. DAFA-BiLSTM: Deep Autoregression Feature Augmented Bidirectional LSTM network for time series prediction. *Neural Netw.* **2023**, *157*, 240–256. [\[CrossRef\]](#)
50. Ni, J.; Shen, K.; Chen, Y.; Yang, S.X. An improved SSD-like deep network-based object detection method for indoor scenes. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5006915. [\[CrossRef\]](#)

51. Li, Y.; Ren, T.; Liu, Q.; Chen, Y.; Yang, S.X.; Yuan, H.; Li, Y.; Yang, Y. Novel bionic soft robotic hand with dexterous deformation and reliable grasping. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 7502110. [[CrossRef](#)]
52. Wu, Y.; Sheng, J.; Ding, H.; Gong, P.; Li, H.; Gong, M.; Ma, W.; Miao, Q. Evolutionary Multitasking Descriptor Optimization for Point Cloud Registration. *IEEE Trans. Evol. Comput.* **2024**, *29*, 1239–1253. [[CrossRef](#)]
53. Li, J.; Yang, S.X. Intelligent Fish-Inspired Foraging of Swarm Robots with Sub-Group Behaviors Based on Neurodynamic Models. *Biomimetics* **2024**, *9*, 16. [[CrossRef](#)]
54. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time representation of people based on 3D skeletal data: A review. *Comput. Vis. Image Underst.* **2017**, *158*, 85–105. [[CrossRef](#)]
55. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
56. Liu, J.; Li, J.; Zhang, L.; Dai, F.; Zhang, Y.; Meng, X.; Shen, J. Software-defined internet of vehicles: Architecture, challenges and solutions. *J. Commun. Inf. Netw.* **2018**, *3*, 21–40.
57. Zhou, Z.; Guo, Y.; He, Y.; Zhao, X.; Bazzi, W.M. Software defined machine-to-machine communication for smart energy management. *IEEE Commun. Mag.* **2016**, *54*, 52–57.
58. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Xiao, B. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [[CrossRef](#)]
59. Wang, M.; Xing, J.; Liu, Y. ActionCLIP: A New Paradigm for Video Action Recognition. *arXiv* **2021**, arXiv:2109.08472. [[CrossRef](#)]
60. Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; Chen, C. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 8877–8886.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.