ADAPTIVE MOTION-AWARE BODY-SURFACE EVIDENCE REPRESENTATION FOR LIDAR-BASED HUMAN MOTION CAPTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

LiDAR-based human motion capture aims to recover temporally coherent 3D human poses from sparse, occluded, and noisy point-cloud observations. Although complementary cross-frame observations can alleviate single-frame incompleteness, existing approaches typically organize temporal evidence through estimated skeletons, making cross-frame association vulnerable to skeletal estimation uncertainty. This limitation becomes particularly severe under sparse measurements and self-occlusion, as unreliable skeletal estimates propagate errors during temporal reasoning. In this paper, we propose a new framework called Adaptive Motion-aware Body-surface Evidence Representation (AMBER) to decouple temporal evidence organization from skeletal reasoning. Specifically, AMBER first constructs a faithful representation by organizing multi-scale cross-frame evidence around adaptive surface anchors. The resulting representation is further incorporated into a Geometry-Motion Refinement (GMR) module to refine the corresponding joint positions and motion dynamics in the skeletal domain. In this way, rather than prematurely compressing evidence into sparse joints, AMBER retains rich geometric details and motion cues on the body surface to ensure robust cross-frame association. Extensive experiments on multiple LiDAR benchmarks demonstrate that AMBER consistently outperforms SOTA methods.

1 INTRODUCTION

Human motion capture (MoCap) aims to recover temporally coherent 3D human pose and motion from sensor observations, and plays an important role in applications such as virtual and augmented reality, sports analysis, rehabilitation, and human–robot interaction (Vlasic et al., 2007; Liu et al., 2013; Rhodin et al., 2016). Among different sensing modalities, LiDAR is particularly attractive for large-scale and unconstrained environments because it provides metric 3D measurements, is insensitive to illumination changes, and naturally supports global motion estimation (Li et al., 2022; Dai et al., 2023; Ren et al., 2024b; 2023). However, human LiDAR observations are inherently sparse and incomplete. Limited sampling density, self-occlusion, clothing deformation, and measurement noise leave many body regions only partially observed or entirely missing, making reliable full-body motion recovery from partial surface measurements highly challenging.

Temporal observations provide complementary evidence for resolving ambiguity. Existing methods exploit temporal context through frame-level feature aggregation (Li et al., 2022; Jang et al., 2023), joint-level spatio-temporal interaction and motion modeling (Ren et al., 2024b;a; An et al., 2026). Recent work incorporates surface representations to provide complementary geometric cues beyond skeletal features (Ren et al., 2026c). Despite these advances, representative structured approaches still follow a skeleton-conditioned reasoning paradigm, where estimated skeletal states guide the organization of temporal observations. As illustrated in Fig. 1(a), this dependency couples temporal evidence organization with skeletal inference, making temporal reasoning dependent on the skeletal state and potentially sensitive to its estimation errors.

This observation motivates us to rethink the reasoning order among LiDAR observations, human surfaces, and skeletons. LiDAR directly observes points on the external body surface, allowing surface anchors to provide observation-grounded local references before skeletal reasoning. Based on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

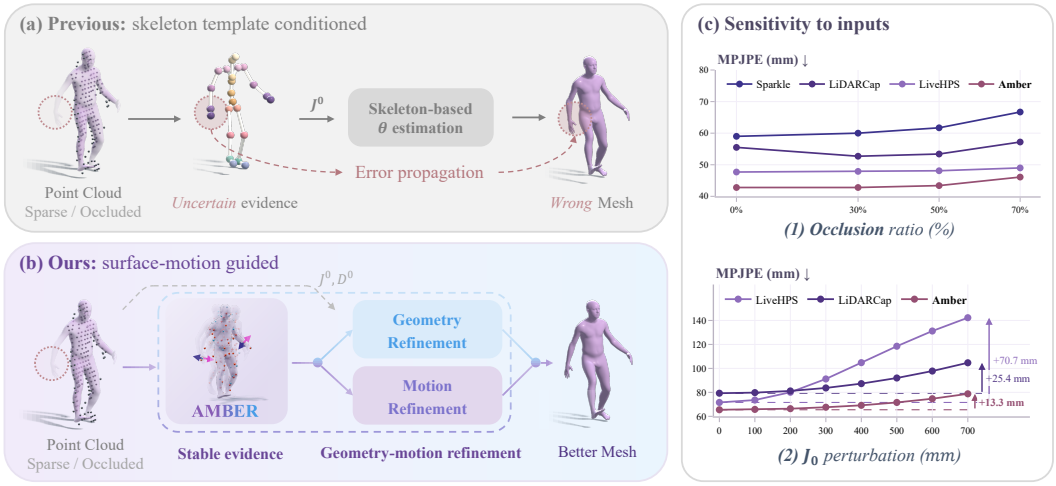

Figure 1: (a) Previous skeleton-conditioned methods use joints as references for evidence association, coupling retrieval with skeletal inference. (b) AMBER reverses the reasoning order by organizing observation-grounded evidence before refining the skeleton through geometry–motion reasoning. (c) Sensitivity to input degradation. The top plot evaluates occlusion ratios, while the bottom perturbs the estimated skeleton with increasing magnitude. AMBER's smaller MPJPE degradation indicates reduced sensitivity to incomplete observations and skeletal-state perturbations.

this insight, we propose **AMBER** (**A**daptive **M**otion-aware **B**ody-surface **E**vidence **R**epresentation), which realizes this reasoning order through two stages, as illustrated in Fig. 1(b). First, **AMBER Construction** adaptively initializes a sparse set of surface anchors from the current LiDAR observation and uses them as local references to aggregate multi-scale geometric observations and directional motion cues from neighboring frames without relying on intermediate skeletal estimates. Second, **Geometry–Motion Refinement** transfers the organized surface evidence to the skeletal domain through separate geometry and motion branches, jointly refining joint locations and temporal dynamics for final SMPL pose recovery.

Extensive experiments on four LiDAR-based human motion capture benchmarks demonstrate that AMBER consistently improves over existing approaches. Further analyses validate the benefit of surface-space evidence organization. The sensitivity results in Fig. 1(c) further show that AMBER degrades more gracefully under observation degradation and skeletal-state perturbations. Our contributions are threefold:

- We propose **AMBER**, a new reasoning paradigm for LiDAR-based human motion capture that decouples cross-frame evidence organization from skeletal inference through an observation-grounded surface representation.
- We introduce an adaptive surface evidence representation that organizes multi-scale geometric and motion cues through sparse surface anchors, together with a Geometry–Motion Refinement module for accurate skeletal recovery.
- We achieve state-of-the-art performance across four challenging benchmarks and conduct extensive controlled studies demonstrating the effectiveness and robustness of AMBER under challenging input conditions.

2 Related Work

Skeleton-centric LiDAR Human Motion Capture Early LiDAR-based motion capture methods exploit temporal context at sequence level to compensate for incomplete observations. LiDARCap (Li et al., 2022) aggregates encoded point-cloud features for skeletal regression, while MOVIN (Jang et al., 2023) incorporates historical observations with pose states into autoregressive motion modeling. These approaches capture temporal continuity through global sequence representations, but do not explicitly establish correspondence between local observations and body regions.

Recent methods progressively shift temporal reasoning toward structured skeletal variables. LiveHPS (Ren et al., 2024b) models joint-wise representations through temporal interaction across consecutive frames and spatial interaction across joints. LiveHPS++(Ren et al., 2024a) further conditions point-cloud interaction on predicted joints to estimate motion cues, which are then used for pose optimization. For longer-term motion modeling, BMLiCap(An et al., 2026) models motion structures with multi-scale Bézier trajectories, while OptimalCap(Ren et al., 2026b) combines hierarchical skeletal representations with kinematics-aware temporal optimization. These developments progressively enhance temporal and motion reasoning, but largely retain the skeleton as the reference space for organizing observations.

Despite these advances, many structured LiDAR motion-capture methods still use skeletal states as references for organizing local or cross-frame observations. This coupling makes evidence association dependent on skeletal inference, which can become unreliable when the input is sparse, noisy, or occluded.

Surface-based Human Representation Learning Surface representations have been explored to preserve geometric information beyond sparse skeletal joints. In human reconstruction, BodyFusion and DoubleFusion (Yu et al., 2017; 2018) combine articulated body structures with deformable surface models to recover geometry and track motion over time. Subsequent approaches extend surface modeling toward human reconstruction from incomplete observations. Continuous implicit representations, such as Occupancy Networks (Mescheder et al., 2019), provide formulations for surface representation, while PIFu and PIFuHD (Saito et al., 2019; 2020) associate image-aligned features with 3D queries for clothed-human digitization. Body-aware reconstruction methods including ICON (Xiu et al., 2022) and ECON (Xiu et al., 2023) incorporate parametric-body guidance with surface reasoning to recover geometry beyond underlying body model.

In 3D observations, IF-Nets (Chibane et al., 2020) and IPNet (Bhatnagar et al., 2020) infer human geometry from incomplete point clouds, while point-cloud pose and shape estimators employ occlusion-aware voting or cascaded regression (Liu et al., 2021; Cai et al., 2026). LiDAR-HMR (Fan et al., 2025) progressively expands sparse pose representations into body meshes using LiDAR features. Across these formulations, surface representations primarily serve as geometric reconstruction targets or shape priors rather than intermediate spaces for temporal evidence organization.

Motion-capture methods introduce surface information as a representation for pose inference. Sparkle (Ren et al., 2026c) combines skeletal joints and surface anchors, where anchors are initialized from predicted joints and refined using geometric features. In contrast, AMBER initializes surface references from current LiDAR observations and uses them to organize multi-scale cross-frame geometric and directional-motion evidence before skeletal refinement. AMBER therefore differs not merely in incorporating surface information, but in changing its role from skeleton-guided enhancement to observation-grounded evidence organization.

3 METHOD

We formulate LiDAR-based human motion capture as recovering temporally coherent 3D human motion from sparse point-cloud observations. Let $\mathcal{X} = \{\mathbf{X}_t\}_{t=0}^{T-1}$ denote a sequence of T frames, where $\mathbf{X}_t \in \mathbb{R}^{N\times 3}$ contains N observed points. The model estimates 6D joint rotations $\mathbf{R}_t \in \mathbb{R}^{K\times 6}$ and recovers pelvis-rooted joint positions $\mathbf{J}_t \in \mathbb{R}^{K\times 3}$ through SMPL (Loper et al., 2015), where $K = 24$ is the number of SMPL joints. Fig. 2 shows the overall framework.

Sparse observations can make coarse skeletal estimates unreliable, especially in partially observed body regions. Using such estimates to organize cross-frame evidence couples evidence construction with skeletal reasoning. We therefore introduce the Adaptive Motion-aware Body-surface Evidence Representation (AMBER). At frame t, AMBER comprises Q_{anc} surface anchors, $\mathcal{A}_t = \{(\mathbf{p}_{t,k}, \mathbf{g}_{t,k}, \mathbf{m}_{t,k})\}_{k=1}^{Q_{\mathrm{anc}}}$, where $\mathbf{p}_{t,k} \in \mathbb{R}^3$ is the anchor position and $\mathbf{g}_{t,k}, \mathbf{m}_{t,k} \in \mathbb{R}^d$ are its geometric and motion descriptors. AMBER Construction organizes cross-frame LiDAR observations without using intermediate skeletal estimates (Sec. 3.1). Geometry–Motion Refinement (GMR) then reads AMBER to refine joint positions and displacements (Sec. 3.2). Training is described in Sec. 3.3.

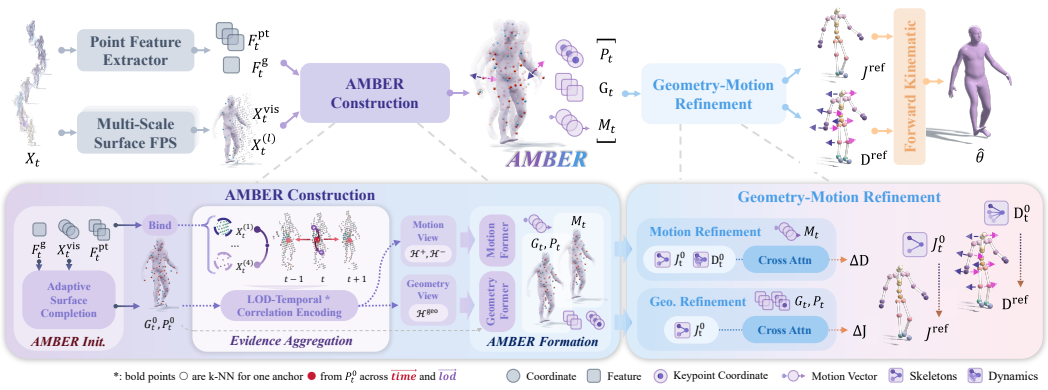

Figure 2: Overview of AMBER. AMBER Construction initializes sparse surface anchors from the current LiDAR observation and aggregates multi-scale geometric and directional-motion cues across neighboring frames. Geometry–Motion Refinement then transfers the organized surface evidence to the skeletal domain for skeleton refinement and final SMPL pose recovery.

Feature Extraction. For each frame $\mathbf{X}_t$, PointNet++ (Qi et al., 2017) extracts point-wise features $\mathbf{F}_t^{\mathrm{pt}} \in \mathbb{R}^{N\times d}$. We combine them with positional encodings and obtain a global feature $\mathbf{F}_t^{\mathrm{g}} \in \mathbb{R}^D$ through adaptive max pooling and a linear projection. We suppress t for single-frame quantities and restore it when needed.

3.1 AMBER Construction

Initialization. For each frame t, we initialize surface anchors to provide spatial references for cross-frame retrieval under sparse and occluded observations. Following (Yu et al., 2023; Ren et al., 2026a), we combine N_{vis} visible points $\mathbf{X}_t^{\mathrm{vis}} \in \mathbb{R}^{N_{\mathrm{vis}}\times 3}$ obtained by farthest-point sampling (FPS) with N_{miss} missing-region candidates predicted from $\mathbf{F}_t^{\mathrm{g}}$. A proximity-scoring network ranks the $N_{\mathrm{vis}} + N_{\mathrm{miss}}$ candidates, and the top Q_{anc} candidates are selected as surface references. During training, the scores are supervised by the distances from the candidates to the ground-truth SMPL surface; no ground-truth surface is used at inference. The selected coordinates are combined with positional encodings and the global feature, and a two-layer Transformer decoder attends to the point-wise features to produce the initial anchor positions and descriptors, $(\mathbf{P}_t^0, \mathbf{G}_t^0)$. The initial anchor representation is defined as $\mathcal{A}_t^0 = (\mathbf{P}_t^0, \mathbf{G}_t^0, \mathbf{M}_t^0)$, where $\mathbf{M}_t^0 = \mathbf{0} \in \mathbb{R}^{Q_{\mathrm{anc}}\times d}$.

Multi-LOD Cross-Frame Retrieval. The initialized AMBER anchors serve as observation-driven references for cross-frame retrieval. For each temporal offset $\delta \in \{-1, 0, +1\}$, we construct an L-level of details (LOD) pyramid using farthest-point sampling (FPS). At level ℓ, $\mathbf{X}_{t+\delta}^{(\ell)} \in \mathbb{R}^{N_\ell\times 3}$ and $\mathbf{F}_{t+\delta}^{(\ell)} \in \mathbb{R}^{N_\ell\times d}$ denote sampled points and point-wise features. For each anchor $\mathbf{p}_{t,k}^0 \in \mathbb{R}^3$, we retrieve n nearest points from every level and offset, providing multi-scale temporal evidence.

Omitting (t, k) from neighbor indices, let $\mathbf{x}_{\ell,i}^\delta \in \mathbb{R}^3$ and $\mathbf{f}_{\ell,i}^\delta \in \mathbb{R}^d$ denote the position and feature of the i-th point. We encode each observation relative to the anchor using $\mathbf{r}_{\ell,i}^\delta = \mathbf{x}_{\ell,i}^\delta - \mathbf{p}_{t,k}^0$, temporal embedding $\mathbf{e}_\delta \in \mathbb{R}^{d_t}$, and consistency cue $\rho_{\ell,i}^\delta = \frac{\langle \mathbf{W}_g \mathbf{g}_{t,k}^0, \mathbf{W}_f \mathbf{f}_{\ell,i}^\delta \rangle}{\|\mathbf{W}_g \mathbf{g}_{t,k}^0\| \|\mathbf{W}_f \mathbf{f}_{\ell,i}^\delta\|}$, where $\mathbf{W}_g, \mathbf{W}_f \in \mathbb{R}^{d_s\times d}$ project features into a shared consistency space. A point-wise encoder $\phi : \mathbb{R}^{d+3+d_t+1} \to \mathbb{R}^d$ maps these cues into tokens:

$$\mathbf{E}_\ell^\delta = \left[\phi\left([\mathbf{f}_{\ell,i}^\delta, \mathbf{r}_{\ell,i}^\delta, \mathbf{e}_\delta, \rho_{\ell,i}^\delta]\right)\right]_{i=1}^n \in \mathbb{R}^{n\times d}. \tag{1}$$

The retrieved tokens are organized into geometry and motion evidence pools. Geometry aggregation integrates observations from all temporal offsets, whereas motion aggregation preserves directional cues from adjacent frames:

$$\begin{aligned}\mathbf{H}^{\mathrm{geo}}_{t,k} &= \mathrm{Concat}^{L}_{\ell=1}[\mathbf{E}^{-1}_{\ell};\mathbf{E}^{0}_{\ell};\mathbf{E}^{+1}_{\ell}] \in \mathbb{R}^{3Ln\times d},\\ \mathbf{H}^{\delta}_{t,k} &= \mathrm{Concat}^{L}_{\ell=1}\,\mathbf{E}^{\delta}_{\ell} \in \mathbb{R}^{Ln\times d}, \qquad \delta\in\{-1,+1\}.\end{aligned} \tag{2}$$

The geometry pool is used to update anchor positions and geometric descriptors, while the two directional motion pools provide temporal motion cues for subsequent refinement.

AMBER Formation. To organize the retrieved evidence into a surface representation for subsequent skeletal reasoning, we use the initial anchor positions and geometric descriptors $(\mathbf{P}^0, \mathbf{G}^0)$ to construct separate geometry and motion queries. Both branches use multi-head cross-attention with branch-specific projections of AMBER evidence into keys and values. The geometry branch jointly aggregates evidence from different temporal offsets, whereas the motion branch preserves directional information from previous and next frames.

The geometry query attends to $\mathbf{H}^{\mathrm{geo}}$ and produces anchor-wise features $\mathbf{U}^{\mathrm{geo}}$. Meanwhile, the motion query separately attends to $\mathbf{H}^{\delta}$ and produces directional features $\mathbf{U}^{\delta}$. The attended features update anchor positions and descriptors through residual refinement:

$$\begin{aligned}\mathbf{p}_k &= \mathbf{p}^0_k + \alpha\tanh(h_p(\mathbf{u}^{\mathrm{geo}}_k)),\\ \mathbf{g}_k &= \mathbf{g}^0_k + h_g(\mathbf{u}^{\mathrm{geo}}_k),\\ \mathbf{m}_k &= \mathrm{Fuse}([\mathbf{u}^{-1}_k;\mathbf{u}^{+1}_k]).\end{aligned} \tag{3}$$

Here, $\mathbf{u}^{\mathrm{geo}}_k, \mathbf{u}^{-1}_k, \mathbf{u}^{+1}_k \in \mathbb{R}^d$ are attended anchor features, $h_p : \mathbb{R}^d \to \mathbb{R}^3$ and $h_g : \mathbb{R}^d \to \mathbb{R}^d$ predict position and feature residuals, and α bounds the position correction. The fusion function $\mathrm{Fuse} : \mathbb{R}^{2d} \to \mathbb{R}^d$ maps directional features into the motion descriptor. Collecting all anchors yields $\mathcal{A} = (\mathbf{P}, \mathbf{G}, \mathbf{M})$, where $\mathbf{P} \in \mathbb{R}^{Q_{\mathrm{anc}}\times 3}$ and $\mathbf{G}, \mathbf{M} \in \mathbb{R}^{Q_{\mathrm{anc}}\times d}$. Then we generate a dense point set $\mathbf{P}^{\mathrm{com}} \in \mathbb{R}^{N_{\mathrm{com}}\times 3}$ from $(\mathbf{P}, \mathbf{G})$ for surface-completion supervision, without passing it to GMR.

3.2 Geometry–Motion Refinement

After cross-frame evidence has been organized in AMBER, skeletal states are introduced for downstream refinement. Following skeletal initialization strategies (Li et al., 2022; An et al., 2026), a GRU predicts the initial joint positions $\mathbf{J}^0 \in \mathbb{R}^{K\times 3}$, the initial bidirectional motion states $\mathbf{D}^0 \in \mathbb{R}^{K\times 6}$, and the hidden state $\mathbf{h} \in \mathbb{R}^{d_h}$. GMR refines these skeletal states by reading the AMBER representation. It contains two independent branches for geometry and motion refinement.

The AMBER representation is treated as a set of anchor tokens. Both branches use multi-head cross-attention with separate parameters, producing refinement features $\mathbf{Z}_b \in \mathbb{R}^{K\times d}$ for $b \in \{\mathrm{geo}, \mathrm{mot}\}$.

Geometry-aware Joint Refinement. The geometry branch refines joint locations using AMBER geometric evidence. Its query represents the current skeletal state, while the memory contains the organized anchor positions and geometric descriptors. Let $\mathbf{e}_{\mathrm{joint}} \in \mathbb{R}^{d_j}$ and $\mathbf{e}_{\mathrm{time}} \in \mathbb{R}^{d_t}$ denote learned joint and temporal embeddings. The query and memory are defined as

$$\begin{aligned}\mathbf{Q}_{\mathrm{geo}} &= \phi^{\mathrm{geo}}_q([\mathbf{e}_{\mathrm{joint}}, \mathbf{h}, \mathrm{PE}(\mathbf{J}^0), \mathbf{e}_{\mathrm{time}}]),\\ \mathbf{S}_{\mathrm{geo}} &= \phi^{\mathrm{geo}}_m([\mathbf{P}, \mathbf{G}, \mathbf{e}^{\mathrm{anc}}_{\mathrm{time}}]).\end{aligned} \tag{4}$$

Here, $\mathbf{e}_{\mathrm{joint}}$ and $\mathbf{h}$ are broadcast to all joint tokens before concatenation, $\mathrm{PE}(\mathbf{J}^0) \in \mathbb{R}^{K\times d_p}$ provides joint-specific positional information, and $\mathbf{e}^{\mathrm{anc}}_{\mathrm{time}} \in \mathbb{R}^{Q_{\mathrm{anc}}\times d_t}$ is obtained by broadcasting the temporal embedding to all anchors. The resulting query and memory satisfy $\mathbf{Q}_{\mathrm{geo}} \in \mathbb{R}^{K\times d}$ and $\mathbf{S}_{\mathrm{geo}} \in \mathbb{R}^{Q_{\mathrm{anc}}\times d}$. The functions ϕ^{geo}_q and ϕ^{geo}_m are token-wise projections. The refined joints are obtained through residual correction:

$$\mathbf{J}^{\mathrm{ref}} = \mathbf{J}^0 + \alpha\tanh(f_{\mathrm{geo}}(\mathbf{Z}_{\mathrm{geo}})), \tag{5}$$

where $f_{\mathrm{geo}} : \mathbb{R}^d \to \mathbb{R}^3$ is a token-wise residual head that predicts a 3D displacement for each joint.

Motion-aware Joint Refinement. The motion branch refines temporal dynamics by transferring AMBER motion evidence to the current skeletal motion state. Its query incorporates skeletal motion context, while the memory combines surface motion and positional information:

$$\begin{aligned}\mathbf{Q}_{\text{mot}} &= \phi_q^{\text{mot}}([\mathbf{e}_{\text{joint}}, \mathbf{D}^0, \text{PE}(\mathbf{J}^0), \mathbf{h}, \mathbf{e}_{\text{time}}]),\\ \mathbf{S}_{\text{mot}} &= \phi_m^{\text{mot}}([\mathbf{M}; \phi_p(\mathbf{P})]).\end{aligned} \tag{6}$$

Here, $\mathbf{Q}_{\text{mot}} \in \mathbb{R}^{K\times d}$ and $\mathbf{S}_{\text{mot}} \in \mathbb{R}^{2Q_{\text{anc}}\times d}$. The position projection $\phi_p : \mathbb{R}^3 \to \mathbb{R}^d$ is applied independently to each anchor token, and ϕ_m^{mot} is a token-wise projection. The semicolon denotes row-wise stacking, allowing the motion memory to preserve both directional motion descriptors and surface positional cues. The refined motion state is predicted by

$$\mathbf{D}^{\text{ref}} = \mathbf{D}^0 + f_{\text{mot}}(\mathbf{Z}_{\text{mot}}), \tag{7}$$

where $f_{\text{mot}} : \mathbb{R}^d \to \mathbb{R}^6$ is a learned motion residual head.

Finally, $\mathbf{D}^{\text{ref}}$ and $\mathbf{J}^{\text{ref}}$ are sent to a commonly used SMPL solver, where an ST-GCN-based inverse kinematics solver is utilized to predict joint rotations $\mathbf{R}_t$.

3.3 Training Objective

Training uses ground-truth joint positions $\widehat{\mathbf{J}}_t$, 6D joint rotations $\widehat{\mathbf{R}}_t$, bidirectional joint motion $\widehat{\mathbf{D}}_t$, and sampled SMPL surface vertices $\widehat{\mathbf{V}}_t$. The objective is

$$\mathcal{L} = \lambda_A \mathcal{L}_{\text{AMBER}} + \lambda_K \mathcal{L}_{\text{K}} + \lambda_S \mathcal{L}_{\text{SMPL}}, \tag{8}$$

where λ_A, λ_K, and λ_S balance the AMBER, kinematic, and SMPL losses, respectively.

AMBER supervision combines surface completion and candidate scoring:

$$\mathcal{L}_{\text{AMBER}} = \lambda_{\text{cd}} \mathcal{L}_{\text{cd}} + \mathcal{L}_{\text{score}}, \tag{9}$$

where the completion term supervises the missing-region candidates, the formed AMBER positions, and the completed point set:

$$\mathcal{L}_{\text{cd}} = \frac{1}{T}\sum_{t=0}^{T-1}\big[\,\text{CD}_2(\mathbf{C}_t^{\text{miss}}, \widehat{\mathbf{V}}_t) + \text{CD}_2(\mathbf{P}_t, \widehat{\mathbf{V}}_t) + \text{CD}_2(\mathbf{P}_t^{\text{com}}, \widehat{\mathbf{V}}_t)\big]. \tag{10}$$

Here, CD_2 is the symmetric squared Chamfer distance. For each scored candidate i, the proximity target is $s_{t,i}^\star = \exp(-d_{t,i}^{\min}/\tau)$, where $\tau > 0$ and $d_{t,i}^{\min}$ is its distance to the nearest ground-truth SMPL vertex. $\mathcal{L}_{\text{score}}$ is the mean squared error between the predicted scores $s_{t,i}$ and these targets.

Then, the kinematic loss contains joint and motion losses supervise both prediction stages. We use

$$\mathcal{L}_K = \frac{1}{T}\sum_{s\in\{0,\text{ref}\}}\sum_{t=0}^{T-1}\Big[\mathcal{L}_{\text{MSE}}(\mathbf{J}_t^s, \widehat{\mathbf{J}}_t) + \mathcal{L}_{\text{smoothL1}}(\mathbf{D}_t^s, \widehat{\mathbf{D}}_t)\Big]. \tag{11}$$

The two stages contribute with equal weights. The motion loss excludes invalid displacement targets. Finally, the SMPL solver loss supervises the forward and inverse kinematics from SMPL outputs:

$$\begin{aligned}\mathcal{L}_{\text{SMPL}} &= \mathcal{L}_{\text{FK}} + \mathcal{L}_{\text{IK}}\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\Big[\mathcal{L}_{\text{MSE}}(\mathbf{R}_t, \widehat{\mathbf{R}}_t) + \mathcal{L}_{\text{MSE}}(\mathbf{J}_t^{\text{FK}}, \widehat{\mathbf{J}}_t)\Big]\end{aligned} \tag{12}$$

Here, $\mathbf{J}_t^{\text{FK}}$ denotes the final SMPL joint positions obtained from the predicted pose.

Table 1: Comparison with state-of-the-art methods on four benchmarks. JPE/VPE are in mm and AE in m/s^2. Best and second-best results are shown in **bold** and underline, respectively. * denotes baseline w/o auxiliary inputs (IMU/neighbor scene scans).

Method	LiDARHuman26M			FreeMotion			NoiseMotion			SLOPER4D		
	JPE↓	VPE↓	AE↓	JPE↓	VPE↓	AE↓	JPE↓	VPE↓	AE↓	JPE↓	VPE↓	AE↓
MOVIN (Jang et al., 2023)	–	–	–	–	–	–	–	–	–	123.8	146.3	10.63
LiDAR-HMR (Fan et al., 2025)	76.2	102.5	–	106.7	130.7	–	50.4	65.2	–	47.7	49.7	–
LiDARCap (Li et al., 2022)	79.3	101.6	4.52	86.3	104.2	6.28	52.6	64.7	4.25	71.6	84.2	4.01
LIP* (Ren et al., 2023)	75.7	96.6	3.28	85.5	90.8	6.16	62.4	78.0	2.53	60.1	74.9	6.16
NE-LiDARCap* (Zhang et al., 2024)	76.8	97.1	3.16	62.5	75.8	2.98	48.8	60.4	2.79	96.8	113.9	3.86
LiveHPS (Ren et al., 2024b)	71.9	92.1	3.41	69.4	83.9	6.99	48.4	60.4	5.78	53.4	63.2	5.88
LiveHPS++ (Ren et al., 2024a)	–	–	–	61.9	75.3	5.42	34.0	42.8	3.48	42.7	50.6	4.34
BMLiCap (An et al., 2026)	66.8	85.4	**2.88**	47.2	59.0	**2.25**	36.9	47.0	2.38	36.5	44.2	1.36
Sparkle (Ren et al., 2026c)	–	–	–	59.0	73.2	–	27.8	36.3	–	41.4	50.5	–
OptimalCap (Ren et al., 2026b)	–	–	–	58.2	72.0	5.43	30.4	38.9	3.21	55.0	69.9	7.01
AMBER(Ours)	**65.5**	**84.4**	2.90	**42.8**	**53.6**	2.33	**27.5**	**36.1**	**2.00**	**32.9**	**40.8**	**1.23**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate AMBER on four challenging LiDAR-based human motion capture benchmarks: LiDARHuman26M (Li et al., 2022), FreeMotion (Ren et al., 2024b), NoiseMotion (Ren et al., 2024a), and SLOPER4D (Dai et al., 2023). These challenging datasets contain extensive occlusion, varying levels of sparsity, and dynamic samples.

Evaluation Metrics We evaluate spatial reconstruction and temporal dynamics using three metrics (Li et al., 2022; Ren et al., 2024b; An et al., 2026): Joint Position Error (JPE, mm), Vertex Position Error (VPE, mm), and Acceleration Error (AE, m/s^2). JPE/VPE measure root-relative reconstruction errors of joints and SMPL surfaces, while AE captures temporal motion variations.

Implementation Details. We process 16-frame clips with 512 LiDAR points per frame, and the feature backbone is pretrained by raycasting following An et al. (2025). AMBER uses 64 surface anchors and four FPS levels containing $(256, 128, 64, 32)$ points, retrieving 16 neighbors per level and temporal offset. Surface and joint features both have 64 channels. We set the refining gating $\alpha = 0.25$, and the hyper-parameters are set to $\lambda_A = \lambda_K = 1, \lambda_{cd} = 10, \lambda_S = 5$. More details are provided in the *appendix*.

4.2 MAIN RESULTS

Quantitative Comparison. We compare AMBER with state-of-the-art LiDAR-based human motion capture methods across four benchmarks in Table 1. AMBER achieves the lowest JPE and VPE on all four datasets while maintaining competitive temporal consistency. On FreeMotion, AMBER reduces JPE by 4.4 mm compared with BMLiCap. On NoiseMotion and SLOPER4D, it further achieves the lowest AE, together with substantial improvements in spatial reconstruction. These results demonstrate that AMBER generalizes effectively across self-occlusion, measurement noise, and sparse long-range observations.

Qualitative Results. Figure 3 compares results across four benchmarks with sparse observations and challenging poses. Missing points make the occluded body regions ambiguous. Several baselines misplace the extended leg in FreeMotion, the limbs in LiDARHuman26M, or the raised arm in NoiseMotion. On SLOPER4D, some fail to preserve the body lean. In contrast, AMBER better retains the overall posture and affected limb configurations, producing reconstructions closer to the ground truth despite incomplete surface observations. Furthermore, Figure 4 demonstrates AMBER's effective completion performance, particularly for deformable parts (arms and legs).

4.3 ABLATION STUDIES

Component Analysis. Table 2 disentangles three contributions of AMBER and GMR. First, AMBER surface reconstruction provides effective auxiliary supervision: on FreeMotion, Init-Aux reduces MPJPE from 52.8 to 49.2 mm, and incorporating cross-frame surface evidence further im-

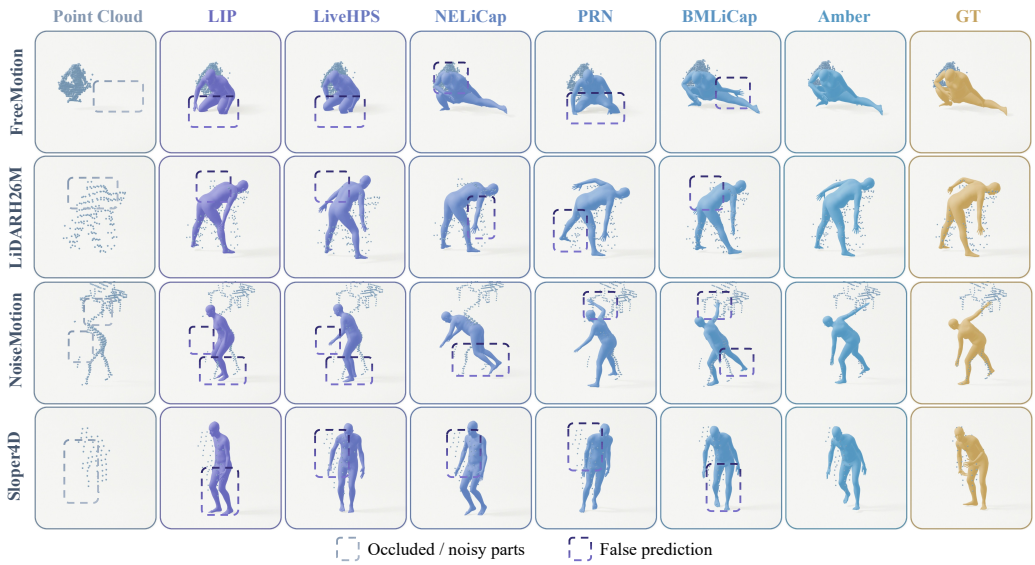

Figure 3: Qualitative comparisons on FreeMotion, LiDARHuman26M, NoiseMotion, and SLOPER4D. We compare reconstructions from competitive baselines. AMBER better preserves challenging limb configurations and overall body posture under self-occlusion, noisy measurements, and sparse long-range observations.

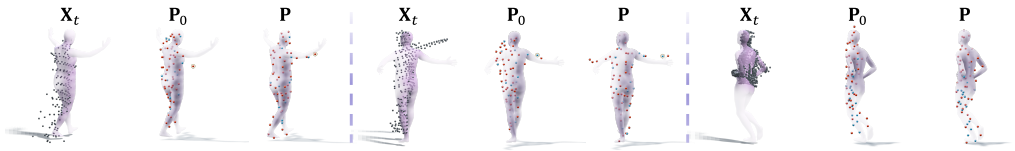

Figure 4: AMBER's completion capabilities. Body parts with occlusions are well-reconstructed.

proves it to 48.7 mm. Second, explicitly transferring the learned surface geometry to skeletal refinement yields substantially larger gains, with CF-GR reaching 43.3 mm on FreeMotion and 33.9 mm on SLOPER4D. Third, the motion branch also improves skeletal estimation by exploiting directional temporal cues, while combining geometry and motion refinement achieves the best MPJPE across all three benchmarks (42.8, 32.9, and 65.5 mm). These results show that surface reconstruction is useful not only as auxiliary supervision, but also as an intermediate representation whose geometric and motion cues complement each other during skeletal refinement.

Robustness across Capture Distances. We further evaluate AMBER across different LiDAR capture distances on SLOPER4D and NoiseMotion. As shown in Fig. 5, the spatial and temporal errors remain relatively stable across distance ranges, indicating that AMBER remains effective under varying capture conditions.

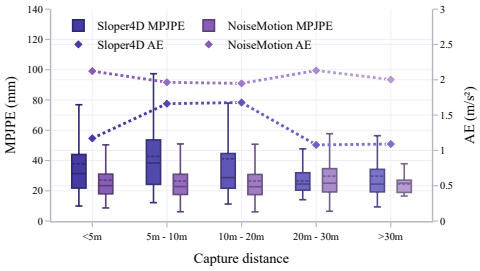

Figure 5: Performance across different LiDAR capture distances on SLOPER4D and NoiseMotion. AMBER maintains stable performance even with sparse inputs.

Retrieval-Space and Robustness Analysis. We first isolate the effect of the retrieval reference space through the controlled comparison in Table 3. Surface-anchored retrieval achieves the best or tied performance across all evaluated dataset–metric pairs, while replacing predicted skeletal references with ground-truth locations produces only marginal changes. Under this controlled setting, the advantage of surface-anchored retrieval is therefore mainly associated with the reference space rather than skeletal estimation accuracy. We further evaluate robustness to observation degradation and skeletal-state perturbations in Fig. 1(c). As the occlusion ratio increases, AMBER maintains lower MPJPE than the compared methods. In the lower plot, we perturb the intermediate skeletal state with increasing magnitude. AMBER exhibits

Table 2: Component ablation across three benchmarks. Results are reported as MPJPE / MPVPE / AE. Baseline uses PointNet++ with skeleton initialization and the same SMPL solver, without AMBER or GMR. Aux uses AMBER only for auxiliary surface reconstruction; Init/CF denote single-frame/cross-frame AMBER, and GR/MR denote geometry/motion refinement.

Dataset	Baseline	AMBER without GMR		Geometry–Motion Refinement (GMR)			Ours
		Init-Aux	CF-Aux	Init-GR	CF-GR	CF-MR	
FreeMotion	52.8/64.9/2.68	49.2/60.9/2.59	48.7/59.8/2.54	45.2/56.4/2.39	43.3/53.9/2.35	44.3/55.2/**2.32**	**42.8/53.6**/2.33
SLOPER4D	55.0/66.7/2.04	42.8/53.2/1.56	41.7/52.1/1.57	36.1/44.5/1.29	33.9/41.6/1.40	34.0/42.2/1.46	**32.9/40.8/1.23**
LiDARHuman26M	70.6/90.7/3.03	66.6/85.8/2.92	66.8/86.0/2.93	65.9/84.8/2.91	66.3/85.3/2.93	66.3/85.1/2.92	**65.5/84.4/2.90**

Table 3: Ablation of cross-frame retrieval reference spaces.

Retrieval reference	LiDARHuman26M			FreeMotion			NoiseMotion			SLOPER4D		
	MPJPE↓	MPVPE↓	AE↓	MPJPE↓	MPVPE↓	AE↓	MPJPE↓	MPVPE↓	AE↓	MPJPE↓	MPVPE↓	AE↓
Skeleton anchors	65.9	84.8	2.93	45.3	56.5	2.39	30.9	39.9	2.25	33.8	41.6	1.29
Oracle skeleton anchors	65.6	84.4	2.92	45.2	56.3	2.38	30.7	39.7	2.25	33.7	41.6	1.29
Surface anchors	**65.5**	**84.4**	**2.90**	**42.8**	**53.6**	**2.33**	**27.5**	**36.1**	**2.00**	**32.9**	**40.8**	**1.23**

smaller MPJPE degradation, indicating greater resilience to errors in intermediate skeletal states. Together, these analyses show that surface space provides a more effective reference for organizing cross-frame LiDAR evidence, while AMBER also remains robust under degraded observations and skeletal-state perturbations.

AMBER Cardinality. We vary the number of surface anchors Q_{anc} on SLOPER4D and FreeMotion to study the effect of surface representation density. As shown in Table 4, increasing Q_{anc} from 16 to 64 consistently improves reconstruction accuracy on both datasets. On FreeMotion, for example, JPE decreases from 52.8 mm to 42.8 mm, indicating that additional anchors provide more complete surface evidence for downstream refinement. Increasing Q_{anc} further to 96 yields no further improvement. We therefore use $Q_{\text{anc}} = 64$ by default.

Table 4: Ablation of the number of AMBER surface anchors.

Q_{anc}	SLOPER4D			FreeMotion		
	JPE↓	**VPE↓**	**AE↓**	**JPE↓**	**VPE↓**	**AE↓**
16	36.4	45.1	1.29	52.8	66.2	2.50
32	33.5	41.3	1.25	46.4	58.6	2.39
64	**32.9**	**40.8**	1.23	**42.8**	**53.6**	**2.33**
96	33.0	40.8	**1.22**	43.4	53.9	2.34

Multi-LOD Cross-Frame Retrieval. Table 5 compares alternative neighborhood construction strategies on LiDARHuman26M. The proposed Multi-LOD FPS Pyramid achieves the best JPE and VPE of 65.5 mm and 84.4 mm, respectively, with an AE of 2.90. Increasing single-scale kNN from 16 to 64 neighbors degrades performance, suggesting that simply enlarging a local neighborhood introduces spatially less relevant observations rather than useful context. Multi-radius and dilated retrieval enlarge the receptive field but remain inferior to the hierarchical FPS design. By retrieving evidence from progressively subsampled point sets, the FPS pyramid captures complementary local and coarse-scale context without excessively enlarging neighborhoods at a single resolution. These results support hierarchical retrieval as a more effective strategy for collecting surface evidence from sparse and incomplete LiDAR observations.

Table 5: Ablation of different neighborhood construction strategies for Multi-LOD cross-frame retrieval.

Variant	**JPE↓**	**VPE↓**	**AE↓**
Single-scale KNN-16	66.0	84.8	2.90
Single-scale KNN-64	66.6	85.8	2.92
Multi-radius Search	66.2	85.2	2.91
Dilated KNN	66.2	85.2	2.90
Multi-LOD FPS Pyramid	**65.5**	**84.4**	**2.90**

5 Conclusion

We presented AMBER, a body-surface evidence representation for LiDAR-based human motion capture from sparse and incomplete point clouds. AMBER organizes multi-scale geometric and directional motion cues around adaptive surface anchors before using them for skeletal refinement, reducing the dependence of cross-frame association on uncertain initial joints. Geometry–Motion Refinement then uses this evidence to correct joint positions and bidirectional motion. Experiments on LiDARHuman26M, FreeMotion, NoiseMotion, and SLOPER4D demonstrate improved reconstruction accuracy and temporal consistency under challenging observations.

AI USE STATEMENT

In this work, we used generative AI tools for translation, language polishing, expression refinement, and the detection of typographical and grammatical errors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies LiDAR-based 3D human motion capture using publicly available datasets released for research purposes. We do not collect new data, recruit human participants, or conduct direct interactions with individuals. The experiments do not use or report personally identifiable information, and all datasets are used in accordance with their licenses and terms of use. We follow standard practices for responsible research and data handling.

REPRODUCIBILITY STATEMENT

To facilitate verification and extension of our work, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in the paper and supplementary materials. The complete source code and trained model weights will be made publicly available after publication to support reproducibility and further research.

REFERENCES

Xiaoqi An, Lin Zhao, Chen Gong, Jun Li, and Jian Yang. Pre-training a Density-Aware Pose Transformer for Robust LiDAR-based 3D Human Pose Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1755–1763, 2025.

Xiaoqi An, Lin Zhao, Jun Li, Chen Gong, and Jian Yang. Bezier degradation modeling for LiDAR-based human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14027–14037, 2026.

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pp. 311–329, 2020.

Zhongang Cai, Liang Pan, Wanqi Yin, Chen Wei, Fangzhou Hong, Mingyuan Zhang, Atsushi Yamashita, Chen Change Loy, Lei Yang, and Ziwei Liu. PointHPS: Cascaded 3D Human Pose and Shape Estimation from Point Clouds. *International Journal of Computer Vision*, 134(5):200, 2026.

Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6970–6981, 2020.

Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 682–692, 2023.

Bohao Fan, Wenzhao Zheng, Jianjiang Feng, and Jie Zhou. LiDAR-HMR: 3D human mesh recovery from LiDAR. *IEEE Transactions on Multimedia*, 27:6962–6975, 2025.

Deok-Kyeong Jang, Dongseok Yang, Deok-Yun Jang, Byeoli Choi, Taeil Jin, and Sung-Hee Lee. MOVIN: Real-time Motion Capture using a Single LiDAR. *Computer Graphics Forum*, 42(7): e14961, 2023.

Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. LiDARCap: Long-Range Marker-Less 3D Human Motion Capture with LiDAR Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20470–20480, 2022.

Guanze Liu, Yu Rong, and Lu Sheng. VoteHMR: Occlusion-Aware Voting Network for Robust 3D Human Mesh Recovery from Partial Point Clouds. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 955–964, 2021.

Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2720–2735, 2013.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Huan Ren, Yihan Chen, Chuxin Wang, Nailong Liu, Wenfei Yang, and Tianzhu Zhang. ComPose: A unified completion-pose framework for robust category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14315–14324, 2026a.

Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2337–2347, 2023.

Yiming Ren, Xiao Han, Yichen Yao, Xiaoxiao Long, Yujing Sun, and Yuexin Ma. LiveHPS++: Robust and Coherent Motion Capture in Dynamic Free Environment. In *European Conference on Computer Vision*, pp. 127–144, 2024a.

Yiming Ren, Xiao Han, Chengfeng Zhao, Jingya Wang, Lan Xu, Jingyi Yu, and Yuexin Ma. LiveHPS: LiDAR-based Scene-level Human Pose and Shape Estimation in Free Environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1281–1291, 2024b.

Yiming Ren, Yujing Sun, Xiao Han, Yichen Yao, Xiaoxiao Long, Xinge Zhu, Siu Ming Yiu, and Yuexin Ma. OptimalCap: Efficient and robust LiDAR-based motion capture in free environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8456–8469, 2026b.

Yiming Ren, Yujing Sun, Aoru Xue, Kwok-Yan Lam, and Yuexin Ma. Sparkle: A robust and versatile representation for point cloud-based human motion capture. In *International Conference on Learning Representations*, 2026c.

Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. EgoCap: Egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics*, 35(6):162:1–162:11, 2016.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020.

Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics*, 26(3):35:1–35:9, 2007.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13296–13306, 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 512–523, 2023.

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 910–919, 2017.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7287–7296, 2018.

Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. AdaPoinTr: Diverse point cloud completion with adaptive geometry-aware transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14114–14130, 2023.

Jingyi Zhang, Qihong Mao, Guosheng Hu, Siqi Shen, and Cheng Wang. Neighborhood-Enhanced 3D Human Pose Estimation with Monocular LiDAR in Long-Range Outdoor Scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7169–7177, 2024.

A Implementation and Network Details

A.1 Implementation Details

Our method is implemented with PyTorch 2.8.0 and CUDA 12.8 and trained on 4 NVIDIA RTX 5090 GPUs for 50 epochs using AdamW (learning rate 5×10^{-4}, weight decay 0.005), a sequence length of 16, and a per-GPU batch size of 8, giving an effective batch size of 32. The intermediate keypoint decoder contains 2 layers with $d_{\text{model}} = 64$, 4 attention heads, and a 256-dimensional feed-forward network; each refinement stage uses two additional 2-layer cross-attention modules with the same 64-dimensional model, 4 heads, and 256-dimensional feed-forward layers. Specifically, the geometry-aware refinement attends from 24 joint queries to 64 surface tokens, while the motion-aware refinement attends to 128 surface-motion tokens for predicting the residual between M_0 and M_{ref}. For LiDAR-based datasets, we follow the dataset settings of BMLiCap and LiveHPS++, training and testing on the corresponding splits. We do not use any Depth-based datasets or the PointHPS training configuration.

A.2 Point Feature Extractor and Multi-Scale Pyramid

To establish a robust perceptual foundation for both body-surface evidence aggregation and skeletal initialization, we design a hierarchical spatio-temporal feature extractor. This module transforms raw, self-occluded LiDAR point sequences $\mathcal{X} = \{\mathbf{X}_t\}_{t=1}^{T}$ into dense local descriptors, abstract global context, and a multi-resolution feature pyramid.

Spatio-Temporal Encoding. Given the input point cloud $\mathbf{X}_t \in \mathbb{R}^{N\times 3}$ ($N = 512$) at frame t, we first employ a PointNet++ backbone consisting of three Set Abstraction (SA) modules and three Feature Propagation (FP) modules to extract hierarchical geometric representations. To resolve spatial ambiguities and non-local dependencies caused by occlusions (e.g., distinguishing inter-limb intersections), the per-point features are enriched with positional encodings and fed into a 2-layer Spatial Transformer Encoder. This self-attention module outputs the dense local point features $\mathbf{F}_t^l \in \mathbb{R}^{N\times d}$ (where $d = 64$), which explicitly serve as the memory bank for initializing surface anchors.

Subsequently, an adaptive spatial max-pooling operation aggregates $\mathbf{F}_t^l$ across the point dimension into a global frame descriptor $\mathbf{F}_t^g \in \mathbb{R}^D$ ($D = 1024$). To capture sequence-level kinematics, the

global features across T frames are processed by a 2-layer Bidirectional GRU (Bi-GRU), yielding the temporal hidden state $\mathbf{h}_t \in \mathbb{R}^{2048}$. This hidden state provides the holistic spatial-temporal context necessary for regressing the initial coarse skeletal joints $\mathbf{J}_t^0$ and bidirectional joint displacements $\mathbf{D}_t^0$.

Multi-Scale Feature Pyramid Construction. For AMBER to effectively perform cross-frame retrieval under varying point densities, we construct a corresponding Multi-Level of Detail (LOD) pyramid. We apply progressive Farthest Point Sampling (FPS) on the spatial coordinates $\mathbf{X}_t$, producing four downsampled coordinate sets: $\mathbf{X}_t^{(1)}$ ($N = 256$), $\mathbf{X}_t^{(2)}$ ($N = 128$), $\mathbf{X}_t^{(3)}$ ($N = 64$), and $\mathbf{X}_t^{(4)}$ ($N = 32$).

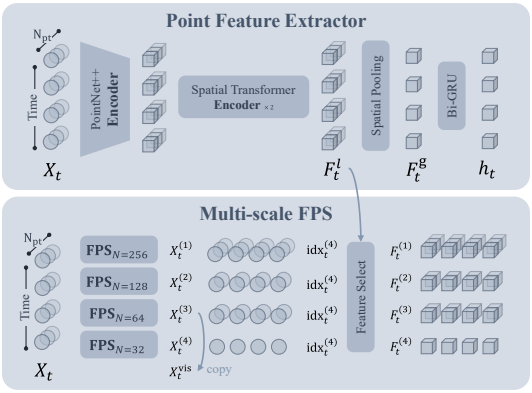

Figure 6: Detailed procedure of point feature extractor and Multi-Scale point pyramid forming.

Notably, the sparsest spatial representation $\mathbf{X}_t^{(4)}$ provides a highly uniform coverage of the observed LiDAR surface. Thus, it is directly copied as the visible candidate set $\mathbf{X}_t^{\text{vis}}$ for the AMBER Initialization stage. Finally, the physical indices extracted by the multi-scale FPS operations ($\text{idx}_t^{(1)}$ to $\text{idx}_t^{(4)}$) are utilized to gather (i.e., feature select) the corresponding features from the dense local tensor $\mathbf{F}_t^l$. This mechanism efficiently produces the multi-scale feature sets $\mathbf{F}_t^{(1)}$ to $\mathbf{F}_t^{(4)}$, which perfectly align with their coordinate counterparts to support the multi-scale k-NN evidence aggregation in AMBER Construction.

A.3 Detailed Formulation of AMBER Construction

To overcome the vulnerability of previous skeleton-conditioned reasoning paradigms under severe sparsity and occlusion, AMBER decouples temporal evidence organization from intermediate skeletal states. AMBER Construction constructs an observation-grounded surface representation $\mathcal{A}_t = \{(\mathbf{p}_{t,k}, \mathbf{g}_{t,k}, \mathbf{m}_{t,k})\}_{k=1}^{Q_{\text{anc}}}$ at each frame t, operating entirely independent of the coarse skeleton $\mathbf{J}_t^0$. As outlined below, this procedure comprises three foundational steps: *Adaptive Anchor Initialization*, *Multi-LOD Evidence Aggregation*, and *AMBER Formation*.

A.3.1 Adaptive Surface Anchor Initialization

To provide reliable spatial reference units across varying degrees of partial visibility, we initialize anchor locations from a combination of surface observations and learned priors. For frame t, the Point Feature Extractor yields point-wise features $\mathbf{F}_t^{\text{pt}} \in \mathbb{R}^{N \times d}$ and a global feature $\mathbf{F}_t^{\text{g}} \in \mathbb{R}^D$ from the input point cloud $\mathbf{X}_t \in \mathbb{R}^{N \times 3}$. We first construct a complementary candidate pool:

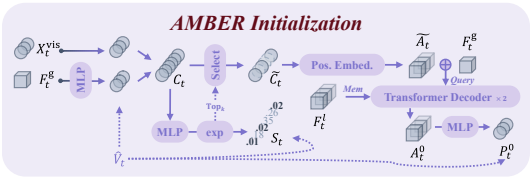

Figure 7: Details of AMBER initialization.

$$\mathbf{X}_t^{\text{vis}} = \text{FPS}_{N_{\text{vis}}}(\mathbf{X}_t) \in \mathbb{R}^{N_{\text{vis}} \times 3}, \quad \mathbf{C}_t^{\text{miss}} = \text{MLP}_{\text{miss}}(\mathbf{F}_t^{\text{g}}) \in \mathbb{R}^{N_{\text{miss}} \times 3}, \tag{13}$$

$$\mathbf{C}_t = [\mathbf{C}_t^{\text{miss}}; \mathbf{X}_t^{\text{vis}}] \in \mathbb{R}^{(N_{\text{vis}}+N_{\text{miss}}) \times 3}, \tag{14}$$

where $N_{\text{vis}} = 32$ visible samples capture confident surface observations, and $N_{\text{miss}} = 64$ generated samples cover missing or self-occluded anatomical regions.

To eliminate outlier candidates far from the actual body manifold, a lightweight proximity scoring network predicts confidence scores $\mathbf{s}_t \in \mathbb{R}^{N_{\text{vis}}+N_{\text{miss}}}$ supervised by Euclidean distances to ground-truth SMPL vertices $\widehat{\mathcal{V}}_t$:

$$s_{t,i} = \exp(-\text{MLP}_{\text{score}}(\mathbf{c}_{t,i})), \quad \text{supervised by} \quad s_{t,i}^{\star} = \exp\left(-\frac{\min_{\mathbf{v} \in \widehat{\mathcal{V}}_t} \|\mathbf{c}_{t,i} - \mathbf{v}\|_2}{\tau}\right), \tag{15}$$

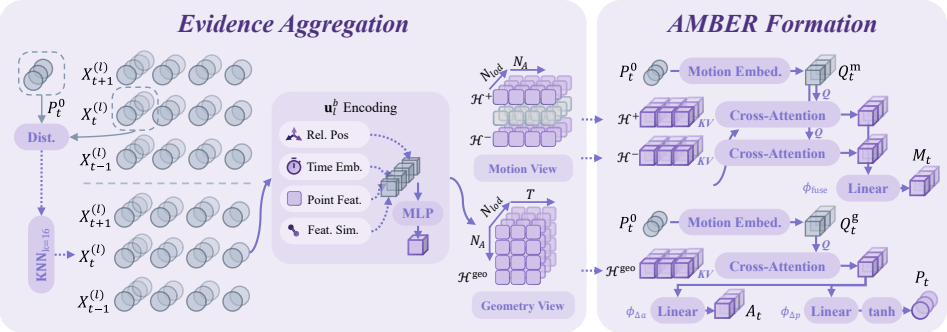

Figure 8: Details of AMBER evidence aggregation and formation.

where $\tau = 0.05\,\mathrm{m}$ is a metric scaling factor. We select the top $Q_{\mathrm{anc}} = 64$ scoring candidates $\widetilde{\mathbf{C}}_t = \mathrm{Top}_k(\mathbf{C}_t, \mathbf{s}_t) \in \mathbb{R}^{Q_{\mathrm{anc}} \times 3}$. These filtered coordinates are embedded with positional encodings $\mathrm{PE}(\widetilde{\mathbf{C}}_t)$ and injected with the global feature $\mathbf{F}_t^{\mathrm{g}}$ to form anchor queries $\mathbf{Q}_t = \mathrm{PE}(\widetilde{\mathbf{C}}_t) \oplus \mathbf{F}_t^{\mathrm{g}}$. A two-layer Transformer decoder then cross-attends $\mathbf{Q}_t$ to the local point feature memory $\mathbf{F}_t^{\mathrm{pt}}$, producing initial surface descriptors $\mathbf{A}_t^0 \in \mathbb{R}^{Q_{\mathrm{anc}} \times d}$. A direct linear head predicts the initial 3D anchor positions:

$$\mathbf{A}_t^0 = \mathrm{Dec}_\theta(\mathbf{Q}_t, \mathbf{F}_t^{\mathrm{pt}}), \qquad \mathbf{P}_t^0 = \phi_{\mathrm{pred}}(\mathbf{A}_t^0) \in \mathbb{R}^{Q_{\mathrm{anc}} \times 3}. \tag{16}$$

The resulting initial representation is $\mathcal{A}_t^0 = \{(\mathbf{p}_{t,k}^0, \mathbf{a}_{t,k}^0, \mathbf{0}_d)\}_{k=1}^{Q_{\mathrm{anc}}}$, where the motion descriptor is initialized to zero prior to temporal retrieval.

A.3.2 Multi-LOD Cross-Frame Evidence Retrieval

Unlike existing frameworks that condition point retrieval on predicted internal joints, AMBER uses the physical anchor coordinates $\mathbf{P}_t^0$ as unbiased local reference centers to retrieve spatio-temporal evidence. For frames $t + \delta$ with $\delta \in \{-1, 0, +1\}$, we construct an L-level FPS pyramid with sizes $N_\ell \in \{256, 128, 64, 32\}$ ($\ell \in \{1, \ldots, 4\}$). Let $\mathbf{X}_{t+\delta}^{(\ell)} \in \mathbb{R}^{N_\ell \times 3}$ and $\mathbf{F}_{t+\delta}^{(\ell)} \in \mathbb{R}^{N_\ell \times d}$ denote the coordinates and point features at level ℓ. For each anchor $\mathbf{p}_{t,k}^0$, we search for its $n = 16$ nearest neighbors across each level ℓ and temporal offset δ:

$$\mathcal{N}_{t,k}^{\ell,\delta} = \mathrm{KNN}_{n=16}\left(\mathbf{p}_{t,k}^0, \mathbf{X}_{t+\delta}^{(\ell)}\right). \tag{17}$$

For each retrieved neighbor point $i \in \mathcal{N}_{t,k}^{\ell,\delta}$, we extract an evidence vector containing four complementary components:

1. **Relative Displacement**: $\mathbf{r}_{\ell,i}^\delta = \mathbf{x}_{\ell,i}^\delta - \mathbf{p}_{t,k}^0 \in \mathbb{R}^3$, preserving local geometric surface offsets;
2. **Point Feature**: $\mathbf{f}_{\ell,i}^\delta \in \mathbb{R}^d$, capturing local neighborhood context;
3. **Temporal Embedding**: $\mathbf{e}_\delta \in \mathbb{R}^d$, indicating relative temporal source $\delta \in \{-1, 0, +1\}$;
4. **Projected Cosine Similarity**: $\rho_{\ell,i}^\delta = \frac{\langle W_q \mathbf{a}_{t,k}^0, W_f^{(\ell)} \mathbf{f}_{\ell,i}^\delta \rangle}{\|W_q \mathbf{a}_{t,k}^0\|_2 \|W_f^{(\ell)} \mathbf{f}_{\ell,i}^\delta\|_2} \in [-1, 1]$, mitigating cross-limb confusion by evaluating feature consistency.

A multi-scale token encoder ϕ_ℓ^b projects this concatenated vector into a unified evidence token:

$$\mathbf{u}_{\ell,i}^b = \phi_\ell^b\left(\left[\mathbf{f}_{\ell,i}^\delta, \mathbf{r}_{\ell,i}^\delta, \mathbf{e}_\delta, \rho_{\ell,i}^\delta\right]\right) \in \mathbb{R}^d, \qquad b \in \{\mathrm{geo}, \mathrm{mot}\}. \tag{18}$$

Collecting all scale levels yields $\mathbf{E}_\ell^\delta \in \mathbb{R}^{n \times d}$. We bifurcate these tokens into two structured evidence views:

$$\mathcal{H}_{t,k}^- = \mathrm{Concat}_{\ell=1}^L \mathbf{E}_\ell^{-1} \in \mathbb{R}^{Ln \times d}, \quad \mathcal{H}_{t,k}^+ = \mathrm{Concat}_{\ell=1}^L \mathbf{E}_\ell^{+1} \in \mathbb{R}^{Ln \times d}, \tag{19}$$

$$\mathcal{H}_{t,k}^{\mathrm{geo}} = \mathrm{Concat}_{\ell=1}^L [\mathbf{E}_\ell^{-1}; \mathbf{E}_\ell^0; \mathbf{E}_\ell^{+1}] \in \mathbb{R}^{3Ln \times d}. \tag{20}$$

The separation of $\mathcal{H}^-$ and $\mathcal{H}^+$ explicitly maintains directional temporal motion flow, while the joint aggregation in $\mathcal{H}^{\mathrm{geo}}$ consolidates multi-frame complementary geometric observations.

A.3.3 AMBER Formation and Bounded Refinement

To distill the organized evidence into the final representation $\mathcal{A}_t$, we generate separate geometry and motion queries:

$$\mathbf{Q}_t^{\mathrm{g}} = \phi_{\mathrm{feat}}^{\mathrm{g}}(\mathbf{A}_t^0) + \phi_{\mathrm{pos}}^{\mathrm{g}}(\mathbf{P}_t^0), \qquad \mathbf{Q}_t^{\mathrm{m}} = \phi_{\mathrm{feat}}^{\mathrm{m}}(\mathbf{A}_t^0) + \phi_{\mathrm{pos}}^{\mathrm{m}}(\mathbf{P}_t^0). \tag{21}$$

Motion Branch. To independently preserve backward and forward kinematics, $\mathbf{Q}_t^{\mathrm{m}}$ attends to $\mathcal{H}^-$ and $\mathcal{H}^+$ through multi-head cross-attentions with distinct weights:

$$\mathbf{u}_{t,k}^{-1} = \mathrm{MCA}_{\mathrm{prev}}(\mathbf{q}_{t,k}^{\mathrm{m}}, \mathcal{H}_{t,k}^-), \qquad \mathbf{u}_{t,k}^{+1} = \mathrm{MCA}_{\mathrm{next}}(\mathbf{q}_{t,k}^{\mathrm{m}}, \mathcal{H}_{t,k}^+). \tag{22}$$

A fusion network ϕ_{fuse} combines the directional features into the anchor motion descriptor:

$$\mathbf{m}_{t,k} = \phi_{\mathrm{fuse}}\left([\mathbf{u}_{t,k}^{-1};\, \mathbf{u}_{t,k}^{+1}]\right) \in \mathbb{R}^d. \tag{23}$$

Geometry Branch. The geometry query $\mathbf{Q}_t^{\mathrm{g}}$ attends to the unified multi-frame evidence $\mathcal{H}^{\mathrm{geo}}$ via multi-head cross-attention, obtaining $\mathbf{u}_{t,k}^{\mathrm{geo}} = \mathrm{MCA}_{\mathrm{geo}}(\mathbf{q}_{t,k}^{\mathrm{g}}, \mathcal{H}_{t,k}^{\mathrm{geo}})$. We update the initial anchor positions and features via zero-initialized residual heads:

$$\mathbf{g}_{t,k} = \mathbf{a}_{t,k}^0 + \phi_{\Delta a}(\mathbf{u}_{t,k}^{\mathrm{geo}}), \qquad \mathbf{p}_{t,k} = \mathbf{p}_{t,k}^0 + \alpha \tanh\left(\phi_{\Delta p}(\mathbf{u}_{t,k}^{\mathrm{geo}})\right), \tag{24}$$

where $\alpha = 0.25\,\mathrm{m}$ bounds the maximum spatial displacement, ensuring stable convergence without exploding coordinate updates under extreme point sparsity.

Finally, an auxiliary local folding module generates dense surface points $\mathbf{P}_t^{\mathrm{com}} \in \mathbb{R}^{1024\times 3}$ by deforming regular 2D grids around each anchor $(\mathbf{p}_{t,k}, \mathbf{g}_{t,k})$, which is supervised via Chamfer distance $\mathrm{CD}_2(\mathbf{P}_t^{\mathrm{com}}, \widehat{\mathcal{V}}_t)$ during training to reinforce anatomical fidelity.

B Additional Qualitative Results

Sequential Visual Comparisons. To evaluate temporal consistency and robustness, we visualize consecutive-frame reconstructions on FreeMotion, LiDARH26M, NoiseMotion, and SLOPER4D in Fig. 9. Across diverse motions and observation conditions, AMBER produces coherent body sequences that closely follow the ground truth. In contrast, competing methods exhibit noticeable pose deviations, temporal inconsistencies, or severe reconstruction failures, particularly under sparse and noisy observations. These results demonstrate that AMBER can effectively exploit cross-frame surface evidence to recover stable and accurate human motion from incomplete LiDAR inputs.

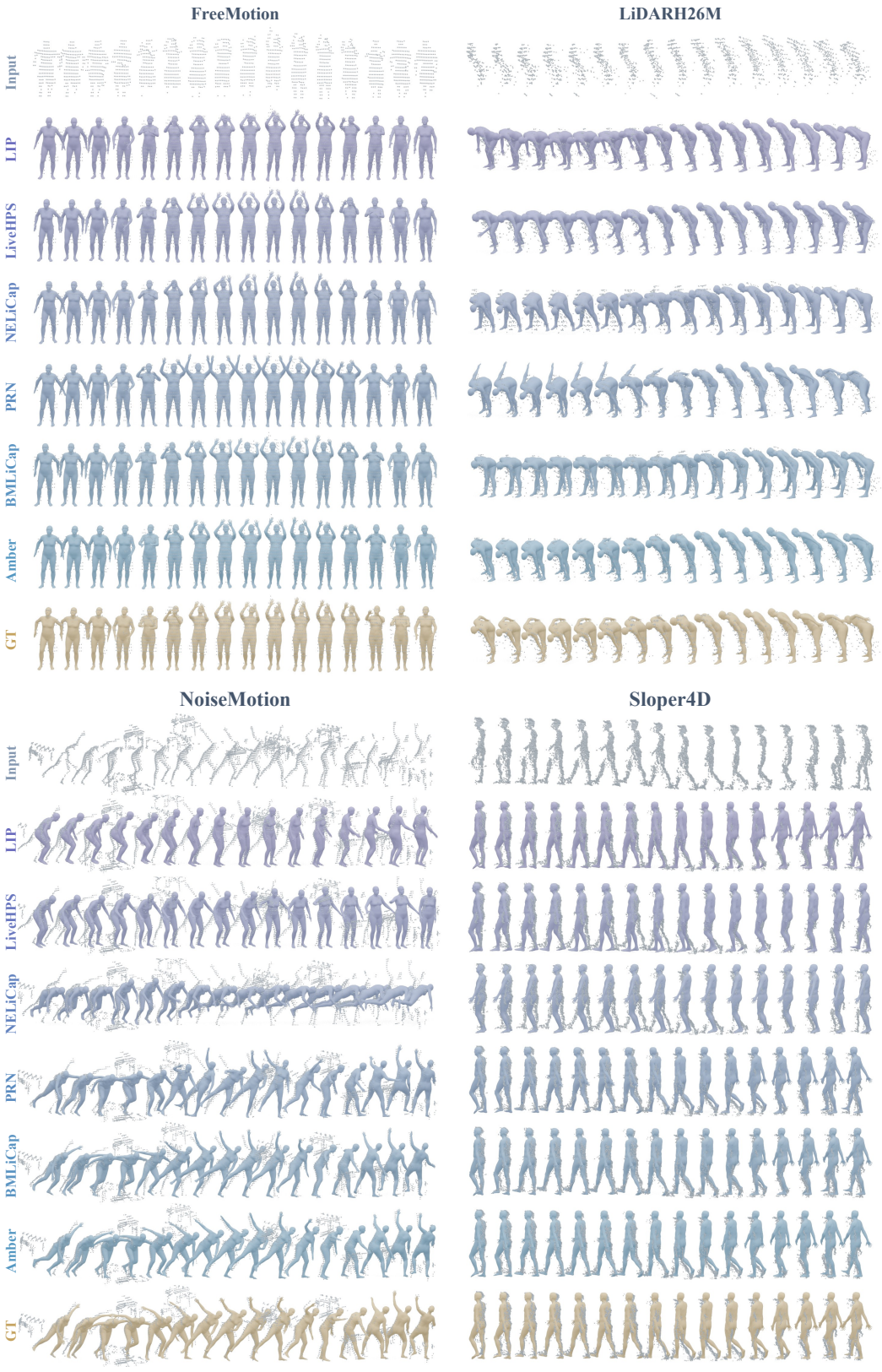

Figure 9: Sequential visualization.