Bézier Degradation Modeling for LiDAR-based Human Motion Capture

Supplementary Material

A. Method Details

Detailed Procedure for TAD We provide the detailed procedure for Trajectory-Aware Bézier Degradation (TAD) in Algorithm 1. Given the finest cubic Bézier chain for joint k, i.e., $\{\mathbf{J}_t^{(k)}\}_{t=0}^{T}$ with control points $\mathbf{C}_{t,1}^{(k)}, \mathbf{C}_{t,2}^{(k)}$, and a step size s, we first resample the timestamps $\mathcal{T}_s$ from 0 to T with step size s. The new joint positions $\widetilde{\mathbf{J}}_i^{(k)}$ are set as the original joint positions at the resampled timestamps. We then extract the tangents $\widehat{\mathbf{d}}_i^{(k)}$ at each new joint position. For each segment between two consecutive resampled timestamps, we sample m points along the segment. If $s < 4$, we upsample the points using the Bézier function to avoid degenerate cases; otherwise, we directly use the keypoints from the finest curve. We then compute the Bernstein basis polynomials and set up a linear system to solve for the new control point offsets $\ell_{i,2}$ and $\ell_{i+1,1}$. Finally, we update the control points accordingly and return the updated Bézier chain.

This ensures that the Bézier chain constructed for each s is as similar as possible to the original curve. When refining the curve at the next level, only minor adjustments are required.

Block-wise Causal Mask in TMT. For the block-wise causal mask implementation of the TMT, we follow [1, 2]. Assume that the TMT receives a concatenated token sequence:

$$\mathbf{Z} = [\mathbf{F}_{\mathcal{P}}\,;\,\mathbf{E}_1\,;\,\mathbf{E}_2\,;\,\cdots;\,\mathbf{E}_L] \in \mathbb{R}^{N\times D},$$

where $\mathbf{F}_{\mathcal{P}}$ denotes the LiDAR feature tokens, and $\mathbf{E}_l \in \mathbb{R}^{M_{s_l}\times D}$ denotes the motion tokens at level l. Let the index ranges of these segments be denoted by

$$\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_L, \qquad \bigcup_{l=0}^{L} \mathcal{I}_l = \{1, \ldots, N\}, \qquad \mathcal{I}_i \cap \mathcal{I}_j = \varnothing.$$

We construct a binary attention mask $\mathbf{A} \in \{0, -\infty\}^{N\times N}$ applied before the softmax operation. The mask enforces that tokens at level l may attend only to LiDAR tokens and all coarser motion levels:

$$\mathbf{A}_{i,j} = \begin{cases} 0, & \text{if } j \in \mathcal{I}_0 \cup \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_{l_i}, \\ -\infty, & \text{otherwise,} \end{cases}$$

where l_i denotes the level index of token i, i.e., the unique l such that $i \in \mathcal{I}_l$ (with $\mathcal{I}_0$ corresponding to $\mathbf{F}_{\mathcal{P}}$). Thus, finer-level tokens are prevented from attending to tokens belonging to the same or finer levels, while LiDAR tokens remain fully visible to all levels.

Algorithm 1 Trajectory-Aware Bézier Degradation

Require: Finest cubic Bézier chain for joint k: $\{\mathbf{J}_t^{(k)}\}_{t=0}^{T}$ with controls $\mathbf{C}_{t,1}^{(k)}, \mathbf{C}_{t,2}^{(k)}$; step size s

Ensure: Updated chain $\{\widetilde{\mathbf{J}}_i^{(k)}, \widetilde{\mathbf{C}}_{i,1}^{(k)}, \widetilde{\mathbf{C}}_{i,2}^{(k)}\}$

Resample $\mathcal{T}_s \leftarrow \{t_0{=}0, \ldots, t_{M-1}{=}T\}$.
Set $\widetilde{\mathbf{J}}_i^{(k)} \leftarrow \mathbf{J}_{t_i}^{(k)}$.
Extract tangents $\widehat{\mathbf{d}}_i^{(k)}$ at each $\widetilde{\mathbf{J}}_i^{(k)}$.
for each $t_i \in \mathcal{T}_s$ **do**
  $m \leftarrow \max(4, s)$
  $\mathbf{t} \leftarrow \mathrm{linspace}(0, 1, m)$
  if $s < 4$ **then** ▷ Upsample to avoid degenerate
    $\widehat{\mathbf{X}} \leftarrow \mathrm{Bezier}(\{\mathbf{J}_t^{(k)}, \mathbf{C}_{t,1}^{(k)}, \mathbf{C}_{t,2}^{(k)}\}_{t=t_i}^{t_{i+1}};\ \mathbf{t})$
  else ▷ Use keypoints from finest curve
    $\widehat{\mathbf{X}} \leftarrow \{\mathbf{J}_t^{(k)}\}_{t=t_i}^{t_{i+1}}$
  end if
  $\mathbf{b}_0 = (1-\mathbf{t})^3$, $\mathbf{b}_1 = 3(1-\mathbf{t})^2\mathbf{t}$
  $\mathbf{b}_2 = 3(1-\mathbf{t})\mathbf{t}^2$, $\mathbf{b}_3 = \mathbf{t}^3$
  $\mathbf{B} \leftarrow (\mathbf{b}_0{+}\mathbf{b}_1) \odot \widetilde{\mathbf{J}}_i^{(k)} + (\mathbf{b}_2{+}\mathbf{b}_3) \odot \widetilde{\mathbf{J}}_{i+1}^{(k)}$
  $\mathbf{q} \leftarrow \mathrm{vec}\left(\widehat{\mathbf{X}} - \mathbf{B}\right)$ ▷ reshape to $(3m)$
  $\mathbf{X} \leftarrow \left[\, \mathbf{b}_1 \odot \widehat{\mathbf{d}}_i^{(k)} \ -\mathbf{b}_2 \odot \widehat{\mathbf{d}}_{i+1}^{(k)} \,\right]$ ▷ reshape to $(3m, 2)$
  $[\ell_{i,2}, \ell_{i+1,1}]^\top \leftarrow \left(\mathbf{X}^\top\mathbf{X} + \varepsilon I_2\right)^{-1}\mathbf{X}^\top\mathbf{q}$
  $\widetilde{\mathbf{C}}_{i+1,1}^{(k)} \leftarrow \widetilde{\mathbf{J}}_{i+1}^{(k)} - \ell_{i+1,1}\,\widehat{\mathbf{d}}_{i+1}^{(k)}$
  $\widetilde{\mathbf{C}}_{i,2}^{(k)} \leftarrow \widetilde{\mathbf{J}}_i^{(k)} + \ell_{i,2}\,\widehat{\mathbf{d}}_i^{(k)}$
end for
return $\{\widetilde{\mathbf{J}}_i^{(k)}, \widetilde{\mathbf{C}}_{i,1}^{(k)}, \widetilde{\mathbf{C}}_{i,2}^{(k)}\}$

During self-attention, the masked attention logits are computed as

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}} + \mathbf{A}\right)\mathbf{V}.$$

This block-wise causal masking ensures a strict information hierarchy, where coarse-level motion provides global trends, and finer levels progressively refine the motion trajectories under guidance from both LiDAR observations and previously reconstructed scales.

We provide a comparison with other forms of masks in Sec.B.

B. Additional Experiments

In this section, we provide additional experimental results, including more ablation experiments on some components and a further exploration of the generalizability of our method.

Table 5. Action-wise 3D pose estimation errors on Human3.6M. We report the MPJPE (mm) for each action category.

Method	Dire.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
MotionBERT [3]	36.3	38.7	38.6	33.6	42.1	50.1	36.2	35.7	50.1	56.6	41.3	37.4	37.7	25.6	26.5	39.2
Bezire	34.7	36.8	37.3	31.6	38.7	45.5	36.0	33.7	47.3	52.7	38.4	36.6	24.1	34.6	24.8	$36.9_{\downarrow 2.3}$
Bezire + TAD	**34.5**	**36.9**	**37.3**	**31.5**	**38.6**	**45.6**	**36.2**	**33.4**	**47.3**	**52.5**	**38.5**	**36.4**	**24.0**	34.5	**24.6**	$\mathbf{36.8}_{\downarrow 2.4}$

Table 6. Cross-dataset generalization performance of BMLiCap. To ensure compatibility between datasets, we use only a subset of the 32-frame time window for each test dataset, so the results differ slightly from the full test.

(a) FreeMotion ↔ NoiseMotion

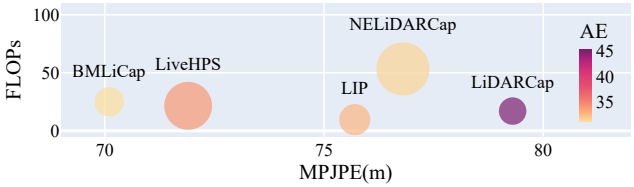
Train	FreeMotion			NoiseMotion		
	MPJPE	MPVPE	AE	MPJPE	MPVPE	AE
FreeMotion	45.9	56.7	22.2	137.0	177.8	43.1
NoiseMotion	80.1	98.5	29.4	37.1	47.3	23.8

(b) LiDARHuman26M ↔ SLOPER4D

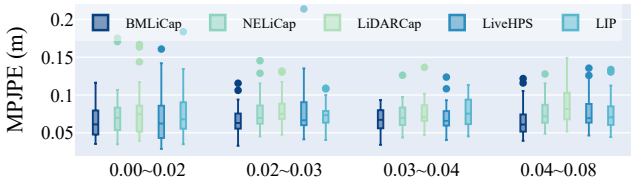
Train	LiDARHuman26M			SLOPER4D		
	MPJPE	MPVPE	AE	MPJPE	MPVPE	AE
LiDARH26M	67.9	87.2	28.3	173.1	207.2	34.3
SLOPER4D	189.7	240.2	48.5	36.2	43.2	20.6

Table 7. Ablation study of attention masking strategies in TMT.

Mode	MPJPE	MPVPE	AE
Global	70.9	90.1	28.8
(a) Diag	68.1	87.0	29.1
(b) Level Independent	67.6	86.1	29.0
(c) Sliding Window	68.2	87.1	29.3
Block-wise Causal	**66.8**	**85.4**	**28.8**



(a) Efficiency analysis. Bubble size indicates the total size of parameters.



(b) Comparison on different joint jerk magnitude (m/s^3).

Figure 10. Efficiency and error distribution analysis of BMLiCap.

Ablation on Attention Masking Strategies. To investigate which information flows are important, we compare it with several common mask structures. We make the following modifications to the masking strategy: **(a)** add unidirectional temporal constraints; **(b)** remove dependencies between levels but keep visibility to point clouds; **(c)** use a sliding window to introduce asymmetric bidirectional interactions between levels. The results are shown in Tab.7. The gains of the block-wise causal mask indicate that enforcing stage-wise temporal ordering and restricting information flow across levels provides a more effective inductive bias for reconstructing coherent motion.

Cross-Dataset Generalization. The cross-dataset experiments evaluate the generalization ability of BMLiCap by training the model on one dataset and directly testing it on the other three without any fine-tuning. As shown in Tab.6, overall, the results show that BMLiCap remains reasonably robust under domain shift, while cross-dataset generalization still poses a challenging setting.

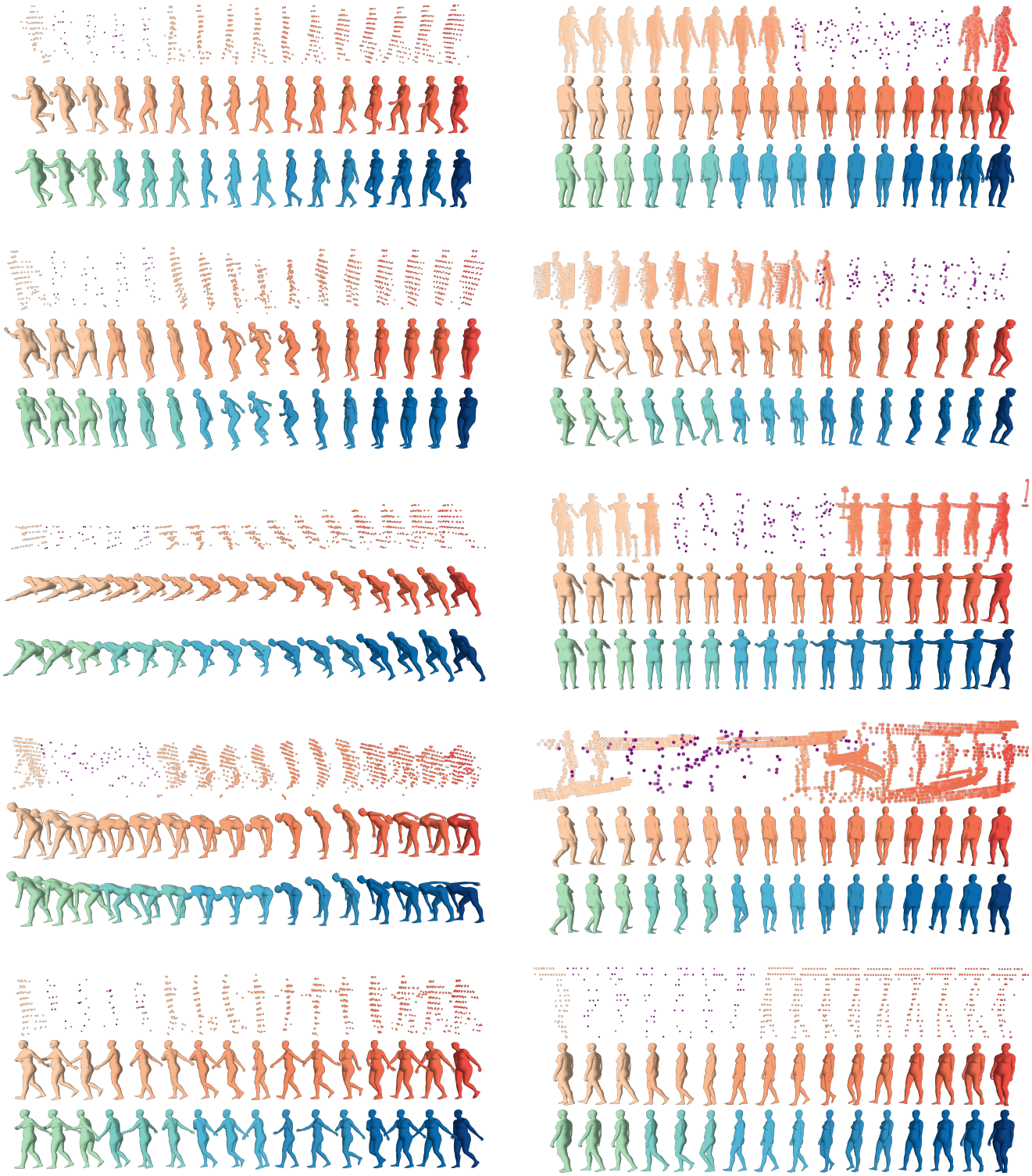
Other Input Modalities. To evaluate the generalizability of our method, we also implement BMLiCap on the commonly used RGB modality. As shown in Tab. 5, based on MotionBERT [3], we implement a long-sequence motion predictor using the configuration $\mathcal{S} = \{40, 81, 243\}$. Our method delivers consistent performance improvements.

Efficiency Analysis. We compare the efficiency of BMLiCap with SOTA methods in Fig. 10a, plotting MPJPE and AE against model FLOPs, with bubble size proportional to parameter count. BMLiCap achieves a better accuracy–efficiency trade-off: AE decreases substantially with only a modest increase in FLOPs and parameters, demonstrating the benefit of multi-scale Bézier modeling.

Motion Complexity Analysis. To assess robustness under varying motion dynamics, we bin test sequences by their average joint jerk magnitude ($\mathrm{m/s^3}$) and report MPJPE for each bin in Fig. 10b. Joint jerk—the third derivative of position—directly reflects the abruptness of motion, making it a natural indicator of reconstruction difficulty. As jerk increases, all methods exhibit higher errors; however, BMLiCap degrades more gracefully than competing approaches. This advantage is attributed to the multi-scale Bézier representation, which explicitly models smooth trajectory priors at coarse levels and reserves fine-level capacity for rapid local deformations, thereby maintaining accuracy even for highly dynamic sequences.

C. Additional Visualization

In this section, we demonstrate more sequential results of our method under complex actions in Fig.11. High-resolution details are visible when the figure is magnified.



(a) LiDARHuman26M (b) NoiseMotion

Figure 11. Randomly selected visualization results of motion sequences from two datasets. For each group, first row: input point clouds with manually incorporated occlusion; second row: ground-truth motion; third row: our reconstructed motion.

References

[1] Zhefei Gong, Pengxiang Ding, Shangke Lyu, Siteng Huang, Mingyang Sun, Wei Zhao, Zhaoxin Fan, and Donglin Wang. CARP: Visuomotor Policy Learning via Coarse-to-Fine Autoregressive Prediction, 2025. 0

[2] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Adv. Neural Inf. Process. Syst.*, 37:84839–84865, 2024. 0

[3] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *ICCV*, pages 15085–15099, 2023. 1