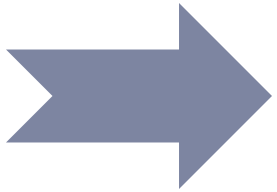
PS-Mamba: Spatial-Temporal Graph **Mamba** for **P**ose **S**equence Refinement

Haoye Dong, Gim Hee Lee
School of Computing, National University of Singapore

Task



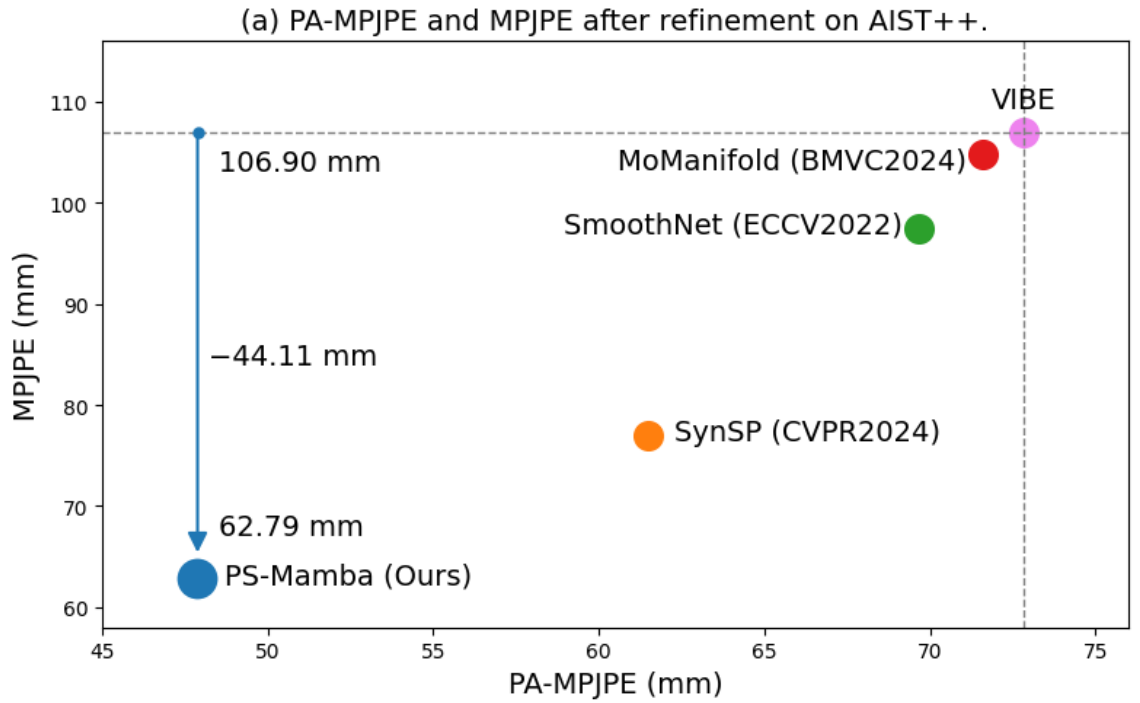
Input Pose Sequence

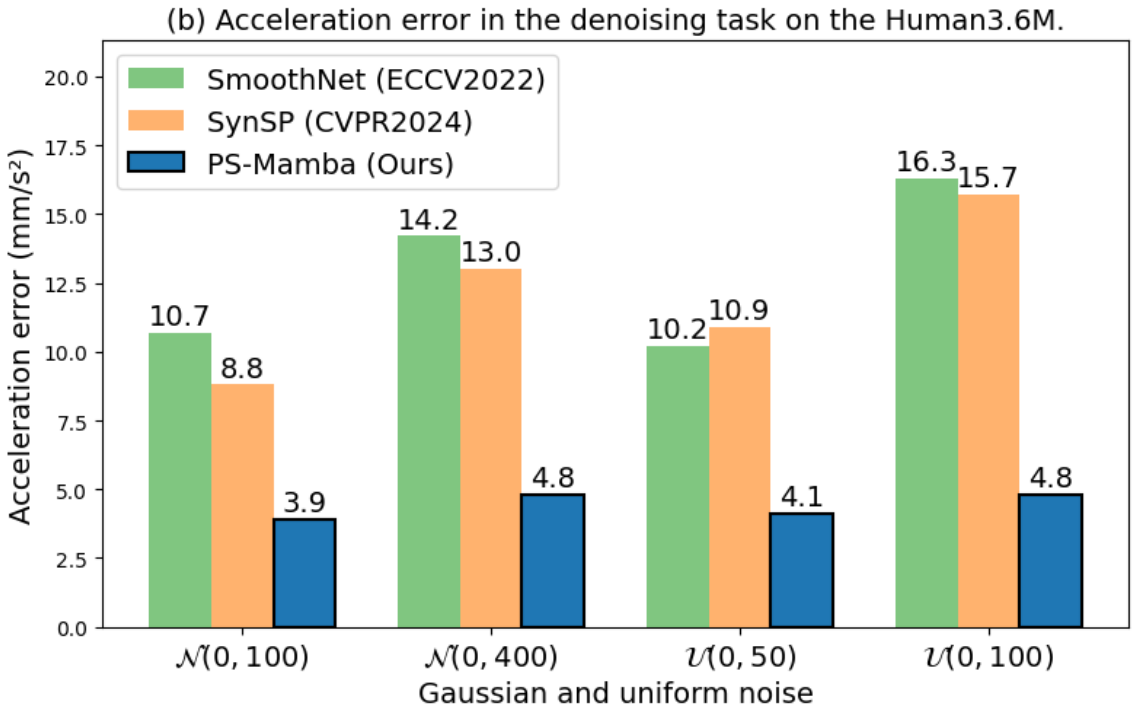
Refined Pose Sequence

Given human pose sequence, we aim to refine them to achieve smooth and accurate sequences.

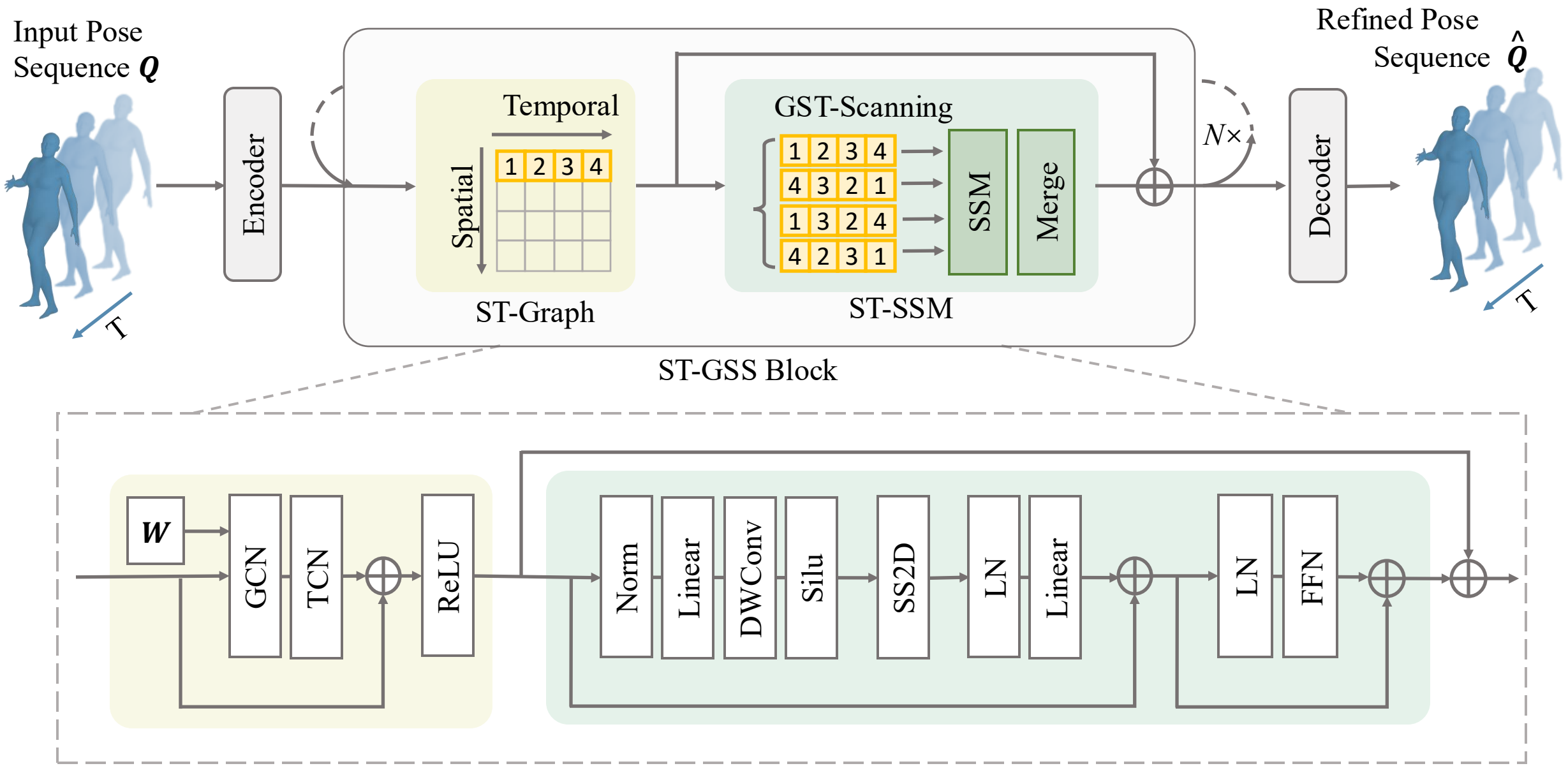
Compare with SOTAs

- We have significant improvements in PA-MPJPE and MPJPE on AIST++.
- We achieve lowest acceleration error on Human3.6M.

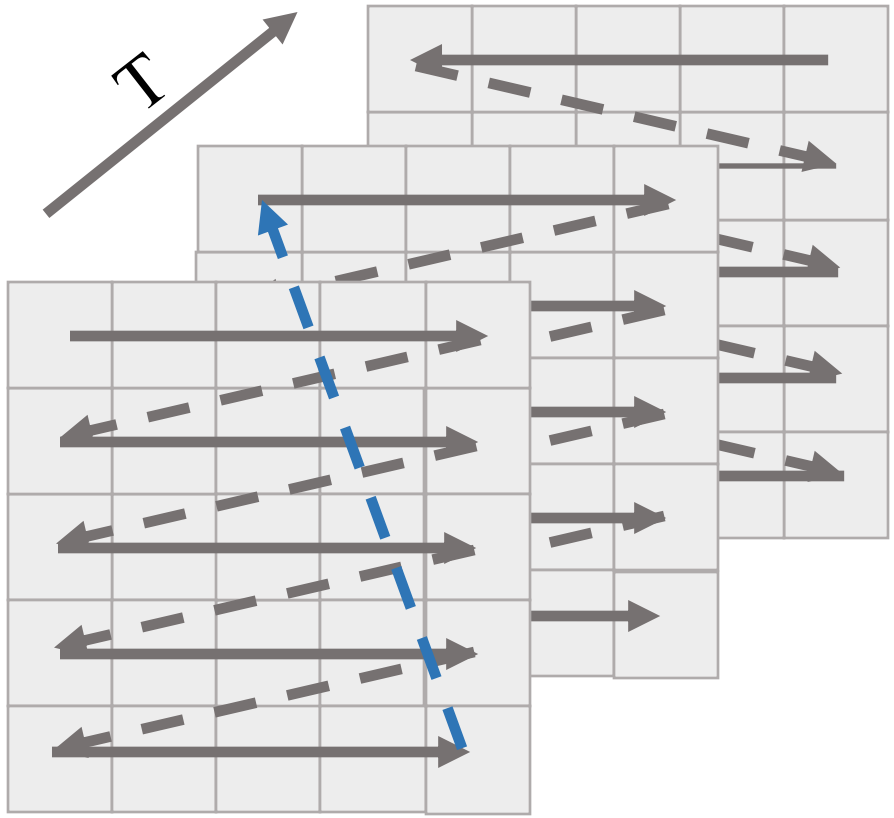




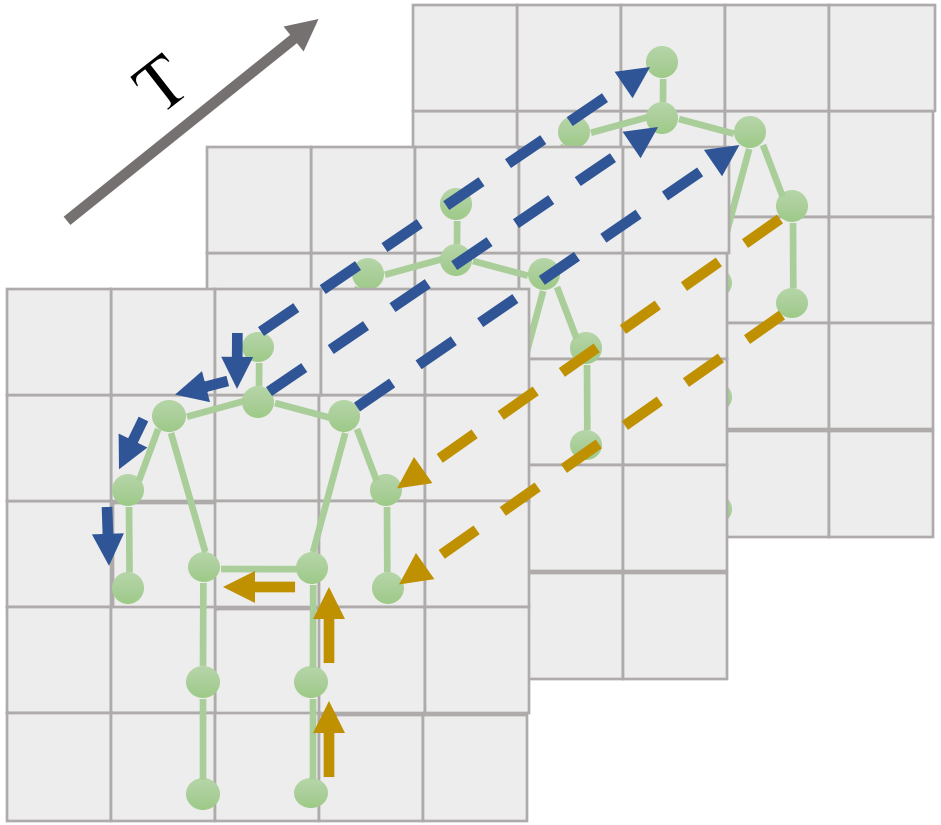
Framework


Input Pose Sequence $\boldsymbol{Q}$
T
Encoder
Temporal
Spatial
1 2 3 4
ST-Graph
GST-Scanning
1 2 3 4
4 3 2 1
1 3 2 4
4 2 3 1
SSM
Merge
ST-SSM
$N\times$
Decoder
Refined Pose Sequence $\hat{\boldsymbol{Q}}$
T
ST-GSS Block
$\boldsymbol{W}$
GCN
TCN
ReLU
Norm
Linear
DWConv
Silu
SS2D
LN
Linear
LN
FFN

Motivation



(a) ST-Scanning



(b) GST-Scanning (Ours)

(a) Traditional ST-Scanning (SpatialTemporal Scanning) involves redundant tokens, failing to capture the dependencies between joints.

(b) Our GST-Scanning (Graph-guided Spatial-Temporal Scanning), under the guidance of the human joint relationship graph, effectively captures joint dependencies while preserving temporal coherence.

Experiments

Comparison in Different Representations

Method	WS	Human3.6M / 2D			Human3.6M / 3D			AIST++ / SMPL		
		MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓
Input	N/A	9.42	7.64	1.54	54.55	42.20	19.17	107.72	74.40	33.20
One-Euro [3]	1	10.69	7.98	0.34	55.20	42.73	3.80	108.97	75.27	14.70
Gaussian1d [48]	32	9.37	7.56	0.51	53.67	41.60	2.43	104.84	72.18	10.05
Savitzky-Golay [29]	32	9.35	7.55	0.17	53.48	41.19	1.34	104.58	72.30	6.07
SmoothNet [50] (ECCV22)	32	9.25	7.57	0.15	52.72	40.92	1.03	103.00	71.19	5.72
SynSP† [41] (CVPR24)	32	12.72	6.69	0.82	62.86	48.11	0.96	89.84	63.68	7.42
PS-Mamba (Ours)	32	7.31	5.74	**0.14**	49.60	38.36	**0.92**	74.04	52.98	**5.51**
SynSP [41] (CVPR24)	8	8.13	6.09	0.15	51.36	40.13	1.02	84.63	59.02	6.08
PS-Mamba (Ours)	8	7.52	5.86	0.15	51.22	39.63	1.72	**72.54**	**52.95**	6.88
SynSP [41] (MV, CVPR24)	8	7.62	5.64	0.15	41.78	33.32	0.98	-	-	-
PS-Mamba (MV, Ours)	8	**6.90**	**4.53**	0.16	**28.63**	**21.83**	1.0	-	-	-

Table 1. Comparison with SOTAs on Human3.6M / 2D, Human3.6M / 3D , and AIST++ / SMPL , respectively. WS means window size. MV denotes Multi-View. Results with † are reproduced from the original paper. Lower values indicate better performance.

PS-Mamba outperforms state-of-the-art methods in 2D, 3D, and SMPL representations, excelling in pose accuracy and temporal smoothness.

Motion Denoising Task

Method	WS	Human3.6M Dataset								CMU-Mocap Dataset							
		$\mathcal{N}(0,100)$		$\mathcal{N}(0,400)$		$\mathcal{U}(0,50)$		$\mathcal{U}(0,100)$		$\mathcal{N}(0,100)$		$\mathcal{N}(0,400)$		$\mathcal{U}(0,50)$		$\mathcal{U}(0,100)$	
		M↓	A↓	M↓	A↓	M↓	A↓	M↓	A↓	M↓	A↓	M↓	A↓	M↓	A↓	M↓	A↓
Noisy input	N/A	65.6	160.3	131.1	320.5	94.8	231.3	189.5	462.7	61.2	143.8	122.5	287.6	88.7	207.9	177.4	415.8
GFPose [6] (CVPR23)	1	42.8	-	64.6	-	50.9	-	89.4	-	40.4	-	60.0	-	45.6	-	84.4	-
SmoothNet [50] (ECCV22)	32	42.4	10.7	57.9	14.2	49.4	10.2	74.3	16.3	35.1	4.3	41.5	4.9	37.7	4.2	44.9	5.1
SynSP [41] (CVPR24)	8	37.6	8.8	56.1	13.0	45.6	10.9	69.4	15.7	20.6	3.7	28.0	4.4	24.4	4.1	32.4	4.6
PS-Mamba (Ours)	8	36.7	8.7	**54.3**	12.9	44.5	10.8	63.6	13.1	11.7	2.5	**17.3**	3.6	**10.7**	2.4	**13.6**	2.9
PS-Mamba (Ours)	32	**32.6**	**3.9**	54.9	**4.8**	**39.1**	**4.1**	**53.8**	**4.8**	**12.7**	**1.3**	**17.3**	**1.5**	11.6	**1.2**	14.2	**1.3**

Table 2. Comparison of 3D motion denoising performance on the Human3.6M [15] and CMU-Mocap [37] under various noise types: $\mathcal{N}(0,100)$, $\mathcal{N}(0,400)$, $\mathcal{U}(0,50)$, and $\mathcal{U}(0,100)$. Results are shown in terms of MPJPE (M↓) and Accel (A↓).

PS-Mamba outperforms SOTAs in accuracy and smoothness for denoising task, demonstrating strong robustness across various noise types and levels.

Comparison with Video-based SOTAs

Dataset	Method	MPJPE↓	Accel↓
AIST++	VIBE [17]	106.9	31.60
	VIBE [17] + MoManifold [7]	104.85	6.34
	VIBE [17] + SmoothNet [50]	97.47	4.15
	VIBE [17] + SynSP [41]	77.00	4.32
	VIBE [17] + PS-Mamba (Ours)	**62.79**	**3.90**
	TCMR [4]	106.72	6.4
	TCMR [4] + SmoothNet [50]	105.51	4.24
	TCMR [4] + SynSP [41]	-	-
	TCMR [4] + PS-Mamba (Ours)	**80.80**	**3.97**
3DPW	PARE [18]	79.00	25.60
	PARE [18] + SmoothNet [50]	78.10	5.91
	PARE [18] + SynSP [41]	76.20	6.16
	PARE [18] + PS-Mamba (Ours)	**75.43**	**5.84**

Table 3. Comparison with different video-based estimators, where "+" denotes the integration of estimators with pose refinement.

Efficiency Analysis

Method	Params(M)	FLOPs(M) / Frame	MPJPE↓	Accel↓
SmoothNet [50]	<u>0.69</u>	<u>1.10</u>	35.1	4.3
PS-Mamba-T (Ours)	**0.64**	**0.44**	<u>14.8</u>	<u>1.6</u>
SynSP [41]	2.79	3.76	20.6	3.7
PS-Mamba (Ours)	1.43	3.52	**12.7**	**1.3**

Table 5. Comparison of Params and FLOPs on CMU-Mocap [37]. PS-Mamba-T denotes the tiny version of our PS-Mamba.

Both PS-Mamba and PS-Mamba-T provide superior performance compared to SmoothNet and SynSP.

Ablation Study

Method	MPJPE↓	PA-MPJPE↓	Accel↓
Input	54.55	42.20	19.17
w/o ST-GSS	370.23	181.88	3.01
w/o ST-SSM	51.97	40.27	1.25
w/o ST-Graph	50.75	38.52	0.97
w/o $\boldsymbol{W}$	50.74	39.40	0.94
w/o Temporal	50.27	38.79	0.98
w/o Residual	50.21	39.24	0.94
w/o mpjpe loss	446.50	181.29	0.98
w/o pa-mpjpe loss	51.02	39.28	0.95
w/o accel loss	50.63	39.49	0.95
PS-Mamba (Full)	**49.60**	**38.36**	**0.92**

Table 6. Ablation study on Human3.6M [15] .

Impact of Window Size

	Input	4	8	16	30	40	50
MPJPE	54.55	53.26	52.79	50.69	51.53	**50.13**	51.91
PA-MPJPE	42.20	40.53	40.25	39.53	38.98	**38.93**	39.16
Accel	19.17	3.22	1.74	1.19	0.94	0.91	**0.90**

Table 4. Impact of window size on Human3.6M.

Increasing the window size generally enhances model performance, particularly in terms of temporal consistency and smoothness

Visual Comparison

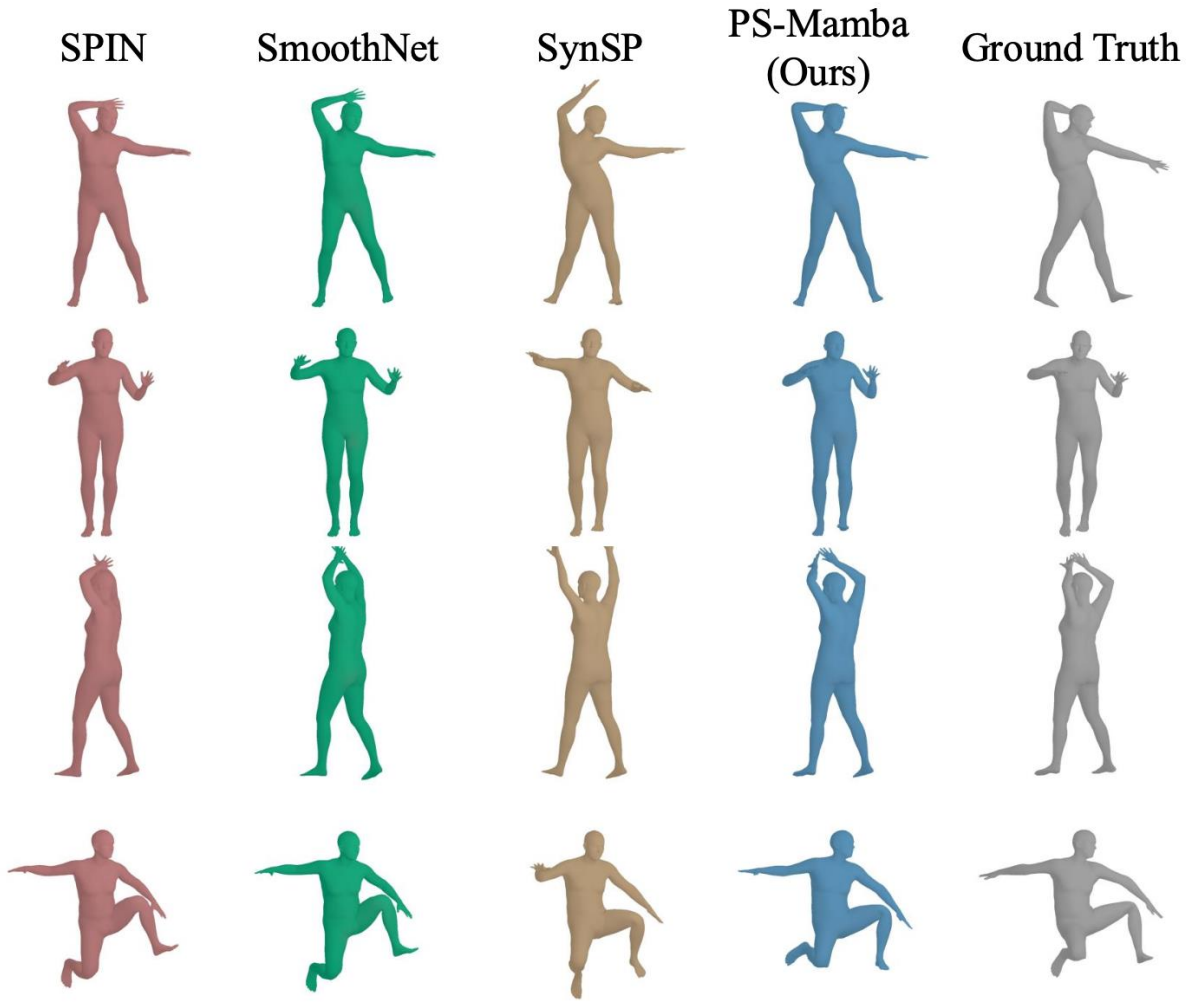
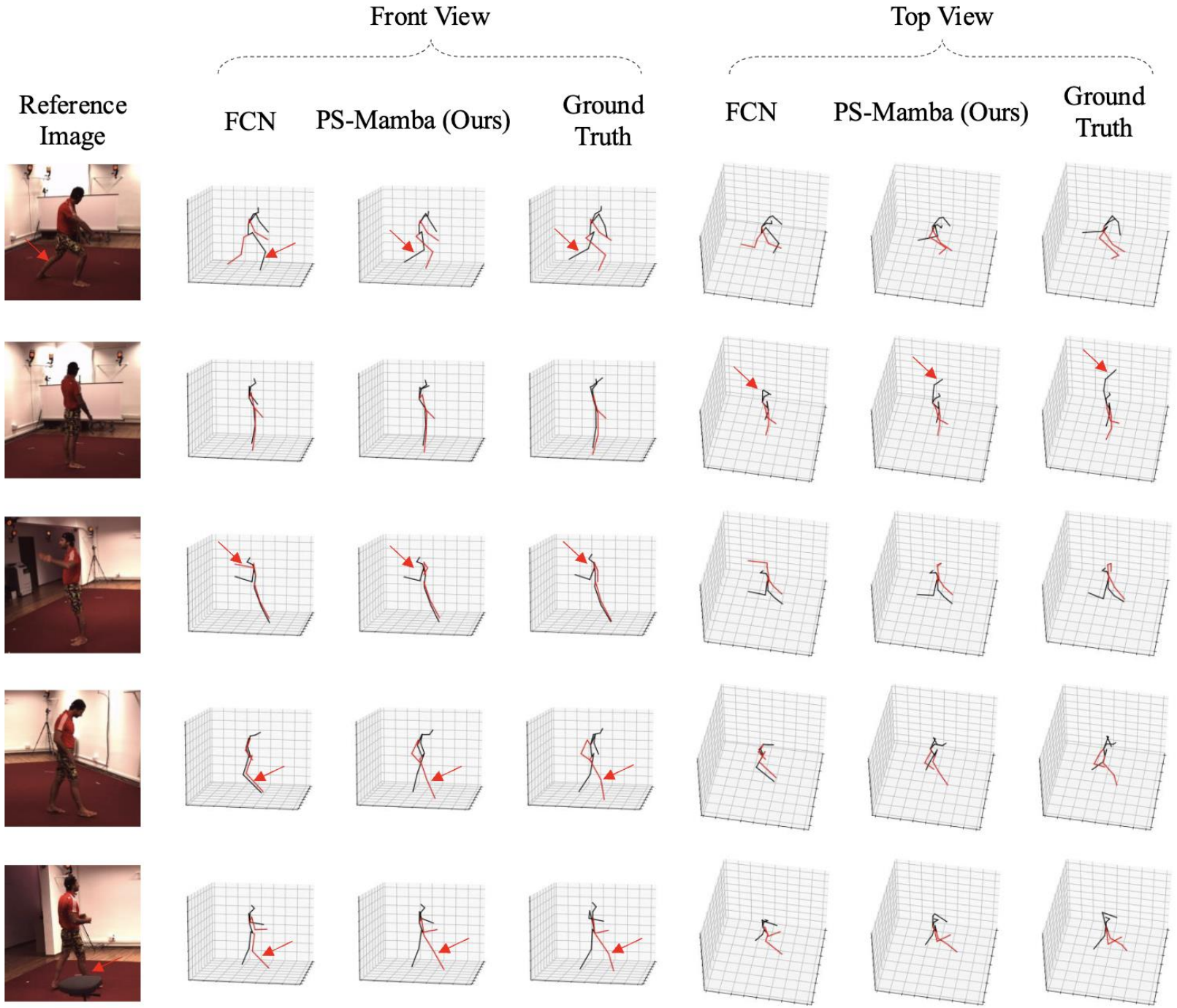


Figure 4. Qualitative comparisons on AIST++ [23, 36]. Our PS-Mamba achieves accurate 3D mesh and outperforms SynSP [41] and SmoothNet [50], all initialized by SPIN [19].

Visual Comparison

In the first row,
FCN misidentifies the left and right feet, while our results align closely with Ground Truth.

In the fifth row,
our model performs robustly even under occlusion.



Failure Cases

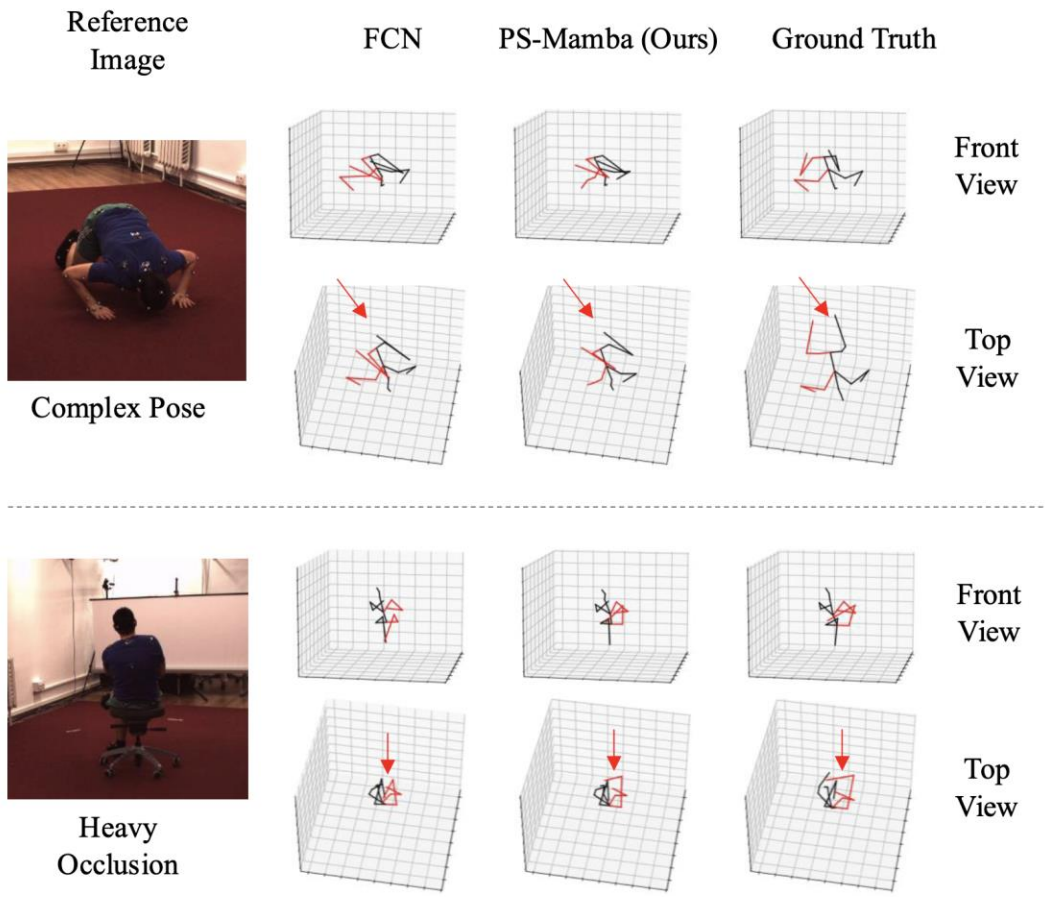
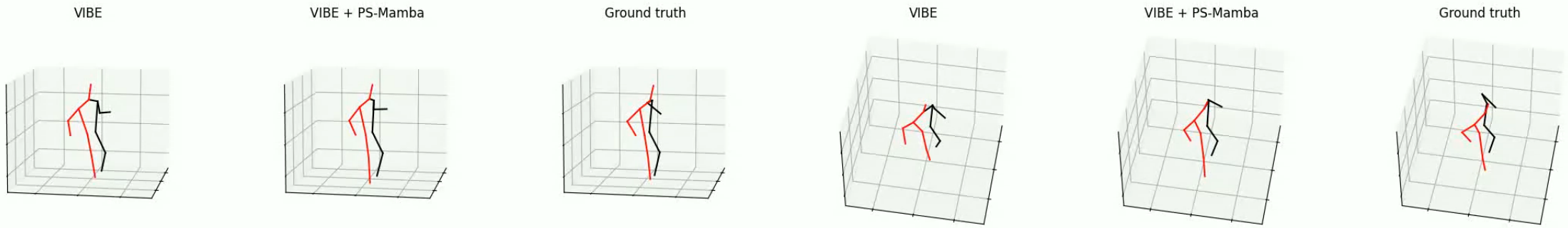
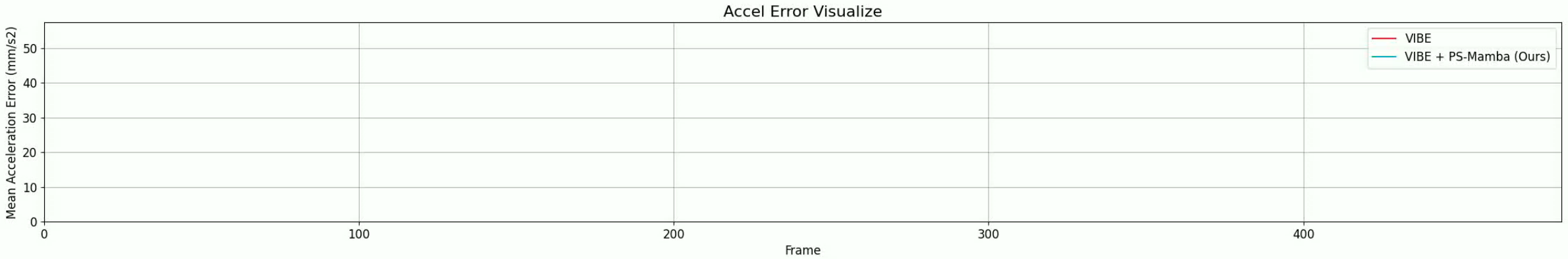
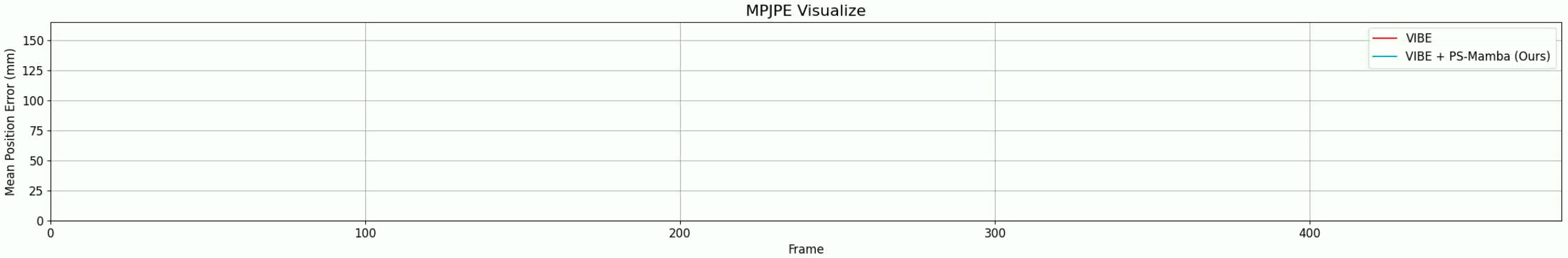


Figure 1. **Failure cases**. When faced with complex poses or heavy occlusions, both our method and the compared method fail to achieve the accurate human pose.

Accel and MPJPE Analysis







Demo Videos

SPIN

PS-Mamba (Ours)

Ground Truth

Demo Videos

SPIN

PS-Mamba (Ours)

Ground Truth

Demo Videos

SPIN + Noise

PS-Mamba (Ours)

Ground Truth

Summary

- Our PS-Mamba is the **first** to incorporate spatial-temporal graph learning with Mamba for the **task of human pose sequence refinement**.
- We design an effective **ST-GSS** block that captures spatial-temporal relationships across frames.
- We introduce a dynamic graph weight matrix that learns the relative influence of edges.
- Experiments show that PS-Mamba outperforms current SOTAs.

Thanks for your time!