Interleaved Selective State Space Models for Efficient WiFi-Based 3D Multi-Person Pose Estimation

Quang-Anh Nguyen Duc[1], Kok-Seng Wong[1,2]

[1]Center for Environmental Intelligence [2]College of Engineering and Computer Science, VinUniversity

INTRODUCTION

Motivation & Background

Aging population
→ 2.1 billion by 2050

Need
continuous health monitoring

Pose estimation applications
fall detection; rehabilitation tracking

Concern
privacy issues with camera-based systems

WiFi sensing advantages
through-wall; works in darkness; privacy-preserving

Camera vs WiFi Sensing

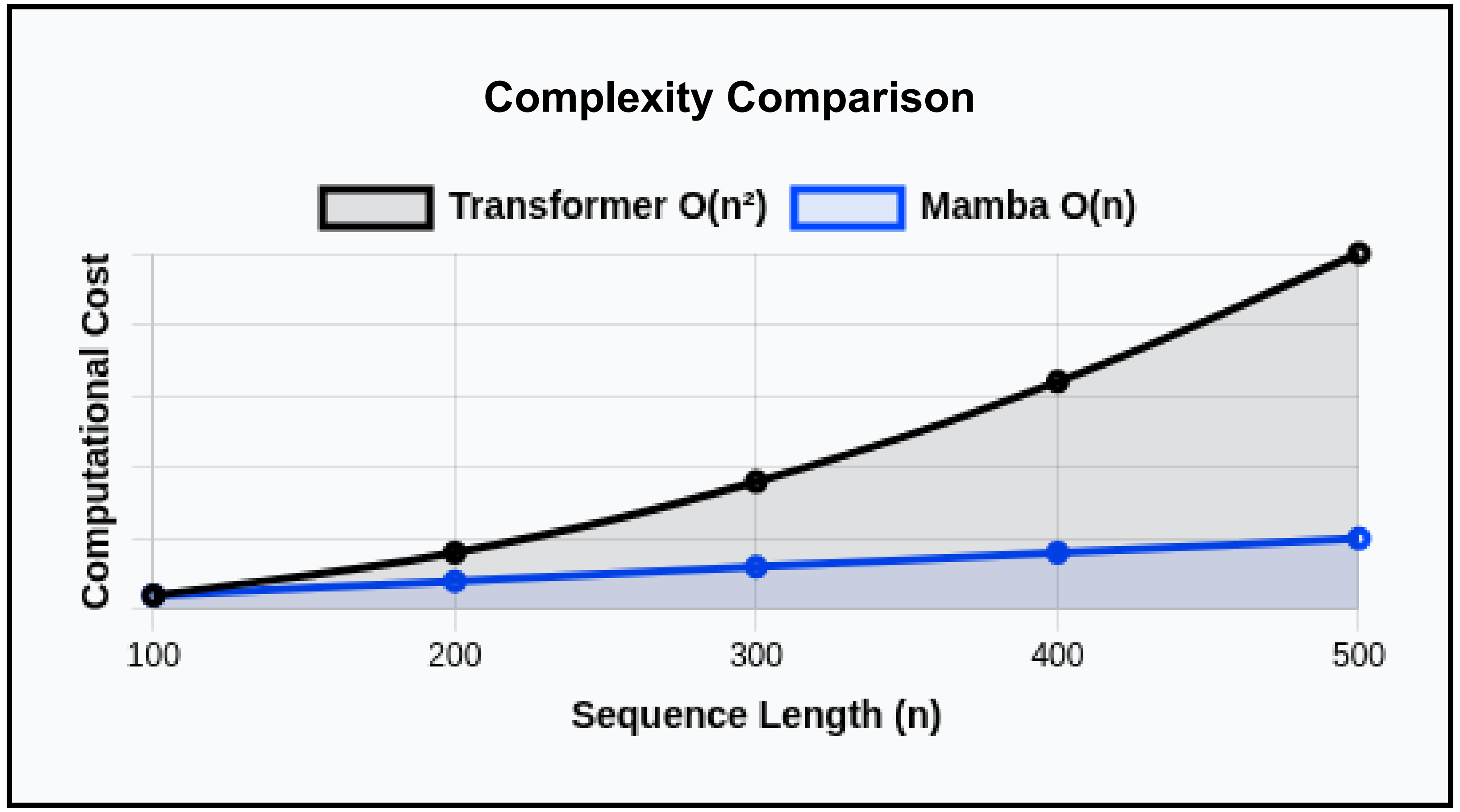
Feature	Camera-based	WiFi-based
Privacy	Intrusive (visual data)	**Privacy-preserving**
Lighting	Requires good light	**Works in darkness**
Obstacles	Blocked by walls/objects	**Through-wall capability**
Applications	Public spaces	**Home health monitoring**

Challenges in WiFi Pose Estimation

- CSI is complex-valued with intricate inter-subcarrier correlations
- Amplitude vs Phase: different physical properties
- Multi-person signal entanglement: signals linearly mix at receive antennas

- **Transformer limitations:**
  - Quadratic complexity $O(n^2)$
  - Huge parameters: 48.2M
  - No explicit temporal consistency

48.2M
Baseline Parameters



Related Work

State Space Models

S4

Mamba

Vision Mamba

WiFi sensing

Person-in-WiFi

WiPose

MetaFi

Method	Accuracy	Complexity	Params
Person-in-WiFi 3D	91.77mm MPJPE	$O(n^2)$	48.2M
WiFi-Mamba (Ours)	**76.75mm MPJPE**	**$O(n)$**	**2.14M**

Proposed Solution: WiFi-Mamba

First State Space Model (SSM) architecture for WiFi-based 3D multi-person pose estimation

DS³M
Dual-Stream Selective State Space Model

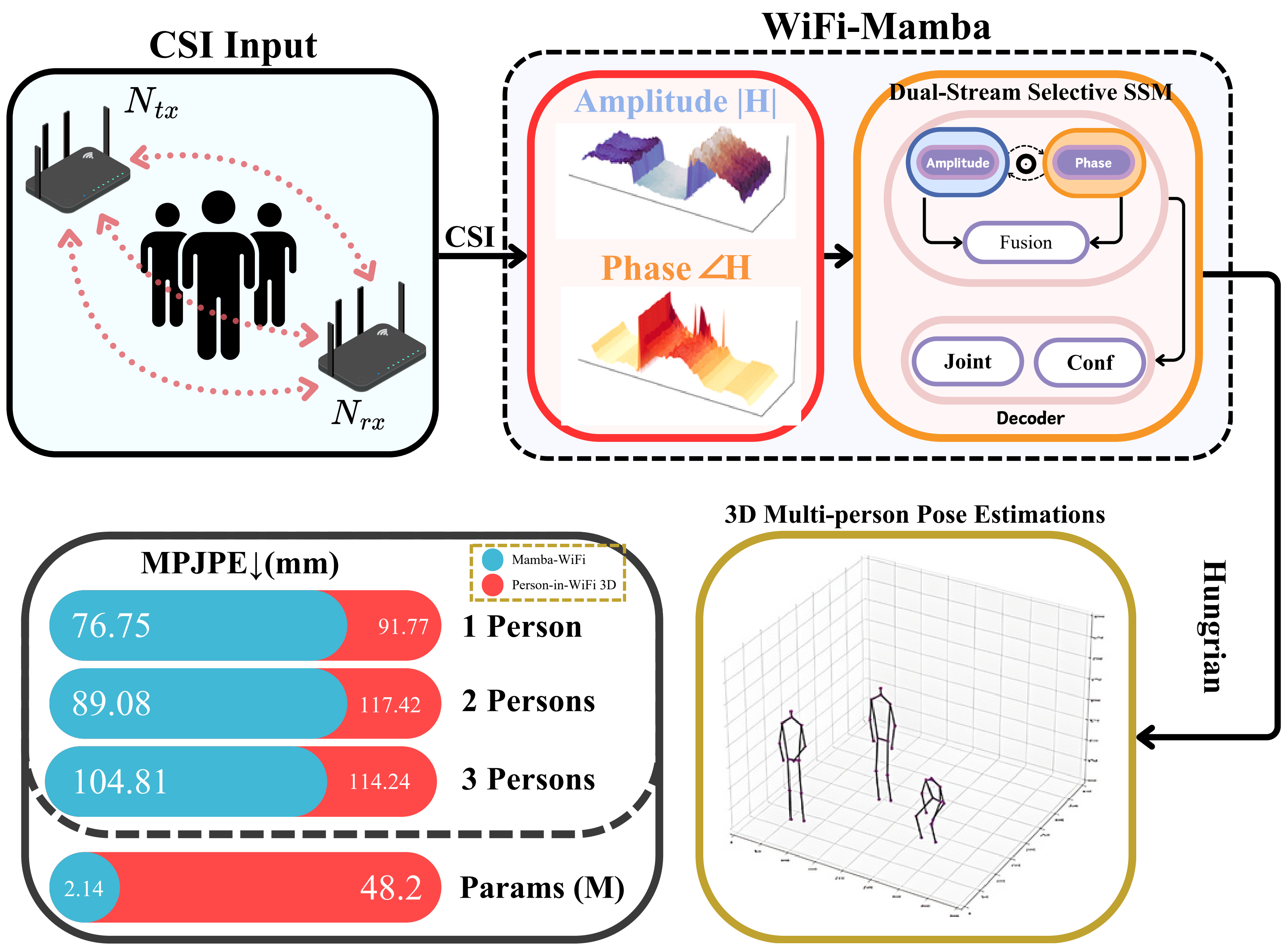
O(n)
Linear complexity

SSA
Selective State Attention

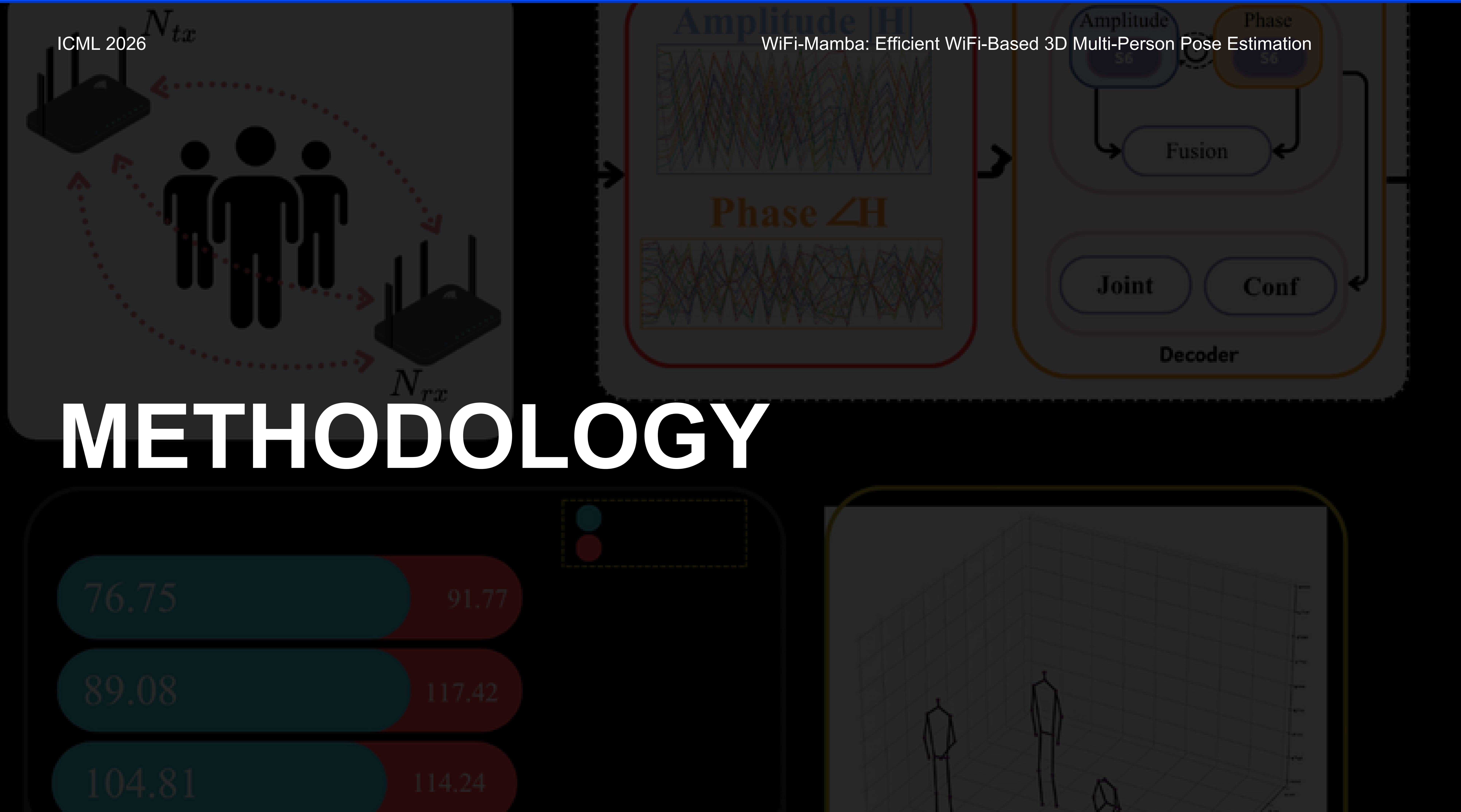
2.14M
Parameters
(4.4% of baseline)

Persistent SSM Memory
for temporal consistency

16–27%
MPJPE improvement

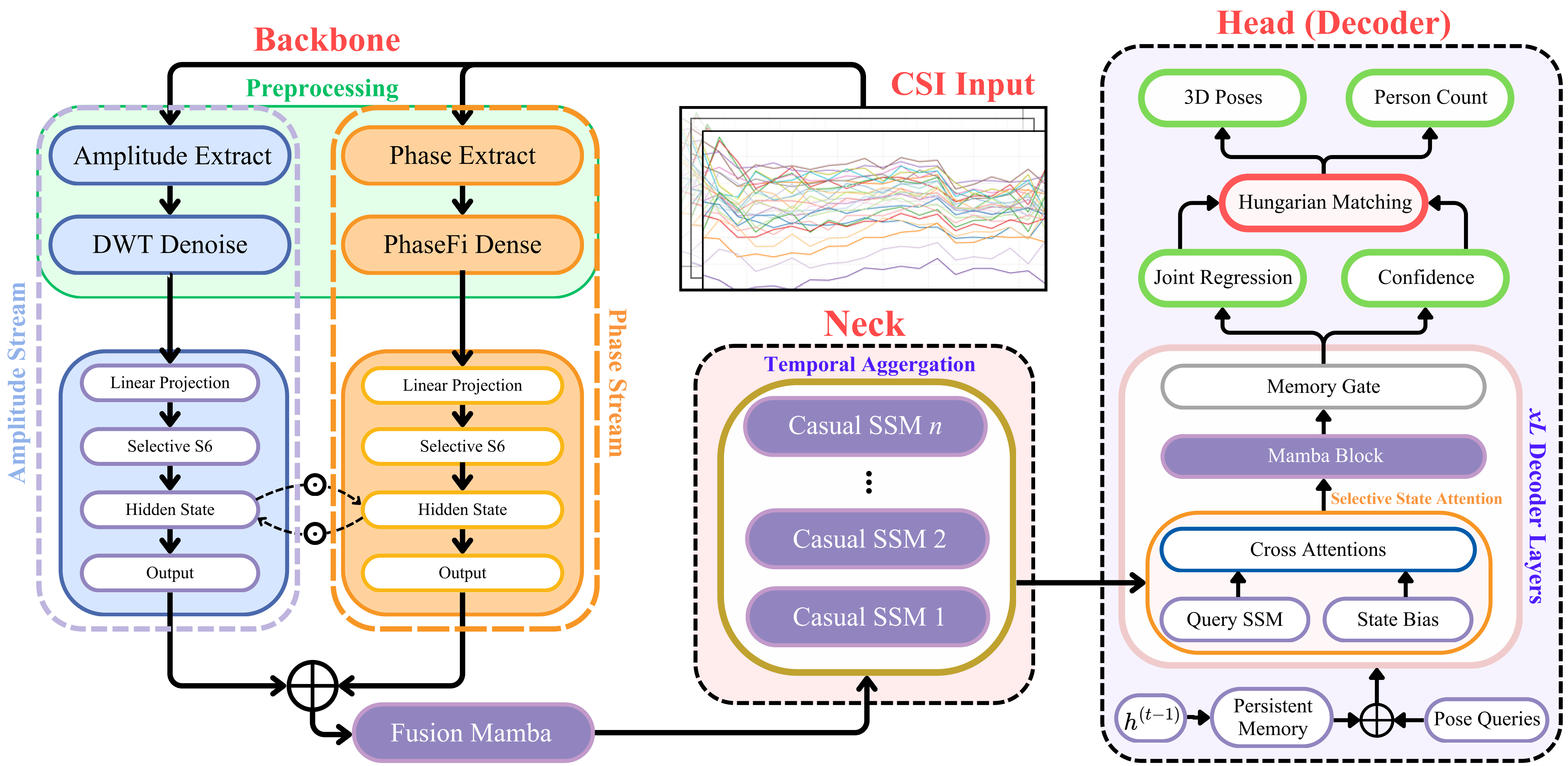


METHODOLOGY



Overall Architecture

1. **Input:** CSI tensor
2. **Preprocessing:** Amplitude/Phase extraction & denoising
3. **Backbone:** DS³M with cross-stream coupling
4. **Neck:** Causal SSM layers for temporal aggregation
5. **Head:** SSA Decoder with Persistent Memory
6. **Output:** 3D poses; confidence scores; person count



Physics-Informed CSI Preprocessing

1. Amplitude & Phase Extraction

$$A[k] = \sqrt{\mathrm{Re}[k]^2 + \mathrm{Im}[k]^2}, \quad \Phi[k] = \mathrm{atan2}(\mathrm{Im}[k], \mathrm{Re}[k]) \tag{9}$$

- $A[k] \in \mathbb{R}^+$: amplitude
- $\Phi[k] \in [-\pi, \pi]$: phase

Preserves the physical properties of the WiFi signal

2. Denoising

Amplitude (DWT + soft thresholding):

$$d_l = \mathrm{sign}(d_l) \cdot \max(|d_l| - \tau_l, 0) \tag{10}$$

where $\tau_i = softplus(\theta_l)$ is a learnable threshold for level l. For phase, we apply PhaseFi to remove CFO/SFO-induced linear distortion:

$$\varphi\_corrected[k] = \varphi[k] - (\alpha k + \beta) \tag{11}$$

- α, β: estimated via linear regression
- Followed by learnable residual refinement

DS³M Core Contribution

- Amplitude & Phase provide complementary information (intensity vs Doppler)
- Two parallel streams (S6) + Cross-stream coupling
- Direct interaction at the hidden state level (not only output fusion)

- **Coupled States**

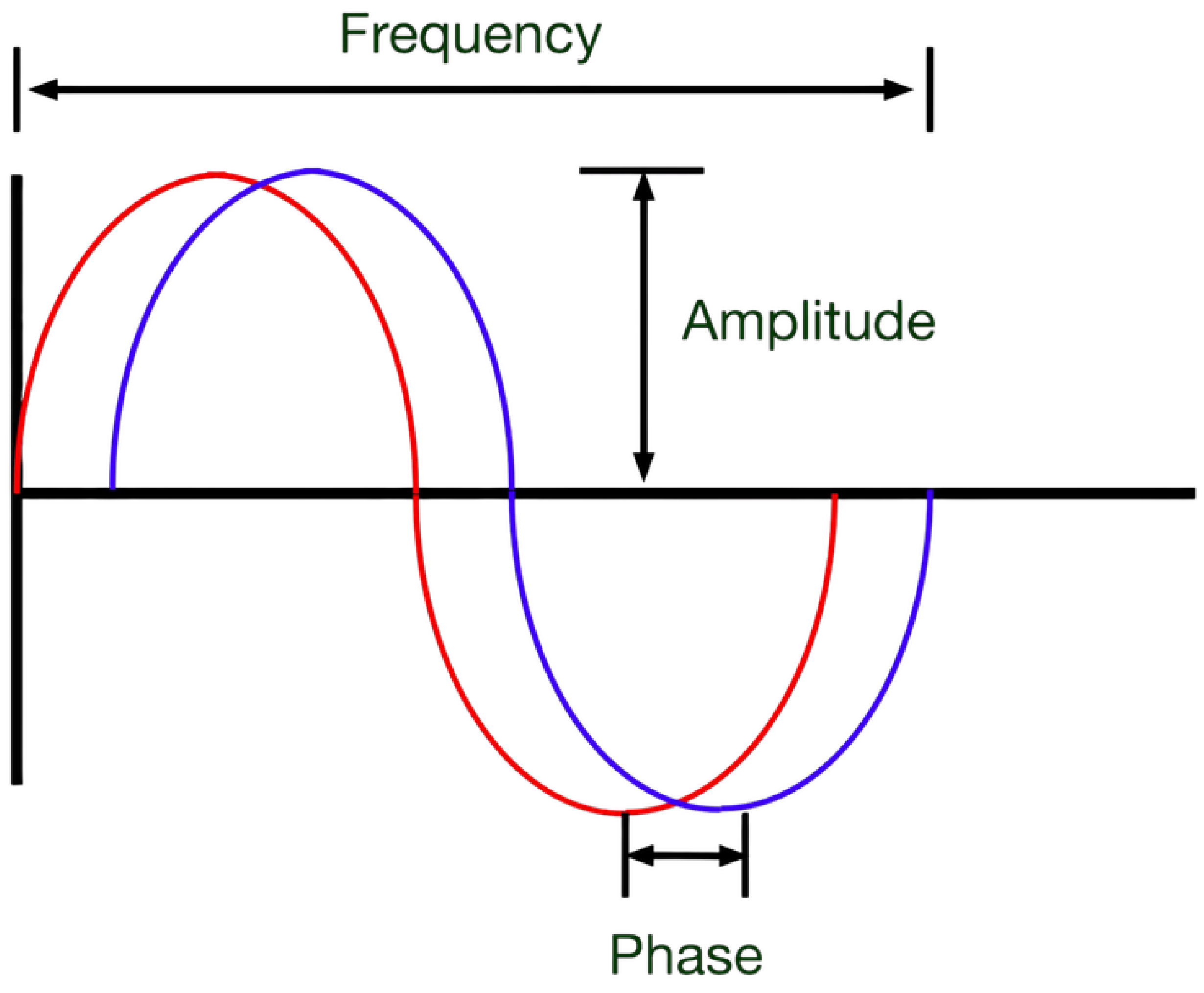
$$\tilde{h}_A^t = h_A^t + \lambda \cdot g_t \odot W_{\phi\to A}(h_\phi^t) \quad (12)$$

$$\tilde{h}_\phi^t = h_\phi^t + \lambda \cdot (1 - g_t) \odot W_{A\to\phi}(h_A^t) \quad (13)$$

- **Adaptive Gate**

$$g_t = \sigma\left(W_g[\mathrm{mean}(h_A^t); \mathrm{mean}(h_\phi^t)]\right) \quad (14)$$

- **Interpretation**
  - λ: controls information exchange strength
  - gt: adaptively modulates information flow direction
  - → Amplitude and Phase reinforce each other → improved representation & accuracy



DS^{3}M Core Contribution

Amplitude – Spatial Domain

$$\tilde{y}^A = y^A \odot \left(1 + \lambda \cdot \sigma(W_m^A \cdot \mathrm{mean}(\tilde{h}_t^A))\right) \quad \textbf{(15)}$$

Phase – Dynamics Domain

(Gating mechanism)

Structural Context

Instead of blindly merging signals, DSM processes amplitude and phase in parallel while allowing them to interact at the hidden-state level.

Cross-stream State Coupling

- When phase signals are clear, they help refine noisy amplitude variations, and vice versa.
- This mutual reinforcement effectively resolves multi-person signal overlap.

Selective State Attention (SSA)

- **Limitation of Standard Attention**
  - Standard cross-attention processes queries independently
  - Cannot model relationships between body joints/persons
  - Leads to fragmented attention in multi-person scenarios
- **Proposed SSA**

  Inject sequential context into pose queries using SSM.

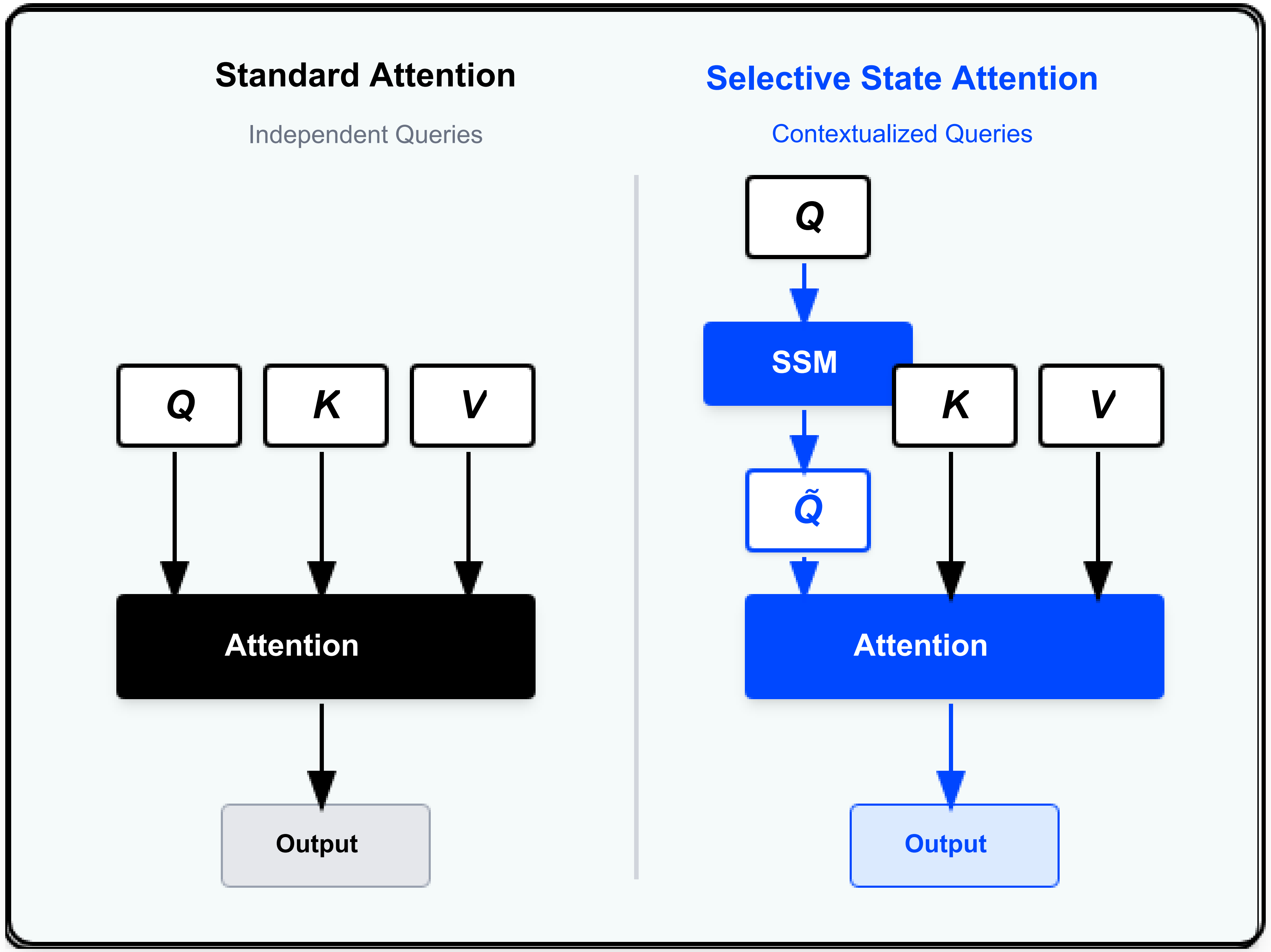
  Query contextualization:

$$\tilde{Q},\ h_Q = \mathrm{SSM}(Q), \quad \tilde{Q} \in \mathbb{R}^{K \times D} \tag{16}$$

  Cross-Attention with Contextualized Queries

$$\mathrm{Attention}(\tilde{Q}, K, V) = \mathrm{softmax}\left(\frac{\tilde{Q}K^T}{\sqrt{d}} + \tau \cdot b\right) \cdot V \tag{17}$$

- **Benefits**
  - Better joint reasoning across body parts/persons
  - More coherent attention over shared CSI features
  - Improved multi-person pose estimation



Persistent SSM Memory

Motivation

Human motion is temporally continuous

Traditional RNNs (GRU/LSTM):

- Vanishing gradients
- High memory cost

→ Need efficient long-range temporal modeling

Key Idea

- Use SSM selective scan to maintain memory across frames
- No explicit gating → stable & scalable temporal modeling

Memory Architecture

- Introduce M learnable memory tokens: $m \in \mathbb{R}^{M \times D}$
- Concatenate with queries:

$$\phi = [m; Q] \in \mathbb{R}^{(M+K) \times D} \tag{18}$$

- Memory tokens store compressed historical context

Loss Functions

We use Hungarian matching for bipartite assignment between predictions $\{(\hat{P}_k, \hat{c}_k)\}_{k=1}^{K}$ and ground truth: $\{P_i\}_{i=1}^{N_p}$

$$\pi^* = \mathrm{argmin}_{\pi} \sum_{i=1}^{N_p} \left(\|P_i - \hat{P}_{\pi(i)}\|_1 - \hat{c}_{\pi(i)} \right) \quad (21)$$

Given π*, the total loss is:

$$\mathcal{L} = \lambda_j \mathcal{L}_{\text{joint}} + \lambda_b \mathcal{L}_{\text{bone}} + \lambda_c \mathcal{L}_{\text{conf}} + \lambda_n \mathcal{L}_{\text{count}}, \quad (22)$$

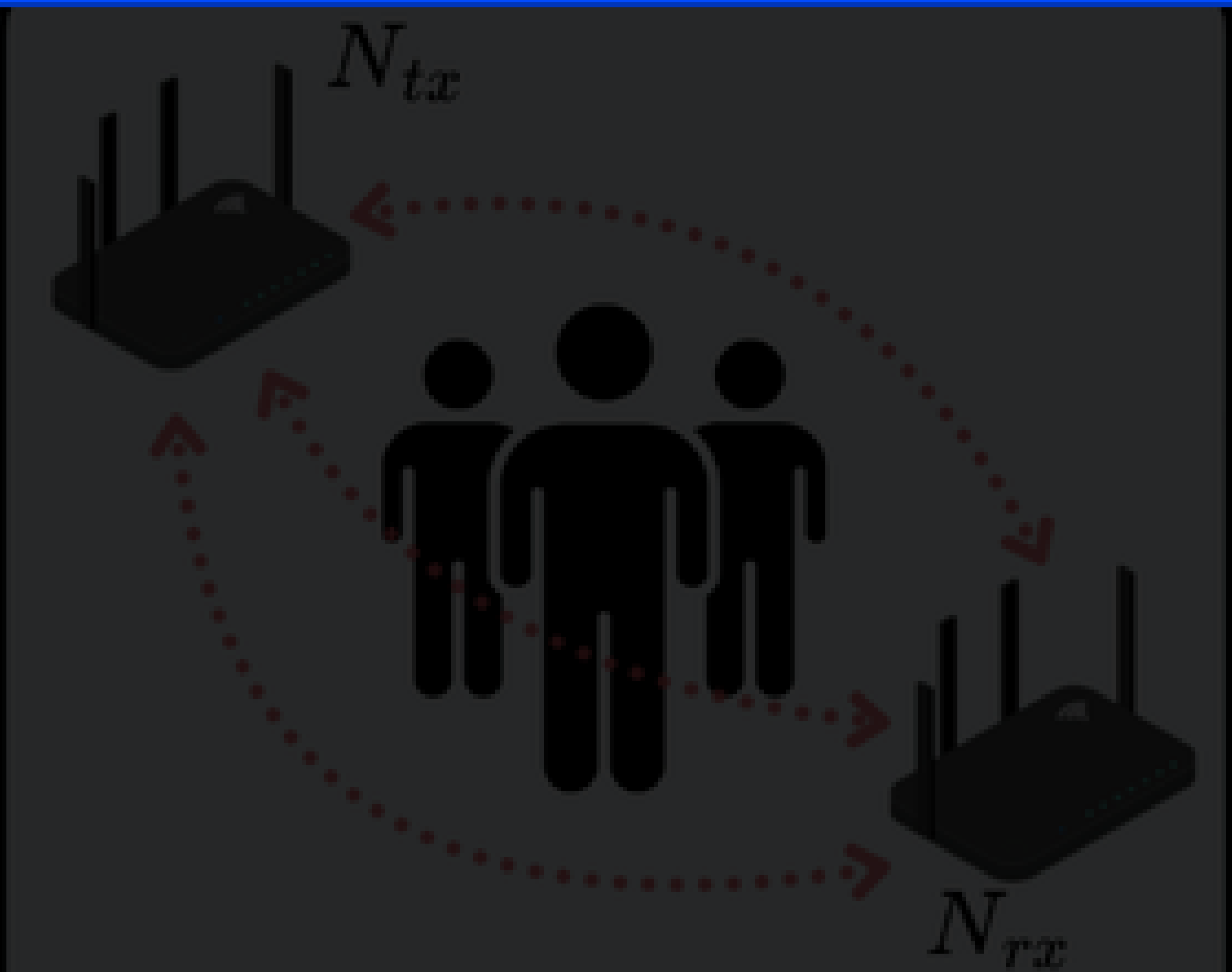
where L_{joint} supervises matched joint positions with l_1

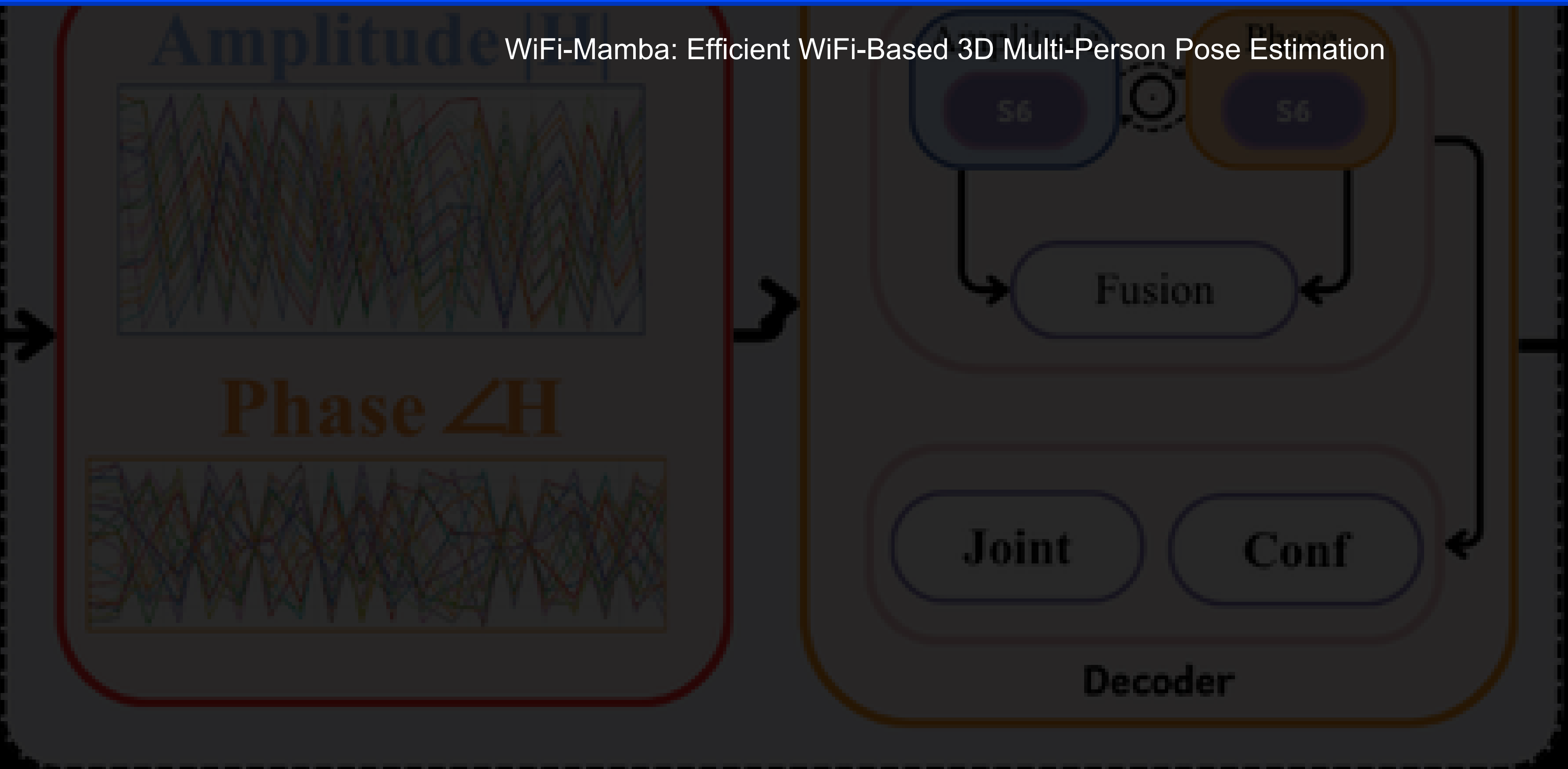
$$\mathcal{L}_{\text{joint}} = \frac{1}{N_p J} \sum_{i=1}^{N_p} \sum_{j=1}^{J} \|P_{i,j} - \hat{P}_{\pi^*(i),j}\|_1, \quad (23)$$

here, $\mathcal{L}_{\text{bone}}$ enforces anatomical bone length consistency:

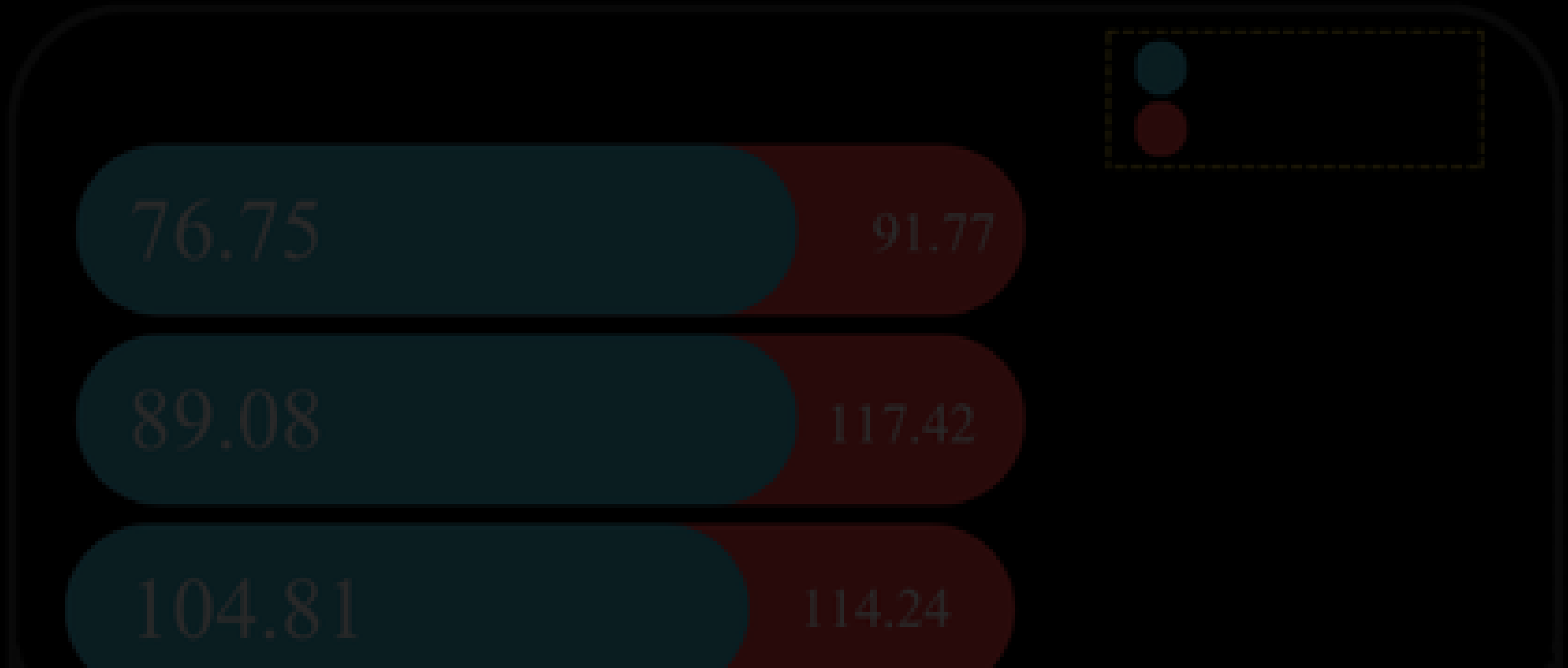
$$\mathcal{L}_{\text{bone}} = \frac{1}{N_p|\mathcal{B}|} \sum_{i=1}^{N_p} \sum_{(j_1,j_2)\in\mathcal{B}} \Big| \|P_{i,j_1} - P_{i,j_2}\|_2 - \|\hat{P}_{\pi^*(i),j_1} - \hat{P}_{\pi^*(i),j_2}\|_2 \Big| \quad (24)$$

the L_{conf} applies Focal Loss on confidence scores with target L_{count} for matched queries, and $y_{k=1}$ uses cross-entropy on person count prediction.





EXPERIMENTS



Experimental Setup

Datasets

Person-in-WiFi 3D

72K frames

9 subjects

WiPose

166,600 frames

12 volunteers

Metrics

- **MPJPE**
- **MPJDLE**
- **PCK@{10, 20, 30}**

Evaluation Results

Method	Params (M)	MPJPE	PA-MPJPE	MPJDLE (h)	MPJDLE (v)	MPJDLE (d)	PCK@30	PCK@20	PCK@10
MetaFi++ (Zhou et al., 2023a)	265	11.640	7.240	5.694	6.101	4.636	7.531	6.045	3.291
DT-Pose (Chen et al., 2025)	48	9.270	6.009	4.166	5.055	3.860	8.195	7.178	5.035
HPE-Li (D. Gian et al., 2025)	8	11.770	6.990	5.364	6.719	5.006	7.527	6.035	3.378
Person-in-WiFi 3D (Yan et al., 2024)	482	9.177	6.451	3.784	5.197	4.500	9.496	8.620	5.970
WiFi-Mamba (Ours)	**214**	**7.675**	**6.285**	**3.211**	**4.777**	**3.141**	**8.629**	**7.807**	**6.109**

Table 1. *Single-person pose estimation on Person-in-WiFi 3D dataset.*

Evaluation Results

Method	Params (M)	MPJPE	PA-MPJPE	MPJPE (h)	MPJPE (v)	PCK@30	PCK@20	PCK@10
MetaFi++ (Zhou et al., 2023a)	265	5.817	3.515	3.245	3.921	5.338	3.665	1.362
DT-Pose (Chen et al., 2025)	48	**3.414**	**2.319**	**1.888**	**2.234**	**7.797**	**6.961**	5.140
HPE-Li (D. Gian et al., 2025)	**8**	3.983	2.552	2.238	2.658	7.058	5.687	3.321
WiFi-Mamba (Ours)	214	3.582	2.406	1.966	2.362	7.592	6.775	**5.199**

***Table 2**. Single-person pose estimation on WiPose dataset. WiFi-Mamba vs. vs. state-of-the-art methods. Lower is better for error metrics. Best results are in bold, second-best are underlined.*

Evaluation Results

Metric	Person-in-WiFi 3D (1P)(Yan et al., 2024) (48.2M)	Person-in-WiFi 3D (2P)(Yan et al., 2024) (48.2M)	Person-in-WiFi 3D (3P)(Yan et al., 2024) (48.2M)	WiFi-Mamba (1P)(Ours) (2.14M)	WiFi-Mamba (2P)(Ours) (2.14M)	WiFi-Mamba (3P)(Ours) (2.14M)
MPJPE	9.177	12.104	14.424	**7.675**	**8.908**	**10.481**
PA-MPJPE	6.451	6.891	6.190	**6.285**	**6.755**	**6.631**
MPJDLE (h)	3.784	6.575	6.141	**3.211**	**4.876**	**3.693**
MPJDLE (v)	5.197	4.843	5.805	**4.777**	**3.792**	**4.758**
MPJDLE (d)	4.500	5.558	9.121	**3.141**	**3.843**	**6.254**
PCK@30	9.496	7.718	6.566	**8.629**	**7.944**	**7.344**
PCK@20	8.620	6.030	4.000	**7.807**	**6.864**	**5.825**
PCK@10	5.970	2.504	1.451	**6.109**	**4.767**	**3.282**

Table 3. *Performance comparison between WiFi-Mamba model and Person-in-WiFi 3D model under different numbers of persons.*

Evaluation Results

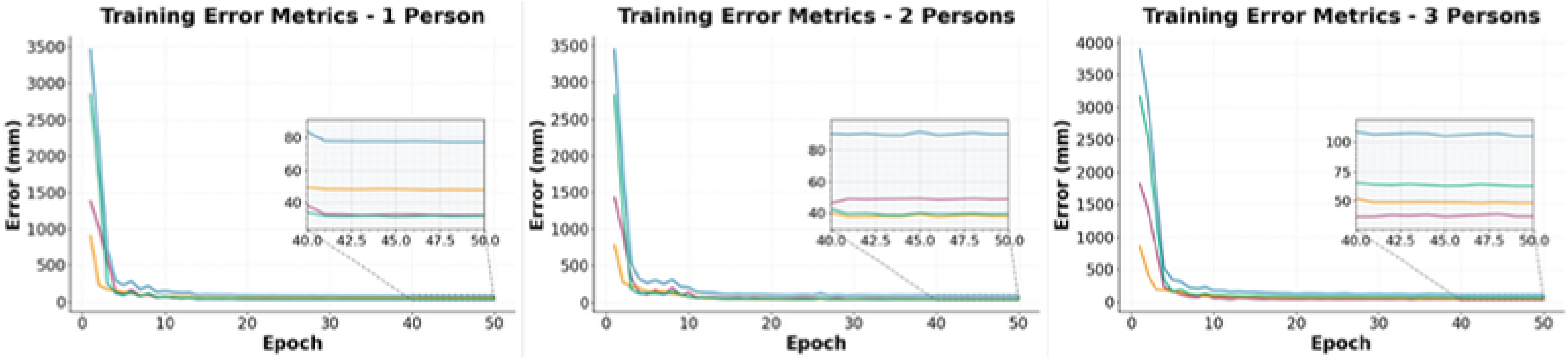


Figure 3. *Training convergence curves for WiFi-Mamba across one, two, and three-person scenarios.*

Ablation Study

Method	MPJPE	MPJDLE(h)	MPJDLE(v)	MPJDLE(d)
Amplitude Only	11.951	3.848	5.780	7.369
Phase Only	12.239	4.186	5.786	7.292
No Separation	11.371	3.958	5.296	6.814
No Memory	11.737	4.138	5.558	6.866
Standard Attention	11.855	4.007	5.497	7.176
WiFi-Mamba	**10.481**	**3.693**	**4.758**	**6.254**

Table 4. *Ablation study on the effect of each component. Lower values are better. Bold and underlined values denote the best and second-best results, respectively.*

Ablation Study

λ	MPJPE	MPJDLE(h)	MPJDLE(v)	MPJDLE(d)
0	10.749	3.627	4.974	6.472
5	10.679	3.774	4.813	6.421
1	10.498	**3.675**	4.772	6.279
5	10.555	3.824	4.922	**6.180**
10	**10.481**	3.693	**4.758**	6.254
20	10.537	3.770	4.969	6.017
30	10.699	4.957	4.957	6.326

Table 5. *Coupling ratio λ sensitivity on 3-person pose estimation performance. Lower values are better. Bold and underlined valuesdenote the best and second-best results, respectively.*

Ablation Study

Setting	Params (M)	MPJPE	MPJPE(h)	MPJPE(v)	MPJPE(d)
Model depth (DS³M + Mamba)					
Blocks = 2	265	12.733	4.136	6.238	7.581
Blocks = 4	366	11.804	4.056	5.699	6.861
SSA layers					
SSA = 1	164	12.790	4.102	6.361	7.597
SSA = 3	264	12.367	4.123	6.066	7.288
SSA = 4	314	12.549	4.400	6.216	7.249
SSM state size (N)					
N = 8	269	12.293	4.091	6.045	7.189
N = 16	278	12.421	3.968	6.039	7.628
N = 32	296	11.984	4.327	5.819	7.136
Memory tokens (M)					
M = 0	214	11.860	4.287	5.762	6.849
M = 8	214	11.880	4.096	5.875	6.861
M = 32	214	12.280	4.327	5.819	7.136

WiFi-Mamba (Final)	**214**	**10.481**	**3.693**	**4.758**	**6.254**

Table 6. *Ablation study on architectural hyperparameter ablation. Lower values are better. Bold and underlined values denote the best and second-best results, respectively.*

Result Under Perbutations

Condition	Paths	Faded (%)	MPJPE	PA-MPJPE	MPJPE(h)	MPJPE(v)	MPJPE(d)	PCK@30
Clean	0	0	**10.136**	**6.072**	**3.591**	**4.824**	**5.814**	**759**
Mild	1	3	11.150	6.435	3.764	5.250	6.577	728
Moderate	2	7	16.732	7.392	7.540	6.857	8.266	605
Severe	3	12	22.809	8.414	11.824	8.909	10.311	485
Extreme	5	20	29.405	9.971	16.636	11.120	12.598	363

Table 8. *Sensitivity analysis under simulated multipath interference.*

Result Under Perbutations

Condition	Reduction (%)	Phase SNR	MPJPE	PA-MPJPE	MPJPE(h)	MPJPE(v)	MPJPE(d)	PCK@30
LOS (Clean)	0	∞	10.136	6.072	3.591	4.824	5.814	**759**
Partial NLOS	10	30	**10.067**	**6.044**	**3.554**	**4.813**	**5.776**	753
Moderate NLOS	25	20	10.572	6.152	3.604	5.187	6.097	729
Full NLOS	45	12	25.189	8.111	13.791	9.339	10.572	462
Extreme NLOS	70	6	62.548	13.091	37.302	18.416	28.235	190

***Table 9**. Sensitivity analysis under simulated NLOS conditions.*

Limitations

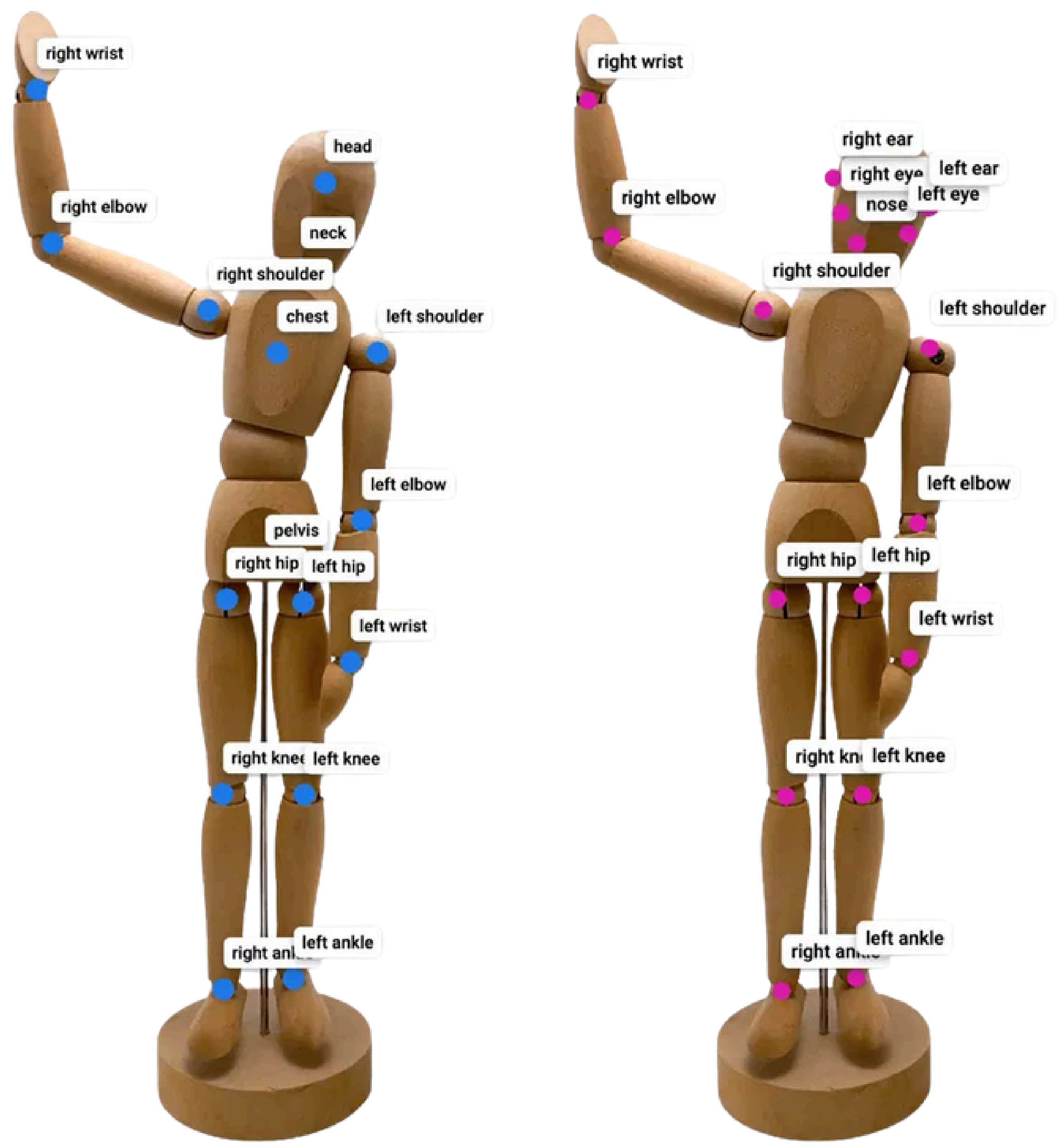
- SSMs may struggle with highly non-local dependencies compared to full attention
- Limited dataset availability:

  Only one public 3D multi-person WiFi dataset

  Maximum 3-person scenarios
- FPS evaluation not standardized due to hardware-dependent CSI acquisition

Future Work

- Extend evaluation to:
  - Larger crowds
  - More diverse indoor environments
  - Severe NLOS and interference conditions
- Develop standardized real-world CSI acquisition benchmarks
- Improve robustness and real-time deployment efficiency
- Explore hybrid SSM-attention architectures for stronger global modeling



Thank You