KOALA: Koopman Predictive Coding for WiFi-Based Anticipatory Human Motion Prediction

Anonymous authors
Paper under double-blind review

Abstract

WiFi Channel State Information (CSI) has emerged as a privacy-preserving alternative to cameras for human pose estimation. However, existing approaches treat pose inference as an instantaneous regression problem and do not model temporal dynamics, making future motion prediction infeasible. Naively applying vision-based prediction methods compounds the estimation noise already present in CSI-derived poses, as autoregressive rollouts amplify errors at every step. We propose KOALA, the framework for human motion prediction directly from WiFi CSI, by lifting noisy CSI-derived pose sequences into a learned Koopman latent space where nonlinear dynamics become linear, enabling multi-horizon prediction via simple matrix-vector products without autoregressive iteration or error accumulation. A residual CSI-conditioned operator resolves the identity attractor problem inherent from Koopman formulations, and an anchor-delta prediction head eliminates the degenerate shortcut of copying the current pose across all horizons. To regularise the lifting and operator jointly, we introduce a Koopman Predictive Coding (KPC) loss that operates in the temporal-encoder feature space, enforcing dynamical consistency across prediction horizons without requiring contrastive, spectral, or auxiliary losses. Experiments on MM-Fi and WiPose show that KOALA achieves robust, consistent performance across both short- and long-term prediction horizons, outperforming all baselines by a substantial margin.

1 Introduction

Anticipating human motion before it fully unfolds is a fundamental capability for intelligent systems operating in human-centric environments, enabling proactive assistance in smart homes, collision avoidance in human-robot collaboration, and early intervention in healthcare monitoring (Lyu et al., 2022). The dominant paradigm for this task relies on skeletal sequences derived from optical motion capture or RGB-D sensors (Martinez et al., 2017; Mao et al., 2019; 2020; Xu et al., 2023; Guo et al., 2023). Despite remarkable progress, camera-based systems impose significant deployment constraints: they require line-of-sight, degrade under occlusion and low illumination, and raise pervasive privacy concerns in domestic and clinical settings. These

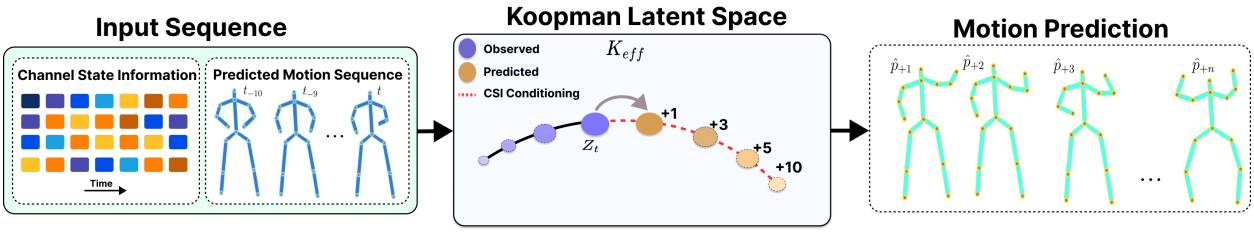

Figure 1: KOALA predicts future human motion from WiFi CSI and observed skeleton poses, these are lifted into a Koopman latent space, where a single CSI-conditioned linear operator propagates the state to multiple future horizons without autoregressive iteration.

barriers motivate a fundamentally different sensing modality, which is ubiquitous, infrastructure-free, and inherently privacy-preserving.

WiFi Channel State Information (CSI) has emerged as a compelling alternative for non-intrusive human sensing (Zhou et al., 2023). By capturing how the human body perturbs multipath radio-frequency propagation, CSI encodes rich motion signatures without requiring any wearable device or visual access to the monitored space (Zhou et al., 2022; Wang et al., 2015; Yang et al., 2023a). A growing line of work has demonstrated the viability of recovering instantaneous 3D skeletal poses from CSI signals, leveraging cross-modal supervision from synchronized cameras to bridge the modality gap (Chen et al., 2023; Yang et al., 2022). Subsequent advances have pushed toward 2D and 3D reconstruction (Zhou et al., 2022; 2023), multi-person scenarios (Yan et al., 2024), and cross-environment generalization (Zhou et al., 2024). These systems, however, share a critical architectural assumption, they operate as instantaneous regressors, mapping a fixed CSI observation window to the current pose state. Modeling forward temporal dynamics, predicting where a person will be rather than where they are, remains largely unexplored. Despite its promise, forecasting human motion from WiFi CSI is fundamentally more challenging than its vision-based counterpart. First, CSI-derived pose observations are inherently noisy and incomplete due to multipath fading, environmental interference, and hardware variability, leading to spatial jitter and missing joints. Second, unlike clean motion capture sequences, the temporal dynamics extracted from CSI are less stable and harder to model. Third, multi-horizon prediction requires maintaining temporal consistency over long sequences, where autoregressive strategies suffer from error accumulation and exposure bias. These challenges call for new modeling paradigms that are robust to noise while capturing long-term motion dependencies.

To address these gaps, we propose **K**oopman **O**perator with **A**ttentive **L**ifting **A**rchitecture (**KOALA**), the first unified framework for WiFi-CSI-based human motion prediction (see Fig. 1). Moving beyond conventional approaches that focus on instantaneous pose estimation or action recognition (Huang et al., 2024; Zhou et al., 2024), KOALA forecasts future human motion directly from ambient Radio Frequency (RF) signals by lifting noisy CSI-derived pose sequences into a learned Koopman latent space where nonlinear dynamics become linear. Multi-horizon prediction is then performed via simple matrix-vector products, without autoregressive iteration or error accumulation. To enforce dynamical consistency and representational richness, KOALA is trained under a bio-inspired Koopman Predictive Coding (KPC) loss (Rao & Ballard, 1999; Friston, 2005) that operates in the temporal-encoder feature space, shaping the lifting by minimizing prediction errors at the timescales relevant to the task. Our contributions are summarized as follows.

- We introduce a novel framework for multi-horizon human pose prediction directly from WiFi CSI, establishing a new problem at the intersection of RF sensing and anticipatory motion understanding.
- We propose a residual CSI-conditioned Koopman operator that resolves the identity attractor problem inherent in prior Koopman-based methods, paired with an anchor-delta prediction formulation that eliminates the degenerate shortcut of copying the current pose across all horizons.
- We introduce a Koopman Predictive Coding loss that subsumes reconstruction, linearity, and stability objectives into a single anchored feature-space prediction term, removing the need for separate contrastive, spectral, and auxiliary losses used in prior work.
- We conduct extensive experiments on MM-Fi (Yang et al., 2023b) and WiPose (Zhou et al., 2022), demonstrating that KOALA consistently outperforms existing baselines across both short- and long-term prediction horizons. Code will be released upon publication.

2 Related Work

2.1 WiFi CSI-Based Human Pose Estimation

Early works established cross-modal supervision as the dominant paradigm for WiFi-based human pose estimation, using synchronized cameras as teacher signals to map CSI amplitude and phase to 2D joint coordinates (Chen et al., 2023; Yang et al., 2022). Subsequent research extended this to full 3D reconstruction, WiPose (Jiang et al., 2020) combines residual CNNs with recurrent models for spatiotemporal feature

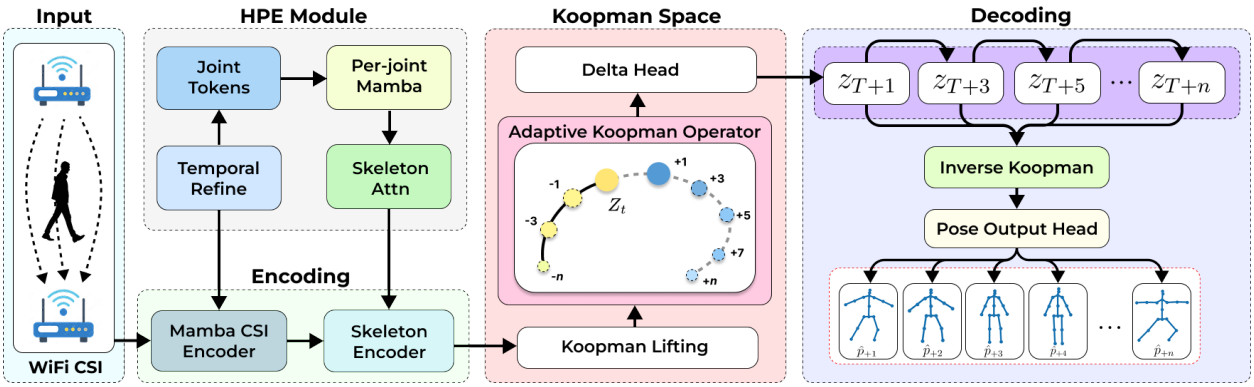

Figure 2: Overview of the KOALA framework. WiFi CSI is encoded by a Mamba CSI encoder whose features are passed to the HPE module and skeleton encoder. The HPE module estimates poses via temporal refinement, joint token construction, per-joint Mamba, and skeleton-biased attention; estimated poses are gradient-detached before entering the skeleton encoder. Dual-stream fusion and a temporal encoder generate contextual motion features, which are lifted into the Koopman latent space. A residual CSI-conditioned operator propagates latent states across prediction horizons without autoregression. Inverse lifting and a delta head decode pose deltas, which are added to an anchor pose for final predictions.

extraction, while MetaFi++ (Zhou et al., 2023) employs Transformer encoders for richer motion representations. Person-in-WiFi 3D (Yan et al., 2024) further scales to multi-person scenarios via set prediction, and AdaPose (Zhou et al., 2024) addresses cross-environment generalization through domain-invariant adaptation. Despite these advances, all existing systems work as instantaneous regressors, mapping a fixed CSI window to the current pose without modeling forward temporal dynamics. Predicting future skeletal states from WiFi sequences remains entirely unaddressed.

2.2 Skeleton-Based Motion Prediction

Predicting future poses from skeleton sequences is a core computer vision task with broad applications (Lyu et al., 2022). Early recurrent methods, including Residual-GRU (Martinez et al., 2017) and Structural-RNN (Jain et al., 2016), suffered from error accumulation and mean-pose collapse at longer horizons. Feed-forward approaches addressed this by predicting all frames simultaneously by LTD (Mao et al., 2019) and HRI (Mao et al., 2020) worked in the frequency domain with motion attention, while MST-GNN (Li et al., 2021) and MSR-GCN (Dang et al., 2021) improved spatial modeling via multi-scale graph convolutions. More recently, AuxFormer (Xu et al., 2023) used auxiliary tasks to strengthen Transformer representations, and siMLPe (Guo et al., 2023) showed that a simple MLP with DCT preprocessing can outperform complex architectures. Despite these, all existing methods operate on clean skeleton observations. To our knowledge, no prior work addresses motion prediction from WiFi CSI, a fundamentally harder setting that requires forecasting future poses from a noisy sensing modality. KOALA is the first to tackle this problem.

3 Methodology

3.1 Preliminaries

Channel State Information. Modern WiFi standards, such as IEEE 802.11, employ Orthogonal Frequency-Division Multiplexing (OFDM) to transmit data over multiple subcarriers simultaneously (792, 2017). The received signal on each subcarrier s between transmit antenna a_t and receive antenna a_r is characterized by a complex-valued channel coefficient:

$$H(s, a_t, a_r) \;=\; |H(s, a_t, a_r)|\, e^{j\angle H(s, a_t, a_r)}, \tag{1}$$

where $|H|$ captures signal attenuation and $\angle H$ encodes the phase shift from multipath propagation. When a person moves, body segments reflect and scatter the signal along multiple paths, each contributing an

attenuated and phase-shifted copy:

$$H(s) \;=\; \sum_{l=1}^{L} \alpha_l \, e^{-j2\pi d_l/\lambda_s}, \tag{2}$$

where α_l and d_l are the attenuation and propagation distance of path l, and λ_s is the wavelength of subcarrier s. As the body moves, path lengths d_l change over time, inducing time-varying fluctuations in both $|H(s, a_t, a_r)|$ and $\angle H(s, a_t, a_r)$ across all subcarrier-antenna pairs that encode the body's spatial configuration. Stacking these coefficients over T time steps, S subcarriers, and A antenna pairs yields the full CSI tensor $\mathbf{X} \in \mathbb{R}^{T\times S\times A}$, where each entry corresponds to $H(s, a_t, a_r)$ at a given time step. In KOALA, $\mathbf{X}$ serves both as direct contextual input for motion dynamics and as the sole sensory source for estimating observed poses, enabling fully end-to-end operation from raw WiFi signals.

Koopman Operator Theory. Human motion is inherently nonlinear, but Koopman theory (Koopman, 1931) provides a principled linearisation. Let $\mathbf{v}_t \in \mathcal{V}$ denote the body state at time t, representing the configuration of the human body in observation space. For a nonlinear body motion model $\mathbf{v}_{t+1} = F(\mathbf{v}_t)$ with transition F, the Koopman operator $\mathcal{K}$ acts on observables $g : \mathcal{V} \to \mathbb{R}^d$ via:

$$(\mathcal{K}g)(\mathbf{v}) \;=\; g\big(F(\mathbf{v})\big), \quad \forall\, \mathbf{v} \in \mathcal{V}. \tag{3}$$

Although F is nonlinear, $\mathcal{K}$ is always linear (Koopman, 1931). For vector-valued observables $\boldsymbol{g} = [g_1, \ldots, g_N]^\top$, the lifted state evolves as $\boldsymbol{g}(\mathbf{v}_{t+1}) = \mathbf{K}\,\boldsymbol{g}(\mathbf{v}_t)$ with $\mathbf{K} \in \mathbb{R}^{N\times N}$, and multi-step prediction at horizon h is $\boldsymbol{g}(\mathbf{v}_{t+h}) = \mathbf{K}^h \boldsymbol{g}(\mathbf{v}_t)$, requiring no autoregressive iteration. The main challenge is learning a lifting $\boldsymbol{g}$ where the finite-dimensional approximation remains accurate (Brunton & Kutz, 2019; Lusch et al., 2018). In KOALA, N corresponds to the Koopman latent dimension D_z and $\boldsymbol{g}$ is the composition of a skeleton-aware pose encoder, a dual-stream fusion module, and a lifting network, while $\mathbf{K}$ is CSI-conditioned in residual form, allowing the WiFi signal to modulate dynamics without the pose model encoding identity information.

Predictive Coding. Predictive coding (PC) posits that the brain minimises hierarchical prediction errors between top-down expectations and bottom-up sensory evidence (Rao & Ballard, 1999), with learning driven by the total energy $\mathcal{F} = \sum_\ell \|\mathbf{r}^{(\ell)} - f_\ell(\mathbf{r}^{(\ell+1)})\|^2_{\boldsymbol{\Sigma}_\ell^{-1}}$. We adopt this principle as a Koopman Predictive Coding (KPC) loss, the Koopman operator predicts future states in the temporal-encoder feature space, and the lifting is refined by minimising the resulting prediction errors. Operating in this internal feature space rather than in raw observation space decouples the KPC signal from CSI noise while maintaining gradient pressure through the lifting and operator at the timescales relevant to the prediction task.

3.2 Problem Formulation

We predict future human poses from WiFi CSI alone at inference. Let $\{\mathbf{x}_t\}_{t=1}^T$ with $\mathbf{x}_t \in \mathbb{R}^{D_c}$ denote T CSI frames, where $D_c = S \times A$ is the product of the number of subcarriers S and antenna pairs A. Here J denotes the number of body joints and $\mathcal{H} = \{1, 3, 5, 10, 15, 20\}$ the set of prediction horizons in frames. Our goal is to predict poses $\hat{\mathbf{p}}_{T+h} \in \mathbb{R}^{J\times 3}$ at each horizon $h \in \mathcal{H}$:

$$\hat{\mathbf{p}}_{T+h} \;=\; f_\theta(\{\mathbf{x}_t\}_{t=1}^T), \quad h \in \mathcal{H}. \tag{4}$$

During training, ground-truth observed poses $\mathbf{P}_{1:T} = \{\mathbf{p}_t\}_{t=1}^T$ and future poses $\mathbf{P}_{T+1:T+\zeta}$ where $\zeta = \max(\mathcal{H})$ are available; at inference, only CSI is required. This task is challenging on three fronts: 1) CSI-to-pose mapping is noisy and environment-dependent due to multipath fading and subject-specific body geometry; 2) multi-horizon prediction requires temporal consistency, and autoregressive methods accumulate errors at long horizons; and 3) pose estimation errors propagate into the dynamics module, creating a cascaded-error problem absent from prior work that assumes clean pose input. We therefore seek a unified model that estimates observed poses from CSI, models their temporal dynamics via a Koopman operator with a residual parametrization that prevents identity collapse, and predicts across all horizons without error accumulation.

3.3 Approach Overview

As illustrated in Fig 2, KOALA operates in four sequential stages. *(i)* **CSI Encoding**: a Mamba-based encoder (Gu & Dao, 2024) extracts per-frame features $\{\mathbf{h}_t\}_{t=1}^{T}$ and a global context vector $\mathbf{c}$ via learned attention pooling. *(ii)* **Integrated Pose Estimation and Encoding**: a Mamba-based HPE module recovers estimated 3D poses from CSI features; a skeleton-aware encoder then converts those poses into compact structured features for the dynamics module. *(iii)* **Koopman Dynamics**: a dual-stream fusion combines CSI and pose features; a lifting network maps the result into a higher-dimensional Koopman latent space; a residual CSI-conditioned operator propagates the state across all prediction horizons in a single rollout from the last observed latent state. *(iv)* **Anchor-Based Decoding**: the inverse lifting recovers per-horizon feature vectors from which an output head produces pose *deltas* relative to an anchor pose, and the final prediction is the sum of anchor and delta.

3.4 CSI Encoding

The CSI encoder processes $\mathbf{X} \in \mathbb{R}^{T \times D_c}$ into both per-frame features and a global context vector. Each frame is first projected by a two-layer MLP with GELU activation, then processed by L_c stacked Mamba blocks with residual connections and layer normalization:

$$\mathbf{h}_t^{(0)} = \mathrm{MLP}_{\mathrm{proj}}(\mathbf{x}_t), \tag{5}$$

$$\mathbf{h}_t^{(\ell)} = \mathrm{LN}\big(\mathrm{Mamba}^{(\ell)}(\mathbf{h}_{1:T}^{(\ell-1)})_t + \mathbf{h}_t^{(\ell-1)}\big). \tag{6}$$

Mamba selective state space mechanism provides linear-time processing suited to the long-range correlations in CSI as the body moves through Fresnel zones. The per-frame representations $\mathbf{h}_{1:T}^{(L_c)}$ are shared downstream by both the HPE module and the dual-stream fusion. A learned attention pooling yields the global context vector:

$$\mathbf{c} = \sum_t \alpha_t \, \mathbf{h}_t^{(L_c)}, \qquad \alpha_t = \mathrm{softmax}_t\big(\mathbf{w}^\top \mathbf{h}_t^{(L_c)}\big), \tag{7}$$

where $\mathbf{w} \in \mathbb{R}^d$ is learnable. The context $\mathbf{c}$ exclusively conditions the Koopman operator, implementing the inductive bias that CSI modulates *how* motion evolves rather than *what* the pose is, while avoiding quadratic cost over long sequences.

3.5 CSI Pose Estimation and Encoding

Raw CSI features are mapped to structured pose representations by two stages: a joint-conditioned Mamba HPE module that recovers 3D joint coordinates, followed by a skeleton-aware encoder that maps coordinates into kinematically structured features for the dynamics module. Both stages share the same learnable anatomical type embeddings $\{\mathbf{e}_j^{\mathrm{type}}\}$ and skeleton adjacency bias $\mathbf{G}$, ensuring consistent joint identity and kinematic connectivity throughout. In both stages, the prediction loss is blocked from backpropagating into the HPE module via gradient detachment, isolating pose localization from the prediction objective.

Stage 1: Joint-Conditioned HPE Module. The HPE module converts per-frame CSI features $\{\mathbf{h}_t^{(L_c)}\}_{t=1}^{T}$ into estimated 3D poses $\{\tilde{\mathbf{p}}_t\}_{t=1}^{T}$, where $\tilde{\mathbf{p}}_t \in \mathbb{R}^{J \times 3}$. It is supervised during training by ground-truth observed poses $\{\mathbf{p}_t\}_{t=1}^{T}$ and used at inference as the sole source of pose information.

The module proceeds through four steps.

Step 1: Temporal refinement. A stack of L_p Mamba blocks with residual connections and layer normalization refines the temporal context of the CSI features:

$$\mathbf{h}_t^{\mathrm{hpe}} \;=\; \mathrm{LN}\big(\mathrm{Mamba}_{\mathrm{hpe}}(\mathbf{h}_{1:T}^{(L_c)})_t + \mathbf{h}_t^{(L_c)}\big). \tag{8}$$

Step 2: Joint token construction. Each frame feature is projected and expanded into J joint-specific tokens by adding a learnable anatomical embedding $\mathbf{e}_j^{\mathrm{type}} \in \mathbb{R}^d$, which encodes the anatomical identity of joint j:

$$\mathbf{m}_{t,j} \;=\; \mathrm{MLP}_{\mathrm{expand}}\big(\mathbf{h}_t^{\mathrm{hpe}}\big) + \mathbf{e}_j^{\mathrm{type}}, \quad j = 1, \ldots, J. \tag{9}$$

These embeddings $\{\mathbf{e}_j^{\text{type}}\}_{j=1}^{J}$ are shared with Stage 2, ensuring both stages reason about joints using the same anatomical vocabulary.

Step 3: Per-joint temporal modeling. Each joint's temporal sequence $\{\mathbf{m}_{t,j}\}_{t=1}^{T}$ is processed independently by a shared-weight Mamba block:

$$\mathbf{m}_{t,j}^{\text{out}} \;=\; \text{LN}\big(\text{Mamba}_{\text{joint}}(\mathbf{m}_{1:T,j})_t + \mathbf{m}_{t,j}\big). \tag{10}$$

Sharing weights across joints encourages learning of universal motion priors such as periodicity and velocity. The anatomical embeddings $\mathbf{e}_j^{\text{type}}$ retain joint-specific identity, so the model effectively uses the same temporal dynamics model for all joints but predicts their coordinates separately based on anatomical position.

Step 4: Spatial self-attention and coordinate regression. Within each frame, the J joint tokens attend to one another via skeleton-biased multi-head self-attention. Let $\mathbf{M}_t = [\mathbf{m}_{t,1}^{\text{out}}, \ldots, \mathbf{m}_{t,J}^{\text{out}}]^\top \in \mathbb{R}^{J\times d}$ collect the J joint tokens for frame t. The attention is modulated by a skeleton graph adjacency bias $\mathbf{G} \in \mathbb{R}^{J\times J}$ built from the human kinematic chain:

$$G_{ij} \;=\; \begin{cases} 0 & \text{if joints } i \text{ and } j \text{ are adjacent in the skeleton,} \\ -\beta & \text{otherwise,} \end{cases} \tag{11}$$

where $\beta > 0$ is a fixed suppression constant. The self-attention with multi-head projection matrices $\mathbf{W}_Q^{\text{hpe}}, \mathbf{W}_K^{\text{hpe}}, \mathbf{W}_V^{\text{hpe}} \in \mathbb{R}^{d\times d_h}$ where d_h is the per-head dimension, is then:

$$\mathbf{M}_t^{\text{attn}} \;=\; \text{softmax}\Big(\frac{(\mathbf{M}_t\mathbf{W}_Q^{\text{hpe}})(\mathbf{M}_t\mathbf{W}_K^{\text{hpe}})^\top}{\sqrt{d_h}} + \mathbf{G}\Big)\mathbf{M}_t\mathbf{W}_V^{\text{hpe}}. \tag{12}$$

After a position-wise feed-forward sublayer and layer normalization, a per-joint two-layer MLP with GELU activation regresses the 3D coordinate of each joint:

$$\tilde{\mathbf{p}}_{t,j} \;=\; \text{MLP}_{\text{coord}}([\mathbf{M}_t^{\text{attn}}]_j) \;\in \mathbb{R}^3, \quad j = 1, \ldots, J, \tag{13}$$

where $[\,\cdot\,]_j$ denotes the j-th row. The full estimated pose is $\tilde{\mathbf{p}}_t = [\tilde{\mathbf{p}}_{t,1}, \ldots, \tilde{\mathbf{p}}_{t,J}] \in \mathbb{R}^{J\times 3}$.

Stage 2: Skeleton-Aware Pose Encoding. The skeleton encoder maps a pose sequence into a compact feature per frame that encodes kinematic relationships across joints, for consumption by the Koopman dynamics module. The estimated poses $\tilde{\mathbf{p}}_t$ are detached from the computation graph before entering this stage. This gradient barrier isolates the HPE module (trained solely on $\mathcal{L}_{\text{est}}$) from the prediction objective, preventing the prediction loss from adapting HPE outputs in ways that help downstream prediction at the expense of geometric pose accuracy.

Spatial encoding. Each joint coordinate $\tilde{\mathbf{p}}_{t,j} \in \mathbb{R}^3$ (or its ground-truth counterpart during teacher forcing) is embedded into $\mathbb{R}^d$ and combined with the shared anatomical type embedding:

$$\mathbf{e}_{t,j} \;=\; \text{MLP}_{\text{embed}}(\tilde{\mathbf{p}}_{t,j}) + \mathbf{e}_j^{\text{type}}, \quad j = 1, \ldots, J. \tag{14}$$

Let $\mathbf{E}_t = [\mathbf{e}_{t,1}, \ldots, \mathbf{e}_{t,J}]^\top \in \mathbb{R}^{J\times d}$, inter-joint dependencies are modelled via multi-head self-attention with the same adjacency bias $\mathbf{G}$ (equation 11), using the skeleton encoder's own projection matrices $\mathbf{W}_Q^{\text{skel}}, \mathbf{W}_K^{\text{skel}}, \mathbf{W}_V^{\text{skel}} \in \mathbb{R}^{d\times d_h}$, distinct from those of the HPE module:

$$\mathbf{E}_t^{\text{attn}} \;=\; \text{softmax}\Big(\frac{(\mathbf{E}_t\mathbf{W}_Q^{\text{skel}})(\mathbf{E}_t\mathbf{W}_K^{\text{skel}})^\top}{\sqrt{d_h}} + \mathbf{G}\Big)\mathbf{E}_t\mathbf{W}_V^{\text{skel}}. \tag{15}$$

After a position-wise feed-forward sublayer, the J attended joint embeddings are concatenated and projected to a single per-frame representation:

$$\mathbf{f}_t^{\text{pose}} \;=\; \text{LN}\big(\mathbf{W}_{\text{pool}}[[\mathbf{E}_t^{\text{attn}}]_1; \ldots; [\mathbf{E}_t^{\text{attn}}]_J] + \mathbf{b}_{\text{pool}}\big) \;\in \mathbb{R}^d, \tag{16}$$

where $[;]$ denotes concatenation, $\mathbf{W}_{\text{pool}} \in \mathbb{R}^{d\times Jd}$, and $\mathbf{b}_{\text{pool}} \in \mathbb{R}^d$.

3.6 Dual-Stream Fusion

To ensure the Koopman operator receives informative input even when pose estimates are imperfect, we combine pose-derived features $\mathbf{f}_t^{\text{pose}}$ with CSI per-frame features $\mathbf{h}_t^{(L_c)}$ through additive residual fusion:

$$\mathbf{f}_t = \text{LN}\Big(\mathbf{W}_c\,\mathbf{h}_t^{(L_c)} + \mathbf{W}_u\,\mathbf{f}_t^{\text{pose}} + \text{MLP}_{\text{fuse}}\big(\mathbf{W}_c\,\mathbf{h}_t^{(L_c)} + \mathbf{W}_u\,\mathbf{f}_t^{\text{pose}}\big)\Big). \tag{17}$$

Additive fusion is intentional; it guarantees that the dynamics module receives CSI information directly, so that inaccurate pose estimates cannot completely dominate the representation. The contribution balance between the two streams is monitored during training via the ratio $\|\mathbf{W}_u\mathbf{f}_t^{\text{pose}}\|_2 \,/\, \|\mathbf{W}_c\mathbf{h}_t^{(L_c)}\|_2$, providing a diagnostic of whether the skeleton stream contributes meaningfully throughout training. The fused sequence $\{\mathbf{f}_t\}_{t=1}^T$ is passed through L_t Mamba blocks with residual connections, yielding temporally contextualised features $\{\tilde{\mathbf{f}}_t\}_{t=1}^T$ that capture motion dynamics for the subsequent Koopman lifting.

3.7 Koopman Lifting and Dynamics

Lifting Network. The lifting $\phi : \mathbb{R}^d \to \mathbb{R}^{D_z}$ maps temporally contextualized fused features into a higher-dimensional Koopman space where linear dynamics hold. Its inverse $\phi^{-1} : \mathbb{R}^{D_z} \to \mathbb{R}^d$ recovers feature-space representations. Both are three-layer MLPs with GELU, dropout, and layer normalization:

$$\mathbf{z}_t = \phi(\tilde{\mathbf{f}}_t) = \text{LN}\big(\text{MLP}_{\text{enc}}(\tilde{\mathbf{f}}_t)\big) \in \mathbb{R}^{D_z}, \tag{18}$$

$$\hat{\tilde{\mathbf{f}}}_t = \phi^{-1}(\mathbf{z}_t) = \text{LN}\big(\text{MLP}_{\text{dec}}(\mathbf{z}_t)\big) \in \mathbb{R}^d. \tag{19}$$

Setting $D_z > d$ is grounded in Koopman theory: the true operator is infinite-dimensional (Koopman, 1931), and any finite-dimensional approximation in $\mathbb{R}^{D_z}$ incurs a projection error that decreases as D_z grows (Brunton & Kutz, 2019; Strässer et al., 2023). We set $D_z = 2d$, providing additional degrees of freedom for the learned eigenfunctions to disentangle nonlinear modes that are coupled in the original feature coordinates.

Proposition 1 (Lifting information preservation)**.** *If the reconstruction error $\|\phi^{-1}(\phi(\tilde{\mathbf{f}}_t)) - \tilde{\mathbf{f}}_t\|^2 \le \epsilon_r$ for all training features, then ϕ is approximately injective on the training manifold: for any two training features $\tilde{\mathbf{f}}_a, \tilde{\mathbf{f}}_b$,*

$$\|\tilde{\mathbf{f}}_a - \tilde{\mathbf{f}}_b\| \;\le\; \|\phi^{-1}(\phi(\tilde{\mathbf{f}}_a)) - \phi^{-1}(\phi(\tilde{\mathbf{f}}_b))\| + 2\sqrt{\epsilon_r}. \tag{20}$$

In particular, if two feature vectors are separated by more than $2\sqrt{\epsilon_r}$, their lifted representations must be distinct (proof in Appendix A).

Residual CSI-Conditioned Operator. A fundamental difficulty in learning Koopman operators with gradient descent is the *identity attractor*. When a dense matrix $\mathbf{K}$ is initialized near identity, the prediction loss for slowly-varying motion is already low, so gradients provide little pressure to push $\mathbf{K}$ away from $\mathbf{I}$. Meanwhile, reconstruction and linearity losses actively penalize deviations from $\mathbf{K} \approx \mathbf{I}$, thereby keeping the operator uninformative throughout training.

KOALA resolves this by parametrizing the operator in **residual form**:

$$\mathbf{K}_{\text{eff}} \;=\; \mathbf{I} + \mathbf{B} + \gamma\,\mathbf{U}(\mathbf{c})\,\mathbf{V}(\mathbf{c})^\top, \tag{21}$$

where $\mathbf{B} \in \mathbb{R}^{D_z \times D_z}$ is the learned residual dynamics matrix (we use $\mathbf{B}$ to distinguish it from the skeleton adjacency bias $\mathbf{G}$ of Section 3.5), $\mathbf{U}(\mathbf{c}), \mathbf{V}(\mathbf{c}) \in \mathbb{R}^{D_z \times r}$ are low-rank CSI adaptation factors produced by two-layer MLPs from the global context $\mathbf{c}$, and $\gamma > 0$ is a learnable scalar parametrised in log-space as $\gamma = \exp(\xi)$. The identity is baked into the architecture rather than enforced during initialization, so $\mathbf{B}$ faces no competing pressure to remain near zero. The log-space parametrization of γ prevents the CSI adaptation from collapsing to zero as gradients weaken, keeping the WiFi conditioning active throughout training.

The rank constraint $r \ll D_z$ ensures CSI modulates a low-dimensional subspace of the dynamics, capturing environment-dependent multipath effects without dominating the base residual dynamics. The low-rank factors $\mathbf{U}(\mathbf{c})$ and $\mathbf{V}(\mathbf{c})$ are computed once per sample and reused across all rollout steps. The one-step update is:

$$\mathbf{z}_{t+1} \;=\; \mathbf{z}_t + \mathbf{B}\,\mathbf{z}_t + \gamma\,\mathbf{U}(\mathbf{c})\big(\mathbf{V}(\mathbf{c})^\top \mathbf{z}_t\big). \tag{22}$$

Proposition 2 (Operator stability)**.** *Let $\mathbf{K}_{\text{eff}}$ be defined as in equation 21 with $\|\mathbf{B}\|_F \leq \eta_B$. Then $\|\mathbf{K}_{\text{eff}}\|_2 \leq 1+\eta_B+\gamma\|\mathbf{U}\|_2\|\mathbf{V}\|_2$, and for h-step rollouts $\|\mathbf{K}_{\text{eff}}^h\|_2 \leq (1+\eta_B+\gamma\|\mathbf{U}\|_2\|\mathbf{V}\|_2)^h$. With η_B small at initialization and $\gamma \approx 0.1$, the amplification factor at $h = 20$ remains bounded during early training, preventing runaway gradient accumulation through long operator chains. The Frobenius norm $\|\mathbf{B}\|_F$ is monitored as a training diagnostic to track operator growth.*

Anchor-Based Multi-Horizon Prediction. To predict at horizons $\mathcal{H}$, KOALA iterates the operator from the last lifted state $\mathbf{z}_T$:

$$\mathbf{z}_{T+k} \;=\; \mathbf{K}_{\text{eff}}\,\mathbf{z}_{T+k-1}, \quad k = 1, \ldots, \zeta. \tag{23}$$

Rather than decoding absolute poses, KOALA predicts pose *deltas* relative to an anchor pose $\bar{\mathbf{p}}$, the last observed pose during teacher forcing or the last estimated pose at inference. Denoting the delta head $\Delta_\theta = \text{MLP}_{\text{out}} \circ \phi^{-1}$:

$$\hat{\mathbf{p}}_{T+h} \;=\; \bar{\mathbf{p}} + \Delta_\theta(\mathbf{z}_{T+h}), \quad h \in \mathcal{H}. \tag{24}$$

This formulation eliminates a common degenerate solution: without the anchor, outputting the current pose at every horizon incurs nearly zero loss for slow motions. Under the anchor formulation, $\Delta_\theta(\mathbf{z}_{T+h}) \approx \mathbf{0}$ for all h incurs proportional loss at long horizons where the true pose has drifted from the anchor, so the operator is forced to encode genuine temporal dynamics. The encoding pipeline executes once per sample; each additional prediction horizon costs only one matrix-vector product.

The following theorem formalizes why linear rollout in the Koopman space produces bounded multi-horizon predictions, while the same rollout in observation space does not.

Theorem 1 (Bounded multi-horizon prediction)**.** *Let $\mathbf{v}_{t+1} = F(\mathbf{v}_t)$ be the nonlinear dynamics governing human motion in observation space. Suppose the lifting ϕ satisfies the approximate Koopman invariance condition*

$$\|\phi(F(\mathbf{v})) - \mathbf{K}_{\text{eff}}\phi(\mathbf{v})\| \;\leq\; \varepsilon \quad \textit{for all observed states } \mathbf{v},$$

and that $\kappa := \|\mathbf{K}_{\text{eff}}\|_2 \leq 1 + \eta$ with $\eta \ll 1$ (an empirical condition monitored via $\|\mathbf{B}\|_F$ during training; see Proposition 2). Let L be the Lipschitz constant of the delta head $\Delta_\theta = \text{MLP}_{\text{out}} \circ \phi^{-1}$. Then the h-step prediction error in observation space satisfies

$$\|\hat{\mathbf{p}}_{T+h} - \mathbf{p}_{T+h}\| \;\leq\; L\,\frac{\kappa^h - 1}{\kappa - 1}\,\varepsilon \;+\; L\,\delta_h \;+\; \|\bar{\mathbf{p}} - \mathbf{p}_T\|, \tag{25}$$

where $\delta_h = \|\phi(\tilde{\mathbf{f}}_{T+h}) - \mathbf{z}^*_{T+h}\|$ is the Koopman invariance defect at horizon h, with $\tilde{\mathbf{f}}_{T+h}$ denoting the temporal-encoder feature at future frame $T + h$ and $\|\bar{\mathbf{p}} - \mathbf{p}_T\|$ is the anchor alignment error (zero under teacher forcing). When $\kappa \leq 1+\eta$ with small η, the geometric factor satisfies $(\kappa^h - 1)/(\kappa - 1) \leq h(1+\eta h/2)$, yielding near-linear growth:

$$\|\hat{\mathbf{p}}_{T+h} - \mathbf{p}_{T+h}\| \;\lesssim\; L\,h\,\varepsilon\,(1 + \eta h/2) \;+\; L\,\delta_h \;+\; \|\bar{\mathbf{p}} - \mathbf{p}_T\|. \tag{26}$$

In contrast, an autoregressive predictor iterating a decoder in observation space with per-step error ε_{ar} incurs

$$\|\hat{\mathbf{p}}^{\text{ar}}_{T+h} - \mathbf{p}_{T+h}\| \;\geq\; h\,\varepsilon_{\text{ar}} - O(h^2\varepsilon_{\text{ar}}^2). \tag{27}$$

The key asymmetry is that ε is measured in a latent space explicitly shaped to minimize Koopman invariance defect via the KPC objective (equation 31), whereas ε_{ar} accumulates in the raw pose observation space where no such regularisation applies. Consequently $\varepsilon \ll \varepsilon_{\text{ar}}$, and the bound tightens as $\mathcal{L}_{\text{KPC}}$ decreases during training. (Proof and supporting remarks in Appendix B.)

3.8 Loss Design

Our training objective comprises three terms that jointly shape the Koopman latent space, supervise pose estimation, and enforce multi-horizon consistency.

Prediction Loss. The prediction loss penalizes pose-space error at each horizon with weights w_h that emphasize long horizons:

$$\mathcal{L}_{\text{pred}} \;=\; \frac{1}{\sum_h w_h} \sum_{h \in \mathcal{H}} w_h \left\| \hat{\mathbf{p}}_{T+h} - \mathbf{p}_{T+h} \right\|_2^2. \tag{28}$$

Training simultaneously across all horizons enforces operator consistency: since multi-step predictions compose the same operator $\mathbf{K}_{\text{eff}}$, gradients from every horizon collectively constrain a single set of parameters.

Koopman Predictive Coding Loss. The KPC loss regularises the lifting and operator by requiring that operator rollouts match the temporal-encoder features the model would produce upon observing future frames. Operating in this feature space $\mathbb{R}^d$ rather than in raw observation or Koopman latent space decouples the supervision signal from CSI noise while maintaining gradient pressure through ϕ and $\mathbf{K}_{\text{eff}}$.

Future feature targets. Let $\mathbf{f}^*_{T+h} \in \mathbb{R}^d$ denote the temporal-encoder feature produced by processing the concatenated observation-and-future pose sequence through the fusion and temporal encoder. Under a no-gradient context, these targets are computed by a single extended forward pass:

$$\left[\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_T, \mathbf{f}^*_{T+1}, \ldots, \mathbf{f}^*_{T+\zeta}\right] \;=\; \text{TemporalEnc}\Big(\text{Fusion}\big(\mathbf{h}^{\text{pad}}_{1:T+\zeta},\, \mathbf{p}^{\text{ext}}_{1:T+\zeta}\big)\Big), \tag{29}$$

where $\mathbf{h}^{\text{pad}}_{1:T+\zeta}$ pads the last CSI memory frame for future positions (no future CSI is available), and $\mathbf{p}^{\text{ext}}_{1:T+\zeta} = [\mathbf{p}_{1:T}, \mathbf{p}_{T+1:T+\zeta}]$ uses ground-truth future poses. The extended forward pass is run under `torch.no_grad()`, so the targets carry no gradient.

Anchored KPC formulation. Let $\mathbf{z}_T = \phi(\tilde{\mathbf{f}}_T)$ be the last lifted observation state, $\mathbf{r}_T = \phi^{-1}(\mathbf{z}_T)$ the no-rollout reconstruction, and $\tilde{\mathbf{f}}^{\text{pred}}_{T+h} = \phi^{-1}(\mathbf{K}^h_{\text{eff}}\, \mathbf{z}_T)$ the feature predicted by rolling out the operator h steps. KPC measures how much the operator-predicted feature *changes* from the current reconstruction and compares this against the actual feature change:

$$\mathbf{f}^{\text{KPC}}_{T+h} \;=\; \tilde{\mathbf{f}}_T + \big(\tilde{\mathbf{f}}^{\text{pred}}_{T+h} - \mathbf{r}_T\big). \tag{30}$$

The KPC loss combines a k=0 reconstruction term with horizon-weighted k>0 prediction terms:

$$\mathcal{L}_{\text{KPC}} \;=\; \frac{1}{W}\left[\|\mathbf{r}_T - \tilde{\mathbf{f}}_T\|_2^2 \;+\; \sum_{h \in \mathcal{H}} w^{\text{KPC}}_h \left\|\mathbf{f}^{\text{KPC}}_{T+h} - \mathbf{f}^*_{T+h}\right\|_2^2\right], \tag{31}$$

where $w^{\text{KPC}}_h = 1/(1{+}h)$ and $W = 1 + \sum_h w^{\text{KPC}}_h$ is a normalising constant. The horizon weighting is intentionally *inverted* relative to $\mathcal{L}_{\text{pred}}$: short horizons dominate the KPC gradient signal. This is critical for stability, as the h-step operator chain $\mathbf{K}^h_{\text{eff}}$ amplifies gradient norms geometrically in h, so up-weighting long horizons in the KPC loss would produce unbounded gradients at initialization. Instead, $\mathcal{L}_{\text{pred}}$ handles long-horizon supervision while KPC provides stable gradient flow through the lifting and operator at all timescales.

The anchoring also eliminates the large-magnitude baseline at initialization. At initialisation $\mathbf{K}_{\text{eff}} \approx \mathbf{I}$, so $\tilde{\mathbf{f}}^{\text{pred}}_{T+h} \approx \mathbf{r}_T$ and $\mathbf{f}^{\text{KPC}}_{T+h} \approx \tilde{\mathbf{f}}_T$. The k>0 loss is then bounded by $\|\mathbf{f}^*_{T+h} - \tilde{\mathbf{f}}_T\|^2$, which is small because adjacent features differ little, preventing the gradient explosion through 20-step operator chains that occurs with an unanchored formulation.

CSI Pose Estimation Loss. The HPE module is supervised with ground-truth observed poses:

$$\mathcal{L}_{\text{est}} \;=\; \frac{1}{T}\sum_{t=1}^{T} \|\tilde{\mathbf{p}}_t - \mathbf{p}_t\|_2^2. \tag{32}$$

This is the sole training signal for the HPE module; because estimated poses are detached before entering the skeleton encoder, the prediction loss cannot bias the HPE module towards features that help prediction at the cost of localization accuracy.

Table 1: Performance comparison across prediction horizons on the WiPose dataset.

Methods	**Metrics**	**100ms**	**300ms**	**500ms**	**1000ms**
MetaFi++ (Zhou et al., 2023)	MPJPE ↓	42.74	41.78	40.62	38.64
	PA-MPJPE ↓	26.62	26.21	26.22	25.74
HPE-Li (D. Gian et al., 2024)	MPJPE ↓	40.00	39.64	39.98	39.69
	PA-MPJPE ↓	26.69	26.60	26.92	27.01
ConvLSTM (Shi et al., 2015)	MPJPE ↓	66.44	65.91	66.02	66.57
	PA-MPJPE ↓	37.00	36.96	36.87	37.15
VMRNN (Tang et al., 2024)	MPJPE ↓	32.20	31.68	31.28	31.22
	PA-MPJPE ↓	23.36	23.37	23.32	23.30
LTD (Mao et al., 2019)	MPJPE ↓	42.07	41.00	40.38	39.35
	PA-MPJPE ↓	28.57	28.41	28.44	28.38
siMLPle (Guo et al., 2023)	MPJPE ↓	39.54	39.41	39.24	39.14
	PA-MPJPE ↓	26.62	26.95	27.07	26.60
KOALA (Ours)	MPJPE ↓	26.14	26.03	26.49	27.28
	PA-MPJPE ↓	20.81	20.77	20.98	21.28

Table 2: Evaluation performance across protocols and settings at different horizons (ms) on the MM-Fi.

Setting	**Metric**	**Protocol 1**						**Protocol 2**						**Protocol 3**					
		100	300	500	1000	1500	2000	100	300	500	1000	1500	2000	100	300	500	1000	1500	2000
S1	MPJPE ↓	54.5	58.2	62.5	66.7	65.0	63.6	53.2	56.9	60.8	64.0	62.1	61.0	55.2	62.3	69.2	76.2	73.3	69.8
	PA-MPJPE ↓	45.5	48.0	51.2	54.1	53.4	52.6	44.3	46.6	49.1	51.0	50.2	49.8	46.5	52.6	58.3	64.3	63.2	60.1
	PCK@20 ↑	91.0	89.5	87.6	85.6	86.5	87.3	91.3	89.9	88.2	86.6	87.5	88.3	90.6	87.8	85.0	82.3	83.7	85.3
	PCK@10 ↑	72.5	69.9	67.3	64.7	65.9	67.1	73.1	70.6	68.1	66.0	67.2	68.3	72.0	68.3	65.1	62.3	64.1	65.8
S2	MPJPE ↓	55.7	62.0	68.4	74.7	71.9	68.7	52.8	58.2	63.5	68.0	65.4	63.5	53.1	59.0	64.5	68.9	66.4	64.4
	PA-MPJPE ↓	47.3	52.5	57.7	62.8	61.8	59.2	44.0	48.0	51.8	54.7	53.5	52.6	44.7	49.0	52.9	55.5	54.4	53.5
	PCK@20 ↑	90.3	87.8	85.2	82.6	84.1	85.6	91.5	89.4	87.2	85.2	86.3	87.4	91.3	89.1	86.7	84.6	85.8	86.9
	PCK@10 ↑	71.7	68.5	65.3	62.6	64.4	66.2	73.5	70.1	67.1	64.2	65.9	67.4	73.4	69.8	66.7	63.8	65.5	67.2
S3	MPJPE ↓	52.1	57.0	61.9	65.7	63.4	61.9	51.5	54.6	58.5	62.2	60.6	59.4	52.6	58.3	63.2	67.2	64.5	62.5
	PA-MPJPE ↓	43.5	47.0	50.5	52.8	51.8	51.1	42.8	44.9	47.6	50.0	49.2	48.9	44.2	48.5	52.0	54.3	52.9	51.7
	PCK@20 ↑	91.8	89.9	87.8	85.9	86.9	87.9	92.0	90.9	89.3	87.6	88.3	88.9	91.6	89.3	87.2	85.4	86.5	87.7
	PCK@10 ↑	73.9	70.7	67.8	65.1	66.5	67.7	74.3	72.1	69.6	67.0	68.1	69.5	73.7	70.2	67.3	64.6	66.3	67.8

Total Loss. The total loss function can be expressed by the following:

$$\mathcal{L} \;=\; \lambda_p \, \mathcal{L}_{\text{pred}} \;+\; \lambda_k \, \mathcal{L}_{\text{KPC}} \;+\; \lambda_e \, \mathcal{L}_{\text{est}}, \tag{33}$$

with $\lambda_p = 1.5$, $\lambda_k = 0.5$, $\lambda_e = 1.0$. This three-term objective is more compact than the seven-term objectives used in prior Koopman-based motion prediction methods, which required separate linearity, contrastive, reconstruction, auxiliary, and spectral-stability losses. The KPC loss subsumes reconstruction (the k=0 term), Koopman linearity (via feature-space prediction), and the stability function previously served by InfoNCE, without requiring batch-negative mining or spectral norm estimation at each training step.

4 Experiments

4.1 Datasets

We evaluate on two WiFi-based datasets: MM-Fi (Yang et al., 2023b) and WiPose (Zhou et al., 2022). MM-Fi provides large-scale multimodal data with synchronized CSI streams across 40 subjects, 27 actions, and 4 environments under three protocols (P1: daily activities; P2: rehabilitation exercises; P3: all 27 actions) and three settings (S1: random split; S2: cross-subject; S3: cross-environment). WiPose provides fine-grained

Table 3: Performance comparison across prediction horizons on the MM-Fi dataset.

Methods	**Metrics**	**100ms**	**300ms**	**500ms**	**1000ms**	**1500ms**	**2000ms**
HPE-Li (D. Gian et al., 2024)	MPJPE ↓	375.2	383.4	351.6	351.2	339.8	368.5
	PA-MPJPE ↓	120.7	122.3	119.4	118.4	119.5	122.8
	PCK@20 ↑	3.8	4.3	7.0	8.6	10.0	6.3
	PCK@10 ↑	0.3	0.2	0.8	1.0	1.1	0.5
MetaFi++ (Zhou et al., 2023)	MPJPE ↓	311.7	308.4	311.2	312.8	316.9	312.0
	PA-MPJPE ↓	105.5	105.4	104.6	104.7	105.1	105.1
	PCK@20 ↑	78.8	80.7	78.4	78.4	75.9	80.1
	PCK@10 ↑	18.1	17.7	18.4	17.2	18.2	16.8
ConvLSTM (Shi et al., 2015)	MPJPE ↓	345.0	340.9	338.7	334.9	334.3	342.1
	PA-MPJPE ↓	102.9	102.8	102.7	102.6	102.8	102.5
	PCK@20 ↑	44.78	47.04	49.17	51.74	51.71	46.89
	PCK@10 ↑	3.7	4.2	4.7	5.5	5.7	4.14
MIM (Wang et al., 2019)	MPJPE ↓	335.6	337.5	340.1	339.8	338.5	337.3
	PA-MPJPE ↓	102.9	102.8	102.6	102.7	102.5	102.1
	PCK@20 ↑	53.85	52.75	49.87	49.44	51.64	52.90
	PCK@10 ↑	6.2	5.7	5.4	5.20	5.9	5.8
SwinLSTM (Tang et al., 2023)	MPJPE ↓	332.6	330.8	330.9	335.5	334.70	340.8
	PA-MPJPE ↓	101.7	101.8	101.9	101.8	101.8	102.3
	PCK@20 ↑	53.4	53.3	53.8	51.1	51.8	49.4
	PCK@10 ↑	5.7	5.8	5.72	4.9	5.4	4.7
VMRNN (Tang et al., 2024)	MPJPE ↓	338.2	340.1	343.1	349.9	343.8	333.7
	PA-MPJPE ↓	102.6	101.9	101.9	102.0	101.8	102.1
	PCK@20 ↑	48.1	48.7	47.5	43.9	48.5	52.4
	PCK@10 ↑	7.2	7.8	7.6	7.4	7.9	8.3
LTD (Mao et al., 2019)	MPJPE ↓	365.4	362.7	359.6	364.7	368.5	368.9
	PA-MPJPE ↓	107.99	108.19	108.41	108.13	107.94	107.88
	PCK@20 ↑	34.3	35.3	36.5	32.92	32.1	33.6
	PCK@10 ↑	2.41	2.72	3.06	2.40	2.23	2.47
siMLPe (Guo et al., 2023)	MPJPE ↓	348.7	357.1	365.7	376.9	383.6	380.4
	PA-MPJPE ↓	106.3	105.8	104.2	107.4	107.4	108.2
	PCK@20 ↑	28.34	28.25	25.07	20.75	18.74	20.14
	PCK@10 ↑	2.82	3.46	3.35	2.98	2.61	2.10
AuxFormer (Xu et al., 2023)	MPJPE ↓	373.2	385.9	382.5	357.2	394.5	367.6
	PA-MPJPE ↓	103.6	101.8	102.2	102.8	101.4	101.49
	PCK@20 ↑	18.6	15.5	18.5	24.75	11.74	21.75
	PCK@10 ↑	1.53	1.89	1.75	2.90	1.06	3.07
KOALA (Ours)	MPJPE ↓	52.1	57.0	61.9	65.7	63.4	61.9
	PA-MPJPE ↓	43.5	47.0	50.5	52.8	51.8	51.1
	PCK@20 ↑	91.8	89.9	87.8	85.9	86.9	87.9
	PCK@10 ↑	73.9	70.7	67.8	65.1	66.5	67.7

CSI data with skeletal annotations for 2D pose understanding. Following the standard evaluation protocols, we evaluate on MPJPE and PA-MPJPE in millimeters, alongside PCKs of 20% and 10%.

4.2 2D Motion Prediction

Table 1 reports results on the WiPose dataset. KOALA achieves the lowest MPJPE and PA-MPJPE at every horizon, reaching 26.14 mm at 100 ms and 27.28 mm at 1000 ms, an improvement of approximately 19% over the strongest baseline VMRNN. Notably, KOALA degradation across horizons is only 1.14 mm

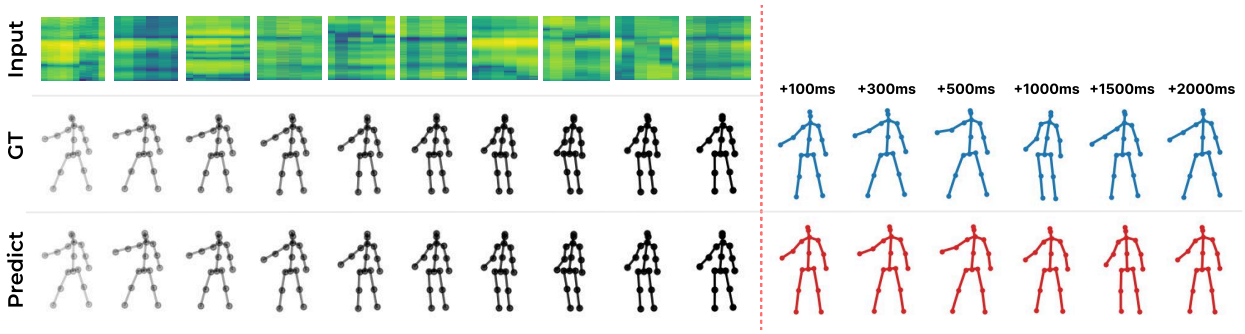

Figure 3: Qualitative results between predicted poses and Ground Truth (GT) for different time steps.

in MPJPE, showing that the Koopman operator propagates latent states without meaningful error accumulation. Among baselines, ConvLSTM performs considerably worse, confirming that naive spatiotemporal convolution struggles in the lower-dimensional 2D pose space, while trajectory-based methods (LTD, siMLPe) and CSI-specialized architectures (MetaFi++, HPE-Li) cluster in the 39-42 mm range, unable to close the gap with state-space approaches.

4.3 3D Motion Prediction

Table 2 reports KOALA on MM-Fi across three protocols, three settings, and six horizons. MPJPE grows moderately with horizon, with PA-MPJPE below 65.3 mm and PCK@20 above 82.3% across all configurations. S3 achieves the lowest short-horizon errors (52.1 mm under P1 at 100 ms), while S2 is the most challenging (74.7 mm at 1000 ms under P1) due to unseen body geometries. Protocol 2 yields the lowest errors under S2 and S3 owing to its constrained action set, whereas Protocol 3 maintains PCK@20 above 85.4% at 1000 ms despite covering all 27 categories. Figure 3 confirms coherent rollouts with anatomically plausible skeletons at all horizons (qualitative results in Appendix D; ablations in Appendix E).

Table 3 compares KOALA against nine baselines on MM-Fi, spanning WiFi pose estimation methods (HPE-Li, MetaFi++), spatiotemporal prediction models (ConvLSTM, MIM, SwinLSTM, VMRNN), and skeleton-based methods (LTD, siMLPe, AuxFormer). KOALA outperforms all baselines by a substantial margin across every metric and horizon: at 100 ms it achieves 52.1 mm MPJPE, roughly 6.4× lower than the best competing method SwinLSTM (332.6 mm) and 73.9% PCK@10 while no baseline exceeds 18.1%. Among WiFi baselines, MetaFi++ consistently outperforms HPE-Li, confirming that stronger pose backbones yield more temporally stable estimates. Skeleton-based methods perform poorly under CSI-derived pose noise, whereas spatiotemporal models exhibit near-flat error profiles across horizons, suggesting mean regression rather than genuine temporal modeling. KOALA, by contrast, exhibits the expected monotonic increase in error with horizon, reflecting real-world temporal dynamics.

5 Conclusion

We presented KOALA, a novel framework for anticipatory human motion prediction directly from WiFi CSI. By lifting noisy CSI-derived pose sequences into a learned Koopman latent space, KOALA achieves multi-horizon prediction through simple matrix-vector products without autoregressive iteration. The residual parametrization resolves the identity attractor, the anchor-delta head eliminates mean-pose collapse, and the KPC loss enforces dynamical consistency without contrastive, spectral, or auxiliary losses. Experiments on MM-Fi and WiPose demonstrate substantial improvements over all baselines, with MPJPE increasing nearly linearly across horizons. The current single-person formulation does not address signal superposition in multi-occupant settings. The CSI-conditioned operator partially accounts for environment variation but has not been validated under zero-shot cross-site transfer. Finally, the Koopman assumption of approximately time-invariant dynamics within the observation window may break down at abrupt action transitions, a direction that is addressable through mixture-of-operators or mode-switching extensions.

References

Ieee standard for information technology–telecommunications and information exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications amendment 2: Sub 1 ghz license exempt operation. *IEEE Std 802.11ah-2016 (Amendment to IEEE Std 802.11-2016, as amended by IEEE Std 802.11ai-2016)*, pp. 1–594, 2017.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

Yi-Chung Chen, Zhi-Kai Huang, Lu Pang, Jian-Yu Jiang-Lin, Chia-Han Kuo, Hong-Han Shuai, and Wen-Huang Cheng. Seeing the unseen: Wifi-based 2d human pose estimation via an evolving attentive spatial-frequency network. *Pattern Recognition Letters*, 171:21–27, 2023. ISSN 0167-8655.

Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXI*, pp. 93–111, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72750-4.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11447–11456, 2021.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:815–36, 04 2005.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4798–4808, 01 2023.

Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A. McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLII*, pp. 72–91, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72945-4.

Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5308–5317, 2016.

Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370851.

B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Transactions on Image Processing*, 30:7760–7775, 2021.

Bethany Lusch, J Nathan Wehmeyer, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018.

Kedi Lyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, and Ruili Wang. 3d human motion prediction: A survey. *Neurocomputing*, 489:345–365, 2022. ISSN 0925-2312.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9489–9497, 2019.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, pp. 474–489, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58567-9.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4674–4683, 2017.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: a machine learning approach for precipitation nowcasting. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 802–810, Cambridge, MA, USA, 2015. MIT Press.

Robin Strässer, Julian Berberich, and Frank Allgöwer. Robust data-driven control for nonlinear systems using the koopman operator*. *IFAC-PapersOnLine*, 56(2):2257–2262, 2023. ISSN 2405-8963. 22nd IFAC World Congress.

Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. SwinLSTM: Improving Spatiotemporal Prediction Accuracy using Swin Transformer and LSTM. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13424–13433, Los Alamitos, CA, USA, October 2023. IEEE Computer Society.

Yujin Tang, Peijie Dong, Zhenheng Tang, Xiaowen Chu, and Junwei Liang. VMRNN: Integrating Vision Mamba and LSTM for Efficient and Accurate Spatiotemporal Forecasting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5663–5673, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, pp. 65–76, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336192.

Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9146–9154, 2019.

Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9475–9486, 2023.

Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 969–978, 2024.

Jianfei Yang, Yunjiao Zhou, He Huang, Han Zou, and Lihua Xie. Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation. *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, pp. 1–6, 2022.

Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing. *Patterns*, 4(3):100703, 2023a. ISSN 2666-3899.

Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023b.

Yue Zhou, Aichun Zhu, Caojie Xu, Fangqiang Hu, and Yifeng Li. Perunet: Deep signal channel attention in unet for wifi-based human pose estimation. *IEEE Sensors Journal*, 22(20):19750–19760, 2022.

Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023.

Yunjiao Zhou, Jianfei Yang, He Huang, and Lihua Xie. Adapose: Toward cross-site device-free human pose estimation with commodity wifi. *IEEE Internet of Things Journal*, 11(24):40255–40267, 2024.

APPENDIX

A Proof of Proposition 1

Proof. Fix any two training features $\tilde{\mathbf{f}}_a, \tilde{\mathbf{f}}_b \in \mathbb{R}^d$. By the triangle inequality:

$$\|\tilde{\mathbf{f}}_a - \tilde{\mathbf{f}}_b\| \leq \|\tilde{\mathbf{f}}_a - \phi^{-1}(\phi(\tilde{\mathbf{f}}_a))\| + \|\phi^{-1}(\phi(\tilde{\mathbf{f}}_a)) - \phi^{-1}(\phi(\tilde{\mathbf{f}}_b))\| + \|\phi^{-1}(\phi(\tilde{\mathbf{f}}_b)) - \tilde{\mathbf{f}}_b\|.$$

The first and third terms are individual round-trip reconstruction errors. Since $\mathcal{L}_{\text{rec}} \leq \epsilon_r$ implies each round-trip error is at most $\sqrt{\epsilon_r}$ on the training manifold, both terms are bounded by $\sqrt{\epsilon_r}$, giving:

$$\|\tilde{\mathbf{f}}_a - \tilde{\mathbf{f}}_b\| \leq \|\phi^{-1}(\phi(\tilde{\mathbf{f}}_a)) - \phi^{-1}(\phi(\tilde{\mathbf{f}}_b))\| + 2\sqrt{\epsilon_r}. \tag{34}$$

Therefore, whenever $\|\tilde{\mathbf{f}}_a - \tilde{\mathbf{f}}_b\| > 2\sqrt{\epsilon_r}$, the middle term must be strictly positive, which implies $\phi(\tilde{\mathbf{f}}_a) \neq \phi(\tilde{\mathbf{f}}_b)$. Hence ϕ is approximately injective on the training manifold, with injectivity guaranteed for any pair separated by more than $2\sqrt{\epsilon_r}$. □

B Proof sketch of Theorem 1

Proof sketch. Telescoping the h-step composition in the Koopman space gives

$$\mathbf{K}_{\text{eff}}^h \mathbf{z}_T - \mathbf{z}_{T+h}^* \;=\; \sum_{k=0}^{h-1} \mathbf{K}_{\text{eff}}^{h-1-k} \Big(\mathbf{K}_{\text{eff}} \mathbf{z}_{T+k} - \mathbf{z}_{T+k+1}\Big) + \Big(\mathbf{z}_{T+h} - \mathbf{z}_{T+h}^*\Big),$$

where $\mathbf{z}_{T+h}^* = \phi(\tilde{\mathbf{f}}_{T+h})$ is the true lifted state. Taking norms, applying the Koopman invariance condition and the spectral bound $\|\mathbf{K}_{\text{eff}}^k\|_2 \leq \kappa^k$, and using the Lipschitz property of Δ_θ yields equation 25. The linear simplification follows from $\kappa^h = 1 + \eta h + O(\eta^2 h^2)$ when ηh is small, giving $(\kappa^h - 1)/(\kappa - 1) \lesssim h$. For the autoregressive lower bound in equation 27, each step feeds its own noisy prediction back as input, so errors accumulate additively at leading order with no latent-space regularisation to suppress ε_{ar}. □

Remark 1. The condition $\kappa \leq 1 + \eta$ is an empirical property of the trained operator rather than an architectural guarantee. Unlike formulations that enforce a hard spectral constraint via projected gradient or power-iteration penalties, KOALA monitors $\|\mathbf{B}\|_F$ as a diagnostic and relies on the KPC loss and residual initialization to keep κ near 1. This is sufficient in practice, as training curves show $\|\mathbf{B}\|_F$ remaining moderate throughout, but the bound in equation 25 should be understood as conditioned on this empirical regularity rather than as a hard guarantee.

C Implementation Details

All models are trained on a single NVIDIA RTX 5090 GPU with the AdamW optimizer, batch size 8, and gradient clipping at 1.0 with the observation window $T = 10$ frames. Multi-horizon predictions are evaluated at $\mathcal{H} = \{1, 3, 5, 10, 15, 20\}$ frames, corresponding to 100-2000 ms at the MM-Fi sampling rate, with maximum horizon $\zeta = \max(\mathcal{H}) = 20$ frames. The model dimension is $d = 128$, Koopman latent dimension $D_z = 256$, and low-rank rank $r = 16$. The CSI encoder uses $L_c = 4$ Mamba layers; the temporal encoder uses $L_t = 2$ Mamba layers; the HPE module uses $L_p = 2$ temporal Mamba layers and $L_j = 1$ per-joint temporal block.

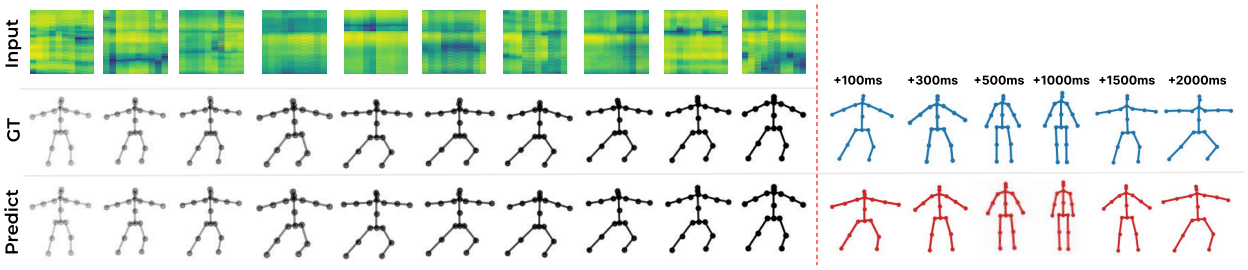

Figure 4: Additional qualitative results between predicted poses and GT for different time steps.

All Mamba blocks use state dimension 16 and convolution width 4. Multi-head attention in both the HPE module and skeleton encoder uses $n_{\text{head}} = 4$ heads with per-head dimension $d_h = d/n_{\text{head}} = 32$. Horizon weights in $\mathcal{L}_{\text{pred}}$ are set to $w_h \in \{0.3, 0.5, 0.8, 1.2, 1.5, 2.0\}$ for $h \in \{1, 3, 5, 10, 15, 20\}$, emphasising long-horizon consistency. The residual dynamics matrix $\mathbf{B}$ is initialized with entries of scale $0.5/D_z$, and its Frobenius norm $\|\mathbf{B}\|_F$ is logged as a diagnostic at every epoch.

D Qualitative Results

Figures 3 and 4 visualise two representative sequences from MM-Fi, showing input CSI spectrograms, ground-truth future poses (blue), and KOALA predictions (red) across horizons from +100 ms to +2000 ms.

In the first sequence in Figure 3, the subject performs a relatively constrained motion involving gradual arm raising. At short horizons (+100 ms to +500 ms), KOALA predictions closely align with the ground truth in both global body configuration and individual joint positions. At longer horizons (+1000 ms to +2000 ms), minor deviations appear in distal joints such as the wrists and ankles, yet the overall body structure remains anatomically plausible with no structural collapse, confirming that the Koopman operator maintains coherent latent dynamics throughout the rollout.

The second sequence, as illustrated in Figure 4 involves more dynamic motion with wider limb displacement and greater inter-frame variability, presenting a substantially harder prediction target. Despite this increased articulation complexity, KOALA accurately tracks the global pose trajectory at short horizons and continues to produce structurally coherent skeletons through to +2000 ms. The predicted poses preserve key kinematic constraints such as joint connectivity and limb proportions, with no catastrophic drift or physically implausible configurations at any horizon.

E Ablations Study

E.1 Architecture Ablations.

Table 4 reports ablation results on MM-Fi P1-S3. Removing the residual operator increases MPJPE by 3-6 mm across all horizons, confirming that the residual parametrization is necessary to escape the identity attractor and learn meaningful dynamics. Removing anchor-delta prediction produces a characteristic flat error profile (64.6-67.7 mm across horizons), consistent with the model collapsing to a near-constant pose output rather than modeling motion,

exactly the degenerate solution the anchor formulation is designed to prevent. Removing dual-stream fusion keeps short-horizon performance competitive (52.7 mm at 100 ms) but degrades rapidly at longer horizons (92.8 mm at 1000 ms), indicating that CSI features provide critical dynamic information beyond what skeletal coordinates alone can supply. Removing HPE pretraining causes severe degradation (∼106 mm throughout), confirming that the two-stage schedule is essential: without a warm initialization of the pose estimator, the skeleton encoder receives uninformative inputs from the outset, preventing the Koopman operator from forming a useful latent space.

Table 4: Ablation study results of KOALA across prediction horizons. Evaluated on MPJPE (mm), lower values indicate better performance.

Variant	**100ms**	**300ms**	**500ms**	**1000ms**	**1500ms**	**2000ms**
w/o Residual Operator	55.8	62.4	67.8	72.1	69.7	67.4
w/o Delta Prediction	67.7	64.8	64.6	65.2	65.7	66.5
w/o Dual Stream Fusion	52.7	67.2	80.2	92.8	90.1	78.8
w/o HPE Pretraining	105.4	106.1	106.7	107.7	107.3	106.7
KOALA (Ours)	**52.1**	**57.0**	**61.9**	**65.7**	**63.4**	**61.9**

E.2 Oracle Ablations.

Table 5 and Table 6 quantify the performance ceiling available to KOALA if the HPE module were replaced by ground-truth poses (oracle). On MM-Fi, the gap narrows substantially with horizon, from 26.8 mm at 100 ms to 7.1 mm at 1500 ms, indicating that the Koopman operator compensates for HPE noise more effectively as the prediction horizon grows and latent dynamics carry more of the burden relative to the initial pose estimate. On WiPose, the gap shrinks to just 5.2 mm MPJPE and 1.7 mm PA-MPJPE at 1000 ms, suggesting the HPE module achieves sufficient accuracy on this dataset that cascaded noise is not a significant bottleneck at long horizons. Across both datasets, the average Δ of 12.1 mm (MM-Fi) and 11.2 mm (WiPose) represents the remaining headroom from end-to-end pose estimation quality, motivating future work on tighter integration of pose estimation and motion prediction.

Table 5: Oracle upper bound comparison on WiPose dataset.

Variant	**Metric**	**100ms**	**300ms**	**500ms**	**1000ms**	**Avg**
Oracle	MPJPE ↓	8.9	13.2	16.4	22.1	15.2
	PA-MPJPE ↓	10.3	13.8	16.1	19.6	15.0
KOALA	MPJPE ↓	26.1	26.0	26.5	27.3	26.5
	PA-MPJPE ↓	20.8	20.8	21.0	21.3	21.0
Δ (Oracle − Ours)	MPJPE ↓	17.2	12.8	9.6	5.2	11.2
	PA-MPJPE ↓	10.5	7.0	5.1	1.7	6.1

E.3 Hyperparameter Settings

Table 7 outlines sensitivity to key hyperparameters on Protocol P1, Setting S3. For model dimension, reducing to $d = 64$ increases MPJPE by 2-2.5 mm while $d = 256$ yields marginal

Table 6: Oracle upper bound comparison on MM-Fi dataset.

Variant	**Metric**	**100ms**	**300ms**	**500ms**	**1000ms**	**1500ms**	**2000ms**	**Avg**
Oracle	MPJPE ↓	25.3	42.9	52.5	58.4	56.3	53.9	48.2
	PA-MPJPE ↓	22.0	37.0	44.5	47.1	46.2	45.1	40.3
KOALA	MPJPE ↓	52.1	57.0	61.9	65.7	63.4	61.9	60.3
	PA-MPJPE ↓	43.5	47.0	50.5	52.8	51.8	51.1	49.5
Δ (Oracle − Ours)	MPJPE ↓	26.8	14.1	9.4	7.3	7.1	8.0	12.1
	PA-MPJPE ↓	21.5	10.0	6.0	5.7	5.6	6.0	9.2

Table 7: Ablation study of KOALA under different hyperparameter settings. MPJPE / PA-MPJPE (mm), lower is better.

Variant	**100ms**	**300ms**	**500ms**	**1000ms**	**1500ms**	**2000ms**
(A) Model dimension d						
$d = 64$	54.16 / 45.75	58.74 / 48.94	63.32 / 52.19	67.55 / 54.72	65.28 / 53.82	63.59 / 52.94
$d = 256$	52.95 / 44.12	56.54 / 46.48	60.49 / 49.30	64.24 / 51.82	62.41 / 50.84	61.31 / 50.40
(B) Latent dimension D_z						
$D_z = 64$	53.75 / 45.31	61.14 / 51.39	67.85 / 56.91	74.60 / 62.82	72.32 / 62.44	68.60 / 59.23
$D_z = 128$	54.72 / 45.94	59.86 / 49.56	64.83 / 52.93	69.13 / 55.67	66.83 / 54.80	64.89 / 53.78
(C) Koopman rank r						
$r = 8$	55.53 / 46.36	59.23 / 48.90	63.66 / 52.10	67.83 / 55.08	65.75 / 53.96	64.44 / 53.46
$r = 32$	53.23 / 44.87	60.79 / 50.93	67.59 / 56.40	74.08 / 61.29	71.23 / 60.27	68.05 / 58.07
(D) Encoder depth L_c						
$L_c = 2$	52.76 / 44.01	58.00 / 47.89	62.58 / 51.15	66.58 / 53.68	64.11 / 52.39	62.44 / 51.51
$L_c = 6$	54.44 / 45.89	61.21 / 51.46	66.77 / 55.79	71.07 / 58.66	68.68 / 57.90	67.04 / 56.86
KOALA (Ours)	**52.10 / 43.50**	**57.00 / 47.00**	**61.90 / 50.50**	**65.70 / 52.80**	**63.40 / 51.80**	**61.90 / 51.10**

gains below 1 mm at roughly double the parameter count, confirming $d = 128$ as the optimal trade-off. Latent dimension has the strongest effect: $D_z = 64$ reaches 74.6 mm at 1000 ms versus 65.7 mm for the default $D_z = 256$, confirming that a sufficiently large Koopman space is necessary to disentangle nonlinear motion modes, with $D_z = 128$ showing intermediate degradation. For Koopman rank, both $r = 8$ and $r = 32$ underperform the default $r = 16$: the former lacks capacity for CSI modulation while the latter over-parameterizes the adaptation and destabilizes long-horizon rollouts (74.1 mm at 1000 ms), identifying $r = 16$ as a sweet spot. Encoder depth is the least sensitive axis: $L_c = 2$ degrades slightly at long horizons while $L_c = 6$ increases error across all horizons, likely from overfitting. Overall, KOALA is most sensitive to latent dimension and rank, and robust to moderate variation in model dimension and encoder depth.