REFINE TOGETHER: MUTUAL POSE-KINEMATIC FEEDBACK FOR 3D HUMAN POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mamba has shown promise for modeling complex spatiotemporal relations in 3D human pose estimation (3D-HPE). However, ambiguous 2D observations can correspond to different 3D configurations, while structural and motion relations remain implicit in latent pose representations. Explicit 3D kinematic relations provide direct target-space cues, yet their reliability depends on the accuracy of the estimated pose. To address these limitations, we propose CoKinPose, a target-space 3D pose modeling framework that jointly refines pose and kinematic representations within Mamba. First, a kinematics-driven state update mechanism derives skeletal and motion relations from an initial 3D pose estimate and incorporates them into feature processing and selective state evolution. A confidence-guided gate adaptively regulates their contributions using 2D observation confidence and the current kinematic representations. Second, an iterative pose-kinematic mutual refinement mechanism applies this state update across successive Mamba blocks and feeds updated pose representations back to refine the kinematic representations used by subsequent blocks. Experiments on Human3.6M and MPI-INF-3DHP demonstrate consistent improvements under standard and challenging conditions, with further analyses validating enhanced structural and motion modeling under spatial skeletal and temporal motion ambiguities. Our code and models will be made publicly available upon publication.

1 INTRODUCTION

Monocular 3D Human Pose Estimation (3D-HPE), as a core task in human motion analysis, has shown broad application prospects in human-centered fields such as human-computer interaction, embodied intelligence, motion analysis, and virtual reality (Guo et al., 2025; Azam & Desai, 2024). Within the prevalent "2D-to-3D lifting" paradigm of 3D-HPE, the primary challenge lies in efficiently modeling complex topological dependencies and dynamic variations (Zhao et al., 2025; Pramono et al., 2025; Zhang et al., 2025a). Such modeling is further challenged by depth ambiguity, perspective effects, and occlusion (Rommel et al., 2024; Hao & Li, 2025; Wang et al., 2024).

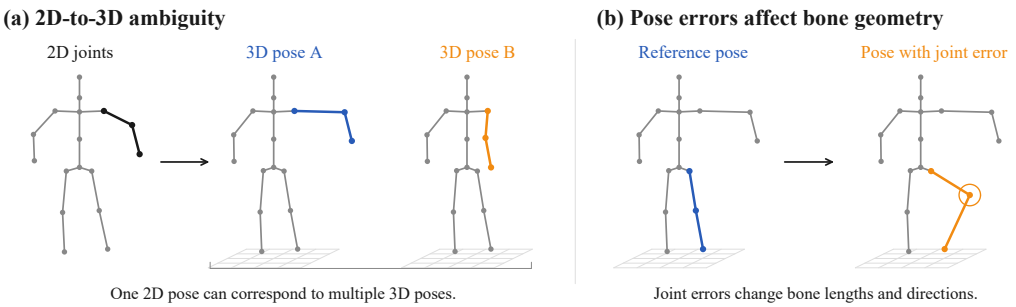


Fig. 1: (a) The same 2D joints can correspond to different 3D arm configurations under orthographic projection. (b) A knee-position error changes the lengths and directions of the connected leg bones.

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

Existing studies on 3D-HPE can be discussed from two related perspectives. One line focuses on 2D to 3D pose representation learning, modeling spatial topology and temporal dynamics to enrich latent pose representations, as exemplified by Mamba-based methods such as PoseMamba (Huang et al., 2025) and SAMA (Lu et al., 2025). Another line focuses on explicit target-space 3D relations, with representative methods exploiting skeletal geometry such as bone lengths and directions. For example, MDTF (Zheng et al., 2026) explicitly estimates 3D bone vectors and fuses bone and joint representations, while DDHPose (Cai et al., 2024) decomposes 3D poses into bone-length and bone-direction representations for pose estimation. These perspectives emphasize latent pose representations and explicit target-space 3D relations, respectively.

However, both perspectives face challenges in 2D-to-3D pose estimation under ambiguous observations. On the one hand, pose representations are learned from 2D joints and their spatiotemporal contexts, allowing rich structural and motion relations to be captured. Nevertheless, similar 2D observations may correspond to different 3D configurations, and the structure and motion information encoded in latent features is not necessarily available as explicit 3D relations to characterize ambiguous observations. For example, Fig. 1(a) shows that the same 2D joints can correspond to different 3D arm configurations. On the other hand, when explicit 3D kinematic relations are derived from a current pose, their reliability depends on the accuracy of the estimated joint positions. Errors in the current pose can consequently propagate into skeletal and motion relations, introducing inaccurate cues for subsequent pose modeling. As illustrated in Fig. 1(b), an error in the estimated knee position changes the lengths and directions of the connected leg bones, showing how pose errors can lead to inaccurate 3D relation descriptors. Our key insight is that pose representations and 3D kinematic relation representations encode the same pose information in different forms, motivating their joint use through mutual refinement. Explicit 3D kinematic relations can provide structural and motion references for refining pose representations, while updated pose representations can provide contextual feedback for refining kinematic relation representations. Therefore, we explore a critical question: ***Can pose representations and 3D kinematic relation representations be mutually refined within Mamba to improve 3D pose estimation under ambiguous 2D observations?***

To answer this question, we propose CoKinPose, which exploits the complementarity between pose representations and 3D kinematic relation representations by grounding their mutual refinement in explicit target-space 3D kinematics. First, we develop a kinematics-driven state update mechanism to incorporate explicit 3D structural and motion information into Mamba's selective state evolution. Specifically, skeletal and motion relations are derived from an initial 3D pose and encoded as kinematic representations. These representations are fused with pose features through a confidence-guided adaptive gate, which regulates their contributions using 2D observation confidence and the current kinematic representations. Second, we introduce an iterative pose-kinematic mutual refinement mechanism that embeds this state update within successive Mamba blocks and enables mutual feedback between the two representations. Kinematic representations guide the update of pose representations with explicit structural and motion information, while the updated pose representations provide contextual feedback to refine the kinematic representations. In this way, CoKinPose enables the two complementary representations of the underlying 3D human state to inform each other throughout state-space modeling, providing explicit structural and motion references for HPE under ambiguous 2D observations. Our contributions are summarized as follows:

- We introduce CoKinPose, a target-space 3D pose modeling framework that integrates latent pose modeling with explicit 3D kinematics to progressively refine pose and kinematic representations for 3D pose estimation. By conditioning Mamba's internal state evolution on relations defined in the target 3D pose space, CoKinPose enables skeletal structure and motion relations to serve as task-aligned guidance for 3D-HPE.
- We develop a kinematics-driven state update mechanism together with an iterative pose-kinematic mutual refinement mechanism. The former incorporates explicit 3D skeletal and motion relations into Mamba's feature processing and selective state evolution, adaptively regulating their contributions using 2D observation confidence and the current kinematic representations. The latter enables kinematic representations to guide pose feature updates, while updated pose features provide contextual feedback to refine the kinematic representations across successive Mamba blocks.
- We evaluate CoKinPose on Human3.6M and MPI-INF-3DHP under standard and challenging conditions. Ablations and targeted experiments assess component effectiveness and the framework's ability to alleviate spatial skeletal and temporal motion ambiguities.

2 RELATED WORK

2.1 POSE REPRESENTATION LEARNING FOR 3D-HPE

Recent 2D-to-3D lifting methods have extensively explored spatiotemporal pose representation learning. MixSTE (Zhang et al., 2022) alternates spatial modeling across joints with temporal modeling along individual joint trajectories. KTPFormer (Peng et al., 2024) incorporates skeletal-connectivity and cross-frame trajectory priors into attention features through fixed local and learnable global topologies. Mamba-based methods further introduce state-space modeling into this process. PoseMamba (Huang et al., 2025) combines bidirectional spatiotemporal modeling with global–local spatial scans, while Pose Magic (Zhang et al., 2025b) adaptively fuses Mamba-based global representations with GCN-based local representations. SAMA (Lu et al., 2025) integrates learnable pose topology into feature and hidden-state aggregation and adapts timescales to joint-specific motion features. SasMamba (Cui et al., 2026) combines structure-aware spatiotemporal convolution with stride-based scanning to capture local interactions and multiscale global dependencies.

These methods primarily improve 3D pose estimation by enriching latent pose representations with spatial topology and temporal dynamics. Such representations capture rich relations among joints and frames, but do not necessarily expose this information as explicit, sample-specific 3D skeletal and motion descriptors for pose refinement. CoKinPose complements this line of research by deriving skeletal and motion relations in the target 3D pose space and using their encoded representations to condition Mamba's feature processing and selective state evolution, providing explicit structural and motion references under ambiguous 2D observations.

2.2 EXPLICIT 3D KINEMATIC RELATION MODELING

Another related perspective explicitly models the target 3D pose through skeletal quantities such as bone lengths, directions, and vectors. The anatomy-aware framework of Chen et al. (2022) decomposes 3D pose estimation into bone-length and bone-direction prediction and additionally incorporates 2D keypoint visibility cues. DDHPose (Cai et al., 2024) introduces diffusion over bone-length and bone-direction representations and recovers joint coordinates with a hierarchical spatiotemporal denoiser. BLAPose (Hsu & Jang, 2024) predicts bone lengths from video-level 2D joint sequences and adjusts existing 3D pose estimates while retaining their predicted bone directions. More recently, MDTF (Zheng et al., 2026) estimates 3D bone vectors in spherical coordinates and adaptively fuses bone and joint representations before spatiotemporal refinement within a Mamba-based framework.

These studies demonstrate the value of introducing explicit 3D skeletal relations into pose estimation through pose decomposition, diffusion modeling, geometric adjustment, or bone-joint fusion. Although they exploit temporal context, their explicit relation representations primarily describe skeletal geometry rather than quantities such as inter-frame 3D velocities. CoKinPose models both 3D skeletal and inter-frame motion relations derived from an initial 3D pose estimate and incorporates their encoded representations into Mamba's selective state evolution. Since pose errors can affect these derived relations, confidence-guided gating adaptively regulates their contributions. Updated pose representations then provide contextual feedback to refine the kinematic representations across successive blocks, enabling mutual refinement grounded in explicit target-space 3D kinematics.

3 METHOD

We propose CoKinPose to jointly refine pose representations and explicit 3D kinematic representations for 3D-HPE under ambiguous 2D observations. The framework consists of two mechanisms. The kinematics-driven state update incorporates explicit skeletal and motion relations into Mamba through reliability-aware feature fusion and state modulation. The iterative pose-kinematic mutual refinement mechanism further enables kinematic representations to guide pose updates, while updated pose representations provide contextual feedback for refining subsequent kinematic representations. Through this interaction, pose and kinematic representations are progressively refined within Mamba, guided by relations defined in the target 3D pose space. Fig. 2 summarizes the framework.

3.1 KINEMATICS-DRIVEN STATE UPDATE

Given a 2D pose sequence $\mathbf{P}_{2\mathrm{D}} \in \mathbb{R}^{T\times N\times 2}$ and the associated joint detection confidence scores $\mathbf{S} \in [0,1]^{T\times N}$, we use an ST-GCN to extract pose features $\mathbf{X}_0 \in \mathbb{R}^{T\times N\times C}$ and decode them into an initial 3D pose $\mathbf{P}_{1\mathrm{3D}} \in \mathbb{R}^{T\times N\times 3}$. Here, T, N, and C denote the number of frames, the number of joints, and the feature dimension, respectively. All 3D poses are expressed relative to the pelvis

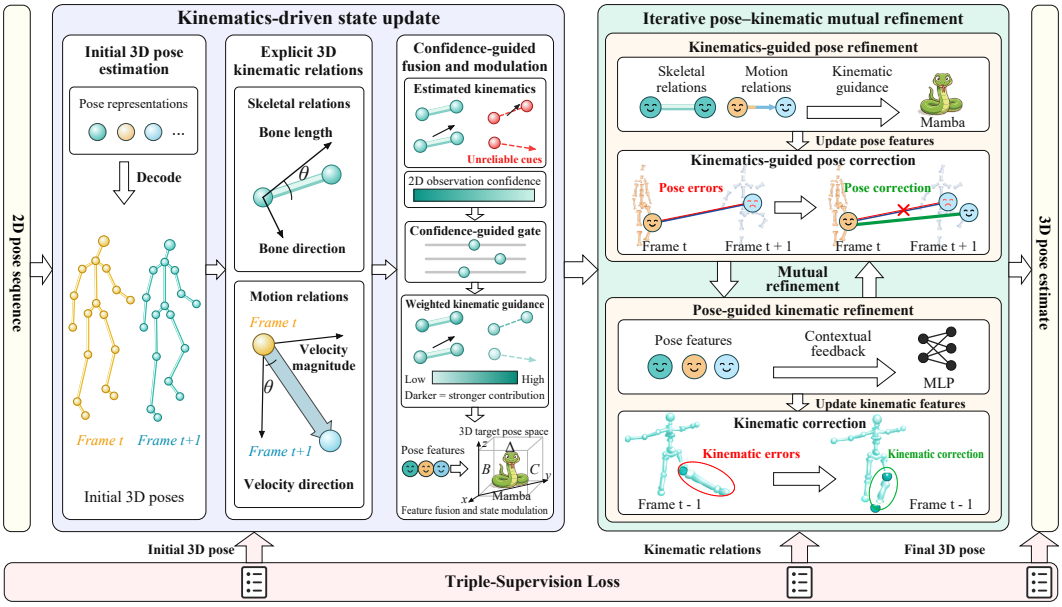


Fig. 2: CoKinPose framework. Darker colors in the weighted guidance indicate greater kinematic contributions. The coordinate frame denotes the 3D target pose space of the guiding relations.

in each frame. The initial pose receives 3D coordinate supervision during training, providing a supervised basis for extracting target-space kinematic relations.

Although pose representations can encode rich spatial and temporal contexts, these relations are not explicitly represented as 3D geometric and motion descriptors. Therefore, we derive skeletal and motion relations from the initial 3D pose estimate to provide explicit target-space kinematic references for subsequent Mamba refinement.

To extract 3D skeletal relations, we use a fixed skeleton tree rooted at the pelvis. Let r denote the root and $\pi(j)$ the parent of a non-root joint j. The parent-to-child bone vector is $\mathbf{b}_{t,j} = \mathbf{P}_{13D}[t, j] - \mathbf{P}_{13D}[t, \pi(j)]$. Collecting all $N-1$ bones over the sequence gives $\mathbf{b} \in \mathbb{R}^{T\times(N-1)\times 3}$. We compute bone lengths $\mathbf{BL} \in \mathbb{R}^{T\times(N-1)\times 1}$ and bone directions $\mathbf{BO} \in \mathbb{R}^{T\times(N-1)\times 3}$ as:

$$\mathbf{BL} = \|\mathbf{b}\|_2, \qquad \mathbf{BO} = \frac{\mathbf{b}}{\mathbf{BL} + \epsilon}, \tag{1}$$

where the norm is taken along the coordinate dimension and $\epsilon > 0$ ensures numerical stability. These descriptors characterize skeletal scale and relative 3D orientation.

For 3D motion relations, we take one inter-frame interval as the time unit and compute $\mathbf{v}_{t,j} = \mathbf{P}_{13D}[t, j] - \mathbf{P}_{13D}[t-1, j]$ for $t = 2, \ldots, T$. This gives the velocity tensor $\mathbf{v} \in \mathbb{R}^{(T-1)\times N\times 3}$. Velocity magnitudes $\mathbf{VM} \in \mathbb{R}^{(T-1)\times N\times 1}$ and directions $\mathbf{VD} \in \mathbb{R}^{(T-1)\times N\times 3}$ are obtained as:

$$\mathbf{VM} = \|\mathbf{v}\|_2, \qquad \mathbf{VD} = \frac{\mathbf{v}}{\mathbf{VM} + \epsilon}, \tag{2}$$

where the velocity describes inter-frame joint displacement without global body translation. Together, the four descriptors explicitly represent spatial skeletal dependencies and temporal motion dependencies in the target 3D pose space.

We align the descriptors with the $T \times N$ joint-frame layout of the pose features. Each bone descriptor is assigned to its corresponding child joint, with zeros at the root joint. Each motion descriptor is assigned to the destination frame of the corresponding frame pair, with zeros in the first frame. This alignment preserves the correspondence between kinematic relations and pose tokens. The aligned length and magnitude descriptors each contain one channel, while the direction descriptors contain three channels. We concatenate the resulting eight channels and apply a nonlinear mapping to obtain the initial kinematic representation $\mathbf{Kin}_0 \in \mathbb{R}^{T\times N\times C/2}$.

Although explicit 3D kinematic representations provide structural and motion references, they are derived from estimated poses and may contain unreliable information caused by inaccurate joint predictions. Therefore, we introduce a confidence-guided gate to adaptively regulate their contribution during pose modeling. Specifically, skeletal confidence is computed as the geometric mean of the confidence scores of the two joints connected by each bone, while motion confidence is computed as the geometric mean of the scores for the same joint in adjacent frames. After applying the same alignment and zero-padding rules as the kinematic descriptors, these confidence cues are concatenated along the channel dimension to form $\mathbf{Q} \in [0,1]^{T\times N\times 2}$. The two channels of $\mathbf{Q}$ provide observation-based reliability cues for skeletal and motion relations, respectively. In block l, $\mathbf{Q}$ and the current kinematic representation jointly generate a gate $\mathbf{G}_l \in \mathbb{R}^{T\times N\times 1}$ for feature fusion:

$$\begin{aligned} \mathbf{G}_l &= \overline{\mathbf{Q}} \odot \sigma\big(\mathbf{W}_g^l \,\mathrm{Cat}(\mathbf{Kin}_{l-1}, \mathbf{Q}) + \mathbf{b}_g^l\big)\,, \\ \widetilde{\mathbf{X}}_{l-1} &= \mathrm{Cat}\Big(\mathbf{G}_l \odot \mathbf{Kin}_{l-1}, \mathrm{Linear}_{\mathrm{x}}^l(\mathbf{X}_{l-1})\Big)\,. \end{aligned} \tag{3}$$

Here, $\overline{\mathbf{Q}}$ is the channel-wise mean of $\mathbf{Q}$, and $\mathbf{W}_g^l$ and $\mathbf{b}_g^l$ are learnable parameters. $\mathrm{Linear}_{\mathrm{x}}^l$ projects pose features to $C/2$ channels, yielding $\widetilde{\mathbf{X}}_{l-1} \in \mathbb{R}^{T\times N\times C}$. Cat, $\odot$, and σ denote channel concatenation, elementwise multiplication, and the sigmoid function, respectively. This design adapts the kinematic contribution using observation confidence and the current relation content.

Beyond feature-level fusion, we further allow kinematic representations to directly influence Mamba's internal state evolution. Specifically, the kinematic representation is used to modulate Mamba's input-dependent parameters, including the input mapping $\mathbf{B}_l$, output mapping $\mathbf{C}_l$, and discretization step size $\mathbf{\Delta}_l$. These parameters control information writing, state reading, and state evolution, respectively. The base parameters are generated from fused pose features, while nonlinear mappings of the kinematic representation produce modulation signals:

$$\widehat{\mathbf{\Theta}}_l = \mathbf{\Theta}_l \odot \big[1 + \mathbf{G}_l \odot \sigma\big(\Phi_{\Theta}^l(\mathbf{Kin}_{l-1})\big)\big]\,, \qquad \mathbf{\Theta} \in \{\mathbf{B}, \mathbf{C}, \mathbf{\Delta}\}. \tag{4}$$

Each Φ_{Θ}^l maps the kinematic representation to the shape of its corresponding parameter, with the gate broadcast over the required dimensions. Through this design, target-space kinematic relations condition both feature processing and selective state evolution, enabling Mamba to adapt its updates according to sample-specific 3D structural and motion patterns.

3.2 Iterative pose–kinematic mutual refinement

The kinematic representation derived from the initial 3D pose provides explicit target-space structural and motion references, but it may still contain errors inherited from the initial estimation. Meanwhile, the pose representation continuously captures contextual information through Mamba refinement, which can provide complementary feedback for improving the kinematic representation. Therefore, we introduce an iterative pose–kinematic mutual refinement mechanism across L successive Mamba blocks, where kinematic representations guide pose updates and updated pose representations refine the kinematic representations for subsequent modeling.

Given the pose representation $\mathbf{X}_{l-1}$ and the kinematic representation $\mathbf{Kin}_{l-1}$, block l first constructs the reliability-aware fused features using Eq. (3) and modulates the input-dependent parameters using Eq. (4). The pose representation is then updated through a residual Mamba block:

$$\mathbf{X}_l = \mathbf{X}_{l-1} + \mathrm{Mamba}_l\Big(\widetilde{\mathbf{X}}_{l-1}\Big)\,, \tag{5}$$

where Mamba_l uses the modulated parameters $\widehat{\mathbf{B}}_l$, $\widehat{\mathbf{C}}_l$, and $\widehat{\mathbf{\Delta}}_l$. By incorporating kinematics-guided feature fusion and state modulation, this update allows the pose representation to be refined under explicit 3D skeletal and motion references rather than relying solely on implicit relations learned from 2D observations. We then feed the updated pose representation back to the kinematic branch through a residual refinement process:

$$\mathbf{Kin}_l = \mathbf{Kin}_{l-1} + \mathrm{MLP}_l\Big(\mathbf{Kin}_{l-1} + \mathrm{Linear}_{\mathrm{fb}}^l(\mathbf{X}_l)\Big)\,. \tag{6}$$

Here, $\mathrm{Linear}_{\mathrm{fb}}^l$ projects pose features into the $C/2$-dimensional kinematic feature space, and the MLP produces a residual update by combining the existing kinematic representation with the contextual

information from the refined pose representation. This update operates in representation space, combining encoded kinematic information with pose-dependent contextual feedback.

The refined kinematic representations are passed to subsequent Mamba blocks for gate generation, feature fusion, and state modulation. The observation reliability cues $\mathbf{Q}$ remain unchanged, while the kinematic contribution is dynamically adjusted according to the updated representation. Through this repeated interaction, kinematic representations continuously provide target-space structural and motion guidance for pose refinement, while pose representations progressively refine the kinematic encoding used in later stages. This establishes a closed-loop refinement process in which the two complementary representations inform each other throughout Mamba state-space modeling.

After the final block, we decode $\mathbf{X}_L$ into the final 3D pose $\mathbf{P}_{23\mathrm{D}} \in \mathbb{R}^{T\times N\times 3}$ and $\mathbf{Kin}_L$ into relation predictions $\mathbf{Kin}_{\mathrm{dec}} \in \mathbb{R}^{T\times N\times 8}$. The decoded relations contain bone lengths, bone directions, velocity magnitudes, and velocity directions in the same aligned layout as the initial kinematic descriptors. Let $\mathbf{P}^* \in \mathbb{R}^{T\times N\times 3}$ denote the ground-truth 3D pose sequence and $\mathbf{D}^* \in \mathbb{R}^{T\times N\times 8}$ its kinematic descriptors constructed using the same extraction and alignment rules. We jointly supervise the initial pose, final pose, and decoded kinematic relations:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{P_2}(\mathbf{P}_{23\mathrm{D}}, \mathbf{P}^*) + \lambda_1 \mathcal{L}_{P_1}(\mathbf{P}_{13\mathrm{D}}, \mathbf{P}^*) + \lambda_2 \mathcal{L}_{\mathrm{kin}}(\mathbf{Kin}_{\mathrm{dec}}, \mathbf{D}^*). \tag{7}$$

Here, $\mathcal{L}_{P_1}$ and $\mathcal{L}_{\mathrm{kin}}$ supervise the initial pose and decoded kinematic relations, respectively, while $\mathcal{L}_{P_2}$ follows PoseMamba (Huang et al., 2025). We set $\lambda_1 = \lambda_2 = 1$ and exclude padded entries from kinematic relation supervision. Joint supervision encourages the model to preserve target-space skeletal and motion information while mutually refining pose and kinematic representations.

4 Experiments

4.1 Datasets with Evaluation Metrics and Experimental Settings

We comprehensively evaluate CoKinPose on two standard benchmarks: Human3.6M and the more challenging MPI-INF-3DHP dataset. *Detailed descriptions of the datasets, evaluation metrics, and specific training configurations are provided* **in Appendix A**.

4.2 Comparison with State-of-the-Art Methods

We evaluate CoKinPose from five complementary perspectives. Standard benchmark comparisons establish its accuracy and computational cost, followed by an examination of low-confidence observations and intense motion. Evaluation on MPI-INF-3DHP extends the comparison to more varied environments and poses. Temporal and action-wise analyses then examine whether the benefits extend across input lengths and movement types. Together, these comparisons assess the practical value of jointly refining pose and explicit 3D kinematic representations.

4.2.1 Evaluation on the Standard Benchmark

Tab. 1: Quantitative comparisons on Human3.6M. T is the number of input frames. CE denotes estimating center frame only. MACs/frame is the integer-truncated cost per input frame in millions. P1: MPJPE. P2: P-MPJPE. P1†: P1 on 2D ground truth.

Model size	Method	T	CE	Params(M)	MACs(G)	MACs/frame(M)	P1↓	P2↓	P1†↓
Small variants	P-STMO (Shan et al., 2022)	243	✓	6.2	0.7	2	42.8	34.4	29.3
	GLA-GCN (Yu et al., 2023)	243	✓	1.3	1.5	6	44.4	34.8	21.0
	HDFormer (Chen et al., 2023)	96	×	3.7	0.6	6	42.6	33.1	21.6
	PoseMamba-S (Huang et al., 2025)	243	×	0.9	3.6	14	41.8	35.0	20.0
	SAMA-S (Lu et al., 2025)	243	×	1.1	3.9	16	40.6	34.0	20.2
	CoKinPose-S	**243**	×	**1.0**	**3.8**	**15**	**39.2**	**33.1**	**19.6**
Base variants	STCFormer (Tang et al., 2023)	243	×	4.7	19.6	80	41.0	32.0	21.3
	PoseFormerV2 (Zhao et al., 2023)	243	✓	14.4	4.8	19	45.2	35.6	–
	PoseMamba-B (Huang et al., 2025)	243	×	3.4	13.9	57	40.8	34.3	16.8
	SasMamba-L (Cui et al., 2026)	243	×	4.1	8.6	35	39.8	33.6	21.0
	SAMA-B (Lu et al., 2025)	243	×	3.3	11.7	48	37.7	32.0	13.6
	CoKinPose-B	**243**	×	**3.7**	**12.2**	**50**	**37.9**	**32.1**	**13.6**
Large variants	MixSTE (Zhang et al., 2022)	243	×	33.6	139.0	572	40.9	32.6	21.6
	STCFormer-L (Tang et al., 2023)	243	×	18.9	78.2	321	40.5	31.8	–
	MotionBERT (Zhu et al., 2023)	243	×	42.3	174.8	719	39.2	32.9	17.8
	MAGFormer-L (Mehraban et al., 2024)	243	×	19.0	78.3	322	38.4	32.5	17.3
	KTPFormer (Peng et al., 2024)	243	×	35.2	76.1	313	40.1	31.9	19.0
	TCPFormer (Liu et al., 2025)	243	✓	35.1	109.2	449	37.9	31.7	–
	PoseMagic (Zhang et al., 2025b)	243	×	14.4	20.3	83	37.5	–	–
	PoseMamba-X (Huang et al., 2025)	243	×	26.5	109.9	452	37.1	31.5	14.8
	SAMA-L (Lu et al., 2025)	243	×	17.3	53.2	218	36.9	31.3	11.9
	CoKinPose-L	**243**	×	**18.0**	**54.4**	**223**	**36.7**	**31.1**	**11.9**

Tab. 1 summarizes the Human3.6M results across model capacities. CoKinPose-S obtains 39.2 mm P1 and 33.1 mm P2 with 1.0M parameters, improving on PoseMamba-S by 2.6 and 1.9 mm, respectively. CoKinPose-B reaches 37.9 and 32.1 mm, reducing the corresponding PoseMamba-B errors by 2.9 and 2.2 mm. CoKinPose-L achieves 36.7 mm P1 and 31.1 mm P2 with 18.0M parameters and 54.4G MACs. With ground-truth 2D inputs, the three variants obtain 19.6, 13.6, and 11.9 mm P1. These results show benefits across model sizes and input qualities. Since overall accuracy averages observations of different difficulty, we next examine where the gains are concentrated.

4.2.2 Performance under Low-confidence Observations and Intense Motion

We examine observation quality and motion intensity separately using unsupervised K-means clustering. Detection-confidence features define Easy, Medium, and Hard groups, while motion descriptors define Low and High groups. All methods share the resulting partitions, with overall errors weighted by group proportions. **Appendix B** provides the protocol, with the clusters visualized in Fig. 6. Low confidence indicates less reliable 2D evidence rather than a verified occlusion label. In Tab. 2, CoKinPose-L has lower errors than SAMA-L on Hard observations and High motion by 4.1 and 3.2 mm, respectively. These margins exceed the 0.2 mm overall difference, while SAMA-L has lower errors on Easy observations and Low motion. This pattern is consistent with the intended role of kinematic guidance under unreliable observations and intense motion.

Tab. 2: Human3.6M MPJPE (mm) by observation confidence and motion intensity.

Method	Overall	Easy	Med.	Hard	Low	High
PoseMamba-X (Huang et al., 2025)	37.1	32.6	43.0	56.8	34.5	44.9
PoseMagic (Zhang et al., 2025b)	37.5	32.8	44.0	57.4	34.8	45.6
DiffPose (Gong et al., 2023)	37.0	32.6	42.5	56.8	34.4	44.8
SAMA-L (Lu et al., 2025)	36.9	32.1	43.4	57.5	33.9	45.9
CoKinPose-L	**36.7**	**32.8**	**42.0**	**53.4**	**34.7**	**42.7**

4.2.3 Evaluation on MPI-INF-3DHP

MPI-INF-3DHP provides a complementary evaluation across varied poses and indoor and outdoor test environments. With an 81-frame input, CoKinPose obtains 99.1 PCK, 88.5 AUC, and 14.2 mm MPJPE (Tab. 3). Its position error is 0.5 mm below PoseMagic and 0.3 mm below PoseMamba. The table reports input lengths. The results indicate that joint pose–kinematic refinement also benefits MPI-INF-3DHP. These metrics aggregate the test environments without isolating outdoor performance or the ambiguity types underlying the gains.

Tab. 3: MPI-INF-3DHP comparisons.

Method	T	PCK ↑	AUC ↑	MPJPE ↓
MixSTE (Zhang et al., 2022)	27	94.4	66.5	54.9
P-STMO (Shan et al., 2022)	81	97.9	75.8	32.2
GLA-GCN (Yu et al., 2023)	81	98.5	79.1	27.8
STCFormer (Tang et al., 2023)	81	98.7	83.9	23.1
DiffPose (Gong et al., 2023)	81	98.0	75.9	29.1
MAGFormer (Mehraban et al., 2024)	81	98.2	85.3	16.2
KTPFormer (Peng et al., 2024)	81	98.9	85.9	16.7
TCPFormer (Liu et al., 2025)	81	99.0	87.7	15.0
PoseMagic (Zhang et al., 2025b)	81	98.8	87.6	14.7
PoseMamba (Huang et al., 2025)	81	–	–	14.5
CoKinPose	**81**	**99.1**	**88.5**	**14.2**

4.2.4 Temporal Accuracy across Input Sequence Lengths

We evaluate temporal accuracy using MPJVE, which compares predicted and ground-truth root-relative joint displacements at the same sampling interval. Fig. 3 reports both position and velocity errors. At $T = 243$, CoKinPose-L obtains 1.52 mm/frame MPJVE, compared with 1.62 for PoseMamba-X. Extending the input to 351 frames reduces CoKinPose-L's MPJPE from 36.7 to 36.3 mm and MPJVE to 1.44 mm/frame. The variant without pose-to-kinematic feedback has higher errors at both lengths, supporting the contribution of repeated interaction between the representations. These comparisons measure motion accuracy rather than smoothness alone.

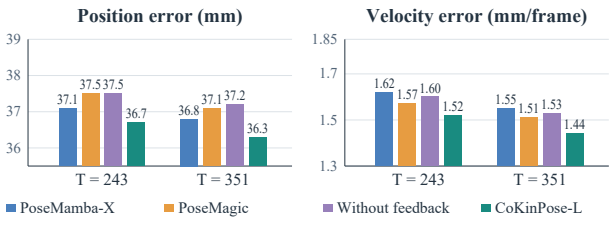


Fig. 3: Pose and velocity errors for 243- and 351-frame inputs on Human3.6M.

4.2.5 Action-wise Performance Analysis

We evaluate action-wise performance to examine whether joint pose-kinematic refinement benefits both complex body configurations and sustained motion. These action types place different demands on structural and temporal modeling, allowing us to assess the effectiveness of explicit skeletal and motion guidance across varied posture and movement patterns. CoKinPose-S improves over

PoseMamba-S on representative actions from both settings, supporting the usefulness of joint refinement under diverse structural and motion conditions. Detailed results are provided **in Appendix C**.

4.3 Ablation Studies

We conduct ablations with CoKinPose-S on Human3.6M to assess explicit skeletal and motion relations, kinematic state modulation, pose-to-kinematic feedback, and auxiliary supervision. Removing these components under a shared setup increases errors (Tab. 4). The gate comparisons **in Appendix D** favor combining confidence and current kinematics, especially at low confidence. Loss-weight sensitivity and additional component and relation analyses appear **in Appendices E and F**, respectively.

Tab. 4: CoKinPose-S ablations. Mod.: state modulation; FB: pose feedback. Static kinematics also disables kinematic self-update.

Variant	Relation inputs BL	BO	VM	VD	Updates Mod.	FB	Aux. losses $\mathcal{L}_{P_1}$	$\mathcal{L}_{\text{kin}}$	MPJPE	MPJVE
Pose-only backbone	×	×	×	×	×	×	✓	×	41.8	1.90
No bone length	×	✓	✓	✓	✓	✓	✓	✓	39.8	1.67
No bone direction	✓	×	✓	✓	✓	✓	✓	✓	40.2	1.70
No velocity magnitude	✓	✓	×	✓	✓	✓	✓	✓	40.0	1.78
No velocity direction	✓	✓	✓	×	✓	✓	✓	✓	39.9	1.77
No skeletal relations	×	×	✓	✓	✓	✓	✓	✓	40.5	1.69
No motion relations	✓	✓	×	×	✓	✓	✓	✓	40.3	1.84
No state modulation	✓	✓	✓	✓	×	✓	✓	✓	40.0	1.72
No pose feedback	✓	✓	✓	✓	✓	×	✓	✓	40.2	1.76
Static kinematics	✓	✓	✓	✓	✓	×	✓	✓	40.6	1.81
No adaptive gate	✓	✓	✓	✓	✓	✓	✓	✓	40.1	1.75
No initial-pose loss	✓	✓	✓	✓	✓	✓	×	✓	39.8	1.69
No relation loss	✓	✓	✓	✓	✓	✓	✓	×	40.7	1.79
No auxiliary losses	✓	✓	✓	✓	✓	✓	×	×	41.2	1.84
CoKinPose-S (Full)	✓	✓	✓	✓	✓	✓	✓	✓	39.2	1.65

4.4 Reliability-aware Kinematic State Responses

To examine how CoKinPose adapts to observation reliability and motion, we compare the normalized transition and input-mapping responses shown in Fig. 4. The bars report responses for individual joints ordered by confidence or motion intensity, allowing us to inspect how state responses vary across observation and motion conditions. We use joint confidence to represent observation reliability. In Fig. 4(a), joints are ordered from low to high confidence, and lower-confidence joints exhibit stronger transition responses. This association is consistent with increased reliance on contextual state when observations are less reliable. Motion intensity combines normalized velocity, acceleration, directional changes, and FFT amplitude to describe displacement and temporal variation. Fig. 4(b) orders joints from high to low motion intensity; stronger input-mapping responses occur at the high-intensity end. The sequence in Fig. 4(c) provides a temporal comparison of the same behavior. During higher-motion intervals (red), input-mapping responses are stronger; during lower-motion intervals (blue), they are weaker. These joint-wise and temporal comparisons show that the displayed state responses vary with observation reliability and motion, consistent with the role of kinematic guidance in selective state updates.

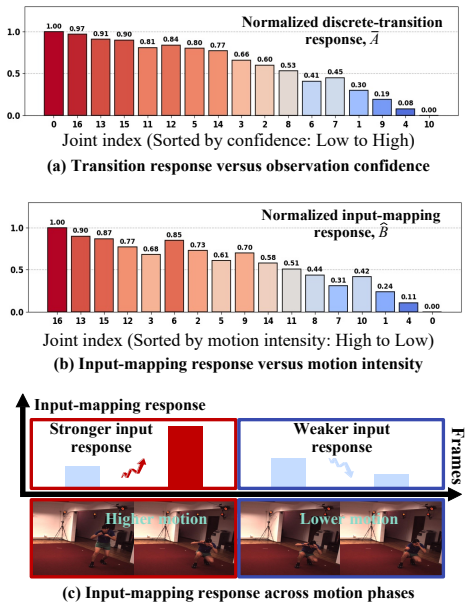


Fig. 4: Discrete-transition and input-mapping responses across observation confidence and motion conditions.

4.5 Kinematics-guided Joint Dependency Modeling

To reveal how CoKinPose models skeletal dependencies under explicit kinematic guidance, we probe Mamba's input encoding and state retrieval through its $\mathbf{B}$ and $\mathbf{C}$ mappings. We compute their interaction map as a proxy for joint dependency strengths in the pose-only backbone and CoKinPose, where brighter colors indicate stronger responses. As shown in Fig. 5(a), the backbone exhibits a strong association between joints 3 and 4, which have adjacent sequence

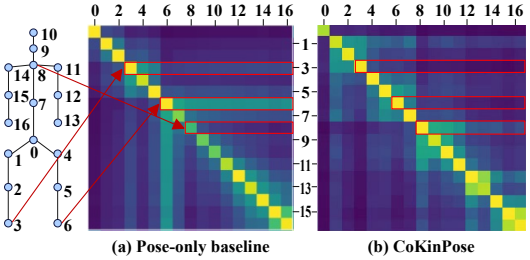


Fig. 5: Joint-dependency proxies from $\mathbf{B}$ and $\mathbf{C}$.

indices but no direct skeletal connection. In contrast, CoKinPose in Fig. 5(b) reduces this association. Furthermore, joint 8 serves as a structural hub connecting multiple body branches. However, Fig. 5(a) shows limited responses beyond its adjacent sequence position, joint 9. Fig. 5(b) shows broader dependencies between joint 8 and its branches. These patterns support skeletal dependency modeling through target-space kinematic guidance.

4.6 Refinement under Ambiguous Observations and Kinematic Errors

To test whether explicit skeletal and motion relations help when projection cues are limited, we compare CoKinPose-L and its ablations on fixed groups defined by ground-truth limb foreshortening and depth-dominant motion. All methods share the same 2D observations and 243-frame context, and group assignment is independent of prediction errors. Tab. 5 shows lower pose, direction, and velocity errors in the difficult groups, with larger degradations when the corresponding skeletal or motion relations are removed. These results support the role of target-space kinematic guidance under limited projection cues. Group definitions and evaluation details are provided **in Appendix G**.

Tab. 5: Projection-related evaluation. Depth is joint-depth MAE. P1/Depth: mm; BO: degrees; velocity: mm/frame.

(a) Spatial: other 70% / difficult 30%

		Other			Difficult		
Variant	Overall	P1	BO	Depth	P1	BO	Depth
PoseMamba-X	37.1	33.2	6.2	18.1	46.2	12.2	30.5
w/o skeletal relations	37.8	33.6	6.5	18.9	47.6	13.2	32.0
w/o motion relations	37.6	34.3	6.3	19.0	45.3	10.8	28.8
w/o pose feedback	37.5	33.9	6.3	18.8	45.9	11.9	30.1
w/o adaptive gating	37.3	33.7	6.2	18.7	45.7	11.6	29.7
CoKinPose-L	**36.7**	**33.4**	**6.0**	**18.3**	**44.4**	**10.2**	**27.2**

(b) Temporal: other 75% / depth-dominant 25%

		Other			Depth-dominant		
Variant	Overall	P1	MPJVE	Depth vel.	P1	MPJVE	Depth vel.
PoseMamba-X	37.1	33.8	1.40	0.70	47.0	2.28	1.57
w/o skeletal relations	37.8	34.9	1.40	0.71	46.5	2.16	1.46
w/o motion relations	37.6	34.4	1.44	0.73	47.2	2.48	1.72
w/o pose feedback	37.5	34.5	1.40	0.71	46.5	2.20	1.53
w/o adaptive gating	37.3	34.3	1.40	0.71	46.3	2.24	1.55
CoKinPose-L	**36.7**	**34.0**	**1.36**	**0.68**	**44.8**	**2.00**	**1.31**

To further assess robustness to inaccurate kinematic guidance, we perturb a copy of the initial 3D pose only for relation extraction in frozen CoKinPose-S variants. All variants share a perturbation field scaled more strongly for low-confidence joints, while the 2D observations, confidence cues, and initial pose features remain unchanged. Tab. 6 compares error increases relative to each variant's unperturbed baseline. The full model degrades less than the variants without adaptive gating or pose feedback and the static-kinematics control, indicating that contribution regulation and continued refinement improve robustness to confidence-correlated relation errors. The perturbation protocol is provided **in Appendix H**.

Tab. 6: MPJPE under confidence-correlated cue errors (mm). σ: perturbation scale; Δ_{30}: increase from the unperturbed baseline.

Variant	$\sigma = 0$	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	Δ_{30}
No adaptive gate	40.1	41.2	43.4	46.5	+6.4
No pose feedback	40.2	41.1	42.9	45.5	+5.3
Static kinematics	40.6	41.8	44.1	47.3	+6.7
CoKinPose-S (Full)	39.2	39.7	40.6	42.0	+2.8

4.7 Pose Correction under Inaccurate 2D Observations

To assess whether CoKinPose can improve 3D estimation from inaccurate 2D observations, we qualitatively compare predicted poses. CoKinPose recovers more accurate limb configurations and better matches the available ground truth, supporting its ability to reduce the impact of input errors. Detailed comparisons are provided **in Appendix I**.

5 Limitations

CoKinPose remains challenged by extreme limb folding, body inversions, and unfamiliar motions, which can degrade the initial 3D estimate and its extracted relations. Observation confidence regulates kinematic contributions but cannot eliminate depth ambiguity in confident 2D detections; mutual refinement cannot guarantee unique recovery from insufficient observations. **Appendix J** provides examples and discussion. Future work includes reducing dependence on initial estimates and improving uncertainty estimation for predicted 3D relations.

6 Conclusion

CoKinPose jointly refines pose and explicit 3D kinematic representations for 3D-HPE. Skeletal and motion relations from an initial pose guide feature fusion and Mamba's selective state evolution, with confidence-guided gating regulating their contribution. Pose feedback refines kinematic representations, while joint supervision encourages the network to internalize their correspondence. Benchmark comparisons, component and gate ablations, and analyses of projection-related difficulty and unreliable kinematic cues support the complementary roles of these mechanisms.

REFERENCES

Md Mushfiqur Azam and Kevin Desai. A survey on 3D egocentric human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1643–1654, 2024.

Qingyuan Cai, Xuecai Hu, Saihui Hou, Li Yao, and Yongzhen Huang. Disentangled diffusion-based 3D human pose estimation with hierarchical spatial and temporal denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 882–890, 2024. doi: 10.1609/aaai.v38i2.27847.

Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Zhi-Qi Cheng, Wei Liu, Hanbing Liu, Bin Luo, Yifeng Geng, and Xuansong Xie. HDFormer: High-order directed transformer for 3D human pose estimation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 581–589, 2023. doi: 10.24963/ijcai.2023/65.

Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3D human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209, 2022. doi: 10.1109/TCSVT.2021.3057267.

Hu Cui, Wenqiang Hua, Renjing Huang, Shurui Jia, and Tessai Hayama. SasMamba: A lightweight structure-aware stride state space model for 3D human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2721–2730, 2026.

Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. DiffPose: Toward more reliable 3D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051, 2023.

Yan Guo, Tianhan Gao, Aoshuang Dong, Xinbei Jiang, Zichen Zhu, and Fuxin Wang. A survey of the state of the art in monocular 3D human pose estimation: Methods, benchmarks, and challenges. *Sensors*, 25(8):2409, 2025. doi: 10.3390/s25082409.

Xiaoyang Hao and Han Li. PersPose: 3D human pose estimation with perspective encoding and perspective rotation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8110–8119, 2025.

Chih-Hsiang Hsu and Jyh-Shing Roger Jang. Enhancing 3D human pose estimation with bone length adjustment. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 3723–3738, 2024.

Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. PoseMamba: Monocular 3D human pose estimation with bidirectional global-local spatio-temporal state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3842–3850, 2025. doi: 10.1609/aaai.v39i4.32401.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. MHFormer: Multi-hypothesis transformer for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13147–13156, 2022.

Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. TCPFormer: Learning temporal correlation with implicit pose proxy for 3D human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5478–5486, 2025. doi: 10.1609/aaai.v39i5.32583.

Ye Lu, Jie Wang, Jianjun Gao, Rui Gong, Chen Cai, and Kim-Hui Yap. A structure-aware and motion-adaptive framework for 3D human pose estimation with Mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7958–7968, 2025.

Soroush Mehraban, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3D human pose estimation with a Transformer-GCNFormer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6920–6930, 2024.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017. doi: 10.1109/3DV.2017.00064.

Jihua Peng, Yanghong Zhou, and P. Y. Mok. KTPFormer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1123–1132, 2024. doi: 10.1109/CVPR52733.2024.00113.

Andy Pramono, I-Cheng Chang, and Betty Dewi Puspasari. PRAMGCN-Net: 3D human pose estimation with a parameterized routing adjacency modulation graph convolutional network. *IEEE Access*, 13:159050–159063, 2025. doi: 10.1109/access.2025.3607702.

Cédric Rommel, Victor Letzelter, Nermin Samet, Renaud Marlet, Matthieu Cord, Patrick Pérez, and Eduardo Valle. ManiPose: Manifold-constrained multi-hypothesis 3D human pose estimation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107350–107378, 2024. doi: 10.52202/079017-3410.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-STMO: Pre-trained spatial temporal many-to-one model for 3D human pose estimation. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2022. doi: 10.1007/978-3-031-20065-6_27.

Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3D human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4790–4799, 2023.

Weiquan Wang, Jun Xiao, Chunping Wang, Wei Liu, Zhao Wang, and Long Chen. Di^{2}Pose: Discrete diffusion model for occluded 3D human pose estimation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98717–98741, 2024. doi: 10.52202/079017-3133.

Bruce X. B. Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. GLA-GCN: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8818–8829, 2023.

Hehao Zhang, Zhengping Hu, Shuai Bi, Jirui Di, and Zhe Sun. HMSFT: Hierarchical multi-scale spatial-frequency-temporal collaborative transformer for 3D human pose estimation. *Pattern Recognition*, 164:111562, 2025a. doi: 10.1016/j.patcog.2025.111562.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13232–13242, 2022.

Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose Magic: Efficient and temporally consistent human pose estimation with a hybrid Mamba-GCN network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10248–10256, 2025b. doi: 10.1609/aaai.v39i10.33112.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8877–8886, 2023.

Xingwang Zhao, Xuan Zhu, Boda Qie, Yifei Sun, Lin Wang, and Peiling Song. 2-D-to-3-D lifting human pose estimation based on simple graph convolution regression. *IEEE Sensors Journal*, 25 (19):36479–36488, 2025. doi: 10.1109/jsen.2025.3596065.

Kaili Zheng, Feixiang Lu, Yihao Lv, Liangjun Zhang, Chenyi Guo, and Ji Wu. 3D human pose estimation via non-causal retentive networks. In *Computer Vision – ECCV 2024*, pp. 111–128, 2025. doi: 10.1007/978-3-031-73414-4_7.

Zenghao Zheng, Lianping Yang, Jinshan Pan, and Hegui Zhu. Mamba-driven topology fusion for monocular 3D human pose estimation. *Image and Vision Computing*, 168:105927, 2026. doi: 10.1016/j.imavis.2026.105927.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15085–15099, 2023.

APPENDIX

A DATASETS, TRAINING, AND EVALUATION PROTOCOLS

A.1 DATASETS AND METRICS

Human3.6M (Ionescu et al., 2014) contains recordings of 11 subjects performing 15 actions. We train on S1, S5, S6, S7, and S8 and evaluate on S9 and S11. Protocol 1 reports mean per-joint position error (MPJPE) after root alignment, and Protocol 2 reports Procrustes-aligned MPJPE (P-MPJPE), both in millimeters. Per-action errors are averaged with equal action weights. MPI-INF-3DHP (Mehta et al., 2017) is evaluated using MPJPE, 3D PCK, and AUC under its standard evaluation protocol, with PCK and AUC expressed as percentages. The MPI-INF-3DHP comparison covers its test environments jointly; it is not an evaluation restricted to outdoor images.

All poses and velocities in our relation analyses use camera coordinates relative to the pelvis in each frame. For a pose $\mathbf{P}$ and its ground truth $\mathbf{P}^*$, the velocity error is

$$E_{\mathrm{vel}} = \frac{1}{(T-1)N} \sum_{t=2}^{T} \sum_{j} \left\| (\mathbf{P}_{t,j} - \mathbf{P}_{t-1,j}) - (\mathbf{P}^*_{t,j} - \mathbf{P}^*_{t-1,j}) \right\|_2 . \tag{8}$$

This MPJVE is measured in mm/frame at the original sampling interval. It measures agreement with ground-truth motion rather than trajectory smoothness alone. Root-relative velocity does not include global body translation. Predictions use the same sampled frames and temporal alignment across methods; overlapping predictions are averaged before temporal evaluation.

For skeletal relations, bone-length error is the mean absolute difference in millimeters. Bone-direction error is the angle in degrees between predicted and ground-truth parent-to-child bone vectors. Direction vectors are normalized before taking their clipped dot product. Invalid zero-length bones are excluded from the angular metric. For decoded motion relations, the velocity vector is reconstructed by multiplying the predicted magnitude by the normalized predicted direction; its error is evaluated against the ground-truth velocity vector. This relation-output error is distinct from MPJVE computed by differencing final pose predictions. Depth and depth-velocity errors are the mean absolute errors of the root-relative camera-axis position and velocity components, in mm and mm/frame, respectively. These metrics are computed without Procrustes alignment so that an incorrect depth orientation remains visible.

A.2 TRAINING AND CONFIDENCE INPUTS

CoKinPose uses the same 2D inputs, coordinate preprocessing, and augmentation convention across all internal comparisons. Detector confidence scores are clipped to $[0, 1]$ and paired with the corresponding 2D joints. For ground-truth 2D inputs, valid joints receive confidence one. Skeletal and motion confidence cues are then constructed and aligned using the method's rules. Confidence indicates observation quality; it is not a label for the accuracy of a predicted 3D relation.

The small, base, and large variants have 1.0M, 3.7M, and 18.0M parameters, respectively. We train with AdamW for 120 epochs, using an initial learning rate of 5×10^{-4}, weight decay 0.01, and learning-rate decay factor 0.99 per epoch. Component and gate ablations use the small variant, whereas the targeted ambiguity analysis uses the large variant. All internal comparisons retain the same input window, training budget, and final-pose objective unless that setting is explicitly varied. The longer-window comparison uses separately trained 243- and 351-frame models evaluated on the same target frames. The MPI-INF-3DHP experiment uses the large variant with an 81-frame input window and training on that dataset's training split.

Tab. 1 reports parameters and multiply–accumulate operations for the displayed input length. Its MACs/frame column is the integer part of $1000\,\mathrm{MACs(G)}/T$. This per-input-frame cost does not equate the output coverage of center-frame and sequence-to-sequence methods.

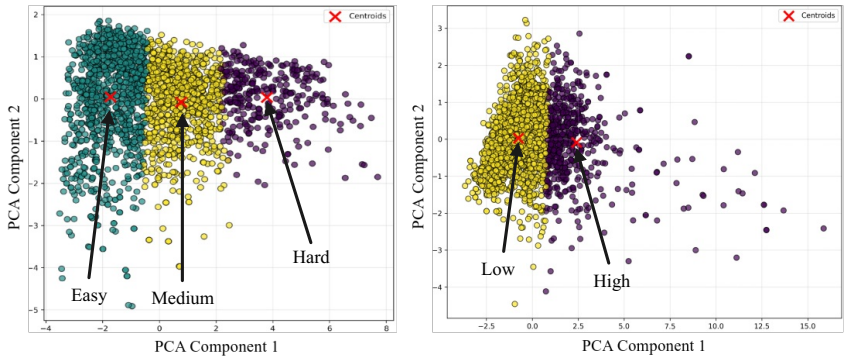
B Confidence and Motion Partitions

We apply unsupervised K-means to the Human3.6M test observations, using non-root joint-confidence vectors for three confidence clusters and motion descriptors for two motion clusters. The motion descriptors combine ground-truth root-relative velocity, acceleration, directional changes, and FFT amplitude. Features are standardized before clustering. Confidence clusters are named Easy, Medium, and Hard by decreasing mean confidence; motion clusters are named Low and High by increasing motion intensity. Cluster sizes are determined by K-means rather than prescribed proportions. The resulting assignments are fixed across methods. Ground-truth motion is used only for evaluation grouping and is not supplied to the model. Fig. 6 displays the clusters in two principal components; PCA is used for visualization.

For action a, let $w_{a,g}$ be the observed fraction of samples in cluster g and $E_{a,g}$ their mean error. With A actions, the action-averaged overall error is

$$E_{\text{all}} = \frac{1}{A} \sum_{a} \sum_{g} w_{a,g} E_{a,g} = \sum_{g} w_g E_g, \tag{9}$$

where $w_g = A^{-1} \sum_a w_{a,g}$ and $E_g = \sum_a w_{a,g} E_{a,g} / \sum_a w_{a,g}$ are the effective cluster proportion and cluster error. Empty action–cluster pairs contribute zero. Confidence groups describe observation reliability, not verified occlusion labels.



(a) Observation-confidence groups **(b) Motion-intensity groups**

Fig. 6: PCA views of K-means clusters for confidence (left) and motion intensity (right).

C Action-wise Pose Estimation

Tab. 7 reports the 15 Human3.6M actions under Protocol 1. The Avg column is the unweighted mean of the displayed per-action errors. This decomposition shows performance across actions; action names do not specify occlusion, observation confidence, or projection ambiguity.

CoKinPose-S improves on PoseMamba-S across the displayed actions, including seated postures and locomotion. The remaining errors in complex seated configurations are larger than those in walking, showing that improvements do not remove the difficulty of all poses. Differences from other methods vary by action. We attribute no action-level gap to a hard skeletal constraint: CoKinPose learns from explicit relation representations but does not impose fixed bone lengths or exact physical dynamics.

Tab. 7: Per-action MPJPE on Human3.6M (mm). Avg is the mean over the 15 actions.

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
MHFormer (Li et al., 2022)	39.2	43.1	40.1	41.0	45.0	51.3	40.7	41.4	53.6	60.4	43.8	41.2	43.8	29.8	30.6	43.0
MixSTE (Zhang et al., 2022)	37.6	40.9	37.3	39.7	42.3	50.0	40.2	39.9	51.8	55.1	42.1	39.8	41.0	27.9	27.9	40.9
P-STMO (Shan et al., 2022)	38.8	42.6	40.3	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.1	42.6	29.3	29.3	42.8
STCFormer (Tang et al., 2023)	39.5	41.5	37.3	38.7	43.1	51.1	39.1	39.7	51.4	57.4	41.7	38.4	40.6	27.0	28.5	41.0
GLA-GCN (Yu et al., 2023)	41.3	44.3	40.8	41.8	45.9	54.2	42.2	41.6	57.9	63.0	45.0	42.8	45.9	29.4	29.9	44.4
KTPFormer (Peng et al., 2024)	37.3	39.2	35.9	37.7	42.6	48.3	38.7	39.1	51.5	56.0	41.7	39.1	40.0	27.0	27.4	40.1
PoseRetNet (Zheng et al., 2025)	36.8	40.1	38.7	38.3	42.9	48.6	38.2	40.0	52.5	55.4	42.3	38.7	39.7	26.1	27.7	40.4
PoseMamba-S (Huang et al., 2025)	39.4	41.5	39.8	35.6	43.4	52.7	40.2	37.9	53.1	63.4	42.3	39.9	39.5	29.2	29.1	41.8
SasMamba (Cui et al., 2026)	39.0	43.0	39.2	36.1	43.2	51.4	40.1	37.3	52.8	61.7	43.1	38.9	39.8	28.3	28.6	41.5
CoKinPose-S	**38.4**	**40.0**	**36.3**	**34.1**	**41.8**	**47.4**	**37.5**	**36.3**	**49.8**	**55.1**	**41.3**	**38.0**	**38.4**	**26.6**	**27.0**	**39.2**

D GATE AND FEEDBACK CONTROLS

For the gate ablations, let the two-channel mask $\mathbf{M}$ indicate valid non-root skeletal positions and valid non-initial temporal positions. Its channel mean preserves the same structural boundary treatment in all variants. No adaptive gating uses $\mathbf{G}_l = \overline{\mathbf{M}}$, which equals one where both relation types are valid. Content-only gating replaces $\mathbf{Q}$ by $\mathbf{M}$ both inside and outside the gate-generating mapping, removing all actual detection confidence. Confidence-only gating uses $\mathbf{G}_l = \overline{\mathbf{Q}}$. The full variant uses Eq. (3). Each gate is used consistently in both feature fusion and explicit state modulation.

Tab. 8: Gate ablations with CoKinPose-S. Overall and confidence-group errors are MPJPE (mm); MPJVE is in mm/frame. The same rule controls fusion and explicit modulation.

Gate	Overall	Easy	Medium	Hard	MPJVE
No adaptive gating	40.1	35.4	46.0	61.2	1.75
Content-only gating	39.7	35.3	45.6	58.7	1.70
Confidence-only gating	39.6	35.4	45.0	58.2	1.71
Full confidence-guided gate	**39.2**	**35.2**	**45.0**	**55.6**	**1.65**

Tab. 8 compares the four gate rules. Combining observation confidence and kinematic content gives the lowest errors among the tested variants, with the clearest benefit on low-confidence observations.

Removing explicit state modulation sets $\widehat{\mathbf{\Theta}}_l = \mathbf{\Theta}_l$ while preserving gated feature fusion. Removing pose-to-kinematics feedback retains the kinematic residual MLP but omits $\mathrm{Linear}^l_{\mathrm{fb}}(\mathbf{X}_l)$. Static kinematics supplies $\mathbf{Kin}_0$ to every pose-update block, and its final relation decoder reads $\mathbf{Kin}_0$ with the same decoder architecture and supervision as the other variants. The pose-only control retains the pose backbone and initial/final pose objectives but removes the kinematic branch; it is an internal control, not a renamed PoseMamba model. Removing an input relation masks its channels before the kinematic encoder while retaining the relation prediction objective. Auxiliary-loss ablations instead change only the indicated loss term. All these distinctions are kept separate in Tab. 4.

E SENSITIVITY TO AUXILIARY LOSS WEIGHTS

The coefficients λ_1 and λ_2 in Eq. (7) balance initial-pose supervision and decoded kinematic-relation supervision against the final-pose objective. Tab. 9 varies one coefficient at a time while fixing the other to one, using CoKinPose-S under the same evaluation protocol as the component ablations.

Both auxiliary objectives contribute to the refinement process. Reducing the weight of initial-pose supervision changes the training signal for the pose from which the initial relations are extracted. Changing the relation-supervision weight changes the emphasis on retaining decodable skeletal and motion information in the kinematic representation. Among the tested configurations, $\lambda_1 = \lambda_2 = 1$ gives the lowest MPJPE, matching the full-model result in Tab. 4. These comparisons assess the selected training weights; the losses do not impose fixed bone lengths or hard motion constraints.

Tab. 9: Sensitivity to the initial-pose weight λ_1 and kinematic-relation weight λ_2 for CoKinPose-S. MPJPE is in millimeters.

λ_1	λ_2	MPJPE
0.25	1	40.0
0.5	1	39.5
1.5	1	39.6
1	0.25	40.3
1	0.5	39.7
1	1.5	39.8
1	**1**	**39.2**

F Additional Component and Relation Analyses

F.1 Effects of Explicit State Modulation

Fig. 7 shows the state-modulation control in Tab. 4, retaining gated feature fusion.

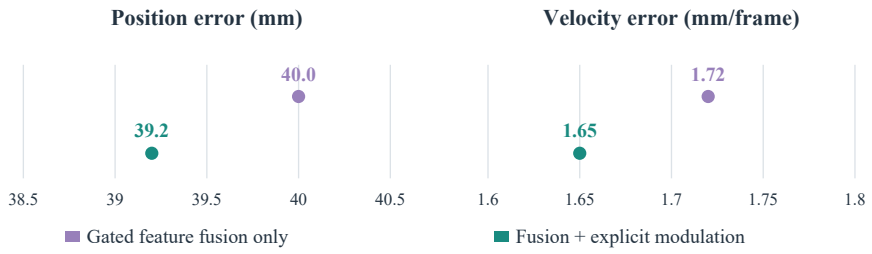


Fig. 7: Effect of explicit parameter modulation with gated feature fusion retained.

F.2 Initial and Refined Kinematic Relations

Tab. 10 compares initial and final pose and relation errors with the feedback and static controls. Fig. 8 additionally relates the confidence groups to initial relation quality. For static kinematics, the final relation decoder reads the unchanged $\mathbf{Kin}_0$.

Tab. 10: Initial and final estimates for CoKinPose-S. Initial relations are extracted from the initial pose; final relations are decoded from the final available kinematic representation. Pose error and bone-length error (BL): mm; bone-direction error (BO): degrees; relation velocity-vector error (VV): mm/frame. VV is distinct from pose-derived MPJVE.

Variant	Stage	Pose error	BL	BO	VV
Full feedback	Initial	44.3	5.8	8.9	2.35
	Final	**39.2**	**3.9**	**6.8**	**1.80**
Without pose feedback	Initial	44.4	5.9	9.0	2.38
	Final	40.2	4.6	7.6	2.04
Static kinematics	Initial	44.2	5.8	8.9	2.36
	Final	40.6	4.9	8.0	2.16

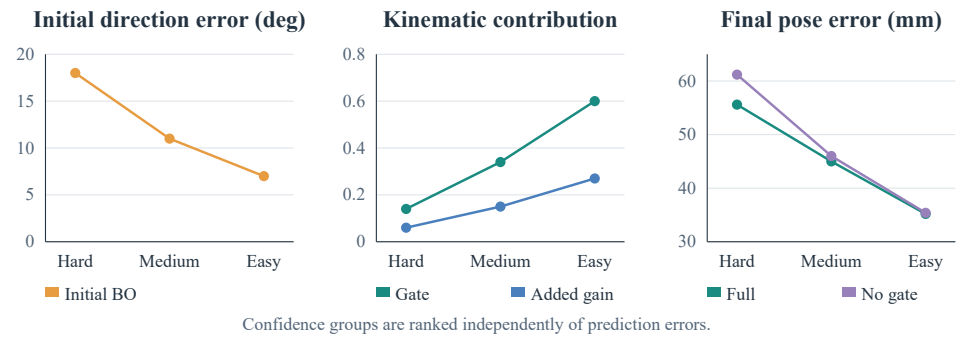


Fig. 8: Initial relation quality, contribution strength, and pose error by confidence group. Gate statistics use locations where both relation types are valid.

G Projection-related Difficulty

The spatial and temporal analyses use the same test observations as the aggregate evaluation. For a non-root ground-truth bone vector $\mathbf{b}_{t,j}$ in camera coordinates, the spatial score is

$$a_s(t) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \frac{b_{t,j,z}^2}{\|\mathbf{b}_{t,j}\|_2^2 + \epsilon}, \tag{10}$$

where $\mathcal{B}$ contains the upper- and lower-arm and upper- and lower-leg bones. A larger score indicates more depth-oriented limbs and weaker image-plane information about their 3D directions. Within each action, the highest-scoring 30% form the spatially difficult group; the remaining 70% form its complementary group. This is a depth-orientation proxy for projection-related difficulty, not a claim that the complete temporal input is identical across distinct poses.

For temporal difficulty, let $\mathbf{v}_{t,j}^*$ be the ground-truth root-relative velocity. Over the full input window $\mathcal{W}$, define

$$a_t = \frac{\sum_{t \in \mathcal{W}, j \neq r} (v_{t,j,z}^*)^2}{\sum_{t \in \mathcal{W}, j \neq r} \|\mathbf{v}_{t,j}^*\|_2^2 + \epsilon}. \tag{11}$$

Within each action, windows are first stratified into quartiles of total root-relative motion energy. Within each stratum, the upper 25% by a_t form the depth-dominant group, and the other 75% form its complement. This controls the coarse distribution of motion magnitude while testing how its orientation affects prediction. Stationary windows receive score zero and remain in the ranked evaluation population; the denominator safeguard prevents undefined ratios. The group comparison remains a stress test of projection cues rather than a demonstration of mathematical non-identifiability.

All compared methods use the same 2D observations, difficulty partitions, coordinate convention, and temporal windows. Confidence scores are supplied only to models whose input formulation uses them. The full-window difficulty score is assigned to its center-frame pose estimate and the velocity ending at that frame; overlapping predictions are merged before evaluation. Velocity metrics exclude sequence-initial frames lacking a predecessor. For these difficulty partitions, fractional weights at quantile boundaries preserve the specified proportions on each metric's common eligible population. After action averaging, the spatial and temporal MPJPE values satisfy

$$E_{\text{all}} = 0.7E_{\text{spatial,other}} + 0.3E_{\text{spatial,hard}} = 0.75E_{\text{temporal,other}} + 0.25E_{\text{temporal,hard}}. \tag{12}$$

The corresponding temporal velocity errors use the same weights. Subset construction is independent of predicted errors, and no group-specific fine-tuning is used.

H Controlled Perturbation of Initial Kinematic Cues

The stress test in Tab. 6 uses frozen CoKinPose-S variants with detected 2D inputs and $T = 243$. Each variant retains its own unperturbed pose representation $\mathbf{X}_0$ and initial pose $\mathbf{P}_{13D}$. For constructing

the initial kinematic descriptors only, we perturb a copy of that pose:

$$\widetilde{\mathbf{P}}_{t,j} = \mathbf{P}_{13\mathrm{D},t,j} + \sigma(1 - S_{t,j})\mathbf{z}_{t,j}, \qquad j \neq r, \tag{13}$$

where $S_{t,j}$ is the original detector confidence and $\mathbf{z}_{t,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$. The root remains zero. The same fixed perturbation field is shared across variants and scaled by $\sigma \in \{0, 10, 20, 30\}$ mm. All four descriptor types are re-extracted together from the perturbed pose using the method's normalization and alignment rules. Independent frame-wise perturbations affect both skeletal geometry and inter-frame motion; the test does not isolate either relation type.

The 2D inputs, confidence cues $\mathbf{Q}$, initial pose features $\mathbf{X}_0$, targets, and evaluation protocol remain unchanged; models are not retrained. Later pose and kinematic representations are updated normally using the perturbed initial cues. At zero perturbation, every variant recovers its own original evaluation result. We compare increases relative to those baselines to separate sensitivity from existing accuracy differences. Because corruption is explicitly coupled to low observation confidence, this experiment tests resilience to confidence-correlated cue errors. It does not establish that observation confidence detects arbitrary 3D errors, particularly incorrect depth estimates from confident 2D observations.

I Qualitative Comparisons under Imperfect Observations

These comparisons expand the visual analysis outlined in Section 4.7. They show how different methods respond to imperfect 2D detections and challenging postures. Input noise and projection ambiguity are related sources of difficulty but are not interchangeable, so examples without 3D ground truth are used to inspect visible structural differences rather than to establish recovery accuracy.

I.1 Pose Refinement with Inaccurate 2D Joints

Fig. 9 presents the input-error examples alongside MotionAGFormer (Mehraban et al., 2024), PoseMamba (Huang et al., 2025), and CoKinPose. The marked arm and torso regions illustrate differences in the reconstructed joint configurations when the corresponding 2D observations are unreliable. CoKinPose uses skeletal and motion relations to guide pose updates while regulating their contribution with observation confidence. The examples visualize this setting; the component and gate comparisons in Tabs. 4 and 8 evaluate the associated prediction errors.

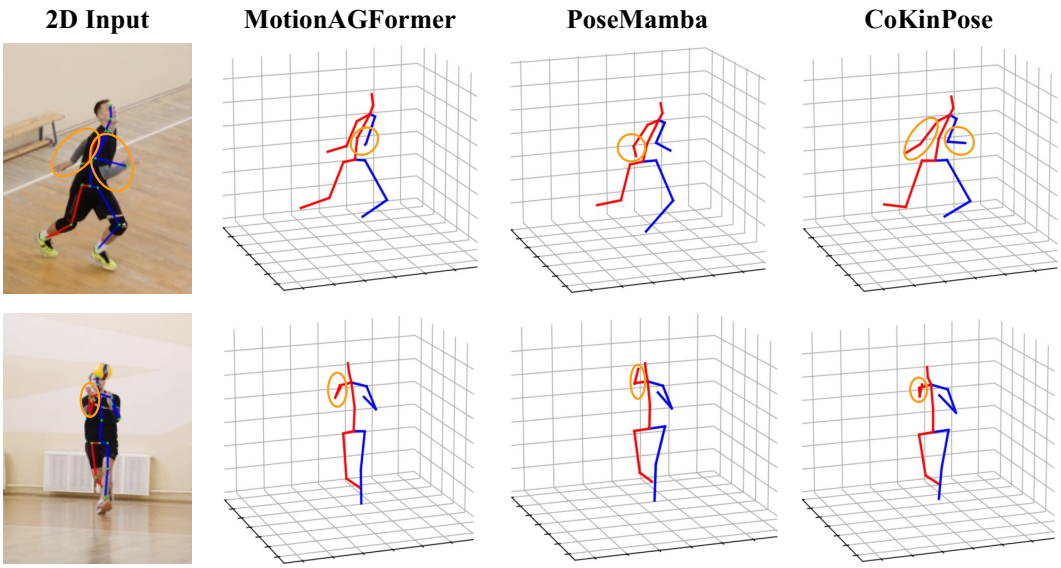


Fig. 9: Qualitative pose comparisons with inaccurate 2D detections. Highlighted regions indicate differences in reconstructed limb configurations.

I.2 Challenging Poses in Unconstrained Scenes

Fig. 10 extends the comparison to unusual motions, self-occlusion, and changes in viewpoint. Some limb and torso configurations differ substantially across methods. These cases illustrate the behavior of kinematics-guided refinement beyond the standard test images, while also exposing sensitivity to the quality of the initial 2D observations.

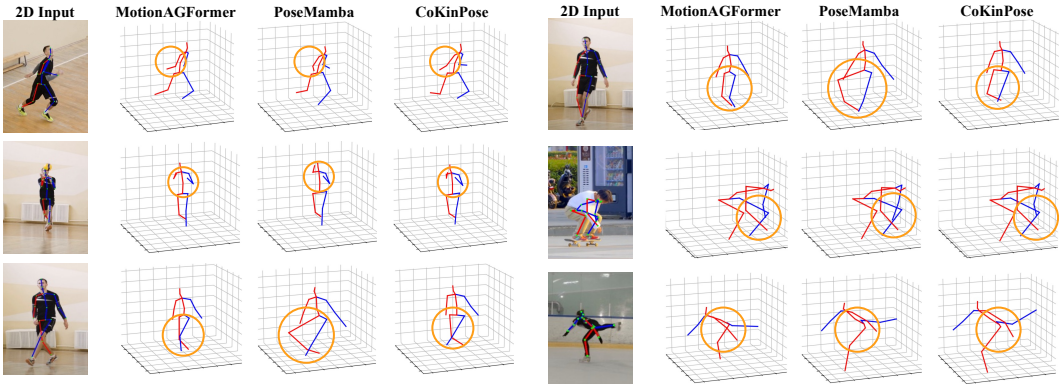


Fig. 10: Qualitative comparisons in unconstrained scenes with imperfect observations.

I.3 Alignment with Ground-Truth Poses

Fig. 11 overlays the predicted poses with ground truth in a common coordinate frame. This comparison makes errors in limb direction and relative joint placement visible. CoKinPose's predictions more closely follow the displayed reference poses in several highlighted regions, consistent with the purpose of incorporating target-space skeletal relations. The quantitative ambiguity-related evaluation is reported separately in Tab. 5.

J Failure Cases and Remaining Limitations

Fig. 12 shows unusual body orientations and severe limb folding. Such poses can depart from training patterns and degrade initial 2D detections. Errors in the initial 3D estimate then affect the skeletal and motion descriptors used for refinement.

CoKinPose relies on an initial 3D estimate and on confidence cues from 2D observations. A confident detection can still admit an incorrect depth configuration. Confidence-guided gating regulates the contribution of relation cues but does not directly measure 3D correctness. Moreover, modeled motion is relative to the pelvis and excludes global body translation. Strong self-occlusion, unusual viewpoints, and out-of-distribution motions remain challenging. The ambiguity-related evaluations measure accuracy under reduced projection cues; they do not imply unique recovery when distinct 3D states produce indistinguishable complete inputs. Severely corrupted observations can leave insufficient context for mutual refinement. Richer visual evidence and uncertainty estimates for predicted 3D relations are possible directions for improvement.

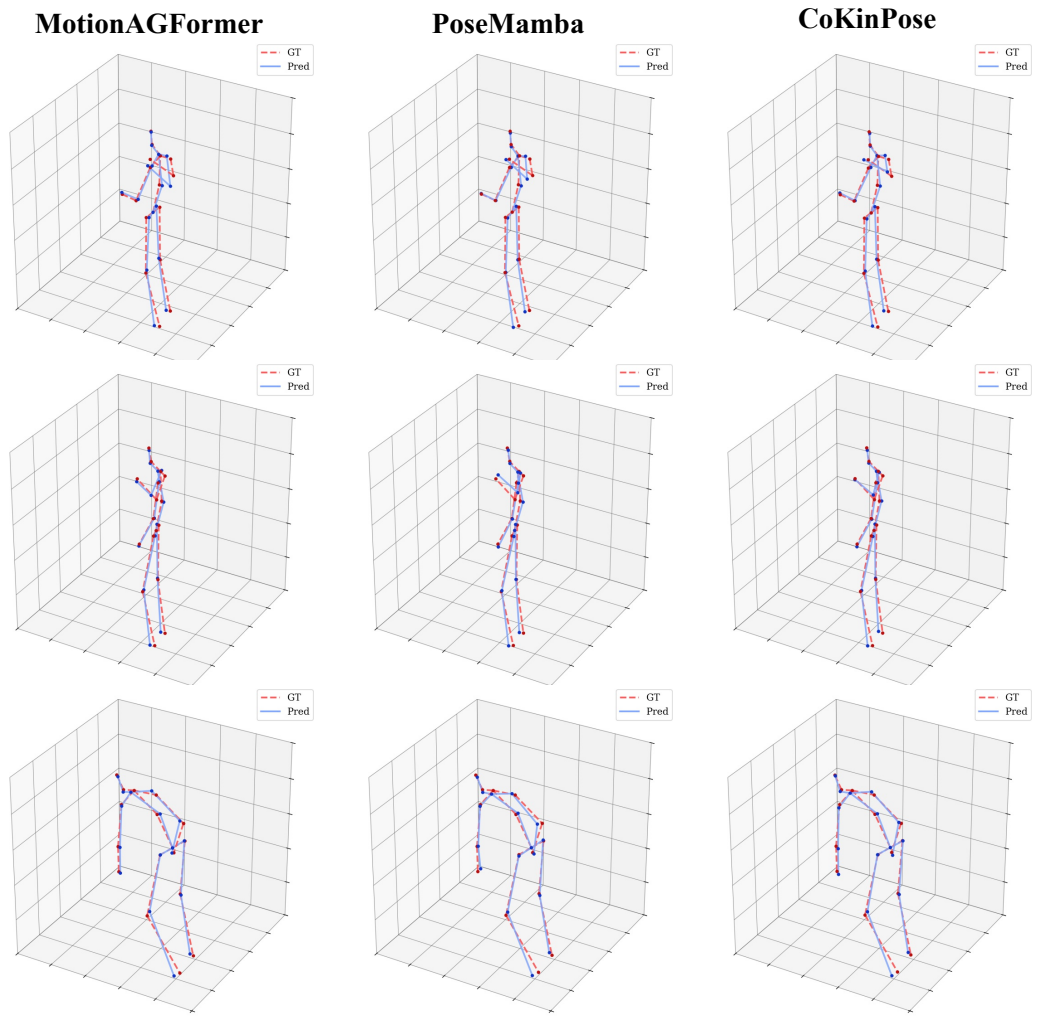


Fig. 11: Predicted poses (solid blue) overlaid with ground truth (dashed red) to compare relative joint placement and limb direction.

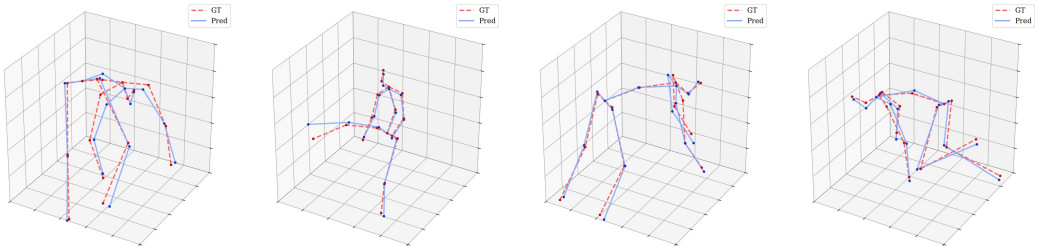


Fig. 12: Failure cases involving unusual body orientations and severe limb folding. Differences between predicted and ground-truth poses show remaining reconstruction errors.

AI Use Statement

Generative AI assisted with translation and text revision, typesetting, and logical checks. The authors retain responsibility for all submitted content.