KIJACOMMP: KINEMATICS-AWARE JACOBIAN ITERATIVE SOLVER FOR MULTI-PERSON MOTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-person motion prediction aims to infer the future 3D poses of all individuals in a scene from historical observations. Existing methods typically predict future joint coordinates directly by learning spatiotemporal correlations. However, small coordinate errors do not guarantee that a predicted pose can be continuously reached from the current motion state. To address these issues, we propose KiJacoMMP, a **ki**nematics-aware **jaco**bian iterative solver for **m**ulti-person **m**otion **p**rediction. KiJacoMMP firstly exploits acceleration field encoder characterizing the evolution of individual velocities and interaction relation encoder modeling spatial dependencies among different individuals. Subsequently, the model takes the last observed pose as the initial condition, integrates the preliminary velocities frame by frame into joint positions, and constructs constraint residuals from predefined skeletal length. We further derive a sparse skeletal Jacobian matrix and an edge coupling matrix, transforming each nonlinear skeletal correction into a damped linearized system. Through multiple iterations, the system jointly updates shared joints, thereby producing kinematically feasible future trajectories while preserving the preliminary motion trend as much as possible. The proposed method requires neither body mass nor an external physics simulator. Experiments on current multi-person motion prediction datasets demonstrate the effectiveness of our proposed method.

1 INTRODUCTION

Multi-person motion prediction (MMP) is one of the core tasks in the interdisciplinary field of computer vision. This technology plays an irreplaceable role in key application scenarios (e.g., autonomous driving Hu et al. (2021); Ghorai et al. (2022); Liu et al. (2025b;a), human-computer interaction Tang et al. (2025); Ro et al. (2019); Yang et al. (2025); Chen et al. (2023), and virtual reality Kim et al. (2022); Zhang et al. (2025); Mitra et al. (2025); Hermawan et al. (2023)). With the development of deep learning technologies, existing methods Zheng et al. (2025); Jeong et al. (2024); Vendrow et al. (2022) have made significant progress in generating diverse motion sequences and achieving visual realism. However, human motion, as a highly articulated dynamic system, is inherently constrained by both kinematics or physics that has not received sufficient attention in existing prediction research Vendrow et al. (2022); Peng et al. (2023).

As illustrated in Figure. 1a, earlier methods (e.g. TBIFormer Peng et al. (2023), T2P Jeong et al. (2024)) primarily model and predict one-person motion, then used relative position to infer spatial coordinates. The methods of recent years (e.g. JRFormer Xu et al. (2023), IAFormer Xiao et al. (2024)) presented in Figure. 1b refined the potential connections between people by leveraging joint interactions among multi-individuals after modeling one-person motion. Most existing methods focus solely on fitting the conditional probability between historical and predicted sequences, rather than addressing the physics feasibility within the skeletal system. Some approaches have also investigated physical considerations in human motion prediction, with many proposing simulation models Curreli et al. (2025); Zhang et al. (2024b); Lu et al. (2026) based on dynamical theorems. But the acquisition of such dynamical parameters(e.g.mass, inertia, and muscle forces) are often challenging to get, limiting their practical applicability. Consequently, how can we leverage read-

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.

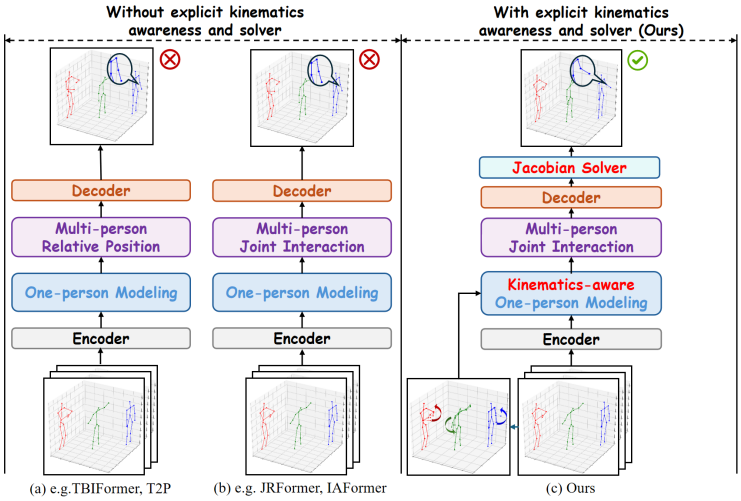


Figure 1: Insight of our work. (a)-(b): Limitations of prior methods ignore kinematics constraints. (c): Our KiJacoMMP has explicit kinematics awareness and solver.

ily observable kinematic information to achieve future motion predictions that satisfy full skeletal coupling constraints?

To address the above issue, we propose KiJacoMMP, a kinematics-aware jacobian iterative solver for multi-person motion prediction. KiJacoMMP employs an interaction relation encoder to extract interaction associations between different individuals, and an acceleration field encoder to characterize the temporal evolution of velocity; subsequently, a velocity field decoder integrates individual motion data with multi-person interaction information to generate preliminary joint velocities for future time steps. The model uses the historical pose as its initial condition and performs frame-by-frame integration of the predicted velocities to obtain preliminary joint positions. For each future time step, the system computes constraint residuals based on predefined skeletal edges, constructs a sparse skeletal Jacobian matrix and an edge coupling matrix, and solves a damped linearized system of equations. The Jacobian iterative solver then refines all shared joints, updating the joint velocities after each refinement step. After multiple iterations, the model ultimately outputs a future motion sequence that accounts for multi-person interaction trends, temporal continuity, and consistency with skeletal kinematics.

Our contributions are summarized as follows:

- KiJacoMMP proposes a kinematics-aware framework for multi-person motion prediction. The framework exploits the discrete time relationships among position, velocity, and acceleration to jointly model individual motion evolution and multi-person interaction information, thereby predicting the future velocity field of each joint.

- KiJacoMMP formulates future motion generation as a discrete constrained integration problem driven frame by frame by the preliminary velocity field. Taking the last observed pose as the initial condition, the system first produces a position at each future time step according to the predicted preliminary velocity. With the constraints imposed by the rigid bones, the originally independent joint predictions are transformed into a joint high-order equation iterative solution problem for full-skeleton coupled motions under shared bone constraints.

- KiJacoMMP derives the sparse Jacobian matrix and edge coupling matrix associated with the skeletal constraints and transforms each nonlinear constraint correction into a damped linearized system. Through multiple iterations, this system simultaneously correct the preliminary predicted velocity field, producing kinematically feasible predictions.

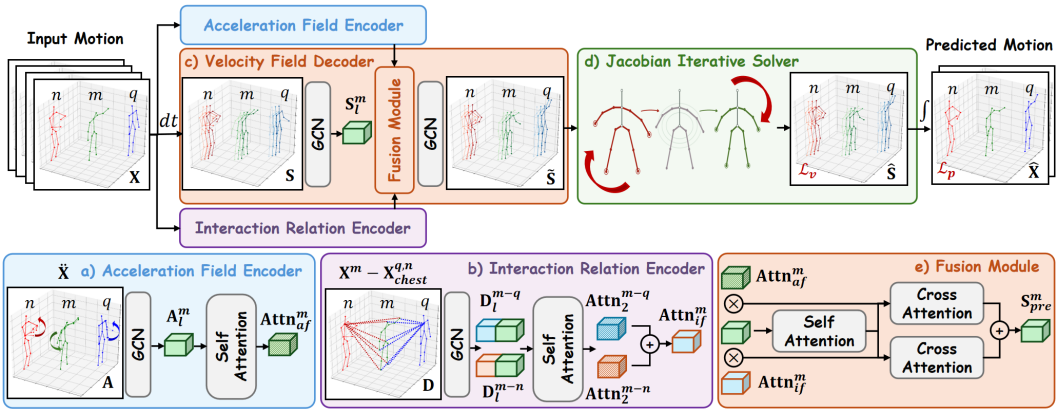


Figure 2: The pipeline of KiJacoMMP, a kinematics-aware Jacobian for multi-person motion prediction. KiJacoMMP contains four core components: a) Acceleration Field Encoder (AFE), b) Interaction Relation Encoder (IRE), c) Velocity Field Decoder (VFD), and d) Jacobian Iterative Solver (JIS). e) Fuison Module is a part of VFD.

2 Related Work

2.1 Multi-person Motion Prediction

Multi-person motion prediction aims to predict the individual motions while capturing their complex interactions with others. Current methods almost rely on Transformer regression using complex strategies to capture spatiotemporal information. MRT Wang et al. (2021) introduces local-range encoder for individual motion and global range encoder for social interactions, while TBIFormer Peng et al. (2023) transforms all the pose sequences into a multi-person bodypart sequence to retain spatial and temporal information. JRFormer Xu et al. (2023) utilizes relation information to enhance interaction modeling and improve future motion prediction. T2P Jeong et al. (2024) introduces a graph-based agent wise module for a reciprocal forecast of local motion conditioned global trajectory and trajectory conditioned local pose. IAFormer Xiao et al. (2024) introduces the interaction-aware pose forecasting transformer to better learn the interaction features. ST-MoE Yin et al. (2026) incorporates four distinct types of spatiotemporal experts, each specializing in capturing different spatial or temporal dependencies. In contrast, KiJacoMMP is inspired by kinematics and use Jacobian iteration to optimize solution in discrete skeleton system, improving the kinetic plausibility of motion sequences.

2.2 Physics-informed Motion Modeling

To achieve a more realistic and reasonable motion modeling, existing methods have utilized physics engines and laws to model human motion. PhysGenLiu et al. (2024) integrates model based kinetic simulation with a data driven video generation process, enabling plausible image dynamics. PhysDiff Yuan et al. (2023) also uses motion imitation in a physics simulator to project the denoised motion of a diffusion step to a kineticly plausible motion. LARP Andriluka et al. (2024) propose a network to support the construction of a recurrent neural architecture to accurately simulate articulated rigid body dynamics. Morph Li et al. (2025) learn a motion imitator within a physics simulator, enforcing kinetic constraints to project the noisy motions into a kineticly plausible space. Additionally, direct application of kinetic equations has also proved to be effective Zhang et al. (2024a); Wei et al. (2023); Zhang et al. (2024b). PIMNet Zhang et al. (2022) calculates human dynamic equations for future motions through lagrangian equation. PhysMop Zhang et al. (2024c) use neural networks to find lagrangians. PhysPT Zhang et al. (2024b) build a physics based body representation and leverage them to impose novel physics inspired training losses. However, not all datasets have information on the mass or momentum of objects, making it difficult to apply lagrangian dynamics or other unified equations under such constraints. KiJacoMMP uses velocity and acceleration, which are not limited by acquisition, to construct kinematic fields and jacobian iterative solver to predict multi-person motion.

3 METHODOLOGY

As illustrated in Figure. 2, KiJacoMMP contains four core components: a) Acceleration Field Encoder (AFE), b) Interaction Relation Encoder (IRE), c) Velocity Field Decoder (VFD), and d) Jacobian Iterative Solver.

3.1 ACCELERATION FIELD ENCODER

Given the input motion $\mathbf{X} \in \mathbb{R}^{P\times T\times 3J}$, which means the motion for P person, T frames, and 3-axis J joint nodes, we calculate the instantaneous velocity and acceleration of each frame:

$$(\mathbf{S}_i^m)_j = ((\mathbf{X}_i^m)_j - (X_{i-1}^m)_j)/t, \quad (\mathbf{A}_i^m)_j = ((\mathbf{S}_i^m)_j - (S_{i-1}^m)_j)/t, \tag{1}$$

where t is the time between each frame. $(\mathbf{X}_i^m)_j$, $(\mathbf{S}_i^m)_j$ and $(\mathbf{A}_i^m)_j$ respectively represent the coordinates, velocity and acceleration of m-th person's j-th joint nodes in the t-th frame. The sets of each of $(\mathbf{S}_i^m)_j$ and $(\mathbf{A}_i^m)_j$ represent the m-th person required velocity $\mathbf{S}^m$ and acceleration $\mathbf{A}^m$:

$$\mathbf{S}^m=\{(\mathbf{S}_i^m)_j|i=1,2,...,T,\ j=1,2,...,J\}, \quad \mathbf{A}^m=\{(\mathbf{A}_i^m)_j|i=1,2,...,T,\ j=1,2,...,J\}. \tag{2}$$

In order to accurately determine which joint nodes should be paid more attention in the acceleration evolution, we use graph convolutional network (GCN) to embed the acceleration to obtain the latent representation $\mathbf{A}_l=GCN(\mathbf{A})$ and input it to the self attention layer to model the acceleration field sequence:

$$\mathbf{Q}_1^m = \mathbf{W}_1^Q\mathbf{A}_l^m, \mathbf{K}_1^m = \mathbf{W}_1^K\mathbf{A}_l^m, \mathbf{V}_1^m = \mathbf{W}_1^V\mathbf{A}_l^m, \mathbf{Attn}_{af}^m = \text{softmax}(\mathbf{Q}_1^m{\mathbf{K}_1^m}^\top/\sqrt{d_{K_1}}), \tag{3}$$

where $\mathbf{Attn}_{af}^m$ represents the acceleration field attention of m-th person's each node.

3.2 INTERACTION RELATION ENCODER

Then, we take the human chest node as the overall semantic point of the whole body, and calculate the distance between each person's joints and other people's chest to model the interaction relation field. Concretely, given the input motion $\mathbf{X}$, we divide the all nodes of each person from the chest nodes of other people as the spatial interaction distance between this person and other people. As illustrated in Figure. 3, in terms of the m-th and n-th person, it can be calculated as follows:

$$(\mathbf{D}_i^{m-n})_j = (\mathbf{X}_i^m)_j - (\mathbf{X}_i^n)_{chest}, \tag{4}$$

where $(\mathbf{D}_i^{m-n})_j$ represents the distance between the m-th person's j-th joint node and the n-th person's chest node in the i-th frame. As the same with velocity and acceleration, the set of $(\mathbf{D}_i^{m-n})_j$ represents the required interpersonal distance $\mathbf{D}^{m-n}$:

$$\mathbf{D}^{m-n}=\{(\mathbf{D}_i^{m-n})_j|i=1,2,...,T,\ n=1,2,...,P,\ j=1,2,...,J,\ m\neq n\}. \tag{5}$$

In order to accurately determine which distance should be paid more attention at each moment, we use graph convolutional network to embed the input coordinates to obtain the implicit spatial representation $\mathbf{D}_l=GCN(\mathbf{D})$ and input each person's interaction distance into the self attention layer to model the spatial change trend of interaction:

$$\begin{aligned}\mathbf{Q}_2^{m-n} = \mathbf{W}_2^Q\mathbf{D}_l^{m-n}, \mathbf{K}_2^{m-n} = \mathbf{W}_2^K\mathbf{D}_l^{m-n}, \mathbf{V}_2^{m-n} = \mathbf{W}_2^V\mathbf{D}_l^{m-n},\\ \mathbf{Attn}_2^{m-n} = \text{softmax}(\mathbf{Q}_2^{m-n}{\mathbf{K}_2^{m-n}}^\top/\sqrt{d_{K_2}}),\end{aligned} \tag{6}$$

where $\mathbf{Attn}_2^{m-n}$ denotes the interaction relation trend in terms of space, which represents the influence of the n-th person with the m-th person from the perspective of the m-th person. Since there are multi-person interactions, we sum and average the score of the m-th person with other people as the total interaction relation attention $\mathbf{Attn}_{if}^m$ of m-th person's each node:

$$\mathbf{Attn}_{if}^m = \frac{1}{P-1}(\sum_{n=1,n\neq m}^{P}\mathbf{Attn}_2^{m-n}), \tag{7}$$

3.3 VELOCITY FIELD DECODER

We use graph convolutional network to embed the input velocity field to obtain latent representation $\mathbf{S}_l$=$GCN(\mathbf{S})$. In section. 3.1 and 3.2, we obtain the acceleration evolution attention and interaction relation attention, so,we can calculate the acceleration field enhanced velocity prediction $\mathbf{S}_a^m = \mathbf{Attn}_{af}^m \mathbf{S}_l^m$ and relation field enhanced velocity prediction $\mathbf{S}_i^m = \mathbf{Attn}_{if}^m \mathbf{S}_l^m$.

In order to better integrate with the two enhanced predictions, we obtain context dependencies of the original latent representation through self attention layer:

$$\mathbf{Q}_3^m = \mathbf{W}_3^Q \mathbf{S}_l^m, \mathbf{K}_3^m = \mathbf{W}_3^K \mathbf{S}_l^m, \mathbf{V}_3^m = \mathbf{W}_3^V \mathbf{S}_l^m, \mathbf{S}_o^m = \text{softmax}(\mathbf{Q}_3^m {\mathbf{K}_3^m}^\top / \sqrt{d_{K_3}}) \mathbf{V}_3^m, \quad (8)$$

where $\mathbf{S}_o^m$ denotes the self enhanced prediction for the m-th person. Then, we use parallel cross attention layer to get the latent representation of preliminary velocity prediction $\mathbf{S}_{pre}^m$ by fusing the acceleration evolution attention and interaction relation attention :

$$\begin{aligned} \mathbf{Q}_4^m = \mathbf{W}_4^Q \mathbf{S}_o^m, \mathbf{K}_4^m = \mathbf{W}_4^K \mathbf{S}_i^m, \mathbf{V}_4^m = \mathbf{W}_4^V \mathbf{S}_i^m, \mathbf{S}_{io}^m = \text{softmax}(\mathbf{Q}_4^m {\mathbf{K}_4^m}^\top / \sqrt{d_{K_4}}) \mathbf{V}_4^m, \\ \mathbf{Q}_5^m = \mathbf{W}_5^Q \mathbf{S}_o^m, \mathbf{K}_5^m = \mathbf{W}_5^K \mathbf{S}_a^m, \mathbf{V}_5^m = \mathbf{W}_5^V \mathbf{S}_a^m, \mathbf{S}_{ao}^m = \text{softmax}(\mathbf{Q}_5^m {\mathbf{K}_5^m}^\top / \sqrt{d_{K_5}}) \mathbf{V}_5^m, \\ \mathbf{S}_{pre}^m = \frac{1}{2}(\mathbf{S}_{io}^{mm} + \mathbf{S}_{ao}^m), \end{aligned} \quad (9)$$

Ultimately, we use graph convolutional network to complete the decoding of multi-person motion velocity field from latent representation to predicted the preliminary predicted velocity value $\widetilde{\mathbf{S}} = GCN(\mathbf{S}_{pre})$.

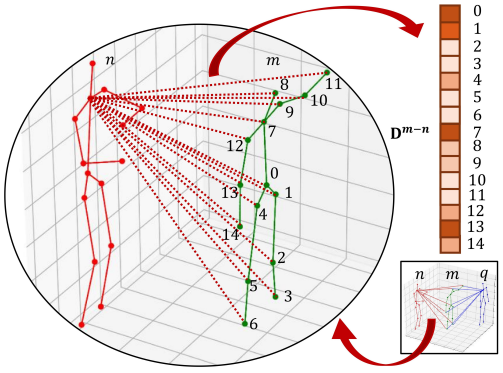


Figure 3: The interaction relation field illustration. The Interaction relation field of the m-th person is the distance between all the joint nodes of the m-th person and the chest nodes of other people. The darker the color, the corresponding nodes' distance with other's chest is closer.

3.4 JACOBIAN ITERATIVE SOLVER

The preliminary velocity $\widetilde{\mathbf{S}}^m$ produced by VFD may violate skeletal constraints after temporal integration. We therefore introduce a Jacobian iterative solver (JIS) that uses this prediction as the frame-wise driving signal and jointly corrects all constrained skeletal edges. Here, $\widetilde{(\cdot)}$, $\overline{(\cdot)}$, and $\widehat{(\cdot)}$ denote the preliminary prediction, intermediate solver state, and final corrected prediction, respectively.

Starting from the last observed pose $(\widehat{\mathbf{X}}^m)_{50} = (\mathbf{X}_{\text{gt}}^m)_{50}$, the initial state at each future time t is obtained by

$$(\overline{\mathbf{X}}^{m,(0)})_t = (\widehat{\mathbf{X}}^m)_{t-1} + \Delta t\, (\widetilde{\mathbf{S}}^m)_t, \quad (10)$$

which explicitly preserves the motion trend predicted by the network. For a constrained skeletal edge $e = (p_e, c_e) \in \mathcal{E}$, its reference length is extracted from the last observed pose:

$$\ell_e^m = \left\| \left[(\mathbf{X}_{\text{gt}}^m)_{50}\right]_{c_e} - \left[(\mathbf{X}_{\text{gt}}^m)_{50}\right]_{p_e} \right\|_2. \quad (11)$$

At iteration r, the residual of edge e is

$$(g_e^{m,(r)})_t = \frac{1}{2}\left(\left\| \left[(\overline{\mathbf{X}}^{m,(r)})_t\right]_{c_e} - \left[(\overline{\mathbf{X}}^{m,(r)})_t\right]_{p_e} \right\|_2^2 - (\ell_e^m)^2 \right). \quad (12)$$

All edge residuals form $(\mathbf{g}^{m,(r)})_t$. With the signed incidence matrix $\mathbf{B}$, the corresponding sparse Jacobian is analytically computed as

$$\left[(\mathbf{J}^{m,(r)})_t\right]_{e,j} = B_{e,j} \left(\left[(\overline{\mathbf{X}}^{m,(r)})_t\right]_{c_e} - \left[(\overline{\mathbf{X}}^{m,(r)})_t\right]_{p_e} \right)^\top. \quad (13)$$

Table 1: Comparisons of performance with state-of-the-art on the **JPE** metric. The lower values indicate better performance. The best results in each setting are **bold**.

Methods	Venue	CMU-Mocap				Mix1				Mix2			
		0.2s	0.6s	1.0s	Avg	0.2s	0.6s	1.0s	Avg	0.2s	0.6s	1.0s	Avg
HRIMao et al. (2020)	*ECCV'20*	49	130	207	129	51	141	233	142	52	140	224	139
MSR-GCNDang et al. (2021)	*CVPR'21*	53	146	231	143	49	132	220	134	60	153	243	152
MRTWang et al. (2021)	*NeuIPS'21*	36	115	193	114	37	122	212	124	38	126	214	126
TBIFormerPeng et al. (2023)	*CVPR'23*	30	109	182	107	34	121	209	121	34	118	198	117
JRFormerXu et al. (2023)	*ICCV'23*	32	104	161	99	32	109	184	108	36	125	211	124
IAFormerXiao et al. (2024)	*ECCV'24*	32	96	159	96	36	112	193	114	36	108	181	108
ST-MoEYin et al. (2026)	*AAAI'26*	31	95	158	95	34	108	187	110	34	106	179	107
KiJacoMMP	Ours	**26**	**90**	**153**	**90**	**29**	**98**	**175**	**101**	**28**	**96**	**164**	**96**

The edge multiplier $(\boldsymbol{\mu}^{m,(r)})_t$ is obtained by solving the damped linearized system

$$\Big[(\mathbf{J}^{m,(r)})_t \mathbf{D} (\mathbf{J}^{m,(r)})_t^\top + \epsilon \mathbf{I}\Big] (\boldsymbol{\mu}^{m,(r)})_t = (\mathbf{g}^{m,(r)})_t, \tag{14}$$

where $\mathbf{D}$ denotes joint correction mobility and $\epsilon > 0$ provides numerical damping. The root mobility is set to zero to preserve its predicted translation. Using $\mathrm{vec}(\cdot)$ to denote joint coordinate vectorization, all joints are updated simultaneously by

$$\mathrm{vec}\Big((\overline{\mathbf{X}}^{m,(r+1)})_t\Big) = \mathrm{vec}\Big((\overline{\mathbf{X}}^{m,(r)})_t\Big) - \mathbf{D}(\mathbf{J}^{m,(r)})_t^\top (\boldsymbol{\mu}^{m,(r)})_t. \tag{15}$$

The residual and Jacobian are recomputed after every update. After R iterations, the final position and its kinematically consistent velocity are

$$(\widehat{\mathbf{X}}^m)_t = (\overline{\mathbf{X}}^{m,(R)})_t, \qquad (\widehat{\mathbf{S}}^m)_t = \frac{(\widehat{\mathbf{X}}^m)_t - (\widehat{\mathbf{X}}^m)_{t-1}}{\Delta t}. \tag{16}$$

The corrected state is then propagated to the next frame. Consequently, JIS retains the preliminary velocity field as the driving motion while enforcing coupled skeletal consistency through a minimum correction. The corresponding detailed derivation is provided in the supplementary material.

3.5 Training Strategy

In order to constrain the predicted value of velocity, inspired by the loss function proposed by IAFormer Xiao et al. (2024), we calculate the velocity loss function of joint velocity error and time consistency between the velocity prediction $\widehat{\mathbf{S}}$ and the velocity ground truth $\mathbf{S}_{gt}$:

$$\mathcal{L}_v = MSE(\widehat{\mathbf{S}}, \mathbf{S}_{gt}) + MSE(conv(\widehat{\mathbf{S}}), conv(\mathbf{S}_{gt})), \tag{17}$$

We also calculate the coordinates loss function of joint position error and time consistency between the coordinates prediction $\widehat{\mathbf{X}}$ and the coordinates ground truth $\mathbf{X}_{gt}$:

$$\mathcal{L}_p = MSE(\widehat{\mathbf{X}}, \mathbf{X}_{gt}) + MSE(conv(\widehat{\mathbf{X}}), conv(\mathbf{X}_{gt})). \tag{18}$$

Finally, the whole model will be optimized by $\mathcal{L} = \mathcal{L}_v + \mathcal{L}_p$.

4 Experiments

4.1 Datasets

CMU-Mocap CMU is a mixed three-person motion dataset. To ensure fair and consistent comparisons with prior work, we adopt the same data setting as in IAFormer. The training set contains 13,000 sequences, the test set contains 3,000 sequences, and the input sequence length is set to 50 frames, and the prediction horizon is set to 25 frames.

Table 2: Comparisons of performance with state-of-the-art on the **APE** metric. The lower values indicate better performance. The best results in each setting are **bold**.

Methods	Venue	CMU-Mocap				Mix1				Mix2			
		0.2s	0.6s	1.0s	Avg	0.2s	0.6s	1.0s	Avg	0.2s	0.6s	1.0s	Avg
HRIMao et al. (2020)	*ECCV'20*	41	97	130	89	38	92	122	84	41	100	133	91
MSR-GCNDang et al. (2021)	*CVPR'21*	46	106	137	96	41	92	120	84	48	110	148	102
MRTWang et al. (2021)	*NeuIPS'21*	36	108	159	101	36	109	166	104	38	115	178	110
TBIFormerPeng et al. (2023)	*CVPR'23*	27	84	118	76	28	81	113	74	30	89	124	81
JRFormerXu et al. (2023)	*ICCV'23*	20	78	114	71	21	73	105	66	22	82	120	75
IAFormerXiao et al. (2024)	*ECCV'24*	23	71	103	66	23	71	101	65	24	76	108	69
ST-MoEYin et al. (2026)	*AAAI'26*	22	70	104	65	22	69	100	64	23	75	109	69
KiJacoMMP	Ours	**19**	**68**	**101**	**62**	**20**	**65**	**95**	**60**	**20**	**71**	**104**	**65**

Mix1 and Mix2 are two crowded multi-person datasets involving a larger number of individuals. These two datasets are constructed by combining MuPoTS-3D Cheng et al. (2022), 3DPW von Marcard et al. (2018). and are test from CMU-Mocap pretrain model. Specifically, Mix1 contains 6 individuals and Mix2 contains 10 individuals, including cases in which some individuals exhibit weak or no explicit interactions with others. Each of these two mixed datasets contains 1,000 motion sequences, and each sequence is composed of 75 frames.

4.2 Metrics

Following IAFormer Xiao et al. (2024), we adopt the mean per Joint Position Error (JPE) and Aligned Position Error (APE) as evaluation metrics. The calculation formula of JPE and APE are as follows:

$$JPE = \frac{1}{P \cdot J} \sum_{p=1}^{P} \sum_{j=1}^{J} ||(\widehat{\mathbf{X}}^p)_j - (\mathbf{X}^p_{gt})_j||^2, \quad APE = JPE(\widehat{\mathbf{X}} - \widehat{\mathbf{X}}_r, \mathbf{X} - \mathbf{X}_r), \tag{19}$$

where $\widehat{\mathbf{X}}_r$ and $\mathbf{X}_r$ are predicted and ground truth of reference human body. For these metrics, lower values indicate better performance.

4.3 Implementation Details

We implement our framework in PyTorch, and the experiments are performed on Nvidia RTX 4090. We train our model for 100 epochs using the Adam optimizer with a batch size of 96 and a dropout of 0.1. In the experiment with CMU-Mocap, CHI3D, Mix1 and Mix2, the historical sequence is 50 frames, the predicted sequence is 25 frames. Each person in above datasets has 15 joint nodes.

4.4 Main Results

Results on CMU-Mocap. Tables. 1 and 2 compare KiJacoMMP with existing methods. KiJacoMMP achieves the lowest JPE at all horizons, obtaining 26, 90, and 153 at 0.2 s, 0.6 s, and 1.0 s, respectively. Compared with the recent ST-MoE, the error is consistently reduced by 5 mm at every horizon, while the average JPE decreases from 95 to 90. It also outperforms IAFormer by 6, 6, and 6 mm. For APE, KiJacoMMP achieves 19, 68, and 101, reducing the average error from 65 to 62 compared with ST-MoE. The consistent gains in both JPE and APE indicate that KiJacoMMP improves global motion evolution while preserving accurate relative body configurations, rather than relying only on a better global trajectory estimate.

Results on Mix1 and Mix2. The advantage becomes more evident in crowded scenes. On Mix1, KiJacoMMP achieves an average JPE of 101, compared with 110 for ST-MoE and 114 for IAFormer. Its long-term JPE at 1.0 s is reduced from 187 and 193 to 175. Meanwhile, the average APE decreases to 60, outperforming ST-MoE and IAFormer by 4 and 5 mm, respectively. On Mix2, which contains ten people and more complex relation patterns, KiJacoMMP further reduces the average

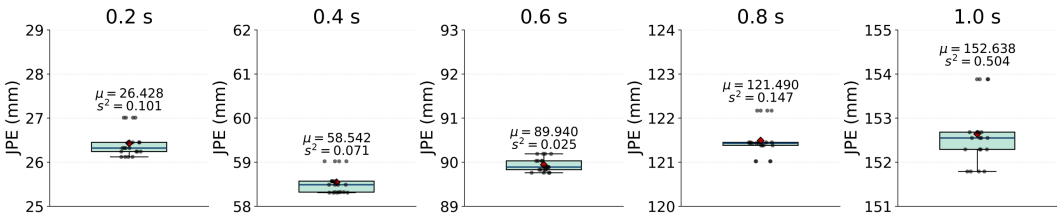


Figure 4: The statistical test of KIJacoMMP on the CMU-Mocap dataset. The experiment was repeated 20 times, and the variance and mean of the experimental results at each time point were labeled above the box plot.

Table 3: Ablation study on different modules in KiJacoMMP. We use JPE metrics on CMU-Mocap datasets with different settings to evaluate the effectiveness of the modules. The best results in each setting are **bold**.

Methods	Modules				0.2s	0.6s	1.0s
	VFD	AFE	IRE	JIS			
Base	✓				29	101	162
w/o AFE	✓		✓	✓	27	95	158
w/o IRE	✓	✓		✓	27	94	157
w/o JIS	✓	✓	✓		28	97	159
KiJacoMMP	✓	✓	✓	✓	**26**	**90**	**153**

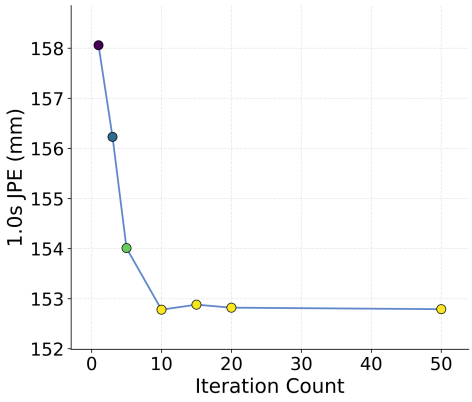


Figure 5: Comparison of different iteration count in JIS.

JPE from 107 to 96 relative to ST-MoE, with improvements of 6, 10, and 15 mm over the three horizons. The corresponding average APE is reduced from 69 to 65. Notably, the long-term JPE gain over ST-MoE increases from 5 mm on CMU-Mocap to 12 mm on Mix1 and 15 mm on Mix2. This trend suggests that explicitly combining inter-person relations with individual kinematic evolution is particularly beneficial when prediction errors and interaction ambiguity accumulate over time.

4.5 Statistical Test

To examine whether the reported performance is stable across repeated optimization runs, we repeat the CMU-Mocap experiment 20 times and summarize the resulting JPE distributions in Figure. 4. The mean errors at 0.2 s, 0.4 s, 0.6 s, 0.8 s, and 1.0 s are 26.428, 58.542, 89.940, 121.490, and 152.638 mm, respectively, which are consistent with the rounded results reported in Table 1. The corresponding sample variances are only 0.101, 0.071, 0.025, 0.147, and 0.504. Although the long-term errors are more sensitive to accumulated uncertainty, its standard deviation is still only 0.710 mm, corresponding to less than 0.5% of the mean JPE. These observations confirm that the accuracy improvements of KiJacoMMP are accompanied by stable optimization.

5 Ablation Study

5.1 Effectiveness of Different Modules

Table 3 shows that the VFD-only Base model obtains JPEs of 29, 101, and 162 at 0.2 s, 0.6 s, and 1.0 s, whereas the complete KiJacoMMP reduces them to 26, 90, and 153. Removing AFE increases the errors to 27, 95, and 158, confirming that acceleration evolution helps VFD estimate how the future velocity field changes, particularly at medium and long horizons. Without IRE, the errors increase to 27, 94, and 157, demonstrating the importance of multi-person interaction cues for resolving individual motion evolution. Removing JIS produces JPEs of 28, 97, and 159, representing the largest overall degradation. Unlike AFE and IRE, which improve the preliminary velocity field,

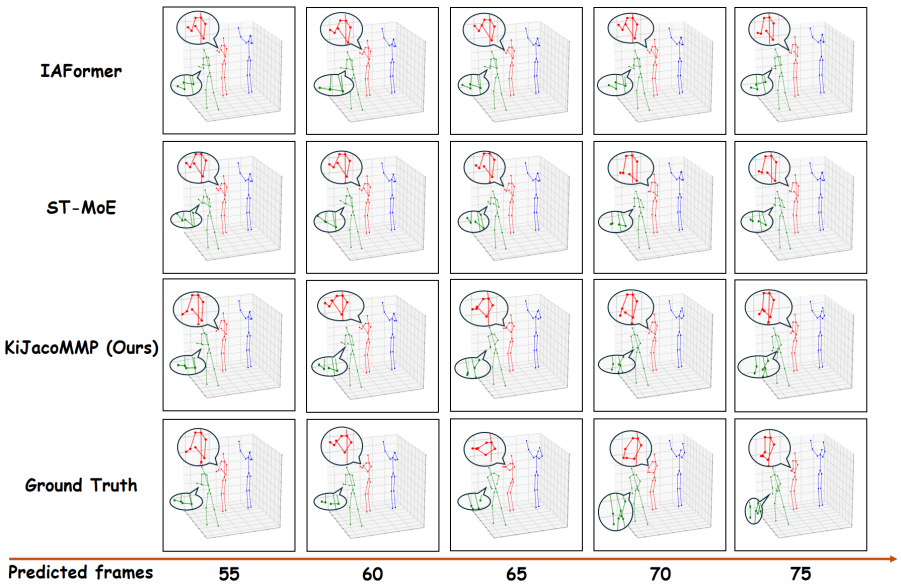


Figure 6: The comparison visualization results of KiJacoMMP with IAFormer and ST-MoE on the CMU-Mocap dataset. The five columns show the forecast results at different frames.

JIS jointly corrects shared joints during frame-wise integration and prevents skeletal inconsistencies from accumulating over time. These results demonstrate that AFE captures temporal motion trends, IRE models inter-person dependencies, and JIS converts the preliminary prediction into a kinematically consistent trajectory; combining the three components yields the strongest performance.

5.2 Effectiveness of Different Iteration Counts in the Solver

Figure. 5 evaluates the influence of the JIS iteration count. Increasing the number of iterations from 1 to 5 rapidly reduces the error from approximately 158.1 to 154.0 mm, indicating that the preliminary prediction initially contains substantial constraint residuals that can be removed through joint Jacobian correction. Increasing the iteration count to 10 further reduces the error to approximately 152.8 mm. Beyond ten iterations, the results remain nearly unchanged: the variation between 10 and 50 iterations is below 0.1 mm. This saturation indicates that the solver reaches a near-consistent skeletal configuration within a small number of updates, after which additional projections provide negligible correction.

6 Visualization

Figure. 6 represents the comparison visualization results of KiJacoMMP with IAFormer and ST-MoE on the CMU-Mocap dataset. From the ground truth, we can see that the hands of the green and red individuals gradually change from folded up to vertical down. All other methods show a folded up stance, with only the red individual's two hands in KiJacoMMP showing a downward pose at the last, which is the closest method to ground truth in the red part.

7 Conclusion

This paper presents KiJacoMMP, a kinematics-aware jacobian iterative solver for multi-person motion prediction. The acceleration field encoder models velocity evolution, the interaction relation encoder captures inter-person dependencies, and the velocity field decoder produces preliminary future velocities. Starting from the last observed pose, these velocities are integrated frame by frame, after which a Jacobian iterative solver jointly corrects shared joints through a damped linearized system derived from fixed skeletal edges. Without requiring mass, muscle forces, or contact forces, KiJacoMMP preserves interaction trends, temporal continuity, and kinematic consistency. Experiments verify the complementary roles of its components and the stable convergence of the solver.

REFERENCES

Cmu graphics lab motion capture database. In *http://mocap.cs.cmu.edu/*.

Mykhaylo Andriluka, Baruch Tabanpour, C. Daniel Freeman, and Cristian Sminchisescu. Learned neural physics simulation for nbsp;articulated 3d human pose reconstruction. In *ECCV*, 2024.

Xiaowei Chen, Xiao Jiang, Lishuang Zhan, Shihui Guo, Qunsheng Ruan, Guoliang Luo, Minghong Liao, and Yipeng Qin. Full-body human motion reconstruction with sparse joint tracking using flexible sensors. *ACM TOMM*, 2023.

Yu Cheng, Bo Wang, and Robby Tan. Dual networks based 3d multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Cecilia Curreli, Dominik Muhle, Abhishek Saroha, Zhenzhang Ye, Riccardo Marin, and Daniel Cremers. Nonisotropic gaussian diffusion for realistic 3d human motion prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1871–1882. IEEE, 2025.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *ICCV*, 2021.

Prasenjit Ghorai, Azim Eskandarian, Young-Keun Kim, and Goodarz Mehr. State estimation and motion prediction of vehicles and vulnerable road users for cooperative autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

Airlangga Adi Hermawan, Yakub Fahim Luckyarno, Isfan Fauzi, Derek Kwaku Pobi Asiedu, Tae-Wook Kim, Deok-Young Jung, Jin Sam Kwak, and Ji-Hoon Yun. Velocity- and error-aware switching of motion prediction models for cloud virtual reality. *IEEE Access*, 2023.

Xuemin Hu, Bo Tang, Long Chen, Sheng Song, and Xiuchi Tong. Learning a deep cascaded neural network for multiple motion commands prediction in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *CVPR*, 2024.

Jinwoo Kim, Heeseok Oh, Woojae Kim, Seonghwa Choi, Wookho Son, and Sanghoon Lee. A deep motion sickness predictor induced by visual stimuli in virtual reality. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Zhuo Li, Mingshuang Luo, Ruibing Hou, Xin Zhao, Hao Liu, Hong Chang, Zimo Liu, and Chen Li. Morph: A motion-free physics optimization framework for human motion generation. *ICCV*, 2025.

Haichao Liu, Kai Chen, Yulin Li, Zhenmin Huang, Ming Liu, and Jun Ma. Udmc: Unified decision-making and control framework for urban autonomous driving with motion prediction of traffic participants. *IEEE Transactions on Intelligent Transportation Systems*, 2025a.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *ECCV*, 2024.

Sheng Liu, Shaobo Zhang, Fei Gao, and Yuan Feng. Highly condensed all-mlp architecture for long-term human motion prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

Ye Lu, Jie Wang, Tianyi Liu, Jianjun Gao, and Kim-Hui Yap. Kinemadiff: Towards diffusion for coherent and physically plausible human motion prediction. In *International Conference on Learning Representations*, volume 2026, pp. 40460–40479, 2026.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *ECCV*, 2020.

Mukund Mitra, Gyanig Kumar, Partha Chakrabarti, and Pradipta Biswas. Investigating inverse reinforcement learning during rapid aiming movement in extended reality and human-robot interaction. *ACM THRI*, 2025.

Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *CVPR*, 2023.

Hyocheol Ro, Yoon Jung Park, Jung-Hyun Byun, and Tack-Don Han. Display methods of projection augmented reality based on deep learning pose estimation. In *ACM SIGGRAPH*, 2019.

Jianwei Tang, Hong Yang, Tengyue Chen, and Jian-Fang Hu. Stochastic human motion prediction with memory of action transition and action characteristic. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. SoMoFormer: Multi-Person Pose Forecasting with Transformers. *arXiv e-prints*, 2022.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018.

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *NeurIPS*, 34, 2021.

Xiu-Shen Wei, Yu-Yan Xu, Chen-Lin Zhang, Gui-Song Xia, and Yu-Xin Peng. Cat: a coarse-to-fine attention tree for semantic change detection. *Visual Intelligence*, 2023.

Peng Xiao, Yi Xie, Xuemiao Xu, Weihong Chen, and Huaidong Zhang. Multi-person pose forecasting with individual interaction perceptron and prior learning. In *The 18th European Conference on Computer Vision (ECCV)*, pp. 402–419, 2024.

Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. Joint-relation transformer for multi-person motion prediction. In *ICCV*, 2023.

Wenrui Yang, Xinchun Yu, and Xiao-Ping Zhang. Glst-gcn: Global-local spatio-temporal graph convolutional network for skeleton-based hand motion prediction. In *ICASSP*, 2025.

Zheng Yin, Chengjian Li, Xiangbo Shu, Meiqi Cao, Rui Yan, and Jinhui Tang. Spatiotemporal-untrammelled mixture of experts for multi-person motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12044–12052, 2026.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023.

Jiawei Zhang, Haonan Zhang, Weitao Zhang, Liang Pu, Zesen Feng, and Jie Guo. Decoupled motion prediction for real-time g-buffer free frame extrapolation. In *ACM MM*, 2025.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. Physdreamer: Physics-based interaction with nbsp;3d objects via nbsp;video generation. In *ECCV*, 2024a.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *CVPR*, 2024b.

Yufei Zhang, Jeffrey O. Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *WACV*, 2024c.

Zhibo Zhang, Yanjun Zhu, Rahul Rai, and David Doermann. Pimnet: Physics-infused neural network for human motion prediction. *IEEE Robotics and Automation Letters*, 2022.

Yuanhong Zheng, Ruixuan Yu, and Jian Sun. Efficient multi-person motion prediction by lightweight spatial and temporal interactions. *ICCV*, 2025.