Efficient Multi-Person Motion Prediction by Lightweight Spatial and Temporal Interactions

Supplementary Material

A. Additional Details on Dataset and Metric

A.1. Dataset Sources

All datasets utilized in this study are sourced from publicly available open-source repositories, including the CMU-Mocap, MuPoTS-3D, and 3DPW datasets. For the 3DPW dataset, we adhere to the protocol outlined in the SoMoF Benchmark [1, 2], which has been used in previous studies [5, 7]. For the 3DPW-RC dataset, we apply the same scripts as in [7] to remove camera movement, thereby enhancing the realism of human motion. For the CMU-Mocap and MuPoTS-3D datasets, we synthesize data based on the original datasets, following approach in [6]. Since synthesized datasets from prior works are not publicly available, all experiments were conducted on our synthesized versions by running official codes of compared methods.

A.2. Metric Formulations

Given the predicted motion $Y = \{y_t^{p,j}\} \in \mathbb{R}^{3J\times P\times T'}$ for P persons across T' time frames with J joints per person, along with the corresponding ground truth $\hat{Y} = \{\hat{y}_t^{p,j}\} \in \mathbb{R}^{3J\times P\times T'}$, the following metrics are used for evaluation.

MPJPE. The Mean Per Joint Position Error (MPJPE) measures the overall joint prediction accuracy by averaging errors across all time frames:

$$\textbf{MPJPE} = \frac{1}{P\cdot T'\cdot J}\sum_{p=1}^{P}\sum_{t=1}^{T'}\sum_{j=1}^{J}||y_t^{p,j} - \hat{y}_t^{p,j}||_2. \quad \text{(A.1)}$$

VIM. The Visibility-Ignored Metric (VIM) focuses on the average joint error for a specific time frame, and the VIM score at timestep t is given by:

$$\textbf{VIM@t} = \frac{1}{P}\sum_{p=1}^{P}\sqrt{\sum_{j=1}^{J}(y_t^{p,j} - \hat{y}_t^{p,j})^2}. \quad \text{(A.2)}$$

JPE. The Joint Precision Error (JPE) assesses both global and local joint predictions using the mean L_2 distance of all joints for timestep t:

$$\textbf{JPE@t} = \frac{1}{P\cdot J}\sum_{p=1}^{P}\sum_{j=1}^{J}||\mathbf{y}_t^{p,j} - \hat{\mathbf{y}}_t^{p,j}||_2. \quad \text{(A.3)}$$

APE. The Aligned Mean Per Joint Position Error (APE) evaluates the forecasted local motion by computing the L_2 distance for each joint, averaged across all joints at a given timestep t, with global displacement removed by subtracting the hip joint:

$$\textbf{APE@t} = \frac{1}{P\cdot J}\sum_{p=1}^{P}\sum_{j=1}^{J}||(y_t^{p,j} - y_{t,\textbf{hip}}^{p}) - (\hat{y}_t^{p,j} - \hat{y}_{t,\textbf{hip}}^{p})||_2. \quad \text{(A.4)}$$

FDE. The Final Distance Error (FDE) quantifies the accuracy of the forecasted global trajectory by computing the L_2 distance for a specific timestep t:

$$\textbf{FDE@t} = \frac{1}{P\cdot J}\sum_{p=1}^{P}\sum_{j=1}^{J}||y_{t,\textbf{hip}}^{p} - \hat{y}_{t,\textbf{hip}}^{p}||_2. \quad \text{(A.5)}$$

These metrics provide a comprehensive evaluation of the accuracy for 3D motion prediction task, capturing both local joint-wise pose errors and global trajectory deviations.

B. Additional Network and Training Details

Details on PIPS and IPIPS Stages. In order to maintain the invariance of our model to the order of individuals, we introduce Permutation-Invariant Person Sorting (PIPS). Given the input $X \in \mathbb{R}^{3J\times P\times T}$, we calculate the sum of distances between each individual and all others as:

$$d_{p_j} = \sum_{p_k\neq p_j}^{P}\|x_{1,\textbf{hip}}^{p_j} - x_{1,\textbf{hip}}^{p_k}\|_2, \quad j,k = 1,\ldots,P. \quad \text{(B.1)}$$

We then sort individuals in descending order based on these values, modifying the input X. After processing the input through the model to obtain output Y', we apply Inverse PIPS to recover the original person order, resulting in the final output $Y \in \mathbb{R}^{3J\times P\times T'}$. These two stages ensures that the output of our model is unchanged for different orders of individuals in the same scene as the input.

Network for Pre-training. For the experiments in the settings of *AMASS/3DPW-Ori* and *AMASS/3DPW-RC*, the network is pre-trained on the AMASS dataset and fine-tuned on the 3DPW dataset. Given that the number of parameters has a significant impact on network performance, especially for lightweight architectures, we utilize a model with 0.65M parameters to increase its capacity. This is achieved by incorporating additional spatial feature updates, as illustrated in Fig. B.1. Specifically, we introduce Local/Global Spatial Feature Update, which extend the Local/Global Temporal Feature Update. These new components maintain a similar architectural structure to the original components but operate along the spatial dimension rather than the temporal

dimension. As demonstrated in Tab. 1 of the main paper, our model with 0.65M parameters achieves the best performance for the *AMASS/3DPW-RC* setting and competitive results for the *AMASS/3DPW-Ori* setting.

Details for Pre-training Settings. For the experiments in the settings of *AMASS/3DPW-Ori* and *AMASS/3DPW-RC*, we pre-train our network on AMASS dataset for 100 epochs with an initial learning rate of 1×10^{-4}, which decays by a factor of 0.8 every 10 epochs. The strategy employed for network fine-tuning is the same as described in Sect. 3.5 of the main paper.

Details on Attention-based Designs. In Sect. 4.3 of the main paper, we compared our architecture with Attention-based alternatives by replacing our ME block and CI block with two distinct Attention-based blocks: one employing Multi-Head Attention (Self-Attention) for local/global temporal feature update and the other utilizing Cross Attention for local/glocal refinement. The detailed architectures are illustrated in Fig. B.2. Our results in Tab. 5 of the main paper demonstrate that our proposed model outperforms traditional Attention-based architectures, which have more parameters and require greater computational resources.

C. More Experimental Results

In addition to the results evaluated using MEJPE, VIM, and APE presented in the main paper, we provide additional results based on the metrics of JPE and FDE in this section. We also include the results across different key frames for VIM, JPE, APE, and FDE, along with an analysis of the computational cost.

Results Evaluated by JPE and FDE. In Tab. C.3, we present the results evaluated using the JPE and FDE metrics across all the settings discussed in the main paper, along with comparisons to previous works. Our EMPMP model achieves the best JPE results, demonstrating significant superiority over the other methods. It also performs exceptionally well in FDE, ranking first in seven out of eight evaluations. Notably, the T2P model [3] predicts multiple trajectories (F) and for a fair comparison, we compute the FDE with $F = 1$, while using $F = 3$ for the other metrics.

Detailed Results Across Key Frames. In the main paper, for all models evaluated using VIM, JPE, APE, and FDE, we report their average results across key frames. Additionally, we provide the detailed results for each individual frame, following the frame selection scheme shown in Tab. C.1. For the VIM metric on the 3DPW dataset, we adopt the same frame selection scheme as used in [5, 7] to ensure consistency and fair comparisons. For other datasets and metrics, frames are selected at reasonable intervals to ensure comprehensive evaluation. Tab. C.4 presents the detailed VIM results across multiple settings of the 3DPW dataset, while Tab. C.5 reports the corresponding JPE, APE, and FDE results for the same dataset. Similarly, Tab. C.6 shows the VIM results for various settings of the CMU-Syn and MuPoTS-3D datasets, and Tab. C.7 provides the JPE, APE, and FDE results for these datasets. These tables provide detailed results for specific key frames and their average values, where our model achieves dominant superior performance in most comparisons, proving its effectiveness through a comprehensive evaluation of performance metrics across different timesteps.

Datasets	3DPW	CMU-Syn & MuPoTS	
Out Length	14frames (900ms)	30frames (2s)	15frames (1s)
VIM	2, 4, 8, 10, 14	2, 6, 11, 21, 30	2, 4, 8, 10, 15
JPE&APE&FDE	7, 14	10, 20, 30	3, 9, 15

Table C.1. Frame selection scheme for different datasets.

Comparison on Computational Cost. In Tab. C.2, we present the detailed computational costs of our EMPMP model and the compared methods, including GPU memory usage, computational FLOPs, and the number of parameters. Our model demonstrates superior performance across various settings while maintaining a significantly lower number of parameters and FLOPs.

Metrics	Memory (MB)	FLOPs (G)	Params (M)
MRT [6] '2021	**2281**	27.55	6.29
SoMoFormer [5] '2022	6308	113.37	12.91
TBIFormer [4] '2023	2826	15.64	7.31
JRT [7] '2023	15544	767.74	3.68
T2P [3] '2024	4304	51.67	4.60
Ours	2674	**1.67**	**0.17**

Table C.2. Comparisons of computational cost for different models in the *CMU-Syn* (2s/2s) setting.

D. Additional Ablation Study

Effectiveness of Learned Affine Transformations. In the CI block, we incorporate both scale and translation for local representation refinement, while employing translation alone for global representation refinement. To assess the effectiveness of this design, we conducted an ablation study exploring various transformation choices for local/global representation refinement. As presented in Tab. D.1, the model that utilizes scale and translation for local refinement, and translation for global refinement, yields the best performance compared to alternative configurations.

Ablation Study on Loss Function. Our loss function comprises two components: mean joint loss and velocity loss. As demonstrated in Tab. D.2, the network that excludes velocity loss performs worse in the *3DPW-Ori*, *3DPW-RC*, and *CMU-Syn* settings, thereby highlighting the necessity of incorporating velocity loss.

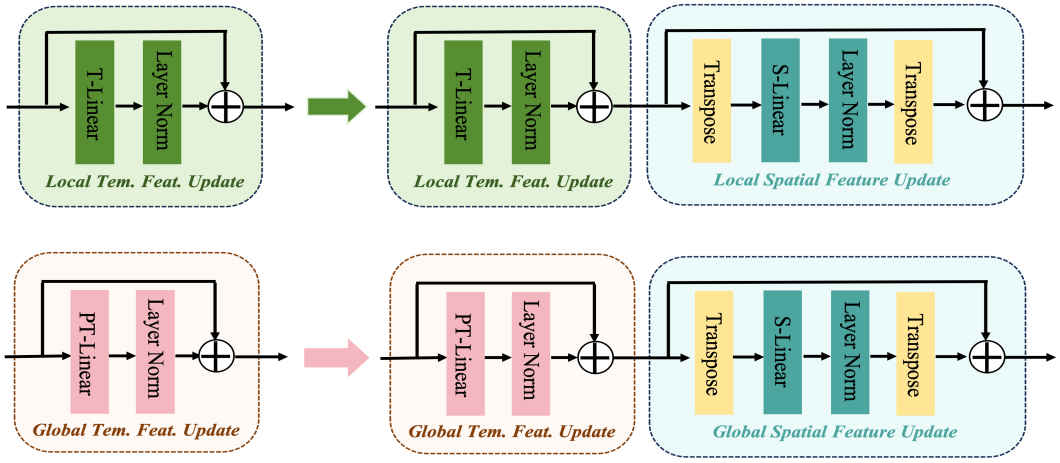

Figure B.1. Architectural modifications for our 0.65M-parameter EMPMP involve enhancing the Local/Global Temporal Feature Update (left) with the addition of the Local/Global Spatial Feature Update (right). These modifications update features along the spatial dimension, thereby increasing the model's expressive capacity.

Local Refinement		Global Refinement		MPJPE
Scale	**Translation**	**Scale**	**Translation**	
×	✓	✓	✓	129.9
✓	×	✓	✓	129.4
✓	✓	×	✓	**128.0**
✓	✓	✓	×	131.3
✓	✓	✓	✓	129.0

Table D.1. Ablation study on learned affine transformations for local and global representation refinement.

Settings	3DPW-Ori	3DPW-RC	CMU-Syn	
In/Out Length	1030ms/900ms		2s/2s	1s/1s
EMPMP-w/o-$\mathcal{L}^{Vel}$	137.7	103.7	130.0	74.7
EMPMP	**131.8**	**99.5**	**128.0**	**73.5**

Table D.2. Ablation study on velocity loss in *3DPW-Ori*, *3DPW-RC* and *CMU-Syn* settings, evaluated using MPJPE.

Ablation Study on Combination Parameter. The combination parameter α in Eq. (11) of the main paper controls the balance between global and local representations. We present the results of our EMPMP model with varying values of α in the *3DPW-RC* setting. As shown in Tab. D.4, the optimal results for MPJPE, VIM, and APE were achieved with $\alpha = 0.2$.

Ablation Study with Fixed-Parameter. To further validate the effectiveness of our architectural design choices, we conducted a comprehensive ablation study with fixed parameter count (0.17M parameters) by systematically varying the number of ME blocks (N), CI blocks (M), and layers (K) in our EMPMP model. As in Tab. D.3, this approach ensures that performance improvements are attributed to architectural design rather than increased model capacity.

LTFU	GTFU	LRwGR	GRwLR	DE	MPJPE	N/M/K
×	✓	×	×	×	150.0	0/5/4
✓	×	×	×	×	134.9	40/0/4
✓	✓	×	×	×	138.8	25/2/4
✓	✓	×	✓	×	135.6	16/1/5
✓	✓	✓	×	×	133.1	18/1/5
✓	✓	✓	✓	×	129.2	18/1/4
✓	✓	✓	✓	✓	**128.0**	16/1/4

Table D.3. Ablation study with fixed-parameter.

α	0	0.1	0.2	0.3	0.4	0.6	0.8	1.0
MPJPE	100.6	102.9	**99.5**	99.8	102.1	100.3	100.4	101.0
VIM	41.7	42.7	**41.3**	41.5	42.3	41.5	41.8	42.6
APE	97.3	99.2	**96.6**	99.4	96.9	97.0	97.4	97.9

Table D.4. Ablation study of the combination parameter α in the *3DPW-RC* setting.

E. Experiments on Larger Group.

As we can not access the dataset with 9-15 interacting persons in MRT [6], we follow TBIFormer [4] to synthesize a 10-person dataset (CMU-Mixed10) on CMU-Mocap. As in following table, our method maintains SOTA in large-group settings.

Metric	MPJPE			VIM			APE			FDE		
Frames	10	20	30	10	20	30	10	20	30	10	20	30
TBIF [4]	71	120	160	57	89	110	79	98	104	91	174	238
T2P [3]	68	117	157	56	88	110	76	96	104	86	167	222
Ours	**63**	**108**	**145**	**52**	**81**	**102**	**70**	**89**	**98**	**78**	**157**	**217**

Table E.1. Results on the CMU-Mixed10 dataset.

F. More Experiments on Long Sequences.

To further verify the ability of our model on long sequences, we add the experiment results with 1/2 ratio on 3DPW-Ori and 3DPW-RC in Tab. F.1, and add the FDE results with 1s/3s on Mocap-Syn in Tab. F.2. Compared with recent (2023-2024) works, our model still achieves excellent results for these experiments.

Datasets	3DPW-Ori								3DPW-RC							
Metric	MPJPE		VIM		APE		FDE		MPJPE		VIM		APE		FDE	
TBIF [4]	113	216	81	158	121	163	151	359	99	168	70	111	118	159	123	212
T2P [3]	104	**187**	**73**	**136**	122	167	**120**	**288**	95	152	66	97	123	167	**96**	170
Ours	**102**	207	77	152	**107**	**148**	140	346	**85**	**142**	**61**	**93**	**100**	**136**	109	**168**

Table F.1. The result of 10-in/20-out frames on the 3DPW dataset. Each metric records the results of frames 10 and 20 respectively.

Methods	TBIF [4]	JRT [7]	T2P [3]	Ours
1s	139	126	116	**107**
2s	247	226	172	**168**
3s	339	316	244	**243**

Table F.2. FDE results in Mocap-Syn on 1s/3s setting.

G. More Qualitative Results

In this section, we provide additional visualization results. Fig. E.1 presents qualitative results in the *CMU-Syn* (1s/1s) setting, while Fig. E.2 showcases results in the *CMU-Syn/MuPoTS* (1s/1s) setting. These results highlight the effectiveness of our EMPMP model, as it generates motion sequences that exhibit a closer alignment with the ground truth, thereby preserving realistic motion dynamics.

Metric	Settings	3DPW-Ori	3DPW-RC	CMU-Syn		AMASS/ 3DPW-Ori	AMASS/ 3DPW-RC	CMU-Syn/MuPoTS	
	In/Out Length	1030ms / 900ms		2s / 2s	1s / 1s	1030ms / 900ms		2s / 2s	1s / 1s
JPE	MRT [6] '2021	236.4	182.0	197.0	92.2	208.8	169.4	228.3	106.6
	SoMoFormer [5] '2022	207.0	143.7	184.1	86.2	167.9	130.3	200.5	98.5
	TBIFormer [4] '2023	202.8	163.2	214.8	103.4	197.1	149.6	218.0	111.4
	JRT [7] '2023	223.8	154.3	193.7	94.2	196.5	138.1	199.1	102.9
	T2P [3] '2024	190.9	150.4	167.5	84.6	173.4	128.8	199.1	109.7
	Ours	**184.8**	**128.1**	**155.2**	**79.5**	**164.1**	**118.6**	**193.7**	**98.3**
FDE	MRT [6] '2021	192.8	136.6	168.9	69.0	167.2	127.5	178.4	77.2
	SoMoFormer [5] '2022	166.0	95.3	157.7	63.6	133.0	87.2	158.8	72.6
	TBIFormer [4] '2023	156.9	117.8	184.3	79.3	154.5	108.4	168.5	78.7
	JRT [7] '2023	181.3	103.3	162.8	70.5	158.0	88.5	**144.7**	75.8
	T2P$^{F=1}$ [3] '2024	165.3	106.9	169.8	73.4	146.6	93.7	164.4	78.6
	Ours	**148.8**	**86.9**	**128.1**	**57.9**	**132.2**	**79.8**	150.6	**70.6**

Table C.3. Results in multiple settings for JPE and FDE. Our EMPMP network achieves the best performance in most comparisons.

Settings	3DPW-Ori						3DPW-RC					
Selected Frames	2	4	8	10	14	AVG	2	4	8	10	14	AVG
MRT [6] '2021	19.6	36.5	68.9	86.4	123.1	66.9	18.5	33.8	59.3	71.5	93.2	55.2
SoMoFormer [5] '2022	13.0	28.5	59.5	76.4	111.7	57.8	12.3	26.5	49.9	59.5	74.0	44.4
TBIFormer [4] '2023	17.4	33.5	63.4	78.2	108.5	60.2	15.7	30.3	56.1	67.6	86.7	51.2
JRT [7] '2023	12.5	29.0	61.6	78.1	111.5	58.5	12.6	28.7	53.6	63.1	77.1	47.0
T2P [3] '2024	16.6	31.6	59.6	73.0	**100.7**	56.3	15.2	29.3	51.5	60.9	77.8	46.9
Ours	**12.3**	**26.2**	**55.1**	**70.5**	102.6	**53.3**	**11.7**	**24.5**	**46.3**	**55.2**	**69.1**	**41.3**
Settings	AMASS/3DPW-Ori						AMASS/3DPW-RC					
Selected Frames	2	4	8	10	14	AVG	2	4	8	10	14	AVG
MRT [6] '2021	21.8	39.1	65.1	75.9	94.1	59.2	20.8	36.4	58.2	66.6	79.4	52.3
SoMoFormer [5] '2022	**9.1**	**21.3**	**47.5**	**61.6**	**91.9**	**46.3**	10.6	22.8	44.5	54.0	68.4	40.0
TBIFormer [4] '2023	13.3	28.4	58.9	74.7	106.8	56.4	13.7	27.0	52.1	62.8	81.4	47.4
JRT [7] '2023	9.5	22.1	48.7	62.8	92.8	47.2	**9.5**	21.7	44.1	53.4	68.8	39.5
T2P [3] '2024	11.0	23.2	50.8	65.7	96.3	49.4	12.0	24.3	46.4	58.1	71.2	42.4
Ours	10.5	23.1	49.8	64.6	95.2	48.6	9.8	**21.6**	**42.6**	**51.8**	**66.2**	**38.4**

Table C.4. Detailed VIM results on the 3DPW dataset across different settings are presented. Our EMPMP demonstrates the best performance in the majority of comparisons across three out of the four settings.

Metric	Settings	3DPW-Ori			3DPW-RC			AMASS/3DPW-Ori			AMASS/3DPW-RC		
	Selected Frames	7	14	AVG	7	14	AVG	7	14	AVG	7	14	AVG
JPE	MRT [6] '2021	150.0	322.9	236.4	128.1	235.9	182.0	133.5	284.2	208.8	121.7	217.1	169.4
	SoMoFormer [5] '2022	125.3	288.8	207.0	105.0	182.5	143.7	101.0	234.9	167.9	92.3	168.4	130.3
	TBIFormer [4] '2023	132.1	273.5	202.8	116.1	210.3	163.2	124.2	270.0	197.1	105.3	193.9	149.6
	JRT [7] '2023	138.7	308.9	223.8	116.6	192.0	154.3	116.3	276.7	196.5	99.7	176.5	138.1
	T2P [3] '2024	126.9	**255.0**	190.9	110.2	190.6	150.4	111.1	235.7	173.4	90.2	167.5	128.8
	Ours	**110.9**	258.7	**184.8**	**95.5**	**160.7**	**128.1**	**98.5**	**229.7**	**164.1**	**85.3**	**151.9**	**118.6**
APE	MRT [6] '2021	103.3	146.9	125.1	102.3	145.0	123.6	95.1	135.6	115.3	90.9	130.7	110.8
	SoMoFormer [5] '2022	92.0	144.7	118.3	91.8	138.0	114.9	74.9	120.2	97.5	78.4	124.4	101.4
	TBIFormer [4] '2023	94.9	137.0	115.9	94.3	136.7	115.5	87.9	132.8	110.3	85.3	130.8	108.0
	JRT [7] '2023	99.0	147.0	123.0	97.6	143.5	120.5	87.0	141.3	114.1	85.4	139.7	112.5
	T2P [3] '2024	92.0	138.3	115.1	92.0	138.3	115.1	83.1	137.1	110.1	82.1	135.3	108.7
	Ours	**78.0**	**119.3**	**98.6**	**75.4**	**117.9**	**96.6**	**73.8**	**116.2**	**95.0**	**70.9**	**110.4**	**90.6**
FDE	MRT [6] '2021	105.7	280.0	192.8	87.6	185.7	136.6	90.9	243.6	167.2	82.4	172.7	127.5
	SoMoFormer [5] '2022	86.9	245.2	166.0	63.4	127.3	95.3	70.9	195.2	133.0	59.4	115.1	87.2
	TBIFormer [4] '2023	89.5	224.4	156.9	74.7	160.9	117.8	87.1	221.9	154.5	70.4	146.4	108.4
	JRT [7] '2023	99.7	263.0	181.3	74.8	131.9	103.3	83.1	233.0	158.0	60.6	116.4	88.5
	T2P$^{F=1}$ [3] '2024	84.5	246.1	165.3	68.6	145.3	106.9	75.4	217.9	146.6	63.7	123.7	93.7
	Ours	**78.3**	**219.4**	**148.8**	**59.0**	**114.8**	**86.9**	**69.6**	**194.8**	**132.2**	**54.1**	**105.6**	**79.8**

Table C.5. Detailed JPE,APE, and FDE results on the 3DPW dataset across different settings, with our model demonstrating dominant superiority over the compared methods.

Settings	CMU-Syn (2s/2s)						CMU-Syn (1s/1s)					
Selected Frames	2	6	11	21	30	AVG	2	4	8	10	15	AVG
MRT [6] '2021	14.6	39.3	58.6	87.8	107.6	61.5	11.4	22.2	39.5	46.4	62.4	36.3
SoMoFormer [5] '2022	9.8	31.9	51.6	84.6	105.6	56.7	**8.4**	19.0	37.2	44.9	62.7	34.4
TBIFormer [4] '2023	14.6	39.0	59.9	93.5	116.4	64.6	11.9	24.2	43.8	51.6	70.4	40.3
JRT [7] '2023	**9.2**	**29.3**	52.7	81.9	109.9	56.6	9.3	20.3	37.6	44.7	58.4	34.0
T2P [3] '2024	14.4	36.4	52.7	75.1	94.6	54.6	10.4	21.1	37.6	43.5	58.8	34.2
Ours	11.8	32.8	**48.4**	**69.6**	**88.8**	**50.2**	8.9	**18.8**	**35.0**	**41.8**	**57.8**	**32.4**
Settings	CMU-Syn/MuPoTS (2s/2s)						CMU-Syn/MuPoTS (1s/1s)					
Selected Frames	2	6	11	21	30	AVG	2	4	8	10	15	AVG
MRT [6] '2021	13.3	35.3	61.6	104.2	136.3	70.1	12.8	23.7	43.8	53.3	74.4	41.6
SoMoFormer [5] '2022	**12.3**	32.6	**56.3**	94.2	123.2	63.7	**12.1**	**22.1**	**41.3**	**50.5**	70.9	**39.3**
TBIFormer [4] '2023	14.1	37.0	62.0	99.1	129.4	68.3	13.4	25.1	46.7	56.7	78.2	44.0
JRT [7] '2023	13.6	34.5	56.4	93.7	125.5	64.7	14.4	26.0	44.0	52.6	**69.2**	41.2
T2P [3] '2024	14.5	37.6	60.2	**91.8**	**116.1**	64.0	13.6	25.0	44.6	53.6	73.3	42.0
Ours	12.7	**32.5**	**56.3**	92.6	120.7	**62.9**	12.6	23.2	43.2	52.4	71.9	40.6

Table C.6. Detailed VIM results on the CMU-Syn and MuPoTS-3D dataset across different settings. Our model achieves the best average results across three out of four settings.

Metric	Settings	CMU-Syn (2s/2s)				CMU-Syn (1s/1s)				CMU-Syn/MuPoTS (2s/2s)				CMU-Syn/MuPoTS (1s/1s)			
	Selected Frames	10	20	30	AVG	3	9	15	AVG	10	20	30	AVG	3	9	15	AVG
JPE	MRT [6] 2021	125.5	203.8	261.7	197.0	35.6	95.8	145.4	92.2	129.1	231.6	324.4	228.3	41.2	108.7	169.9	106.6
	SoMoFormer [5] 2022	105.0	193.6	253.6	184.0	**26.9**	88.2	143.5	86.2	111.8	205.4	284.4	200.5	**37.8**	**100.6**	157.1	98.5
	TBIFormer [4] 2023	132.9	223.4	288.2	214.8	38.1	106.8	165.3	103.4	126.6	221.0	306.5	218.0	42.8	114.1	177.5	111.4
	JRT [7] 2023	115.1	200.6	265.5	193.7	29.9	97.0	155.7	94.2	117.5	203.0	277.0	199.1	42.1	105.1	161.7	102.9
	T2P [3] 2024	113.2	167.9	221.6	167.5	32.3	87.3	134.2	84.6	127.4	203.0	**267.1**	199.1	46.6	114.3	168.4	109.7
	Ours	**100.8**	**154.6**	**210.4**	**155.2**	27.4	**81.5**	**129.8**	**79.5**	**109.3**	**196.3**	275.6	**193.7**	38.5	101.3	**155.3**	**98.3**
APE	MRT [6] 2021	82.5	99.0	105.1	95.5	30.5	67.0	82.6	60.0	97.1	143.5	164.8	135.1	38.8	86.6	118.4	81.2
	SoMoFormer [5] 2022	68.0	91.6	101.7	87.1	24.2	62.5	78.9	55.2	93.7	137.8	161.1	130.8	38.6	86.1	117.2	80.6
	TBIFormer [4] 2023	83.3	100.1	107.3	96.9	32.0	72.1	86.7	63.6	101.8	144.6	166.8	137.7	39.9	90.1	123.2	84.4
	JRT [7] 2023	80.5	99.4	107.6	95.8	26.3	68.8	89.9	61.6	93.1	132.2	**151.9**	125.7	37.2	84.2	113.9	78.4
	T2P [3] 2024	79.0	98.7	105.9	94.5	28.2	67.4	84.1	59.9	109.2	155.1	178.9	147.7	44.1	99.7	134.8	92.8
	Ours	**67.6**	**86.1**	**95.9**	**83.2**	**24.0**	**58.4**	**74.6**	**52.3**	**89.0**	**131.9**	153.9	**124.9**	**37.0**	**81.6**	**110.0**	**76.2**
FDE	MRT [6] 2021	94.3	177.2	235.4	168.9	21.7	69.0	116.3	69.0	87.8	177.5	270.1	178.4	28.8	77.2	125.8	77.2
	SoMoFormer [5] 2022	77.5	167.5	228.1	157.7	**14.2**	60.9	115.8	63.6	81.3	159.7	235.4	158.8	28.4	73.4	116.2	72.6
	TBIFormer [4] 2023	99.8	193.2	259.9	184.3	24.1	78.1	135.8	79.3	84.7	169.7	251.2	168.5	29.9	78.1	128.2	78.7
	JRT [7] 2023	85.8	169.3	233.5	162.8	17.3	68.7	125.6	70.5	**76.3**	**144.4**	**213.4**	**144.7**	32.8	76.2	118.6	75.8
	T2P$^{F=1}$ [3] 2024	85.8	178.4	245.2	169.8	17.8	70.8	131.7	73.4	90.5	166.8	235.9	164.4	30.3	79.9	125.7	78.6
	Ours	**72.3**	**127.3**	**184.8**	**128.1**	15.0	**55.9**	**103.0**	**57.9**	78.4	151.1	222.3	150.6	**27.4**	**70.4**	**114.0**	**70.6**

Table C.7. Detailed JPE,APE, and FDE results on the CMU-Syn and MuPoTS-3D dataset across different settings, and our model present dominant superiority over most of the compared methods.

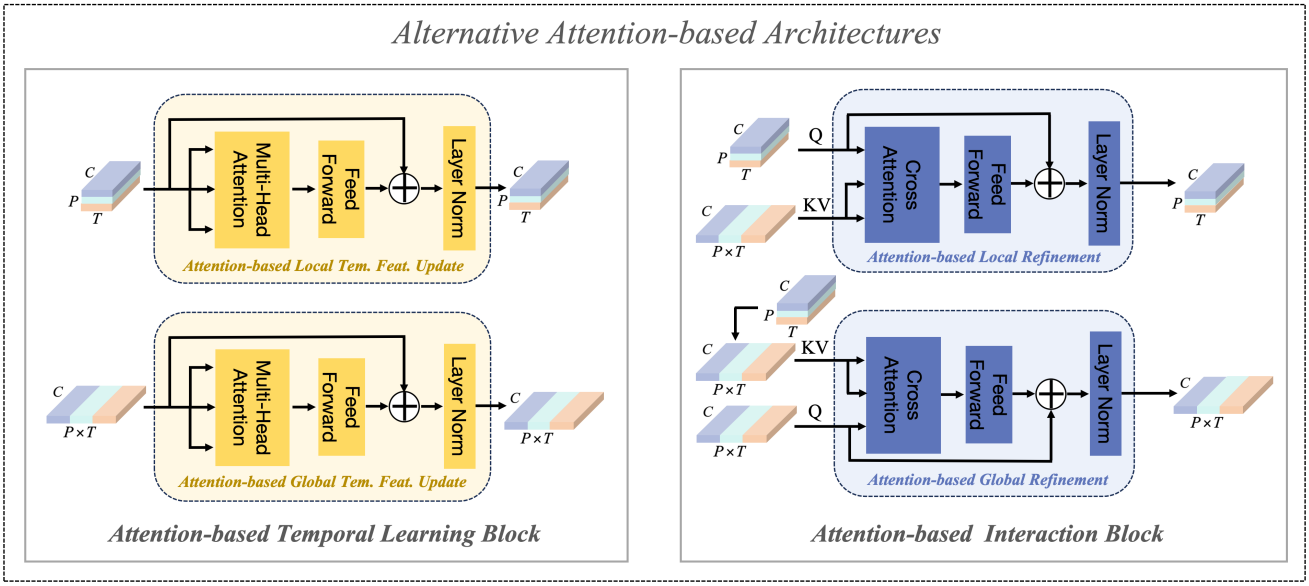

Figure B.2. Attention-based architectures used in our ablation study. Our ME block and CI block are replaced with Attention-based Temporal Learning block containing Multi-Head Attention (Self-Attention) module and Attention-based Interaction block containing Cross Attention module,respectively.

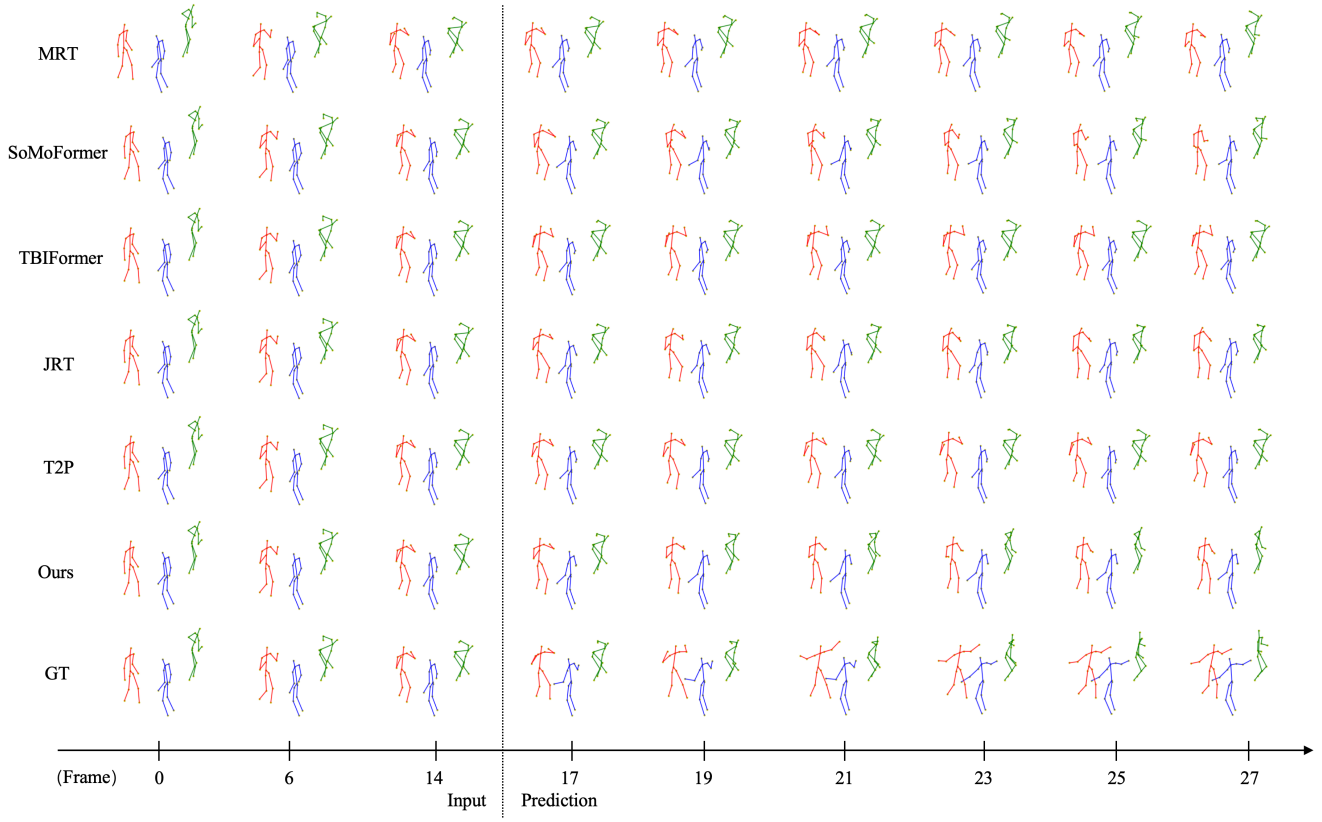

Figure E.1. Qualitative results in the *CMU-Syn* (1s/1s) setting. Different colors indicate different individuals. The model predicts 15 frames based on the input of 15 frames.

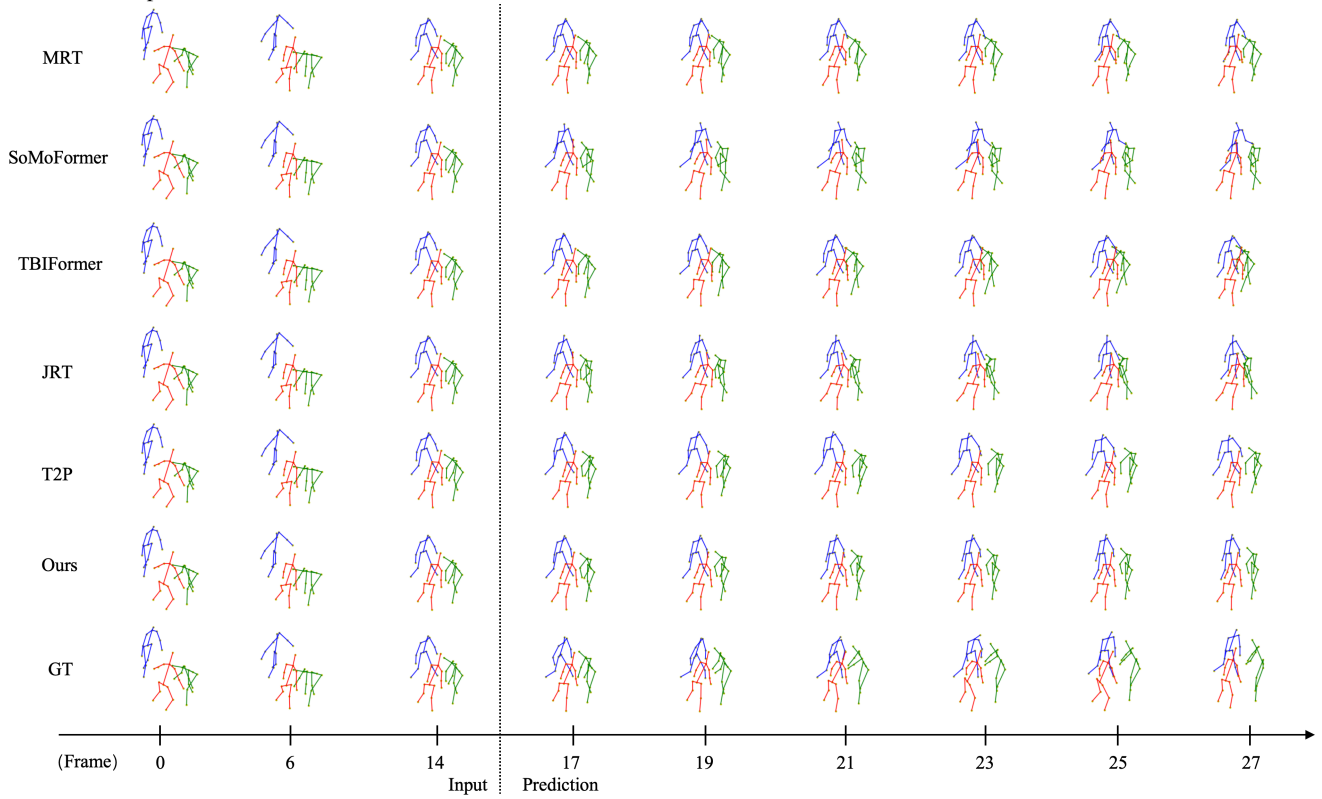

Figure E.2. Qualitative results in the *CMU-Syn/MuPoTS* (1s/1s) setting. Different colors indicate different individuals. The model predicts 15 frames based on the input of 15 frames.

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 1

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 1

[3] Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1617–1628, 2024. 2, 3, 4, 5, 6, 7

[4] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17121–17130, 2023. 2, 3, 4, 5, 6, 7

[5] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. Somoformer: Multi-person pose forecasting with transformers. *arXiv preprint arXiv:2208.14023*, 2022. 1, 2, 5, 6, 7

[6] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 1, 2, 3, 5, 6, 7

[7] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9816–9826, 2023. 1, 2, 4, 5, 6, 7