NeRMo: Learning Implicit Neural Representations for 3D Human Motion Prediction

Dong Wei, Huaijiang Sun, Xiaoning Sun, Shengxiang Hu

Nanjing University of Science and Technology, Nanjing, China

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Quick Preview

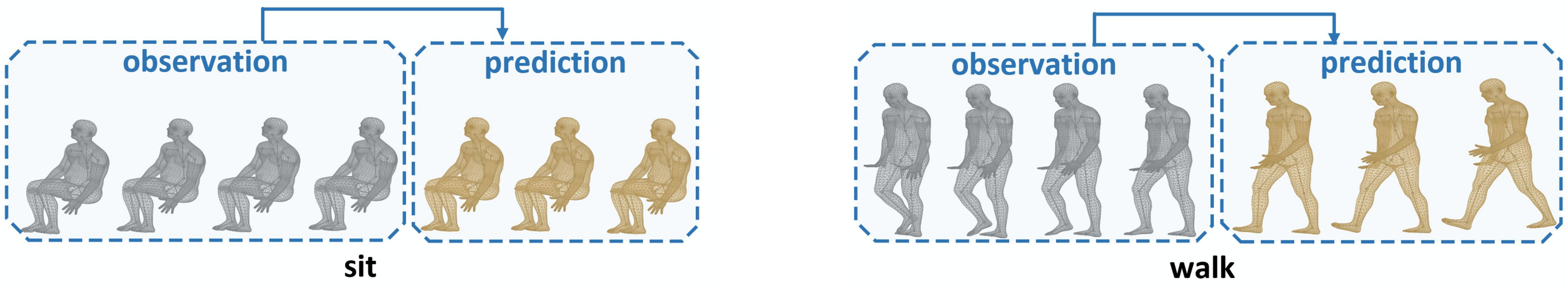
- Introduction to the background and challenges in human motion prediction
- Reformulate human motion prediction from continuous perspective
- Propose an efficient meta-optimization to learn strong inductive bias
- Achieve improved prediction performance based on complete and incomplete observations

Self-Driving Cars

Human-Robot Interaction

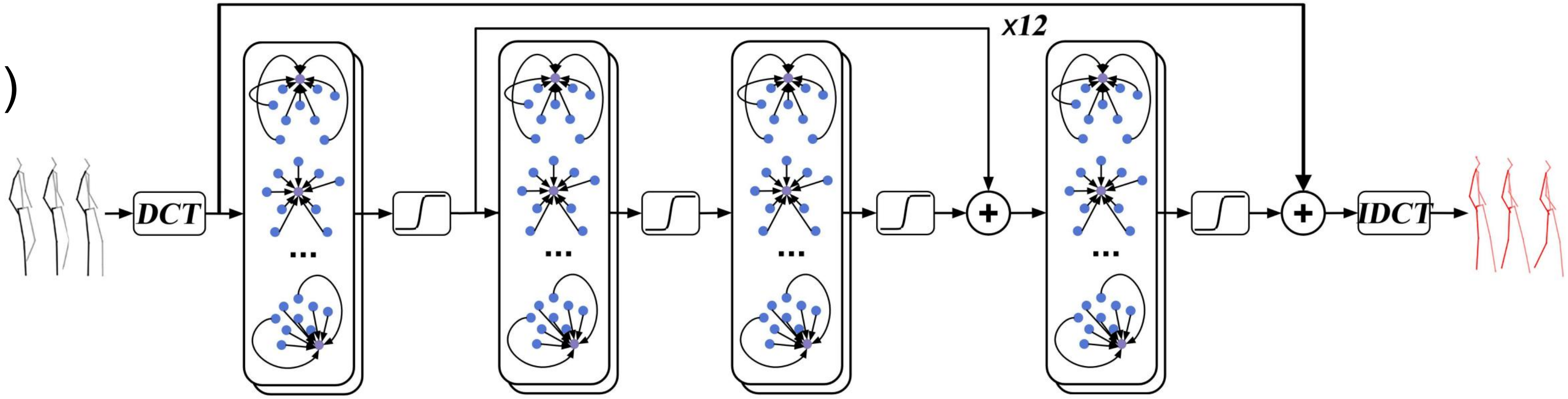
Problem Definition

Human Motion Prediction aims to accurately forecast the future motion conditioned on the historical observed movements.

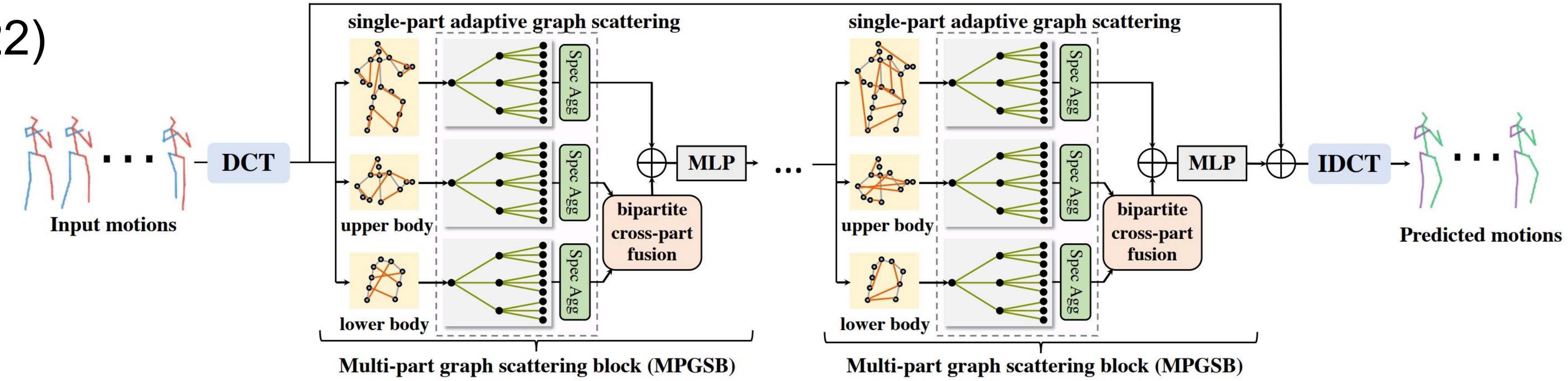


Related Work

LTD (Mao et al. 2019)



SPGSN (Li et al. 2022)



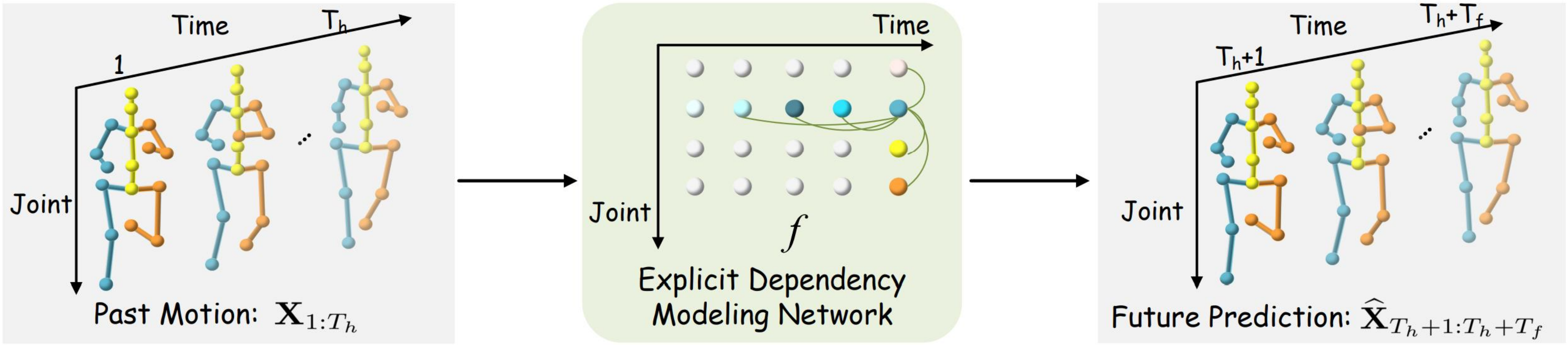
Mao Wei, Miaomiao Liu, et al. Learning Trajectory Dependencies for Human Motion Prediction. ICCV, 2019.
Maosen Li, Siheng Chen, et al. Skeleton Graph Scattering Networks for 3D Human Motion Prediction. ECCV, 2022.

Challenge

Existing human motion prediction methods: Predicting future human poses as a function of the historical observations (named historical-value models), that is,

$$\widehat{\mathbf{X}}_{T_h+1:T_h+T_f} = f(\mathbf{X}_{1:T_h})$$

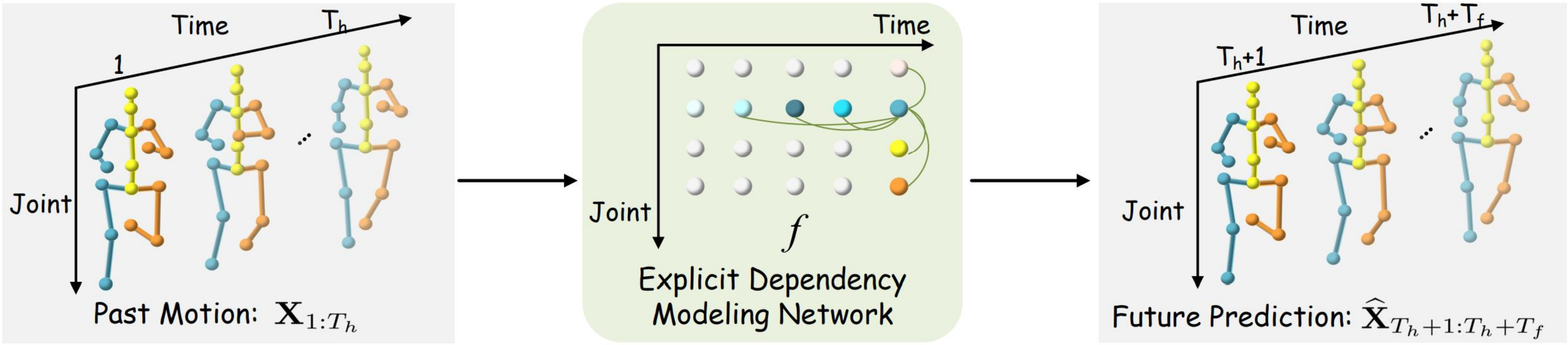
They always focus on dedicated network structure to capture the spatial-temporal relations.



(a) Previous historical-value models

Challenge

$$\widehat{\mathbf{X}}_{T_h+1:T_h+T_f} = f(\mathbf{X}_{1:T_h})$$

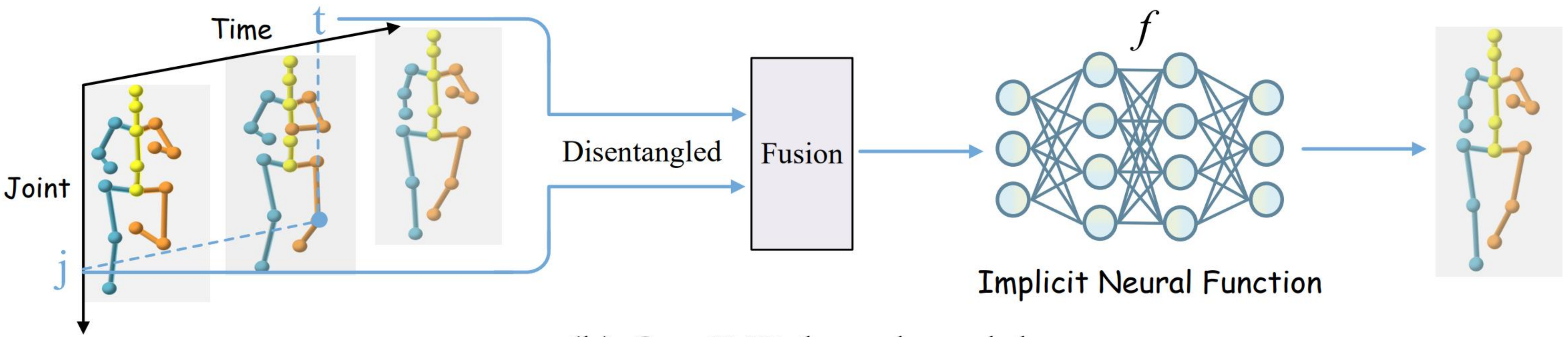


(a) Previous historical-value models

- Such modeling ignores the underlying continuous temporal dynamics.
- It suffers from considerable perforamnce degardation when handling incomplete observations, which is frequently encountered in real-world scenarios.

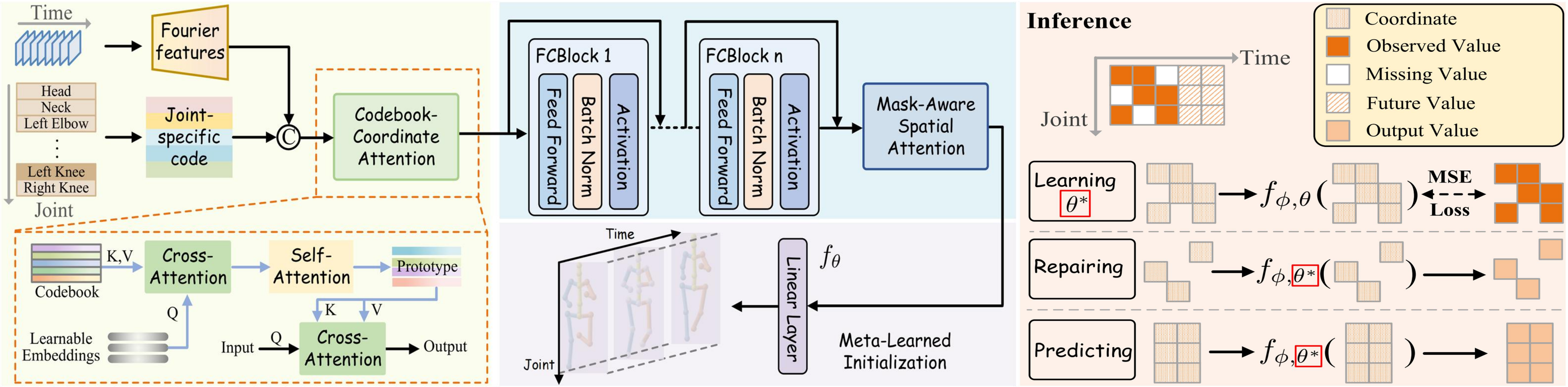
Key Insight

Our key idea is to learn an ***Implicit Neural Representation*** (INR), which represents motions as a continuous function to approximate the temporal dynamics.



(b) Our INR-based model

Overview

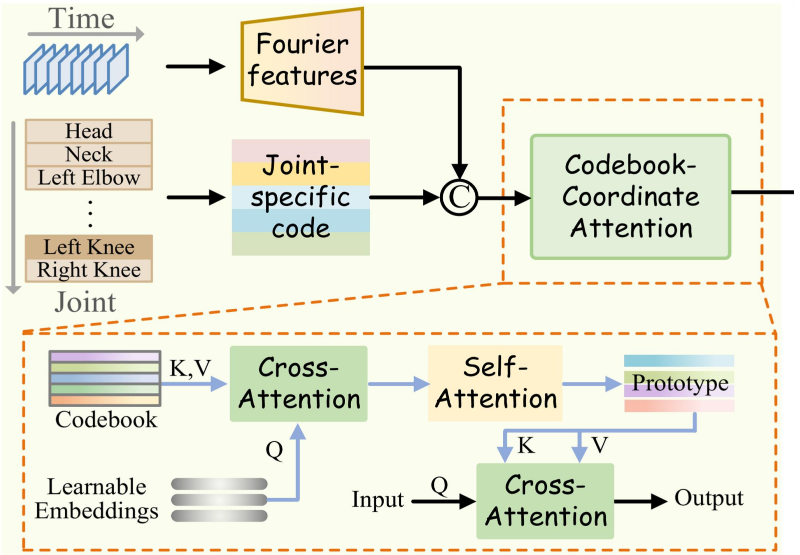


(Left) The disentangled spatial-temporal representations are fed into a codebook-coordinate attention.

(Middle) NeRMo consists of several MLPs and a mask-aware spatial attention. The meta-optimization framework is designed to learn strong inductive bias by a bi-level optimization.

(Right) At inference, NeRMo can simultaneously handle missing values and predict future poses.

Reformulation



➤ Neural Motion Representation

$$f_\theta : (t, \boldsymbol{z}) \longmapsto \boldsymbol{x}_t$$

where $\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j=1}^{J}$ is a set of learnable joint-specific latent codes,

$\boldsymbol{z}_j \in \mathbb{R}^d$ is indexed by a discrete joint variable j.

➤ Fourier Mapping

$$\gamma(t) = (\cdots, \sin(2^l \pi t), \cos(2^l \pi t), \cdots)$$

where frequency embedding maps the temporal coordinates t from $\mathbb{R}^1$ into higher diemnsional space $\mathbb{R}^{2L}$, $l \in \{0, 1, 2, \cdots, L-1\}$.

➤ Codebook-Coordinate Attention

We exploit the knowledge contained in codebook to enrich the feature representation by cross-attention machanism.

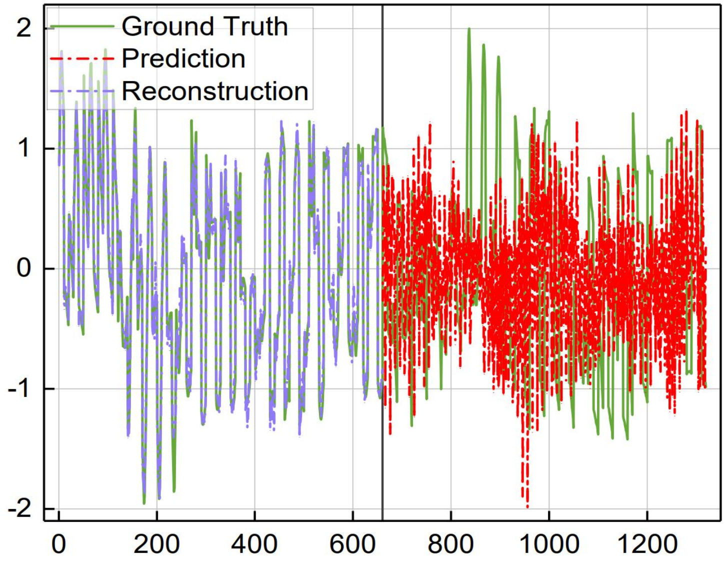
Meta-Optmization

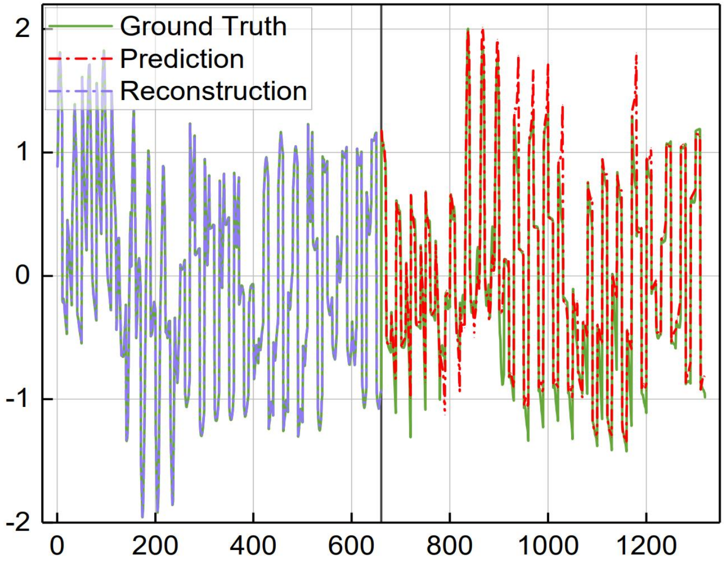
Motivation:

1) Vanilla INRs are required to encode each motion into a separate continuous function, which is not optimal when confront with a large number of diverse motions.

2) Vanilla INRs struggle with extrapolation across the forecast horizon.

Ours





Meta-Optmization

The Meta-optimization framework is carefully designed to learn the strong inductive bias by a bi-level optimization by categorizing the parameters of INRs into two types:

1) Personalized Modulation: as instance-specific parameter ϕ

2) Generic Rule: as instance-agnostic parameter θ

Loss Function:

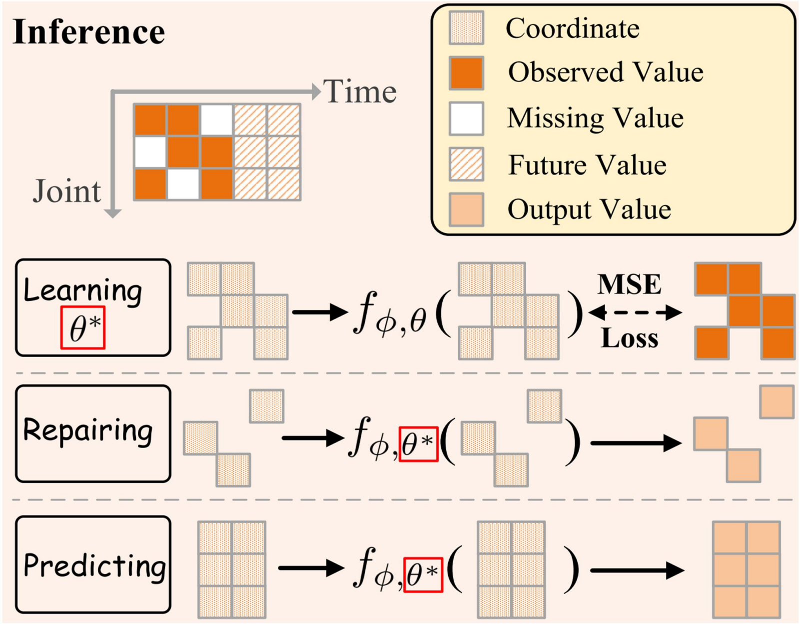
$$\phi^* = \arg\min_{\phi} \sum_{i=1}^{N} \sum_{t=T_h+1}^{T_h+T_f} \mathcal{L}(f_{\phi,\theta_i^*}(t, \boldsymbol{z}), \boldsymbol{x}_t^{(i)}),$$

$$\text{s.t.} \quad \theta_i^* = \arg\min \sum_{t=1}^{T_h} \mathcal{L}(f_{\phi,\theta_i}(t, \boldsymbol{z}), \boldsymbol{x}_t^{(i)}),$$

Outer Loop: learn a strong inductive bias for extrapolation

Inner Loop: act as the standard supervised learning process;

Inference



During the inference process, our goal is to estimate the value for future timestamps based on the incoming observations, including:

1) Conventional setting;

2) Incomplete past motion.

Process:

1) Fix the generic rule parameters,

2) Compute the personalized modulation for new data.

Experiments

- Datasets

Human3.6M

CMU-MoCap

3DPW

- Evaluaton Metric

MPJPE: Mean Per Joint Position Error on 3D human joint coordinates.

- Baselines

Res-sup, LTD, DMGNN, MSR-GCN, PGBIG, SPGSN, DeFeeNet, MT-GCN

Experiments

- Conventional Motion Prediction

Results on Human3.6M dataset

scenarios	walking				eating				smoking				discussion			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup. [31]	29.4	50.8	76.0	81.5	16.8	30.6	56.9	68.7	23.0	42.6	70.1	82.7	32.9	61.2	90.9	96.2
DMGNN [26]	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8
LTD [30]	12.3	23.0	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9	12.5	27.4	58.5	71.7
MSR-GCN [11]	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2	12.0	26.8	57.1	69.7
PGBIG [28]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN [25]	10.1	19.4	34.8	41.5	7.1	14.9	**30.5**	**37.9**	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
DeFeeNet [43]	10.4	20.0	34.7	42.2	7.0	15.2	31.4	38.4	6.8	14.5	29.0	35.8	11.1	25.4	55.8	68.2
Ours	**9.7**	**18.6**	**33.2**	**39.8**	**6.8**	**14.6**	30.9	40.3	**6.1**	**11.7**	**26.5**	**33.9**	**9.4**	**21.0**	**49.8**	**65.2**

scenarios	directions				greeting				phoning				posing			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup. [31]	35.4	57.3	76.3	87.7	34.5	63.4	124.6	142.5	38.0	69.3	115.0	126.7	36.1	69.1	130.5	157.1
DMGNN [26]	13.1	24.6	64.7	81.9	23.3	50.3	107.3	132.1	12.5	25.8	48.1	58.3	15.3	29.3	71.5	96.7
LTD [30]	9.0	19.9	43.4	53.7	18.7	38.7	77.7	93.4	10.2	21.0	42.5	52.3	13.7	29.9	66.6	84.1
MSR-GCN [11]	8.6	19.7	43.3	53.8	16.5	37.0	77.3	93.4	10.1	20.7	41.5	51.3	12.8	29.4	67.0	85.0
PGBIG [28]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN [25]	7.4	17.2	39.8	50.3	**14.6**	**32.6**	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
DeFeeNet [43]	7.0	17.0	40.0	50.9	16.8	33.0	**68.5**	**83.2**	11.6	19.9	41.0	50.1	14.7	28.3	65.0	81.1
Ours	**6.7**	**16.8**	**39.5**	**48.8**	15.2	34.3	73.2	91.7	**8.0**	**17.7**	**37.9**	**48.0**	**9.4**	**22.5**	**56.1**	**72.1**

scenarios	purchases				sitting				sittingdown				takingphoto			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup. [31]	36.3	60.3	86.5	95.9	42.6	81.4	134.7	151.8	47.3	86.0	145.8	168.9	26.1	47.6	81.4	94.7
DMGNN [26]	21.4	38.7	75.7	92.7	11.9	25.1	44.6	**50.2**	15.0	32.9	77.1	93.0	13.6	29.0	46.0	58.8
LTD [30]	15.6	32.8	65.7	79.3	10.6	21.9	46.3	57.9	16.1	31.1	61.5	75.5	9.9	20.9	45.0	56.6
MSR-GCN [11]	14.8	32.4	66.1	79.6	10.5	22.0	46.3	57.8	16.1	31.6	62.5	76.8	9.9	21.0	44.6	56.3
PGBIG [28]	**12.5**	28.7	**60.1**	**73.3**	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN [25]	12.8	**28.6**	61.0	74.4	9.3	19.4	**42.3**	53.6	14.2	27.7	**56.8**	**70.7**	8.8	18.9	41.5	52.7
DeFeeNet [43]	16.8	32.7	67.9	80.3	14.2	23.6	47.7	58.7	**10.1**	29.4	62.0	70.8	**7.8**	**16.9**	**38.3**	**47.9**
Ours	13.6	30.5	64.6	78.1	**8.5**	**18.7**	42.5	54.4	13.4	**27.3**	58.2	73.5	8.1	18.1	40.9	51.7

scenarios	waiting				walkingdog				walkingtogether				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res-sup. [31]	30.6	57.8	106.2	121.5	64.2	102.1	141.1	164.4	26.8	50.1	80.2	92.2	34.7	62.0	101.1	115.5
DMGNN [26]	12.2	24.2	59.6	77.5	47.1	93.3	160.1	171.2	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
LTD [30]	11.4	24.0	50.1	61.5	23.4	46.2	83.5	96.0	10.5	21.0	38.5	45.2	12.7	26.1	52.3	63.5
MSR-GCN [11]	10.7	23.1	48.3	59.2	20.7	42.9	80.4	93.3	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
PGBIG [28]	**8.9**	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN [25]	9.2	**19.8**	43.1	54.1	17.8	37.2	71.7	**84.9**	8.9	18.2	33.8	40.9	10.4	22.3	**47.1**	**58.3**
DeFeeNet [43]	9.6	**19.8**	**42.3**	**53.6**	17.6	41.1	72.7	**84.9**	8.8	19.0	36.1	41.8	11.3	23.7	48.8	59.2
Ours	10.8	23.5	50.6	63.1	**17.3**	**36.7**	**71.4**	85.0	**8.1**	**16.7**	**32.9**	**40.3**	**9.9**	**21.8**	**47.1**	59.1

Experiments

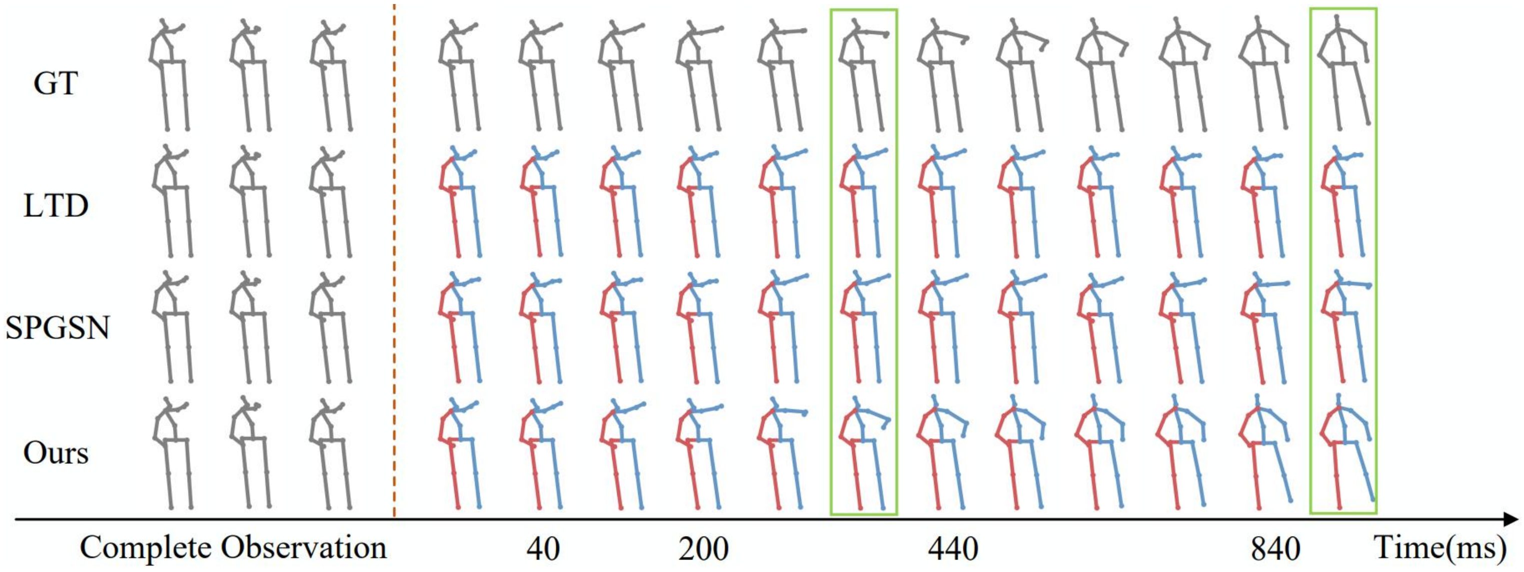
➢ Conventional Motion Prediction

dataset	CMU-Mocap				3DPW			
millisecond	80ms	160ms	320ms	400ms	200ms	400ms	600ms	800ms
Res-sup. [31]	24.74	44.21	76.30	88.73	113.9	173.1	191.9	201.1
DMGNN [26]	14.07	24.44	45.90	55.45	37.3	67.8	94.5	109.7
LTD [30]	9.94	18.02	33.55	40.95	35.6	67.8	90.6	106.9
MSR-GCN [11]	8.72	15.83	30.57	38.10	37.8	71.3	93.9	110.8
PGBIG [28]	<u>8.20</u>	15.41	30.13	37.27	35.3	67.8	<u>89.6</u>	**102.6**
SPGSN [25]	8.30	<u>14.80</u>	**28.64**	**36.96**	<u>32.9</u>	<u>64.5</u>	91.6	104.0
DeFeeNet [43]	-	-	-	-	33.7	65.9	90.1	<u>103.9</u>
Ours	**8.05**	**14.14**	<u>29.43</u>	<u>37.15</u>	**30.8**	**63.2**	**89.4**	106.2

Results on CMU-Mocap and 3DPW datasets

Experiments

- Conventional Motion Prediction

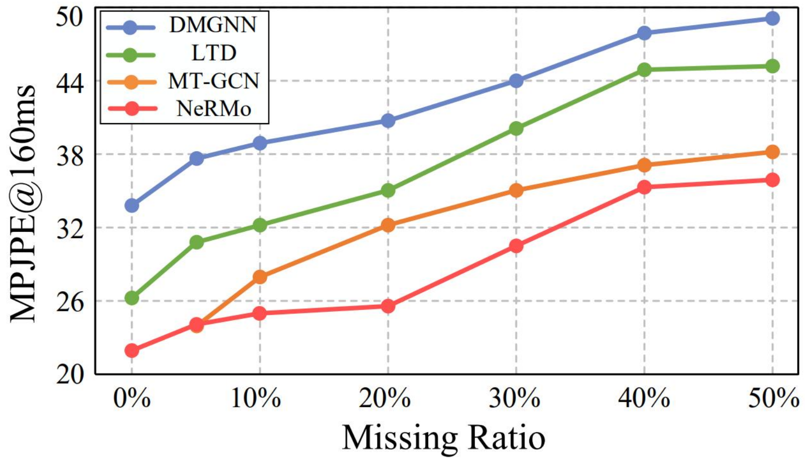


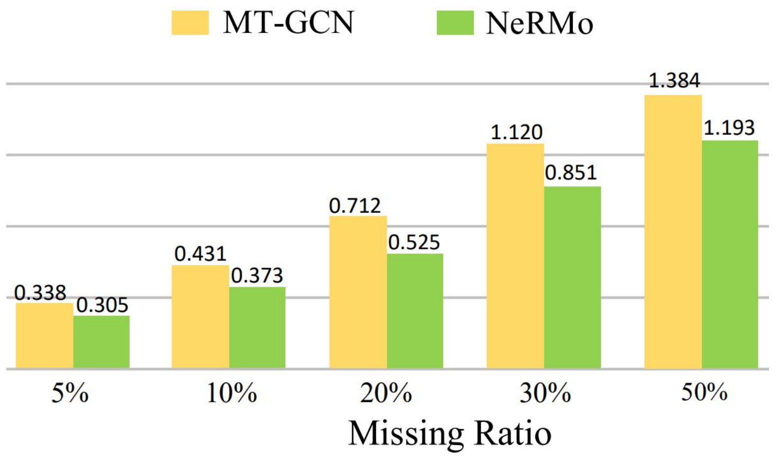
Visualization Results

Experiments

➢ Prediction based on Incomplete Observations

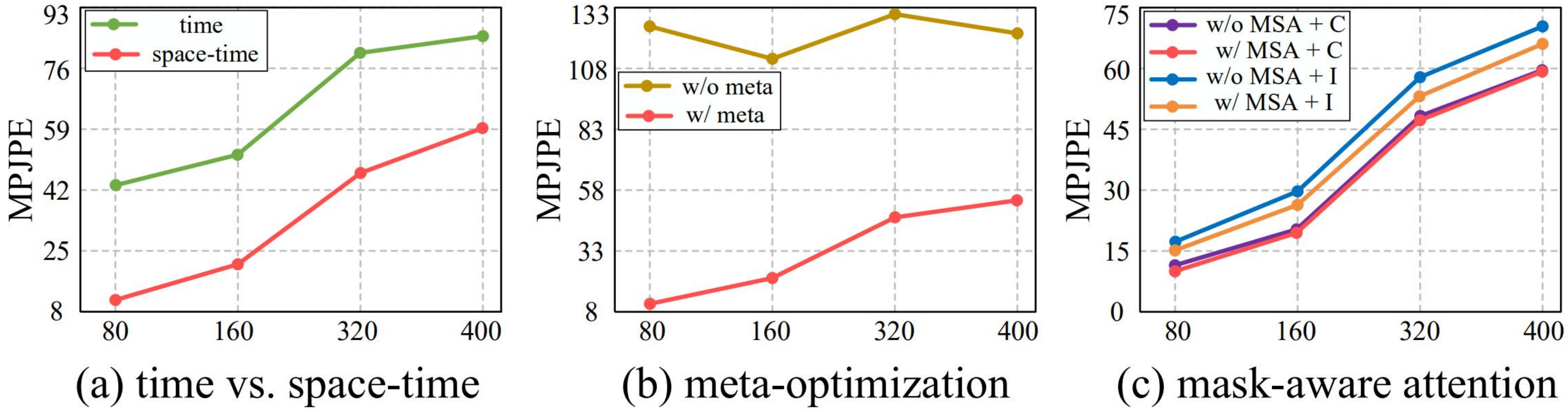
scenarios		walking				eating				smoking				discussion			
method	time cost	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
DMGNN [26]	*35.19s*	25.7	38.4	60.9	75.1	23.2	35.4	48.9	60.9	18.5	24.6	45.0	62.2	29.1	48.3	74.4	85.2
LTD [30]	*24.85s*	24.5	35.5	51.0	57.7	20.8	30.0	45.8	53.2	21.4	29.9	43.9	49.8	24.6	40.5	70.2	81.6
R+DMGNN [26]	*76.24s*	21.8	36.1	58.9	74.0	16.5	26.2	43.4	52.1	14.4	20.0	40.8	53.7	20.9	39.9	65.8	73.3
R+LTD [30]	*65.82s*	19.9	30.8	47.5	54.3	12.3	22.7	37.3	**45.1**	13.6	18.9	37.7	50.6	15.4	33.5	65.7	74.9
MT-GCN [9]	*61.35s*	16.4	24.8	40.8	48.1	11.3	19.8	38.4	47.0	11.4	16.8	34.3	**42.8**	13.3	33.1	67.6	76.5
TCD [37]	*1923.32s*	18.4	29.8	46.4	53.1	11.7	20.9	38.7	46.9	14.5	21.1	42.0	51.4	14.6	33.1	66.5	75.9
Ours	*41.27s*	**14.2**	**23.7**	**38.1**	**45.7**	**10.5**	**17.9**	**37.0**	46.6	**10.6**	**15.4**	**32.8**	43.7	**11.2**	**31.9**	**62.0**	**70.1**





Experiments

- Ablation Studies



Importance of our design, including spatial-temporal representation, meta-optimization, and model architecture

Thank you for tuning in!