HumanMAC: Masked Motion Completion for Human Motion Prediction

https://lhchen.top/Human-MAC

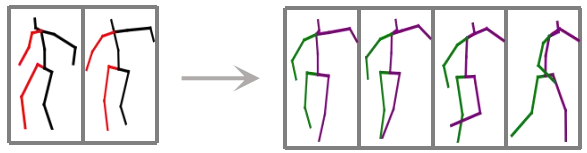
Ling-Hao Chen[1*], Jiawei Zhang[2*], Yewen Li[3], Yiren Pang[2], Xiaobo Xia[4], and Tongliang Liu[4]

[1]Tsinghua University, [2]Xidian University
[3]Nanyang Technological University, [4]The University of Sydney

{thu.lhchen}@gmail.com {zjw}@stu.xidian.edu.cn {yewen001}@e.ntu.edu.sg
{yrpang}@outlook.com {xiaoboxia.uni}@gmail.com {tongliang.liu}@sydney.edu.au

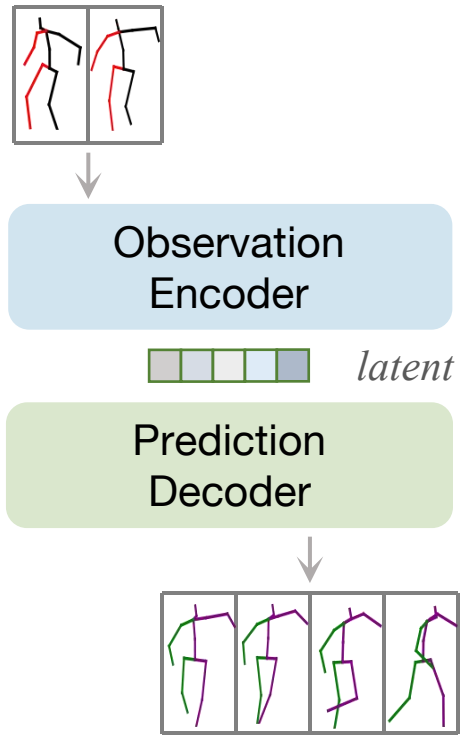
This work is led by Ling-Hao Chen and Xiaobo Xia jointly.

Motivation



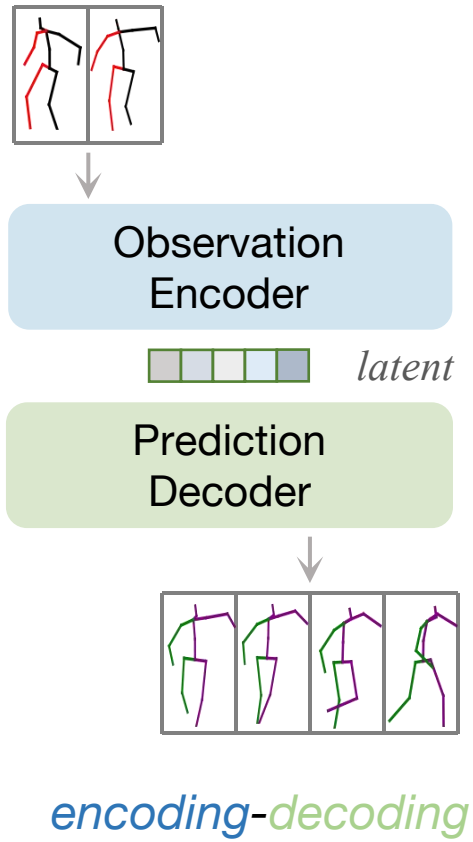
Observation Prediction

Motivation



encoding-decoding

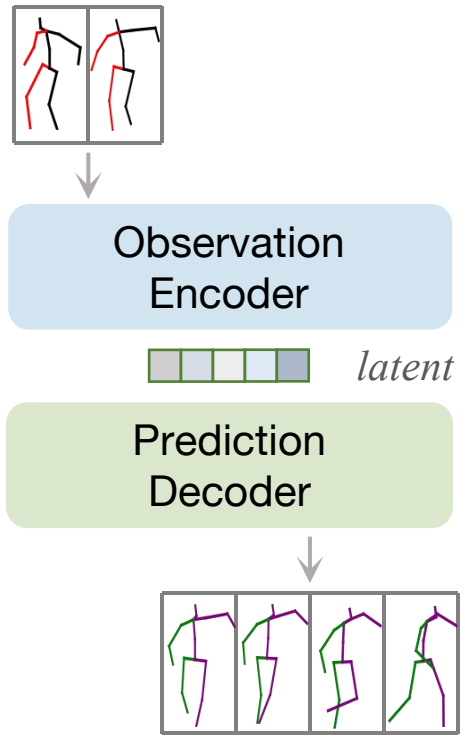
Motivation



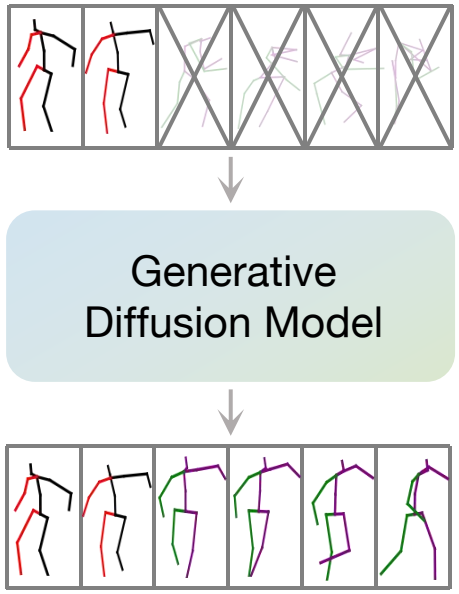
Drawbacks:

- Rely on ***multiple loss*** constraints for high-quality prediction results.
- Need ***multi-stage training***.
- It is ***hard to*** realize the ***switch*** of different categories of motions.

Motivation



encoding-decoding

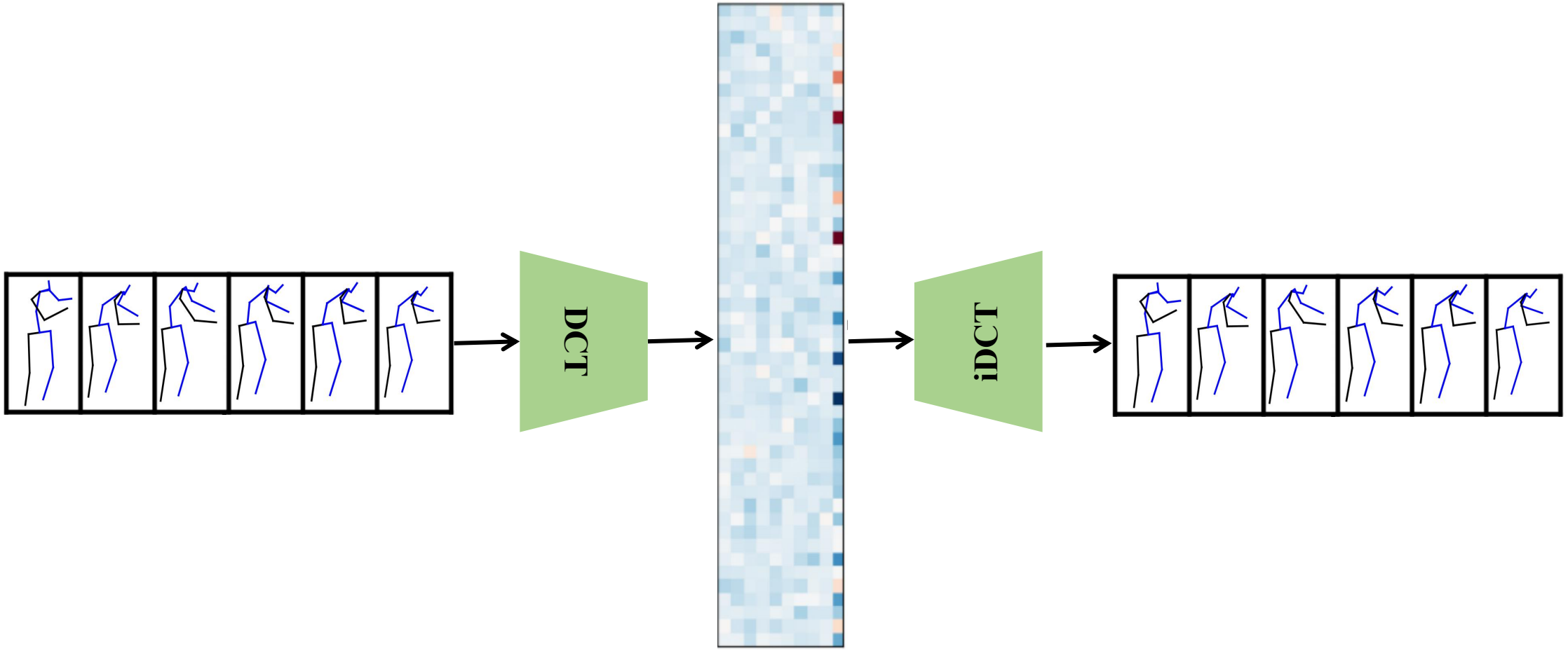


***MA**sk-**C**ompletion*
(HumanMAC)

Properties:

- ✓ Only ***one loss*** function during training.
- ✓ Trained in an ***end-to-end*** manner.
- ✓ Achieve more diverse prediction results that ***contain the switch*** of different categories of motions.

Preliminaries (*Discrete Cosine Transform*)



Methodology

☐ Model Training

Algorithm 1: Training procedure of HumanMAC

Input: motion $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$, noising steps T,
the initialized noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$,
maximum iterations $I_{\max}$.

Output: the noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

for $I = 0, 1, \ldots, I_{\max}$ **do**

$\mathbf{y}_0 = \texttt{DCT}(\mathbf{x}) \sim p(\mathbf{y}_0)$;

$t = \texttt{Uniform}(\{1, 2, \cdots, T\})$;

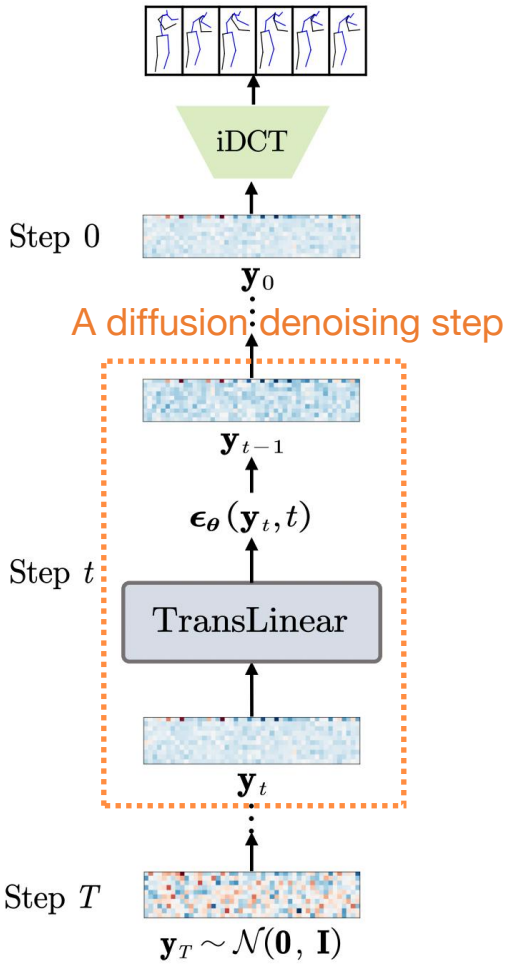
$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

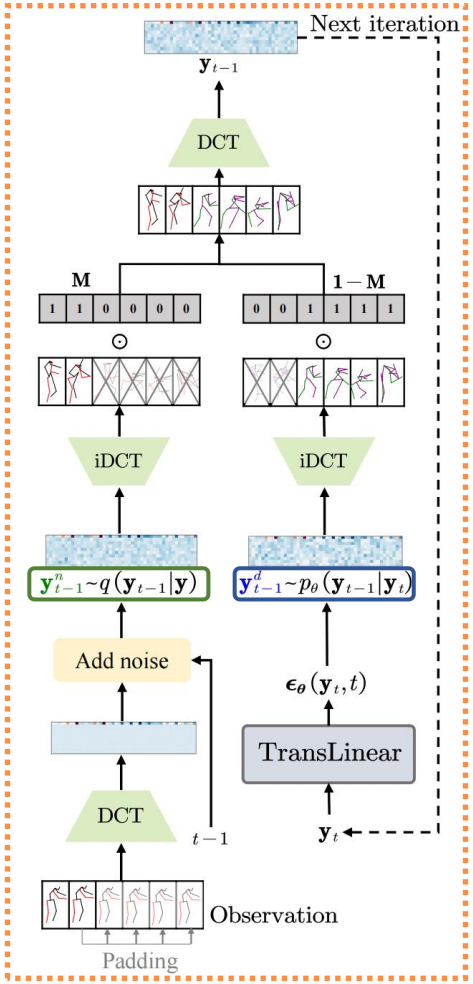
$\boldsymbol{\theta} = \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}} \left(\sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t \right) \right\|^2$;

return the noise prediction network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

Methodology

◘ DCT-Completion in Inference





A diffusion DCT-Completion step

Algorithm 2: Inference procedure of HumanMAC

Input: observed motion $\mathbf{x}^{(1:H)} \in \mathbb{R}^{H\times 3J}$, the mask of the observation $\mathbf{M}$, noising steps T, the trained noise prediction network ϵ_{θ}.

Output: competed motion $\mathbf{x} \in \mathbb{R}^{(H+F)\times 3J}$.

$\mathbf{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

$\mathbf{x} := \texttt{Pad}(\mathbf{x}) \in \mathbb{R}^{(H+F)\times 3J}$ // observation padding;

$\mathbf{y} = \texttt{DCT}(\mathbf{x}) \sim p(\mathbf{y})$;

for $t \in T, T-1, ..., 1$ **do**

$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$;

$\mathbf{y}^n_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{y} + \sqrt{1-\bar{\alpha}_{t-1}}\mathbf{z}$;

$\mathbf{y}^d_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\theta}(\mathbf{y}_t, t)\right) + \sigma_t\mathbf{z}$;

$\mathbf{y}_{t-1} = \texttt{DCT}[\mathbf{M} \odot \texttt{iDCT}(\mathbf{y}^n_{t-1}) + \texttt{iDCT}((\mathbf{1} - \mathbf{M}) \odot \mathbf{y}^d_{t-1}])$;

return $\texttt{iDCT}(\mathbf{y}_0)$.

$$\mathbf{M} = [\underbrace{1,1,\ldots,1}_{H-\text{dim}}, \underbrace{0,0,\ldots 0}_{F-\text{dim}}]^\top \Rightarrow \mathbf{M} = [\underbrace{1,1,\ldots,1}_{H-\text{dim}}, \underbrace{0,0,\ldots 0}_{(F-M)-\text{dim}}, \underbrace{1,1,\ldots 1}_{M-\text{dim}}]^\top$$

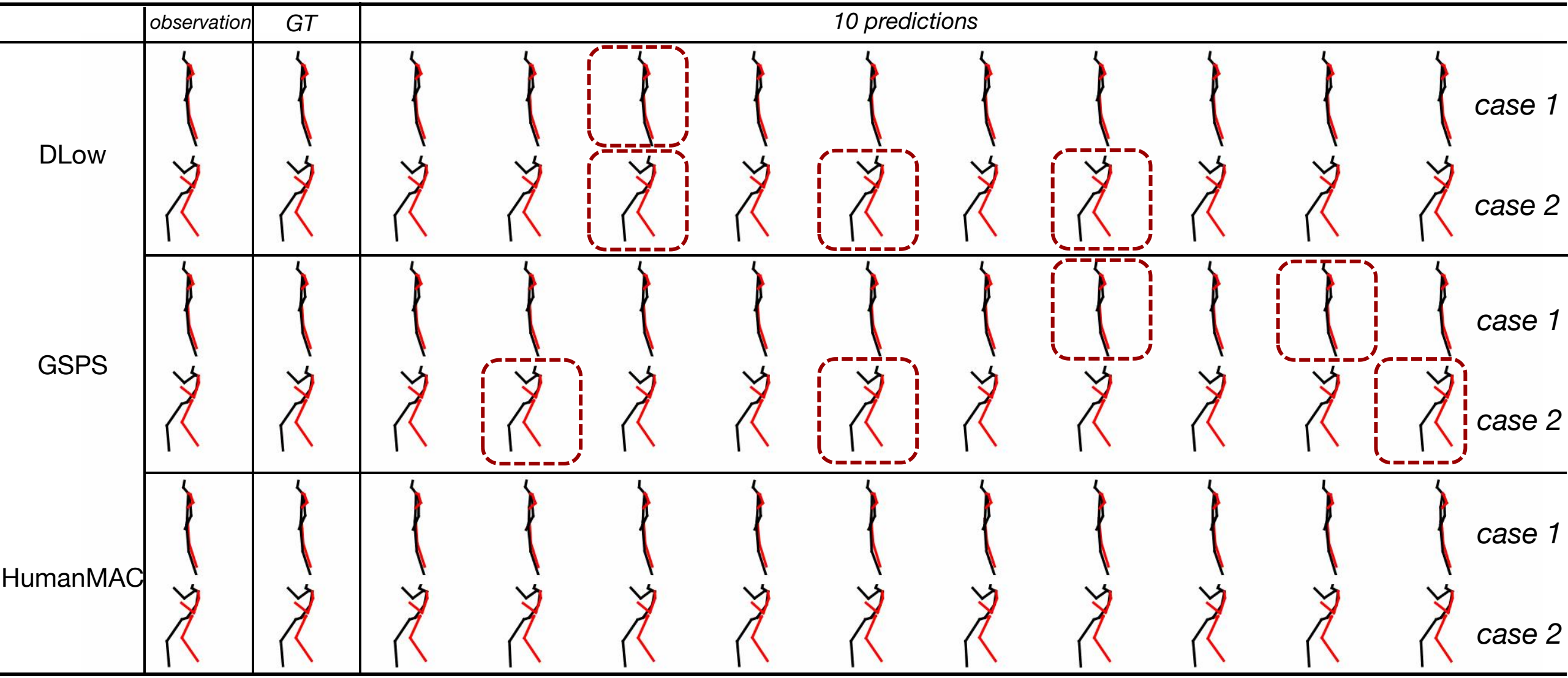
Experiments

◘ Quantitative results

	One-Stage	# Loss	Human3.6M					HumanEva-I				
			APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
acLSTM [84]	✓	1	0.000	0.789	1.126	0.849	1.139	0.000	0.429	0.541	0.530	0.608
DeLi GAN [19]	✓	1	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
MT-VAE [75]	✓	3	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
BoM [6]	✓	1	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
DSF [79]	✗	2	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
DLow[78]	✗	3	11.741	0.425	0.518	0.495	0.531	4.855	0.233	0.244	0.343	0.331
GSPS [40]	✗	5	14.757	0.389	0.496	0.476	0.525	5.825	0.233	0.244	0.343	0.331
MOJO [82]	✗	3	12.579	0.412	0.514	0.497	0.538	4.181	0.234	0.244	0.369	0.347
BeLFusion [4]	✗	4	7.602	0.372	0.474	**0.473**	**0.507**	-	-	-	-	-
DivSamp [13]	✗	3	15.310	0.370	0.485	0.475	0.516	6.109	0.220	0.234	**0.342**	**0.316**
MotionDiff [68]	✗	4	**15.353**	0.411	0.509	0.508	0.536	5.931	0.232	0.236	0.352	0.320
HumanMAC	✓	1	6.301	**0.369**	**0.480**	0.509	0.545	**6.554**	**0.209**	**0.223**	**0.342**	0.335

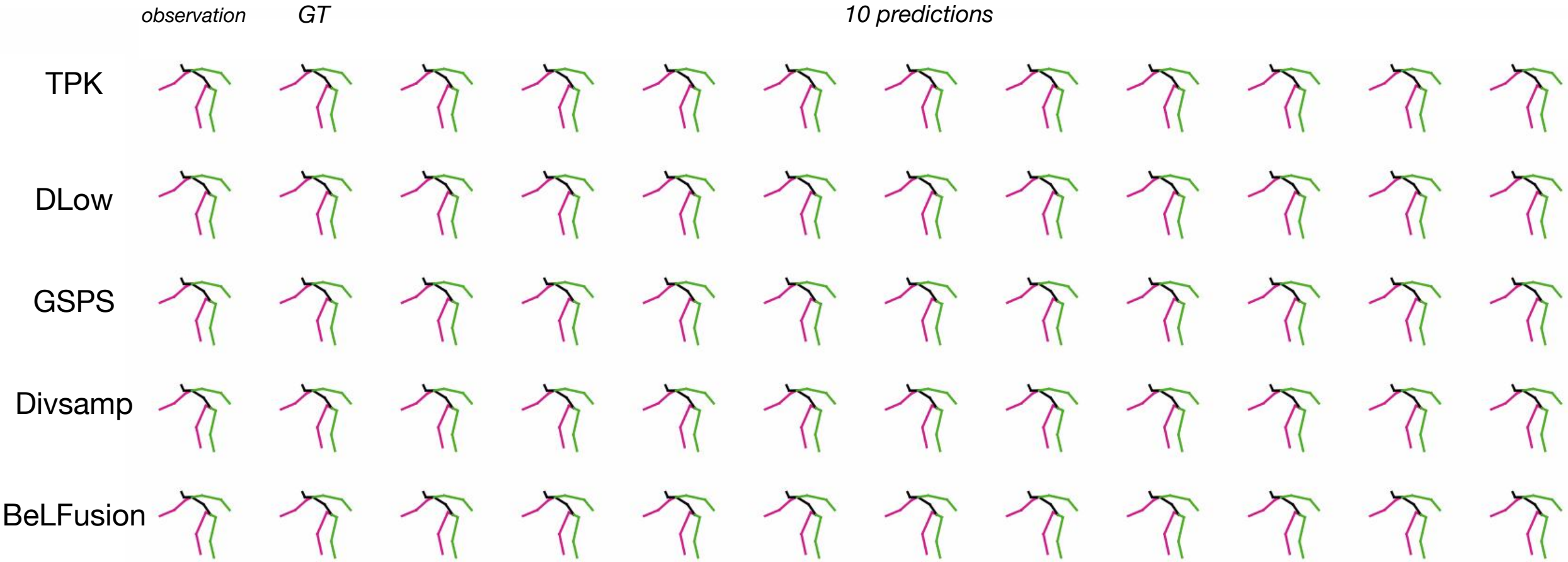
Experiments

☐ Visualization results



Experiments

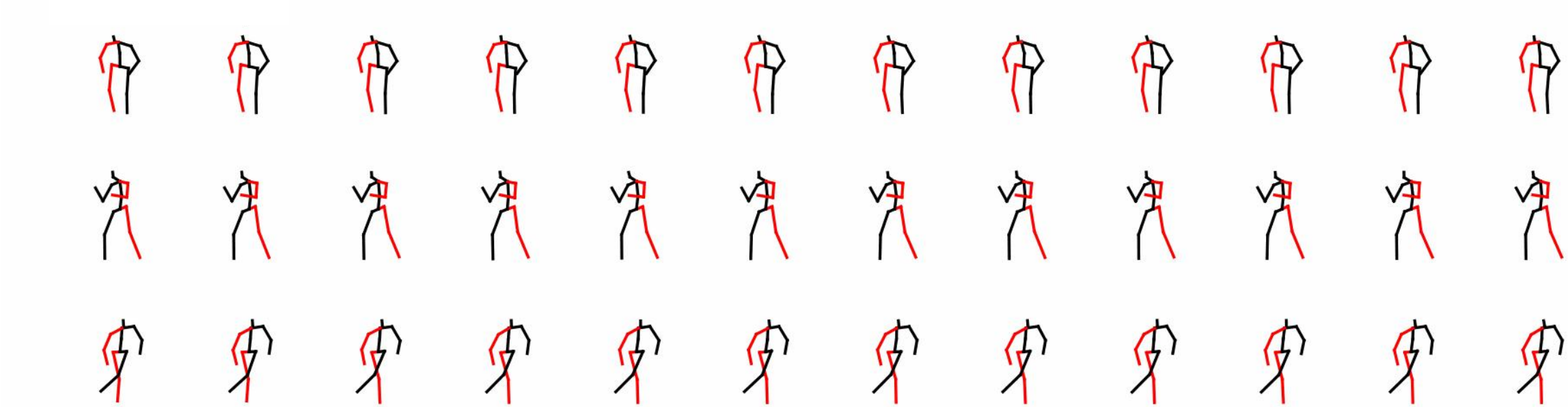
☐ Visualization results (WalkDog)



Experiments

☐ Visualization results (WalkDog)

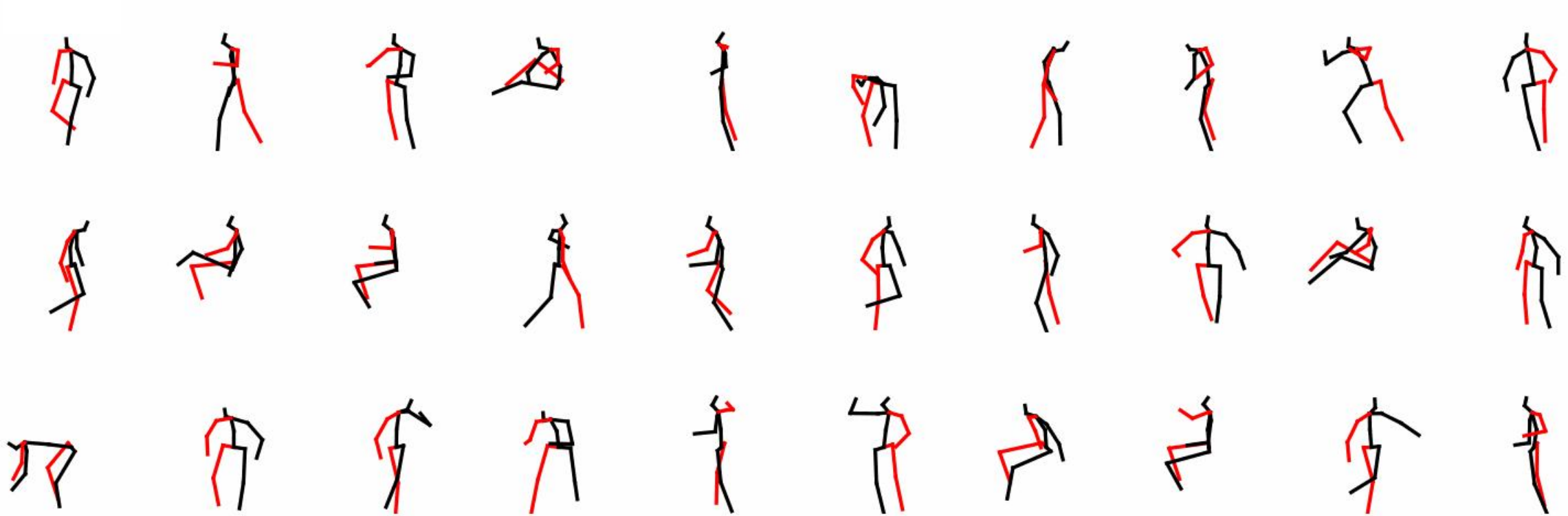
observation *GT* *10 predictions*



HumanMAC (Ours)

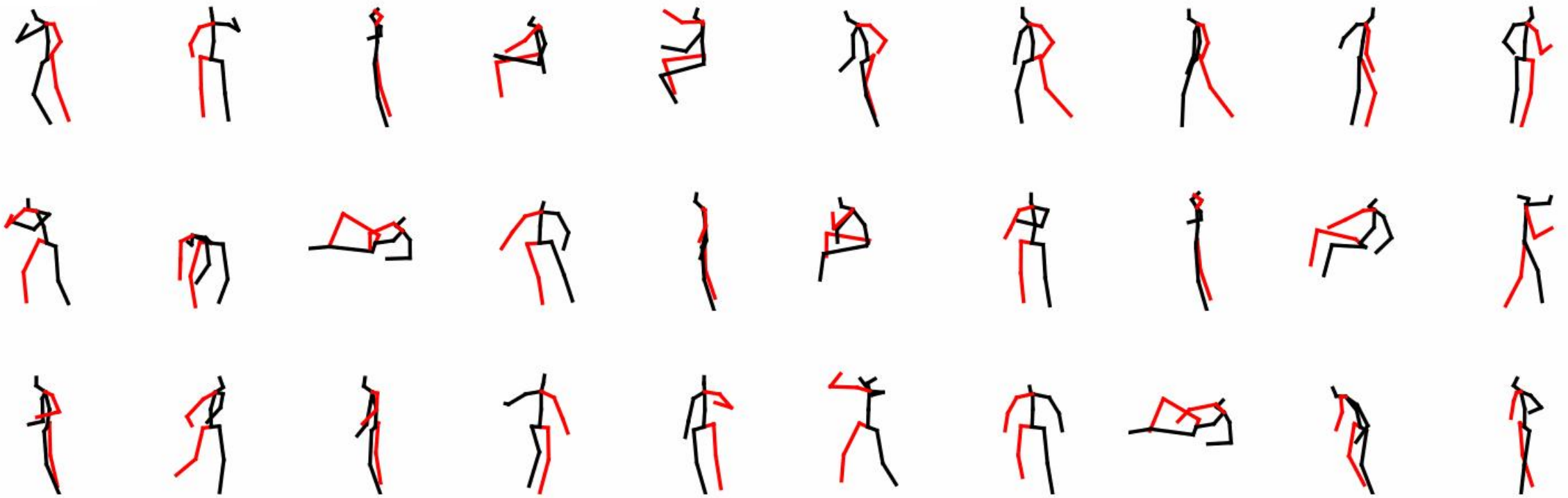
Experiments

- **Motion Switch Ability**



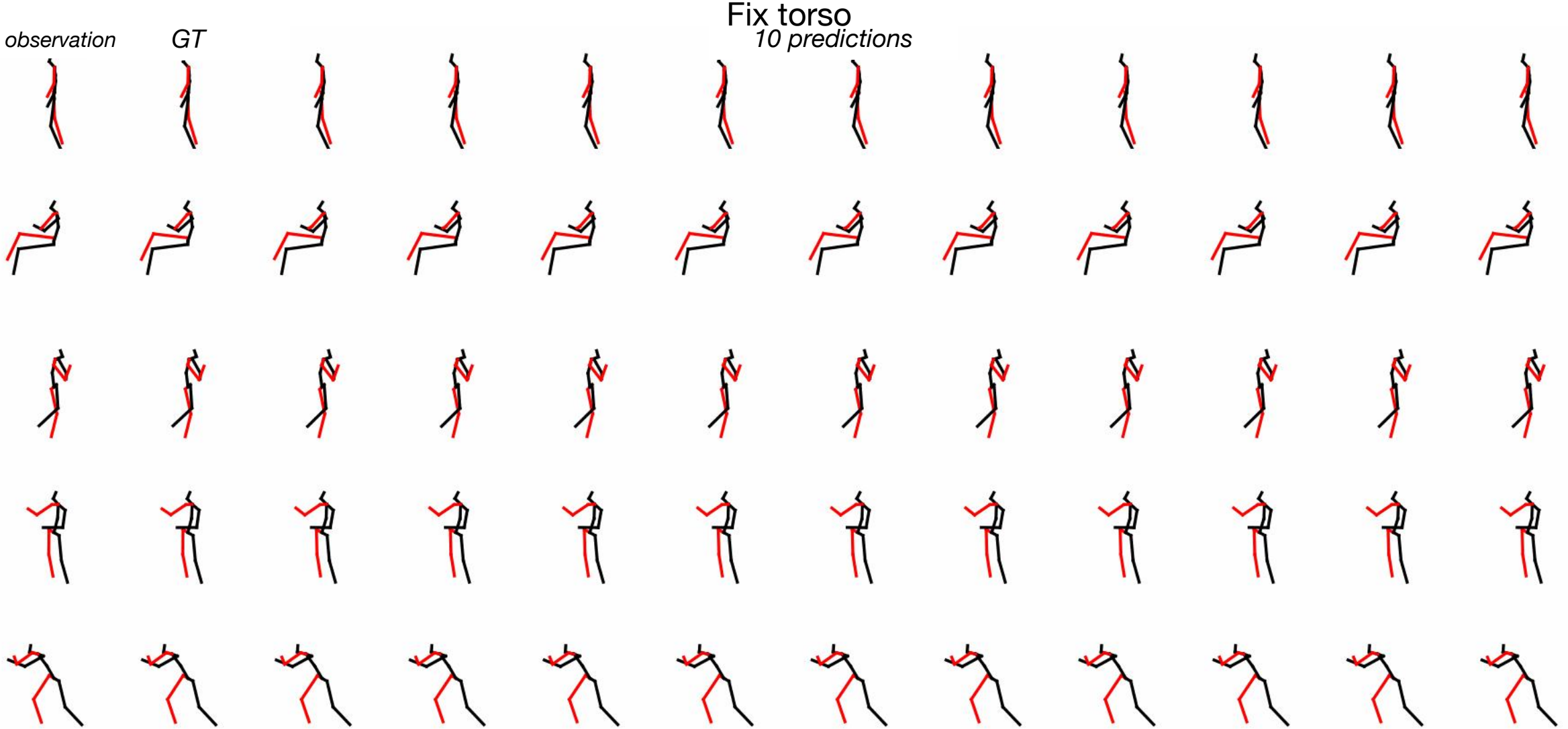
Experiments

- **Motion Switch Ability**



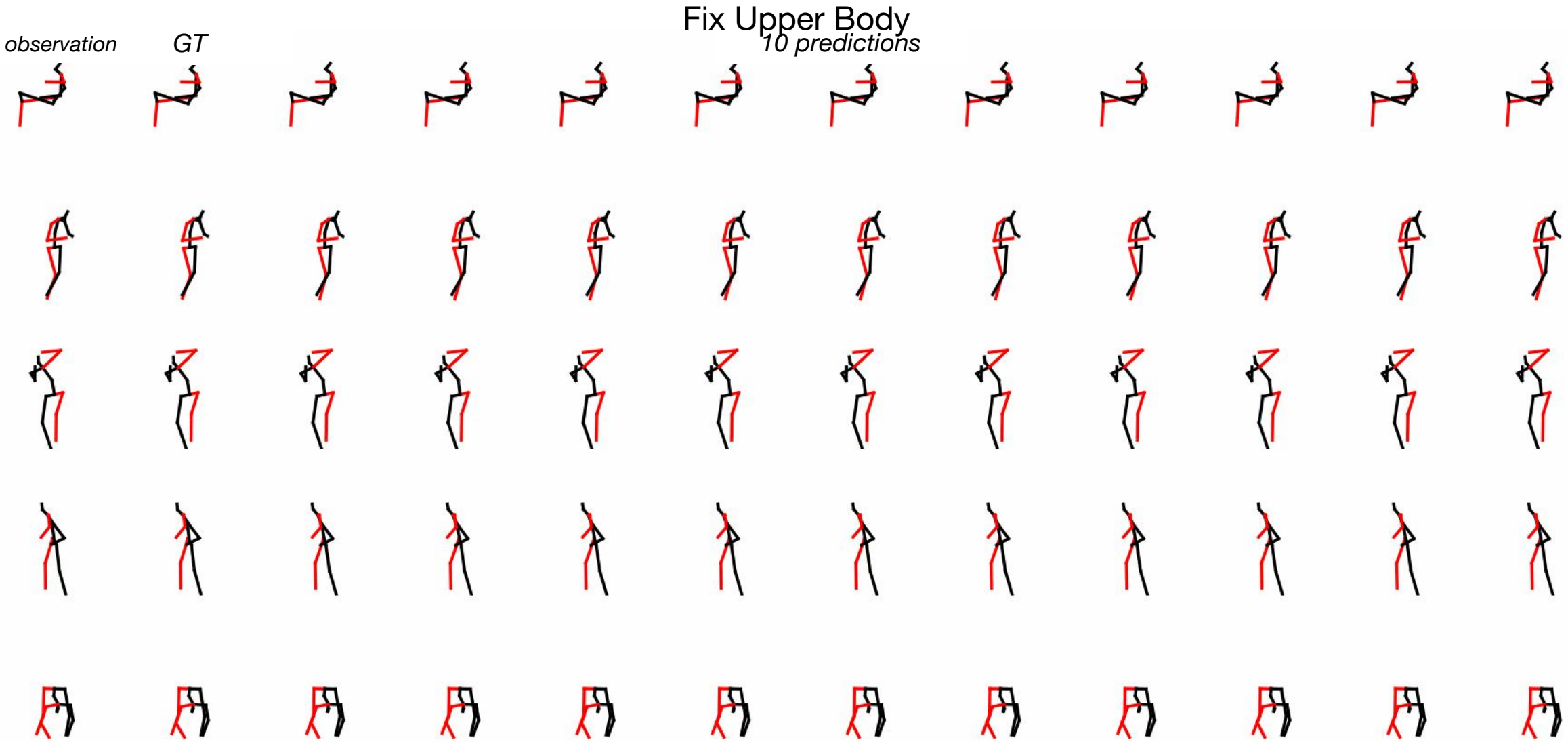
Experiments

☐ Part-body Controllable Prediction



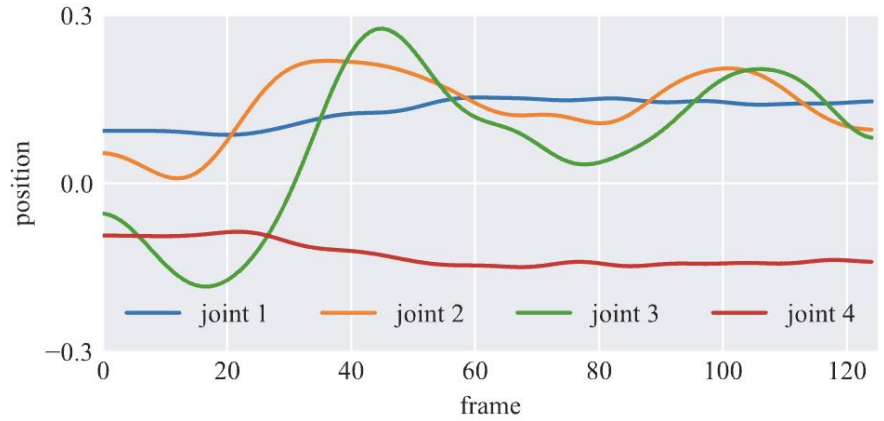
Experiments

☐ Part-body Controllable Prediction

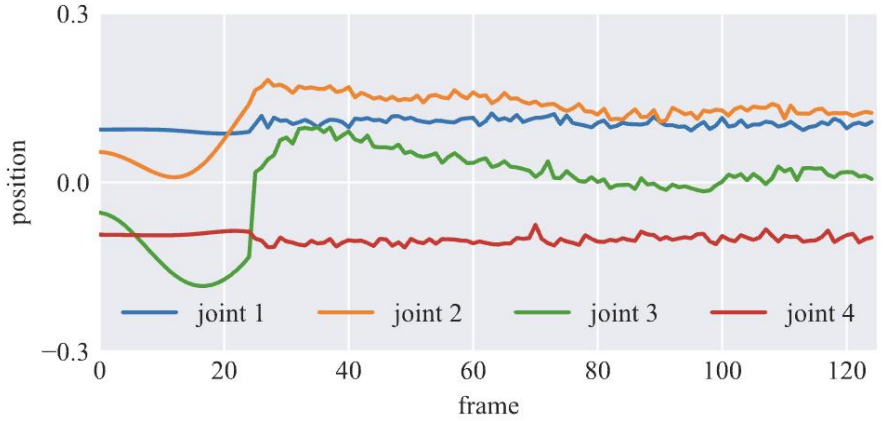


Experiments

◘ Ablation



(a) w/ DCT.

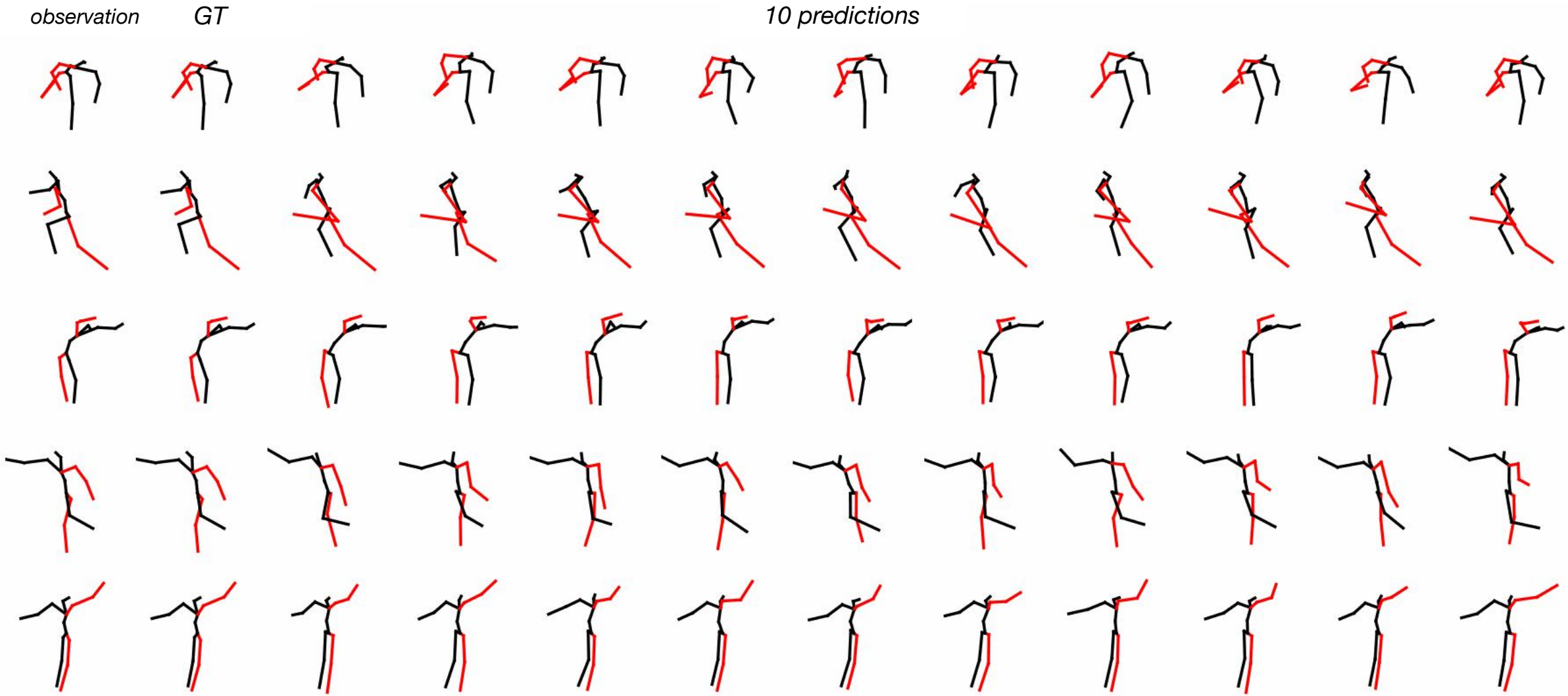


(b) w/o DCT.

	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
w/o DCT	**7.191**	0.444	0.521	0.521	0.550
w/ DCT	6.301	**0.369**	**0.480**	**0.509**	**0.545**

Experiments

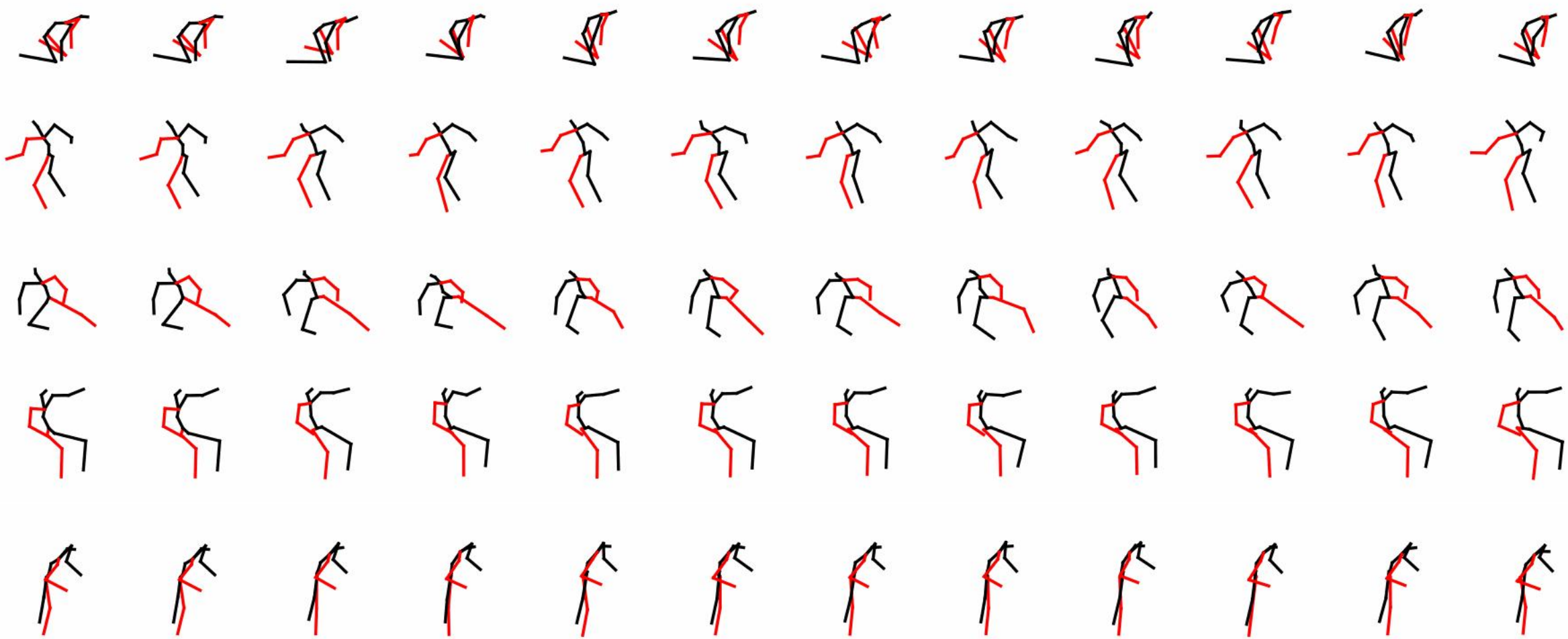
◘ Zero-shot Motion Prediction



Experiments

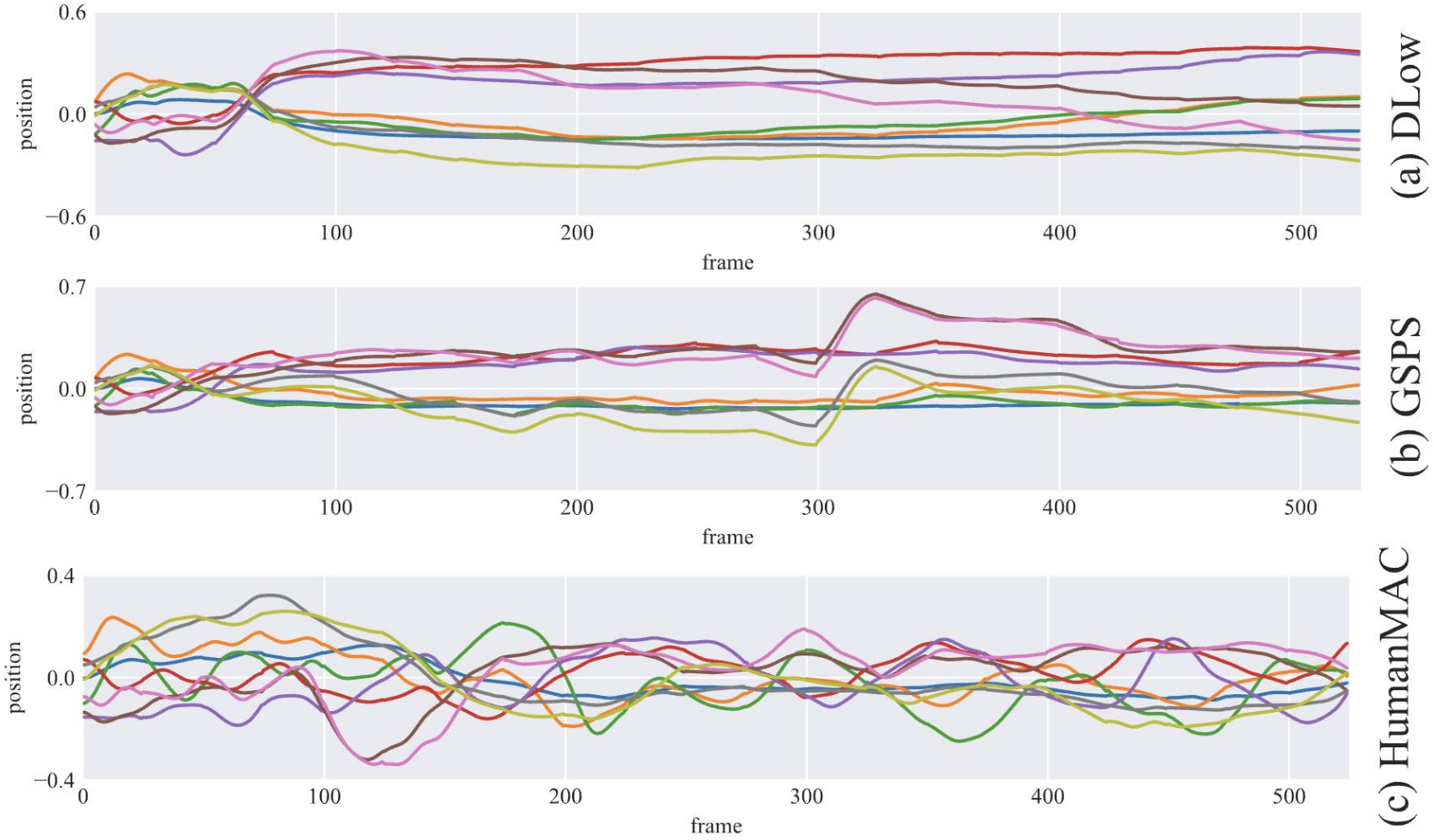
◘ Zero-shot Motion Prediction

observation *GT* *10 predictions*



Experiments

- **Long time series prediction**



HumanMAC: Masked Motion Completion for Human Motion Prediction

https://lhchen.top/Human-MAC

Ling-Hao Chen[1*], Jiawei Zhang[2*], Yewen Li[3], Yiren Pang[2], Xiaobo Xia[4], and Tongliang Liu[4]

For commercial, research, or co-operation purpose, please contact at: thu.lhchen@gmail.com.

[1]Tsinghua University, [2]Xidian University

[3]Nanyang Technological University, [4]The University of Sydney

Watch it!

Try it!