Diverse Human Motion Prediction Guided by Multi-Level Spatial-Temporal Anchors Supplementary Material

Sirui Xu, Yu-Xiong Wang*, and Liang-Yan Gui*

University of Illinois at Urbana-Champaign
{siruixu2, yxw, lgui}@illinois.edu

In this supplementary material, we first provide a visualization video for additional qualitative comparisons. Please refer to the video[1] and Sec. A for details. In Sec. B, we demonstrate the additional information of our approach including the network architecture, and implementation details for both stochastic and deterministic prediction. In Sec. C, for completeness, we provide additional qualitative results and comparisons, including a user study, and quantitative analysis for both stochastic and deterministic predictions on Human3.6M and HumanEva-I.

A Visualization Video

We include a video here to provide more comprehensive visualizations of 3D human motion prediction. These visualizations show that indeed our approach produces more diverse sequences, which we attribute to our STARS strategy, where anchors can explicitly locate diverse modes. We further show the visualization of controllable motion prediction and illustrate the motion variation at both spatial and temporal levels, suggesting our novel manipulation of future motion in the *native space* and *time*.

B Additional Details of Methodology

B.1 Multi-Level Spatial-Temporal Anchor-Based Sampling

Architecture. Here we detail our STARS w/ IE-STGCN by formulating the incorporation of backbone and anchor-based sampling, as illustrated in Fig. 3 of the main paper. For instance, we sample $\mathbf{z} \in p(\mathbf{z})$ and select i-th spatial anchor $\mathbf{a}_i^s \in \mathbb{R}^{M \times V \times C^{(l)}}$ and j-th temporal anchor $\mathbf{a}_j^t \in \mathbb{R}^{M \times V \times C^{(l)}}$ at each level, where (i, j) is the 2D spatial-temporal anchor index corresponding to the 1D index k. (1) The 4th layer in Fig. 3 of the main paper is denoted as

$$\mathbf{a}_i^{s_1} \in \mathcal{A}_s^{(1)}, \mathbf{a}_j^{t_1} \in \mathcal{A}_t^{(1)}, \ \mathbf{H}_k^{(4)} = \sigma(\mathbf{Adj}_s^{(3)} \mathbf{Adj}_f^{(3)} (\mathbf{H}^{(3)} + \mathbf{a}_i^{s_1} + \mathbf{a}_j^{t_1}) \mathbf{W}^{(3)}). \quad (1)$$

* Yu-Xiong Wang and Liang-Yan Gui contributed equally to this work.
[1] Video: https://youtu.be/ibYfsvCg7tQ

(2) The 5th layer in Fig. 3 of the main paper is denoted as

$$\mathbf{z} \sim p(\mathbf{z}),\ \mathbf{H}_k^{(5)} = \sigma((\mathbf{M}_s \odot \mathbf{Adj}_s^{(4)})\mathbf{Adj}_f^{(4)}[\mathbf{H}_k^{(4)} : \mathbf{z}]\mathbf{W}^{(4)}), \tag{2}$$

where the 5th layer is the pruned layer, and $\mathbf{M}_s$ is the predefined mask used for spatial interaction pruning.

(3) The 6th layer in Fig. 3 of the main paper is denoted as

$$\mathbf{a}_i^{s_2} \in \mathcal{A}_s^{(2)}, \mathbf{a}_j^{t_2} \in \mathcal{A}_t^{(2)},\ \mathbf{H}_k^{(6)} = \sigma(\mathbf{Adj}_s^{(5)}\mathbf{Adj}_f^{(5)}(\mathbf{H}_k^{(5)} + \mathbf{a}_i^{s_2} + \mathbf{a}_j^{t_2})\mathbf{W}^{(5)}). \tag{3}$$

Training. Recall that in Sec. 3 of the main paper, we divide the loss functions into the following three categories: (1) reconstruction losses, including reconstruction error and multi-modal reconstruction error; (2) diversity promoting loss; (3) motion constraint losses, including history reconstruction error, pose prior, limb loss, and angle loss. Here, we provide detailed formulations of these loss functions.

(1) Reconstruction error, encouraging the best prediction close to the ground truth, thus prompting the corresponding anchor capture a mode, denoted as

$$\mathcal{L}_r = \min_k \|\widehat{\mathbf{Y}}_k - \mathbf{Y}\|^2. \tag{4}$$

(2) Multi-modal reconstruction error [4], which encourages predictions to cover multi-modal ground truth and thus promoting anchors to capture more modes, is denoted as

$$\mathcal{L}_{mm} = \frac{1}{N}\sum_{n=1}^{N} \min_k \|\widehat{\mathbf{Y}}_k - \mathbf{Y}_n\|^2. \tag{5}$$

The multi-modal ground truth [11] is defined as $\{\mathbf{Y}_n\}_{n=1}^N$, representing the possible future motions in multiple modes. Specifically, given a threshold ϵ, we cluster the future motions with similar start pose, as $\{\mathbf{Y}_n\}_{n=1}^N = \{\mathbf{Y}_n | \|\mathbf{X}_n[T_h] - \mathbf{X}[T_h]\| \leq \epsilon\}$, where $\mathbf{X}_n$ is the historical pose sequence of $\mathbf{Y}_n$.

(3) Historical reconstruction error [5] alleviates the discontinuity between prediction and history by bringing the recovered historical motion $\widehat{\mathbf{X}}_k$ close to the past sequence of ground truth $\mathbf{X}$. Recall that our model recovers the past motion via Inverse DCT (IDCT) as $\widehat{\mathbf{X}}_k = (\mathbf{C}^\mathsf{T}\widetilde{\mathbf{Y}}_k)_{1:T_h}$. We denote this loss as

$$\mathcal{L}_h = \frac{1}{K}\sum_{k=1}^{K} \|\widehat{\mathbf{X}}_k - \mathbf{X}\|^2. \tag{6}$$

(4) Diversity-promoting loss [12], which explicitly promotes pairwise distances of predictions to ensure that the anchors do not collapse to the same, is defined

as

$$\mathcal{L}_d = \frac{2}{K(K-1)} \sum_{j=1}^{K} \sum_{k=j+1}^{K} \exp(-\frac{\|\widehat{\mathbf{Y}}_j - \widehat{\mathbf{Y}}_k\|_1}{\alpha}). \tag{7}$$

(5) Pose prior, using a pretrained normalizing flow p_{nf} to measure the likelihood of the generated human poses $\widehat{\mathbf{Y}}_k$. We use this module to constrain that the generated poses have a high probability in p_{nf},

$$\mathcal{L}_{nf} = -\sum_{k=1}^{K} \log p_{nf}(\widehat{\mathbf{Y}}_k). \tag{8}$$

(6) Limb loss, constraining the limb length to be consistent with the ground truth, is denoted as

$$\mathcal{L}_l = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{\mathbf{L}}_k - \mathbf{L}\|^2, \tag{9}$$

where the limb length is defined as the distance between two physically connected joints, and the vector $\widehat{\mathbf{L}}_k$ includes limb lengths of all poses in $\widehat{\mathbf{X}}_k$.

(7) Angle loss constrains the angles of human skeleton to be in some valid ranges. Please refer to [4] for more details on the pose prior, limb loss, and angle loss.

Additional Implementation Details. First, the number of channels of the 8 STGCN layers $C^{(l)}$ starts from $C^{(0)} = 3$, then $128, 64, 128, 64, 128, 64, 128$, and finally $C^{(8)} = 3$. Accordingly, we insert 128-dimensional anchors in the fourth and sixth layers and use a 64-dimensional random noise in the fifth layer. Our code is based on PyTorch [6], and we use ADAM [3] to train the model. The learning rate is set to 0.001 and decayed after the 100 epochs as

$$lr = 0.001 \times (1.0 - \frac{\max(0, epoch - 100)}{400}). \tag{10}$$

For Human3.6M, the weight of each loss term $(\lambda_r, \lambda_{mm}, \lambda_h, \lambda_d, \lambda_{nf}, \lambda_l, \lambda_a)$ is $(2, 1, 50, 160, 0.01, 500, 100)$. And the first 20 DCT coefficients are used.

For HumanEva-I, the weight of each loss term $(\lambda_r, \lambda_{mm}, \lambda_h, \lambda_d, \lambda_{nf}, \lambda_l, \lambda_a)$ is $(2, 1, 10, 32, 0.002, 50, 10)$. Only the first 8 DCT coefficients are used.

We set the distance threshold ϵ for generating multi-modal ground truth mentioned above to 0.5 on both datasets.

Limitation. One potential limitation of our STARS is that the number of spatial and temporal anchors is decided manually via cross-validation. We conduct several ablation studies in Fig. 4 of the main paper to investigate the impact of anchor number. However, a better strategy might be learning it directly from data.

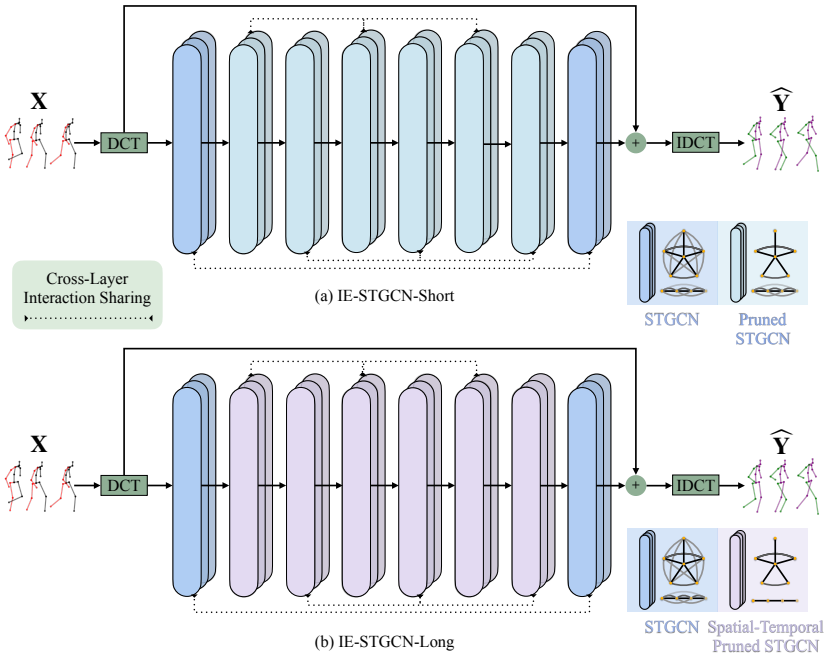

Fig. I: Overview of deterministic models. For IE-STGCN-Short, we set all middle layers as pruned STGCN layers, while we further apply temporal pruning to intermediate pruned layers for IE-STGCN-Long

B.2 Deterministic Human Motion Prediction

Architecture. We illustrate our architecture for deterministic prediction *without* STARS, as shown in Fig. I. Note that there are only 4 pruned STGCN layers for stochastic prediction model. Here, we design all middle layers to be the pruned layers from layer 2 to layer 7. For IE-STGCN-Long, we further apply temporal interaction pruning to the these middle layers.

Temporal Interaction Pruning. We emphasize the locality of frequency components by leveraging a temporal mask $\mathbf{M}_f$ to the frequency adjacency matrix of the spatial-temporal graph,

$$\widehat{\mathbf{Adj}}_f^{(l)} = \mathbf{M}_f \odot \mathbf{Adj}_f^{(l)}, \ \mathbf{M}_f[i][j] = \begin{cases} 1, & \text{for } |f_i - f_j| = 1, v_i = v_j \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Training. We retain only two loss terms from Sec. B.1 *i.e.*, , the reconstruction error, and the history reconstruction error. We denote the unique output as $\widehat{\mathbf{Y}}$, and the recovered historical motion as $\widehat{\mathbf{X}}$. These two loss terms are adapted as follows.

Table I: Quantitative results on Human3.6M and HumanEva-I for $K = 50$. Our model significantly outperforms two additional baselines in all metrics. The results of two baselines are directly reported from [7, 10]

Method	Human3.6M [2]					HumanEva-I [8]				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
ProTran [10]	/	0.381	0.491	/	/	/	0.258	0.255	/	/
Motron [7]	7.168	0.375	0.488	/	/	/	/	/	/	/
STARS (Ours)	**15.884**	**0.358**	**0.445**	**0.442**	**0.471**	**6.031**	**0.217**	**0.241**	**0.328**	**0.321**

Table II: Quantitative results of short-term prediction on Human3.6M. We compare our IE-STGCN-Short with deterministic prediction baselines. We report the MPJPE error of 3D joint positions in *millimeter*. Our model outperforms all baselines

Actions	Walking				Eating				Smoking				Discussion			
msec	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*
LTD [5]	11.1	21.2	37.4	43.6	7.3	15.1	29.6	36.8	7.1	14.9	29.5	36.2	11.5	25.5	55.8	69.2
STS-GCN [9]	14.5	26.0	43.8	51.0	9.2	18.0	35.5	43.3	9.5	18.1	34.9	42.3	13.3	27.9	57.5	71.9
MSR-GCN [1]	10.8	20.9	36.9	42.4	6.9	14.6	29.0	36.0	7.5	15.4	30.6	37.5	10.4	23.5	51.9	65.0
IE-STGCN-Short	**10.0**	**19.4**	**35.1**	**41.6**	**6.3**	**13.8**	**28.3**	**35.9**	**6.4**	**13.1**	**25.8**	**32.3**	**9.0**	**19.9**	**42.8**	**55.2**

Actions	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
msec	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*
LTD [5]	8.3	19.1	43.5	**54.5**	14.6	30.8	64.5	79.9	9.0	18.6	39.3	49.2	11.2	25.8	59.6	76.5	13.7	30.2	63.4	77.7	9.8	20.3	44.1	55.7
STS-GCN [9]	10.2	23.0	50.6	63.2	17.0	36.6	72.5	86.4	11.0	21.9	44.0	53.8	13.7	30.4	67.3	84.7	16.3	35.9	70.5	83.1	11.9	23.8	49.3	60.8
MSR-GCN [1]	7.7	18.9	44.7	56.2	15.1	33.1	70.9	85.4	9.1	18.9	39.9	50.0	10.3	24.6	59.2	75.9	13.3	30.1	63.6	77.8	9.8	20.6	44.2	55.5
IE-STGCN-Short	**7.1**	**17.8**	**43.1**	54.7	**12.8**	**28.5**	**61.1**	**75.9**	**8.2**	**17.5**	**37.3**	**47.3**	**8.3**	**18.9**	**43.6**	**57.7**	**12.2**	**28.3**	**60.0**	**74.1**	**8.9**	**19.4**	**42.9**	**54.4**

Actions	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				**Average**			
msec	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*	*80*	*160*	*320*	*400*
LTD [5]	14.8	**29.5**	**57.2**	**71.2**	9.1	19.1	41.1	51.8	9.4	19.7	43.2	54.6	21.1	41.2	75.1	88.8	9.6	19.2	36.0	43.1	11.2	23.3	47.9	59.3
STS-GCN [9]	18.2	37.2	66.2	79.4	10.8	22.3	47.7	59.4	11.7	24.0	49.6	62.0	24.3	48.0	85.1	97.3	11.7	22.7	41.7	49.1	13.5	27.7	54.4	65.8
MSR-GCN [1]	15.4	32.0	60.7	73.8	8.9	19.5	43.1	54.4	10.4	22.4	50.7	62.4	24.9	51.5	100.3	112.9	9.2	18.7	35.7	43.2	11.3	24.3	50.8	61.9
IE-STGCN-Short	**14.6**	30.8	60.2	72.8	**7.9**	**17.8**	**39.7**	**50.3**	**7.9**	**17.6**	**40.2**	**51.6**	**18.4**	**38.2**	**74.4**	**87.9**	**8.3**	**17.2**	**33.3**	**41.2**	**9.7**	**21.2**	**44.5**	**55.5**

(1) Reconstruction error:

$$\mathcal{L}_r = \min_k \|\widehat{\mathbf{Y}}_k - \mathbf{Y}\|^2 = \|\widehat{\mathbf{Y}} - \mathbf{Y}\|^2. \tag{12}$$

(2) History reconstruction error:

$$\mathcal{L}_h = \frac{1}{K}\sum_{k=1}^{K} \|\widehat{\mathbf{X}}_k - \mathbf{X}\|^2 = \|\widehat{\mathbf{X}} - \mathbf{X}\|^2. \tag{13}$$

Implementation Details. We adopt the same number of channels as demonstrated in Sec. B.1. We increase the batch size to 256 and train the model for only 50 epochs. The learning rate is set to 0.01 and decays by a factor of 0.1 every 5 epochs after the 20th epoch. For IE-STGCN-Short, we use the first 20 DCT coefficients. For IE-STGCN-Long, we use the first 35 DCT coefficients.

Table III: User study on Human3.6M. Pairwise human voting results for predicted motions. Under human evaluation, our predictions significantly outperform the baseline in terms of diversity, considering the motion fidelity

Model pair	Motion diversity	
	Ours	GSPS
STARS (Ours) vs.	n/a	62.1%
GSPS vs.	37.9%	n/a

Table IV: Quantitative results of long-term prediction on Human3.6M. We compare our IE-STGCN-Long with deterministic prediction baselines. We report the MPJPE error of 3D joint positions in *millimeter*. Our model outperforms all baselines

Actions	Walking				Eating				Smoking				Discussion			
msec	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*
LTD [5]	52.3	55.7	58.1	**59.2**	**49.9**	**60.9**	69.2	74.2	50.1	59.8	67.4	72.1	90.7	105.4	114.9	120.4
STS-GCN [9]	60.3	64.6	65.9	70.2	57.2	68.3	75.5	82.6	54.2	63.8	70.8	76.1	91.8	105.2	113.8	118.9
MSR-GCN [1]	53.3	55.4	58.1	63.7	50.8	61.4	69.7	75.4	50.5	59.5	67.1	72.1	87.0	101.9	111.4	116.8
IE-STGCN-Long	**49.3**	**53.5**	**57.4**	61.1	50.2	61.1	**69.1**	**74.1**	**44.2**	**51.8**	**59.0**	**64.3**	**74.0**	**85.1**	**94.1**	**100.4**

Actions	Directions				Greeting				Phoning				Posing				Purchases				Sitting			
msec	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*
LTD [5]	76.0	91.2	103.0	108.8	105.0	120.6	133.2	139.4	67.9	**82.2**	95.0	**103.3**	111.2	137.3	159.2	172.8	100.3	115.2	127.7	135.5	79.5	98.4	113.5	122.8
STS-GCN [9]	79.5	92.9	102.2	109.6	111.2	122.4	131.8	136.1	72.5	87.9	99.7	108.3	115.8	142.4	161.7	178.4	104.6	119.4	132.7	141.0	82.0	97.6	110.9	121.4
MSR-GCN [1]	**75.8**	**89.9**	**100.5**	**105.9**	106.3	120.0	131.5	136.3	67.9	82.5	95.8	104.7	112.5	140.1	162.8	176.5	**99.2**	**114.0**	**126.9**	**134.4**	77.6	94.0	107.7	115.9
IE-STGCN-Long	76.5	91.6	102.4	107.6	**101.2**	**116.5**	**129.5**	**135.8**	**67.2**	**82.1**	95.1	103.8	**79.7**	**97.2**	**113.9**	**129.3**	100.3	117.5	131.1	139.4	**77.5**	95.0	109.1	117.6

Actions	Sitting Down				Taking Photo				Waiting				Walking Dog				Walking Together				**Average**			
msec	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*	*560*	*720*	*880*	*1000*
LTD [5]	**98.2**	**119.1**	**136.1**	147.1	**76.8**	**95.0**	**110.3**	**120.4**	76.8	91.0	102.3	109.5	108.3	121.2	135.8	146.3	56.3	61.9	65.5	68.2	79.9	94.3	106.1	113.3
STS-GCN [9]	104.1	121.4	137.6	148.4	81.2	99.6	111.6	126.3	80.3	95.0	105.9	113.6	119.0	129.0	143.9	151.5	61.9	65.4	69.1	72.5	85.0	98.3	108.9	117.0
MSR-GCN [1]	102.4	122.7	139.6	149.3	77.7	96.9	112.3	121.9	74.8	87.8	98.2	105.5	**107.7**	**120.8**	**135.7**	**145.7**	56.2	60.9	65.0	69.5	80.0	93.8	105.5	112.9
IE-STGCN-Long	100.9	120.7	136.6	**146.8**	78.1	96.9	112.1	122.1	**71.7**	**85.0**	**96.0**	**103.6**	111.9	126.3	142.9	153.1	**54.1**	**59.5**	**63.5**	**67.5**	**75.8**	**89.3**	**100.8**	**108.4**

C Additional Results

C.1 Quantitative Results

Comparison with Additional Baselines. Motron [7] provides a flexible output structure, which can produce deterministic predictions by weighting each mode by a confidence value. Another baseline is ProTran [10], a deep probabilistic method combining transformer architectures and state space models. However, these two baselines do not provide results for all standard metrics in the literature. In Table I, we compare our method with them. For all metrics reported, our method still consistently outperforms baselines.

C.2 Qualitative Results

User study on Human3.6M. In Tables 1 and 2 of the main paper, we compare our method with the state-of-the-art method GSPS [4]. We measure the quality of the best predictions and overall diversity according to the metrics demonstrated in Sec 4.1 of the main paper. Here, we conduct a user study to evaluate the diversity of predicted human motions under consideration of motion fidelity. The reason for this user study is that there may be unrealistic predictions, or

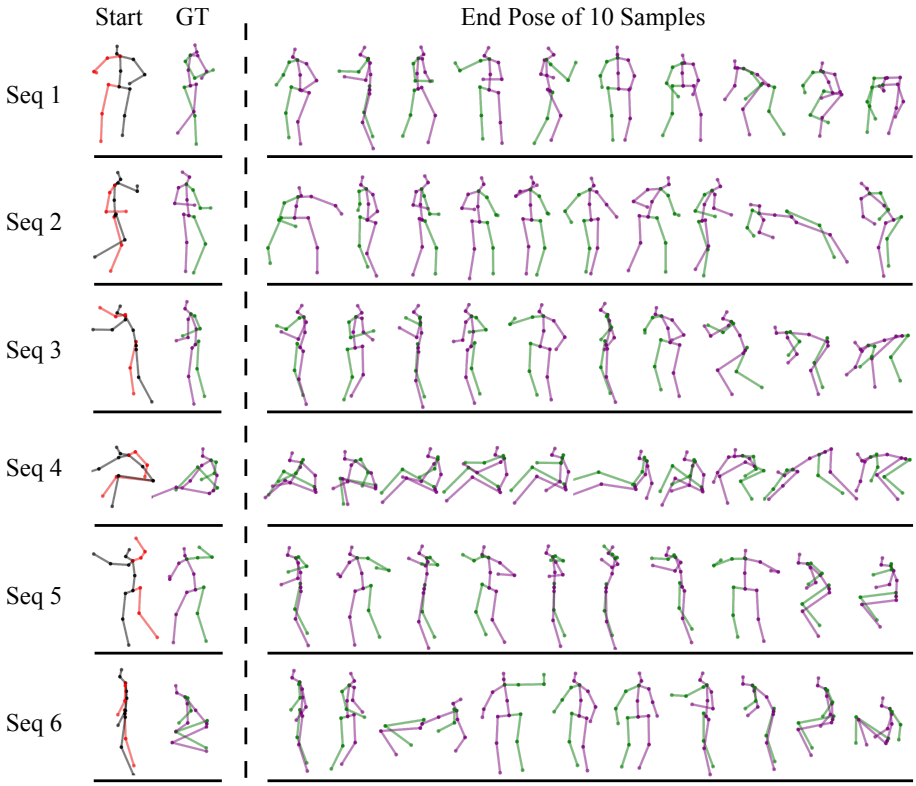

Fig. II: Additional visualization on Human3.6M. We show the start pose, the end pose of ground truth future motion, and the end poses of ten samples predicted by our approach

outliers, that result in very large APD, but do not affect the ADE since ADE measures only the quality of the best prediction. We evaluate and rule out such "cheating" behavior through user studies.

We conduct a *double-blind* user study. We randomly sample 20 input sequences on Human3.6M. For GSPS, we randomly sample two predicted sequences. For our STARS w/ IE-STGCN, we randomly sample two predictions *generated by different spatial-temporal anchors*. We design pairwise evaluations. Considering one pair from ours and another one from GSPS, human judges are asked to determine which pair is more diverse given the action labels, taking into account motion fidelity.

From the results of the human evaluations in Table III, our approach has a success rate of 62.1% against GSPS, verifying that our method generates more diverse and valid motions than the baseline method.

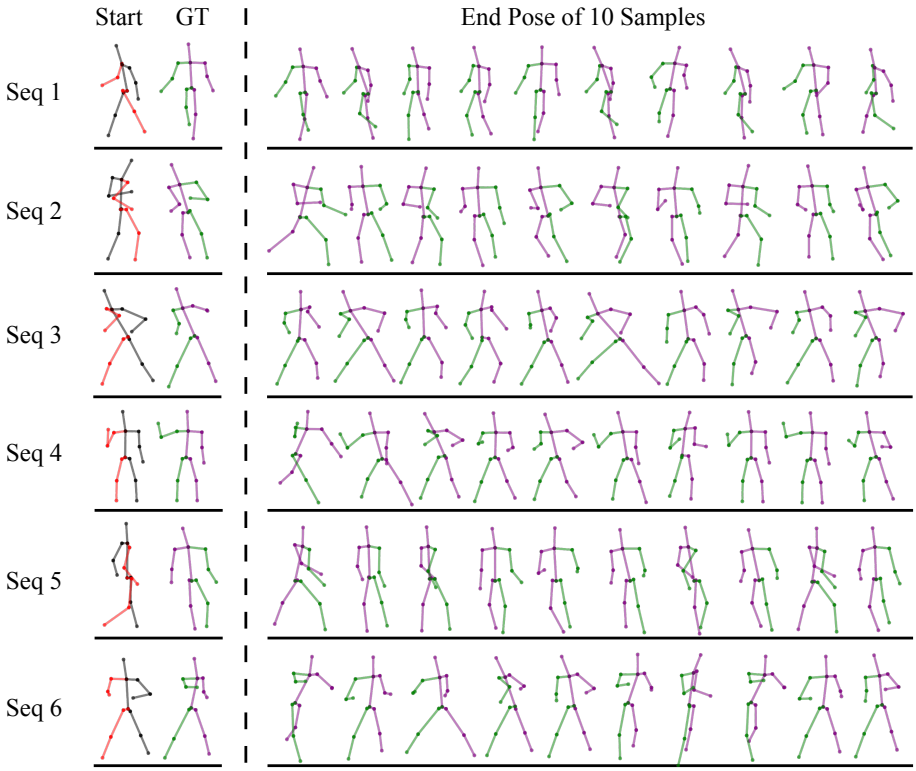

Fig. III: Additional visualization on HumanEva-I. We show the start pose, the end pose of ground truth future motion, and the end poses of ten samples predicted by our approach

Additional Visualizations on Human3.6M and HumanEva-I. In Fig. II and Fig. III, we provide additional qualitative results by visualizing the end poses of 10 samples.

Additional Qualitative Comparisons with GSPS Baseline. We provide comparisons with GSPS in Fig. IV, in addition to Fig. 5 of the main paper. Our model still successfully generates predictions that are closer to the ground truth.

Additional Visualizations on Controllable Motion Prediction. In addition to Fig. 7 of the main paper, we present the visualization of the controllable motion prediction in Fig. V.

C.3 Additional Results on Deterministic Prediction

Effectiveness on Deterministic Prediction. To evaluate the deterministic prediction, we randomly select 256 sequences for each action category. We *re-evaluate* the pretrained models provided by baseline approaches [1,5,9] with the

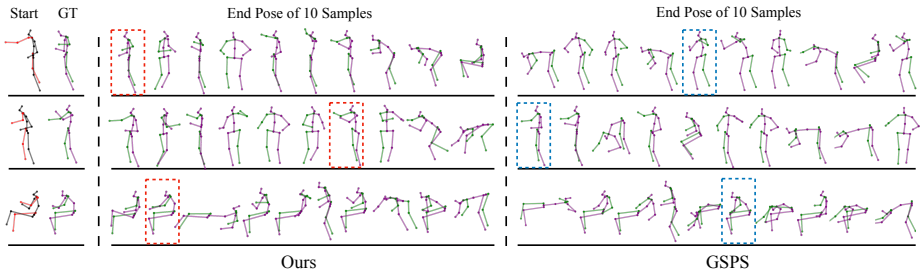

Fig. IV: Additional comparisons with the baseline on Human3.6M. As highlighted by the red and blue dashed boxes, the best predictions of our method are closer to the ground truth than the state-of-the-art baseline GSPS [4]

Table V: Average MPJPE error in mm on Human3.6M, comparing different spatial-temporal enhancement techniques. "TP" indicates whether we prune the temporal interaction. "TS" means if we share the temporal adjacency matrix across some of the layers (See Figure I). "TF" stands for start GCN and end GCN containing unpruned fully temporal interactions. "SP", "SS" and "SF" have the similar meaning but apply to temporal interaction

TP	TS	TF	SP	SS	SF	Short-term prediction				Long-term prediction			
						80	*160*	*320*	*400*	*560*	*720*	*880*	*1000*
		✓			✓	10.1	21.7	44.9	55.8	79.8	94.3	106.3	113.9
		✓	✓	✓	✓	9.8	21.3	44.8	55.8	77.0	91.1	102.7	110.1
✓			✓	✓	✓	10.2	22.1	46.2	57.2	77.1	90.8	101.9	109.6
	✓	✓	✓	✓	✓	**9.7**	**21.2**	**44.5**	**55.5**	77.1	91.1	102.6	110.1
✓	✓		✓	✓	✓	23.8	44.4	76.1	88.2	76.9	90.2	101.3	109.2
✓		✓	✓	✓	✓	10.0	21.8	45.7	56.9	75.8	89.3	100.8	108.4
✓	✓	✓	✓	✓	✓	10.0	21.8	45.7	56.9	**75.7**	89.4	**100.8**	108.5
✓		✓	✓	✓	✓	10.0	21.8	45.7	56.9	75.8	**89.3**	**100.8**	**108.4**
✓		✓	✓		✓	9.9	21.6	45.1	56.3	**75.7**	89.5	**100.8**	**108.4**
✓		✓		✓	✓	9.9	21.7	45.3	56.3	76.3	89.9	101.3	109.0
✓		✓	✓	✓		10.3	23.2	48.5	59.8	79.0	92.5	103.7	111.2

same selection of test data. Note that we re-evaluate STS-GCN using a standard evaluation metric for a fair comparison, reporting MPJPE in millimeters at each frame, rather than the average MPJPE over all frames as in [9].

Here, we provide detailed results and comparisons across all actions. As shown in Table II and Table IV, our model achieves state-of-the-art short- and long-term prediction performance.

Effectiveness of interaction enhancements. We conduct an ablation study in Table V to demonstrate the effectiveness of our proposed interaction enhancements in deterministic predictions. The results show that with proper use of interaction enhancements, we could improve the accuracy for both the short- and long-term deterministic predictions. We observe that the spatial interaction enhancements are effective for both short-term and long-term horizon, while temporal enhancements are helpful only for long-term prediction. The reason may be that in the long-term prediction, we use more DCT coefficients, which may contain redundancies that need to be pruned.

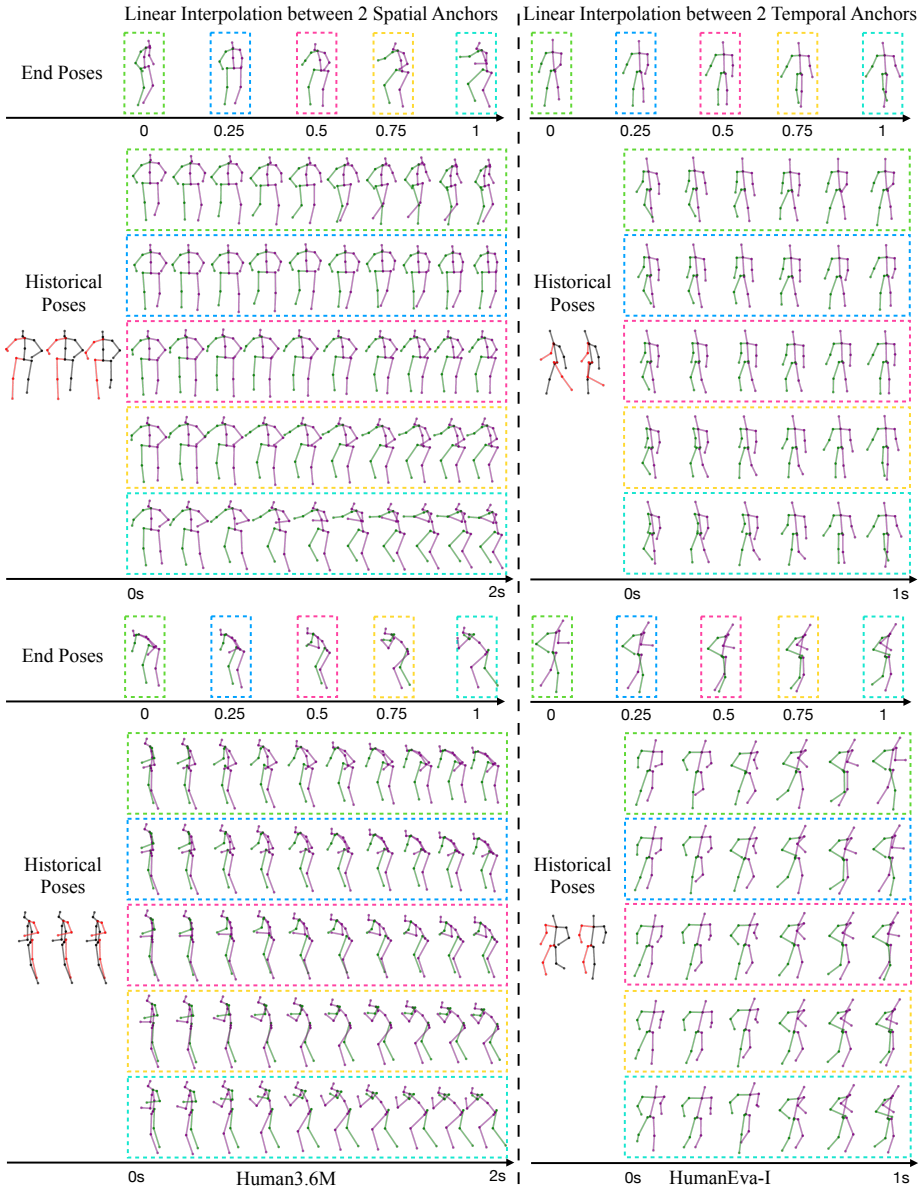

Fig. V: Linear interpolation of anchors. We show additional visualizations on Human3.6M and HumanEva-I. We provide seamless control over directions and speeds of future motion by linear interpolation of the spatial and temporal anchors

References

1. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11467–11476 (2021) 5, 6, 8
2. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2013) 5
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 3
4. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13309–13318 (2021) 2, 3, 6, 9
5. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9489–9497 (2019) 2, 5, 6, 8
6. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019) 3
7. Salzmann, T., Pavone, M., Ryll, M.: Motron: Multimodal probabilistic human motion forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6457–6466 (2022) 5, 6
8. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision **87**(1), 4–27 (2010) 5
9. Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-Time-Separable Graph Convolutional Network for pose forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11209–11218 (2021) 5, 6, 8, 9
10. Tang, B., Matteson, D.S.: Probabilistic transformer for time series analysis. Advances in Neural Information Processing Systems **34**, 23592–23608 (2021) 5, 6
11. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. arXiv preprint arXiv:1907.04967 (2019) 2
12. Yuan, Y., Kitani, K.: DLow: Diversifying latent flows for diverse human motion prediction. In: European Conference on Computer Vision. pp. 346–364 (2020) 2