ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Diverse Human Motion Prediction via Gumbel-Softmax Sampling from an Auxiliary Space

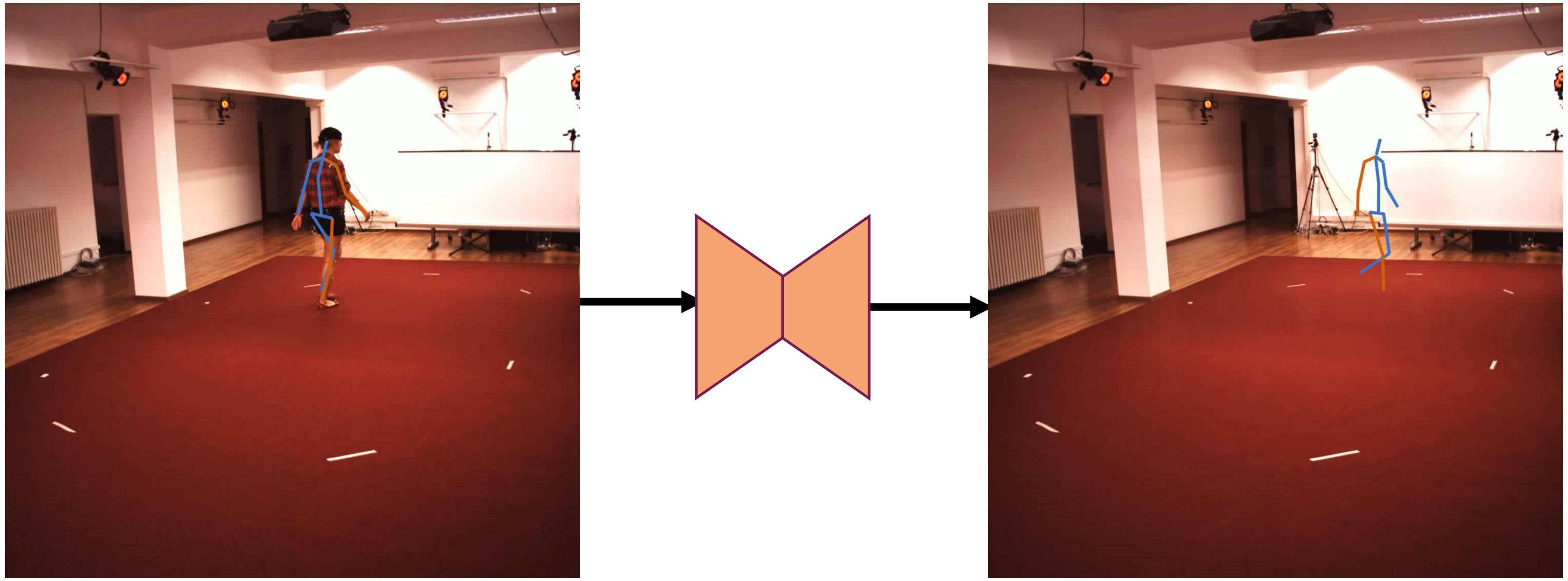
Lingwei Dang [1], Yongwei Nie [1*], Chengjiang Long [2], Qing Zhang [3], Guiqing Li [1]

[1] South China University of Technology, [2] Meta Reality Lab, [3] Sun Yat-sen University

csdanglw@mail.scut.edu.cn, {nieyongwei, ligq}@scut.edu.cn, clong1@fb.com, zhangqing.whu.cs@gmail.com

Human Motion Prediction (HMP)

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

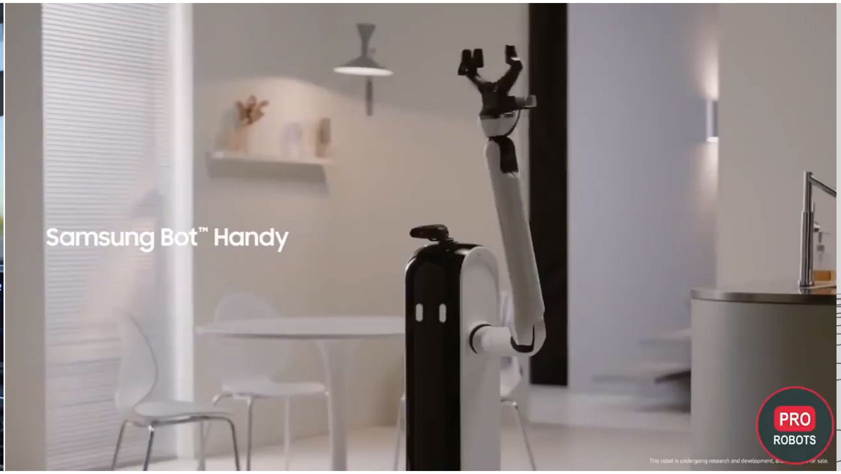


Observed Seq Future Seq

Applications

ACM
Multimedia 2022
Lisbon, Portugal | 10-14 October

autonomous driving



human-computer interaction

animation creation

Deterministic HMP

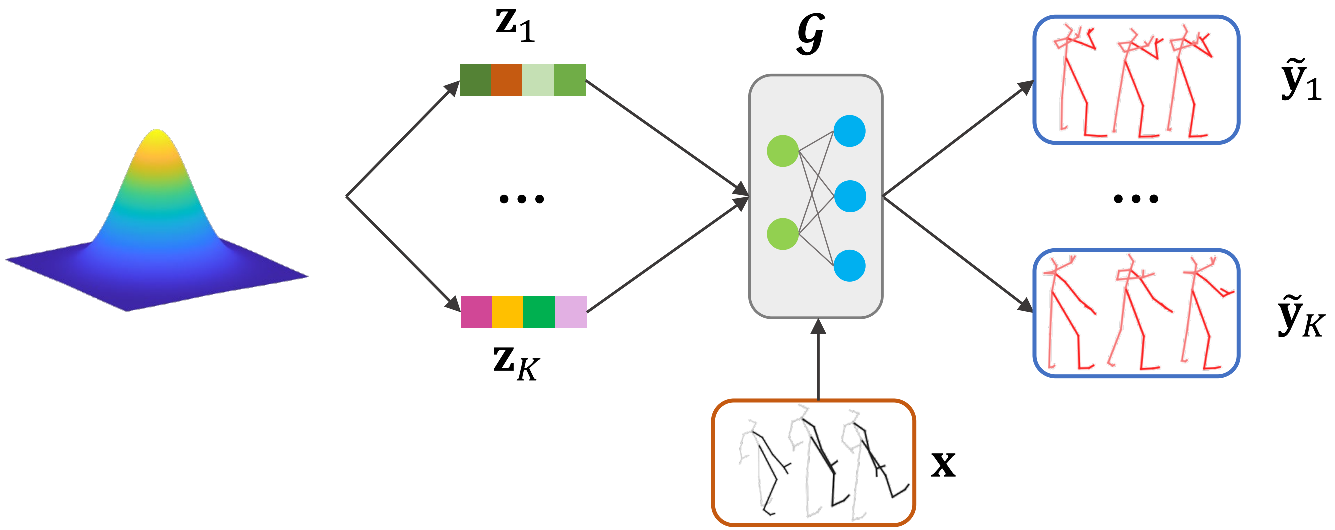
ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

- Predict only one motion in the future
- Ignore the uncertainty of human motion
- Cannot satisfy some safety-critical applications

LTD Mao W, *et al.*, ICCV 2019
HisReps Mao W, *et al.*, ECCV 2020
LDR Cui Q, *et al.*, CVPR 2020
MSR Dang L, *et al.*, ICCV 2021
MLMA Mao W, *et al.*, IJCV 2021
PGBIG Ma T, *et al.*, CVPR 2022

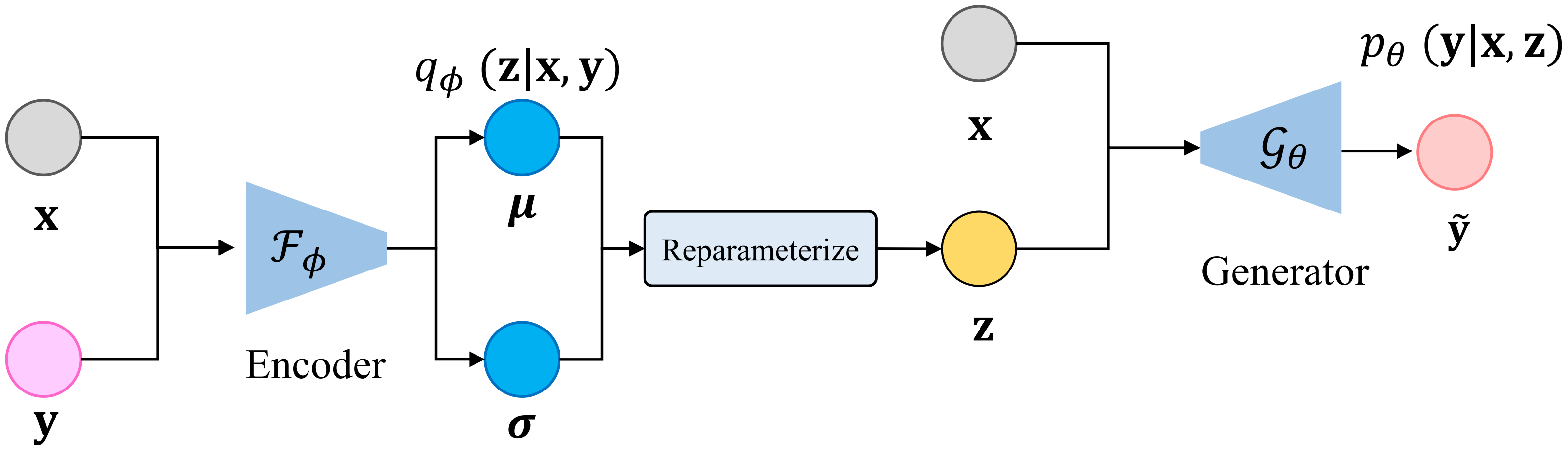
Stochastic and Diverse HMP

- Adopt deep generative networks, *e.g.* GAN or CVAE to model the conditional distribution of future poses given previous ones
- Sample random noises and decode them into multiple possible future motions



Hp-gan Barsoum E, *et al.*, CVPR 2018 **Bihmp-gan** Kundu J, *et al.*, AAAI 2019 **MNM** Li B, *et al.*, TIP 2020

Example: CVAE-based Approach



- Introduce an approximate posterior $q(\mathbf{z}|\mathbf{x},\mathbf{y})$ and maximizing the ELBO:

$$\log p(\mathbf{y} \mid \mathbf{x}) = \log \int p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} = \log \int \frac{p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z} \mid \mathbf{x}, \mathbf{y})} q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) d\mathbf{z} \geq \mathbb{E}_q \log \frac{p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z} \mid \mathbf{x}, \mathbf{y})}$$

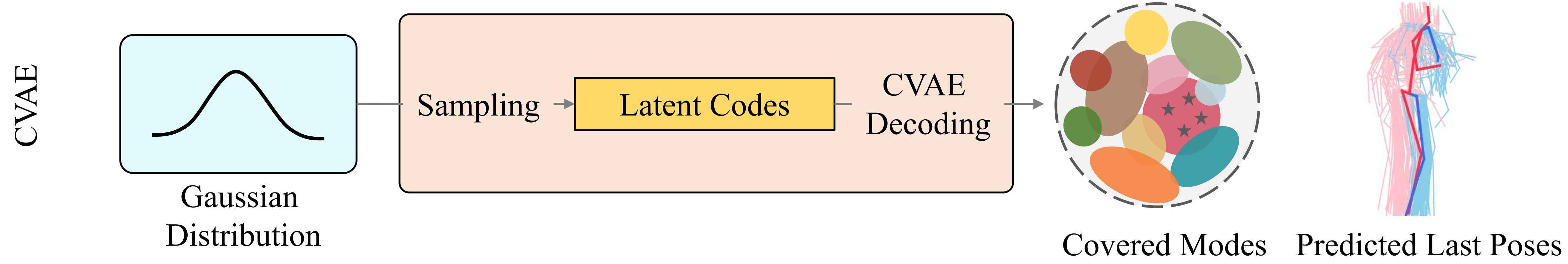
- Model the two distributions of $q\ (\mathbf{z}|\mathbf{x},\mathbf{y})$ and $p\ (\mathbf{y}|\mathbf{x},\mathbf{z})$ by two neural networks $\mathcal{F}_\phi$ and $\mathcal{G}_\theta$, and estimates their parameters by optimizing the following loss function

$$\mathcal{L}(\mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) = -\mathcal{KL}(q_\phi(\mathrm{z} \mid \mathbf{x}, \mathbf{y}) \| p(\mathrm{z})) + \mathbb{E}_{q_\phi} \log p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$$

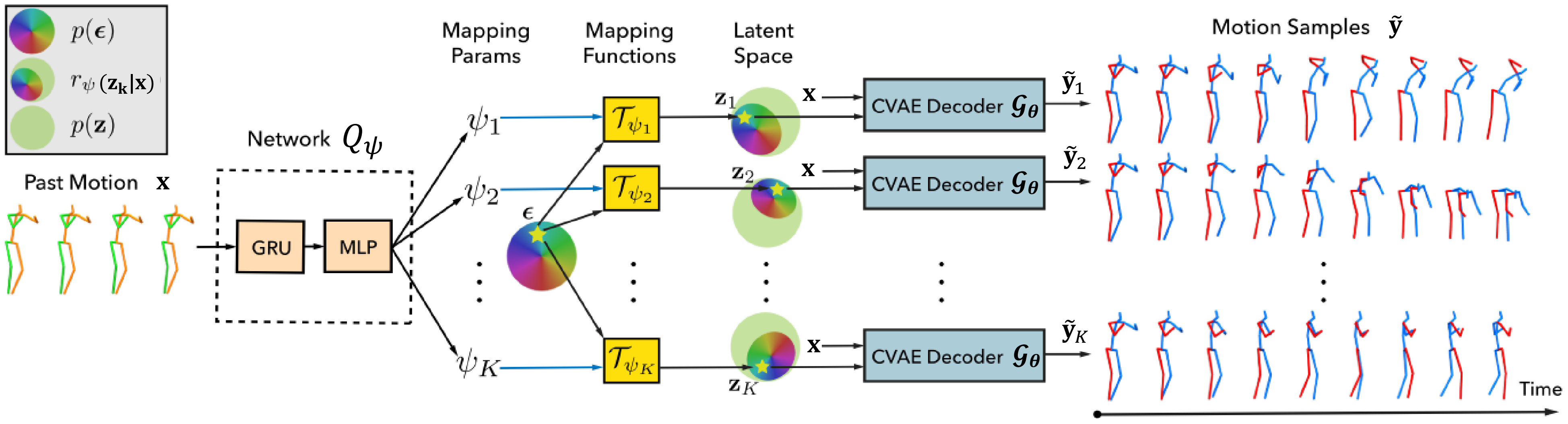
AEVB Kingma D, *et al.*, arXiv 2013

Limitations

- learns an imbalanced multimodal conditional distribution
- Latent codes correspond to the dominant modes, while ignoring other minor modes



Stochastic and Diverse HMP: DLow



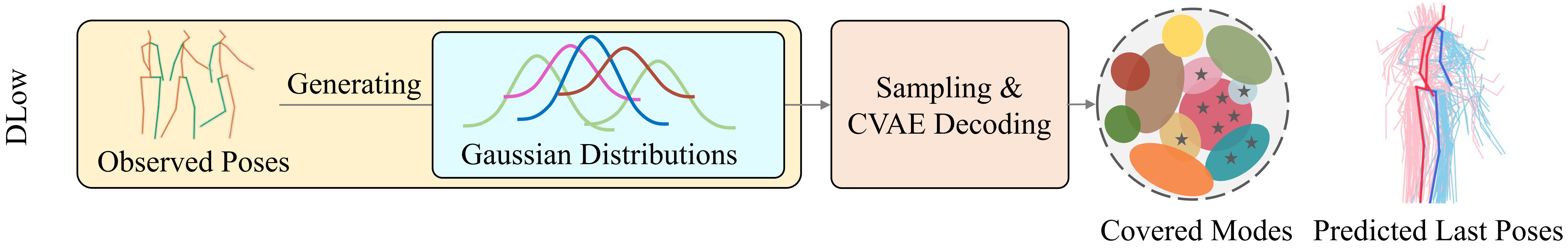
- Given x, they used a network Q_ψ to generate *K* Gaussian distributions $\{(\mathbf{A}_k, \mathbf{b}_k)\}_{k=1}^{K} = Q_\psi(\mathbf{x})$
- Then, they predicted *K* results by

$$\epsilon \sim \mathcal{N}(0,1)$$
$$\mathbf{z}_k = \mathbf{A}_k \epsilon + \mathbf{b}_k, \quad 1 \le k \le K$$
$$\mathbf{y}_k = \mathcal{G}_\theta(\mathbf{x}, \mathbf{z}_k), \quad 1 \le k \le K.$$

DLow Ye Y, *et al.*, ECCV 2020

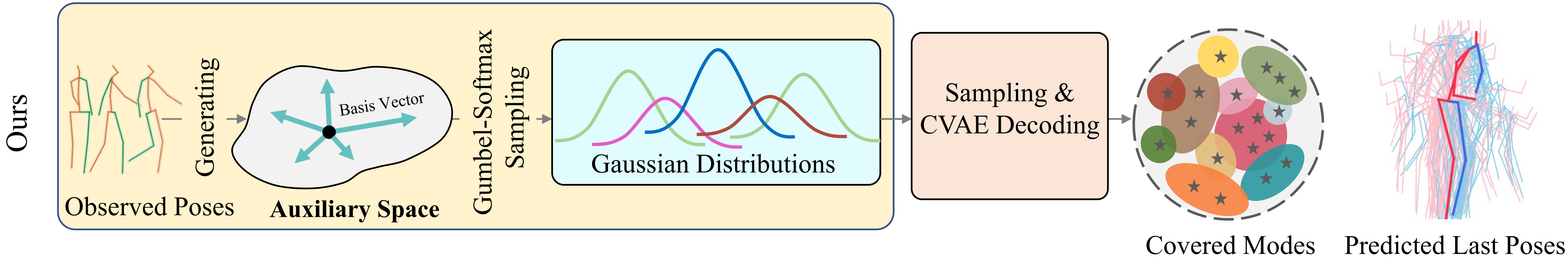
Limitations

- Can only generate a fixed number of Gaussian distributions
- Entangles the performance of diverse prediction with the learning of the network parameters, requiring to consider all training data and make tradeoffs

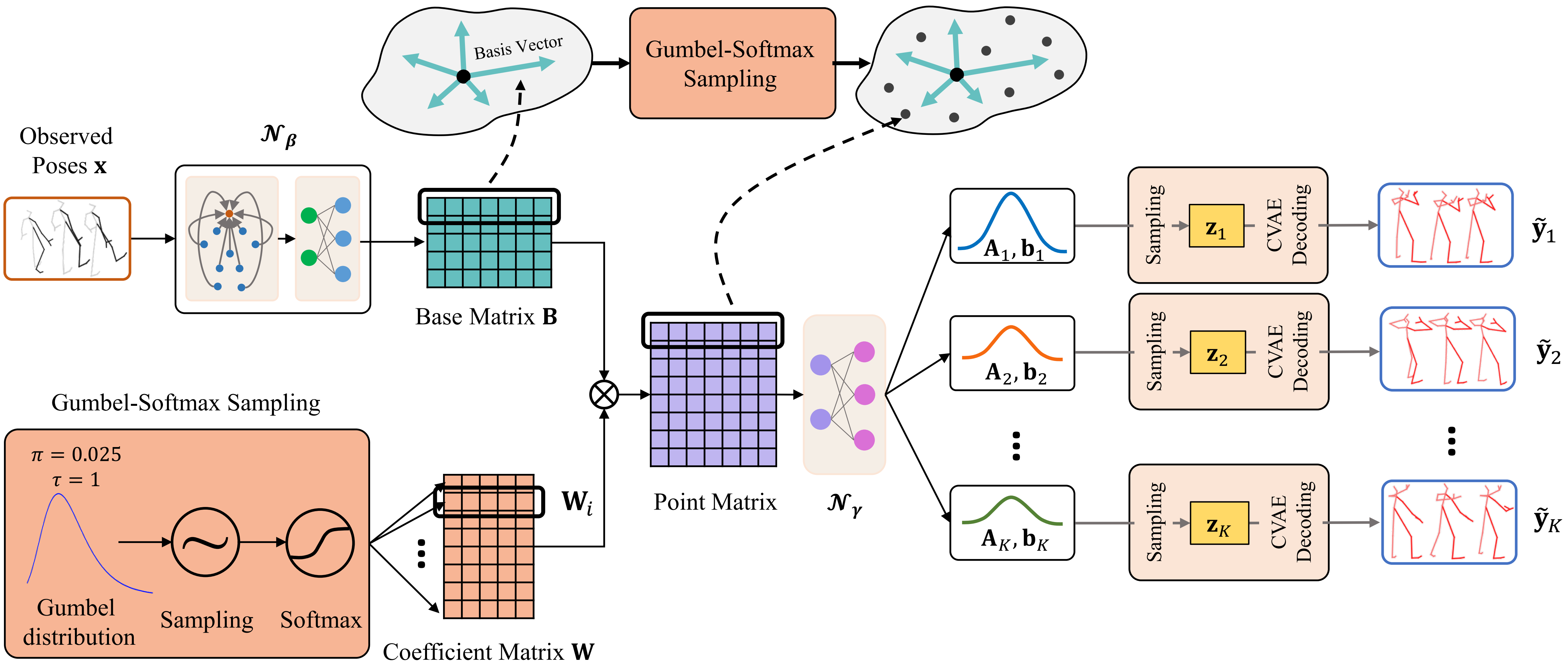


Our Method: Key Insights

- Learn an auxiliary space then sample points from the it and map them to Gaussian distributions which finally correspond to different modes of the target distribution
- Disentangles the correlation between diverse prediction and the network parameter learning
- Arbitrary number of samples can be generated after the auxiliary space is built



Our Method



Basis Vector
Gumbel-Softmax Sampling
Observed Poses x
$\mathcal{N}_\beta$
Base Matrix **B**
Gumbel-Softmax Sampling
$\pi = 0.025$
$\tau = 1$
Gumbel distribution
Sampling
Softmax
$\mathbf{W}_i$
Coefficient Matrix **W**
Point Matrix
$\mathcal{N}_\gamma$
$\mathbf{A}_1, \mathbf{b}_1$
$\mathbf{A}_2, \mathbf{b}_2$
$\mathbf{A}_K, \mathbf{b}_K$
Sampling
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_K$
CVAE Decoding
$\tilde{\mathbf{y}}_1$
$\tilde{\mathbf{y}}_2$
$\tilde{\mathbf{y}}_K$

Training Losses

- Hinge-diversity loss

$$\mathcal{L}_{hdiv} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathbf{max}\left(0, \eta - \left\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}_j\right\|_2\right)$$

- Accuracy loss

$$\mathcal{L}_{acc} = \min_{k} \|\mathbf{y} - \tilde{\mathbf{y}}_k\|_2, k \in [1, K]$$

- KL loss

$$\mathcal{L}'_{KL} = \mathcal{KL}(r_{\beta,\gamma}(\mathbf{z}_k \mid \mathbf{x}) \| p(\mathbf{z})), k \in [1, K]$$

Quantitative Results

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

All the results are calculated by sampling 50 times for each input historical pose sequence.

	Method	Human3.6M [22]					HumanEva-I [44]				
		APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
deterministic	LTD [37]	0.000	0.516	0.756	0.627	0.795	0.000	0.415	0.555	0.509	0.613
	MSR [15]	0.000	0.508	0.742	0.621	0.791	0.000	0.371	0.493	0.472	0.548
stochastic	Pose-Knows [49]	6.723	0.461	0.560	0.522	0.569	2.308	0.269	0.296	0.384	0.375
	MT-VAE [50]	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
	HP-GAN [6]	7.214	0.858	0.867	0.847	0.858	1.139	0.772	0.749	0.776	0.769
	BoM [7]	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
	GMVAE [16]	6.769	0.461	0.555	0.524	0.566	2.443	0.305	0.345	0.408	0.410
	DeLiGAN [21]	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
	DSF [51]	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
	DLow [52]	11.741	0.425	0.518	0.495	0.531	4.855	0.251	0.268	0.362	0.339
	GSPS [36]	14.757	0.389	0.496	0.476	0.525	5.825	0.233	0.244	0.343	0.331
	Ours	**15.310**	**0.370**	**0.485**	**0.475**	**0.516**	**6.109**	**0.220**	**0.234**	**0.342**	**0.316**

MSR Dang L, *et al.*, ICCV 2021 **GSPS Mao W**, *et al.*, ICCV 2021 (Oral) **DLow** Ye Y, *et al.*, ECCV 2020

Qualitative Results

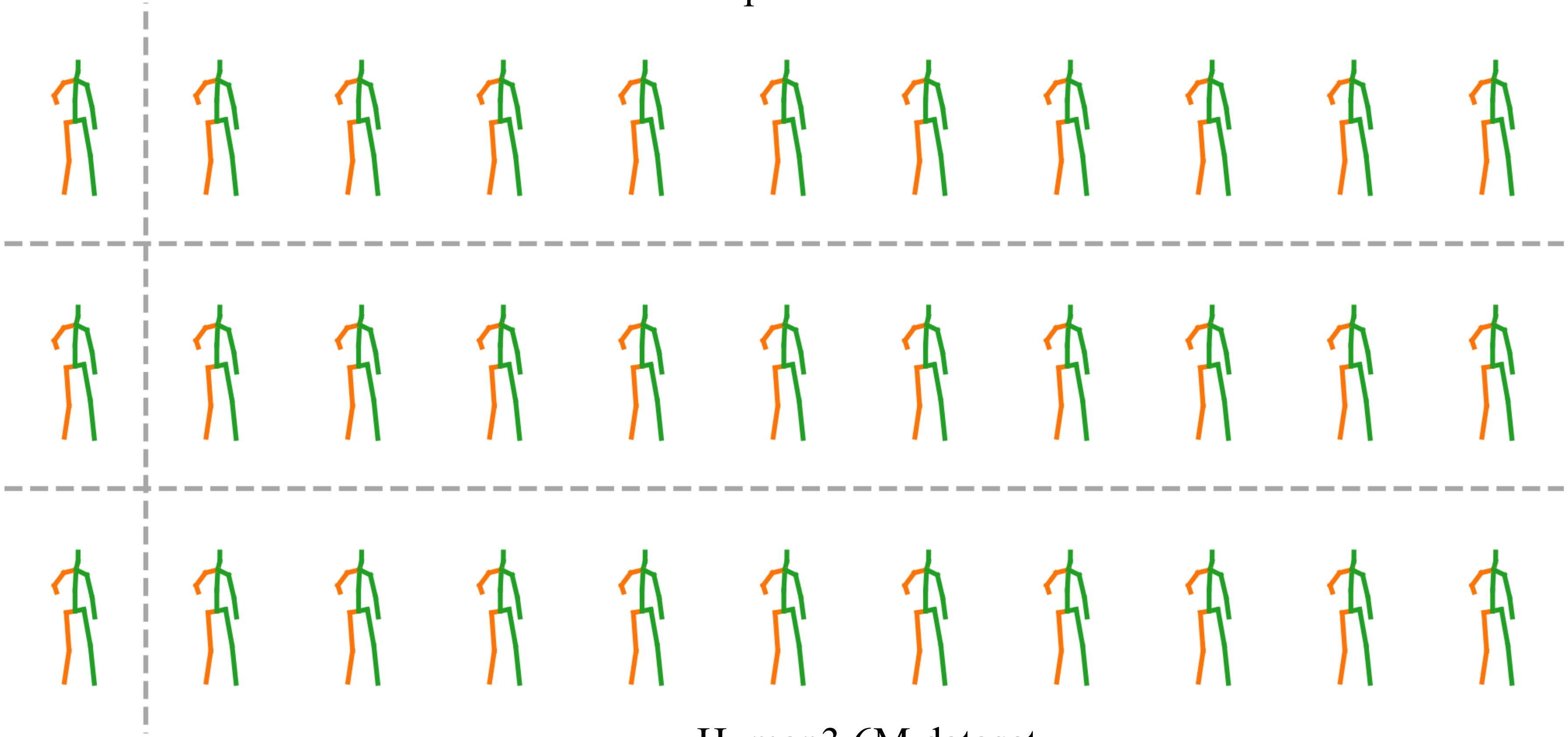
GT

10 predicted results

DLow
(12.367)

GSPS
(15.177)

Ours
(16.463)

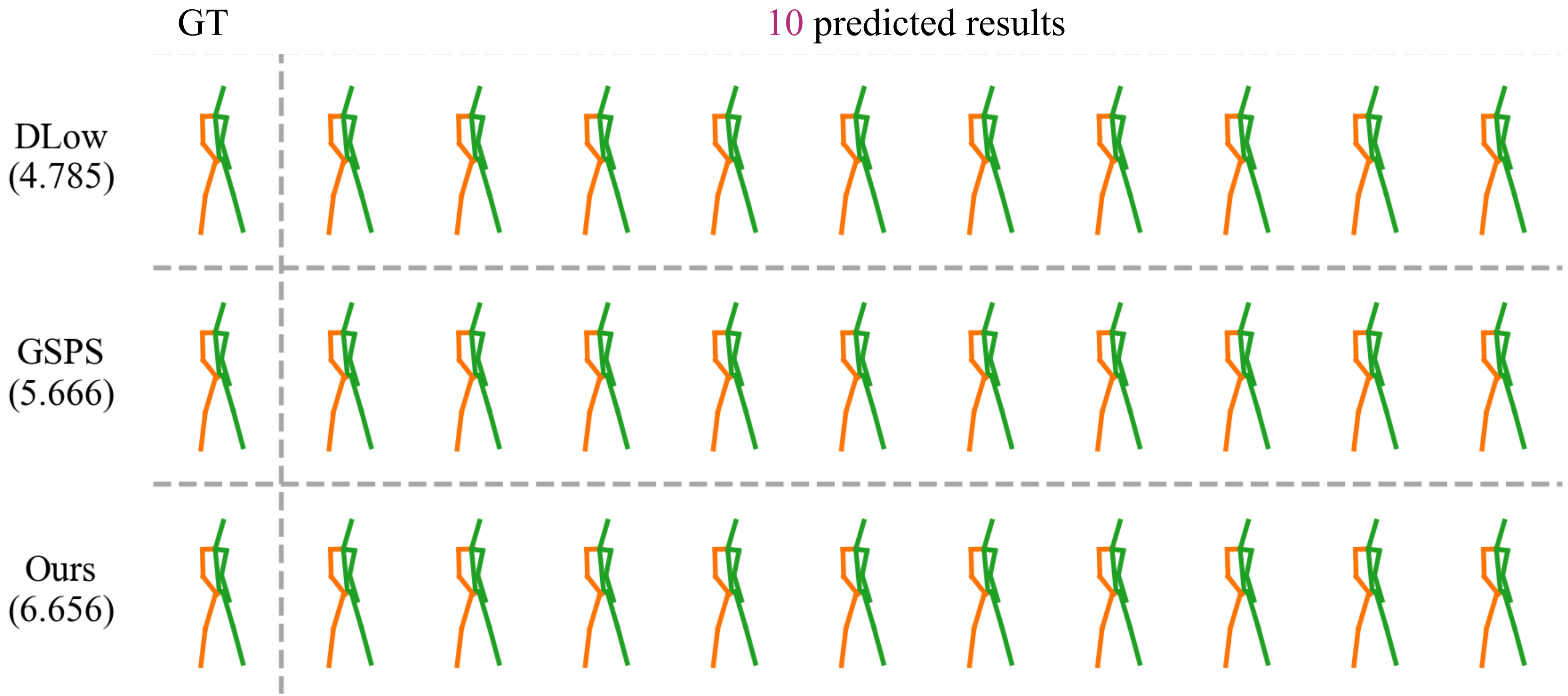


Human3.6M dataset

DLow Ye Y, *et al.*, ECCV 2020

GSPS Mao W, *et al.*, ICCV 2021 (Oral)

Qualitative Results



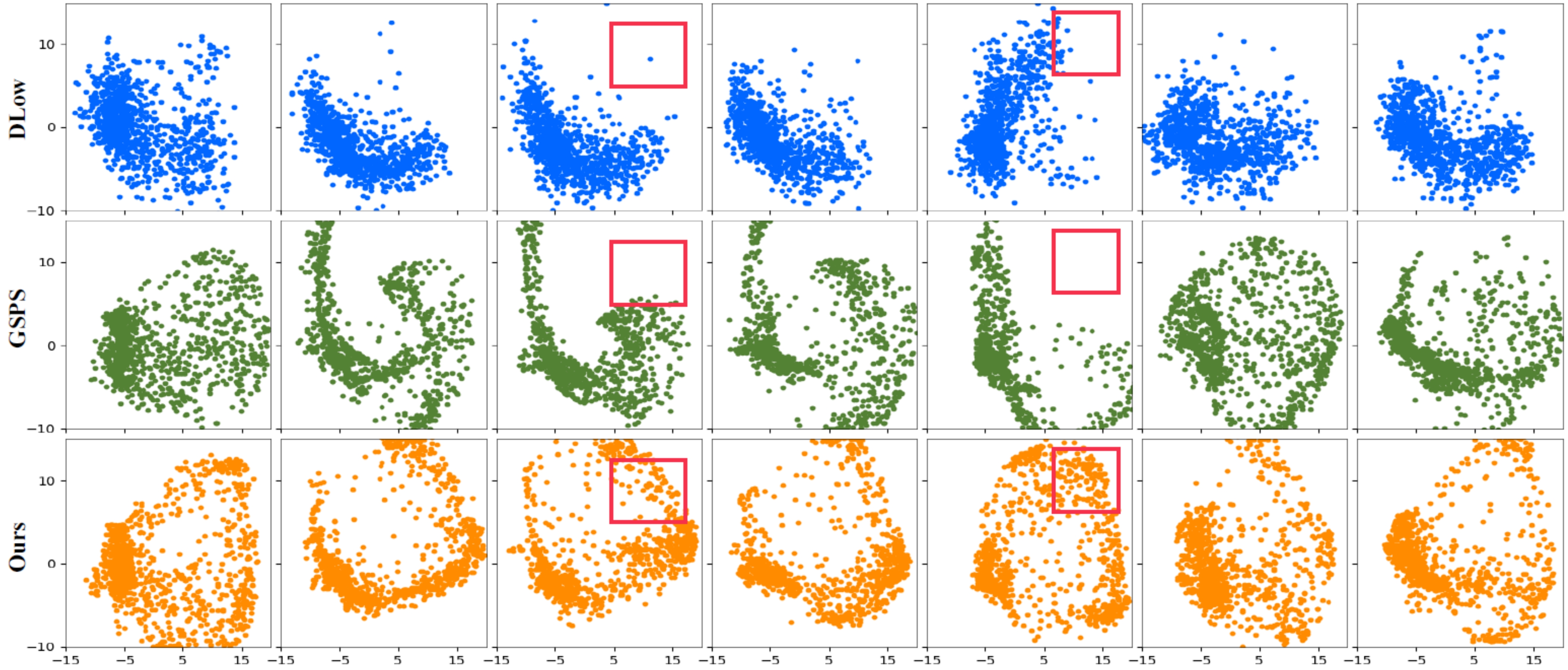
HumanEva-I dataset

DLow Ye Y, *et al.*, ECCV 2020

GSPS Mao W, *et al.*, ICCV 2021 (Oral)

Qualitative Results

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October



DLow Ye Y, *et al.*, ECCV 2020 **GSPS** Mao W, *et al.*, ICCV 2021 (Oral)

QR Code for our project:

https://github.com/Droliven/diverse_sampling

Thank you!

Acknowledgement

This work is sponsored by Prof. Yongwei Nie's and Prof. Guiqing Li's Natural Science Foundation of China projects (62072191, 61972160), and their Natural Science Foundation of Guangdong Province projects (2019A1515010860, 2021A1515012301).