Motion Diversification Networks

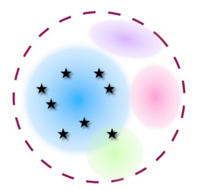
Hee Jae Kim and Eshed Ohn-Bar
Boston University

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivation

Challenge: Simplistic latent code sampling strategies hinder diversity in current generative modeling techniques

$$\epsilon \sim p_0(\epsilon) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$



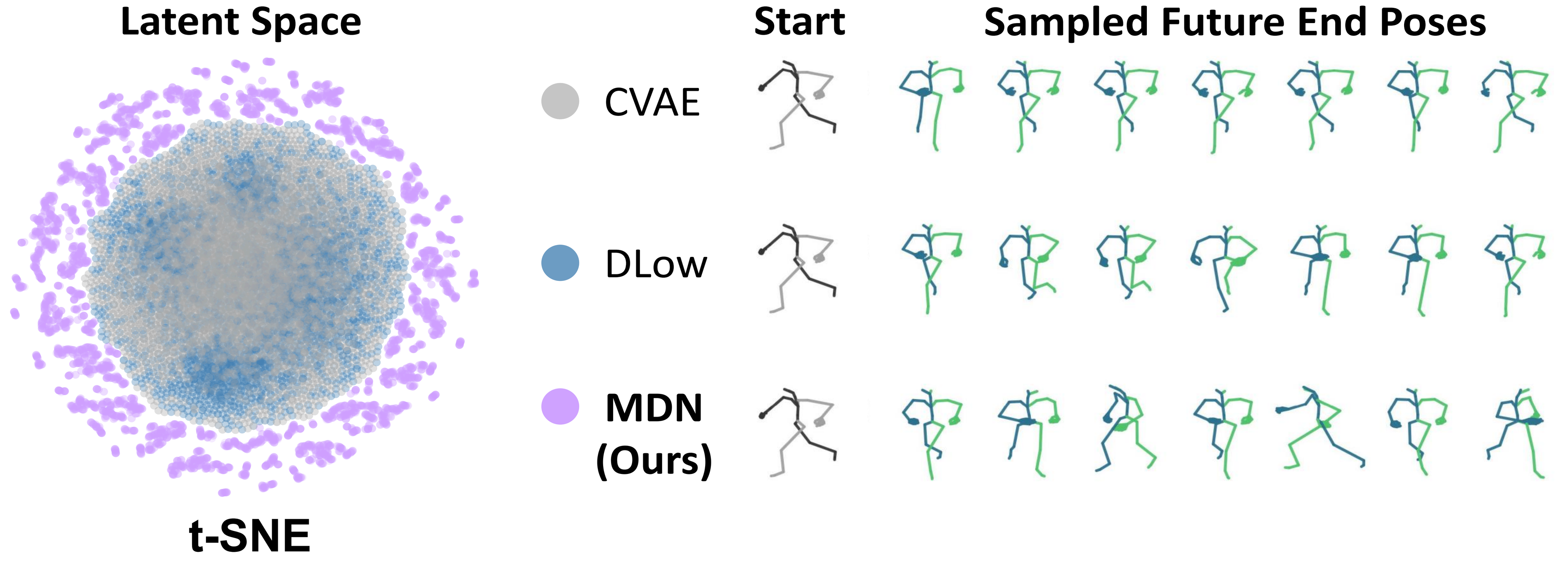
Related Work: DLow, ECCV 2020, STARS, ECCV 2022 – introduce an affine transformation to circumvent ***mode collapse*** in generative models

$$\mathrm{Z} = \mathrm{A}\epsilon + \mathrm{b}$$

However, this limited capacity transformation cannot capture complex sample correlations, i.e., ineffective for uncertainty *across the modes* and rare modes.

Key Idea: A ***transformer-based diversification mechanism*** for highly realistic and diverse 3D motion generation.

Overview



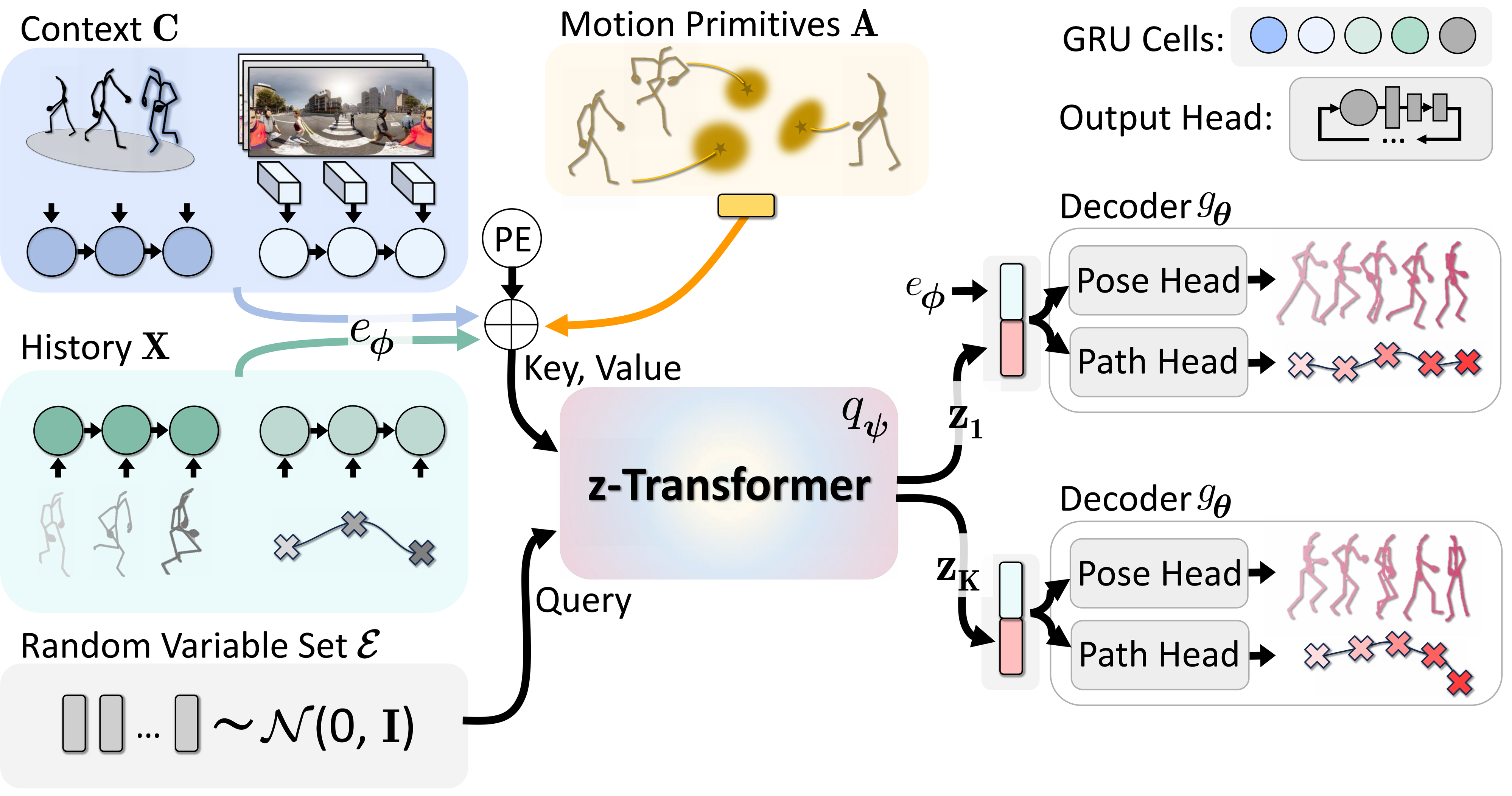
z-transformer: We employ an ***attention-based diversification module*** to produce a diverse set of latent vectors that expressively model correlations among multiple samples and modes.

Motion Primitives: To guide sample diversity and reduce modeling complexity in diverse scenarios, we incorporate deterministic motion primitives (centroids of clusters in the 3D pose space).

Scene and Social Context: We use the transformer architecture to easily fuse in additional context, i.e., as keys and values in the z-transformer.

Dense Urban Navigation Benchmark: Prior datasets (e.g., Human3.6M, AMASS, and HumanML3D) are limited to static indoor settings. We introduce **DenseCity**, a simulation benchmark with dense pedestrians. We also use YouTube, which helps further address the current gap between simulated, generated, and realistic 3D human motion.

Method



Model Architecture of MDN

Loss Functions:

(Stage 1) $\mathcal{L}_{\mathrm{CVAE}} = \mathrm{KL}(e_\phi(\mathrm{z}|\mathrm{X},\mathrm{C}) \,\|\, p_0(\mathrm{z})) + \|g_\theta(\mathrm{z}) - \mathrm{Y}\|_2^2$

(Stage 2) $\mathcal{L}_{\mathrm{MDN}} = \mathcal{L}_r + \mathcal{L}_d$

Reconstruction loss, $\mathcal{L}_r = \min_k \|\hat{\mathrm{Y}}_k - \mathrm{Y}\|^2$

Diversity promoting loss, $\mathcal{L}_d = \frac{2}{K(K-1)} \sum_{j=1}^{K} \sum_{k=j+1}^{K} \exp\left(-\frac{\|\hat{\mathrm{Y}}_j - \hat{\mathrm{Y}}_k\|_1}{\alpha}\right)$

Latent Variable Transformer (q_ψ):
We transform the K set of random variables ($\mathcal{E} \in \mathbb{R}^{K \times N_z}$) into a diverse set of latent codes ($\mathbf{Z} \in \mathbb{R}^{K \times N_z}$) by utilizing motion primitives ($\mathbf{A} \in \mathbb{R}^{K \times N_f}$), scenes and social contexts ($\mathbf{C} \in \mathbb{R}^{T_h \times N_c}$) as keys ($\mathbf{K} \in \mathbb{R}^{K \times N_z}$) and values ($\mathbf{V} \in \mathbb{R}^{K \times N_z}$):

$$\mathcal{E} \sim p_0(\epsilon) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbf{Z} = q_\psi(\mathbf{X}, \mathbf{C}, \mathcal{E}, \mathbf{A}) \;\; = \mathrm{Attn}(\mathcal{E}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathcal{E}\mathbf{K}^{\mathrm{T}}}{\sqrt{N_z}}\right)\mathbf{V}$$

Diversified latent codes ($\mathrm{z}_k \in \mathbb{R}^{N_z}$) are decoded into highly diverse motion samples ($\hat{\mathbf{Y}}_k \in \mathbb{R}^{T_f \times N_f}$) using the pre-trained decoder (g_θ):

$$\{\hat{\mathbf{Y}}_k\}_{k=1}^{K} = g_\theta(\mathrm{z}_k, \mathbf{X}, \mathbf{C})$$

T_h : Time horizon in history

T_f : Time horizon in future

Quantitative Results

Dataset

Evaluation on DenseCity

Method	3D Pose			2D Path		
	APD ↑	ADE ↓	FDE ↓	APD ↑	ADE ↓	FDE ↓
CVAE [32]	7.451	0.610	0.932	-	-	-
PoseGPT [38]	9.099	0.913	0.980	-	-	-
HuMoR [48]	11.134	0.705	1.030	-	-	-
DLow [79]	11.980	0.596	0.899	-	-	-
Cao et al. [7]	5.810	0.858	1.285	0.978	0.701	0.675
MDN	16.799	0.584	**0.879**	1.065	0.666	0.621
Cao et al. [7]+YouTube	5.593	0.805	1.096	0.782	0.697	0.632
MDN+YouTube	**16.812**	**0.578**	0.921	**1.331**	**0.646**	**0.589**

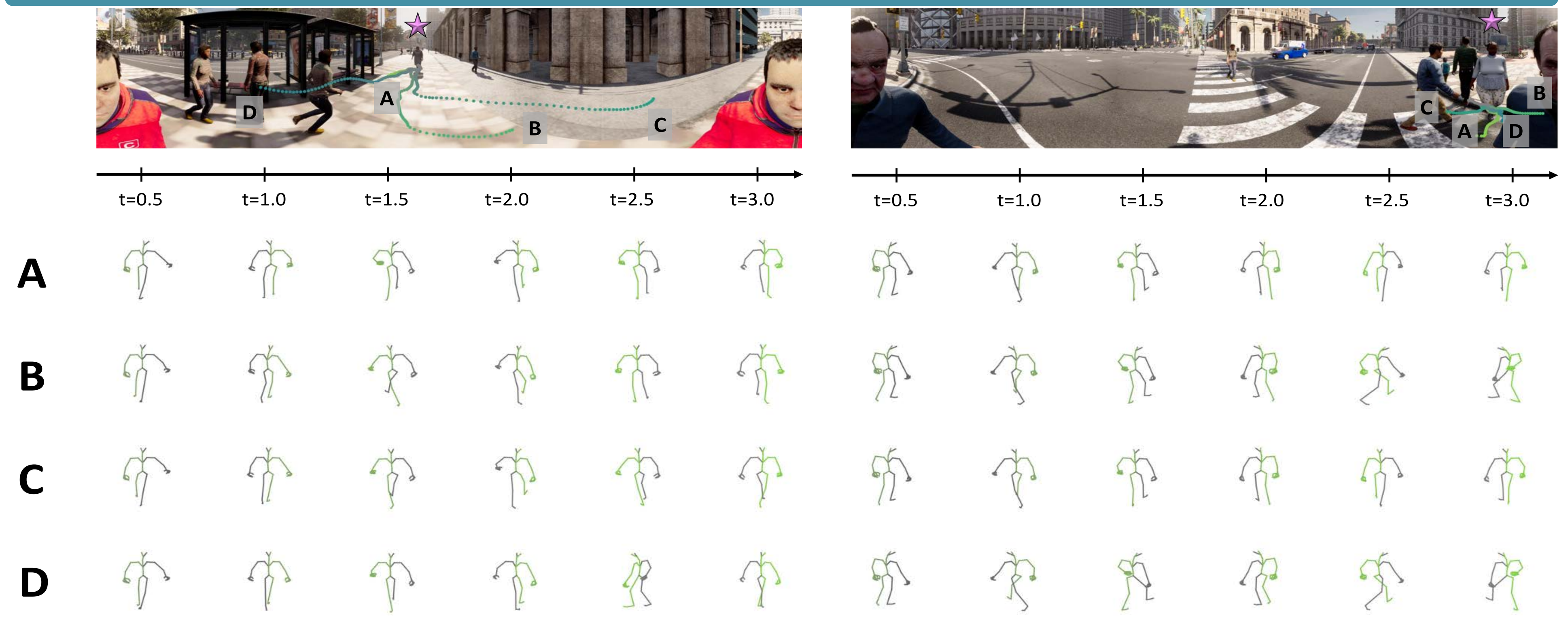
Evaluation on Human3.6M

Method	APD ↑	ADE ↓	FDE ↓
DLow [74]	11.741	0.425	0.518
MOJO [81]	12.579	0.412	0.514
GSPS [40]	14.757	0.389	0.496
BeLFusion [3]	7.602	0.372	0.474
STARS [68]	15.884	0.358	0.445
MDN (Ours)	**17.450**	**0.355**	**0.442**

Evaluation on 3DPW & HPS

Method	3DPW [61]			HPS [24]		
	APD ↑	ADE ↓	FDE ↓	APD ↑	ADE ↓	FDE ↓
CVAE [32]	3.068	1.032	1.096	3.592	0.970	1.019
DLow [79]	4.111	1.010	1.069	4.835	0.958	1.005
MDN	6.355	0.982	1.032	**6.943**	0.916	0.971
CVAE [32]+YouTube	3.100	0.991	1.060	3.634	0.913	0.958
DLow [79]+YouTube	4.160	0.969	1.034	5.010	0.898	0.938
MDN+YouTube	**8.266**	**0.918**	**0.986**	6.925	**0.886**	**0.930**

Qualitative Results



Perceptual User Study