NeRM: Learning Neural Representations for High-Framerate Human Motion Synthesis

Dong Wei[1], Huaijiang Sun[1], Bin Li[2], Xiaoning Sun[1], Shengxiang Hu[1], Weiqing Li[1], Jianfeng Lu[1]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology
2. Tianjin AiForward Science and Technology Co., Ltd

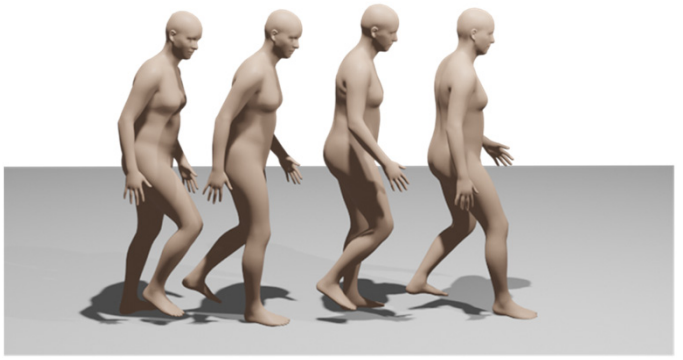
Problem Definition

Given an arbitrary condition *c*, such as:

- text description
- class label
- …

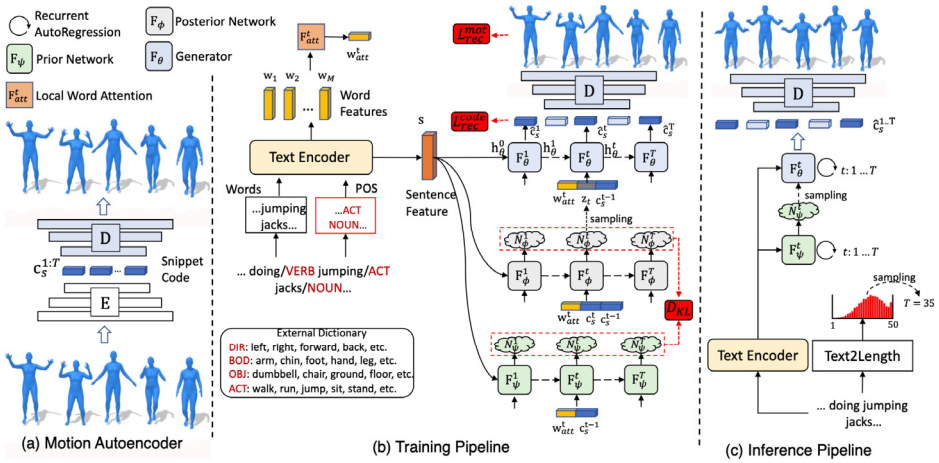
Goal: Generate plausible and diverse human motions.

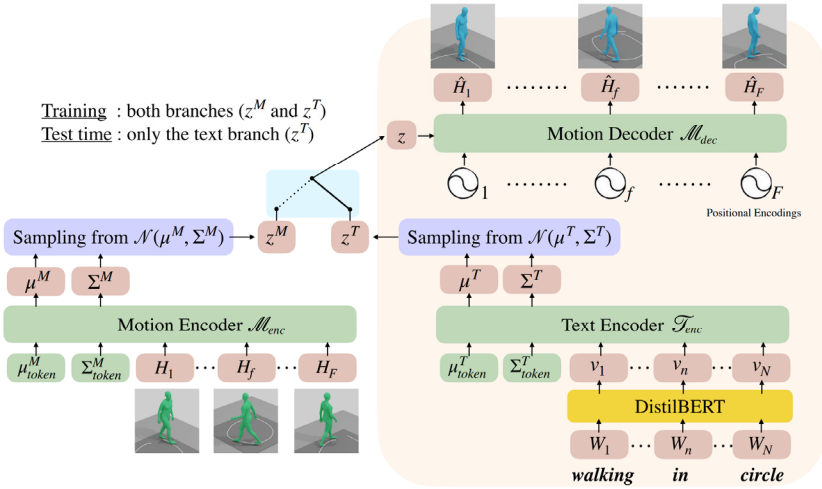
Text: *"A person walks to the right slowly."*



Related Work

T2M (Guo et al. 2022) & TEMOS (Petrovich et al. 2022): a conditional VAE model





Problem:

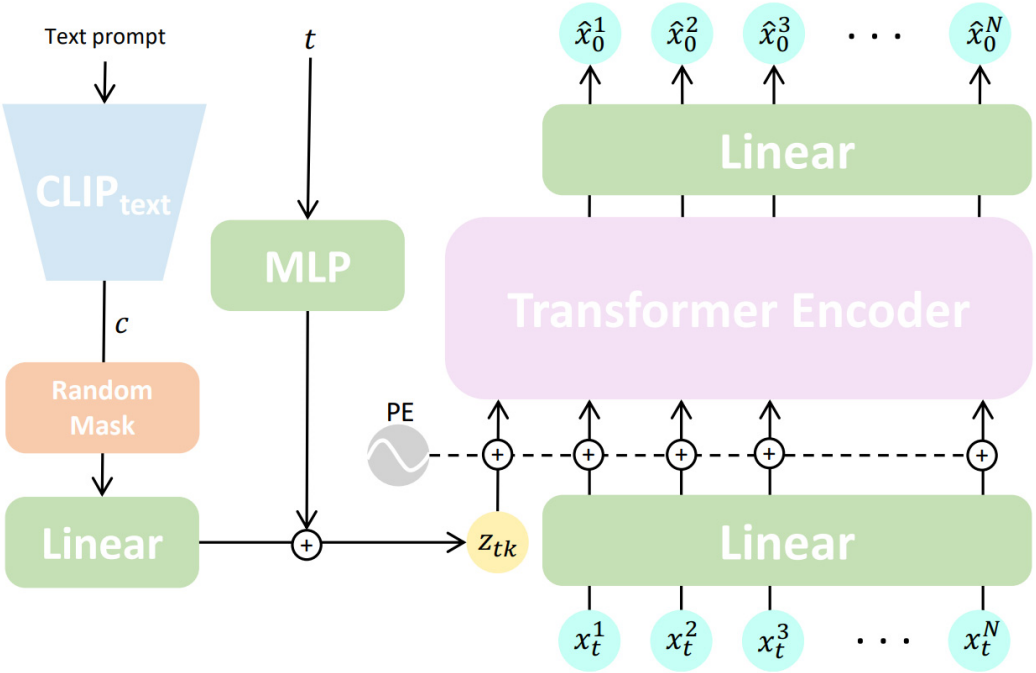
- VAE brings about posterior collapse with a powerful decoder → not diverse enough
- Limited generative capability → not plausible enough

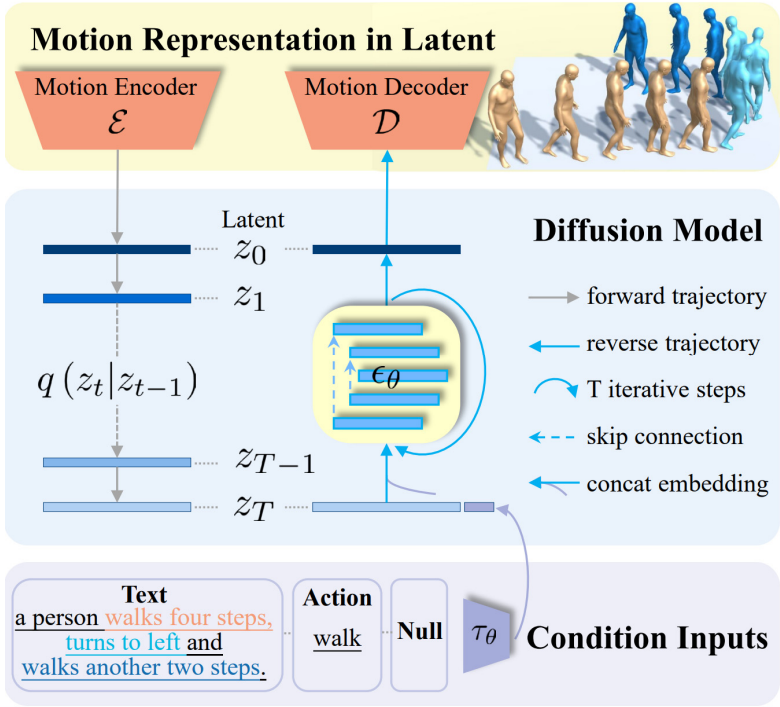
Chuan Guo, Zhihao Zou, Xinxin Zuo, et al. Generating diverse and natural 3D human motions from text. CVPR, 2022.

Mathis Petrovich, Michael J. Black, Gul Varol. TEMOS: Generating diverse human motions from textual descriptions. ECCV, 2022.

Related Work

MDM (Tevet et al. 2023) & MLD (Chen et al. 2023): a diffusion model





Address the limitations of VAE models.

Guy Tevet, Sigal Raab, Brian Gordon, et al. Human motion diffusion model. ICLR, 2023.

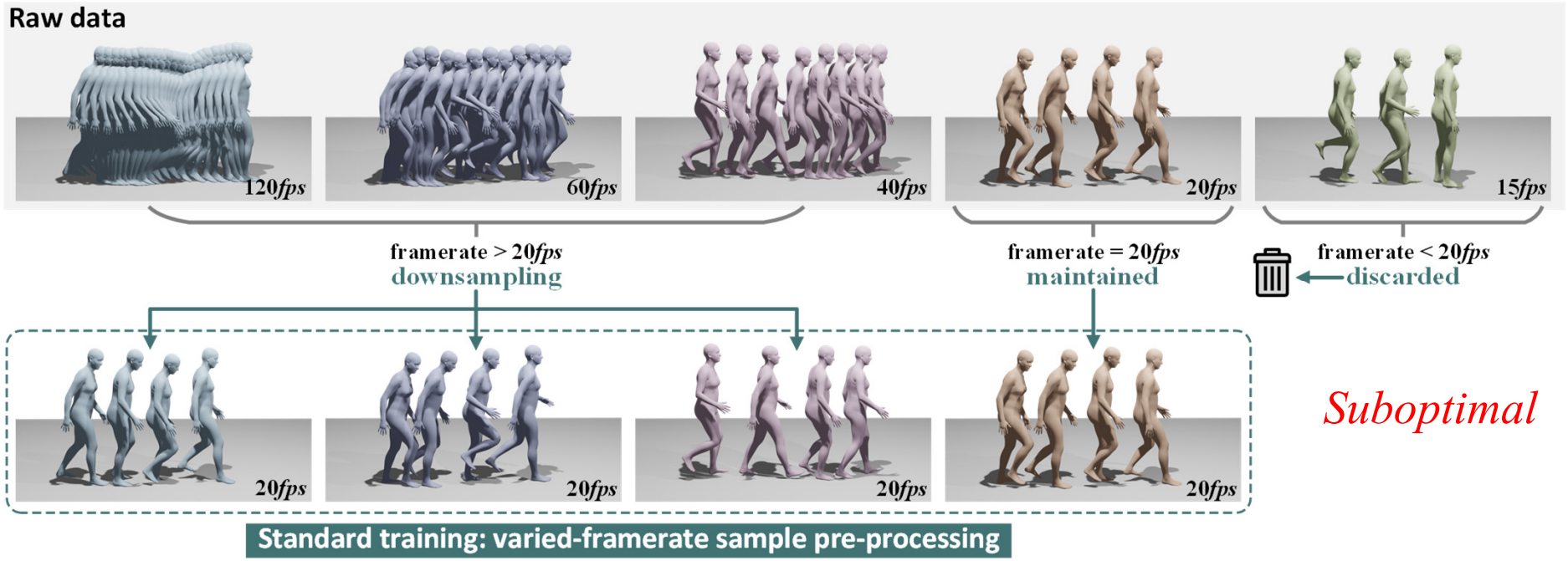
Xin Chen, Biao Jiang, Wen Liu, et al. Executing your commands via motion diffusion in latent space. CVPR, 2023.

Challenge

Existing methods cannot support high-framerate motion generation, due to:

- Memory overloads
- Varied-size training dataset

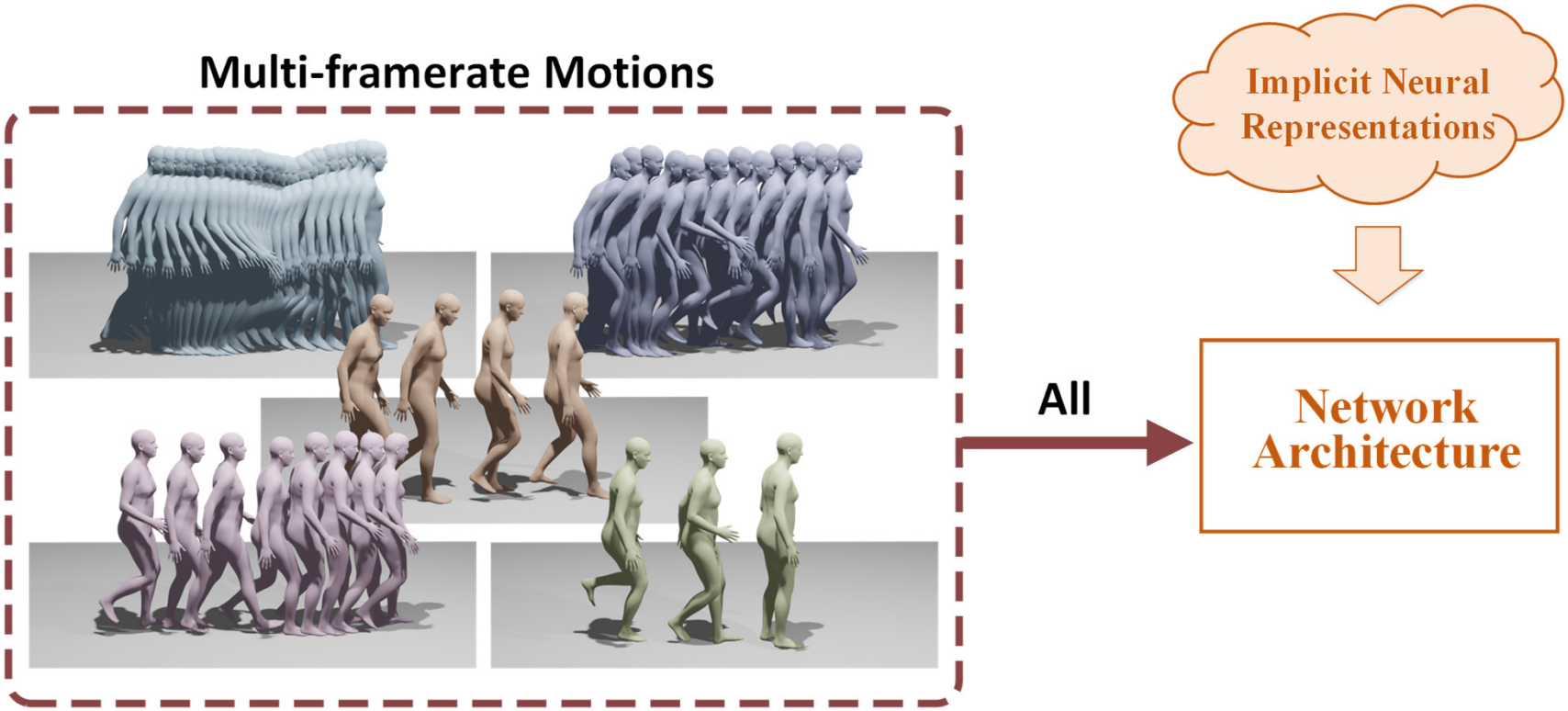
Therefore, they employ a preprocessing step.



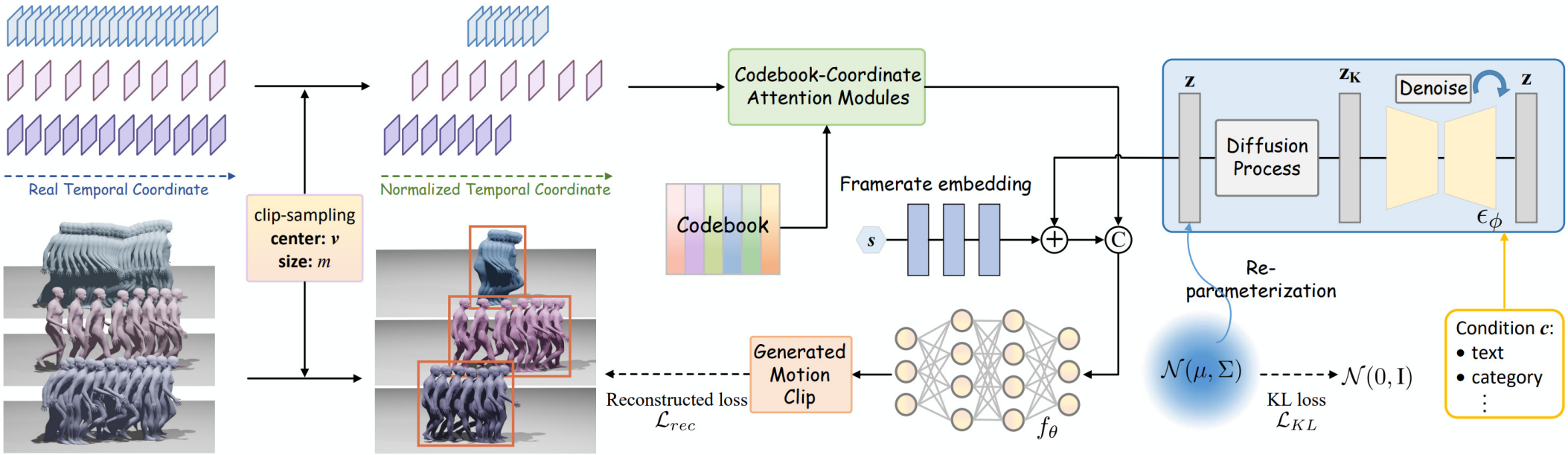
Suboptimal

Key Insight

It is desired to embrace the natural diversity of motion framerates and process them at their **native** framerate.



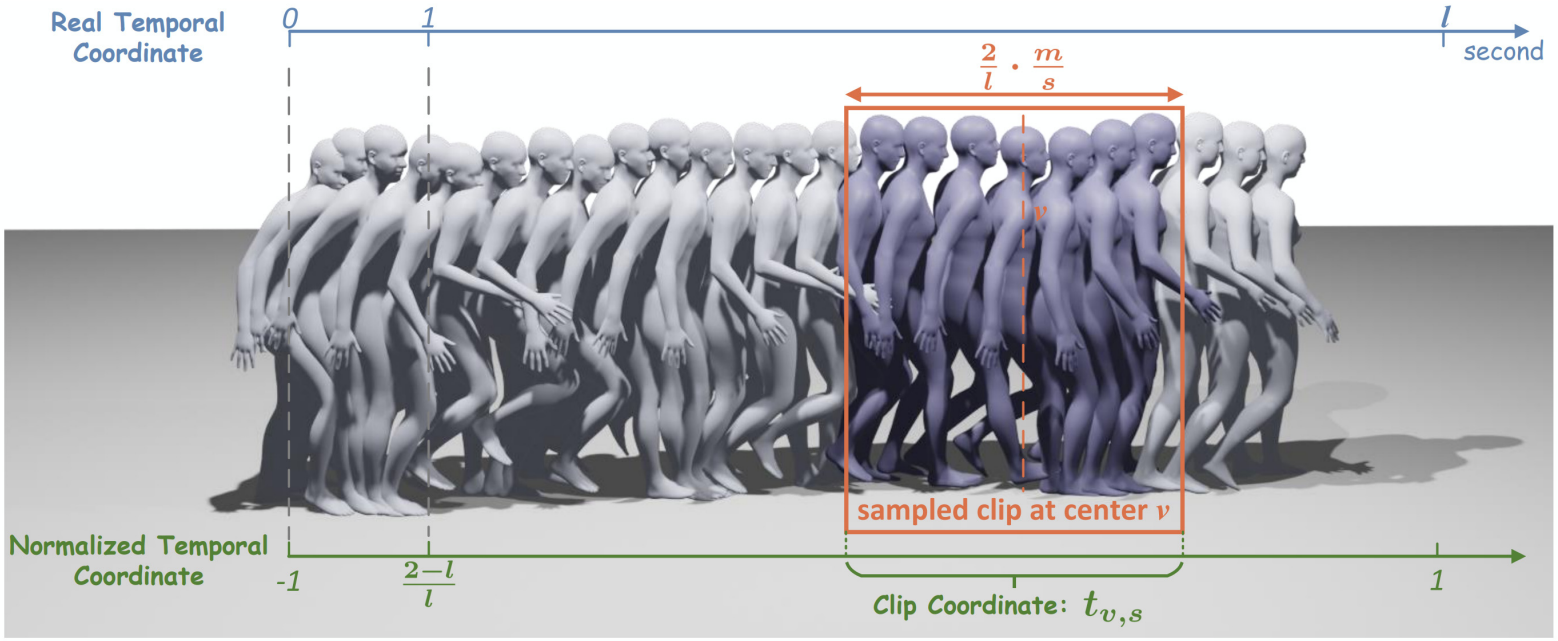
Overview



Two-stage pipeline of our NeRM. In the first stage (left), we sample motion clips at random framerates from full-size motion sequences for training. The second stage (right) takes the latent variables as input to our diffusion model and can be guided by various conditions.

Multi-framerate Training

To learn from multi-framerate motions in training sets, we design our approach by exploiting the temporal consistency in motions to generate motion *clips*.



Loss Function

The first stage: Variational implicit neural representation

$$\mathcal{L}^i = \mathcal{L}^i_{rec} + \lambda_{KL}\mathcal{L}^i_{KL} = \|\hat{x}^i_{clip} - x^i_{clip}\|^2 + \lambda_{KL} D_{\mathrm{KL}}(\mathcal{N}(\mu_i, \Sigma_i)||p(z))$$

The first stage: Model parameters optimization

$$\{(\mu_i^*, \Sigma_i^*)\}_{i=1,2,\cdots,n} = \arg\min_{\mu_i,\Sigma_i} \mathcal{L}^i, \qquad \theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \min_{\mu_i,\Sigma_i} \mathcal{L}^i$$

The second stage: Motion diffusion in latent space

$$\min_{\phi} \mathbb{E}_{k,\boldsymbol{z}\sim Z,\epsilon\sim\mathcal{N}(0,\mathrm{I})}\left[\|\epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_k}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_k}\epsilon, k)\|^2\right]$$

The second stage: Classifier-free guidance

$$\epsilon_\phi(\boldsymbol{z}_k, k, c) = r\epsilon_\phi(\boldsymbol{z}_k, k, c) + (1-r)\epsilon_\theta(\boldsymbol{z}_k, k, \emptyset)$$

Experiments

Text-to-Motion

Method	HumanML3D (Guo et al., 2022)					KIT (Plappert et al., 2016)				
	FID ↓	R-Precision (Top-3) ↑	Multimodal Dist ↓	Diversity →	MM ↑	FID ↓	R-Precision (Top-3) ↑	Multimodal Dist ↓	Diversity →	MM ↑
Real	$0.002^{\pm.000}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-	$0.031^{\pm.004}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
JL2P (Ahuja & Morency, 2019)	$11.02^{\pm.046}$	$0.486^{\pm.002}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-	$6.545^{\pm.072}$	$0.483^{\pm.005}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Hier (Ghosh et al., 2021)	$6.532^{\pm.024}$	$0.552^{\pm.004}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-	$5.203^{\pm.107}$	$0.531^{\pm.007}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	-
T2M (Guo et al., 2022)	$1.067^{\pm.002}$	$0.740^{\pm.003}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$	$2.770^{\pm.109}$	$0.693^{\pm.007}$	$3.401^{\pm.008}$	$\underline{10.91}^{\pm.119}$	$1.482^{\pm.065}$
MoFusion (Dabral et al., 2023)	-	0.492	-	8.82	$\underline{2.521}$	-	-	-	-	-
MDM (Tevet et al., 2023)	$0.544^{\pm.044}$	$0.611^{\pm.007}$	$5.566^{\pm.027}$	$\underline{9.559}^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$	$0.497^{\pm.021}$	$0.396^{\pm.004}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
PhysDiff (Yuan et al., 2023)	$\underline{0.433}$	0.631	-	-	-	-	-	-	-	-
MLD (Chen et al., 2023)	$0.473^{\pm.013}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$	$\mathbf{0.404}^{\pm.027}$	$\underline{0.734}^{\pm.007}$	$\underline{3.204}^{\pm.027}$	$10.80^{\pm.117}$	$\mathbf{2.192}^{\pm.071}$
NeRM (fixed-framerate train)	$0.489^{\pm.013}$	$\underline{0.774}^{\pm.003}$	$\underline{3.186}^{\pm.015}$	$9.692^{\pm.082}$	$2.330^{\pm.075}$	$0.522^{\pm.029}$	$0.727^{\pm.005}$	$3.843^{\pm.027}$	$10.90^{\pm.094}$	$\underline{1.931}^{\pm.133}$
NeRM (native-framerate train)	$\mathbf{0.389}^{\pm.011}$	$\mathbf{0.779}^{\pm.003}$	$\mathbf{3.178}^{\pm.016}$	$\mathbf{9.547}^{\pm.073}$	$2.193^{\pm.081}$	$\underline{0.472}^{\pm.019}$	$\mathbf{0.736}^{\pm.007}$	$\mathbf{3.189}^{\pm.031}$	$\mathbf{10.94}^{\pm.114}$	$1.785^{\pm.082}$

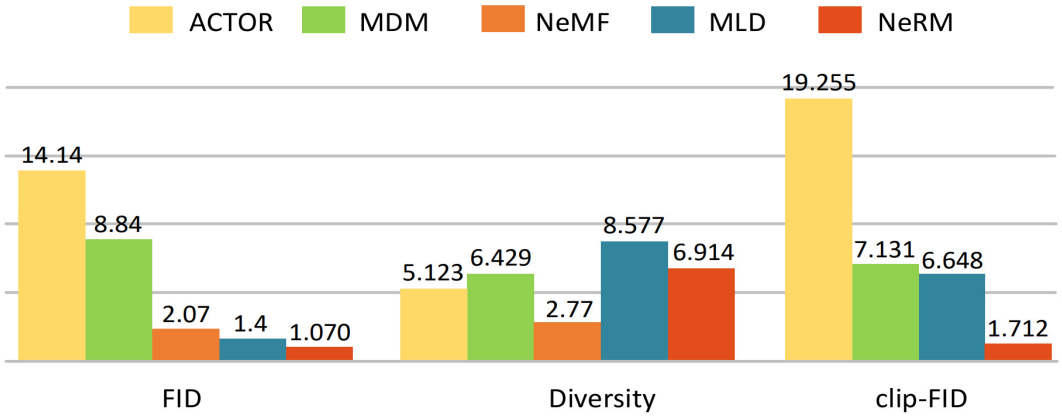
- SOTA FID
- Using original multi-framerate datasets can achieve much better performance.

Experiments

Action-to-Motion

Method	UESTC (Ji et al., 2018)					HumanAct12 (Guo et al., 2020)			
	FID_{train} ↓	FID_{test} ↓	Accuracy ↑	Diversity →	MM →	FID_{train} ↓	Accuracy ↑	Diversity →	MM →
Real	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
ACTOR (Petrovich et al., 2021)	$20.5^{\pm2.3}$	$23.43^{\pm2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$6.840^{\pm.030}$	$2.530^{\pm.020}$
MDM (Tevet et al., 2023)	$9.98^{\pm1.33}$	$\mathbf{12.81}^{\pm\mathbf{1.46}}$	$0.950^{\pm.000}$	$33.02^{\pm.28}$	$\mathbf{14.26}^{\pm\mathbf{.12}}$	$0.100^{\pm.000}$	$\mathbf{0.990}^{\pm\mathbf{.000}}$	$6.680^{\pm.050}$	$2.520^{\pm.010}$
MLD (Chen et al., 2023)	$12.89^{\pm.109}$	$15.79^{\pm.079}$	$0.954^{\pm.001}$	$33.52^{\pm.14}$	$13.57^{\pm.06}$	$\mathbf{0.077}^{\pm\mathbf{.004}}$	$0.964^{\pm.002}$	$6.831^{\pm.050}$	$2.824^{\pm.038}$
INR-MLP (Cervantes et al., 2022)	$\mathbf{9.55}^{\pm\mathbf{.06}}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$	$0.114^{\pm.001}$	$0.970^{\pm.001}$	$6.786^{\pm.057}$	$2.507^{\pm.034}$
NeRM (Ours)	$11.75^{\pm.31}$	$14.23^{\pm.174}$	$\mathbf{0.956}^{\pm\mathbf{.001}}$	$\mathbf{33.20}^{\pm\mathbf{.21}}$	$14.41^{\pm.06}$	$0.106^{\pm.000}$	$0.977^{\pm.001}$	$\mathbf{6.866}^{\pm\mathbf{.032}}$	$\mathbf{2.492}^{\pm\mathbf{.048}}$

Unconditional Generation



Experiments

Ablation Study

Effectiveness of time encoding

Simple	Codebook	Motion Reconstruction (MRE)					Motion Synthesis (clip-FID)				
		20	40	60	100	120	20	40	60	100	120
✗	✗	0.134	0.141	0.164	0.094	0.091	0.471	0.803	1.070	1.769	2.944
✓	✗	0.053	0.049	0.039	0.057	0.043	0.397	0.519	0.701	0.142	1.717
✗	✓	**0.041**	**0.036**	**0.034**	**0.035**	**0.038**	**0.389**	**0.493**	**0.680**	**0.903**	**1.315**

Effectiveness of variational INRs

Variational INRs	Motion Reconstruction (MRE)					Motion Synthesis (clip-FID)				
	20	40	60	100	120	20	40	60	100	120
✗	**0.032**	**0.030**	**0.031**	**0.036**	**0.027**	1.280	2.924	7.012	10.482	14.654
✓	0.041	0.036	0.034	0.035	0.038	**0.389**	**0.493**	**0.680**	**0.903**	**1.315**

Experiments

Ablation Study

Effectiveness of different framerates

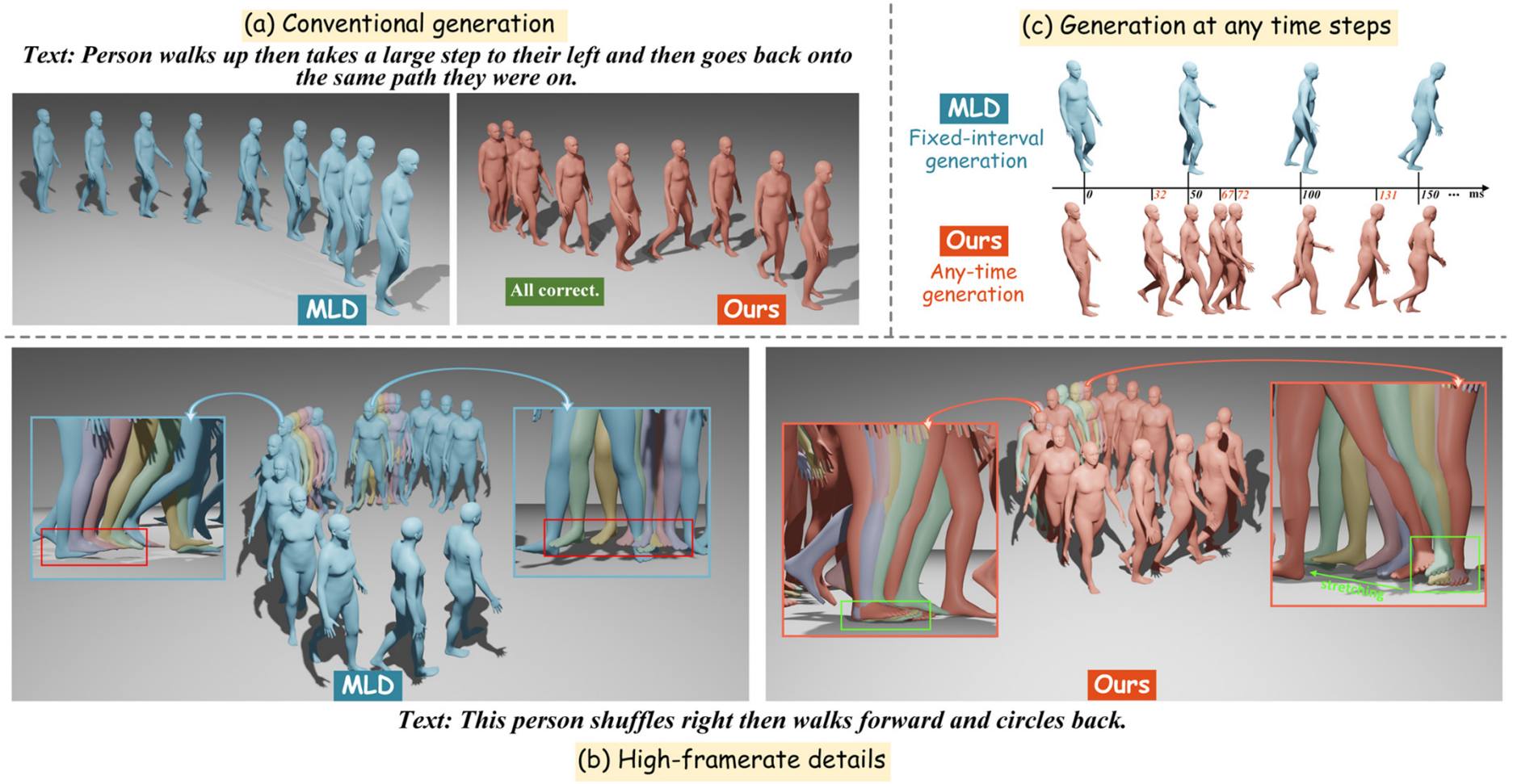
Method	20	40	60	100	120
T2M	1.067	2.831	6.442	9.182	11.264
MDM	0.544	1.882	4.605	5.966	8.401
MLD	0.473	1.465	3.816	5.138	7.878
NeRM	**0.389**	**0.493**	**0.680**	**0.903**	**1.315**

Effectiveness of time normalization

Time Normalization	**MRE@20fps**	**FID**	**Diversity**
✗	0.118	0.958	9.892
✓	**0.041**	**0.389**	**9.547**

Experiments

Visualization



Thank You!