Motion Mamba: Efficient and Long Sequence Motion Generation

ECCV 2024


Zeyu Zhang[12*†]


Akide Liu[1*]


Ian Reid[3]


Richard Hartley[2]


Bohan Zhuang[1]


Hao Tang[4✉]

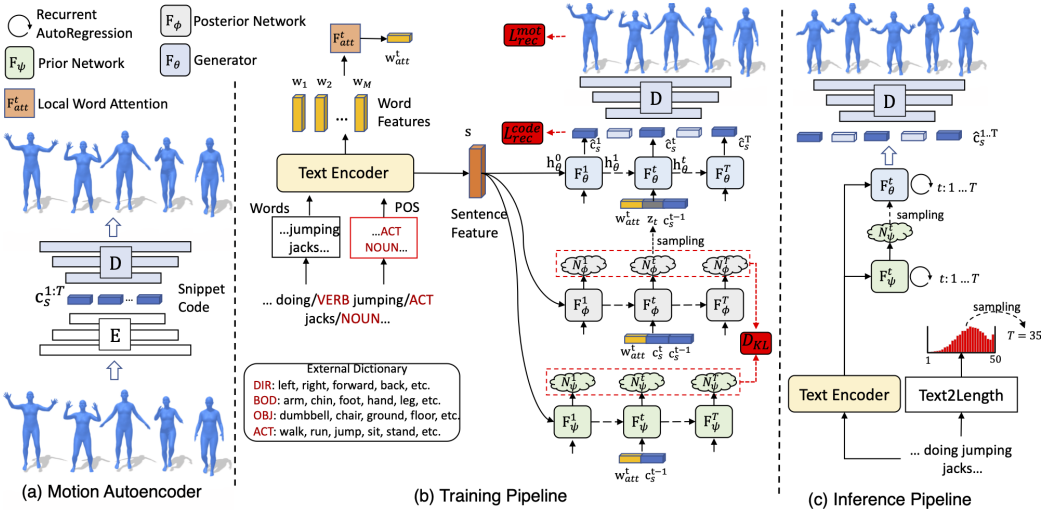
[*]Equal Contribution. ✉Corresponding author.
[†]Work done while being a research assistant at Monash University.

[1]Monash University [2]The Australian National University

[3]Mohamed bin Zayed University of Artificial Intelligence [4]Peking University

Talk @ miHoYo Shanghai, July 22, 2024

Text-to-Motion

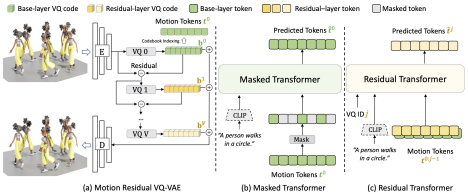


Guo et al. *Generating Diverse and Natural 3D Human Motions from Text* (CVPR 2022)

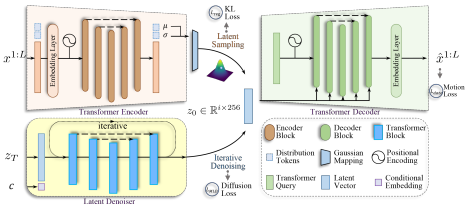
T2M for Now

- Autoregressive Method
  - Jiang et al. *MotionGPT: Human Motion as a Foreign Language* (NIPS 2023)
  - Pinyoanuntapong et al. *MMM: Generative Masked Motion Model* (CVPR 2024)
  - Guo et al. *MoMask: Generative Masked Modeling of 3D Human Motions* (CVPR 2024)
- Diffusion-based Method
  - Zhang et al. *MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model* (TPAMI 2024)
  - Tevet et al. *MDM: Human Motion Diffusion Model* (ICLR 2023)
  - Chen et al. *MLD: Executing your Commands via Motion Diffusion in Latent Space* (CVPR 2023)
  - Zhang et al. **Motion Mamba: Efficient and Long Sequence Motion Generation** (ECCV 2024)

MoMask vs MLD

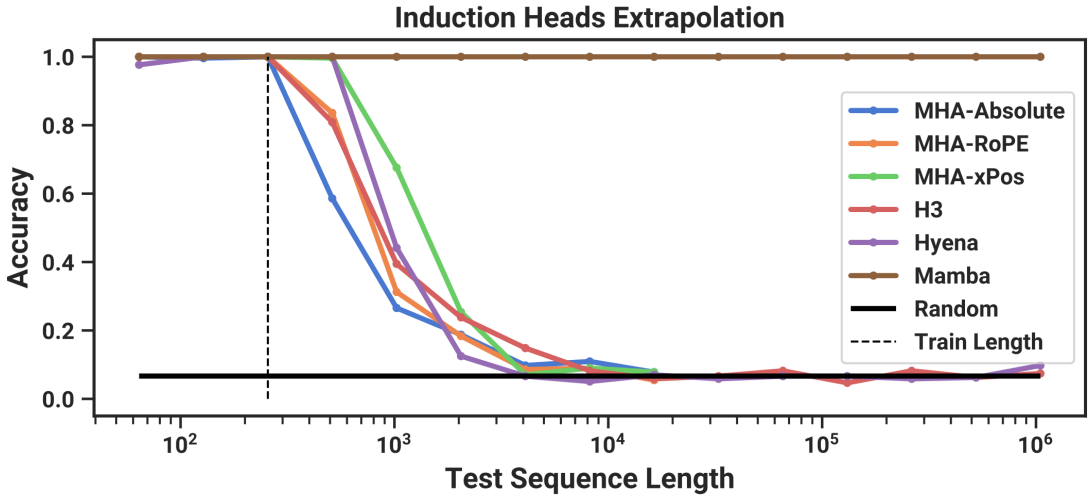


Guo et al. *MoMask: Generative Masked Modeling of 3D Human Motions* (CVPR 2024)



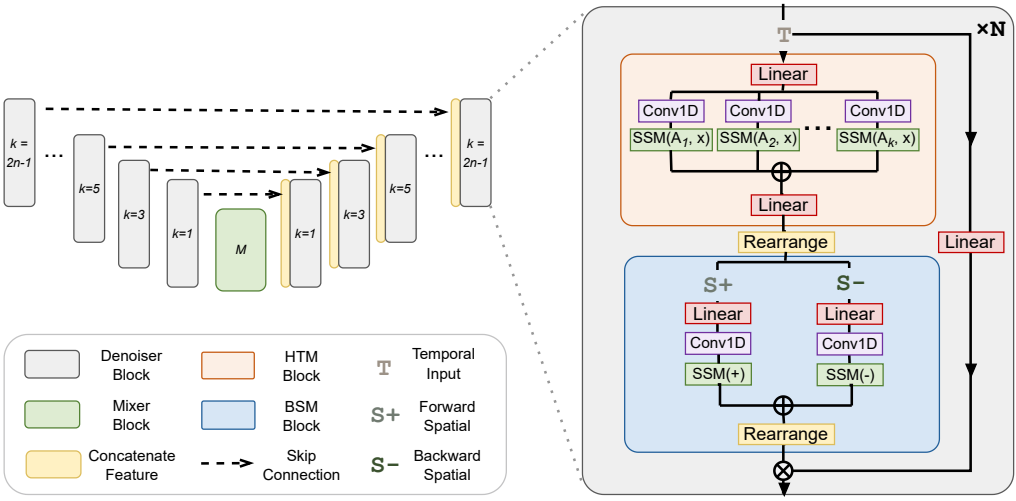
Chen et al. *MLD: Executing your Commands via Motion Diffusion in Latent Space* (CVPR 2023)

Motivation: Efficient and Long Sequence Modeling



Gu et al. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* (COLM 2024)

Motion Mamba Architecture



Zhang et al. *Motion Mamba: Efficient and Long Sequence Motion Generation* (ECCV 2024)

Comparative Studies on HumanML3D

Comparison of text-conditional motion synthesis on **HumanML3D**. The motion encoder from MLD evaluates these metrics. The evaluation results are sorted by descending FIDs. The right arrow $\rightarrow$ refers to that closer to real motion. **Bold** and underline indicate the best and second best result.

Method	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Seq2Seq	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
LJ2P	$0.246^{\pm.001}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2G	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
Hier	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
TEMOS	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
T2M	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
MotionDiffuse	$\underline{0.491}^{\pm.001}$	$\underline{0.681}^{\pm.001}$	$\underline{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$\underline{3.113}^{\pm.001}$	$\underline{9.410}^{\pm.049}$	$1.553^{\pm.042}$
MLD	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$\underline{0.473}^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$\underline{2.413}^{\pm.079}$
Motion Mamba (Ours)	$\mathbf{0.502}^{\pm.003}$	$\mathbf{0.693}^{\pm.002}$	$\mathbf{0.792}^{\pm.002}$	$\mathbf{0.281}^{\pm.009}$	$\mathbf{3.060}^{\pm.058}$	$9.871^{\pm.084}$	$2.294^{\pm.058}$

Comparative Studies on KIT-ML

We involve the **KIT-ML** dataset and evaluate the SOTA methods on the text-to-motion task. The evaluation results are sorted by descending FIDs. The right arrow $\rightarrow$ refers to that closer to real motion. **Bold** and underline indicate the best and second best result.

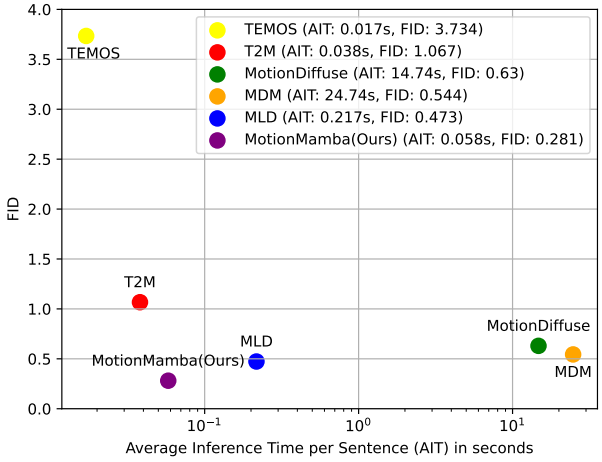
Method	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
Seq2Seq	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
T2G	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
LJ2P	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Hier	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	$2.090^{\pm.083}$
TEMOS	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$	$0.532^{\pm.034}$
T2M	$0.370^{\pm.005}$	$0.569^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$3.401^{\pm.008}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$
MDM	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
MotionDiffuse	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$\mathbf{2.958}^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
MLD	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$\mathbf{2.192}^{\pm.071}$
Motion Mamba (Ours)	$\mathbf{0.419}^{\pm.006}$	$\mathbf{0.645}^{\pm.005}$	$\mathbf{0.765}^{\pm.006}$	$\mathbf{0.307}^{\pm.041}$	$3.021^{\pm.025}$	$11.02^{\pm.098}$	$1.678^{\pm.064}$

Long Sequence Motion Generation

In order to evaluate the models' capability in long sequence motion generation, we compared our method with an existing approach on the recently introduced HumanML3D-LS dataset. This dataset comprises motion sequences longer than 190 frames from the original evaluation set. Our model demonstrates superior performance compared to other methods.

Method	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.437^{\pm.003}$	$0.622^{\pm.004}$	$0.721^{\pm.004}$	$0.004^{\pm.000}$	$3.343^{\pm.015}$	$8.423^{\pm.090}$	-
MDM	$0.368^{\pm.005}$	$0.553^{\pm.006}$	$0.672^{\pm.005}$	$\underline{0.802}^{\pm.044}$	$3.860^{\pm.025}$	$\underline{8.817}^{\pm.068}$	-
MotionDiffuse	$0.367^{\pm.004}$	$0.521^{\pm.004}$	$0.623^{\pm.004}$	$2.460^{\pm.062}$	$3.789^{\pm.005}$	$\mathbf{8.707}^{\pm.143}$	$1.602^{\pm.013}$
MLD	$\underline{0.403}^{\pm.005}$	$\underline{0.584}^{\pm.005}$	$\underline{0.690}^{\pm.005}$	$0.952^{\pm.020}$	$\underline{3.580}^{\pm.016}$	$9.050^{\pm.085}$	$\mathbf{2.711}^{\pm.104}$
Motion Mamba (Ours)	$\mathbf{0.417}^{\pm.003}$	$\mathbf{0.606}^{\pm.003}$	$\mathbf{0.713}^{\pm.004}$	$\mathbf{0.668}^{\pm.019}$	$\mathbf{3.435}^{\pm.015}$	$9.021^{\pm.070}$	$\underline{2.373}^{\pm.084}$

Efficiency


TEMOS (AIT: 0.017s, FID: 3.734)
T2M (AIT: 0.038s, FID: 1.067)
MotionDiffuse (AIT: 14.74s, FID: 0.63)
MDM (AIT: 24.74s, FID: 0.544)
MLD (AIT: 0.217s, FID: 0.473)
MotionMamba(Ours) (AIT: 0.058s, FID: 0.281)
TEMOS
T2M
MLD
MotionMamba(Ours)
MotionDiffuse
MDM
FID
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
10^{-1}
10^{0}
10^{1}
Average Inference Time per Sentence (AIT) in seconds

Visualization

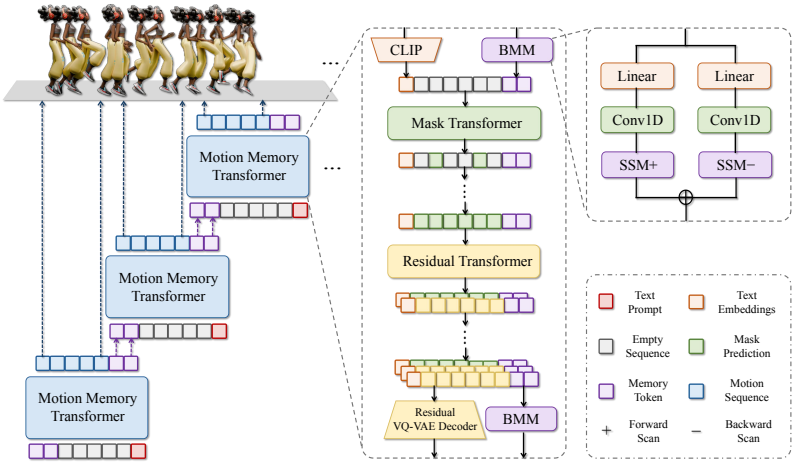
MotionDiffuse

MDM

MLD

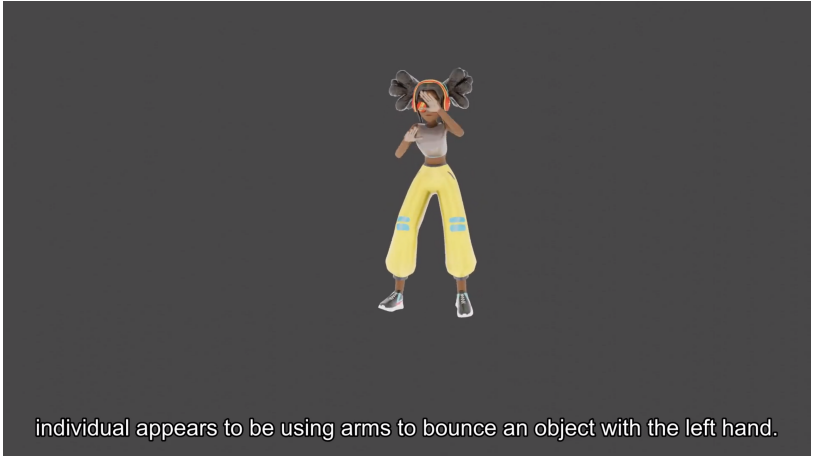
Motion Mamba (Ours)

Further Application: Arbitrary Long Motion Generation



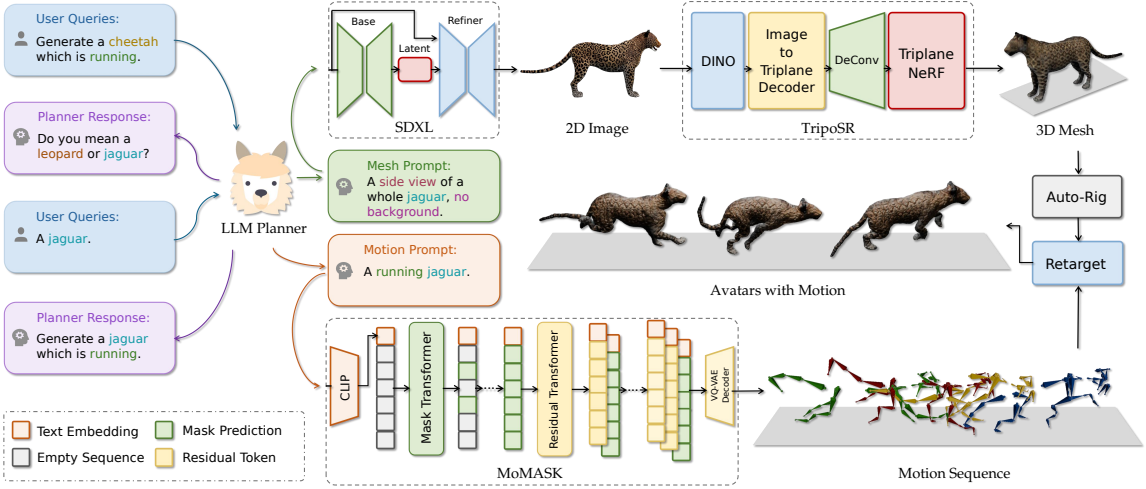
Zhang et al. *InfiniMotion: Mamba Boosts Memory in Transformer for Arbitrary Long Motion Generation*

Further Application: Arbitrary Long Motion Generation



Zhang et al. *InfiniMotion: Mamba Boosts Memory in Transformer for Arbitrary Long Motion Generation*

Further Application: Dynamic Avatar Generation



Zhang et al. *Motion Avatar: Generate Human and Animal Avatars with Arbitrary Motion* (BMVC 2024)

Further Application: Dynamic Avatar Generation

A demon girl with blue hair is running fast.

A demon girl with blue hair spins rapidly.

A demon girl with blue hair holds her left foot with her left hand.

Luigi is bending his leg.

Luigi is doing kung fu.

Luigi is rolling his hands.

Zhang et al. *Motion Avatar: Generate Human and Animal Avatars with Arbitrary Motion* (BMVC 2024)

Further Application: HOI & HSI

Jiang et al. *TRUMANS: Scaling Up Dynamic Human-Scene Interaction Modeling* (CVPR 2024)

Limitation on Text-to-Motion Generation

- Motion generation requires much stronger conditions for controllable and precise generation, such as trajectory, video, and even joint-level control.
- Current methods still produce unrealistic joint angles and jitters. Future work should focus on generating physically compatible motions for broader real-world applications, such as robot manipulation.
- Despite the existence of works like Motion Avatar and OmniMotionGPT for animal motion generation, significant challenges remain due to the lack of datasets and the variability in animal motion representation.

Q&A

Feel free to ask any questions. Thank you!

Motion Mamba Project Website