JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Executing your Commands via Motion Diffusion in Latent Space

Session: THU-AM-145

https://chenxin.tech/mld/

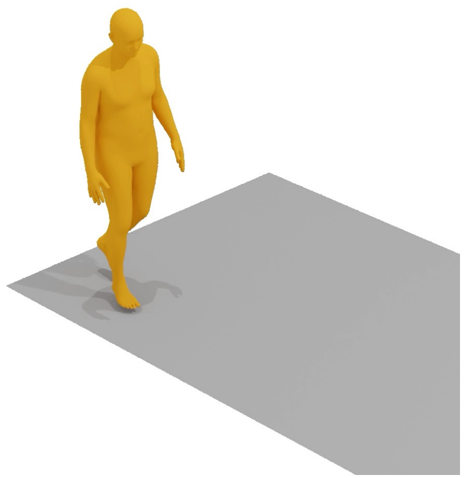
Xin Chen[1*], Biao Jiang[2*], Wen Liu[1], Zilong Huang[1], Bin Fu[1]

Tao Chen[2], Jingyi Yu[3], Gang Yu[1†]

Tencent腾讯 [1]Tencent PCG [2]Fudan University [3]ShanghaiTech University

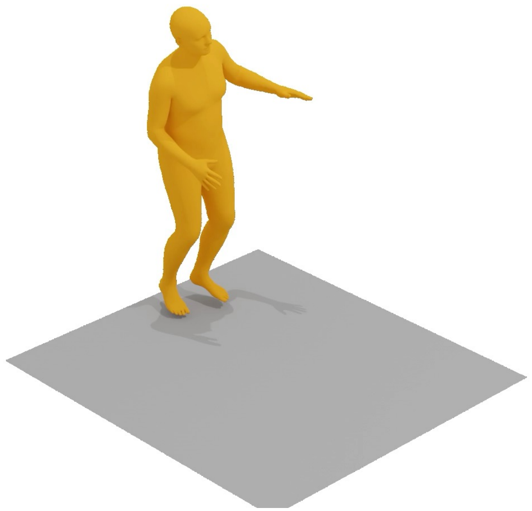
[*] Equal contribution.
[†] Corresponding author.

Motion Latent Diffusion



The person was doing a cool walk.

Warm up

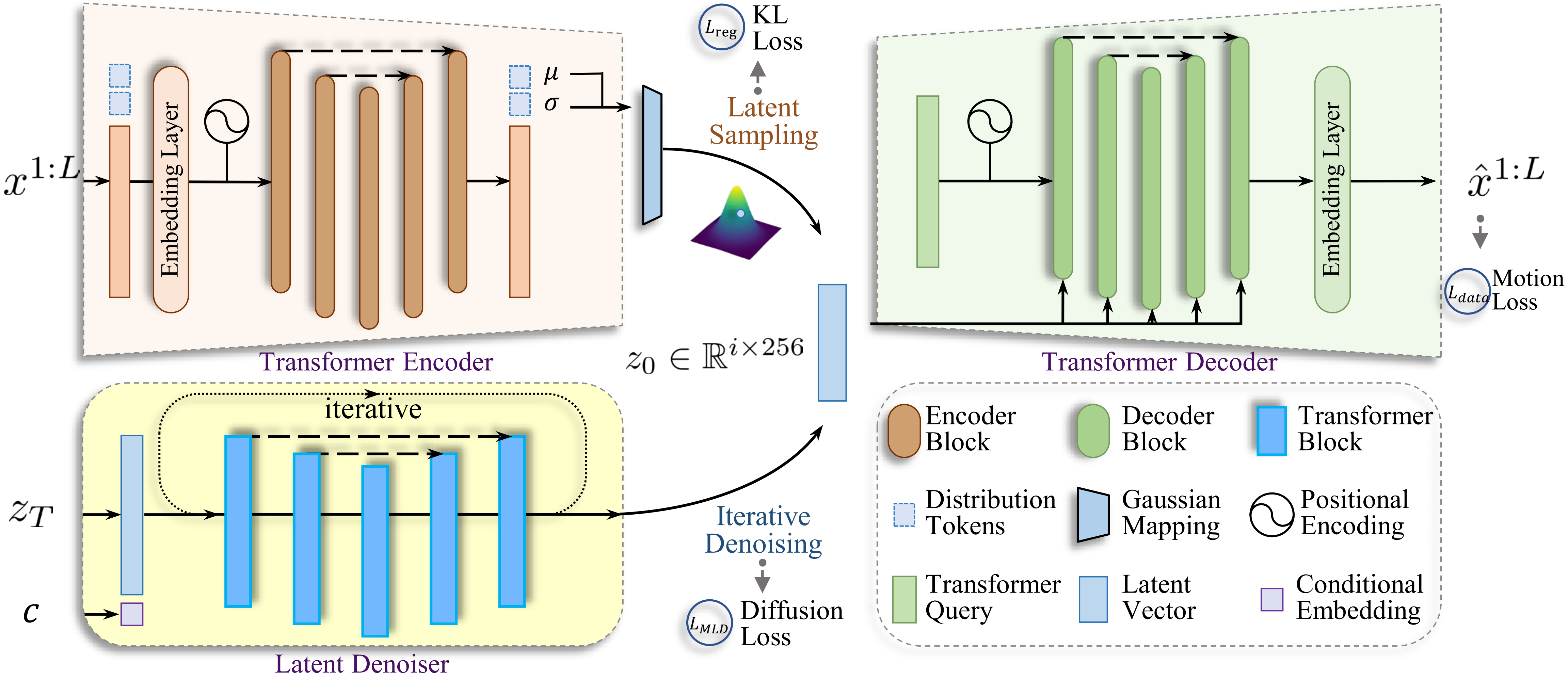


Text-Conditioned

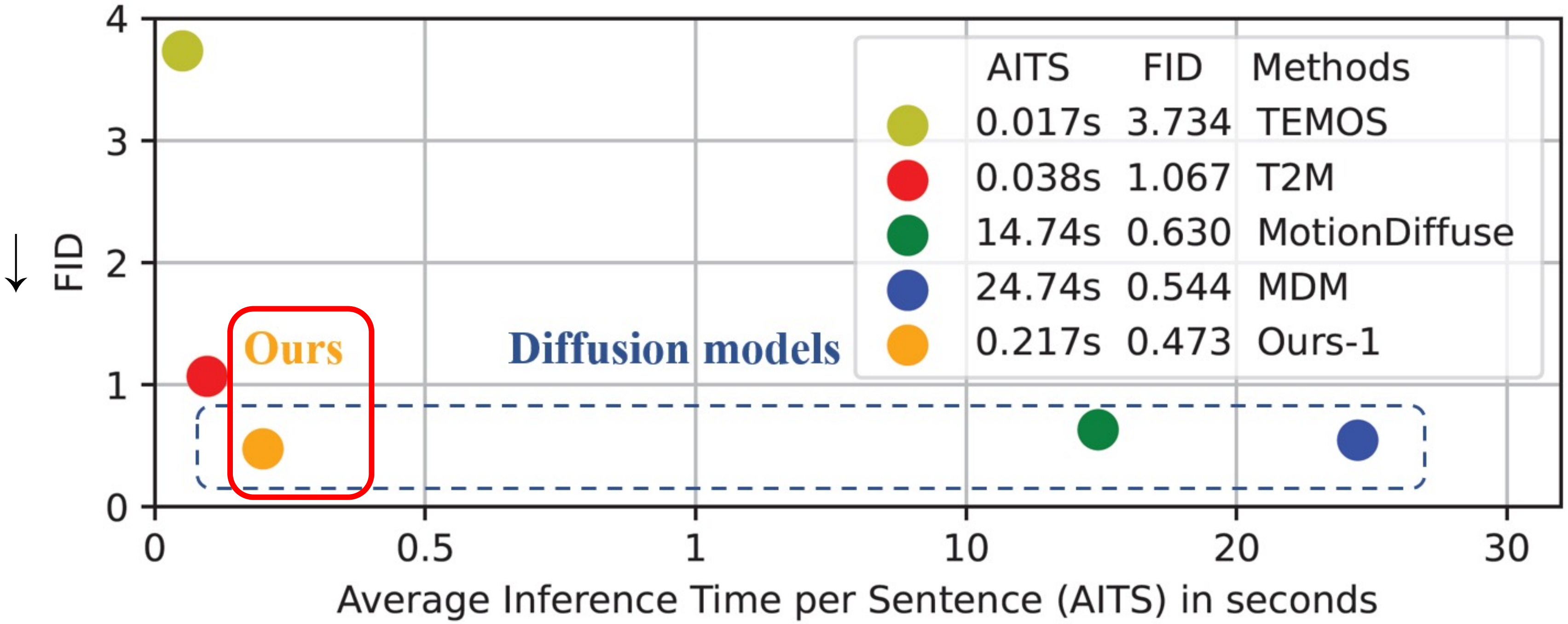
Action-Conditioned

Unconditioned

Motion Latent Diffusion



Quantitative Comparison



	AITS	FID	Methods
●	0.017s	3.734	TEMOS
●	0.038s	1.067	T2M
●	14.74s	0.630	MotionDiffuse
●	24.74s	0.544	MDM
●	0.217s	0.473	Ours-1

MLD requires less computational overhead (horizontal axis to the left), which is two orders of magnitude faster than other diffusion model-based methods, and has better motion quality (vertical axis to the bottom)

Prior Work

1. Unified hidden space

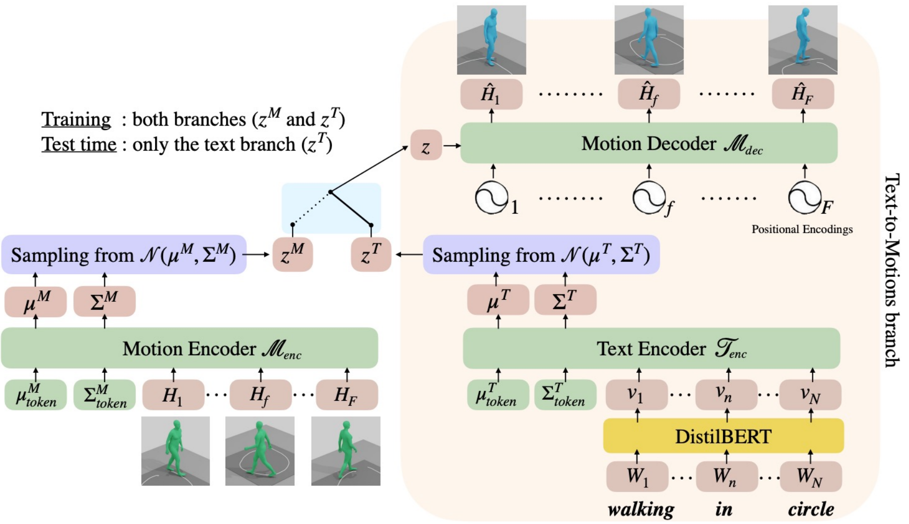
Limited to highly different distributions

2. Diffusion models on raw motion

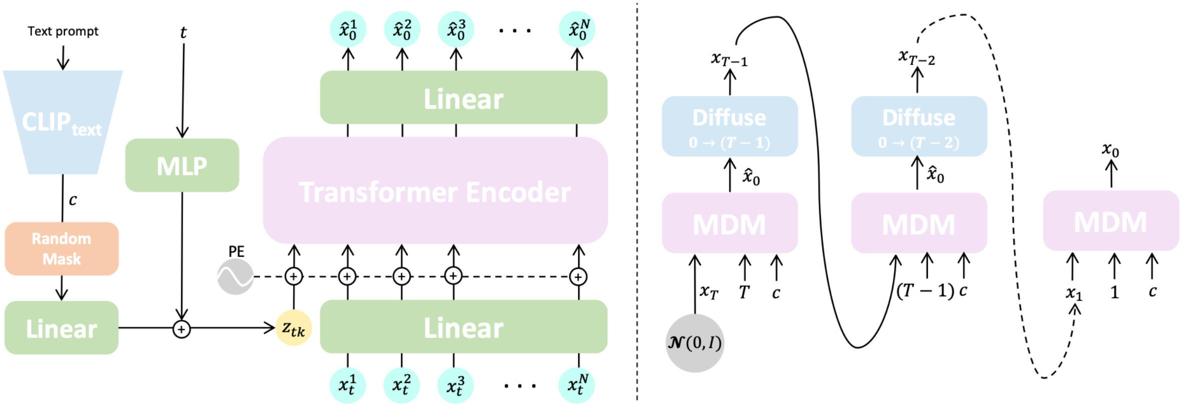
High computational complexity

Susceptible to artifacts

[TEMOS: Petrovich et al. ECCV 2022]

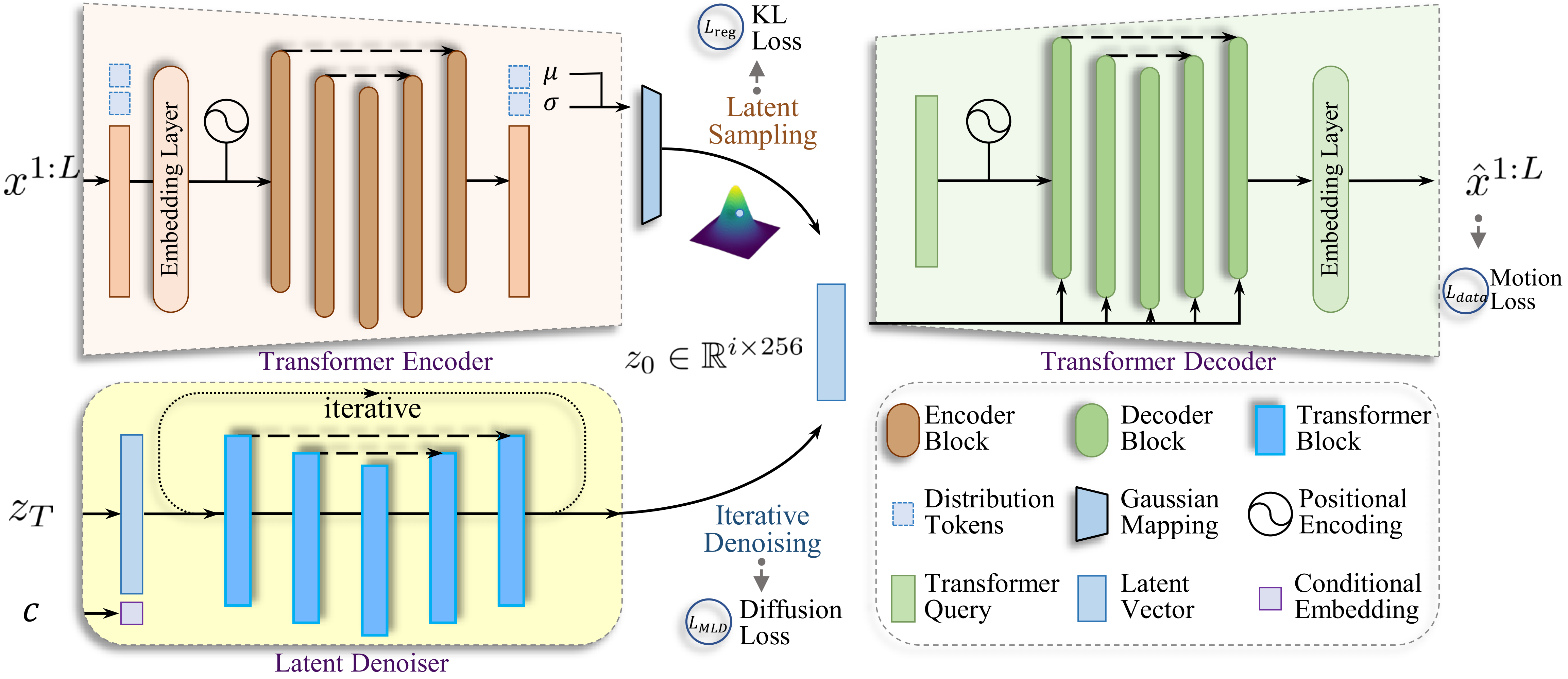

Training : both branches (z^M and z^T)
Test time : only the text branch (z^T)
$\hat{H}_1$
$\hat{H}_f$
$\hat{H}_F$
z
Motion Decoder $\mathcal{M}_{dec}$
Positional Encodings
Text-to-Motions branch
Sampling from $\mathcal{N}(\mu^M, \Sigma^M)$
z^M
z^T
Sampling from $\mathcal{N}(\mu^T, \Sigma^T)$
μ^M
Σ^M
μ^T
Σ^T
Motion Encoder $\mathcal{M}_{enc}$
Text Encoder $\mathcal{T}_{enc}$
μ^M_{token}
Σ^M_{token}
H_1
H_f
H_F
μ^T_{token}
Σ^T_{token}
v_1
v_n
v_N
DistilBERT
W_1
W_n
W_N
walking
in
circle

[MDM: Tevet et al. ICLR 2023]



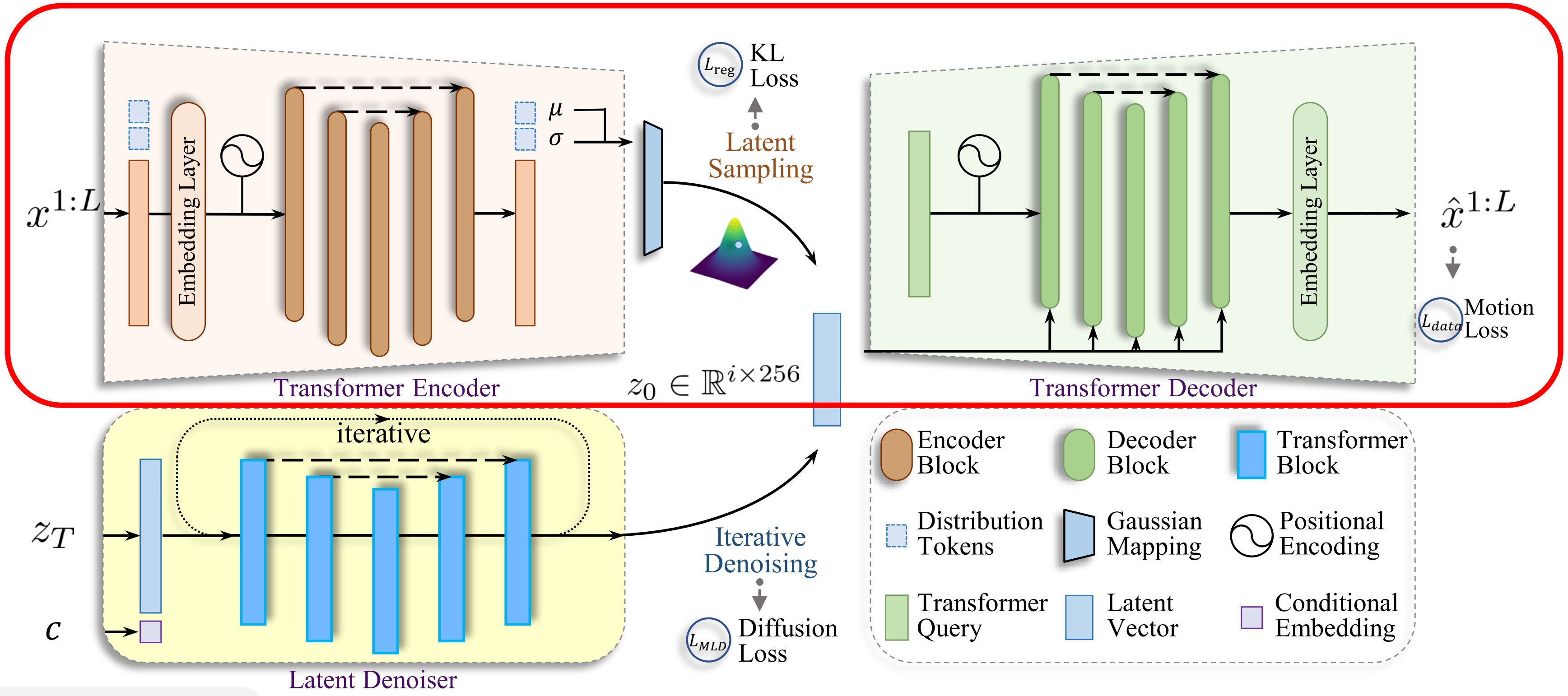
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motion Latent Diffusion



VAE

Learning latent motion representation



Advantages

1. Improve diversity

2. Reduce the effect of noise in the raw data

3. Reducing the amount of data facilitates the computational cost of learning subsequent text-to-action mappings

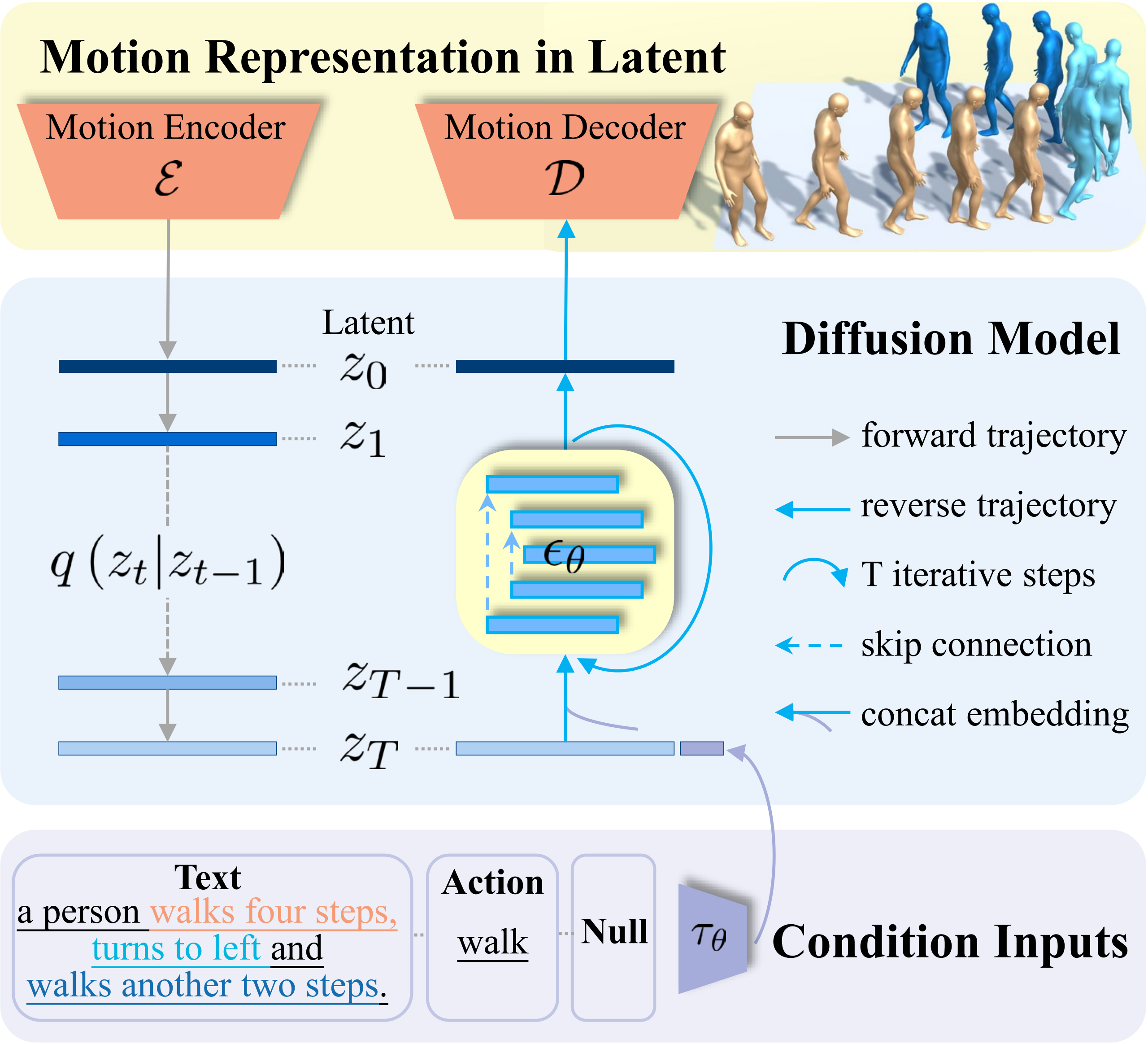
Latent Diffusion

Learning probabilistic mappings from input conditions (text, label, etc.) to hidden representation

Advantages

1. Vivid motions matching conditions

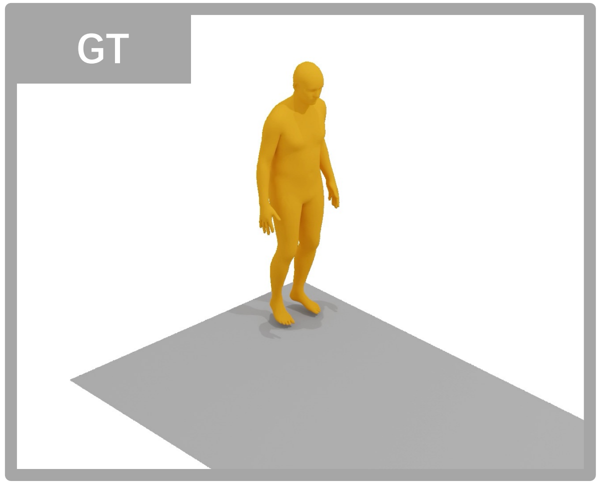
2. Reduced computational overhead

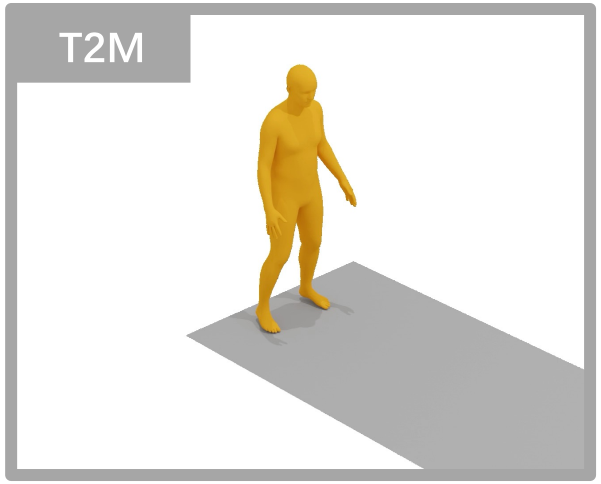


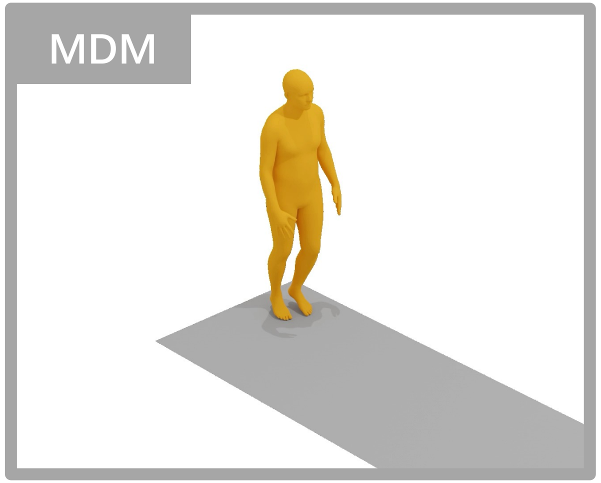
Comparison

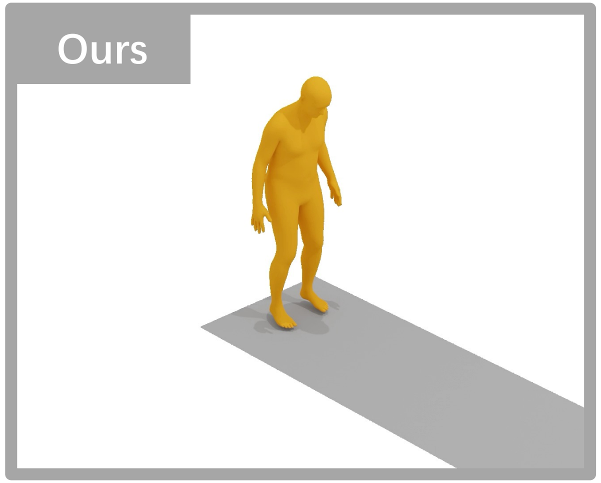
Text-to-Motion

“walking forward with legs wide apart.”





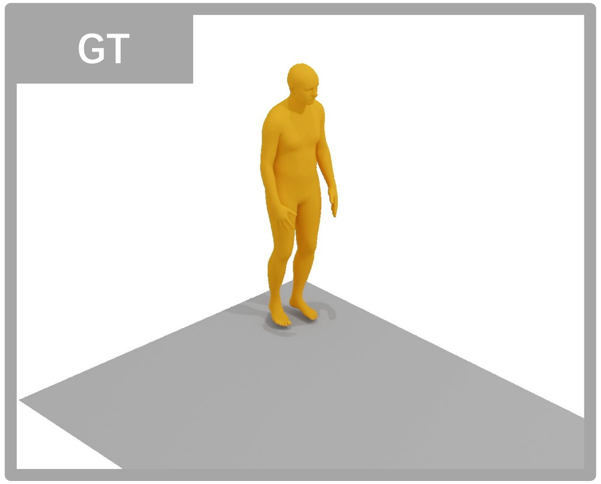




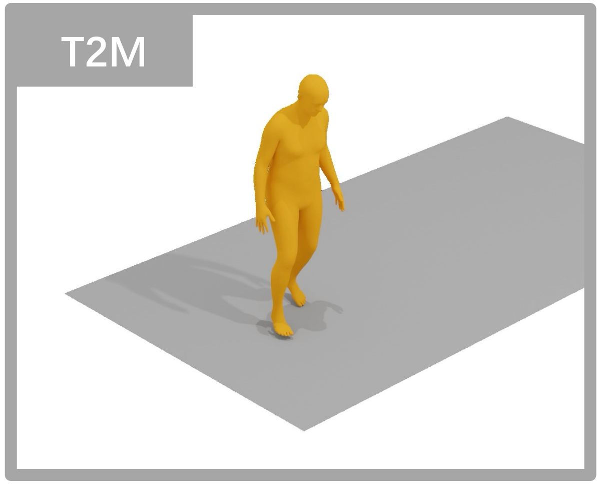
Comparison

Text-to-Motion

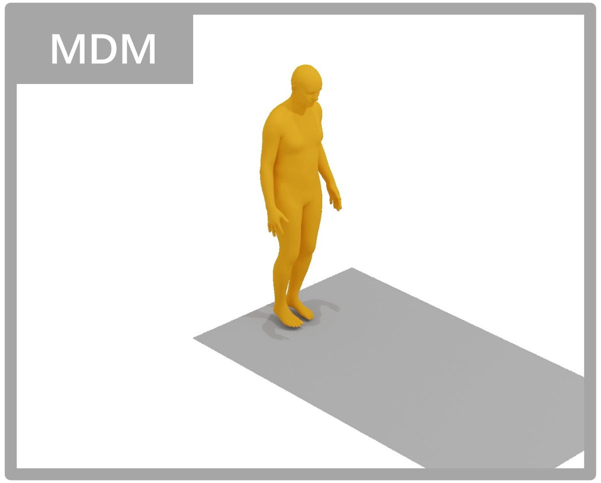
"the person was doing a cool walk"



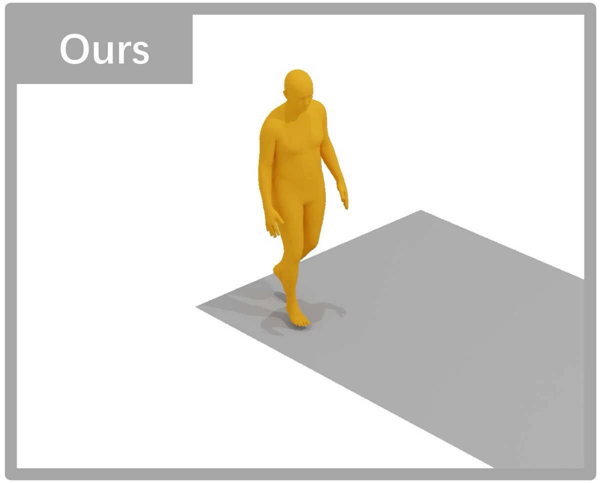
GT



T2M



MDM

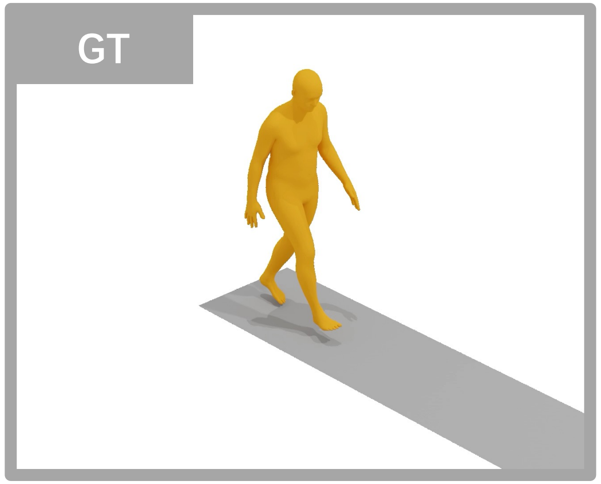


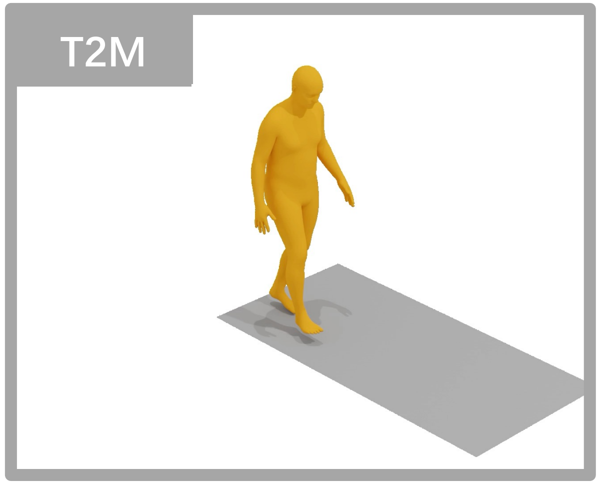
Ours

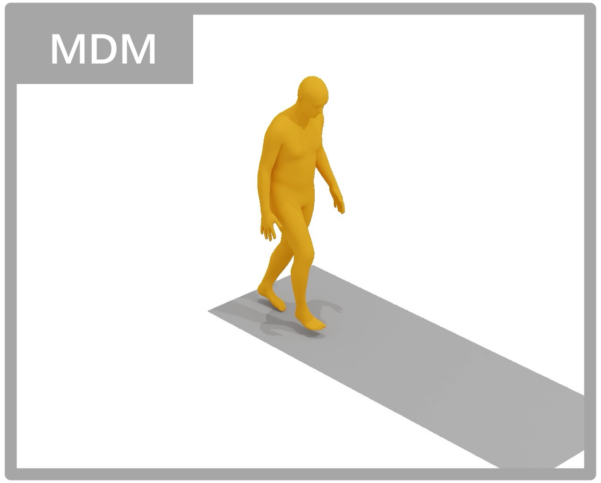
Comparison

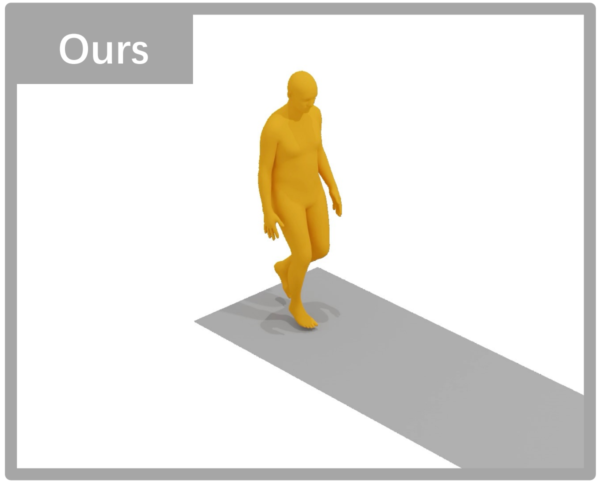
Text-to-Motion

“a person walks forward, turns, then sits, then stands and walks back”


GT


T2M

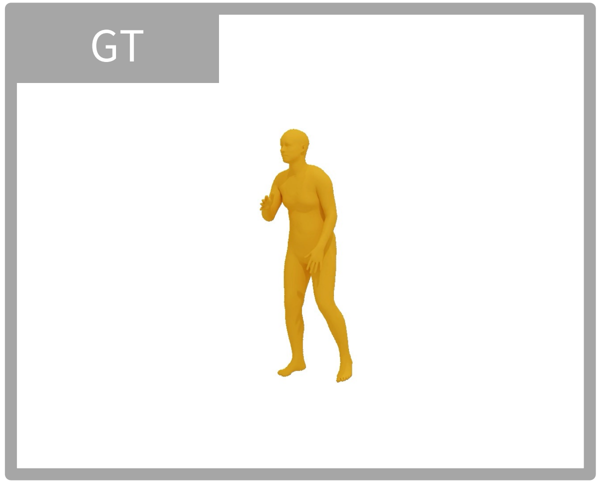

MDM


Ours

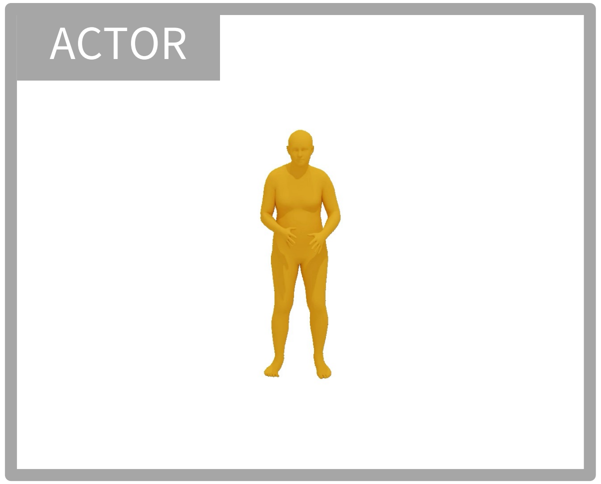
Comparison

Action-to-Motion

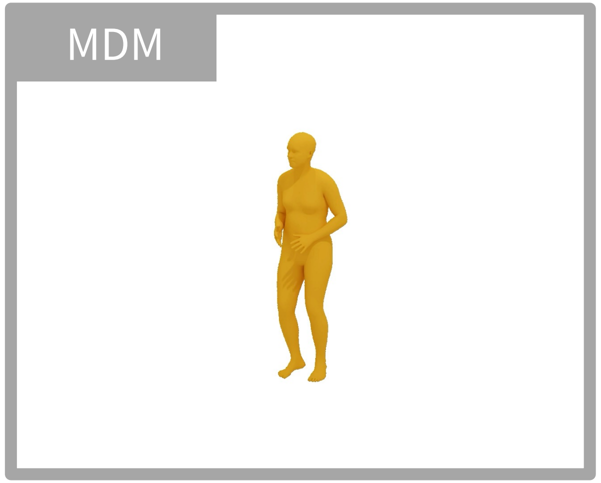
“**Drink**”



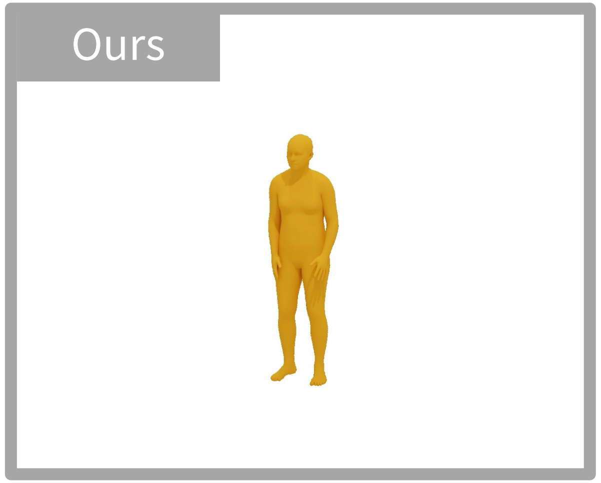
GT



ACTOR



MDM

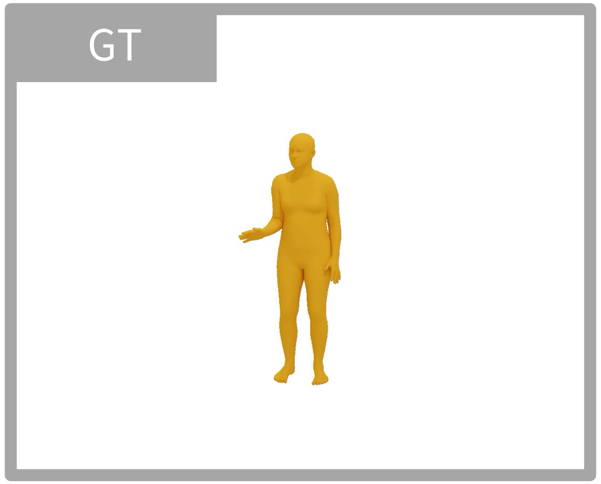


Ours

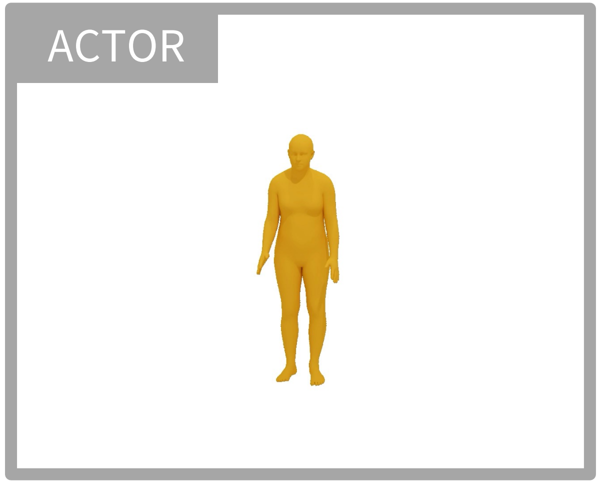
Comparison

Action-to-Motion

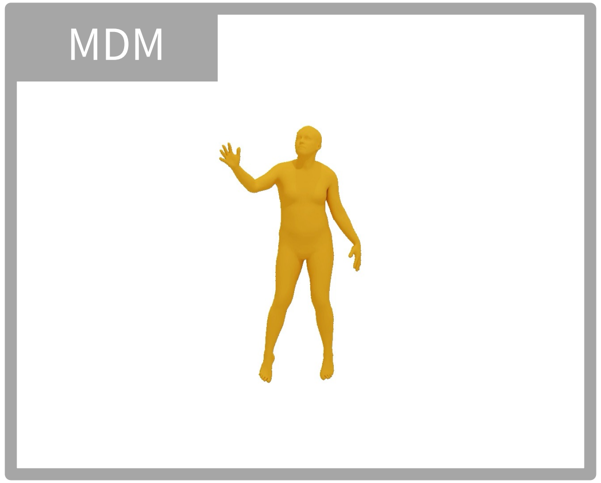
"Lift dumbbell"



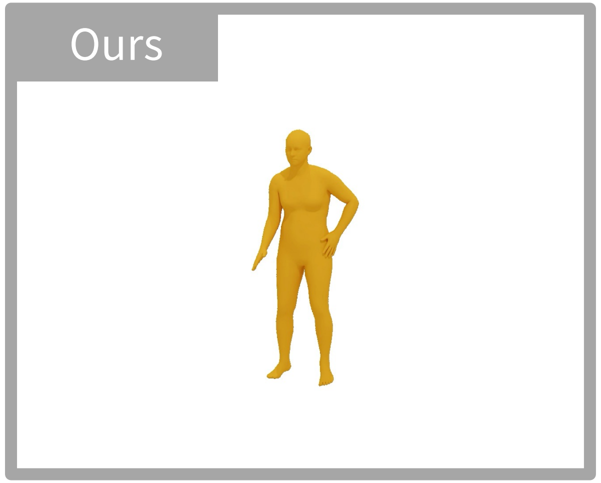
GT



ACTOR

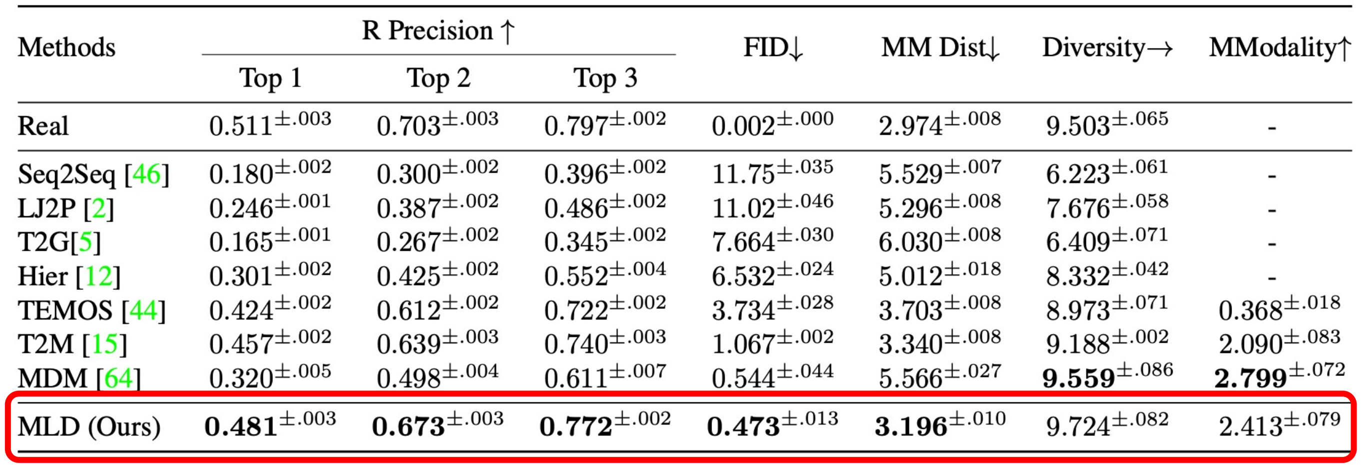


MDM



Ours

Quantitative Comparison



Methods	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 [46]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
LJ2P [2]	$0.246^{\pm.001}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2G[5]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
Hier [12]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
TEMOS [44]	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
T2M [15]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [64]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	**9.559**$^{\pm.086}$	**2.799**$^{\pm.072}$
MLD (Ours)	**0.481**$^{\pm.003}$	**0.673**$^{\pm.003}$	**0.772**$^{\pm.002}$	**0.473**$^{\pm.013}$	**3.196**$^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$

Performance on text-to-motion generation tasks

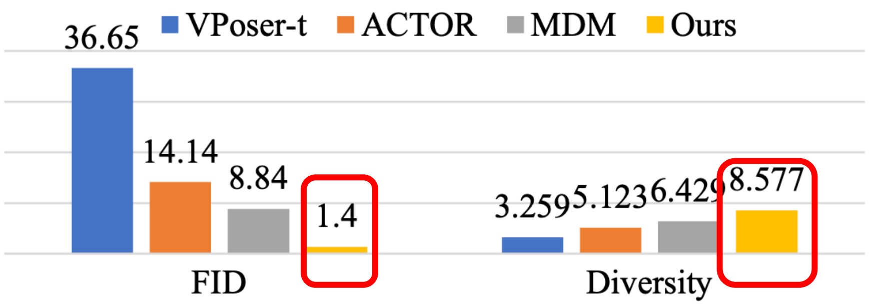
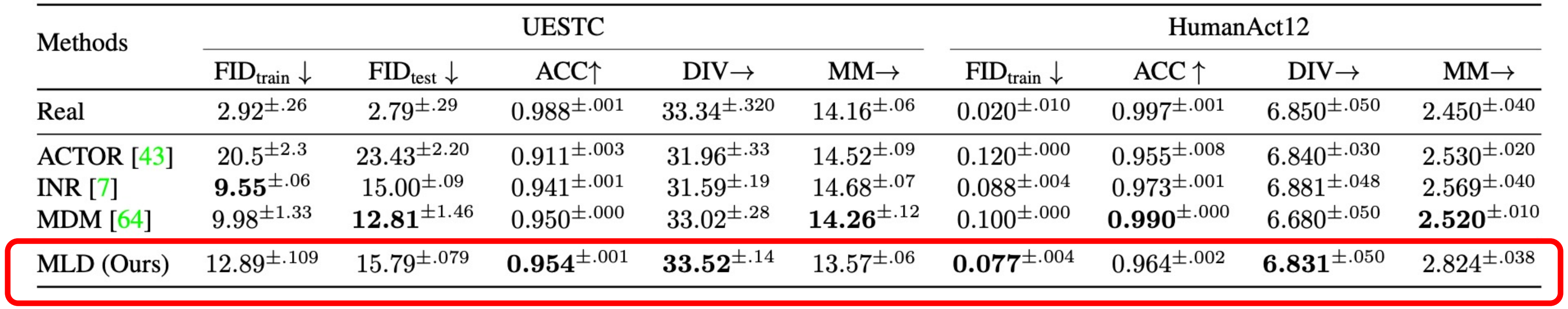


Figure 5. Comparison of unconditional motion generation on part of AMASS [39] dataset with the state-of-the-art methods. We provide both FID and Diversity to evaluate generated motions.

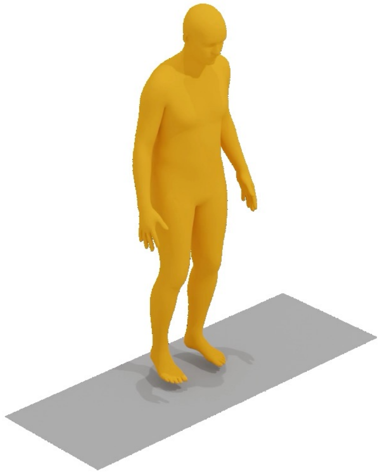


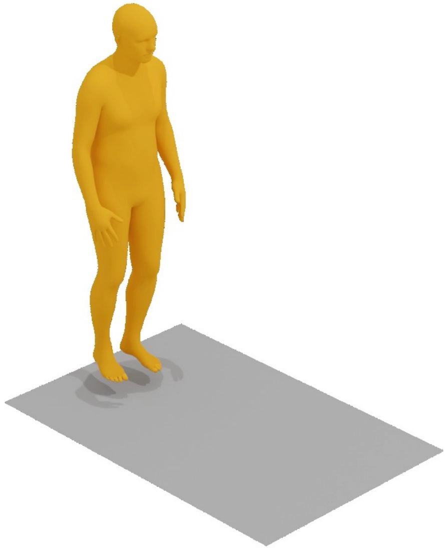
Methods	UESTC					HumanAct12			
	$\text{FID}_{\text{train}}$ ↓	FID_{test} ↓	ACC↑	DIV→	MM→	$\text{FID}_{\text{train}}$ ↓	ACC ↑	DIV→	MM→
Real	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
ACTOR [43]	$20.5^{\pm2.3}$	$23.43^{\pm2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$6.840^{\pm.030}$	$2.530^{\pm.020}$
INR [7]	**9.55**$^{\pm.06}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$
MDM [64]	$9.98^{\pm1.33}$	**12.81**$^{\pm1.46}$	$0.950^{\pm.000}$	$33.02^{\pm.28}$	**14.26**$^{\pm.12}$	$0.100^{\pm.000}$	**0.990**$^{\pm.000}$	$6.680^{\pm.050}$	**2.520**$^{\pm.010}$
MLD (Ours)	$12.89^{\pm.109}$	$15.79^{\pm.079}$	**0.954**$^{\pm.001}$	**33.52**$^{\pm.14}$	$13.57^{\pm.06}$	**0.077**$^{\pm.004}$	$0.964^{\pm.002}$	**6.831**$^{\pm.050}$	$2.824^{\pm.038}$

Performance on action-to-motion generation tasks

More Results

Text-to-Motion





"A person doing jumping jacks."

"A person walks in a circle to their right."

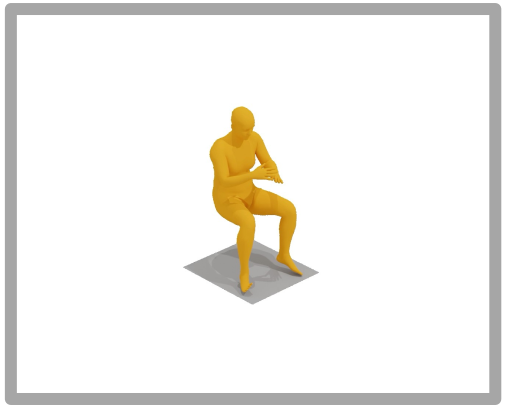
"A person jumps forwards and turns left in mid air"

Tencent 腾讯

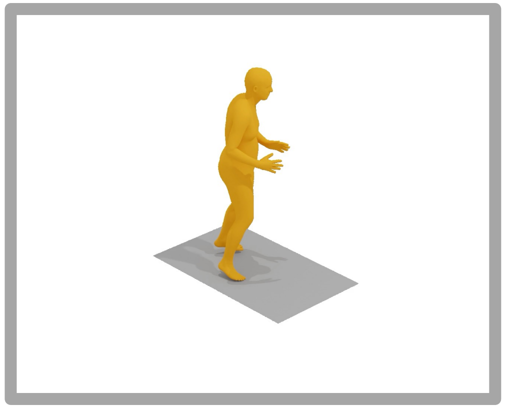
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

More Results

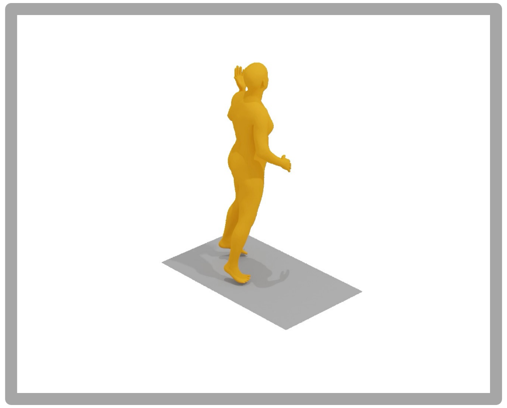
Action-to-Motion



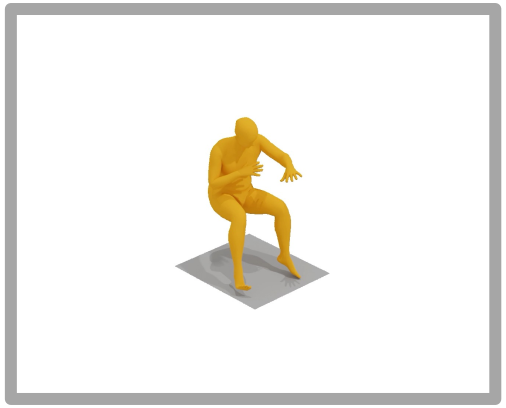
“Eat”



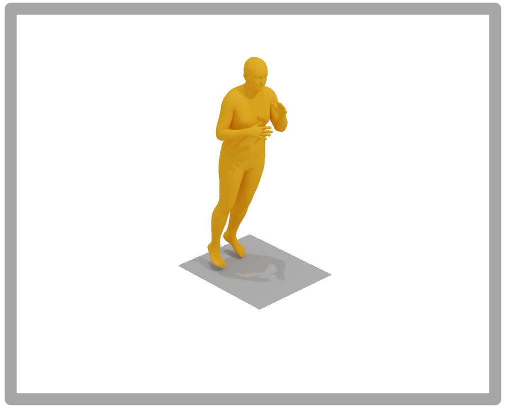
“Boxing”



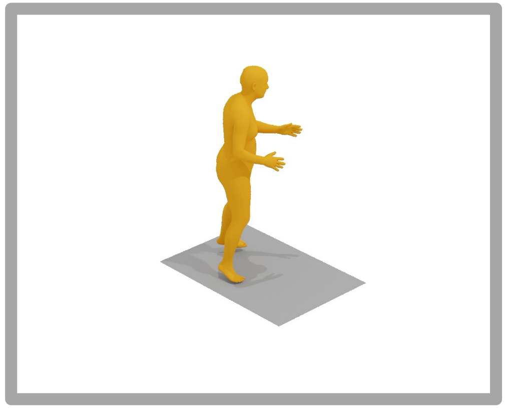
“Throw”



“Eat”



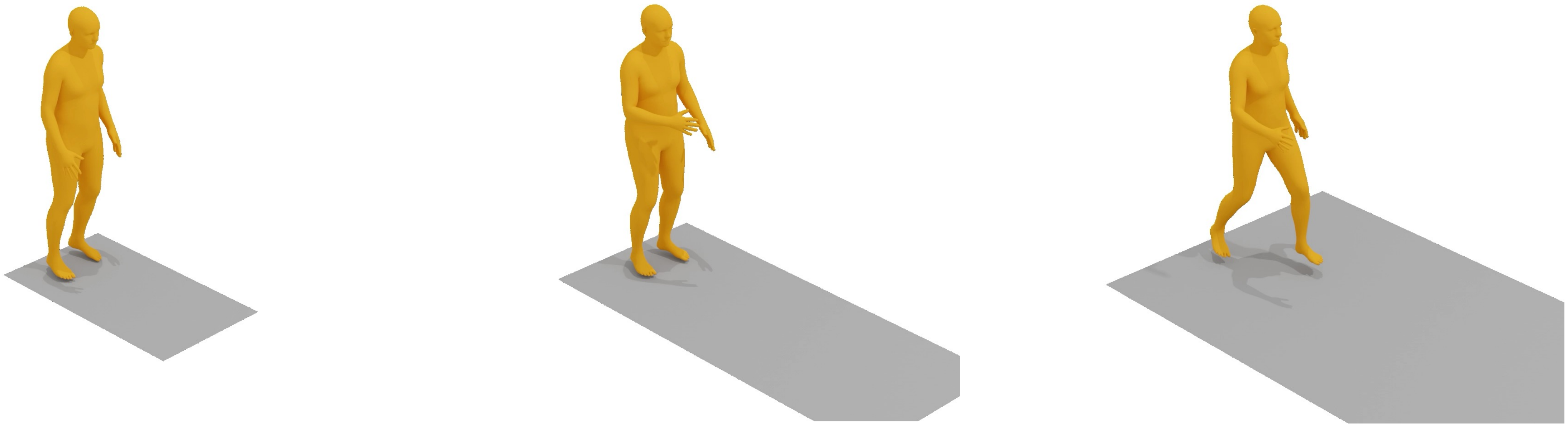
“Boxing”



“Throw”

More Results

Unconditioned



JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

Executing your Commands via Motion Diffusion in Latent Space

Thanks for Watching!

More details please check our paper and project

Project

Paper

Code

FUDAN UNIVERSITY 1905

上海科技大学 立志成才 报国裕民 2013 SHANGHAITECH UNIVERSITY

Tencent 腾讯