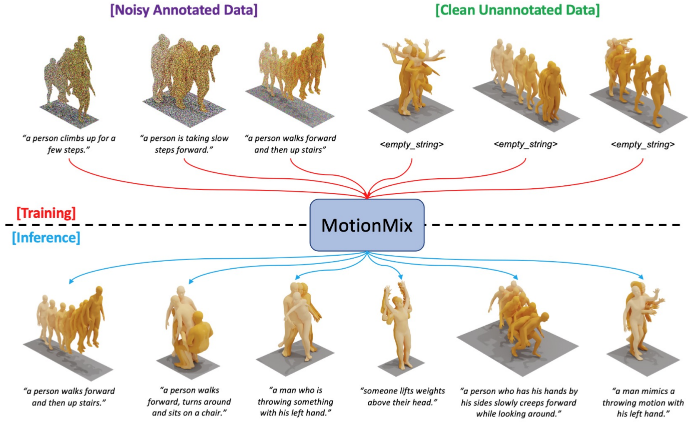
MotionMix: Weakly-Supervised Diffusion for Controllable Motion Generation

Nhat M. Hoang[1], Kehong Gong[2], Chuan Guo[2], Michael Bi Mi[2]
[1]Nanyang Technological University, [2]Huawei Technologies Co., Ltd.

Scan for motion videos

Overview

- Controllable generation of 3D human motions becomes an important topic as the world embraces digital transformation.
- Existing methods heavily rely on costly, annotated high-quality motion data.
- We propose MotionMix, a simple yet effective weakly-supervised approach for diffusion model to utilize both noisy and unannotated motion sequences.



Setup

Motion generation tasks and benchmarks:

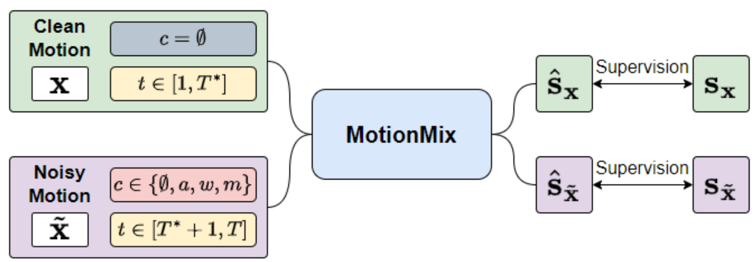
- **Text-to-Motion (T2M):** HumanML3D and KIT-ML datasets
- **Action-to-Motion (A2M):** HumanAct12 and UESTC datasets
- **Music-to-Dance (M2D):** AIST++ dataset

MotionMix was applied to *sota* diffusion model of different designs:

- **MDM** for T2M, A2M: x0-parameterization, classifier-free guidance.
- **MotionDiffuse** for T2M : epsilon-parameterization.
- **EDGE** for M2D : x0-parameterization with classifier-free guidance.

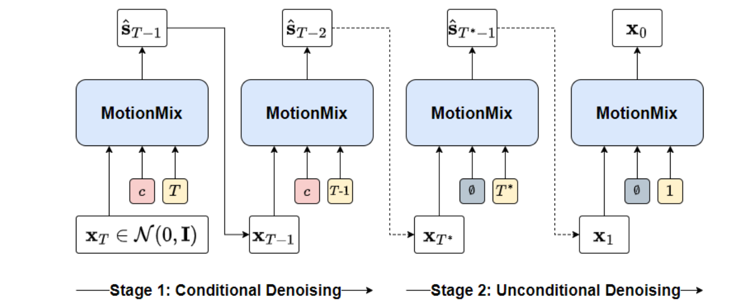
Methodology

(a) Training: Separate the diffusion steps into two distinct ranges regarding the data type. Clean samples are guided by an empty condition. Ground-truth s_x refers to both x-prediction and epsilon-prediction, two fundamental objective of diffusion models.



(b) Inference (two stages):

- Stage 1: generate rough motion guided by the conditional input.
- Stage 2: refine these rough motions to high-quality ones while the conditional input is masked.



Evaluation Results

Text-to-Motion

Dataset	Method	R Precision (top 3)↑	FID↓	Multimodal Dist.↓	Diversity→	Multimodality↑
ML3D	Real Motion	0.797$^{\pm.002}$	0.002$^{\pm.000}$	2.974$^{\pm.008}$	9.503$^{\pm.065}$	-
ML3D	MDM	0.611$^{\pm.007}$	0.544$^{\pm.440}$	5.566$^{\pm.027}$	9.559$^{\pm.860}$	2.799$^{\pm.072}$
ML3D	MDM (**MotionMix**)	0.632$^{\pm.006}$ (↑3.4%)	0.381$^{\pm.042}$ (↑30.0%)	5.325$^{\pm.026}$ (↑4.3%)	9.520$^{\pm.090}$ (↑69.6%)	2.718$^{\pm.019}$ (↓2.9%)
KIT-ML	Real Motion	0.779$^{\pm.006}$	0.031$^{\pm.004}$	2.788$^{\pm.012}$	11.080$^{\pm.097}$	-
KIT-ML	MDM	0.396$^{\pm.004}$	0.497$^{\pm.021}$	9.191$^{\pm.022}$	10.847$^{\pm.109}$	1.907$^{\pm.214}$
KIT-ML	MDM (**MotionMix**)	0.404$^{\pm.005}$ (↑2.0%)	0.322$^{\pm.020}$ (↑35.2%)	9.068$^{\pm.019}$ (↑1.3%)	10.781$^{\pm.098}$ (↓28.3%)	1.946$^{\pm.019}$ (↑2.0%)
KIT-ML	MotionDiffuse	0.739$^{\pm.004}$	1.954$^{\pm.062}$	2.958$^{\pm.005}$	11.100$^{\pm.143}$	0.730$^{\pm.013}$
KIT-ML	MotionDiffuse (**MotionMix**)	0.742$^{\pm.005}$ (↑0.4%)	1.192$^{\pm.073}$ (↑39.0%)	3.066$^{\pm.018}$ (↓3.6%)	10.998$^{\pm.072}$ (↓310%)	1.391$^{\pm.111}$ (↑90.5%)

Real | **MDM** | **MDM(MotionMix)** | **MotionDiffuse** | **MotionDiffuse (MotionMix)**

"a man mimics a throwing motion with his left hand."

Qualitative samples from HumanML3D test set. Please view videos on our webpage.

Action-to-Motion

Dataset	Method	FID ↓	Accuracy ↑	Diversity →	MultiModality →
HumanAct12	Real Motion	0.053$^{\pm.003}$	0.995$^{\pm.001}$	6.835$^{\pm.045}$	2.604$^{\pm.040}$
HumanAct12	Action2Motion	0.338$^{\pm.015}$	0.917$^{\pm.001}$	6.850$^{\pm.050}$	2.511$^{\pm.023}$
HumanAct12	ACTOR	0.120$^{\pm.000}$	0.955$^{\pm.008}$	6.840$^{\pm.030}$	2.530$^{\pm.020}$
HumanAct12	MLD	0.077$^{\pm.004}$	0.964$^{\pm.002}$	6.831$^{\pm.050}$	2.824$^{\pm.038}$
HumanAct12	MDM	0.100$^{\pm.000}$	0.990$^{\pm.000}$	6.860$^{\pm.050}$	2.520$^{\pm.010}$
HumanAct12	MDM (**MotionMix**)	0.196$^{\pm.007}$ (↓96%)	0.930$^{\pm.003}$ (↓6.1%)	6.836$^{\pm.062}$ (↑96%)	3.043$^{\pm.054}$ (↓422.6%)
UESTC	Real Motion	2.790$^{\pm.290}$	0.988$^{\pm.001}$	33.349$^{\pm.320}$	14.160$^{\pm.060}$
UESTC	ACTOR	23.430$^{\pm2.200}$	0.911$^{\pm.003}$	31.960$^{\pm.330}$	14.520$^{\pm.090}$
UESTC	MLD	15.790$^{\pm.079}$	0.954$^{\pm.001}$	33.520$^{\pm.140}$	13.570$^{\pm.060}$
UESTC	MDM	12.810$^{\pm1.460}$	0.950$^{\pm.000}$	33.100$^{\pm.290}$	14.260$^{\pm.120}$
UESTC	MDM (**MotionMix**)	11.400$^{\pm.393}$ (↑11%)	0.960$^{\pm.003}$ (↑1.1%)	32.806$^{\pm.176}$ (↓118%)	14.277$^{\pm.094}$ (↓17%)

Music-to-Dance

Method	PFC ↓	Beat Align. ↑	$Dist_k$ →	$Dist_g$ →
Real Motion	1.380	0.314	9.545	7.766
Bailando	1.754	0.23	10.58	7.72
FACT	2.2543	0.22	10.85	6.14
EDGE†	1.605$^{\pm.224}$	0.224$^{\pm.025}$	5.549$^{\pm.783}$	4.831$^{\pm.752}$
EDGE (**MotionMix**)	1.988$^{\pm.120}$ (↓21.32%)	0.256$^{\pm.013}$ (↑13.3%)	10.103$^{\pm2.039}$ (↑95.0%)	6.595$^{\pm.173}$ (↑15.1%)

Ablation Studies

Ablation on Denoising Pivot

→ More clean data does not lead to better performance while more annotated data

Method	R Precision (top 3)↑	FID↓	Multimodal Dist.↓	Diversity→	Multimodality↑
Real Motion	0.797$^{\pm.002}$	0.002$^{\pm.000}$	2.974$^{\pm.008}$	9.503$^{\pm.065}$	-
MDM	<u>0.611$^{\pm.007}$</u>	0.544$^{\pm.440}$	5.566$^{\pm.027}$	<u>9.559$^{\pm.860}$</u>	2.799$^{\pm.072}$
50% noisy, T_1=20, T_2=60					
MDM (**MotionMix**) *(T^*=0)*	0.598$^{\pm.006}$	0.714$^{\pm.045}$	<u>5.503$^{\pm.036}$</u>	9.750$^{\pm.123}$	**3.044$^{\pm.054}$**
MDM (**MotionMix**) *(T^*=20)*	0.601$^{\pm.005}$	0.497$^{\pm.048}$	5.562$^{\pm.026}$	9.414$^{\pm.092}$	<u>2.935$^{\pm.059}$</u>
MDM (**MotionMix**) *(T^*=40)*	0.604$^{\pm.008}$	<u>0.402$^{\pm.032}$</u>	5.524$^{\pm.033}$	9.396$^{\pm.094}$	2.747$^{\pm.070}$
MDM (**MotionMix**) *(T^*=60)*	**0.632$^{\pm.006}$**	**0.381$^{\pm.042}$**	**5.325$^{\pm.026}$**	**9.520$^{\pm.090}$**	2.718$^{\pm.019}$
MDM (**MotionMix**) *(T^*=80)*	0.594$^{\pm.005}$	0.589$^{\pm.059}$	5.670$^{\pm.033}$	9.242$^{\pm.086}$	2.602$^{\pm.057}$

Ablation on Noisy Ratio

→ More clean data ≠ better, more annotated data (even noisy) = better

Method	R Precision (top 3)↑	FID↓	Multimodal Dist.↓	Diversity→	Multimodality↑
Real Motion	0.797$^{\pm.002}$	0.002$^{\pm.000}$	2.974$^{\pm.008}$	9.503$^{\pm.065}$	-
MDM	0.611$^{\pm.007}$	0.544$^{\pm.440}$	5.566$^{\pm.027}$	9.559$^{\pm.860}$	2.799$^{\pm.072}$
T_1=20, T_2=60, T^*=60					
MDM (**MotionMix**) *(30% noisy)*	0.601$^{\pm.007}$	0.898$^{\pm.045}$	5.581$^{\pm.030}$	9.080$^{\pm.092}$	<u>2.856$^{\pm.074}$</u>
MDM (**MotionMix**) *(50% noisy)*	**0.632$^{\pm.006}$**	<u>0.381$^{\pm.042}$</u>	**5.325$^{\pm.026}$**	**9.520$^{\pm.090}$**	2.718$^{\pm.019}$
MDM (**MotionMix**) *(70% noisy)*	<u>0.615$^{\pm.006}$</u>	**0.359$^{\pm.030}$**	<u>5.545$^{\pm.031}$</u>	<u>9.457$^{\pm.098}$</u>	**2.867$^{\pm.107}$**

Ablation on Noisy Range

→ MotionMix is robust on different levels of corrupted motions

Method	R Precision (top 3)↑	FID↓	Multimodal Dist.↓	Diversity→	Multimodality↑
Real Motion	0.797$^{\pm.002}$	0.002$^{\pm.000}$	2.974$^{\pm.008}$	9.503$^{\pm.065}$	-
MDM	0.611$^{\pm.007}$	0.544$^{\pm.440}$	5.566$^{\pm.027}$	9.559$^{\pm.860}$	2.799$^{\pm.072}$
50% noisy, $T^* = T_2$					
MDM (**MotionMix**) *(T_1=20, T_2=40)*	<u>0.616$^{\pm.006}$</u>	<u>0.451$^{\pm.033}$</u>	<u>5.459$^{\pm.027}$</u>	9.585$^{\pm.101}$	2.585$^{\pm.076}$
MDM (**MotionMix**) *(T_1=20, T_2=60)*	**0.632$^{\pm.006}$**	**0.381$^{\pm.042}$**	**5.325$^{\pm.026}$**	**9.520$^{\pm.090}$**	<u>2.718$^{\pm.019}$</u>
MDM (**MotionMix**) *(T_1=20, T_2=80)*	0.604$^{\pm.004}$	0.614$^{\pm.060}$	5.540$^{\pm.024}$	<u>9.554$^{\pm.104}$</u>	**2.768$^{\pm.095}$**
50% noisy, $T^* = T_2$					
MDM (**MotionMix**) *(T_1=10, T_2=30)*	0.592$^{\pm.008}$	0.713$^{\pm.048}$	5.633$^{\pm.028}$	9.567$^{\pm.109}$	2.783$^{\pm.139}$
MDM (**MotionMix**) *(T_1=20, T_2=40)*	**0.616$^{\pm.006}$**	<u>0.451$^{\pm.033}$</u>	**5.459$^{\pm.027}$**	9.585$^{\pm.101}$	2.585$^{\pm.076}$
MDM (**MotionMix**) *(T_1=40, T_2=60)*	<u>0.598$^{\pm.004}$</u>	0.554$^{\pm.076}$	5.600$^{\pm.031}$	**9.479$^{\pm.100}$**	<u>2.815$^{\pm.094}$</u>
MDM (**MotionMix**) *(T_1=60, T_2=80)*	0.597$^{\pm.008}$	**0.437$^{\pm.039}$**	<u>5.554$^{\pm.033}$</u>	<u>9.452$^{\pm.092}$</u>	**2.895$^{\pm.079}$**