M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Mingshuang Luo[1,2,3], Ruibing Hou[1*] , Zhuo Li[4], Hong Chang[1,3], Zimo Liu[2], Yaowei Wang[2], Shiguang Shan[1,3]

[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),

Institute of Computing Technology, CAS, China

[2]Peng Cheng Laboratory, China, [3]University of Chinese Academy of Sciences, China, [4]WeChat, Tencent Inc

mingshuang.luo@vipl.ict.ac.cn, {houruibing,changhong,sgshan}@ict.ac.cn

albertzli@tencent.com, {liuzm, wangyw}@pcl.ac.cn

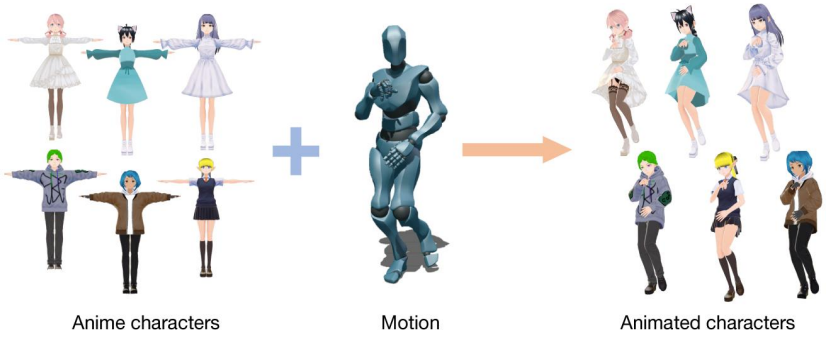
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

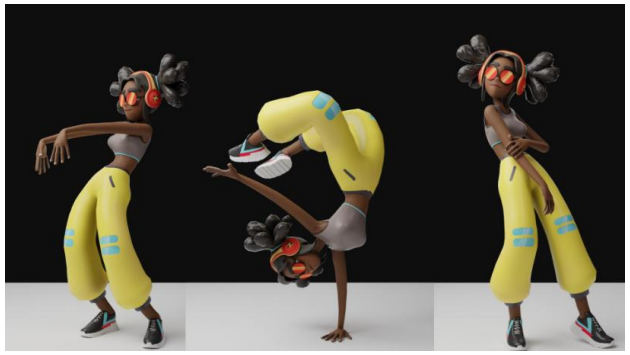
鹏城实验室

M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Background： Human motion understanding and generation have applications in content creation, digital avatars, AR/VR, and robotics. The development of generative AI and large language models, along with the availability of extensive motion datasets, has made it feasible to train a unified motion comprehension and generation model. Such advancements could significantly accelerate progress in game development, VR/AR, digital avatars, and robotics.



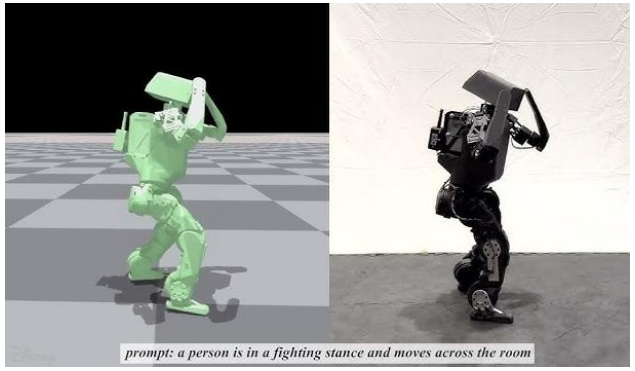
3D Anime Generation



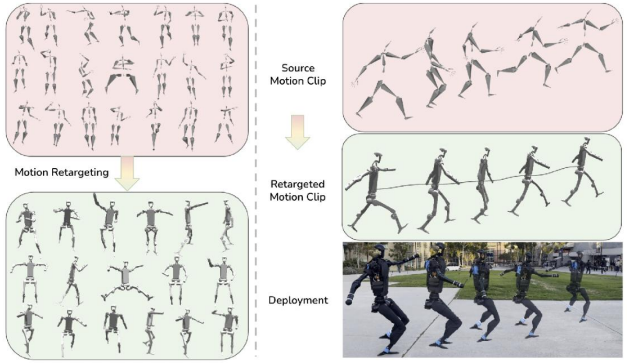
3D Dance Generation

Visual Reality

Digital Avatars



Humaniod-Robotics Control



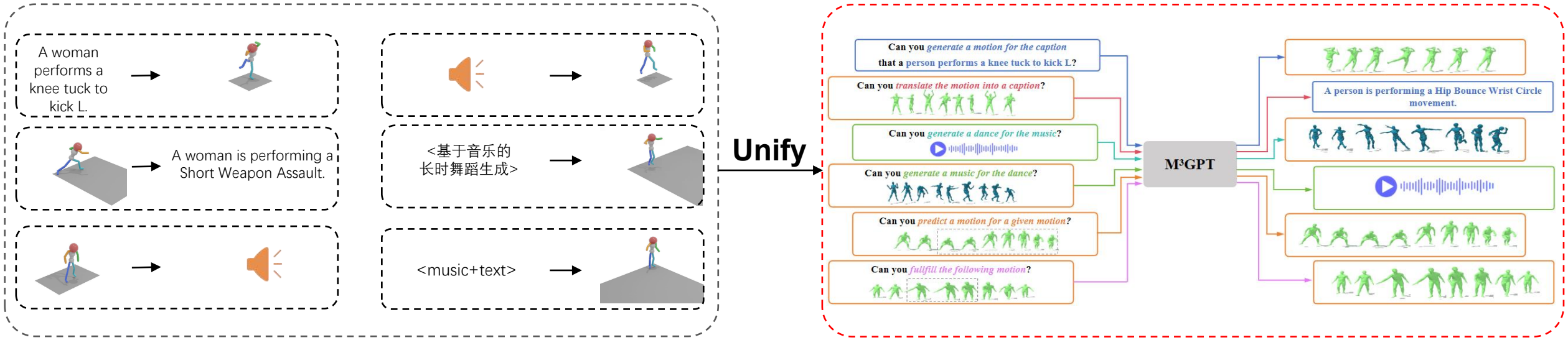
Humaniod-Robotics Control

M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Problems:

① High cost for collecting and annotating 3D human motion data
② Most existing datasets are collected for single task
③ Most existing methods are designed for single task
④ Datasets from different tasks are not well-utilized across tasks
⑤ A lack of effective synergy between different tasks

Aim: Build a unified multimodal, multitask LLM-based framework for motion comprehension and generation



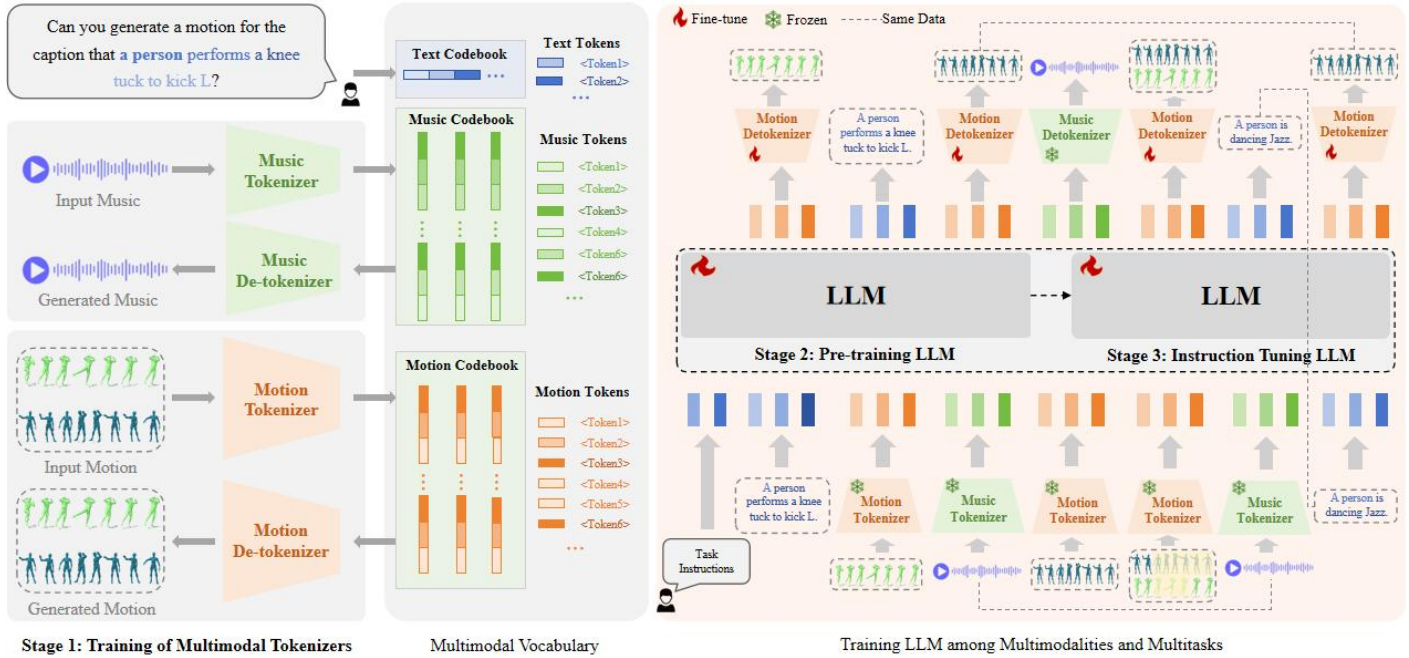
M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Challenges and Solutions for build a unified framework:

① Unified Representation of Different Modalities: For motion/dance sequences and audio, use modality-specific VQ-VAE for discretization, maintaining a unified representation format consistent with text.

② Collaborative Training of Different Tasks: Introduce two additional tasks, text-to-dance and music-to-text, to achieve tri-modal alignment of text-motion/dance-music, enabling effective unification and conversion of data across different modalities for each task.

③ Optimization of Motion/Dance Generation: Fully optimize the Motion-Detokenizer, and during LLM training, update the discrete representation of motion sequences by optimizing the loss of real motion reconstruction. The updated discrete motion encoding is then used to supervise the continuous optimization of the LLM.



Three training stages:

Stage 1: Multimodal tokenizer training (Text、Motion/Dance、Music)

Stage 2: Pre-training LLM (Multimodal multitask pretraining based on a fixed single instruction)

Stage 3: Instruction Tuning LLM (Multimodal multitask fine-tuning based on diverse text instructions)

Quantitative Results (Ablation Study):

Fixed vs. Re-optimized Motion-Detokenizer; With vs. Without the Introduction of Two Additional Collaborative Tasks; Instruction Fine-Tuning vs. No Instruction Fine-Tuning

Methods	Re-Optimizing motion de-tokenizer	Text-to-Motion			Music-to-Dance		
		R TOP1 ↑	FID ↓	Div↑	FID_k ↓	Div_k ↑	BAS ↑
Ground Truth	-	0.675	0.009	2.316	17.10	8.19	0.2374
Trained single task		0.645	0.081	2.124	83.33	5.18	0.1835
Trained single task	✓	**0.656**	**0.078**	2.133	75.47	5.57	0.1884
T2M+A2D		0.564	0.094	2.080	51.26	6.73	0.2037
T2M+A2D	✓	0.578	0.092	2.106	47.71	7.47	0.1958
T2M+A2D+T2D+A2T		0.617	0.093	2.110	42.70	7.54	0.2084
T2M+A2D+T2D+A2T	✓	0.626	0.088	2.197	25.24	**7.63**	**0.2217**
M^3GPT (Pretrained without T2D and A2T)		0.526	0.105	2.058	40.71	7.47	0.2030
M^3GPT (Pretrained without T2D and A2T)	✓	0.547	0.104	2.099	37.14	7.61	0.2005
M^3GPT (Pre-trained)		0.598	0.089	2.218	32.71	7.43	0.2090
M^3GPT (Pre-trained)	✓	0.601	0.092	2.251	27.65	7.52	0.2105
M^3GPT (Instruction-tuned)		0.606	0.091	2.251	28.46	7.49	0.2052
M^3GPT (Instruction-tuned)	✓	0.615	0.093	**2.253**	**24.34**	7.50	0.2056

① Re-Optimizing Motion-Detokenizer can improve the performance of generated motion.
② The two additional tasks (text-to-dance and music-to-text) can help align the data modalities of different tasks, thereby improving the model's performance.
③ Instruction Fine-Tuning can help model focus on specific task.

M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Quantitative Results (Comparison with SOTAs):

Comparison for Text-to-Motion, Motion-to-Text, Motion Prediction and Motion In-between on Motion-X dataset

Methods	Text-to-Motion			Motion-to-Text			Motion Prediction		Motion In-between	
	R TOP1↑	FID↓	Div↑	R TOP3↑	Bleu@4↑	CIDEr↑	FID↓	Div↑	FID↓	Div↑
Real	$0.675^{\pm 0.003}$	$0.009^{\pm 0.000}$	$2.316^{\pm 0.011}$	0.881	-	-	0.009	2.316	0.009	2.316
MLD [4]	$0.612^{\pm 0.003}$	$0.122^{\pm 0.008}$	$2.267^{\pm 0.018}$	-	-	-	-	-	-	-
T2M-GPT [47]	$0.647^{\pm 0.002}$	$0.101^{\pm 0.005}$	$2.270^{\pm 0.033}$	-	-	-	-	-	-	-
MotionDiffuse [49]	$0.659^{\pm 0.002}$	$\mathbf{0.075}^{\pm 0.004}$	$2.220^{\pm 0.022}$	-	-	-	-	-	-	-
TM2T [13]	$0.581^{\pm 0.002}$	$0.148^{\pm 0.003}$	$2.005^{\pm 0.034}$	0.806	**12.13**	20.16	-	-	-	-
MDM [37]	$0.472^{\pm 0.008}$	$0.078^{\pm 0.000}$	$2.133^{\pm 0.012}$	-	-	-	1.028	1.746	0.831	1.768
Trained single task	$0.656^{\pm 0.002}$	$0.078^{\pm 0.002}$	$2.133^{\pm 0.012}$	0.767	10.14	22.92	0.774	1.778	0.692	1.810
M^3GPT (Pre-trained)	$0.601^{\pm 0.002}$	$0.092^{\pm 0.002}$	$2.251^{\pm 0.012}$	0.834	11.00	24.12	0.707	1.874	0.604	1.879
M^3GPT (Instruction-tuned)	$0.615^{\pm 0.003}$	$0.093^{\pm 0.002}$	$2.253^{\pm 0.026}$	**0.845**	11.50	**42.93**	**0.682**	**1.838**	**0.612**	**1.900**
M^3GPT (Instruction-tuned only T2M)	$\mathbf{0.661}^{\pm 0.003}$	$0.076^{\pm 0.002}$	$\mathbf{2.273}^{\pm 0.026}$	-	-	-	-	-	-	-

Comparison of Text-to-Motion and Motion-to-Text with different size of T5

Methods on Motion-X	LLM	Training time	Text-to-Motion			Motion-to-Text		
			R-TOP1 ↑	FID ↓	Div ↑	R-TOP3 ↑	Bleu4 ↑	CIDEr ↑
M³GPT	T5-small (60M)	5 days	0.598	0.096	2.202	0.822	10.43	38.22
M³GPT	T5-base (220M)	7 days	0.615	0.093	2.253	0.845	11.50	42.93
M³GPT	T5-large (770M)	8 days	**0.619**	**0.090**	**2.256**	**0.848**	**11.64**	**43.05**

Comparison of Music-to-Dance and Dance-to-Music with different size of T5

Methods on AIST++	LLM	Training time	Music-to-Dance			Dance-to-Music	
			FID_k ↓	DIV_k ↑	BAS ↑	BCS ↑	BHS ↑
M³GPT	T5-small (60M)	5 days	28.05	5.96	0.2091	89.1	91.2
M³GPT	T5-base (220M)	7 days	24.34	7.50	0.2056	93.6	94.0
M³GPT	T5-large (770M)	8 days	**23.26**	**7.54**	**0.2061**	**93.8**	**94.1**

Comparison for Music-to-Dance and Dance-to-Music on AIST++ and FineDance dataset

Methods	Music-to-Dance on AIST++			Music-to-Dance on FineDance			Dance-to-Music on AIST++		
	FID_k ↓	Div_k ↑	BAS ↑	FID_k ↓	Div_k ↑	BAS ↑	BCS↑	BHS↑	F1↑
Real	17.10	10.60	0.2374	-	-	0.2120	-	-	-
FACT [18]	35.35	5.94	0.2209	113.38	3.36	0.1831	-	-	-
Bailando [35]	28.16	**7.83**	0.2332	82.81	7.74	0.2029	-	-	-
EDGE [40]	42.16	3.96	0.2334	94.34	**8.13**	0.2116	-	-	-
Lodge [20]	37.09	5.58	**0.2423**	**45.56**	6.75	**0.2397**	-	-	-
Foley [9]	-	-	-	-	-	-	96.4	41.0	57.5
CMT [8]	-	-	-	-	-	-	97.1	46.2	62.6
D2MGAN [55]	-	-	-	-	-	-	95.6	88.7	93.1
CDCD [56]	-	-	-	-	-	-	96.5	89.3	92.7
LORIS [44]	-	-	-	-	-	-	**98.6**	90.8	94.5
Trained single task	75.47	5.57	0.1884	128.37	6.48	0.2036	93.9	93.6	92.8
M^3GPT (Pre-trained)	27.65	7.52	0.2105	92.35	7.67	0.2134	93.4	93.8	94.2
M^3GPT (Instruction-tuned)	**24.34**	7.50	0.2056	86.47	7.75	0.2158	93.6	**94.0**	**94.9**

① M³GPT can get competitive performance among different motion-related tasks compared to existing SOTAs.

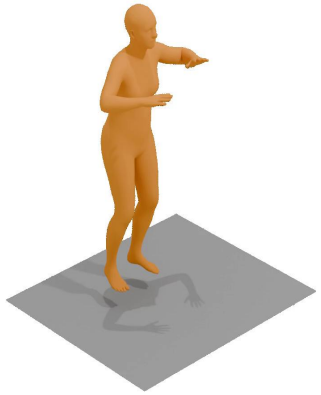
② Pre-training can help improve the performance of each single task.

③ M³GPT can get better performance with the increasing of LLM size.

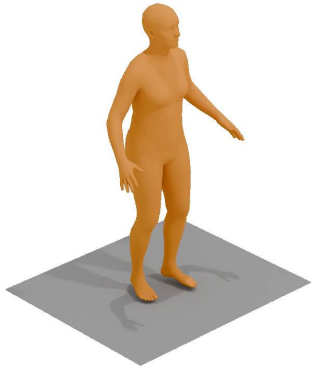
M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Qualitative Results (text-to-motion and music-to-dance):

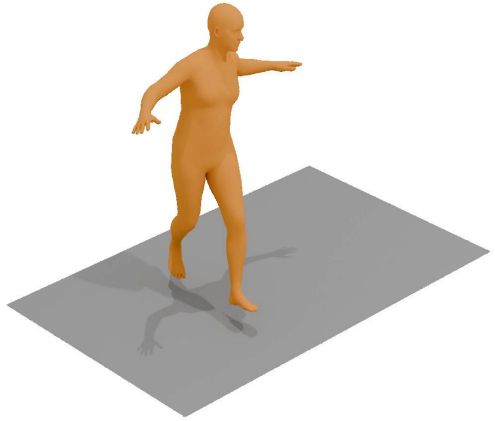
person bends over to grab something and acts like they are fighting off other people from grabbing the thing.

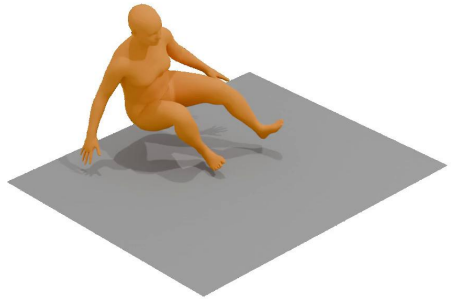


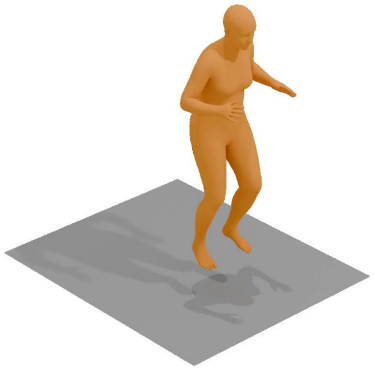
a person hunches fowrad and swings their arms to act like a monkey.



A group of people are doing the Circle In Circle Out dance move.

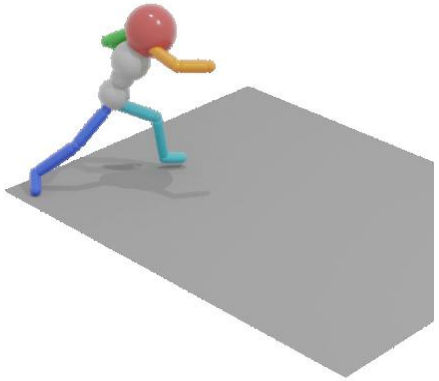




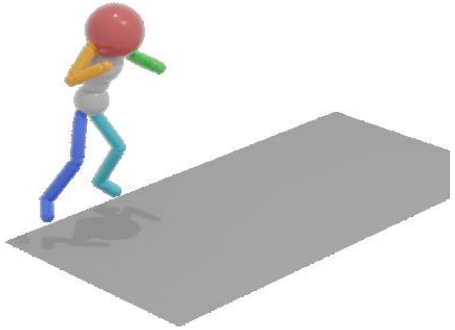


M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Qualitative Results (motion-to-text and dance-to-music):

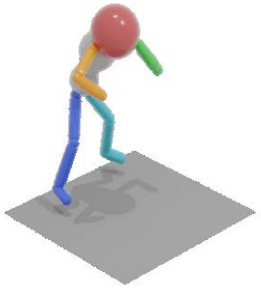


A woman is performing a Short Weapon Assault.

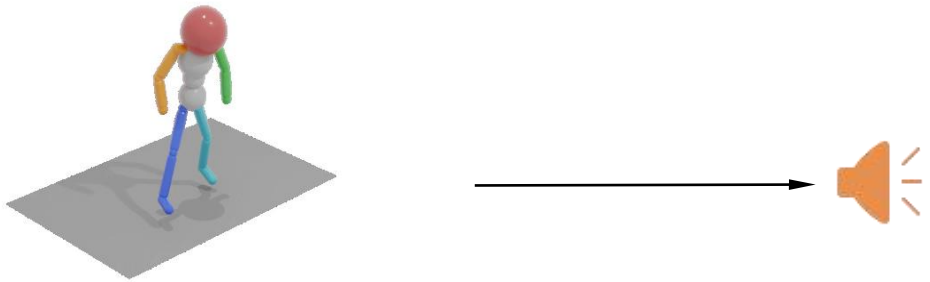


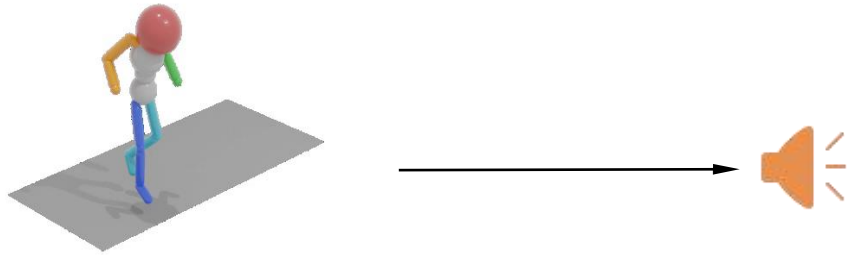
The person is doing the Hand Ausweichhi.

The person covers their ears with their hands while sitting.

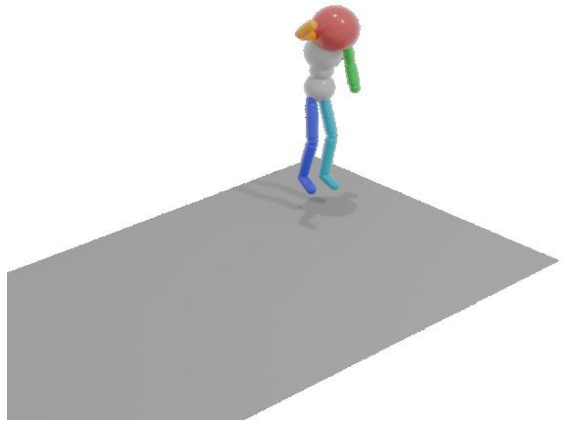


a person is sitting and throwing their arms around.

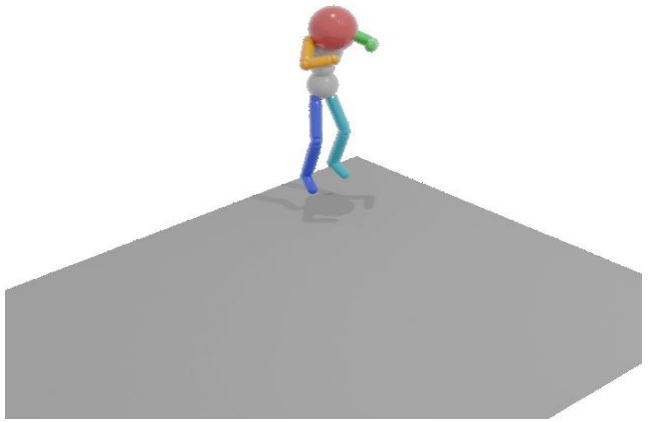




Qualitative Results (music-text to dance and long-term dance generation):

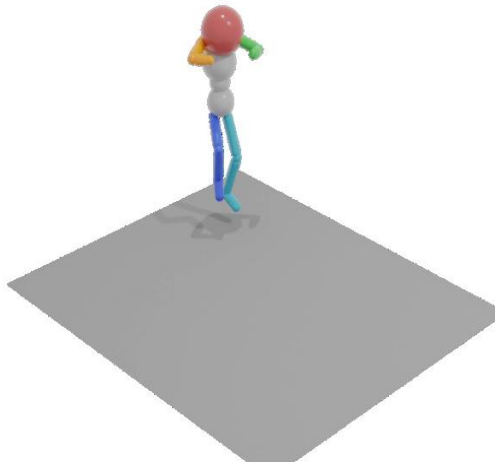


<music>+a person is spinning in circles

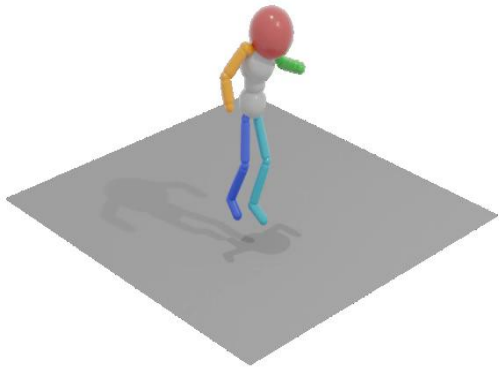


<music>+a person performs a figure-eight jump

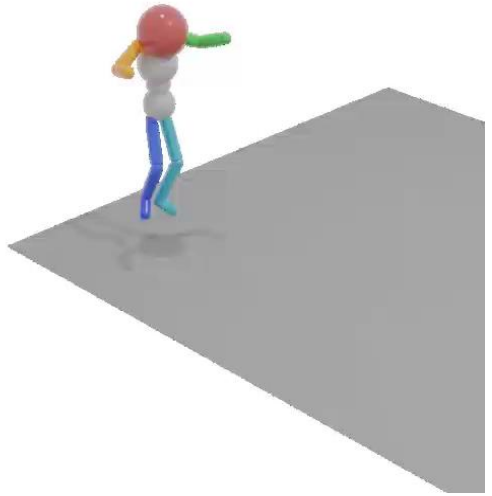
<music>+a_person does a cartwheel



long-term dance generation-aistpp-sample 0



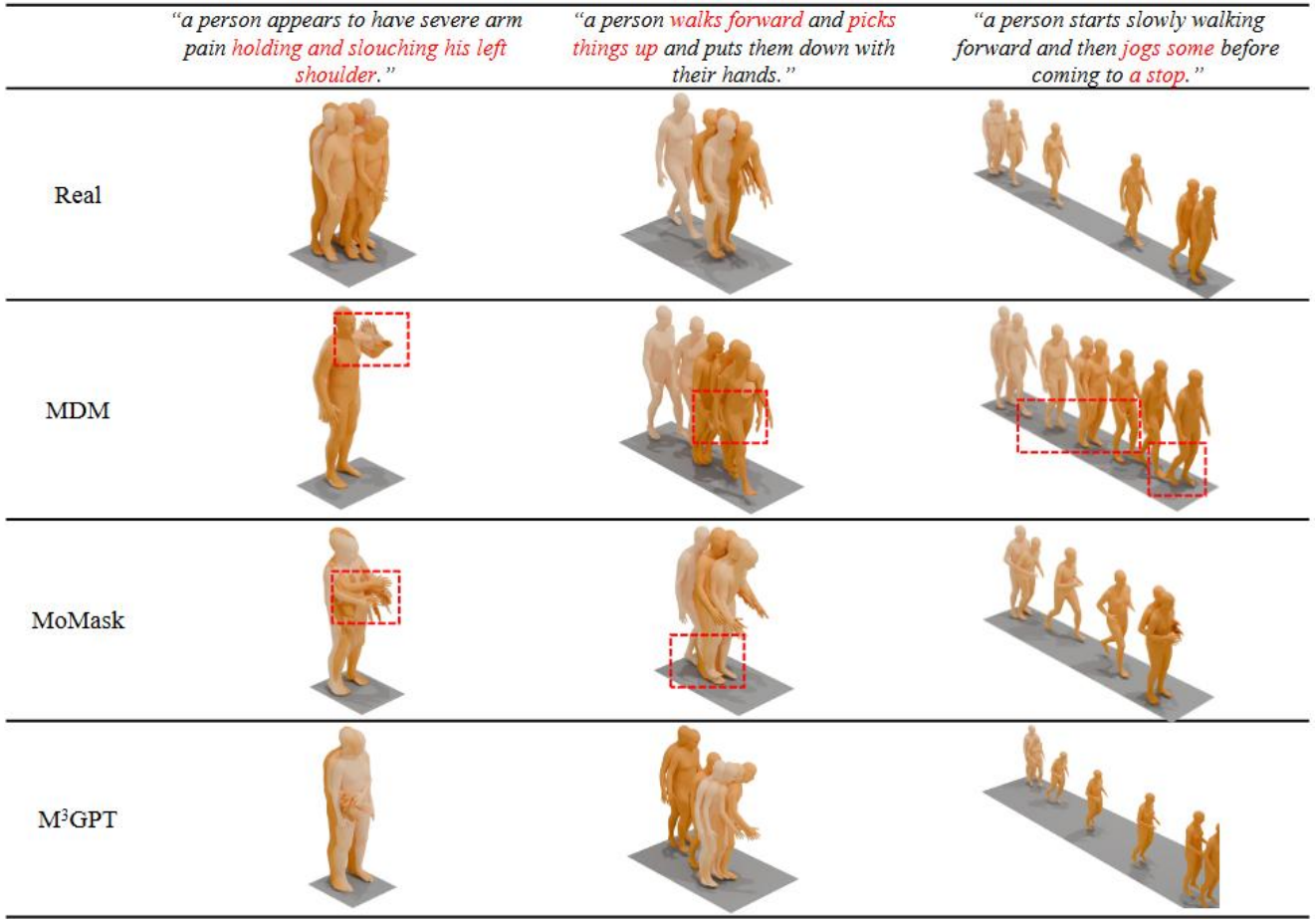
long-term dance generation-aistpp-sample 1

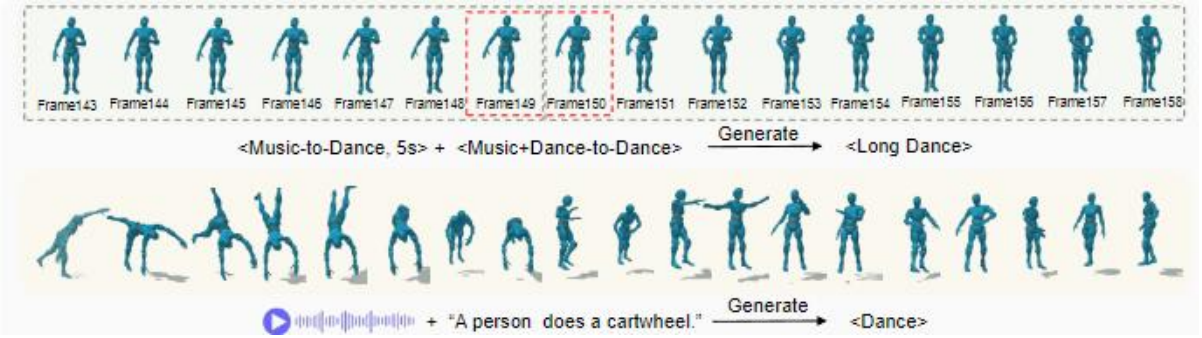


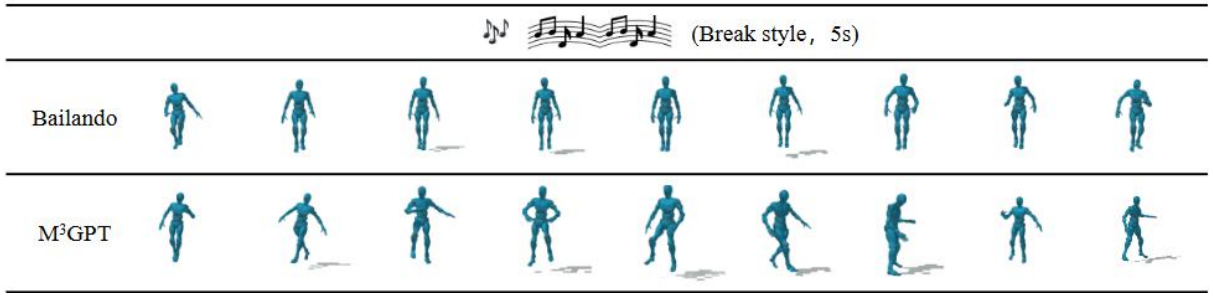
long-term dance generation-finedance-sample 0

M³GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation

Qualitative Results (Comparison with Other Methods):







Conclusion：

① We present M³GPT, a unified framework for comprehending and generating motion aligned with both text and music modalities.

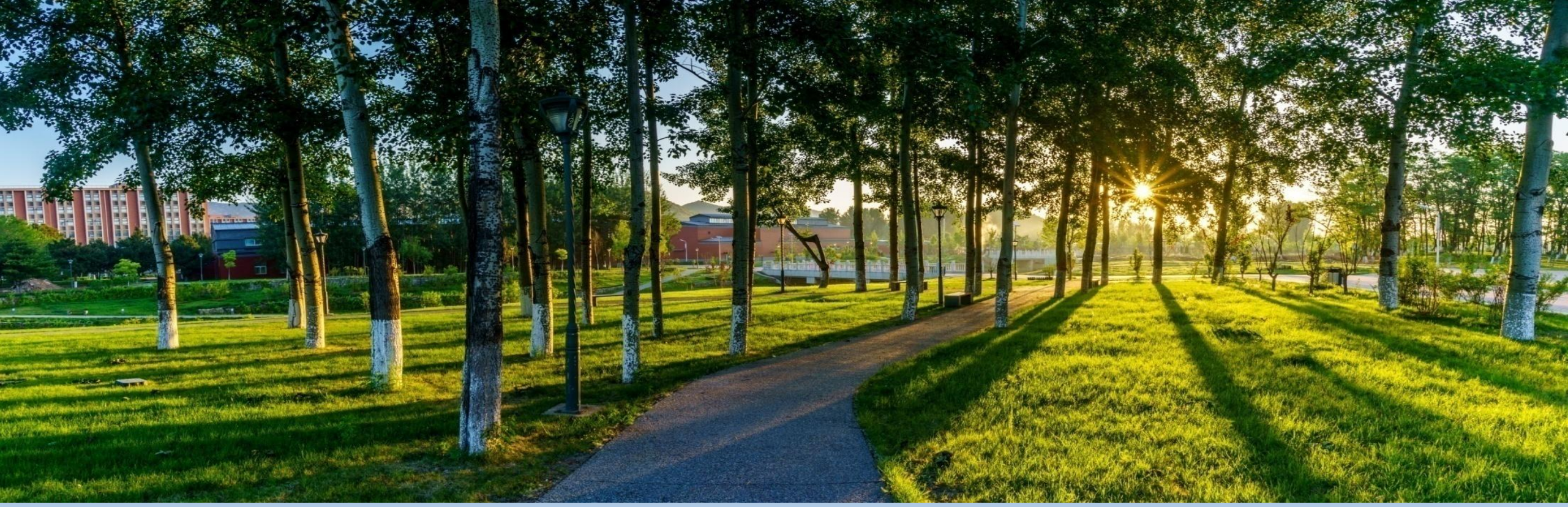
② We build a multimodal vocabulary, including text tokens, music tokens, and motion/dance tokens.

③ We leverage text as a bridge to connect and synergy different motion-relevant tasks.

④ We re-optimize the motion-detokenizer to enhance the performance of generated motions.

⑤ We evaluate our M³GPT in both comprehension and generation tasks, achieving competitive performance.

⑥ We also assess our M³GPT in music-text conditioned dance generation and long-term dance generation, demonstrating strong zero-shot generation abilities.



THANKS

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

鹏城实验室
PENGCHENG LABORATORY