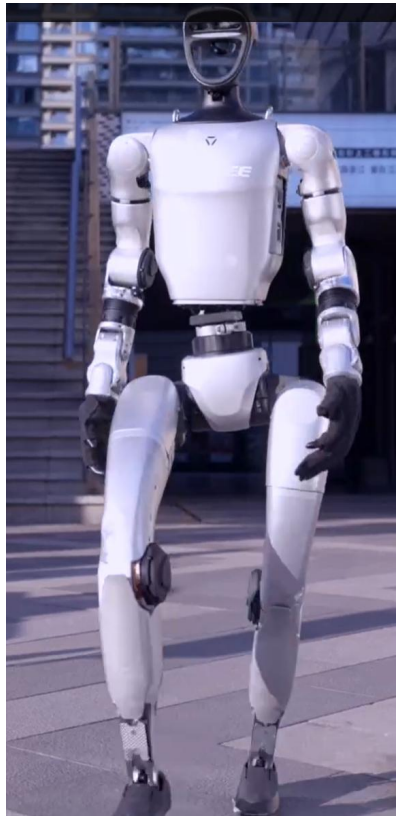
Scaling Large Motion Models with Million-Level Human Motions

RUC BAAI PKU BeingBeyond

Introduction

- Humanoid robots are receiving more and more attention
- We hope that robots can move like humans.

Unitree G1



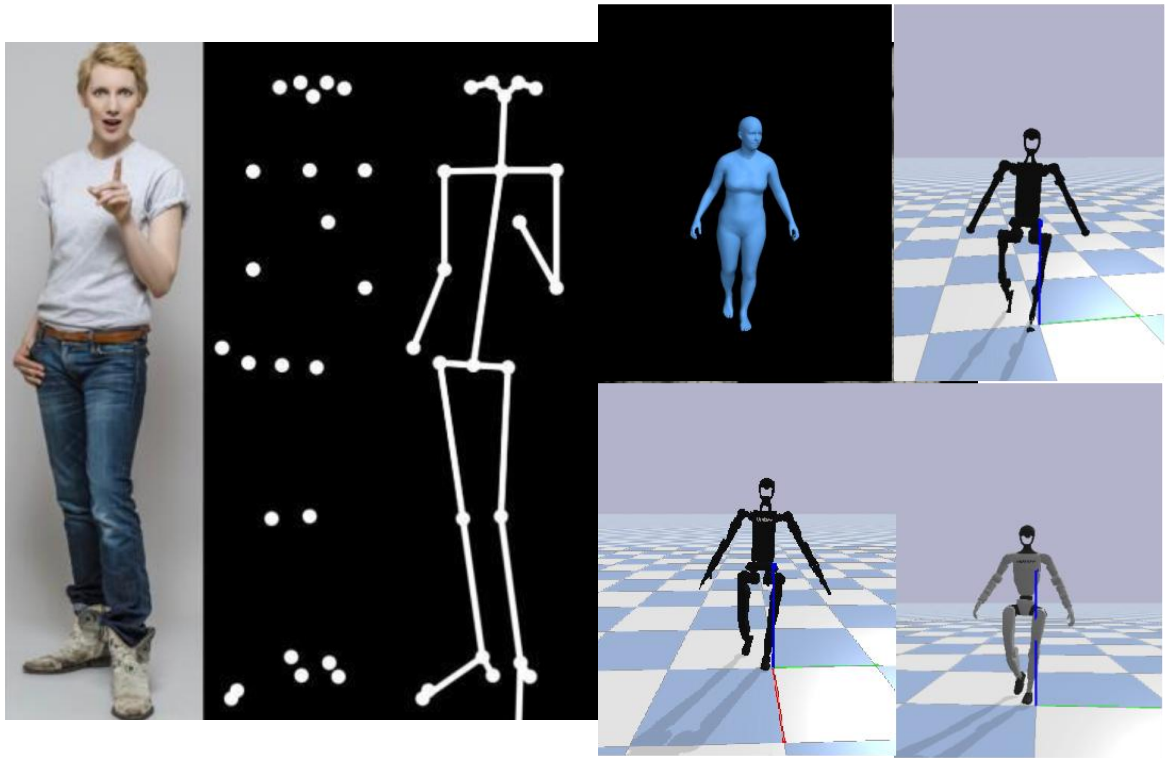
Unitree H1

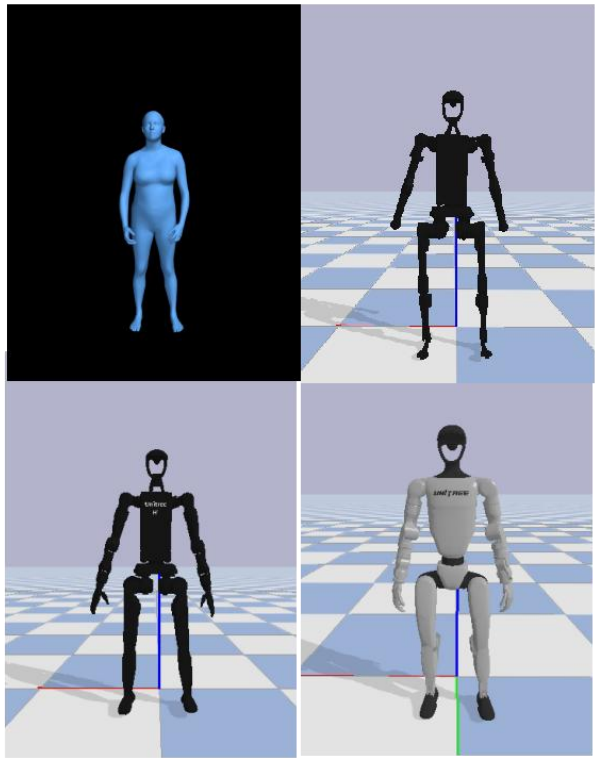
Tesla Optimus

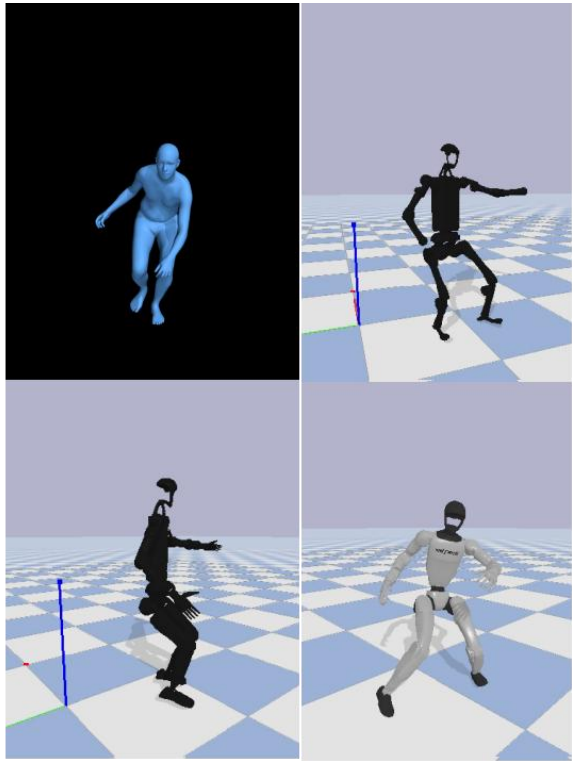
BostonDynamics Atlas

Introduction

- Human motion is a good starting point.
- Human motion is represented by keypoints and can be transferred to robots through retargeting methods.

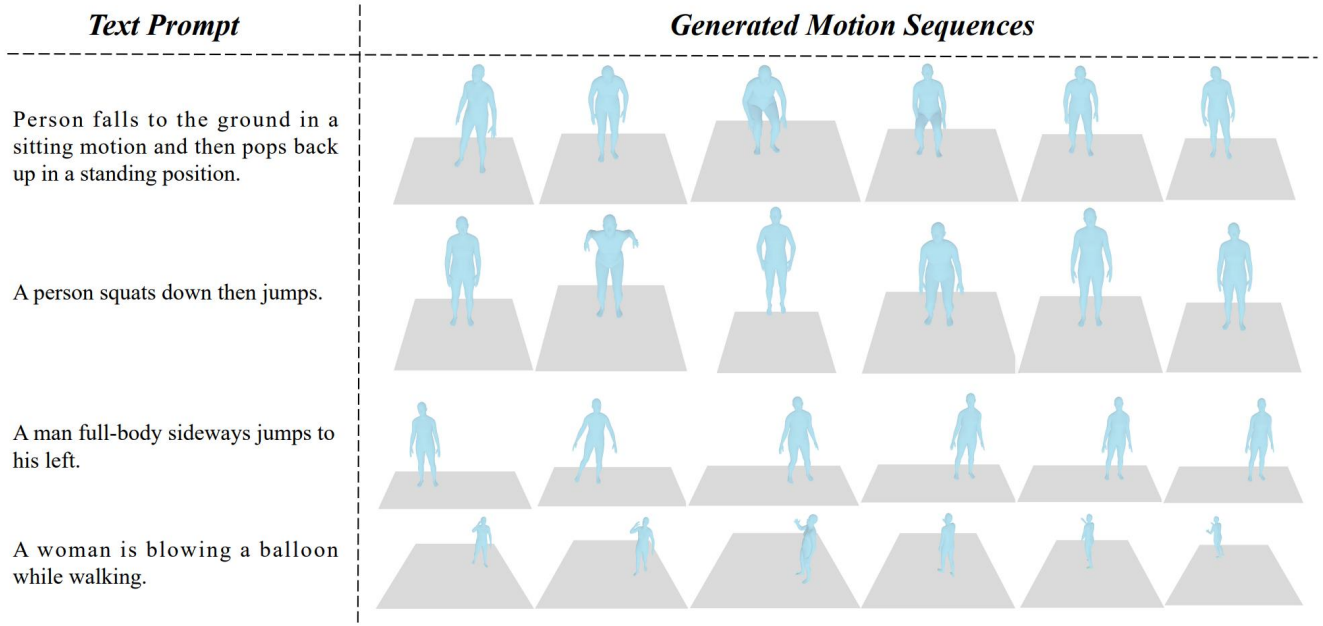






Scaling Motion Generation

- We hope to train motion generation models that can understand human instructions and generate human motion.
- **The core problem:** a lack of large-scale human motion datasets.

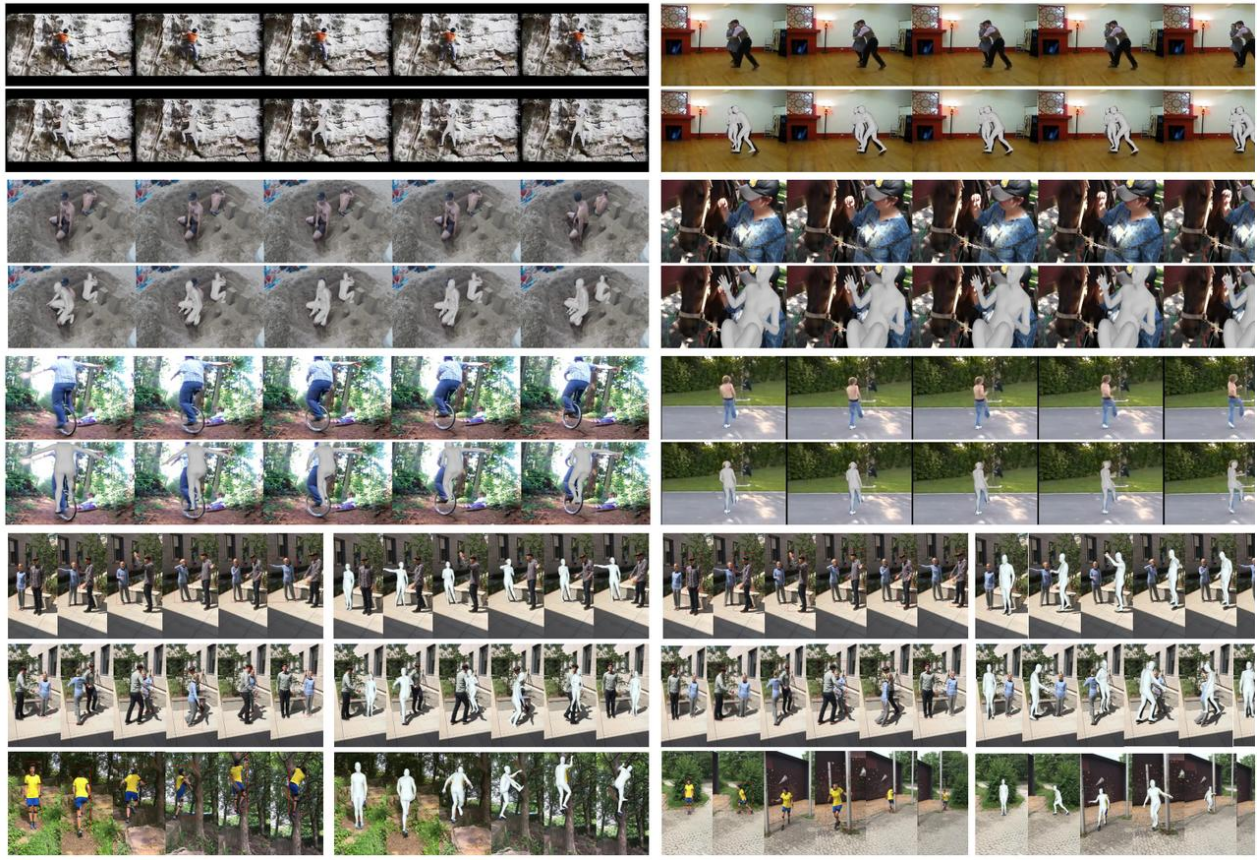


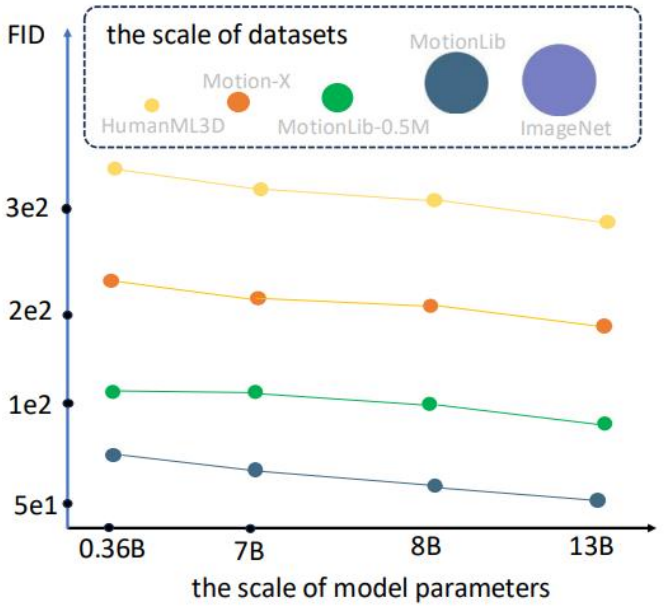
	SEQ NUM	TEXT NUM	HOURS	MOTION	TEXT	RGB	DEPTH	BBOX	PERSON
KIT (Plappert et al., 2016)	5.7K	5.7K	11.2	B	body	✗	✗	✗	single
HumanML3D (Guo et al., 2022a)	29.2K	89K	28.6	B	body	✗	✗	✗	single
MotionX (Lin et al., 2024)	81.1K	142K	144.2	B,H,F	body	✓	✗	✗	single
MotionVerse (Zhang et al., 2024a)	320k	373k	-	B,H,F	body	✓	✗	✗	single
MotionLib	1.21M	2.48M	1456.4	B,H	hier	✓	✓	✓	single & multi

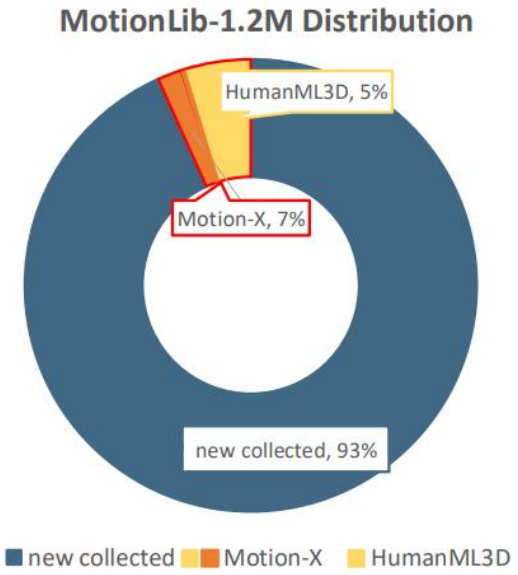
MotionLib

MotionLib Dataset

- Building the first million-scale motion generation dataset (over 1.2 million sequences, over 2.4 million text snippets, over 1456 hours).

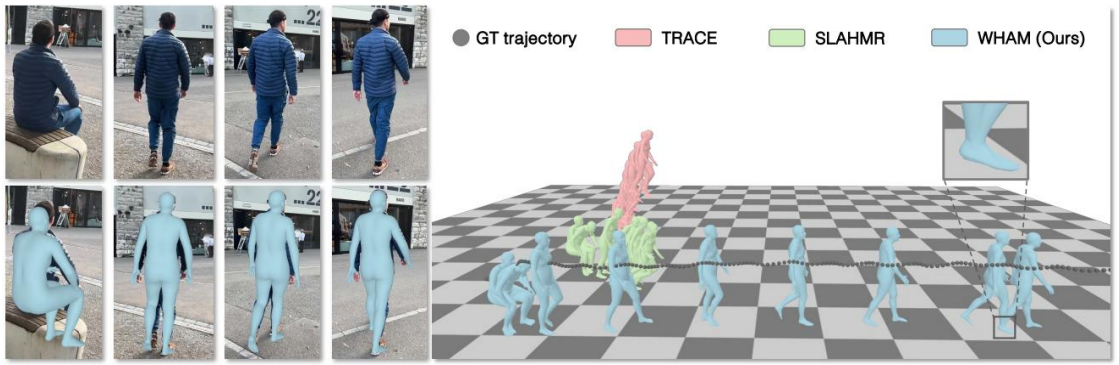


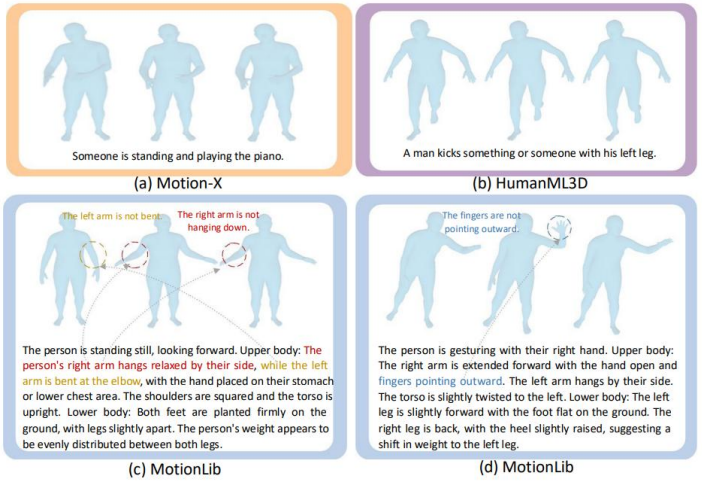


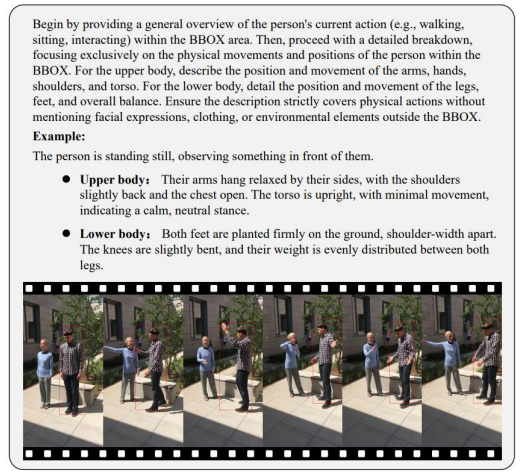


MotionLib

- **Video collection:** Collecting a large amount of publicly available video data and online videos.
- **Motion reconstruction:** Estimating motion from videos using motion reconstruction methods.
- **Text annotation:** Utilizing video understanding large multimodal models for fine-grained text annotation of motion.
- **Further refinement:** Utilizing an RL policy for motion refinement, using GPT-4o for text quality assessment.



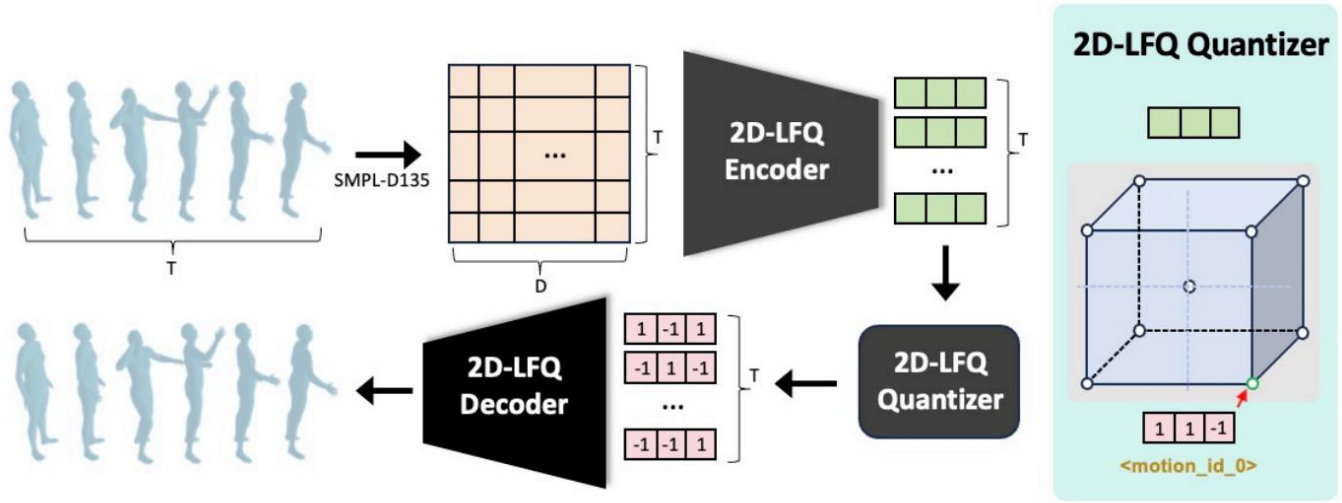




MotionBook

MotionBook Motion Encoding

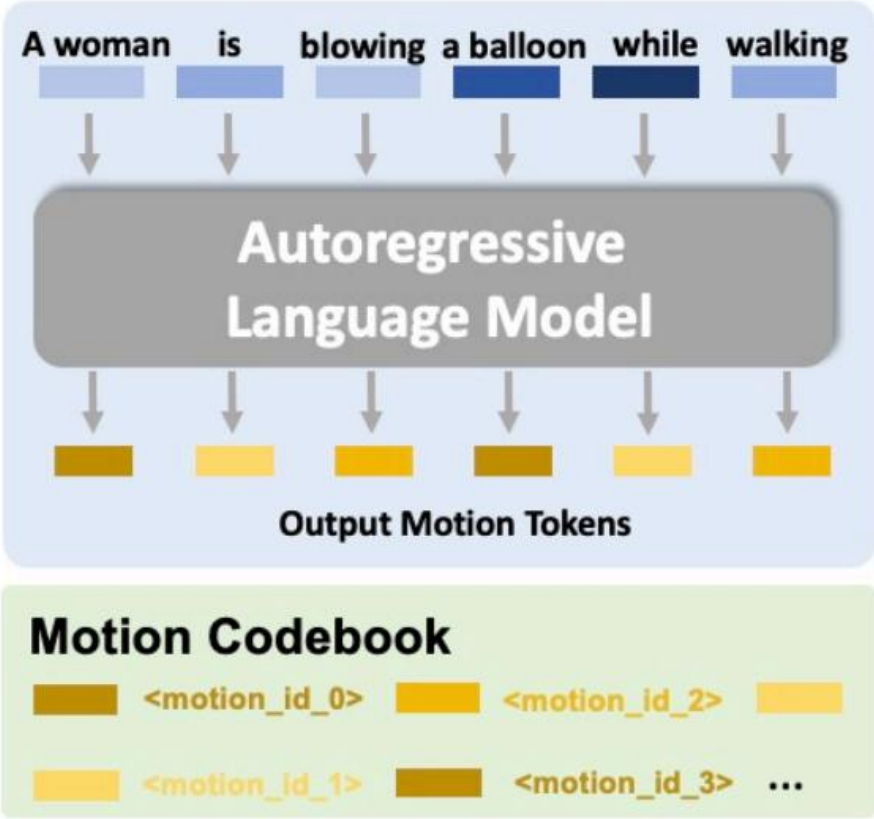
- **Objective:** To address the issues of inefficient and information-losing motion representation.
- **Motion Representation:** Uses the lossless and compact SMPL-d135 model.
- **2D-LFQ Discretization:**
  - **Structuring:** Temporal motion data (a short sequence of continuous actions) is reorganized into a 2D image format (Time x Feature), preserving spatiotemporal continuity.
  - **Compression and Quantization:** An encoder compresses this "image" into a d-dimensional feature vector z. We apply sign quantization to each dimension of z: +1 for positive, -1 for negative/zero.
  - **Codebook Construction:** This maps motion to a sequence of d values (+1/-1), implicitly creating a 2^d combination motion codebook for efficient, compact representation.



Large Motion Model

Motion Generation Model Framework (Being-M0)

- **Architecture:** Built upon a pre-trained Large Language Model (LLM) of the decoder-only type (e.g., GPT-2, LLaMA).
- **Integration:** The discrete motion tokens generated by 2D-LFQ are directly added to the LLM's vocabulary. Special tokens <mot> and </mot> are used to mark the start and end of the motion sequences.
- **Pre-training:** Continued pre-training on a large amount of text-motion data results in a powerful Large Motion Model.



Experiment

Scaling Laws

- **Larger Models Are Better:** Increasing model parameters (i.e., model size) leads to a stable improvement in performance (e.g., LLaMA-13B outperforms 7B, and 7B outperforms GPT-2).
- **More Data Is Better:** Increasing the amount of data (i.e., training data size) significantly improves performance.
- **First Verification:** This is the first systematic demonstration that "scaling laws" also apply to the field of motion generation.

			Motion-X-eval			MotionLib-eval		
Decoder	#Inst.	#Param.	R@1 ↑	R@3 ↑	FID ↓	R@1 ↑	R@3 ↑	FID ↓
Real	-	-	0.514	0.831	0.046	0.297	0.634	0.004
GPT-2	0.02M	355M	0.213	0.426	47.319	0.058	0.152	30.612
GPT-2	0.08M	355M	0.468	0.792	**0.083**	0.114	0.281	22.077
GPT-2	0.5M	355M	0.463	0.793	0.121	0.161	0.354	9.157
GPT-2	1.2M	355M	0.472	0.791	0.112	0.166	0.375	6.936
LLaMA-2	0.02M	7B	0.216	0.433	47.538	0.059	0.158	29.643
LLaMA-2	0.08M	7B	0.472	0.798	0.166	0.118	0.294	21.593
LLaMA-2	0.5M	7B	0.468	0.799	0.178	0.164	0.369	9.146
LLaMA-2	1.2M	7B	0.475	0.798	0.156	0.171	0.380	6.632
LLaMA-3	0.02M	8B	0.216	0.435	47.906	0.059	0.162	29.257
LLaMA-3	0.08M	8B	0.483	0.815	0.122	0.120	0.301	21.295
LLaMA-3	0.5M	8B	0.483	0.817	0.113	0.166	0.368	8.973
LLaMA-3	1.2M	8B	0.486	0.820	0.117	0.173	0.386	**6.029**
LLaMA-2	0.02M	13B	0.223	0.446	47.210	0.061	0.169	29.143
LLaMA-2	0.08M	13B	0.488	0.820	0.156	0.124	0.314	21.001
LLaMA-2	0.5M	13B	0.490	0.819	0.145	0.174	0.374	8.824
LLaMA-2	1.2M	13B	**0.491**	**0.823**	0.133	**0.185**	**0.391**	6.221

Experiment

Motion Generation

- The state-of-the-art in motion generation.

	Decoder	R@1 ↑	R@3 ↑	FID ↓	MMDist ↓
Real	-	0.511	0.797	0.002	2.974
MLD	-	0.481	0.772	0.473	3.196
MotionDiffuse	-	0.491	0.782	0.630	3.113
ReMoDiffuse	-	0.510	0.795	0.103	2.974
Fg-T2M++	-	0.513	0.801	0.089	2.925
LMM	-	0.525	0.811	**0.040**	2.943
T2M-GPT	GPT-2	0.492	0.775	0.141	3.121
DiverseMotion	GPT-2	0.510	0.802	0.072	2.941
MoMask	-	0.521	0.807	0.045	2.958
MotionGPT[1,*]	T5	0.409	0.667	0.162	3.992
MotionGPT[1]	T5	0.492	0.778	0.232	3.096
MotionGPT[2,*]	LLaMA2-13B	0.367	0.654	0.571	3.981
MotionGPT[2,*]	LLaMA-13B	0.363	0.633	0.592	4.029
MotionGPT[2]	LLaMA-13B	0.411	0.696	0.542	3.584
MotionLLM	Gemma-2b	0.482	0.770	0.491	3.138
AvatarGPT	LLaMA-13B	0.389	0.623	0.567	-
MotionGPT-v2	LLaMA3.1-8B	0.496	0.782	0.191	3.080
Being-M0-VQ	LLaMA2-13B	0.519	0.803	0.166	2.964
Being-M0-LFQ	LLaMA2-13B	**0.528**	**0.820**	0.141	**2.875**

- More data leads to stronger OOD generalization capability.

TRAIN SET	R@1 ↑	R@3 ↑	FID ↓
Real	0.176	0.379	0.076
HumanML3D	0.034	0.112	82.674
MotionX	0.051	0.141	70.547
MotionLib-#11	0.098	0.218	11.930

Experiment

Motion Tokenizer

- Our designed motion tokenizer method achieves better performance and less information loss when processing larger-scale datasets.

			HumanML3D		Motion-X		MotionLib	
Tokenizer	#Num.	#Param.	FID ↓	MPJPE ↓	FID ↓	MPJPE ↓	FID ↓	MPJPE ↓
VQ-VAE	512	19.43M	0.078	69.2	0.852	106.4	5.324	123.6
H^2VQ	512	-	-	-	-	62.34	-	-
RQ-VAE	512	19.43M	**0.052**	**37.5**	0.568	56.9	4.026	78.2
2D-LFQ	16384	108.35M	0.092	45.6	**0.295**	**54.1**	**2.315**	**64.1**

Thank You