RoMo: A Large-Scale, Richly Organized Dataset and Semantic Taxonomy for Human Motion Generation

Jiahao Zhang[1,2] Joseph Liu[2] Young-Yoon Lee[2] Seonghyeon Moon[2] Victor Zordan[2] Guy Tevet[3]
C. Karen Liu[3] Stephen Gould[1] Oren Jacob[2] Haomiao Jiang[2] Mubbasir Kapadia[2,4] Yizhak Ben-Shabat[2]

Australian National University · ROBLOX · Stanford University · RUTGERS

Project Website

CVPR JUNE 3-7, 2026 · DENVER COLORADO

1. Introduction

Text to Motion Generation (T2M)

T2M aims to synthesize realistic **3D human motion** from natural language descriptions. Enabling scalable animation for games, VR, humanoids, etc. However, high-quality human motion data is difficult to collect **at scale**.

Text to Motion Datasets

Dataset	Hierarchical Semantic Taxonomy	Category Sub-category	Clip Number Core (Total)	Hour
KIT-ML	✗	-	3.9K (3.9K)	11.2 (11.2)
BABEL	✓	8 / 260	13K (13K)	43.5 (43.5)
HumanML3D	✗	-	0 (15K)	0 (28.6)
SnapMoGen	✗	-	20K (20K)	43.7 (43.7)
Motion-X	✗	-	48.6K (81.1K)	86 (144.2)
Motion-X++	✗	-	39.4K (120.5K)	59 (180.9)
MotionMillion	✗	-	560K (2M)	726.5 (2000)
RoMo (ours)	✓	**54 / 2065**	**820K (2.58M)**	**1237.8 (3023)**

Recent advances in monocular motion estimation make web-scale motion collection possible. RoMo is not only **large-scale**, but also **richly organized** with a **semantic taxonomy** and multi-level captions.

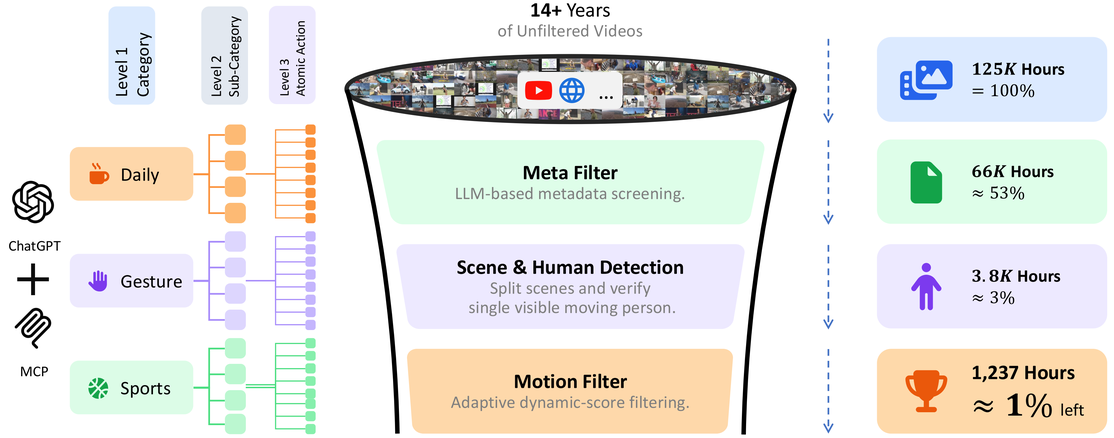
2. Contributions

820K Clips | **1,237** Hours
Large-Scale Motion Dataset
Rich captions and high-quality filtered motion.

54 Category | **2065** Subcategory
Taxonomy-Aware Curation
Adaptive filtering for quality and diversity.

Motion Toolbox
Open Source Motion Toolbox
Standardized evaluation and motion analysis.

3. Taxonomy-Aware Adaptive Filtering Pipeline



Video Segmentation
Long videos are segmented into temporally coherent motion clips using **Qwen3-VL**.

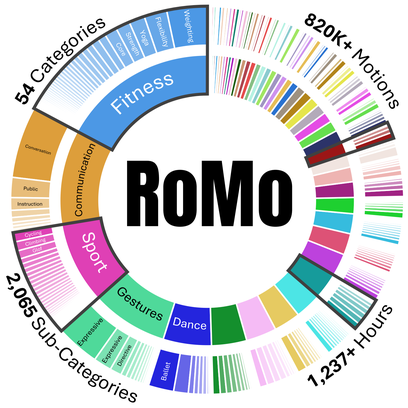
Multi-Level Captioning
5 level captions are generated to describe actions, objects, and interaction contexts using **Qwen3-VL**.

Motion Estimation
GVHMR reconstructs world-grounded SMPL motion sequences from monocular videos.

Taxonomy Mapping
Each clip is mapped into our hierarchical motion taxonomy using **LLM-assisted** matching.

Fitness → Weighing → Lift barbell overhead
*"A person **lifts** a barbell from the ground to above his head in a smooth motion, lowering it back down repeatedly."*

Sport → Bowling → Roll Bowling Ball
*"A person **picks** up the ball, takes a few steps forward, and **releases** it with a smooth motion."*

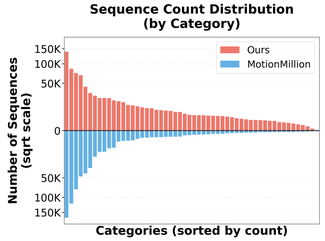


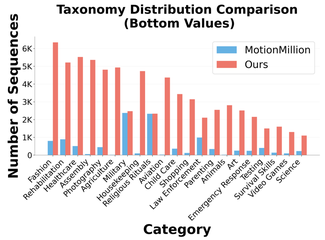
Video Games → Motion Gaming → Swing
*"A person in a VR headset **swings** a virtual sword at floating objects, hitting them and causing them to shatter."*

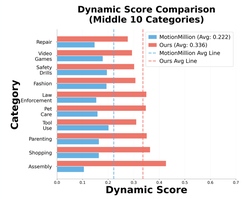
Outdoors → Rafting → Push raft with paddle
*"A person in a raft on the riverbank uses a paddle to **push off** from the shore."*

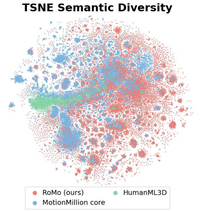
4. Scaling High-Quality Human Motion

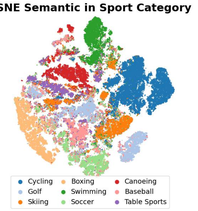
Superior Diversity, Coverage and Motion Dynamics











+41% higher dynamic score and +61.7% subcategory coverage over MotionMillion.

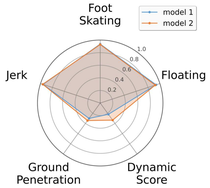
Training Modern Motion Generators on RoMo

Method	Diversity ↑	FID ↓	Matching Score ↑	Dynamic Score ↑	Ground Penetration ↓ ($\times 10^{-5}$)	Foot Skating ↓ ($\times 10^{-3}$)	Floating ↓ ($\times 10^{-2}$)
MDM	27.67	20.63	12.06	0.2138	0.0	1.70	1.67
MMGPT	16.68	12.80	22.08	0.3268	3.55	92.0	0.0311

5. Open-Source Motion Toolbox

Interactive Visualization
Inspect and compare motion sequences directly in the browser.

Unified Benchmarking
Reproducible evaluation tools for large-scale human motion research.



Unified Conversion
Convert between popular motion formats and research datasets.