RoMo

A Large-Scale, **R**ichly **O**rganized Dataset and Semantic Taxonomy for Human **Mo**tion Generation

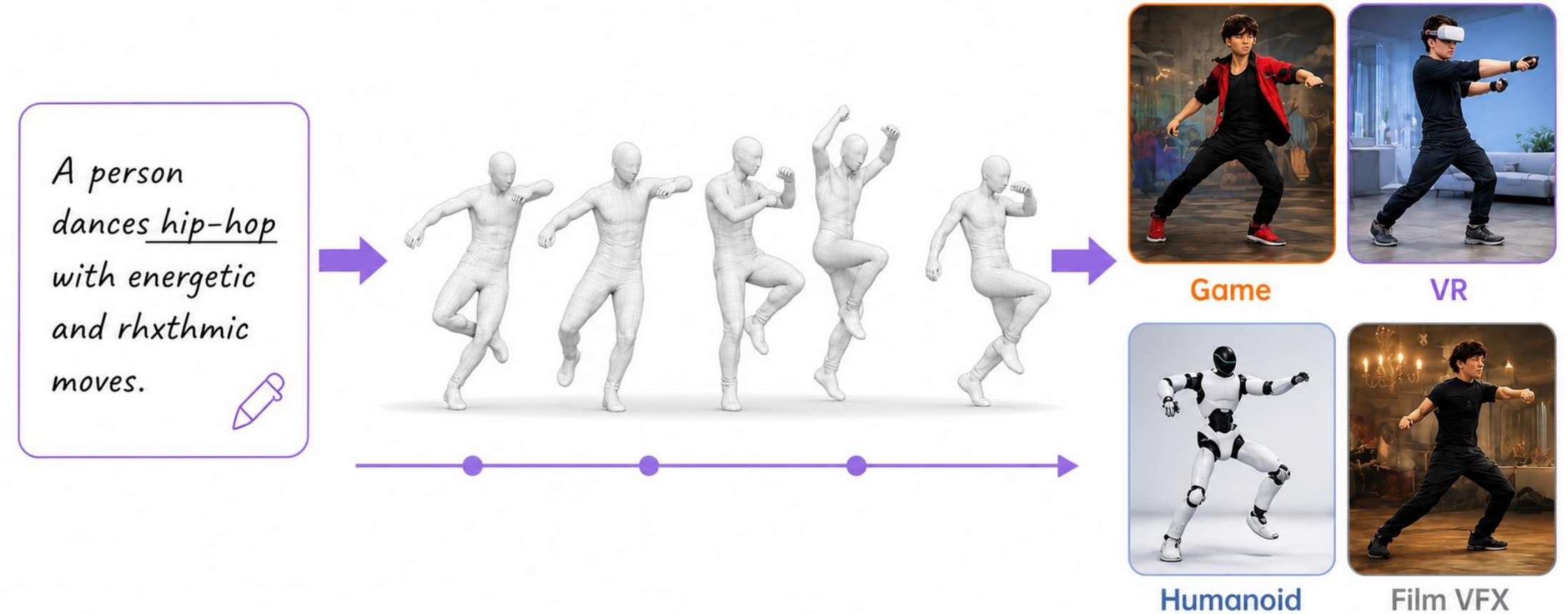
Jiahao Zhang[1,2], Joseph Liu[2], Young-Yoon Lee[2], Seonghyeon Moon[2]

Victor Zordan[2], Guy Tevet[3], C. Karen Liu[3], Stephen Gould[1]

Oren Jacob[2], HaomiaoJiang[2], Mubbasir Kapadia[2,4], Yizhak Ben-Shabat[2]

[1]Australian National University [2]Roblox [3]Stanford University [4]Rutgers University

Text to Motion Generation (T2M)

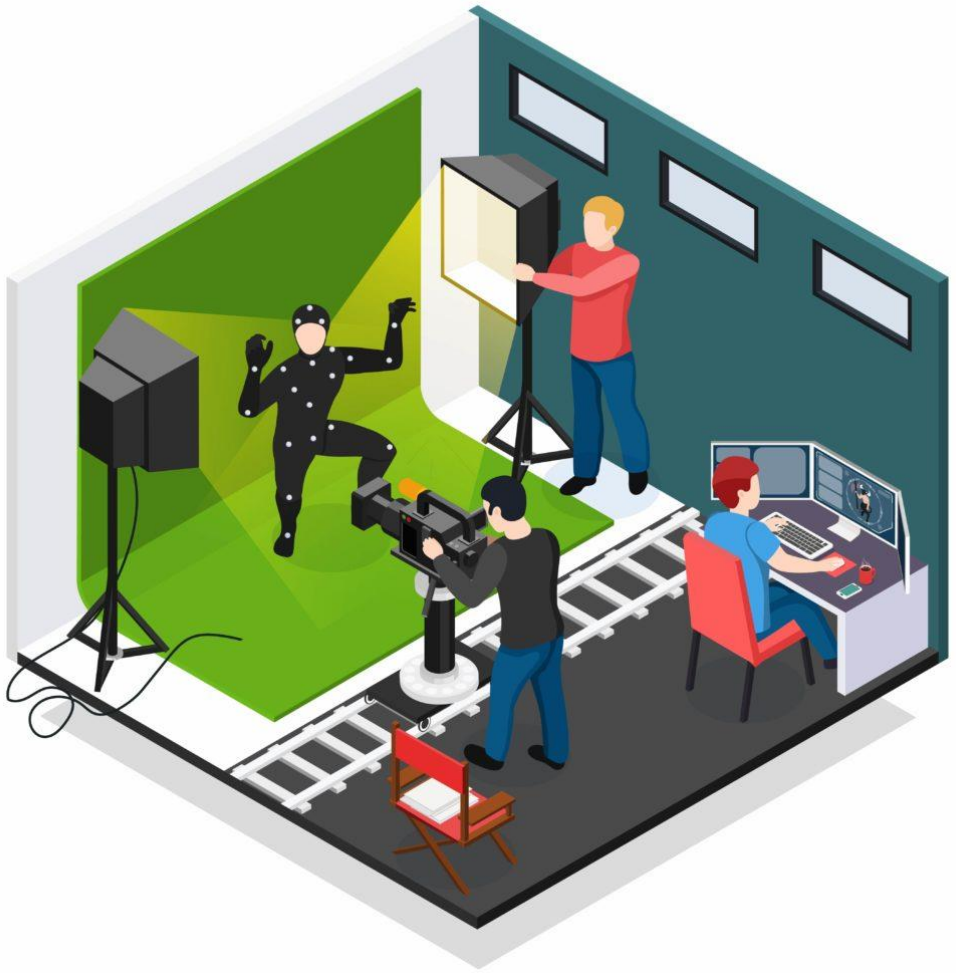


Motivation

- For images
  - ImageNet-21K (2012): **14 million** images for classification
  - LAION-5B (2022): **5 billion** text-image pairs for image generation
- What about T2M 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

Motion Capture is Complicated and Expensive



Our Solution

Same as **ImageNet** and **LAION** use the internet as the data source!

RoMo

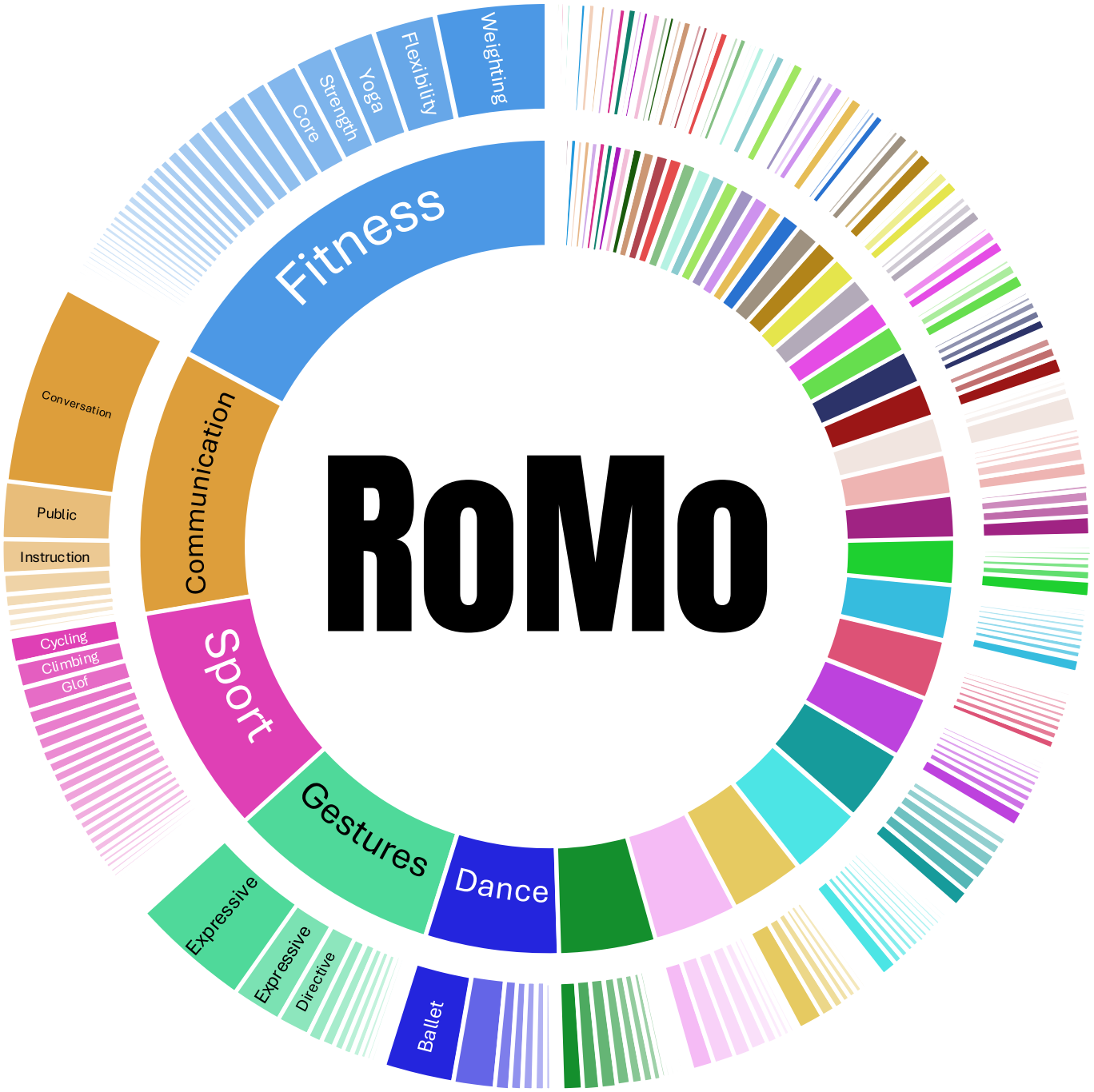
A Large-Scale, **R**ichly **O**rganized Dataset and Semantic Taxonomy for Human **Mo**tion Generation

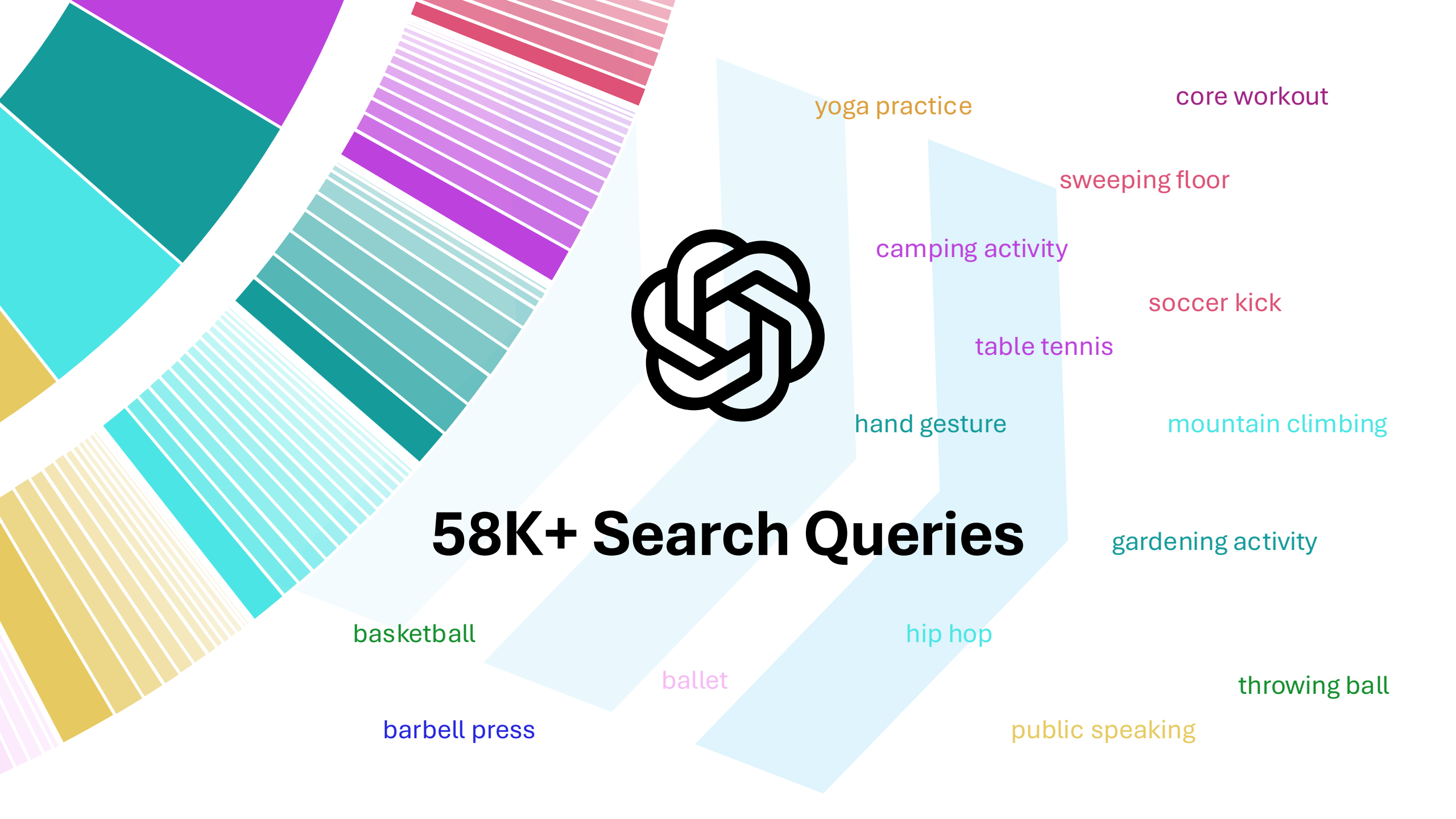
RoMo is organized into

54 Categories

Further divided into

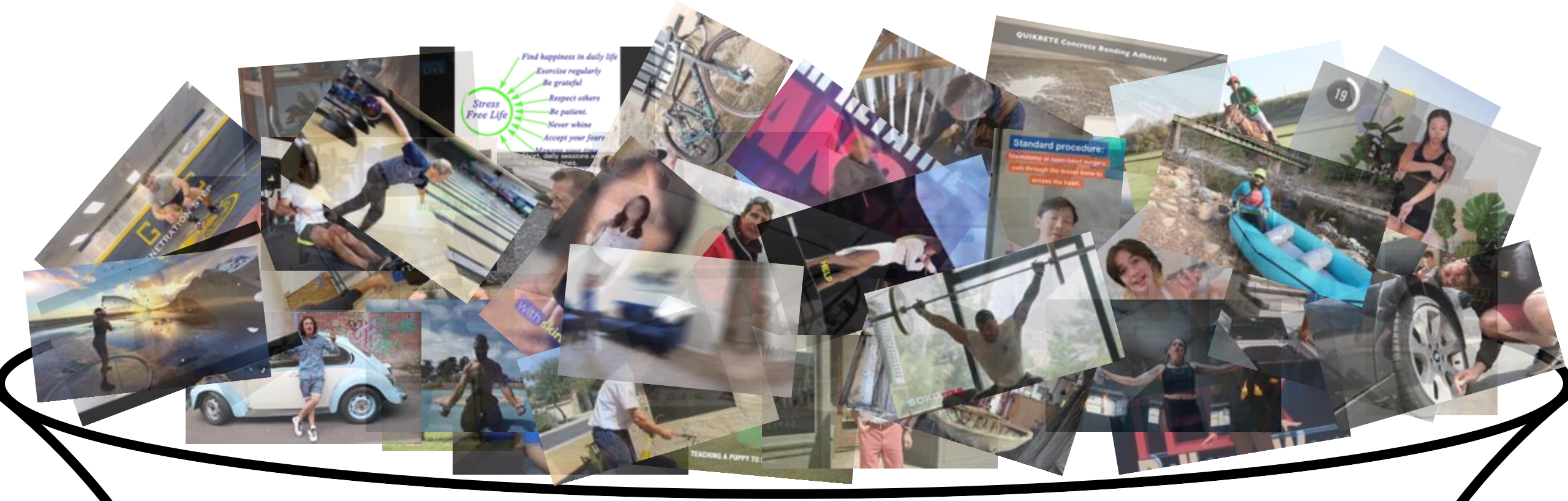
2,065 SubCategories




yoga practice
core workout
sweeping floor
camping activity
soccer kick
table tennis
hand gesture
mountain climbing
58K+ Search Queries
gardening activity
basketball
hip hop
ballet
throwing ball
barbell press
public speaking

14+ Years

of Unfiltered Videos



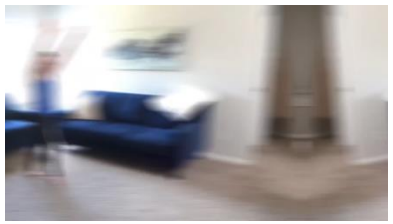
Filters

Only **1%** remaining

1,237+ Hours

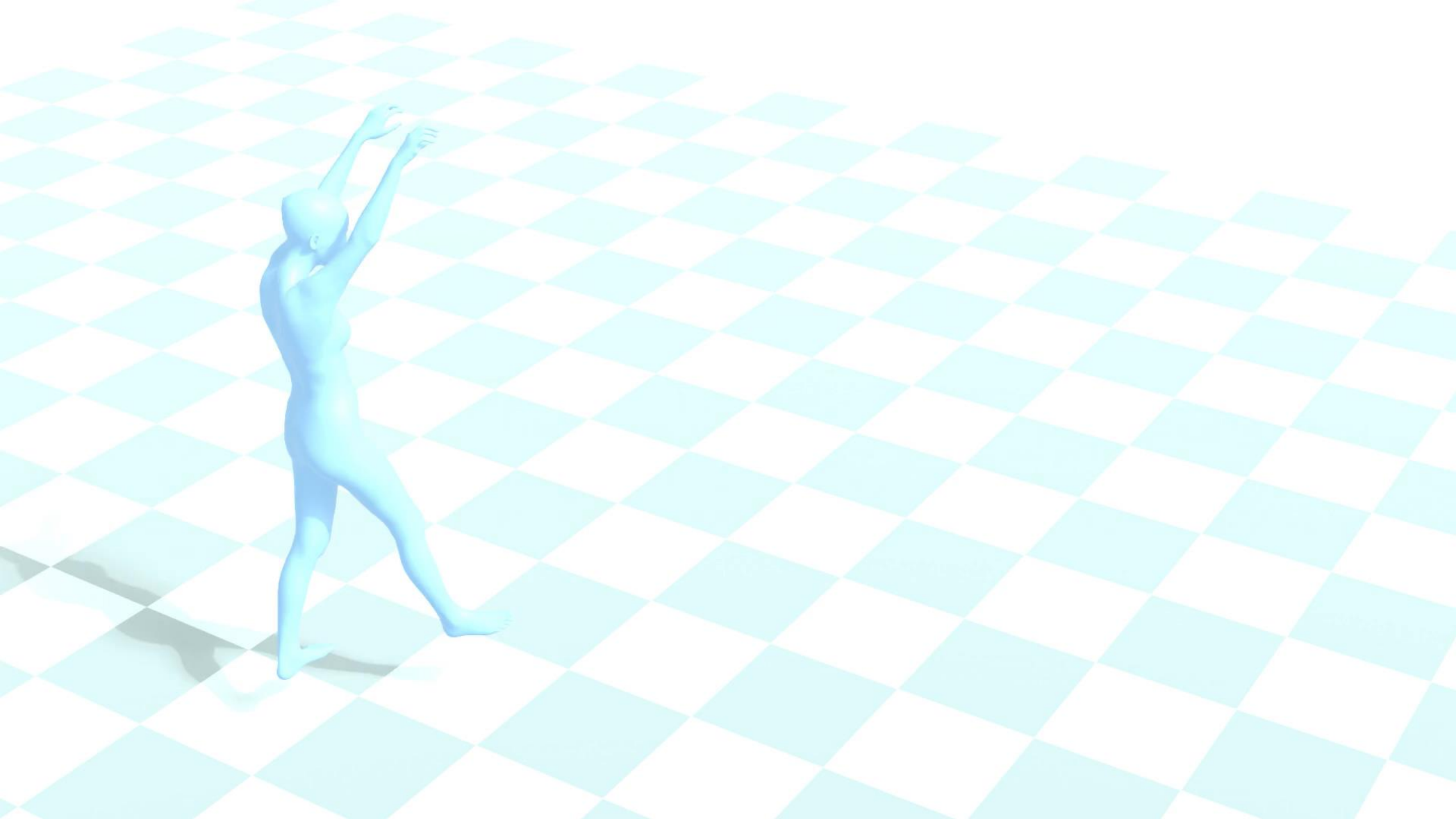
of Human Motion Videos

Qwen3-VL



```
[{
  "start": 0,
  "end": 16,
  "action": "lift barbell"
},
{
  "start": 16,
  "end": 19.8,
  "action": "walk away"
}]
```


GVHM



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."*

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."*

Sport → Bowling → Roll Bowling Ball

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

Outdoors → Rafting → Push raft with paddle

*"A person in a raft on the riverbank uses a paddle to **push off** from the shore."*

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

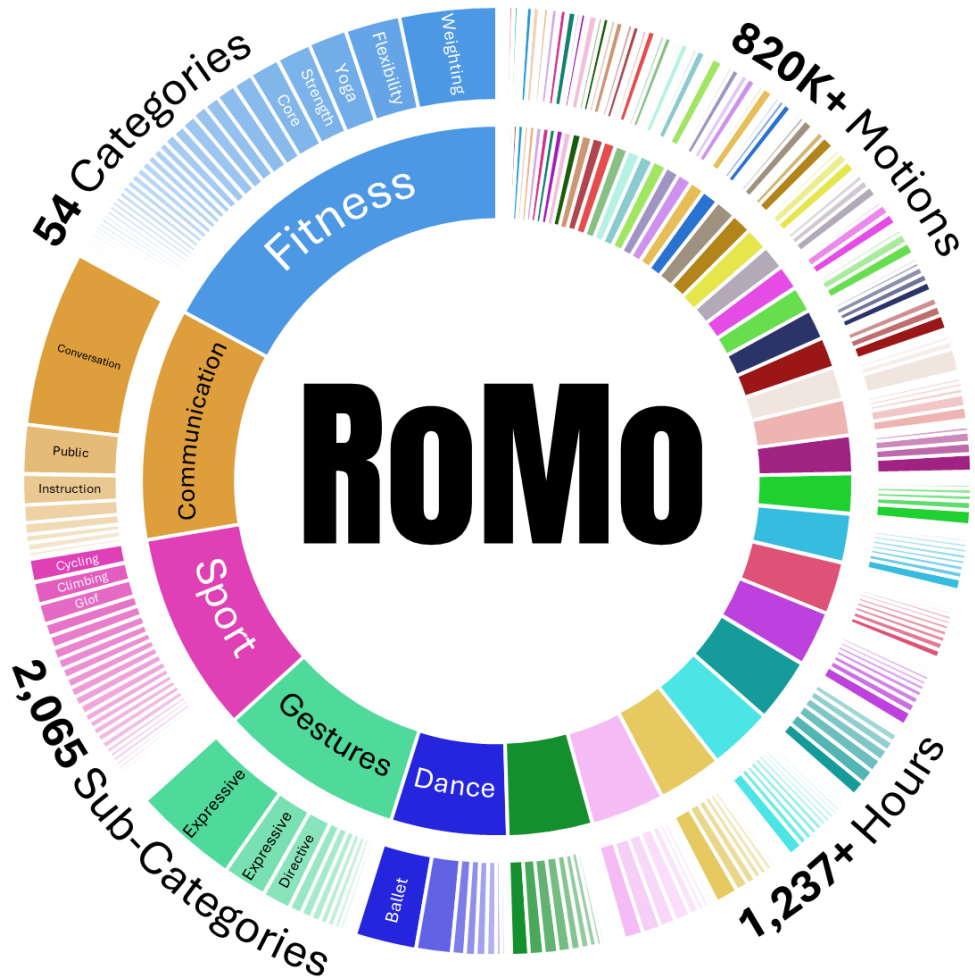
*"A person **picks** up the ball, takes a few steps forward, and **releases** it with a smooth motion."*

Video Games → Motion Gaming → Swing

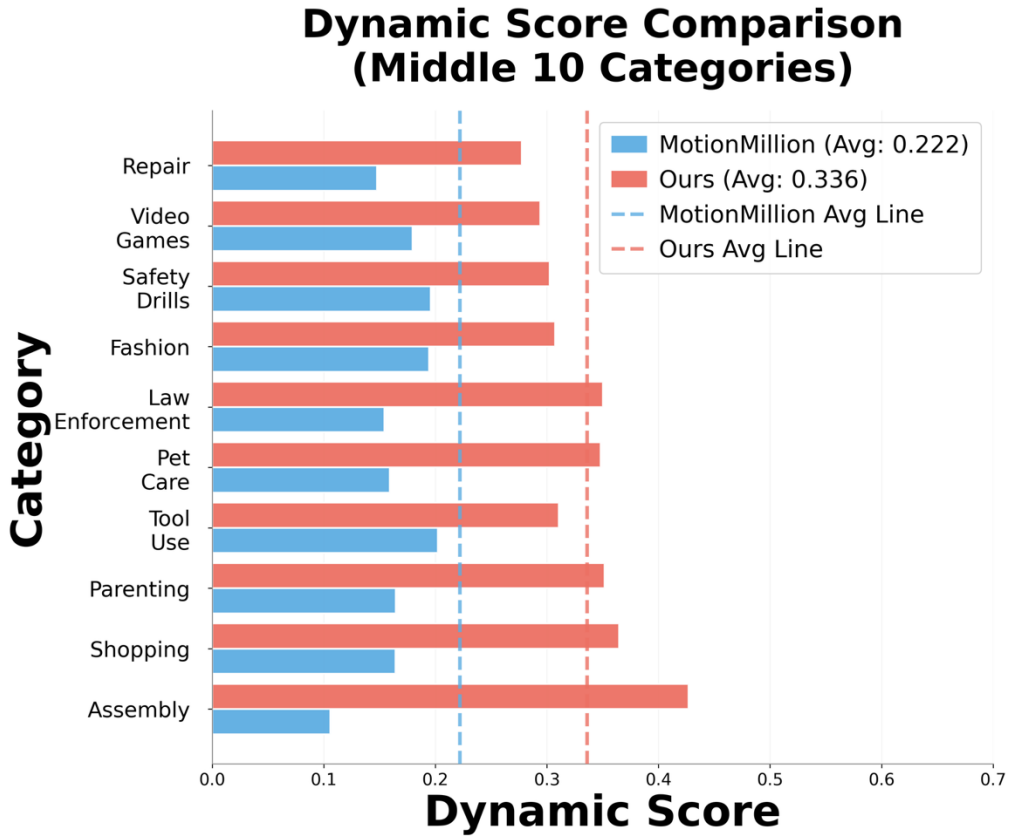
*"A person in a VR headset **swings** a virtual sword at floating objects, hitting them and causing them to shatter."*

Outdoors → Rafting → Push raft with paddle

*"A person in a raft on the riverbank uses a paddle to **push off** from the shore."*

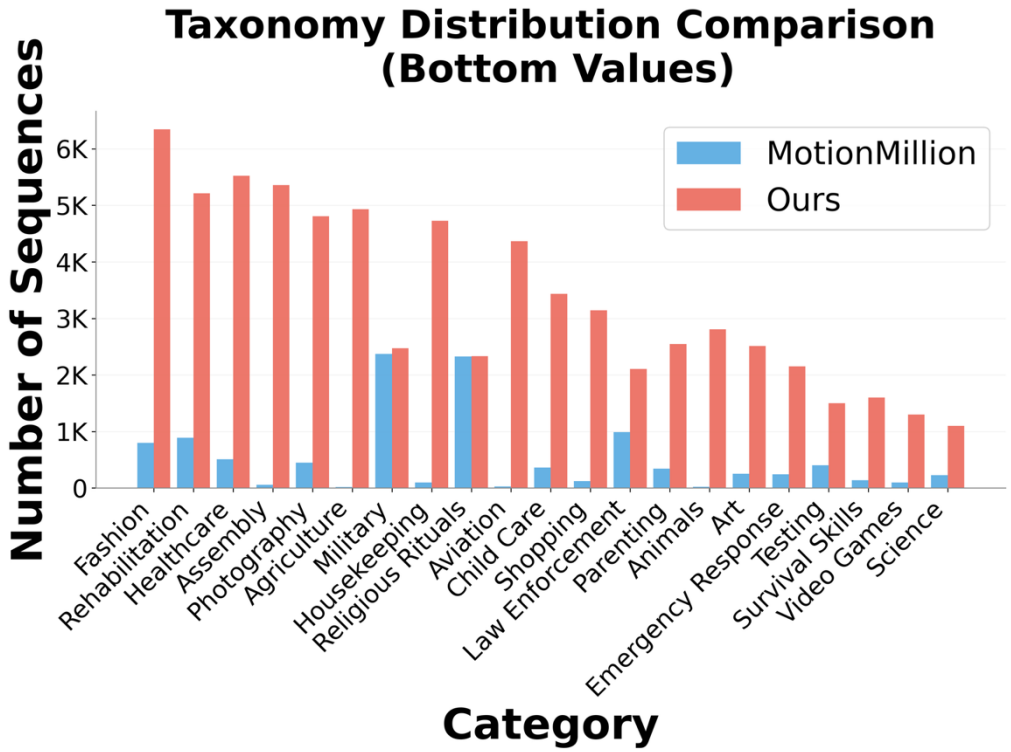

RoMo
54 Categories
820K+ Motions
2,065 Sub-Categories
1,237+ Hours
Fitness
Communication
Sport
Gestures
Dance
Weighting
Flexibility
Yoga
Strength
Core
Conversation
Public
Instruction
Cycling
Climbing
Golf
Expressive
Expressive
Directive
Ballet

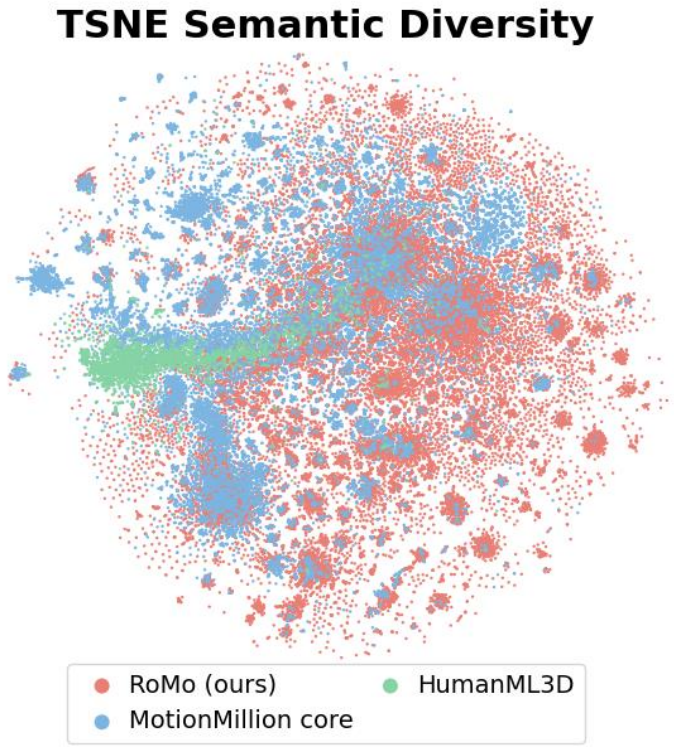
Superior Diversity, Coverage and Motion Dynamics



+41% higher dynamic score over MotionMillion.

Superior Diversity, Coverage and Motion Dynamics





+61.7% subcategory coverage over MotionMillion.

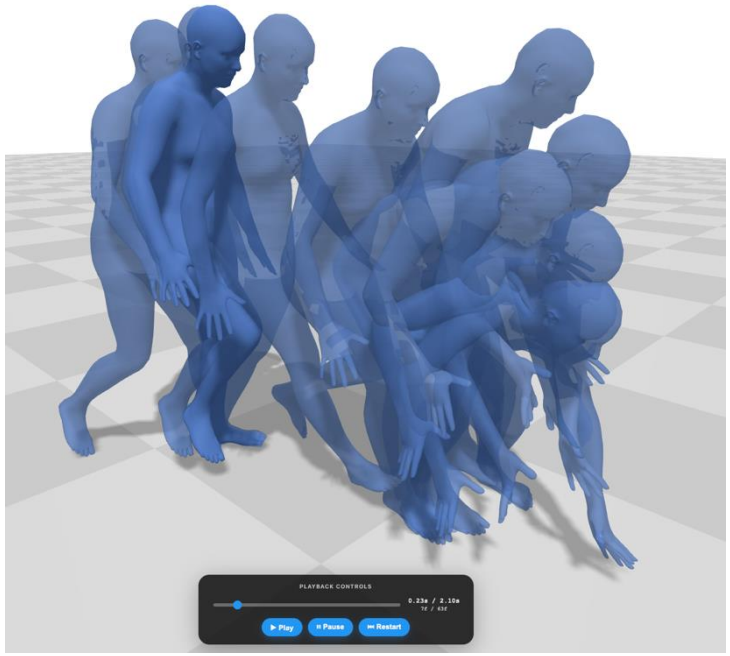
Training Modern Motion Generators on RoMo

Method	**Diversity** $\uparrow$	**FID** $\downarrow$	**Matching Score** $\uparrow$	**Dynamic Score** $\uparrow$	**Ground Penetration** $\downarrow$ $(\times 10^{-5})$	**Foot Skating** $\downarrow$ $(\times 10^{-3})$	**Floating** $\downarrow$ $(\times 10^{-2})$
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

Open-Source Motion Toolbox

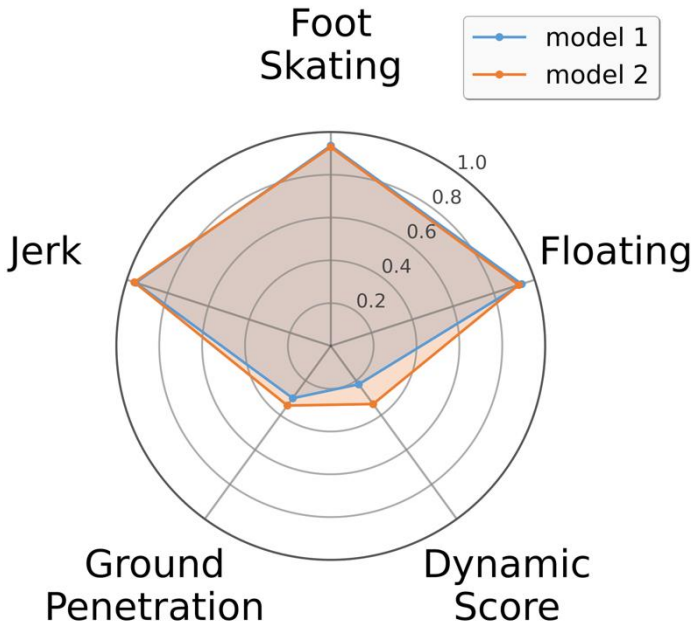
Interactive Visualization

Inspect and compare motion sequences directly in the browser.



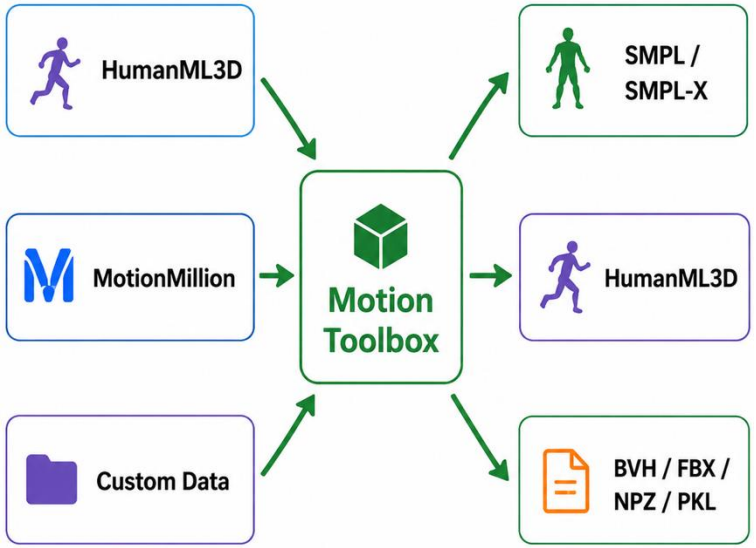
Unified Benchmarking

Reproducible evaluation tools for large-scale human motion research.



Unified Conversion

Convert between popular motion formats and research datasets.



Thanks!




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.