

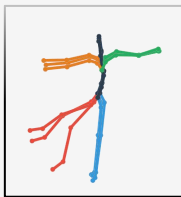
“A person squats down, jumps forward off both feet, and lands on both feet.”

Video

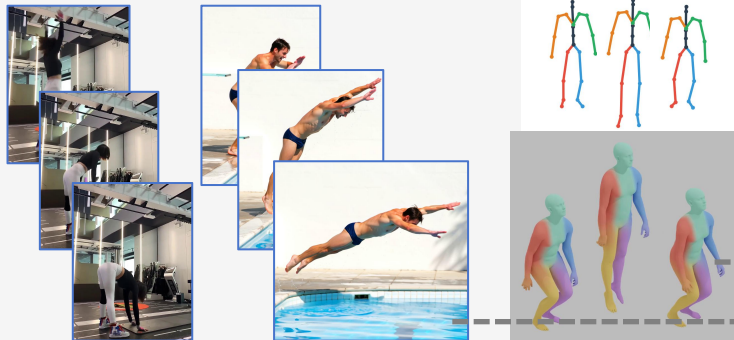


ViMoGen-228K

Motion



Diverse Motion Sources



in the wild

synthetic videos

mocap

Building from ViMogen-228K

1 Text-Guided Candidate Filtering



Rule-based Filtering

- Generic motion captions ✗
- Object/scene-heavy captions ✗



GPT-5.5 scoring & Filtering

- Visual noise score
- Object noise score
- Motion specificity score
- Low ambiguity score



High-quality caption candidate pool

MRBench-Train

- non-selected samples retained for training

5 fps frames



HY-Motion
Supplement
rare categories

3 Semantic Alignment & Disambiguation

Structured Semantic Alignment Evaluation



Does the person in the frames perform the described motion? Answer yes or no based only on visible evidence.

Aligned motion-text pairs

2 Taxonomy-Guided Balanced Sampling



Hierarchical Taxonomy

- Coarse: Locomotion
- Middle: Vertical
- Fine: Jumping forward



Qwen
Review



bge-en-icl
Deduplication



Repeat until
Target
Composition

Diverse sampled subset

4 Multi-granular Text Expansion

Multi-granularity Rewrite

- Concise:** A person squats...
- Standard:** A person squats down...
- Fine-grained:** The person starts from a standing position, bends their knees to lower into a squat...



Manual Validation



- ✓ Motion Diversity
- ✓ Source Diversity
- ✓ Multi-granular Captions
- ✓ Well-Aligned Motion-Text Pairs