Large Movement Models (LMMs)

What Would an LMM Be?

A Large Movement Model (LMM) is a deep neural network trained on high-dimensional human movement data. It learns the statistical and biomechanical patterns of human motion across a range of actions and intentions. Analogous to a Large Language Model (LLM) predicting the next word, an LMM predicts the next posture or joint state based on prior movements, contextual goals, and environmental constraints.

Core characteristics of an LMM:

- Probabilistic priors on joint motion and constraints
- Learned transition patterns between postures
- Intent-aware behavior generation (e.g., reaching, grasping)
- Recovery from perturbations and contextual adaptation

Analogies with LLMs

Mapping LLM structures to LMM equivalents:

Language (LLM) vs Movement (LMM):

- Token = word → Token = joint angle or posture vector
- Sentence = phrase → Sequence = motion segment
- Grammar = syntax → Biomechanics = kinematic rules
- Semantics = meaning → Intent = goal of movement
- Next-token prediction → Next-pose prediction

Built-in Constraints and Shortcuts

Unlike text, human motion is highly structured and constrained:

- Fixed set of joints and degrees of freedom
- Known joint limits and ranges
- Constant skeletal topology (tree-like graph)
- Bilateral symmetry (mirroring reduces training data needs)
- Reusable primitives (e.g., walk cycle, reach, grasp)
- Energy and goal-directed behavior constraints

Applications of LMMs

LMMs have powerful potential across domains:

- Robotics: humanoid motion generation, adaptive control
- Prosthetics: personalized motion prediction
- Animation: lifelike movement for games, VR, and film

- Healthcare: rehabilitation, fall prediction, therapy modeling
- Human-aware AI: collaborative robotics, avatars

How 3D Motion Can Be Captured from Video

There are two primary approaches:

1. Single-View 3D Pose Estimation

This method starts with 2D keypoint detection (e.g., OpenPose, BlazePose) followed by a model that infers 3D joint positions (e.g., VideoPose3D, VIBE).

Pros:
- Accessible from regular video
- No special hardware needed

Cons:
- Depth ambiguity
- Sensitive to occlusions and camera angle

2. Multi-Camera Motion Capture (MoCap)

Multiple synchronized cameras triangulate joint positions in 3D space. This is a standard in film and biomechanics labs.

Pros:
- High accuracy
- Reliable ground truth data

Cons:
- Requires controlled environments
- Expensive and hardware-intensive

Tools and Frameworks

Common tools used in the 3D motion extraction pipeline:

- OpenPose – 2D keypoint detection
- VideoPose3D – Infers 3D pose from 2D keypoints
- VIBE – Estimates full 3D mesh from video
- AMASS – Aggregated MoCap dataset
- SMPL/SMPL-X – Parametric body models
- PARE – Handles occlusion-robust pose estimation

Challenges

Despite advances, several issues remain:

- Ambiguity in depth from monocular video
- Occlusion handling and multi-person scenes
- Capturing fine detail (hands, faces)

- Missing physical dynamics (e.g., forces, torques)

Sports Footage as a Source of 3D Movement Data

Professional sports footage represents a highly valuable and underused source of natural human movement data for training Large Movement Models (LMMs). These events are extensively filmed from multiple synchronized camera angles, making them ideal for 3D motion reconstruction. Athletes also demonstrate extreme and varied movements, offering rich biomechanical learning opportunities.

Key Advantages:

- Multi-camera setups enable 3D triangulation of joint positions
- High-frequency, full-body dynamic movement under real conditions
- Diverse labeled action categories across sports (e.g., tackles, passes, jumps)
- Natural sequences of goal-directed motion and interaction

Training Value for LMMs:

- Predictive motion modeling (e.g., next-frame pose estimation)
- Learning multi-agent interaction and coordination
- Modeling biomechanics under high loads (e.g., collisions, sprints)
- Training synthetic avatars or virtual sport simulations

Suggested Data Pipeline:

- Ingest and synchronize footage from multiple angles
- Extract 2D keypoints (e.g., using OpenPose, HRNet)
- Triangulate to 3D using geometry or learned models
- Segment and clean sequences by movement phase or play
- Label actions and outcomes (e.g., tackle, pass, goal)
- Train LMM on pose sequences using generative architectures

Relevant Public Datasets:

- NBA Courtside 3D – multiview basketball movement data
- SoccerNet – annotated soccer player tracking and actions
- OpenTrack – athletics biomechanics from video and sensors

Predictive Applications of Large Movement Models (LMMs)

LMMs don't just interpret what a person is doing—they can predict what they're about to do. By analyzing temporal sequences of motion, LMMs enable powerful forecasting capabilities across human-aware systems. This predictive layer opens the door to applications that anticipate, adapt, and act proactively in dynamic environments.

Key Predictive Applications:

- Intent Prediction – Identify what someone is about to do based on early-stage motion. Useful in assistive systems, security, and collaborative robotics.
- Motion Completion – Given a partial trajectory, predict how it would plausibly resolve. Enables motion synthesis, recovery systems, and intelligent animation.
- Trajectory Forecasting – Predict where a person or group will move next. Crucial for smart infrastructure, autonomous systems, and evacuation planning.
- Task Anticipation – Forecast which action a person is preparing to take. Used in logistics, sports analytics, military training, and ergonomic support.
- Human-Robot Coordination – Enable robots to synchronize with human movement, preemptively aligning paths or tasks without verbal input.
- Accessibility Enhancement – Detect instability or hesitation, and proactively offer assistance. Can reduce falls, automate doors, and trigger alerts.