Real-Time Non-Verbal Human Robot Interaction with Reachy Mini

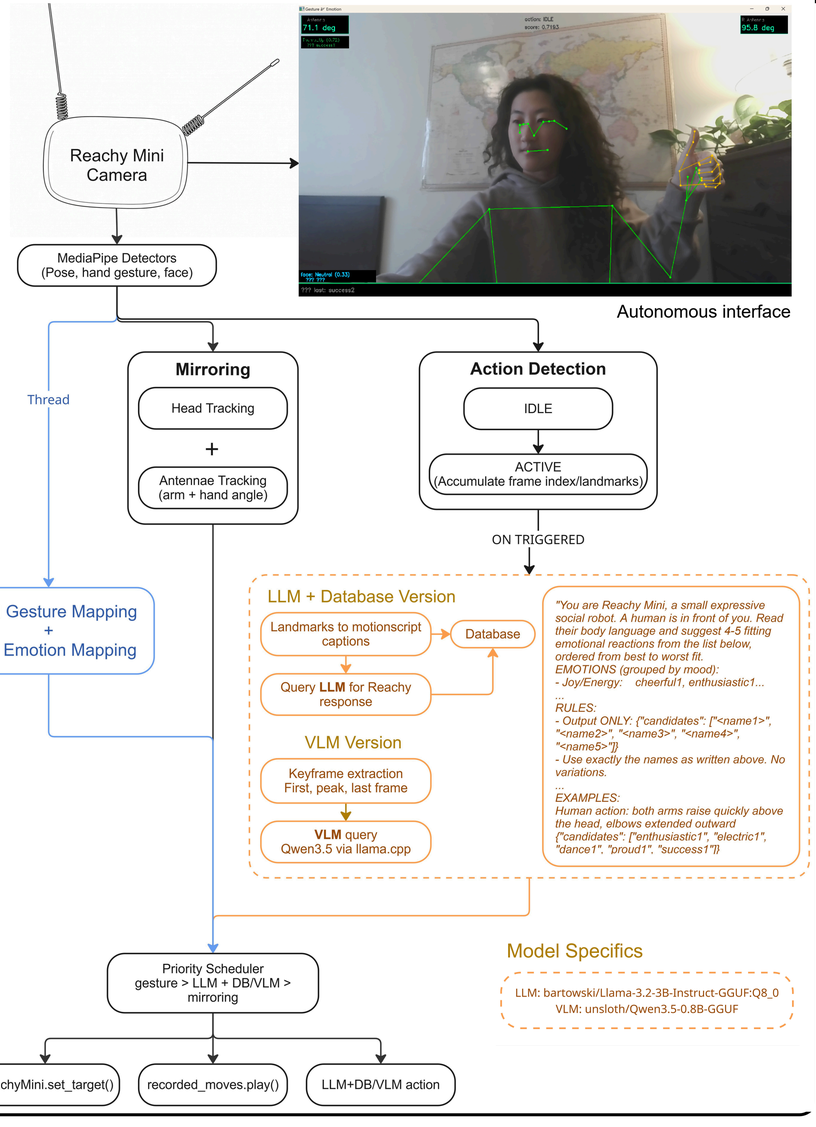
Quang Minh Dinh*, Stella Lin*, Gemmin Sugiura*, Bita Azari, Yasaman Etesam, Chuxuan Zhang, Angelica Lim *equal contribution
School of Computing Science, Simon Fraser University

Motivation

Non-verbal interaction is a fundamental component of natural human communication, yet in human-robot interaction it remains underexplored relative to language-centric approaches. We present a real-time non-verbal interaction system for Reachy Mini, a compact humanoid robot, that perceives and autonomously responds to human emotion and body motion.

Pipeline

The interaction system operates through three concurrent layers: background mirroring, hand gesture and facial emotion mapping, and keypose/keyframe detection for Large Language Model (LLM) [1] and Vision Language Model (VLM) [2] inferencing or action database retrieval. A priority scheduler arbitrates between layers, ensuring higher-priority reactions preempt background behaviours seamlessly.

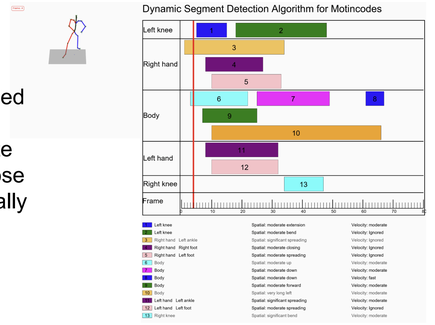


Action Detection

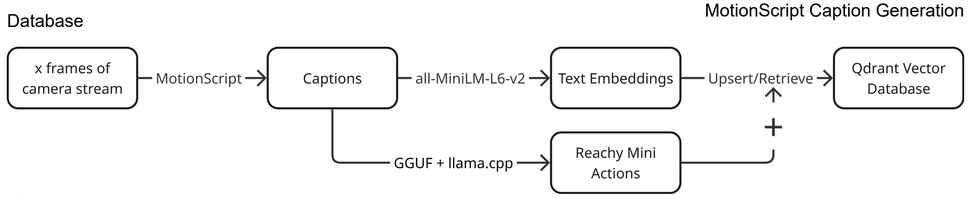
1. Pose, hand, and facial landmarks are extracted per frame using MediaPipe and converted into normalized motion features over a short temporal window.
2. A three-state machine transitions between Idle → Active → Stabilizing based on motion score thresholds.
3. The language model or action database is triggered only once motion settles after sufficient activity, preventing spurious reactions to transient movement.

Reaction Generation

Stabilized action windows are captioned in natural language using MotionScript [3]. Captions are embedded and matched against a vector database of prior interactions, enabling the system to retrieve or generate contextually appropriate robot responses. When no close match exists, a quantized language model running locally generates a novel response, which is then stored alongside its embedding for future retrieval.



MotionScript Caption Generation



Discussion

Effective non-verbal robot behaviour proved to depend more on timing and intentionality than on motion complexity. Responses using only 2-3 degrees of freedom, when timed deliberately, felt more natural than elaborate full-body movements executed poorly. Introducing small delays before a response made the robot feel proactive rather than reactive, as though it had processed and chosen a response, whereas instantaneous reactions read as mechanical. Discrete, bounded motions were also critical: continuous or looping movements prevented users from segmenting robot states, making the interaction feel unresponsive even when the robot was actively reacting. These findings suggest that for compact humanoids, a well-designed motion vocabulary and response timing model may matter more than expressive range.

Teleoperation

We present a web-based teleoperation system for the Reachy Mini robot that enables low-latency remote control at up to 60 Hz through multiple interaction modalities. The system provides real-time camera streaming, mobile-first control via accelerometer-based head pose mapping and dual virtual joysticks, autonomous gaze via MediaPipe-based face tracking, and access to pre-programmed procedural animations.

Human Study

We are currently conducting a Turing test-style user study (N = 50+) in which participants interacted freely with Reachy Mini using natural gestures and expressions. The robot operated under one of two conditions: autonomous mode (our system) or teleoperation (human-controlled baseline). Participants judge which condition they experienced and rate the naturalness of the robot's responses using the Godspeed Questionaire [4] which is one of the most frequently used questionnaires used in the area of Human Robot Interaction.

Teleoperator interface

Results

End-to-end pipeline latency averages ~3s for the MotionScript + LLM version and ~2s for the VLM version, measured from motion stabilization to robot response onset. Hand gesture and facial emotion recognition achieve ~70% and ~50% accuracy respectively across 6 gesture and 6 emotion classes under real-time conditions. Full user study results will be reported upon study completion.

References

[1] Meta AI, "The Llama 3 Herd of Models," arXiv:2407.21783, 2024.
[2] Qwen Team, "Qwen3 Technical Report," arXiv:2505.09388, 2025.
[3] Yazdian et al., "MotionScript: Natural Language Descriptions for Expressive 3D Human Motions," IROS 2025, IEEE.
[4] Bartneck et al., "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," Int. J. Social Robotics, 2009.