Non-Verbal Human-Robot Interaction with Reachy Mini: A Real-Time Multimodal System and Turing Test Evaluation

Quang Minh Dinh*, Stella Lin*, Gemmin Sugiura*, Bita Azari‡, Yasaman Etesam‡, Chuxuan Zhang‡, Angelica Lim†
Department of Computing Science, Simon Fraser University, Canada

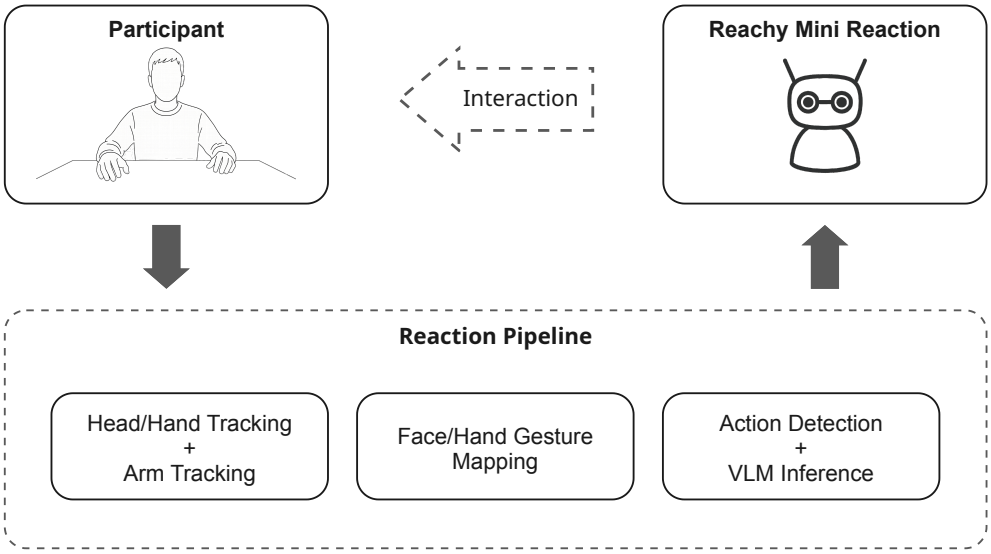

Fig. 1: System overview of our multimodal non-verbal interaction pipeline. A participant's gestures and expressions are captured via RGB camera and processed through four concurrent layers: a mirroring layer that dynamically tracks either head orientation or hand/arm motion, a hand gesture recognition layer, a facial emotion recognition layer, and an action detection module that triggers a locally-hosted VLM to select an appropriate response from a library of 81 pre-recorded expressions, all coordinated by a priority scheduler to drive Reachy Mini's outputs.

***Abstract*—Non-verbal cues such as gesture, posture, and facial expression are central to human communication, yet remain largely unaddressed in social robotics. Existing HRI systems either rely on language or require cloud inference, limiting real-time, expressive non-verbal response. We present a fully offline, real-time non-verbal interaction system for Reachy Mini, a compact humanoid robot, that perceives and autonomously responds to human body motion and emotion. Our system combines continuous head and arm mirroring, hand gesture recognition, facial emotion detection, and a locally-hosted Vision Language Model (VLM) that selects from a library of 81 pre-recorded expressions, coordinated by a priority scheduler across four concurrent perception layers. Running entirely on consumer hardware with $\sim$3 s end-to-end latency, the system produces behaviour convincing enough that participants in a Turing test-style study could not reliably distinguish it from human teleoperation (38% overall accuracy, below the 50% chance baseline). These results suggest that lightweight, offline multimodal perception is sufficient to produce socially credible non-verbal robot behaviour, opening a path toward accessible and deployable social HRI systems.**

***Index Terms*—Human-robot interaction, non-verbal interaction, social robotics, Turing test, emotion recognition, gesture recognition, multimodal perception**

I. Introduction

Nonverbal communication, including gesture, posture, and facial expression, plays an important role in human social interaction, yet human–robot interaction (HRI) research has largely emphasized language-based interfaces [1]. Designing autonomous nonverbal robot behavior requires specifying a prior over human gestures, affect, and interaction timing. When this prior is misspecified, it induces a mismatch between design-time assumptions and deployment-time interaction distributions (analogous to covariate shift [2]). In zero-shot settings, this makes policy design underdetermined and motivates iterative, empirically grounded approaches.

We address this with an distribution-matching evaluation framework inspired by generative adversarial training [3],

*Equal contribution
†Project supervisor
‡Authors listed in alphabetical order

in which human participants act as discriminators judging whether robot behavior is autonomous or controlled via teleoperation [4]. Teleoperation provides an empirical estimate of natural interaction and serves as a proxy for a high-quality behavioral policy. Human feedback is then used to iteratively refine the autonomous policy [5]. The perceptual layer of our system is grounded in real-time body language and affect recognition [6], [7], enabling responses to pose, gesture, and facial expression without speech. To evaluate the resulting behaviors, we measure perceived naturalness using the Godspeed Questionnaire [8].

Our system is implemented on Reachy Mini, a desktop humanoid robot platform designed for expressive human–robot interaction. Reachy Mini provides a total of 9 degrees of freedom (DoF), including 6 DoF in the head, 1 DoF for body rotation, and 2 DoF from independently actuated antennas, enabling rich nonverbal behaviors [9].

This paper contributes a nonverbal interaction pipeline for Reachy Mini, a teleoperation baseline, a formal evaluation framework for nonverbal HRI, and a pilot user study with $N = 17$ participants.

II. Methods

A. Problem Formulation

We define an interaction trajectory $\tau = \{(s_t, a_t)\}_{t=1}^{T}$, where $s_t \in \mathcal{S}$ denotes the observed nonverbal state at time t, comprising pose landmarks, facial expression features, and gesture descriptors, and $a_t \in \mathcal{A}$ denotes the robot's response action. Let $p_{\text{teleop}}(\tau)$ denote the empirical distribution over trajectories induced by human teleoperation, and $p_{\text{auto}}(\tau)$ the distribution induced by an autonomous policy G. We treat teleoperation as a surrogate for the ideal policy, under the assumption that a skilled human operator approximates the ground-truth distribution of natural, contextually appropriate nonverbal responses [4]. Our goal is to match p_{auto} to p_{teleop}.

To compare these distributions, we define a participant-specific discriminator $D_i(\tau) \in \{0, 1\}$ corresponding to a binary judgment of whether a trajectory is perceived as teleoperated. In contrast to standard adversarial settings, D_i is not learned but is given implicitly by human responses.

Conceptually, this corresponds to minimizing the divergence $D_{\text{KL}}(p_{\text{auto}} \,\|\, p_{\text{teleop}})$. As this is intractable, we instead use human discrimination as an empirical proxy and define the generator objective

$$J(G) = \mathbb{E}_i \left[\left| \mathbb{E}_{\tau \sim p_{\text{teleop}}}[D_i(\tau)] - \mathbb{E}_{\tau \sim p_{\text{auto}}}[D_i(\tau)] \right| \right], \quad (1)$$

which measures the distinguishability of the two distributions under human judgment. This formulation cancels participant-specific response biases independent of trajectory type and is minimized when the two conditions are indistinguishable. This is analogous to the adversarial objective in generative modeling [3], where the generator is optimized to fool a discriminator, except here the discriminator is human and fixed. With a single trajectory observation per condition per participant, this reduces to the binary judgments collected in our pilot study.

Indistinguishability alone is insufficient; the policy must also align with human perceptual expectations. We therefore impose moment-matching constraints over the K Godspeed dimensions:

$$\mathbb{E}_{p_{\text{auto}}}[\phi_k] = \mathbb{E}_{p_{\text{teleop}}}[\phi_k], \quad k = 1, \ldots, K, \quad (2)$$

where ϕ_k corresponds to anthropomorphism, animacy, likeability, intelligence, and safety [8]. In practice, these expectations are approximated using participant-level averages.

Thus, our objective is to develop a policy that produces interaction trajectories indistinguishable from teleoperation under human judgment while matching human perceptual expectations along key social dimensions.

B. Mirroring

Behavioral imitation and mirroring construct the basis for social interactions and acknowledgment between individuals [10]. To provide a sense of connectedness and reduce the stillness and delay in interactions with humans, Reachy Mini was designed to perform mirroring as the default behavior when no reaction is ready, or the robot is waiting for action signals. Specifically, at time step t, a face motion score m_t^f and a hand motion score m_t^h are assigned to the corresponding camera frame, which are then used to determine the target for the robot head mirroring. The new head, pose, and antenna positions are then calculated from the target position. Mirroring is employed at a fix rate of $f = 50Hz$.

Motion Estimation and Mirroring Target Selection. To track the human motion across different camera frames, all tracked landmarks are expressed in a body-centric coordinate frame, relative to the absolute shoulder positions lm_t^{s1} and lm_t^{s2} at time t:

$$p_t^f = \frac{1}{\|lm_t^{s1} - lm_t^{s2}\|_2} \left(lm_t^n - \frac{lm_t^{s1} + lm_t^{s2}}{2} \right) \quad (3)$$

$$p_t^h = \frac{1}{\|lm_t^{s1} - lm_t^{s2}\|_2} \left(lm_t^w - \frac{lm_t^{s1} + lm_t^{s2}}{2} \right), \quad (4)$$

where p_t^f is the relative face position, p_t^h is the relative position of the hand with the higher confidence score, and lm_t^n and lm_t^w are the corresponding absolute positions for nose and wrist.

The motion scores are estimated by applying an exponential moving average (EMA) to the displacement of the normalized position in time:

$$m_t = \alpha_m m_{t-1} + (1 - \alpha_m)\|p_t - p_{t-1}\|_2 \quad (5)$$

To prevent rapid switching, we employed hysteresis for motion detection with two thresholds:

$$\text{moving} = \begin{cases} m_t > \tau_m^{\text{low}}, & \text{if already moving} \\ m_t > \tau_m^{\text{high}}, & \text{otherwise} \end{cases} \quad (6)$$

The target for the robot head mirroring is set to the hand if it is active and the face is inactive. Otherwise, the target is set to the face, and the two antennas are mapped to the two arms.

Control mapping. The robot head's rotation is controlled by some selected target's signals:

$$\theta_t^{\text{pitch}} = \theta_p y_t, \quad \theta_t^{\text{yaw}} = -\theta_y x_t, \quad \theta_t^{\text{roll}} = -\lambda_r \theta_t^{\text{t\_roll}}, \tag{7}$$

where θ_p, θ_y and λ_r are tunable hyperparameters, x_t and y_t construct the target's absolute position in the camera frame coordinate system, and $\theta_t^{\text{t\_roll}}$ is the extracted target roll. For hand, θ_t^{roll} is set to 0.

To ensure smooth updates, an EMA is applied to each Euler angle θ_t to get the actual control angle $\tilde{\theta}_t$:

$$\tilde{\theta}_t = \alpha_r \tilde{\theta}_{t-1} + (1-\alpha_r)\theta_t, \quad \tilde{\theta}_0 = 0 \tag{8}$$

If there is no valid signal, the control angle slowly decays to a neutral angle:

$$\tilde{\theta}_t = \gamma \tilde{\theta}_{t-1} \tag{9}$$

For antenna mapping, the angles θ_t^{arm1} and θ_t^{arm2} are extracted from the camera frame, where arm direction is defined as the direction from a shoulder to the corresponding wrist. The antenna angles are also filtered using EMA, following Eq. 8 and Eq. 9.

The head control angles are represented as a unit quaternion $\mathbf{q}_t \in \mathbb{R}^4$. To ensure smooth movements, an EMA is applied in quaternion space using normalized linear interpolation:

$$\tilde{\mathbf{q}}_t = \alpha_q \tilde{\mathbf{q}}_{t-1} + (1-\alpha_q)\mathbf{q}_t, \quad \tilde{\mathbf{q}}_t \leftarrow \frac{\tilde{\mathbf{q}}_t}{\|\tilde{\mathbf{q}}_t\|_2}, \tag{10}$$

where $\tilde{\mathbf{q}}_0 = [0,0,0,1]$.

To extend the robot head's horizontal rotation and maintain a seamless head and body motion, $\theta_{body}^{\text{yaw}}$ is slightly adjusted whenever $\theta_{head}^{\text{yaw}}$ exceeds a threshold θ_{th}:

$$\theta_{body}^{\text{yaw}} \leftarrow \theta_{body}^{\text{yaw}} + \text{sign}(\theta_{head}^{\text{yaw}}) \cdot \frac{\delta}{f}, \tag{11}$$

where δ is a tunable body follow rate.

C. Hand Gesture Recognition

Hand gesture recognition provides a direct, low-latency channel from discrete human signs to robot emotional expressions. MediaPipe's GestureRecognizer runs on every camera frame, returning both a classified gesture label and the corresponding hand landmarks. To suppress spurious detections caused by transient hand positions, recognized labels are accumulated in a sliding debounce buffer of length N_{deb}; a gesture is considered *stable* only when the same label appears in all N_{deb} consecutive frames. Formally, let $g_t^{(i)}$ denote the gesture label at frame $t-i$. A stable gesture $\hat{g}_t$ is declared when:

$$\hat{g}_t = g_t^{(0)} \quad \text{if } g_t^{(0)} = g_t^{(1)} = \cdots = g_t^{(N_{\text{deb}}-1)}, \tag{12}$$

and no stable gesture is emitted otherwise.

Upon detecting a new stable gesture distinct from the previously triggered label $\hat{g}_{t-1}$, a corresponding pre-recorded emotional expression is selected from the library and dispatched to the robot. The gesture-to-emotion mapping covers 6 hand sign classes. Gesture-triggered responses occupy the highest priority level in the scheduler, preempting both face emotion reactions and VLM-driven responses. Under real-time operating conditions, gesture recognition achieves approximately 70% accuracy across the 6 classes.

D. Face Emotion Recognition

To enable Reachy Mini to react to the participant's affective state, we employ EmotiEffLib [11], [12] to predict facial emotions from the camera stream in real time. Emotions are classified into 7 categories: happiness, sadness, anger, surprise, fear, disgust, and neutral. To reduce computational load, inference runs every N_f frames on a face crop extracted from MediaPipe face landmark detections. Predicted labels are accumulated in a debounce buffer of length $N_{f,\text{deb}}$; for each of the 6 non-neutral categories, a reaction is triggered only when the confidence score exceeds threshold τ_f for n_f consecutive frames. Face emotion reactions achieve approximately 50% accuracy across 6 emotion classes under real-time conditions.

Face emotion reactions serve as a secondary channel, activated only when no hand gesture is simultaneously detected. This priority ordering ensures that explicit intentional gestures always take precedence over passive facial expression responses.

E. Action detection

The action detection pipeline estimates the overall motions from the camera frames to signal the trigger of the vision language model (VLM).

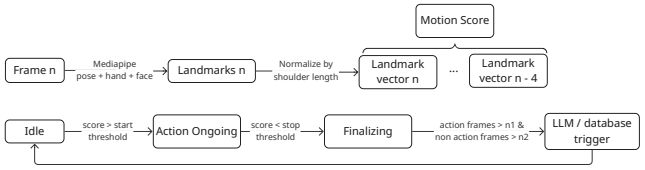

Fig. 2: Action Detection Pipeline. Motion scores are calculated for all normalized landmarks between adjacent time steps. Motion states **Idle**, **Active**, **Stabilizing**, and **Triggered** rotate based on different threshold criterion.

Pose Motion Score Estimation. At time t, each landmark lm_t^i of $\mathbf{n_d}$ pose, hand, or face landmarks detectable with Mediapipe in frame t is converted to the body-centric coordinate frame using the two shoulder landmarks, as described in Sec. II-B. We construct a feature vector $h_t \in \mathbb{R}^{\mathbf{n_d}}$ that contains all detected normalized landmarks and a mask vector $\mathbf{m}_t \in \mathbb{R}^{\mathbf{n_d}}$ that contains 1s for detected landmarks and 0s for missing landmarks. The pose motion score μ_t is calculated as follow:

$$\mu_t = \frac{\|\mathbf{m}_t \odot \mathbf{m}_{t-1} \odot (h_t - h_{t-1})\|_2}{\sqrt{\mathbf{m}_t^T \mathbf{m}_{t-1}}} \tag{13}$$

To reduce noise, we apply a short temporal smoothing on the recent motion values:

$$\tilde{\mu}_t = \sum_{i=1}^{5} \frac{\mu_{t-i+1}}{5} \tag{14}$$

Fig. 2 illustrates our state machine for action detection with 4 stages: **Idle**, **Active**, **Stabilizing**, and **Triggered**. In the **Idle** stage, there is no motion, and the transition to **Action** happens when $\tilde{\mu}_t > \tau_d^a$. **Stabilizing** serves as the signal to conclude the action, which happens after $\tilde{\mu}_t < \tau_d^s$. To avoid glitches and temporary mid-action stops, the number of active frames n_d^a must exceed a threshold η_a and the number of stable frames n_d^s must exceed a threshold η_s to transit to **Triggered**. If the condition is not satisfied, the state moves back to **Active**. **Triggered** signals the VLM call, which will be elaborated in Sec. II-F, and afterward the state moves to **Idle** again.

F. Vision Language Model

When the action detector transitions to the **Triggered** state (Sec. II-E), a set of keyframes extracted from the buffered camera stream is forwarded to a locally-hosted Vision Language Model (VLM) for response selection. Three keyframes are extracted from the detected action clip: the first frame, the peak-motion frame, and the last frame. These are passed to a quantized Qwen3-0.5B model [13] running via `llama.cpp`, operating fully offline on consumer hardware.

Prompt Design. The VLM receives the three keyframes together with a structured prompt describing Reachy Mini's identity as a small expressive social robot and instructing it to read the human's body language and suggest a fitting emotional reaction. The prompt supplies the full list of 81 available pre-recorded expressions, grouped by mood category, and constrains the model to output a strict structured format:

`candidates: [<name1>,'<name2>, <name3>]`

where each candidate name is drawn exactly from the expression library. Requiring multiple ranked candidates provides robustness: if the top candidate fails to execute, the system falls back to the next in order.

Keyframe Extraction and Queuing. Keyframes are extracted from a rolling frame buffer of length L_{buf} maintained throughout the interaction. Given a triggered event with n active frames and a peak-motion frame offset δ_{peak} from the end of the buffer, the three keyframe indices are:

$$i_{\text{first}} = L_{\text{buf}} - n \tag{15}$$

$$i_{\text{peak}} = L_{\text{buf}} - 1 - \delta_{\text{peak}} \tag{16}$$

$$i_{\text{last}} = L_{\text{buf}} - 1 \tag{17}$$

Extracted keyframes are enqueued in a dedicated background `VLMWorker` thread with a bounded queue of depth 4, decoupling inference latency from the main perception loop. If the queue is full or a VLM inference is already in flight, the incoming clip is dropped to prevent backlog accumulation.

Thinking Behaviour. While VLM inference runs in the background, Reachy Mini plays a short "thinking" animation randomly selected from a curated set. This deliberate delay serves two purposes: it prevents instantaneous reactions that participants perceived as mechanical in our pilot study observations, and it fills the 2–3 s inference window with socially legible behaviour. End-to-end latency from motion stabilization to robot response onset averages approximately 3 s for the VLM version.

G. Teleoperation

Fig. 3: Web-based teleoperation interface for Reachy Mini. The operator streams live camera footage while controlling the robot via a mobile device, where physical phone orientation maps directly to the robot's head pose, creating an intuitive first-person embodiment analogous to a VR system.

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, phone-based head pose control via accelerometer mapping, dual virtual joysticks for antenna control, a real-time 3D third-person view of Reachy Mini for spatial awareness, and access to procedural animations. This system serves as the human-controlled baseline condition in our study, providing an empirical approximation of the ideal interaction distribution p_* against which the autonomous policy is evaluated.

III. Experiments and Results

A. Ablation Study

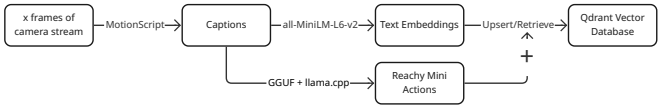

Fig. 4: Pipeline V1 with MotionScript and the LLM. Captions from MotionScript are used to generate actions for Reachy Mini and as the key to insert or retrieve an entry from the Qdrant database.

Fig. 4 shows our first pipeline V1 with a two-stage response generation approach: pose landmarks from the detected action clip were captioned in natural language using MotionScript [14], and the resulting caption was forwarded to a locally-quantized Llama-3.2-3B-Instruct model [15] (Q8_0 GGUF) to select a response. This chain averaged 5–9 s end-to-end latency, causing robot responses to feel contextually decoupled from the triggering interaction. To mitigate the latency issue, we employed Qdrant [16] as the vector database to cache the generated motion. We embedded the MotionScript's caption using all-MiniLM-L6-v2 [17] and used the embedded vector to insert a new motion to the database, and to retrieve an entry if the cosine similarity between the current vector and the entry's

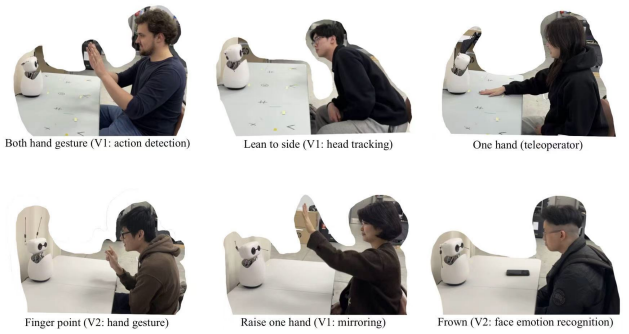

Fig. 5: Sample participant interactions during the pilot study. *Top row, left to right:* a participant performing a two-handed gesture triggering action detection (V2); leaning to the side with head tracking active (V2); a teleoperator controlling the robot with one hand (V2). *Bottom row:* a participant using a finger-point gesture for hand gesture recognition (V3); raising one hand while the mirroring layer tracks arm elevation (V2); exhibiting a frown expression detected by the facial emotion recognition module (V3).

vector is higher than a threshold τ_v. To diversify the responses for similar movements, we adjusted the LLM's output to a subset (4-5) of actions and randomly selected one to perform. For frequent movements such as waving or shaking the head, the motion database reduced the latency to 0 s. However, the latency issue persisted for emergent movements.

We proposed our new pipeline V2, in which a single VLM step replaced the MotionScript + LLM chain and operated directly on RGB keyframes (Sec. II-F). Our new pipeline reduced latency to $\sim$2 s, without the need for a motion database.

Our current pipeline V3 additionally introduced continuous mirroring with hands and face as a persistent background behaviour, reduced response repetition, and re-tuned detection thresholds based on the observations from pipeline V2. These changes are reflected in the pilot study results: teleoperation identification accuracy fell from 25% in V2 to 14% in V3 (Tab. I), indicating that V3 autonomous behaviour was more difficult to distinguish from human-controlled responses.

B. Pilot Study

Study Design. We conducted a Turing test-style pilot study to evaluate whether participants could distinguish autonomous robot behaviour from teleoperated behaviour, and to measure perceived social qualities under each condition using the Godspeed Questionnaire [8]. The study included 17 participants ($M_{\text{age}} = 22.2$, range 19–31). Each participant interacted with Reachy Mini for one minute and was then asked to judge whether the robot they experienced was operating autonomously or was controlled by a human operator.

The study used two between-subject cohorts corresponding to two pipeline versions (V2 and V3), and one within-subject cohort under V3. Between-subject participants were assigned to a single condition (Autonomous or Teleoperated). Within-subject participants ($n = 4$) experienced both conditions in random order and completed the Godspeed Questionnaire after each. Between Pipeline V2 and V3, we introduced several targeted improvements based on observations from V2 participants: repetitive behaviour patterns were reduced, real-time head and hand tracking via mirroring (Sec. II-B) was added, and motion and reaction parameters were re-tuned.

Measures. Participants completed the Godspeed Questionnaire [8] after each interaction. The questionnaire assesses five dimensions on 5-point semantic differential scales: Anthropomorphism (5 items), Animacy (6 items), Likeability (5 items), Perceived Intelligence (5 items), and Perceived Safety (3 items). The Turing test judgment was recorded as a binary forced-choice (Autonomous or Teleoperated).

C. Results

TABLE I: Turing Test Prediction Accuracy by Condition and Pipeline

Pipeline	Condition	Correct	Total	Accuracy
V2	Autonomous	4	5	80%
V2	Teleoperated	1	4	25%
V3	Autonomous	2	5	40%
V3	Teleoperated	1	7	14%
V2 + V3	Autonomous	6	10	60%
V2 + V3	Teleoperated	2	11	18%

Turing Test Accuracy. Tab. I reports prediction accuracy by condition. In the Autonomous condition, 6 of 10 observations (60%) were correctly identified. In the Teleoperated condition, only 2 of 11 observations (18%) were correctly identified, with 9 of 11 participants incorrectly judging the teleoperated robot as autonomous. Overall accuracy across all observations was 38% (8/21), below the 50% chance baseline as shown in Fig. 6. A one-sample binomial test against the chance rate of 0.5 was not significant ($p = 0.38$), indicating that participant performance was indistinguishable from random guessing.

The pronounced asymmetry between conditions reveals a strong *autonomy assumption bias*: participants were far more likely to attribute autonomous behaviour to the robot regardless of the actual condition. This effect was more pronounced in V3 than V2 (Tab. I), suggesting that the V3 pipeline improvements reduced behavioural cues that previously allowed participants to identify teleoperation.

TABLE II: Mean Godspeed Subscale Scores by Condition (1–5 scale)

Condition	Anth.	Anim.	Like.	Intel.	Safety
Autonomous	2.72	3.22	3.96	3.08	3.00
Teleoperated	3.35	3.59	4.38	3.33	3.30

Godspeed Questionnaire. Tab. II reports mean subscale scores by condition across all observations ($N = 21$) as illustrated in Fig. 7.. Teleoperated interactions received higher

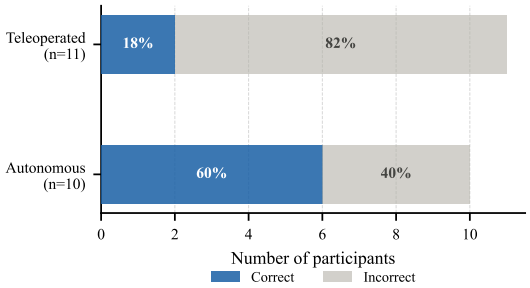

Fig. 6: Turing test prediction accuracy by condition. Participants in the Teleoperated condition identified the correct control mode at only 18% (2/11), well below the 50% chance baseline, indicating a strong autonomy-assumption bias. Participants in the Autonomous condition performed closer to chance at 60% (6/10).

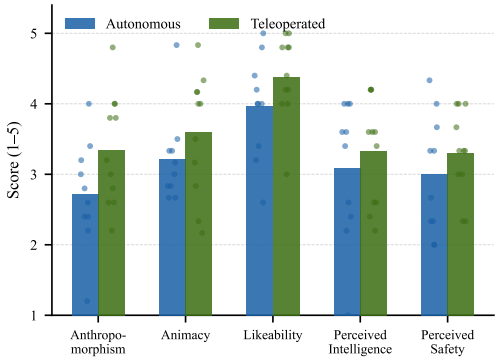

Fig. 7: Mean Godspeed subscale scores by condition. Bars show condition means; overlaid points show individual observations ($N = 21$). Teleoperated interactions received higher scores across all subscales except Perceived Safety, though differences should be interpreted as descriptive trends given the pilot study sample size.

mean scores on Anthropomorphism (3.35 vs. 2.72), Animacy (3.59 vs. 3.22), Likeability (4.38 vs. 3.96), and Perceived Intelligence (3.33 vs. 3.08). Perceived Safety scores were comparable across conditions (Teleoperated: 3.30; Autonomous: 3.00). These descriptive trends are consistent with the expectation that human-controlled movements convey richer social cues; however, given the pilot study sample size ($N = 17$ participants), formal inferential comparisons are underpowered and results should be interpreted as preliminary. Within-subject participants ($n = 4$) showed the same directional pattern on Anthropomorphism ($\Delta = +0.90$) and Animacy ($\Delta = +0.54$) but not on other subscales.

D. Discussion

The Turing test results suggest that the autonomous pipeline produces robot behaviour sufficiently naturalistic that participants cannot reliably distinguish it from teleoperation. The improvement from V2 to V3 in the teleoperated condition (25%→14% accuracy) further indicates that the addition of continuous mirroring, reduced behavioural repetition, and parameter tuning contributed to more convincing and socially present autonomous responses.

Godspeed scores were consistently higher in the Teleoperated condition, reflecting the ceiling that current autonomous non-verbal response systems face when compared to direct human-driven behaviour. Notably, Likeability scores were high in both conditions (≥ 3.96), suggesting that the robot was perceived positively regardless of control mode. The Perceived Safety dimension showed minimal difference between conditions, which may reflect the predictable movement envelope of Reachy Mini rather than control mode per se.

These results should be interpreted with caution given the small sample, mixed between- and within-subject design, and the absence of statistical power for subscale comparisons. A larger follow-up study with balanced between-subject groups and standardised interaction protocols is needed to draw inferential conclusions.

IV. Conclusion

We presented a distribution-matching evaluation framework and a three-layer nonverbal interaction pipeline for Reachy Mini, grounding autonomous behavior design in empirical interaction trajectories. Our pilot study (N = 17) provides an initial signal on policy indistinguishability and perceptual alignment across Godspeed dimensions. Future iterations will leverage discrimination accuracy and Godspeed scores as feedback signals to progressively refine the autonomous policy, closing the gap between p_{auto} and p_{teleop}.

References

[1] J. Urakami and K. Seaborn, "Nonverbal cues in human–robot interaction: A communication studies perspective," *ACM Transactions on Human-Robot Interaction*, 2023.

[2] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[4] C. Zhang *et al.*, "React to this (RTT): A nonverbal Turing test for embodied AI," 2025.

[5] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] D. McColl and G. Nejat, "Affect detection from body language during social HRI," in *2012 IEEE RO-MAN*, 2012, pp. 1013–1018.

[7] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *CVPR Workshop*, 2003, p. 53.

[8] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[9] Pollen Robotics, "Reachy mini: An open source desktop robot for human-robot interaction," 2024. [Online]. Available: https://www.pollen-robotics.com/reachy-mini/

[10] P. Rochat and C. Passos-Ferreira, *From Imitation to Reciprocation and Mutual Recognition.* Totowa, NJ: Humana Press, 2009, pp. 191–212.

[11] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9815154
[12] A. Savchenko, "Facial expression recognition with adaptive frame rate based on multiple testing correction," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 30 119–30 129. [Online]. Available: https://proceedings.mlr.press/v202/savchenko23a.html
[13] A. Yang *et al.*, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388
[14] P. Yazdian *et al.*, "MotionScript: Natural language descriptions for expressive 3D human motions," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.
[15] Meta AI, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
[16] Qdrant Team, "Qdrant," 2025. [Online]. Available: https://qdrant.tech/
[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

V. Appendix

A. Contributions

Minh:

- Implemented face emotion recognition and action detection pipeline.
- Added hand and face rotation and smoothing mechanisms for mirroring in pipeline V3.
- Implemented the motion database for pipeline V1.

Stella:

- Implemented hand gesture mapping, mirroring, head tracking, and VLM module
- Implemented and conducted MotionScript + LLM in V1 pipeline of the ablation study
- Analysis of the participant study data

Gemmin:

- Built the web-based teleop interface and conducted all teleop sessions.
- Developed Reachy infra, i.e., sdk wrapper (e.g., 3d renderer, etc).
- Formalized Turing-style study as an adversarial distribution matching problem.

B. Pilot Study Script

Between Subject Thank you for participating in our study. Today you'll be interacting with Reachy Mini, a small humanoid robot. The session will last about one minute. During the interaction, we'd like you to interact with the robot naturally. You can wave, make gestures, move your body, or show facial expressions. There are no right or wrong things to do. After the interaction, we'll ask you two things: first, whether you think the robot was operating autonomously or was controlled by a human operator; and second, to fill out a short questionnaire about your experience. Please note that we cannot tell you which condition you are in until after the study. Do you have any questions before we begin?

Within Subject Thank you for participating in our study. Today you'll be interacting with Reachy Mini twice, each time for about one minute. In each interaction, we'd like you to interact with the robot naturally: wave, gesture, move your body, or show facial expressions. There are no right or wrong things to do. After each interaction, you'll be asked whether you think the robot was autonomous or human-controlled, and you'll fill out a short questionnaire. The two interactions may feel similar or different, just respond based on what you experienced each time. The order of the two conditions is randomized. Please treat each interaction independently. Do you have any questions before we begin?

C. Teleoperator Manual

Your Role. You are the human operator controlling Reachy Mini's behaviour during the teleoperated condition. Your goal is to make the robot's responses feel as natural and socially appropriate as possible. Participants will interact freely with the robot and afterward judge whether it was autonomous or human-controlled — your job is to respond convincingly.

System Setup. The teleoperation interface runs in a web browser. Before the participant enters, confirm that: (1) the live camera stream from Reachy Mini is visible, (2) head pose control via phone orientation is responsive, and (3) the expression library panel is accessible.

Controls.

- *Head pose* — physically tilt and rotate your phone. Your phone's orientation maps directly to Reachy Mini's head pose. Tilt forward/back controls pitch; rotate left/right controls yaw.
- *Antennas* — use the two virtual joysticks on screen to control the left and right antennas independently.
- *Expressions* — tap any expression from the pre-recorded library to trigger it on the robot. Expressions are grouped by mood category.

Guidelines for Natural Behaviour.

- *Always be reacting.* Do not leave the robot still for more than 2–3 seconds. If nothing notable is happening, use subtle head movements or antenna shifts to maintain social presence.
- *Mirror before responding.* When a participant moves or gestures, briefly track them with the head before triggering an expression — this matches the rhythm of natural social response.
- *Match affect to input.* If a participant waves, respond with a cheerful expression. If they frown or look away, respond with a curious or subdued expression.
- *Avoid mechanical patterns.* Do not use the same expression repeatedly. Vary responses even to similar inputs.
- *Do not react too fast.* Introduce a natural 1–2 second pause before triggering an expression, consistent with the $\sim$3 s latency of the autonomous system.

During the Session. Each participant interaction lasts approximately one minute. You will not be visible to the

participant at any point. Do not communicate with the study administrator during the session.

After the Session. Remain at your station until the administrator confirms the participant has completed their questionnaire and left the interaction area. Do not discuss your role or the study conditions with participants before data collection is complete.