Generation of Real-time Robotic Emotional Expressions Learning from Human Demonstration in Mixed Reality

Anonymous Authors

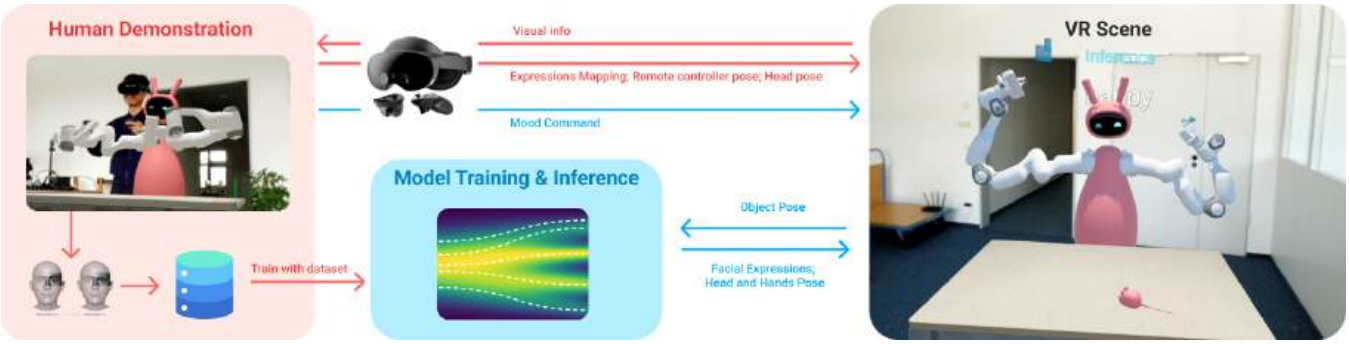

Fig. 1: Mixed-reality pipeline for learning robotic emotions. (left) An expert wearing an HMD teleoperates a virtual robot; the system records facial blend-shapes together with head and hand/controller poses, forming an affect-rich demonstration dataset. (centre) These demonstrations train a flow-matching generative model that maps a desired mood label plus live perceptual cues (the mouse pose) to continuous joint targets. (right) At inference the trained model runs at 120 Hz, taking the operator's high-level mood command and the pose of salient objects to drive the robot's eyes, ears, neck and arms with recognisable emotions inside the MR scene. Red (pink) arrows denote signals used only during data collection/training; blue arrows denote signals present at runtime.

***Abstract*— Expressive behaviors in robots are critical for effectively conveying their emotional states during interactions with humans. In this work, we present a framework that autonomously generates realistic and diverse robotic emotional expressions based on expert human demonstrations captured in Mixed Reality (MR). Our system enables experts to teleoperate a virtual or real robot from a first-person perspective, capturing their facial expressions, head movements, and upper-body gestures, and mapping these behaviors onto corresponding robotic components including eyes, ears, neck, and arms. Leveraging a flow-matching-based generative process, our model learns to produce coherent and varied behaviors in real-time in response to moving objects, conditioned explicitly on given emotional states. A preliminary test validated the effectiveness of our approach for generating autonomous expressions. Supplementary material can be found at `https://drive.google.com/drive/folders/1JAFP2CLSaGa8f1hgq4FEQApEZ6MgluqF?usp=sharing`**

I. Introduction

Expressive behaviour is a cornerstone of engaging human-robot interaction (HRI): people interpret a robot's gaze, posture and motion as cues to its internal state, and emotionally expressive robots are trusted more and even forgiven for mistakes more readily than impassive ones [1], [2], [3]. Designing such behaviour, however, is hard. Hand-crafted rule sets do not scale to the diversity of human affect, and supervised pipelines demand large, carefully annotated datasets that are expensive to gather and struggle to generalise in real time [4], [5]. Recent work has begun to replace rigid scripts with learned generative models that synthesise motion directly from high-level intent, yielding richer expressions but still relying on offline generation or heavy post-processing [6], [7]. In parallel, Mixed-Reality (MR) teleoperation has emerged as an effective way to capture nuanced human demonstrations—including subtle facial cues and upper-body gestures—without constraining the expert's viewpoint or embodiment [8]. At the modelling level, flow-matching generative processes promise fast, stable synthesis and naturally accommodate continuous control, making them attractive for real-time robots [9], [10].

We unite these strands and present a framework that autonomously produces diverse and coherent emotional expressions on a physical robot in real time. Experts teleoperate a virtual robot in MR from a first-person perspective; their facial expressions, head motions and body gestures are mapped onto robotic eyes, ears, neck and arms, providing rich demonstrations. A flow-matching generator then learns to conditioning a desired emotional label and live perceptual cues (e.g., a moving object) into continuous joint poses, yielding behaviour that observers recognise as naturally "fear", "angry", "curious", "sad", "bored" and or "happy" at 10 Hz. The contributions of this study are listed as follows:

1) MR demonstration pipeline: a device-agnostic capture method that records fine-grained affective motion from a first-person operator.
2) Flow-matching emotional generator: the first application of flow matching to real-time robotic expression, conditioning on explicit emotion labels.

II. Our Approach

Our framework is composed of two essential components: a platform for gathering training data and a real-time infer-

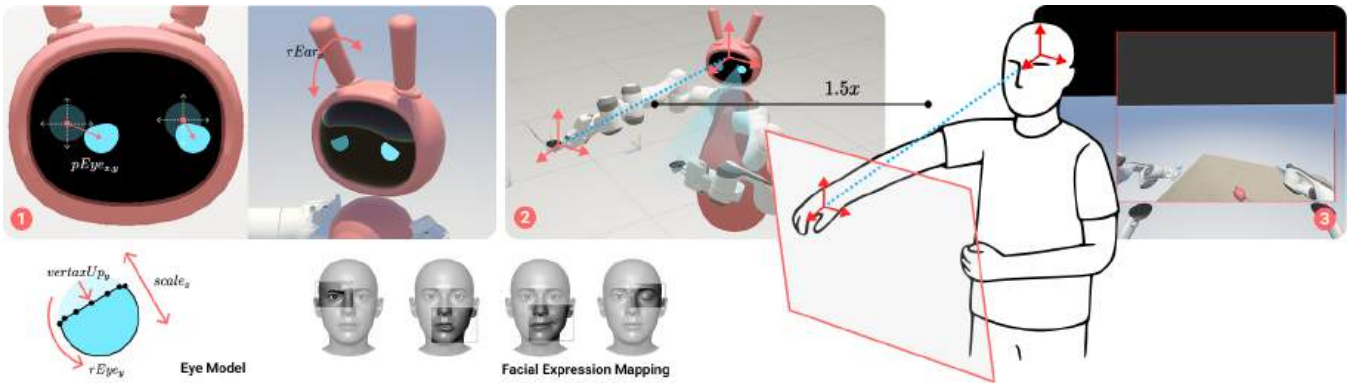

Fig. 2: The XR platform: 1. 7 facial-expression values detected by the XR-headset map the robot's ears angle and shape of the eyes, the gaze direction also maps the position of the eyes on the plane of robot face screen. Some value of the facial expression also maps the movement of the robot's ear. 2. Human's head position and orientation maps the robot's end effector, relative to the operator's head pose as the origin. The positional value is sceled by 1.5 for enhancing operator's reachablitiy. 3. There is an virtual screen floating in front of the operator, which allows the operator to observe the environment from the first person perpective.

ence system for emotional expressions.

A. Mixed–Reality Data-Collection Platform

Our data are gathered in an XR application built in *Unity* using the Meta Quest Developer SDK (v74) and deployed on a Quest Pro headset. When the app launches, a mixed-reality scene appears in which a mobile bimanual robot (base, torso, Kinova arms, head, LED eyes, and actuated ears) is placed on the real floor in front of the operator (Fig. 1, right). A table and a small *mouse* prop that scurries along random trajectories are spawned between the robot and the user. The moving mouse serves as a salient target that elicits richer, direction-specific expressions than a static cue. Because passthrough video keeps the real surroundings visible, the operator retains situational awareness while demonstrating (Fig. 1 right).

1) End effector mapping: The Quest headset provides 6-DoF poses for the user's head and for each hand-held controller. The robot base is fixed; hence we map only the *orientation* of the operator's head to the robot head, discarding translation. For the arms we adopt a relative mapping: the position of each controller expressed in the operator's head frame is scaled by 1.5 and retargeted to the corresponding robot end-effector, while orientations are matched directly. The scale factor compensates for the robot's greater reach, allowing the user to command poses that lie outside their own workspace (step 2 in Fig 2).

2) Facial-Expression Mapping: Quest Pro face tracking outputs 70 per-muscle blend-shape values in $[0, 1]$.[1] We select seven that most affect perceived emotion: *eye closedness* C_{eye}, *lip-corner dimple* D_{lip}, *brow lower* H_{brow} (all left and right), and *chin raise* H_{chin}. Gaze direction (θ_x, θ_y) is also available. The robot eye consists of two cylinders whose shapes, scales, and positions are controlled, and each ear has one rotational DoF. Let vertaxUp_y, vertaxLow_y, $r_{\text{Eye}z}$, $s_{\text{Eye}y}$, $r_{\text{Ear}x}$, and $(p_{\text{Eye}x}, p_{\text{Eye}y})$ denote those DoFs (step 1 in 2). We employ the mapping:

$$
\begin{aligned}
\text{vertaxLow}_y &= \min\big(D_{\text{lip}},\ \text{vertaxLow}_y\big),\\
\text{vertaxUp}_y &= \max\big(-\tfrac{H_{\text{chin}}+H_{\text{brow}}}{2},\ \text{vertaxUp}_y\big),\\
r_{\text{Eye}y} &= (H_{\text{chin}} + H_{\text{brow}})\,\pi/6,\\
r_{\text{Ear}x} &= \pi/2\,(-H_{\text{chin}} + H_{\text{brow}}),\\
s_{\text{Eye}z} &= 1 - 0.9\,C_{\text{eye}},\\
(p_{\text{Eye}x}, p_{\text{Eye}y}) &= \text{clamp}\big(-(\theta_x, \theta_y)/\theta_{\max},\ -1, 1\big).
\end{aligned}
$$

3) First-Person View: A virtual monitor rigidly attached to the headset shows live footage from a camera mounted on the robot's head, giving the operator a first-person preview of how each demonstrated motion will appear in situ. Pressing either controller's trigger sends the current observation–action pair to a back-end server for logging, and the screen border flashes red as confirmation (step 3 in 2).

4) WebSocket Back-End: A lightweight WebSocket server streams observations to disk at $10Hz$ during demonstration and, later, relays live observations to the trained model for closed-loop inference.

B. Real Robot Implementation

Our approach can be directly integrated into a real robotic system for data collection and teleoperation. The operator's head pose and facial expression data are transmitted to the robot using the same mapping strategy described in the previous sections.

The visual feedback is captured by a camera mounted on the robot's neck and streamed to the operator through a WebRTC-based communication pipeline. However, when deploying the system on a real robot, a critical challenge arises from the discrepancy between human head motion and robot neck dynamics. Specifically, the robot neck typically responds more slowly than human head rotations.

[1] https://developers.meta.com/horizon/documentation/unity/move-face-tracking

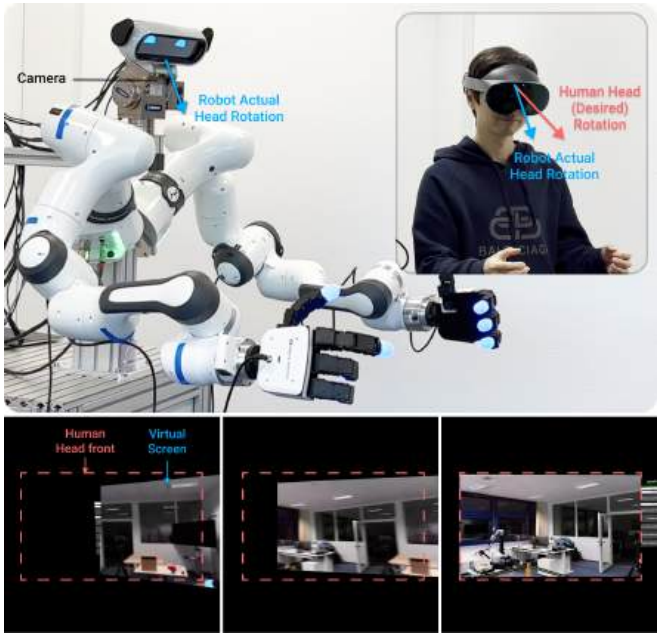

Fig. 3: **Top**: The MR headset subscribes to the actual rotation of the robot head and computes the rotation difference with respect to the human head. **Bottom**: The teleoperation view locally compensates the virtual screen orientation according to this difference, aligning the displayed video stream with the robot head pose and reducing visual mismatch to mitigate motion sickness.

If the virtual screen rigidly follows the human head pose while the robot head lags behind, the displayed visual content no longer corresponds to the robot's actual viewing direction. This inconsistency can lead to a strong visual–vestibular mismatch and significantly increase the risk of motion sickness for the operator.

To address this issue, the MR application continuously subscribes to the robot's actual head rotation and computes its difference relative to the human head orientation. Instead of fixing the virtual screen directly in front of the operator or strictly following the human head motion, the screen orientation is adjusted based on this rotation difference. As a result, the virtual screen rotates more slowly than the human head, remaining aligned with the robot's real viewing direction.

This adaptive compensation strategy not only reduces visual inconsistency and motion sickness, but also allows the operator to better perceive and anticipate the robot's physical head movements during teleoperation.

C. Flow-Matching Policy

We formulate behavioral cloning as conditional flow matching, where a neural vector field transports samples from a simple source distribution to the demonstrated action distribution. Given $\boldsymbol{x}_0 \sim p_0$, $\boldsymbol{x}_1 \sim p_1$, and $t \sim \mathcal{U}[0,1]$, we use the linear interpolation

$$\boldsymbol{x}_t = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0,$$

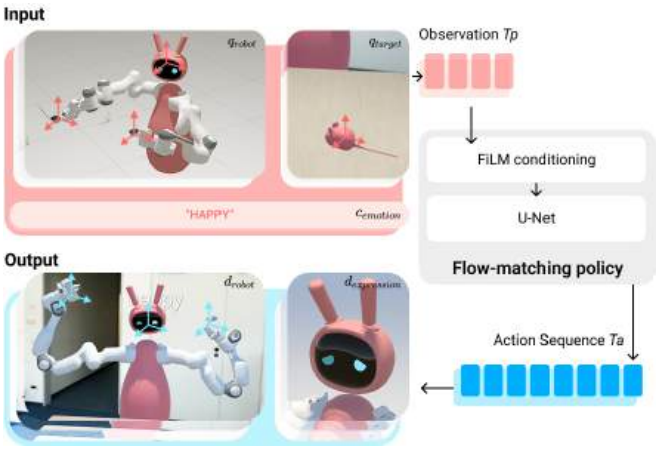

Fig. 4: Overview of flow matching for expression generation. A history window of robot and target poses plus an emotion label (pink) is fed through FiLM-conditioned U-Net to predict the blue action sequence executed on the robot.

and train the model to regress the constant target flow $\boldsymbol{x}_1 - \boldsymbol{x}_0$:

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1} \left\| v_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0) \right\|^2 . \qquad (1)$$

Here, $p_0 = \mathcal{N}(0, I)$ is the base action distribution and p_1 is the demonstration distribution, following the linear conditional flow formulation [11].

To condition the policy on the current scene, we extend the vector field as

$$\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{o}) = v_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t|\boldsymbol{o}),$$

where the observation $\boldsymbol{o}$ concatenates a one-hot emotion label c_{emotion}, the robot head and end-effector poses q_{robot}, and the target pose q_{target}. The vector field is parameterized by a FiLM-conditioned U-Net [12], [13]. The generated action $\boldsymbol{x}$ contains robot head and end-effector targets d_{robot} as well as facial-expression commands $d_{\text{expression}}$.

At inference, we sample $\boldsymbol{x}_0 \sim \mathcal{N}(0, I)$ and integrate the learned flow from $t = 0$ to $t = 1$:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \Delta t\, v_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t|\boldsymbol{o}), \qquad t \in [0, 1]. \qquad (2)$$

The policy predicts T_p future action steps, and the robot executes the first T_a steps before replanning, enabling closed-loop expression generation.

We collected demonstrations for seven emotions (*happy, sad, angry, fear, bored, curious, calm*) toward the moving mouse target (Fig. 5). Each clip contains approximately 10,000 frames at 10 Hz, including robot poses, end-effector trajectories, gaze direction, and eye/ear motions, providing paired full-body and facial-expression supervision for the flow model.

III. Preliminary Results and Future Directions

Training protocol. The flow model was trained for 3000 epochs with a batch size of 256 and a learning rate of 1×10^{-4}. We explored four history-window lengths (1, 2, 4, and

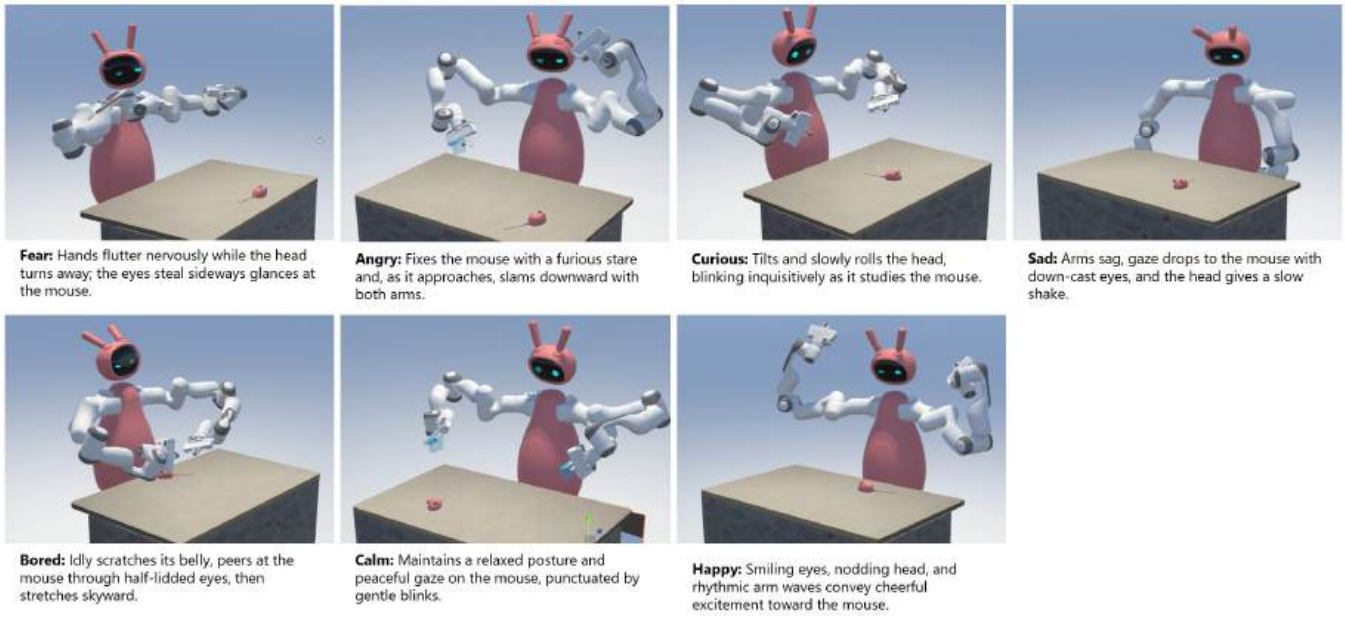

Fig. 5: Generated emotional expressions directed toward the moving target object.

16 frames) in combination with two prediction horizons (16 and 32 frames).

Expert appraisal. A panel of HRI researchers informally inspected roll-outs and provided qualitative feedback:

1) *Temporal context.* A 16-frame history performed noticeably worse than 2–4 frames, suggesting that our FiLM-conditioned U-Net does not fully exploit long temporal correlations. Replacing FiLM with a transformer-based temporal encoder may improve sequence understanding at the cost of heavier training.
2) *Prediction horizon.* Longer horizons (32 frames) produced more complete, fluid gestures, whereas short horizons introduced occasional "jumps" when the policy re-planned. This points to a weak internal notion of phase; additional data or an explicit timing signal could reduce discontinuities.
3) *Emotion coverage.* Six of the seven emotions transferred convincingly; the *curious* behaviour lacked the distinctive "poke" motion present in the demonstrations. We attribute this to data sparsity and will extend the dataset with targeted examples.

Next steps. We will (i) integrate a transformer backbone for richer temporal reasoning, (ii) expand the training corpus to balance under-represented actions, and (iii) conduct a controlled user study to quantify recognisability, naturalness, and preference compared with teleoperation baselines.

REFERENCES

[1] C. Breazeal, "Role of expressive behaviour for robots that learn from people," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3527–3538, 2009.

[2] M. Bretan, G. Hoffman, and G. Weinberg, "Emotionally expressive dynamic physical behaviors in robots," *International Journal of Human-Computer Studies*, vol. 78, pp. 1–16, 2015.

[3] X. Liu, J. Dong, and M. Jeon, "Robots'"woohoo" and "argh" can enhance users' emotional and social perceptions: An exploratory study on non-lexical vocalizations and non-linguistic sounds," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 4, pp. 1–20, 2023.

[4] K. Mahadevan, J. Chien, N. Brown *et al.*, "Generative expressive robot behaviors using large language models," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 482–491.

[5] R. Stock-Homburg, "Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research," *International Journal of Social Robotics*, vol. 14, no. 2, pp. 389–411, 2022.

[6] D. Zhang, J. Peng, Y. Jiao *et al.*, "Exface: Expressive facial control for humanoid robots with diffusion transformers and bootstrap training," *arXiv preprint arXiv:2504.14477*, 2025.

[7] Y. Hu, P. Huang, M. Sivapurapu *et al.*, "Elegnt: Expressive and functional movement design for non-anthropomorphic robot," 2025. [Online]. Available: https://arxiv.org/abs/2501.12493

[8] D. Black, M. Nogami, and S. Salcudean, "Mixed reality human teleoperation with device-agnostic remote ultrasound: Communication and user interaction," *Computers & Graphics*, vol. 118, pp. 184–193, 2024.

[9] Y. Lipman, R. T. Chen, H. Ben-Hamu *et al.*, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[10] F. Zhang and M. Gienger, "Affordance-based robot manipulation with flow matching," *arXiv preprint arXiv:2409.01083*, 2024.

[11] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[13] E. Perez, F. Strub, H. De Vries *et al.*, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.