Integrated Conditional Flow Matching for Automatic Multimodal Generation of Emotional Expressions for a Social Robot

José Manuel Pérez-Molerón[1], Noe Pérez-Higueras[1], Fernando Caballero[1], Randy Gomez[2], Luis Merino[1]

***Abstract*— This paper introduces a novel generative framework for synthesizing expressive robotic routines for the Haru social robot. Our approach leverages a unified Spatio-Temporal Convolutional Neural Network (CNN) trained with Flow Matching to concurrently orchestrate motor trajectories (positions and velocities), mouth LED patterns, and eye video expressions. Unlike previous disjointed models, this unified architecture enables the simultaneous generation of all modalities, conditioned on a continuous emotion space derived from Plutchik's wheel of emotions. By formulating emotion generation as a probability path that bridges Gaussian noise with the data distribution, the model facilitates the production of smooth, diverse, and multi-modally coherent robotic behaviors, avoiding preprogrammed or hand-animated behaviors that can be repetitive and predictable.**

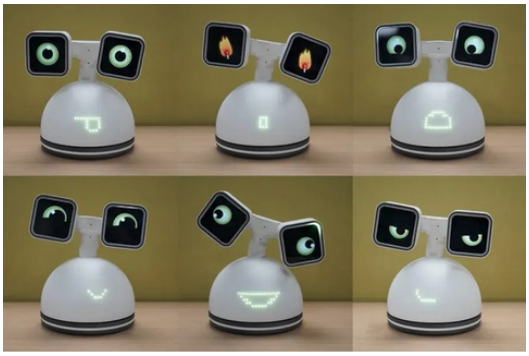

Fig. 1. Haru the robot showing different pre-defined emotional expressions.

I. Introduction

Non-verbal communication is vital for effective human-robot interaction (HRI), improving trust and long-term engagement in social robots like the tabletop assistant Haru [1], [2], [3]. However, traditional expressive behaviors rely on pre-programmed, hand-crafted routines [4], [5], which suffer from a limiting "novelty effect" due to repetitive motions and scale poorly to the complexity of human affect [6], [7], [8].

To overcome these limitations, research has shifted toward deep generative models for affective movement [6], [9]. Diffusion Policies have been proposed to generate complex motion behaviors. However, their iterative denoising causes high latency and jerky, non-smooth trajectories [10], [11], [12]. Conversely, Flow Matching (FM) maps noise distributions directly to target trajectories via continuous vector fields, ensuring computationally efficient, physically smooth, and temporally coherent motions [13], [11], [12].

In this paper, we introduce a multi-modal generative framework that applies Conditional Flow Matching to synthesize real-time emotional routines for Haru. Our approach generates motor trajectories (positions and velocities) with synchronized LED mouth patterns and low resolution videos for the eyes expressions, conditioned on a continuous emotion space derived from Plutchik's wheel [7]. Kinematic evaluations demonstrate that our model produces diverse, smooth and coherent behaviors efficiently.

II. Related Work

The generation of expressive robotic motion has evolved from rigid, pre-programmed routines to dynamic, data-driven frameworks, with recent focus shifting toward deep generative models and Flow Matching.

A. Traditional and LfD Approaches for Affective Motion

Early work in expressive robotics relied on hand-animated routines or interpolation between expert demonstrations. For the tabletop robot Haru, Mier et al. [7] explored Task-Parameterized Gaussian Mixture Models (TPGMM) to interpolate animations based on Plutchik's wheel of emotions. Similarly, Trovato et al. [14] employed polynomial classifiers to select combinations of predefined facial configurations for the KOBIAN-R robot. Other approaches utilized Inverse Reinforcement Learning (IRL) to maintain specific emotive styles during functional tasks [15]. However, these models generally lack the capacity to generate continuous, high-frequency variations and struggle to generalize without extensive data. Our system overcomes these limits by generating continuous and diverse multi-modal trajectories.

B. Deep Generative Models and Latent Spaces

To address the constraints of discrete emotion classes, researchers proposed mapping trajectories to continuous spaces. Sripathy et al. [16] introduced a model mapping motions to a continuous Valence-Arousal-Dominance (VAD) space, while the MoveAE framework utilized Variational Autoencoders (VAEs) to manipulate affective features via latent space arithmetic [6]. Recently, Large Language Models (LLMs) have also been employed, such as the EMOTION

[1]Universidad Pablo de Olavide, Avenida Rectora Rosario Valpuesta, 1. Seville, Spain

[2]Honda Research Institute Japan Co., Ltd., Japan

[3]https://www.rwapsych.com.au/blog/emotional-regulation-Wemotions-as-indicators-and-not-dictators/

This work has been partially supported by the grants COBUILD (PID2024-161069OB-C31), and PICRAH 4.0 (PLEC2023-010353), both funded by the Ministry of Science, Innovation and Universities, the Spanish Research Agency, and the European Regional Development Fund, MICIU /AEI /10.13039/501100011033 / FEDER, UE)).

framework, which uses natural language feedback to refine gesture sequences [17].

However, these approaches face notable challenges. VAE-based models often struggle with training instability and fail to capture nuanced high-frequency variations [11], [7]. By focusing primarily on first-order dynamics, many of these models produce jerky trajectories lacking smooth acceleration [11]. Additionally, LLM-driven frameworks suffer from high computational latency and lack workspace awareness, potentially leading to kinematically infeasible or colliding motions [17].

C. Diffusion and Flow Matching in Robotics

The current state of the art emphasizes diffusion and flow-based generative models. Diffusion Policies capture complex multi-modal action probability distributions effectively [10] but suffer from high inference latency and often produce jerky motions [11]. Flow Matching (FM) [13] maps noise directly to target trajectories, emerging as a faster, smoother alternative. Extensions like FlowMP capture second-order dynamics for physical feasibility [11], while CoLA-Flow and LG-Flow perform FM in latent spaces to improve temporal coherence [12], [18]. In emotional expression, Wang et al. [8] successfully applied FM to generate real-time expressions from human demonstrations. Nevertheless, existing diffusion and flow-based applications in robotics focus on single output modalities (such as joint kinematics or isolated facial gestures). Our work builds on these foundations, introducing a unified Flow Matching framework that jointly generates motor trajectories, mouth LED patterns, and eye video displays within a single probability flow field, ensuring tight temporal synchronization and real-time generation.

III. Dataset

A. Robotic Platform and Dataset

The target platform is Haru, a tabletop social robot with 7 degrees of freedom (DOFs), two LCD screens for eye animations, and an LED matrix for mouth expressions [3]. We curated an initial dataset of 21 pre-programmed routines (sets of ordered in time robot commands, including motion trajectories, eyes and LEDs for the mouth, defining the necessary input in order to express a specific emotion) categorized into four emotions: *Happiness*, *Sadness*, *Anger*, and *Boredom*. Each routine is mapped to continuous Cartesian coordinates (p_x, p_y) on Plutchik's wheel of emotions.

The recorded states for each routine include kinematic data (motor positions and velocities of the base, neck pitch, neck roll, and eye roll) and visual states (mouth LED patterns and eye video latents, see Fig. 1). Data were collected at 100 Hz and linearly interpolated to a fixed length of 512 timesteps per routine ($\sim$5 seconds).

B. Data Augmentation

To improve model generalization and robustness, we synthetically augmented the original dataset to generate multiple variants per routine. The augmentation process only mutates the motor trajectories (positions and velocities) while keeping the mouth LEDs and eye latents unmodified. The pipeline introduces three main sources of variability that are applied simultaneously:

- **Plutchik Jitter**: Small Gaussian noise ($\mathcal{N}(0, \sigma)$), where $\sigma = 0.05$ after experiments results to add small shifts to the emotional coordinates. Variations in the magnitude of this vector compute an intensity scale that dynamically alters the routine's overall amplitude.
- **Temporal Warping**: Smooth temporal drifts are applied using 1D Gaussian filtering to alter the pacing of the movement.
- **Centered Amplitude Scaling and Noise**: We apply a scaling factor to the trajectory relative to a neutral motor position (0.5, as the motor values are normalized to a $[0.0, 1.0]$ range). Finally, smoothed Gaussian noise is injected, modulated by a temporal windowing mask that prevents sudden jerks at the start and end of the motion.

IV. Methodology

We propose a novel multi-modal generative framework that treats robotic routine synthesis as a unified conditional flow generation problem. Unlike previous iterations[19] that employed disjointed models for different modalities, the new architecture unifies the generation of motor trajectories (positions and velocities), mouth LED patterns, and eye video expressions under a single Spatio-Temporal Convolutional Neural Network (CNN) trained with the Flow Matching (FM) methodology [13].

A. Unified Representation and Modality Encoding

To correctly account for the complex interplay and synchronization between different expressive modalities, all information is encoded into a unified latent space.

For the eye video, we first train a spatial autoencoder to reduce the high dimensionality of the video frames into a compact latent representation.

This eye latent representation, along with the motor positions, velocities, and mouth LED sequences, are projected and concatenated into a unified 512-dimensional vector per timestep. This comprehensive 512-dimensional representation ensures that changes in one modality (e.g., a sudden neck movement) are intrinsically linked to corresponding changes in the others (e.g., an eye blink or a change in mouth shape).

B. Unified Flow Matching Architecture

To achieve fluid and expressive multi-modal motion for the social robot Haru, the unified 512-dimensional vectors are processed by a single Spatio-Temporal CNN model through the Flow Matching paradigm. The objective is to learn a time-dependent vector field $\Psi_{\text{unified}}(x_t, t, c)$ that generates a probability path $p_t(x)$ between a base distribution $p_0(x) = p_{noise}$ (Gaussian noise) and the target distribution $p_1(x)$ of complete emotional routines. Here, x_t represents the unified multi-modal state (the 512-dimensional vector) at the current flow step, $t \in [0, 1]$ denotes the continuous integration time,

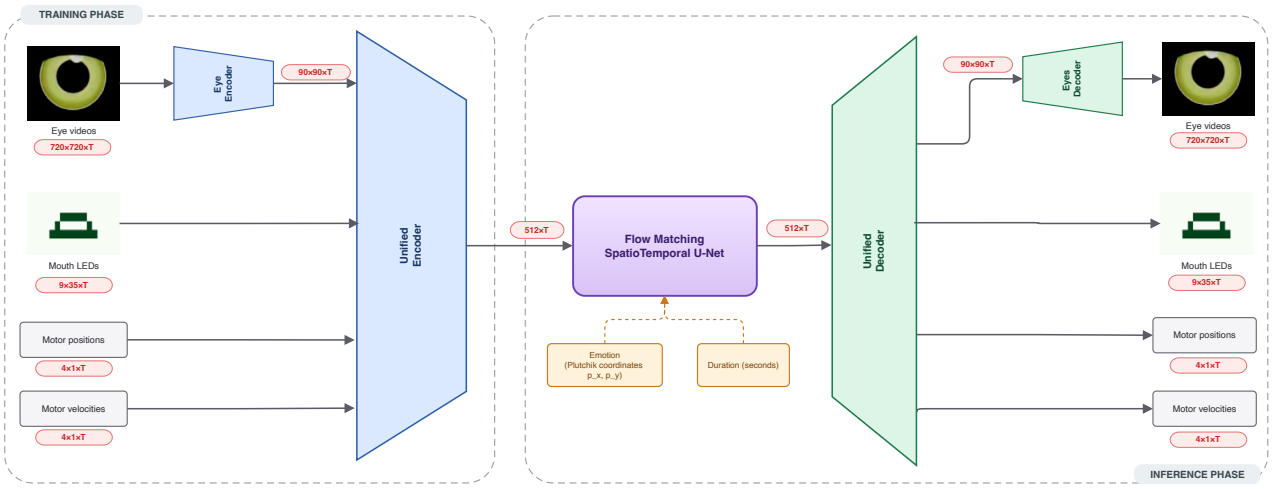

Fig. 2. System architecture diagram. During training, high-dimensional eye videos, mouth LEDs, and motor states are compressed into a unified $512 \times T$ representation. The Spatio-Temporal U-Net learns the vector field conditioned on Plutchik coordinates and duration. During inference, the flow generates the $512 \times T$ latent vector, which is decoded back to reconstruct the full multi-modal routine.

and c is the emotional conditioning vector used to guide the generation towards a specific emotion.

This unified operator Ψ_{unified} directly regresses the specific flow for all modalities simultaneously, discarding the need for separate models or isolated Feature-wise Linear Modulation (FiLM) networks for the LED patterns.

C. Network Architecture and Emotional Conditioning

The network parameterizing the vector field Ψ_{unified} is designed as a Spatio-Temporal U-Net. The input latent representation, reshaped into a spatial format of dimensions $[C, T, H, W]$ (where C denotes the number of feature channels, T is the temporal sequence length, and H and W represent the spatial height and width), is processed through an encoder-decoder structure consisting of multiple *SpatialTemporalBlocks*.

Each *SpatialTemporalBlock* sequentially applies spatial and temporal processing: first, a 2D spatial convolution (kernel size 3) captures spatial patterns across the latent dimensions, followed by a 1D temporal convolution (kernel size 7, dilation 2) to learn long-range temporal dependencies across the trajectory sequence. At the lowest resolution (bottleneck) of the U-Net, we introduce a Spatio-Temporal Self-Attention mechanism that independently performs spatial attention over the latent pixels and temporal attention across frames, significantly enhancing the global coherence of the generated routines.

Emotional and temporal conditioning is injected at every block. The emotional coordinates (x_{pl}, y_{pl}) and the integration time step t are encoded using Sinusoidal Embeddings and processed by an MLP to form the conditioning vector c:

$$c = \text{MLP}_{cond}(\text{Emb}(t) \oplus \text{Emb}(x_{pl}) \oplus \text{Emb}(y_{pl})) \quad (1)$$

Instead of a simple linear projection bias, the conditioning vector c dynamically modulates both the spatial and temporal feature maps within each *SpatialTemporalBlock* via affine transformations (scale and shift).

This mechanism allows the network to adapt the velocity and amplitude of the multi-modal motion—motors, mouth LEDs, and eye latents (see Fig. 4) precisely matching the target emotion. Finally, an output layer maps the features back to the required dimensional space for trajectory integration via an ODE solver. Propagating the initial noise state through the unified motion field over the flow-time horizon can be formulated using the 4th-Order Runge-Kutta method (Eq. 2) for high precision.

$$\begin{aligned} x_{t+1} &= x_t + \frac{\delta}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= \Psi_{\text{unified}}(x_t, t, c) \\ k_2 &= \Psi_{\text{unified}}(x_t + \frac{\delta}{2}k_1, t + \frac{\delta}{2}, c) \\ k_3 &= \Psi_{\text{unified}}(x_t + \frac{\delta}{2}k_2, t + \frac{\delta}{2}, c) \\ k_4 &= \Psi_{\text{unified}}(x_t + \delta k_3, t + \delta, c) \end{aligned} \quad (2)$$

where Ψ_{unified} represents our trained multi-modal model, t is the timestamp from 0 to 1 and δ is the time step.

V. Experimental Results and Discussion

In this section, we evaluate the generative framework through quantitative analysis of the generated routines and qualitative human perceptual studies.

A. Quantitative Analysis

Our primary quantitative validation compares the statistical distribution of motor trajectories against the expert

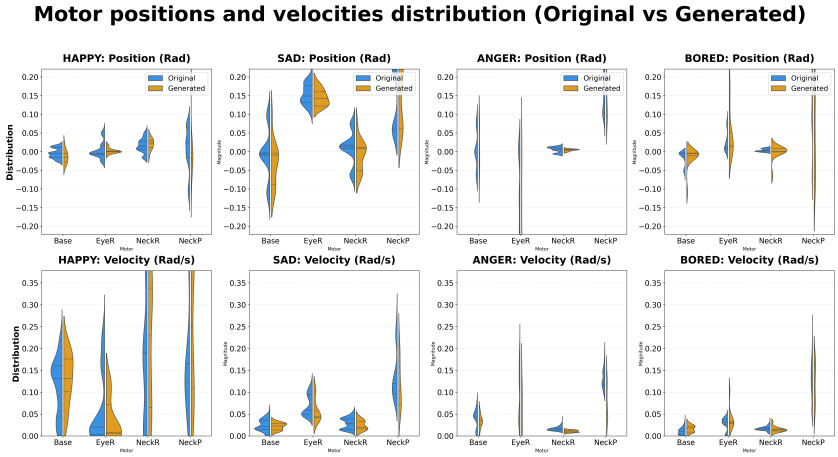

Fig. 3. Comparison of motor position ([Rad]) and velocity ([Rad/s]) distributions using the new unified Spatio-Temporal U-Net. The split-violin plots present a direct comparison between the expert ground truth (left-half) and the generated routines (right-half), demonstrating a highly accurate adjustment of the generated data.

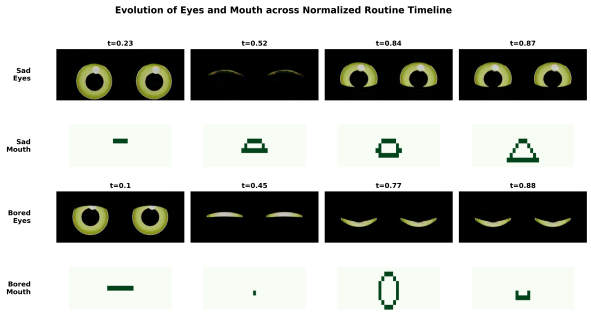

Fig. 4. Temporal evolution of the generated visual features (eyes and mouth) across a normalized timeline. The unified Flow Matching architecture synthesizes complex, coherent multi-modal expressions. The figure illustrates how the *Sad* and *Bored* emotional states evolve during the timeline, maintaining clear emotional distinguishability.

ground truth, and assesses emotional classification accuracy using the generated modalities.

First, the kinematic fidelity of the generated routines is evaluated. As depicted in the split-violin plot (Fig. 3), our new unified Flow Matching system demonstrates exceptional kinematic fidelity. The generated motor velocity and position profiles fit the ground truth distributions remarkably well, showing a notable improvement in data adjustment compared to previous versions of the architecture. For instance, routines labeled as *Sadness* correctly exhibit low-velocity distributions centered on the Neck pitch, whereas *Happy* routines exhibit significantly higher velocity distributions.

Furthermore, we employ automated classifiers to verify if the generated motions and visual features are correctly mapped to their intended categories. Processing raw data involves handling high-dimensional sequences (e.g. 165,000 dimensions per routine). We apply Principal Component Analysis (PCA) to reduce dimensionality while preserving variance, followed by Linear Discriminant Analysis (LDA) to maximize the distance among emotions.

As summarized in Tables I and II, both LDA and Support Vector Machine (SVM) models—trained exclusively on the ground-truth dataset—achieved over 91% accuracy when categorizing the generated trajectories from our initial iteration. This demonstrates that the synthesized routines are statistically aligned with the original emotional spaces.

TABLE I

CLASSIFIERS PERFORMANCE METRICS PER METHOD

Method	Accuracy	Precision	Recall	F1-score
LDA	91.2%	93.8%	90.6%	91.2%
SVM	91.2%	91.7%	91.0%	91.1%

TABLE II

CLASSIFICATION ACCURACY PER CATEGORY

Emotion	LDA	SVM
Angry	87.5%	87.5%
Bored	100.0%	88.9%
Happy	75.5%	87.5%
Sad	100.0%	100.0%

VI. CONCLUSIONS

This paper presents a unified generative framework for synthesizing emotional routines for the Haru social robot. By employing a Spatio-Temporal CNN trained with Flow Matching, our model concurrently generates coherent motor trajectories, mouth LEDs, and eye videos. Previous quantitative analyses demonstrate a high degree of fidelity between the generated data and the motor kinematics (positions and velocities) of the original routines. Although new perceptual surveys for this unified model have not been done yet, we expect emotion recognition to improve significantly. This anticipation relies on the refined movements and LED patterns, as well as the crucial incorporation of eye expressions, which we previously identified as essential for the users. Future work will incorporate Reinforcement Learning from Human Feedback (RLHF) to further enhance emotional clarity and diversity.

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] 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.

[3] R. Gomez, D. Szapiro, K. Galindo, and K. Nakamura, "Haru: Hardware design of an experimental tabletop robot assistant," in *HRI*. ACM, 2018.

[4] R. Gomez, D. Szapiro, L. Merino, and K. Nakamura, "A holistic approach in designing tabletop robot's expressivity," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1970–1976.

[5] J. Gray, G. Hoffman, S. O. Adalgeirsson, M. Berlin, and C. Breazeal, "Expressive, interactive robots: Tools, techniques, and insights based on collaborations," in *Proc. HRI 2010 Workshop: What Do Collaborations Arts Say About HRI*, 2010, pp. 21–28.

[6] M. Suguitan, R. Gomez, and G. Hoffman, "Moveae: Modifying affective robot movements using classifying variational autoencoders," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 481–489.

[7] G. Mier, F. Caballero, K. Nakamura, L. Merino, and R. Gomez, "Generation of expressive motions for a tabletop robot interpolating from hand-made animations," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 213–220.

[8] C. Wang, M. Gienger, and F. Zhang, "Generation of real-time robotic emotional expressions learning from human demonstration in mixed reality," *arXiv preprint arXiv:2508.08999*, 2025.

[9] I. Rodriguez, J. M. Martínez-Otzeta, I. Irigoien, and E. Lazkano, "Spontaneous talking gestures using generative adversarial networks," *Robotics and Autonomous Systems*, vol. 114, pp. 57–65, 2019.

[10] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. C. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[11] K. Nguyen, A. T. Le, T. Pham, M. Huber, J. Peters, and M. N. Vu, "Flowmp: Learning motion fields for robot planning with conditional flow matching," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[12] W. Songwei, J. Zhiduo, X. Guanghu, S. Wandong, L. Hong, and L. Yang, "Cola-flow policy: Temporally coherent imitation learning via continuous latent action flow matching for robotic manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2601.23087

[13] Y. Lipman, R. T. Chen, H. Lamba, P. Esser, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] G. Trovato, T. Kishi, N. Endo, K. Hashimoto, and A. Takanishi, "Development of facial expressions generator for emotion expressive humanoid robot," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, 2012, pp. 303–308.

[15] A. Zhou and A. D. Dragan, "Cost functions for robot motion style," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3632–3639, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52156435

[16] A. Sripathy, A. Bobu, Z. Li, K. Sreenath, D. S. Brown, and A. D. Dragan, "Teaching robots to span the space of functional expressive motion," *arXiv preprint arXiv:2203.02091*, 2022.

[17] P. Huang, Y. Hu, and J. Zhang, "Emotion: Expressive motion sequence generation for humanoid robots with in-context learning," *IEEE Robotics and Automation Letters (RA-L)*, 2025.

[18] S. et al., "Lg-flow policy: Trajectory-level action generation framework for robotic manipulation," 2026.

[19] J. M. Perez-Moleron, N. Perez-Higueras, R. Gomez, and L. Merino, "Automatic generation of emotional expressions for the social robot haru with conditional flow matching," in *Proceedings of the 2026 IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2026, in press.