

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

José Manuel Pérez-Molerón¹ Noe Pérez-Higueras¹ Fernando Caballero¹ Randy Gomez² Luis Merino¹
¹Universidad Pablo de Olavide, Seville, Spain ²Honda Research Institute, Japan jmpermol@upo.es

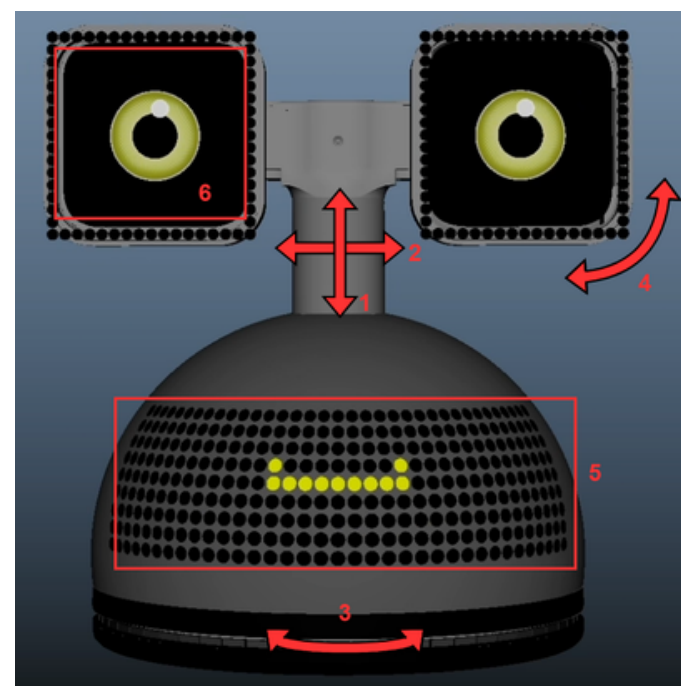
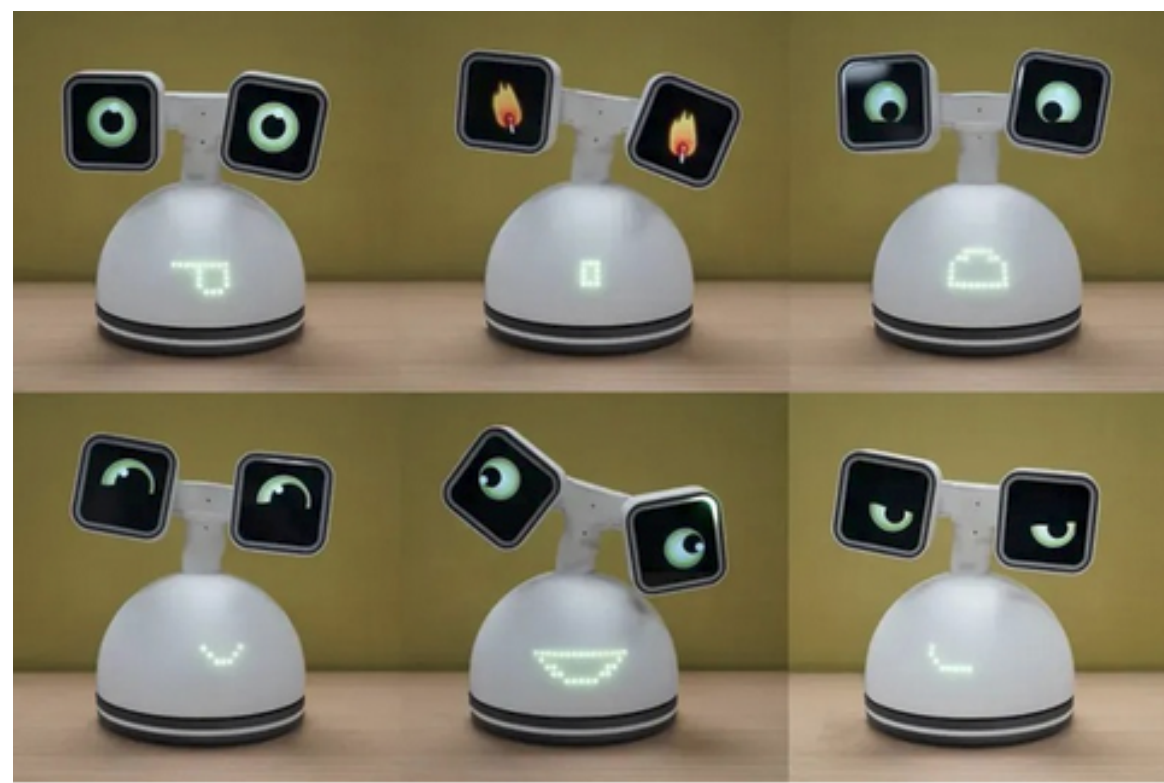
Introduction

Objetives

- Create a generative framework to synthesize expressive robotic routines for the social robot Haru.
- Create a generative framework to synthesize expressive robotic routines for the social robot Haru.

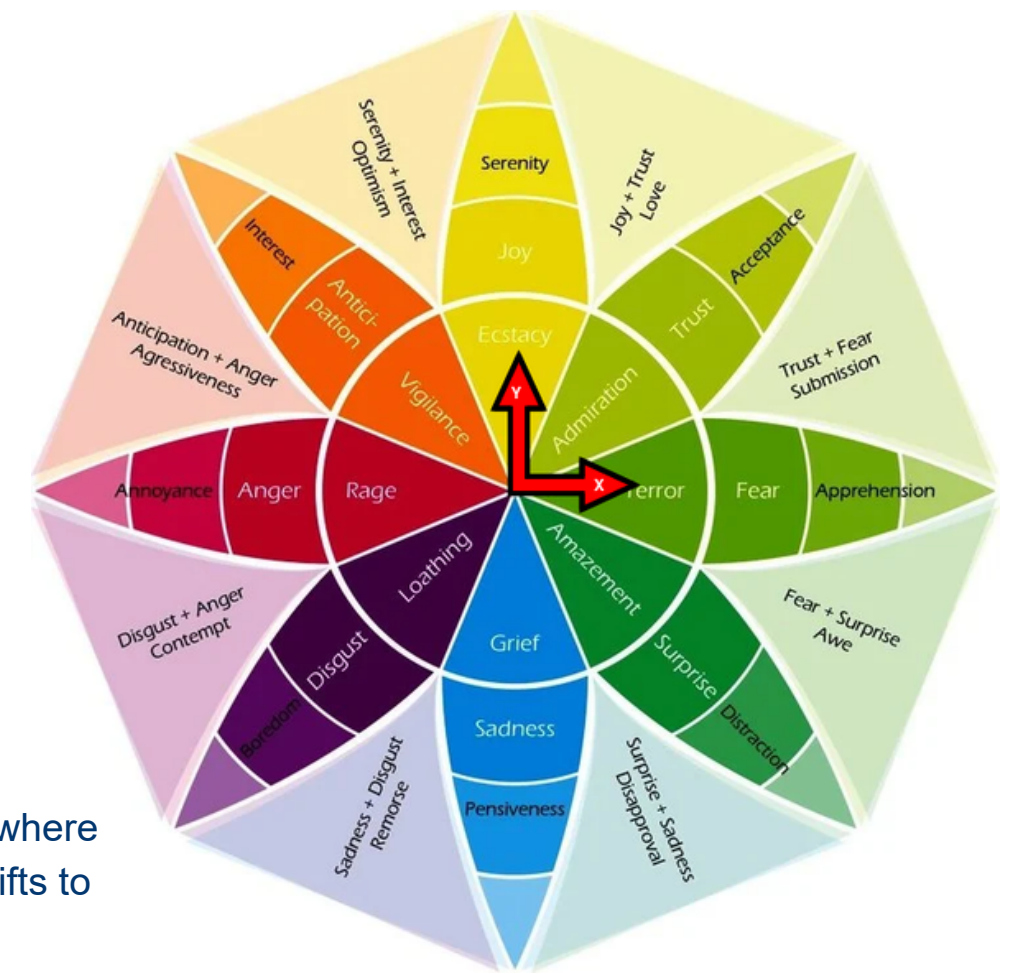
Generated outputs

- Motor positions and velocities
 - Neck pitch and roll (1 & 2)
 - Base Yaw (3)
 - Eye roll (4)
- LED panel patterns (5)
- Eye videos (6)



Emotional encoding and data augmentation

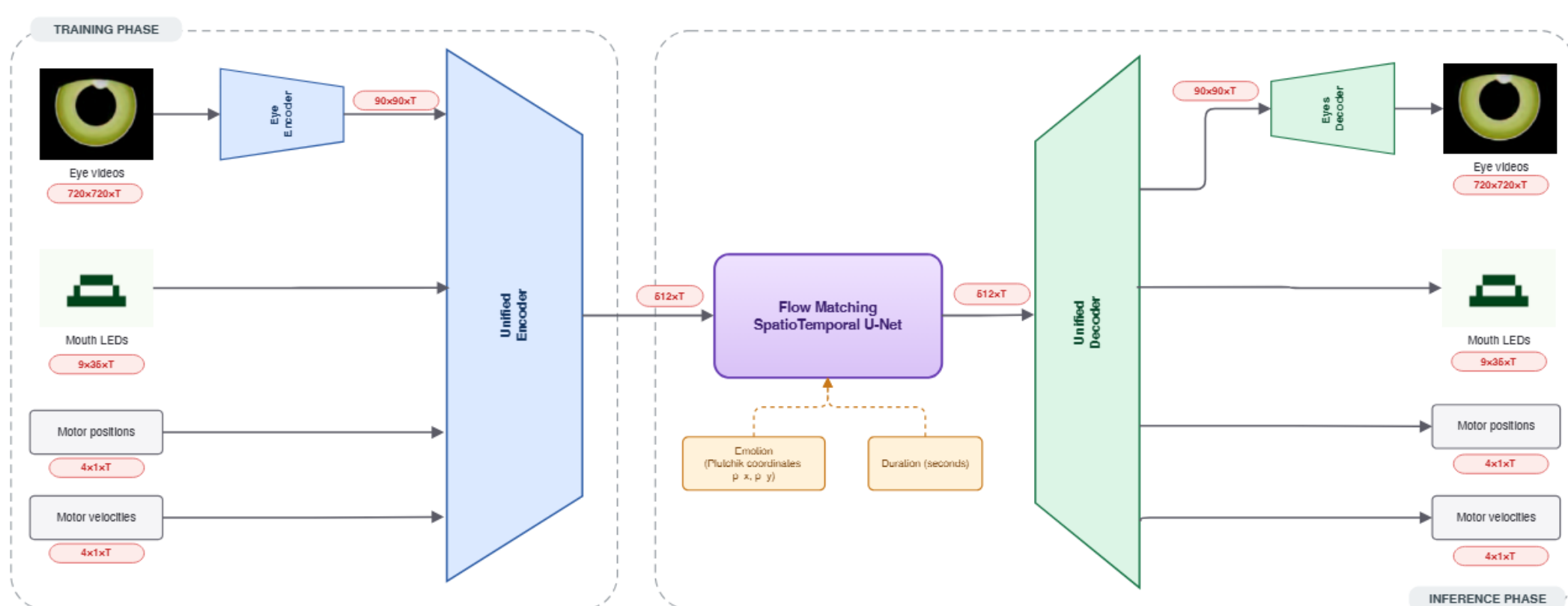
- Each emotional routine is mapped to continuous Cartesian coordinates (px, py) on Plutchik's wheel of emotions.
- The recorded states for each routine include kinematic data (motor positions and velocities of the base, neck pitch, neck roll and eye roll) and visual states (mouth LED patterns and eye video latents, linearly interpolated to a fixed length of 512 timesteps per routine)



- we synthetically augmented the original dataset to generate multiple variants per routine using the following methods:
 - **Plutchik Jitter**: Small Gaussian noise ($N(0, \sigma)$), where $\sigma = 0.05$ after experiments results to add small shifts to the emotional coordinates
 - **Temporal Warping**: Smooth temporal drifts are applied using 1D Gaussian filtering to alter the pacing of the movement.
 - **Centered Amplitude Scaling and Noise**: applied an scaling factor to the trajectory relative to a neutral motor injecting smoothed Gaussian noise

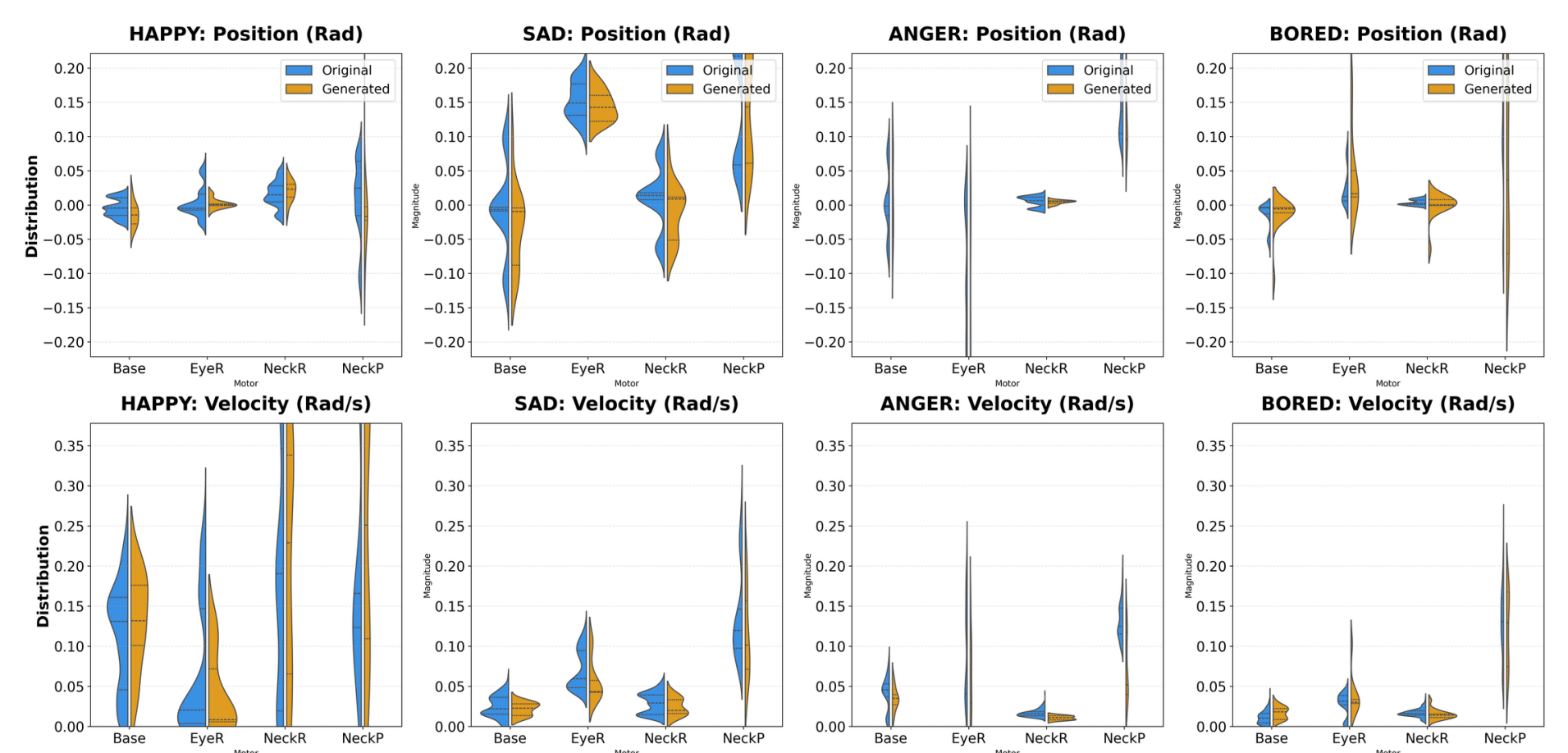
System Overview

- The system architecture consists of the following key components:
 - **Eye encoder/decoder**: autoencoder to compress the information from the eye videos into a smaller representation
 - **Unified Autoencoder**: the main autoencoder to compress all information from different modalities into a single representation of 512 values for each time step.
 - **Flow Matching Spatio Temporal U-Net**: The main model is an SpatioTemporal U-Net representing the vector field of the Flow Matching paradigm.



Motor trajectories evaluation

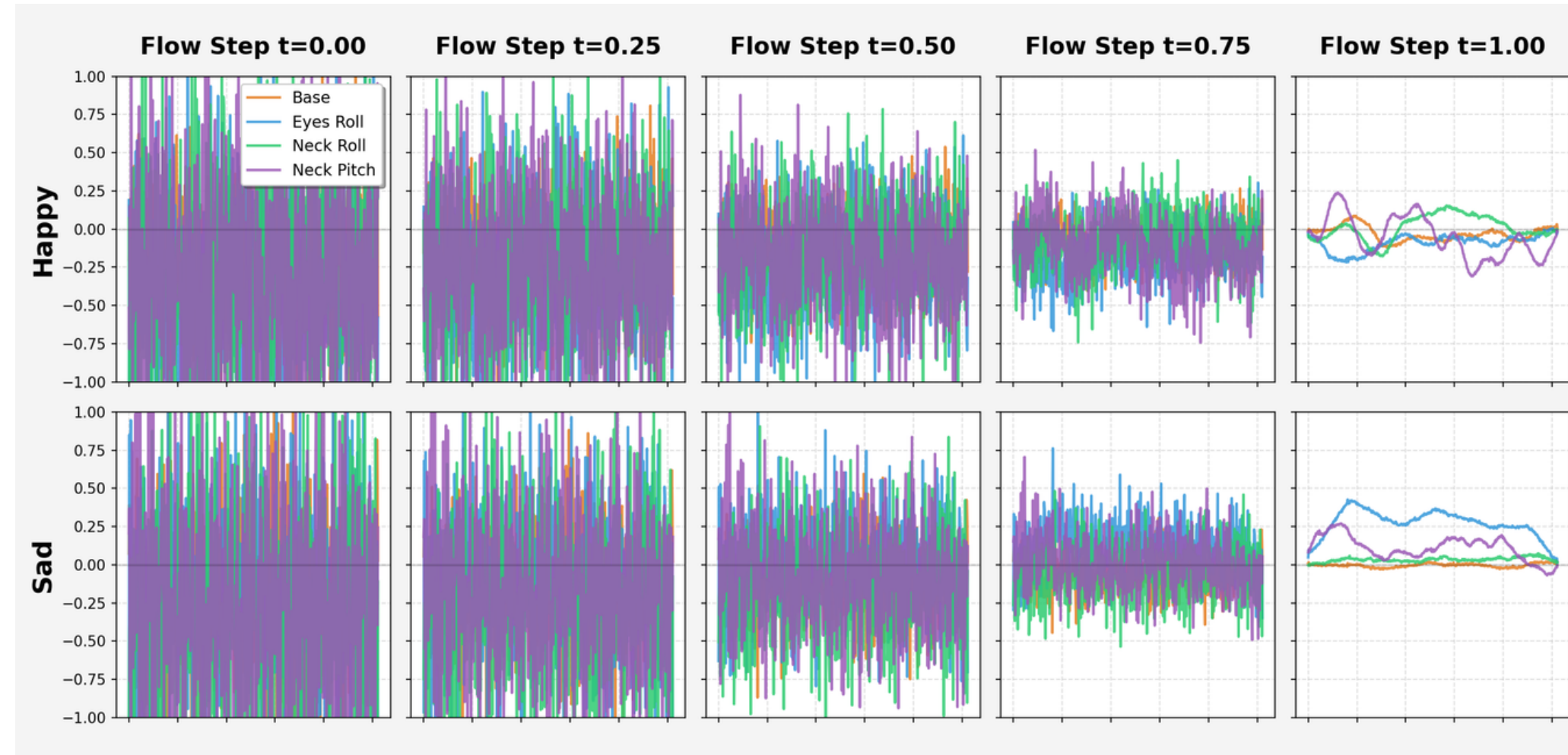
Motor positions and velocities distribution



Note: The orange distributions represent those generated by our model, compared to the original distributions shown in blue

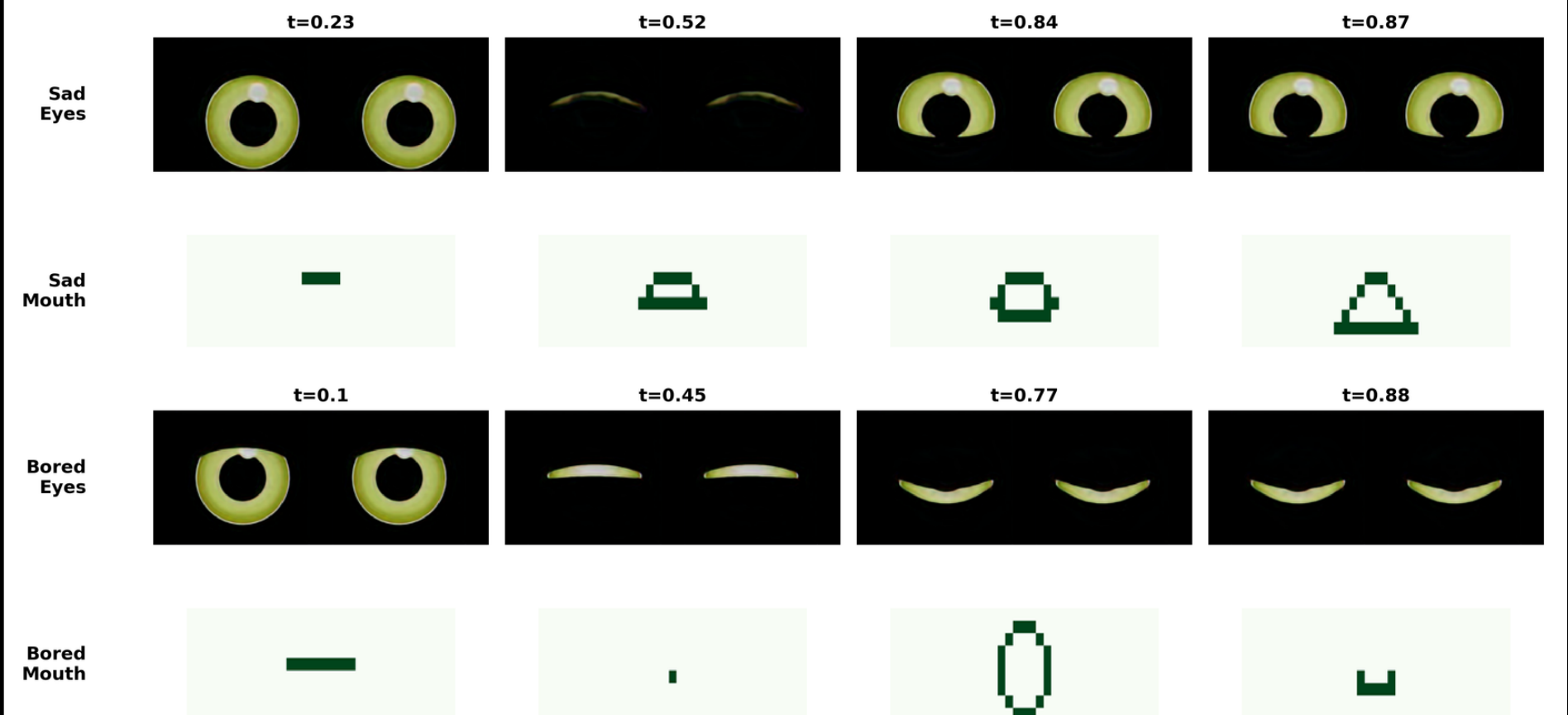
Trajectories generation by different modalities

Motor positions and velocities evolution



► Note: The image above shows the evolution of motor trajectories during the Flow Matching generation process from pure Gaussian noise (t=0.0) to the final trajectory conditioned by the emotion (t=1.0).

Visual patterns evolution



- The image above shows how mouth LEDs pattern and eyes change over time during the execution of the emotional routine.
- Visual modalities evolution shows how the entire face change over the time depending on the emotion generated by our model

Conclusions & next steps

- **Unified Generative Architecture**: Successfully synthesizes synchronized motor trajectories, mouth LEDs, and eye video displays in a single Flow Matching framework
- **Kinematic Continuity & Emotion Alignment**: Produces smooth, non-jerky motions that maintain using Plutchik's continuous emotional space maintaining high accuracy in positions and velocities for each motor distinguished by emotions.
- **RLHF Integration**: Future work will incorporate Reinforcement Learning from Human Feedback (RLHF) to refine emotional clarity and expression diversity
- **HRI Field Trials**: User perceptual studies on the physical Haru platform will evaluate real-time interaction quality and perceived emotional expressiveness.

Acknowledgements

This work was partially supported by the projects AI-FUSE (SAIA202500X163851SV0) and COBUILD (PID2024-161069OB-C31), funded by the Spanish Research Agency and the Ministry of Science and the European Union (MCIN/AEI/10.13039/501100011033) and by ERDF "A way of making Europe"

