Generating Socially-Aware Gestures for Virtual Avatars

David Werkhoven[1,2], Raquel Fernández[1], and Esam Ghaleb[2]

1) MPI for Psycholinguistics, 2) University of Amsterdam

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

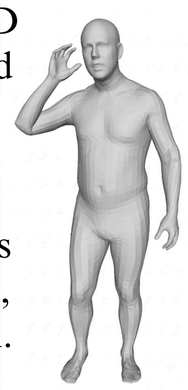
UNIVERSITEIT VAN AMSTERDAM

Current generative models produce monadic gestures tied to static speaker IDs, ignoring the dynamic social and emotional context of face-to-face human interaction.

- How can we generate natural 3D conversational gestures that adapt to both the active speaker and the listening partner?
- Can a continuous social stance vector replace rigid speaker IDs to produce dynamic, personality-driven motion?

Dataset: Seamless Interaction

- Over 4,000 hours of face-to-face conversational interactions.
- Includes speaker audio, VAD timelines, Big Five traits, and interpersonal relationship metadata.
- 30 FPS **SMPL-H** parameters with global pose, translation, and articulated hand motion.

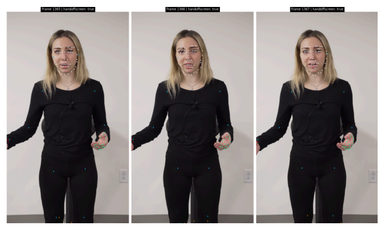


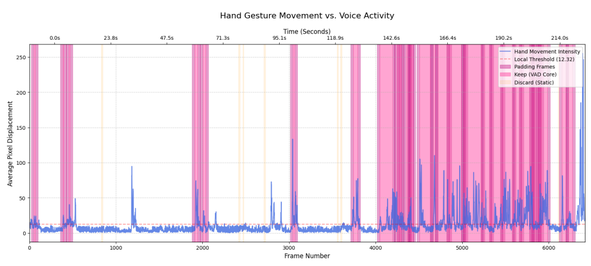
Meta AI

Seamless Interaction

Dataset Pre-Processing

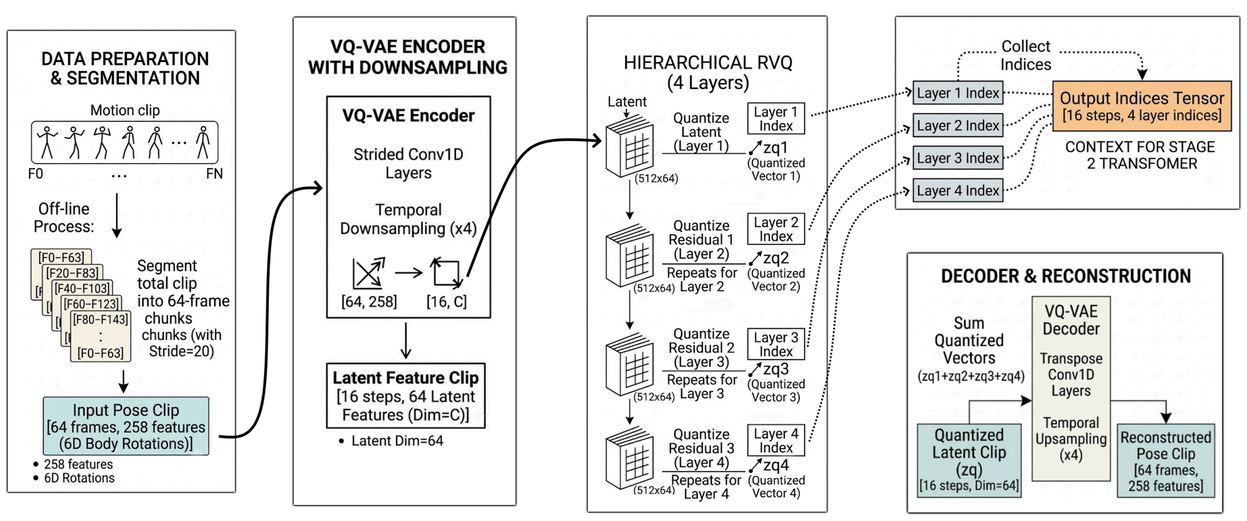
- Exclude windows with off-screen hand occlusions, keeping only visible hand gestures.
- Filter low-motion speech clips using person-specific gesture-intensity thresholds.
- Pad speech segments to capture pre-speech wind-up and post-speech follow-through.





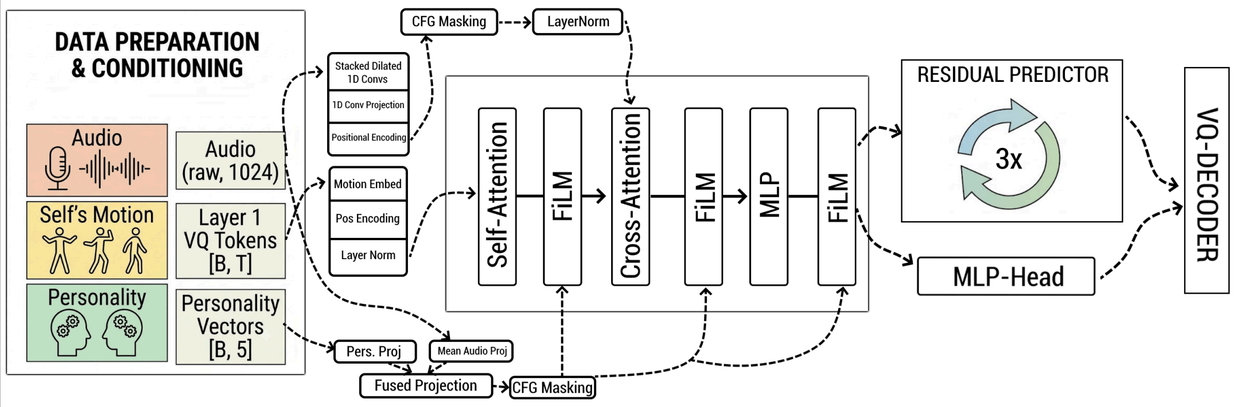
1) Stage 1: Pre-Training for Motion Priors

- **Avoid mean-pose collapse:** Discrete motion tokens support sharp, natural gestures.
- **Learn dynamic motion:** Train the 4-layer downsampled model only on high-intensity segments, treating stationary poses as noise.
- **Enforce pose fidelity:** Optimise VQ-VAE reconstruction with rotation, velocity, acceleration, and full SMPL-H mesh losses.



2) Stage 2: Autoregressive Transformer Approach for Gesture Generation

- **Autoregressive gesture prediction:** Use a causal decoder-only transformer to predict future motion from past gesture context.
- **Audio + identity conditioning:** Condition on single-speaker audio via cross-attention, with speaker identity injected via Feature-Wise Linear Modulation layers.
- **Dyadic extension:** Add speaker and partner audio streams to model interactive speaking and listening gestures.

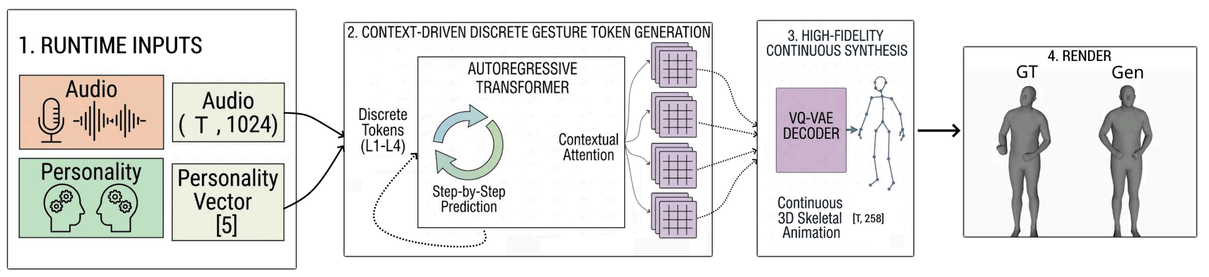


3) Training Strategy for Gesture Token Prediction

- **Dual-objective supervision:** Combine soft-label cross-entropy over codebook distances with MSE on summed continuous latent vectors.
- **Hierarchical masking:** Use cosine-decay block masking across residual layers while keeping Layer 0 fully visible for stability.
- **Conditioning dropout:** Mask audio and speaker vectors with learned null embeddings to enable classifier-free guidance at inference.

4) Autoregressive Motion Inference

- **Audio-driven tokens:** A decoder-only transformer predicts multi-layer motion codes from audio, step by step.
- **Probabilistic gesture sampling:** Sample across four codebook layers with temperature-controlled Top-K/Top-P filtering for varied, natural gestures.
- **Smooth SMPL-H animation:** Use overlapping sliding windows before VQ-VAE decoding to produce temporally smooth 3D motion.



Current Findings

- **Successful Motion Quantisation (Stage 1)**: High-intensity preprocessing enabled robust VQ-VAE compression and high-fidelity motion reconstruction.
- **Autoregressive Generation Bottlenecks (Stage 2)**: The autoregressive transformer captures short-term dependencies but can accumulate errors over long sequences.
- **Architectural Alternatives**: To mitigate long-sequence drift, ongoing exploration is focused on benchmarking alternative generative frameworks, specifically diffusion-based models and encoder-decoder transformer configurations, against the current autoregressive baseline.

Correspondence: esam.ghaleb@mpi.nl • david.werkhoven@student.uva.nl