BEYOND TURN-TAKING: MULTIMODAL CONTINUOUS FULL-DUPLEX EMOTIONAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing multimodal conversational agents follow a turn-taking paradigm and typically remain silent while the user is speaking. This limits their ability to maintain perception while continuously generating multimodal feedback, such as facial expressions, conditioned on the user's evolving state and interaction context. We formulate this setting as multimodal continuous full-duplex emotional interaction, in which perception and response generation proceed concurrently. To support this task, we introduce **ContinuEI**, a dataset containing 30.24 hours of temporally aligned interactions from 1,072 dialogues and 12,303 sentences. It pairs users' visual-acoustic affective cues with system-side text, speech, and 3D facial motion, preserving the temporal correspondence between evolving user states and multimodal system responses. We further develop **FuSE**, an end-to-end **Fu**ll-duplex **S**treaming **E**motional interaction framework built upon Qwen3-Omni to provide a low-latency baseline for this task. FuSE incorporates an Actor module for continuous facial behavior generation, together with spatio-temporal attention and dual-stream conditional modeling. Experiments show that FuSE improves interaction naturalness and reduces response latency compared with conventional cascaded turn-taking baselines. These results support multimodal continuous emotional interaction as a promising task for studying continuous perception and multimodal response generation.

1 INTRODUCTION

Human communication is a continuous, multimodal, and dynamically evolving process. Beyond exchanging information through language, humans continuously perceive and respond to each other's affective states through verbal and non-verbal behaviors, including vocal prosody, facial expressions, and eye gaze. These fine-grained affective signals are fundamental to natural interaction and emotional engagement. Therefore, enabling conversational agents with continuous affective understanding and expressive capabilities has become a key research direction in emotional interaction. However, achieving natural continuous emotional interaction remains challenging. Existing multimodal emotional dialogue systems (Chu et al., 2025; Zhang et al., 2025) predominantly follow a turn-taking paradigm. During continuous user expression, these systems typically wait until users finish speaking before generating a complete verbal response, leaving the agent in a feedback-limited state during listening. This interaction gap disrupts the continuity of emotional feedback and limits the ability of agents to maintain sustained emotional engagement. Recently, full-duplex speech dialogue research (Défossez et al., 2024) has explored mechanisms to overcome traditional turn-taking constraints, such as real-time response generation, interruption handling, and speech overlap management. However, existing studies primarily focus on bidirectional speech synchronization, including turn management and dialogue coordination, while largely leaving unexplored how an agent can continuously generate multimodal affective feedback consistent with the user's evolving state and interaction context while maintaining perception.

Motivated by this limitation, we investigate a new full-duplex paradigm for multimodal emotional interaction. Unlike conventional full-duplex speech communication that emphasizes interruption handling and conversational coordination, our definition of full-duplex primarily focuses on the temporal concurrency between perception and expression. Specifically, an agent should continuously understand users' multimodal states while concurrently generating corresponding verbal and non-verbal feedback without waiting for users to complete their expressions. This paradigm aims to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

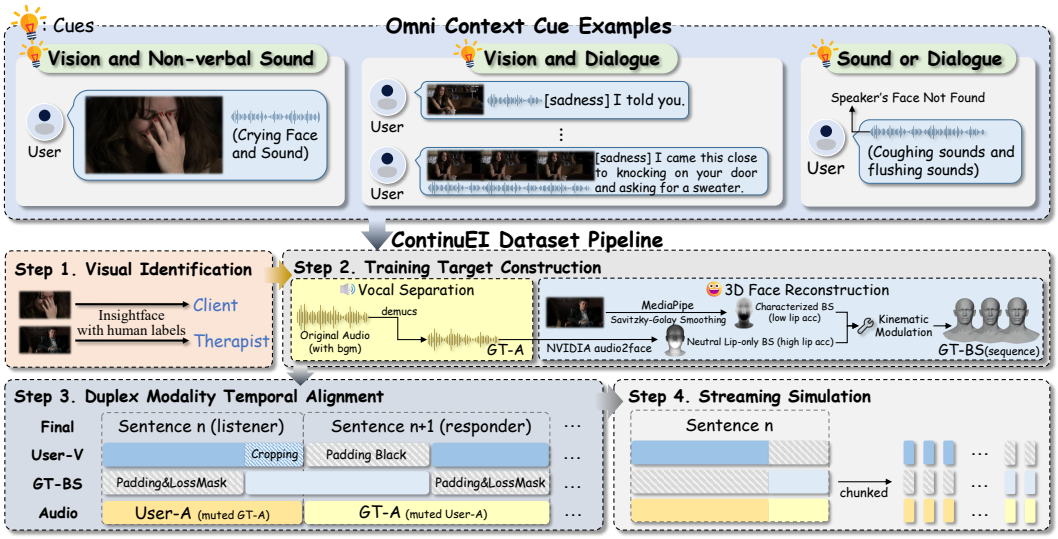


Figure 1: Overview of the proposed **ContinuEI** dataset. **ContinuEI** preserves temporally aligned multimodal contexts and responses, enabling research on continuous emotional interaction. The dataset construction pipeline consists of four steps.

overcome the "silent listening" limitation of turn-based interaction and enable continuous emotional expression throughout the interaction, leading to more natural and emotionally engaging multimodal conversations. Furthermore, for non-verbal expression generation, we adopt ARKit-based[1] 3D facial animation rather than direct video synthesis, due to its lightweight representation, explicit controllability, and compatibility with diverse interactive embodiments. In conclusion, the main contributions of this paper can be summarized as follows:

1. We introduce **ContinuEI**, a dataset for **Continu**ous multimodal **E**motional **I**nteraction. **ContinuEI** provides temporally aligned multimodal contexts and responses, enabling the modeling of temporal dynamics between users' ongoing affective cues and system verbal and non-verbal responses, and facilitating research on continuous emotional interaction.
2. We propose **FuSE**, an end-to-end **Fu**ll-duplex **S**treaming **E**motional interaction framework, enabling agents to continuously understand user states while concurrently generating context-consistent verbal and non-verbal feedback, supporting continuous and emotionally coherent interaction.
3. Experiments support the feasibility of continuous full-duplex emotional interaction, with **FuSE** achieving competitive interaction naturalness and first-packet latency relative to cascaded baselines and providing an end-to-end reference implementation for future research.

2 Related Work

2.1 Multimodal Emotional Conversational Agents

Early research on emotional dialogue systems primarily focused on text-based interactions, including emotion recognition, empathy strategy modeling, and emotionally consistent response generation (Liu et al., 2021; Zhang et al., 2024; Xie et al., 2025). Recent advances in Omni-modal Large Language Models (OLLMs) (Jin Xu, 2025; Xu et al., 2025; Yao et al., 2024) have further extended conversational agents beyond text-based communication by enabling unified modeling of language, speech, and visual modalities. These multimodal capabilities allow agents to better capture users' affective states from diverse signals (Chu et al., 2025) and generate richer multimodal emotional responses for more engaging interactions (Zhang et al., 2025). However, despite their strong multimodal understanding and generation capabilities, existing systems still predominantly follow turn-taking interaction protocols, where response generation begins only after receiving complete user

[1] https://arkit-face-blendshapes.com/

inputs. Therefore, enabling continuous emotional expression throughout user interaction remains an important open challenge.

2.2 Continuous Full-Duplex Interaction

Recently, full-duplex speech dialogue systems (OpenAI, 2024; Défossez et al., 2024; Fu et al., 2025; 2024) have explored mechanisms to overcome traditional turn-taking constraints by enabling concurrent listening and speaking through real-time response generation, interruption handling, and overlap management. These approaches have significantly advanced conversational coordination and speech-level synchronization. Nevertheless, maintaining continuous multimodal affective engagement throughout interaction remains an open direction. Mihoyo's LPM (Zeng et al., 2026) investigates continuous emotional interaction through 2D video generation by introducing dual-stream acoustic diffusion conditioning for expressive response generation. While demonstrating the potential of continuous emotional expression, it primarily focuses on speech-driven video synthesis and does not explicitly consider multimodal user-state understanding or a unified interaction framework. In summary, these developments motivate a new full-duplex interaction paradigm that enables continuous affective engagement through concurrent perception and expression.

2.3 3D-Based Facial Animation

Parametric 3D facial animation has been widely studied for generating explicit facial motion representations, offering controllable and low-latency non-verbal behavior generation for interactive agents. Recent advances have evolved from early deep learning approaches (Cudeiro et al., 2019; Richard et al., 2021) to Transformer-based methods (Fan et al., 2022) and diffusion-based models (Ma et al., 2024; Yang et al., 2025; NVIDIA et al., 2025). Meanwhile, VQ-VAE based latent representations (Xing et al., 2023) have further improved the modeling of complex facial dynamics. These developments have significantly enhanced the naturalness and expressiveness of generated facial motions. However, existing methods primarily model facial animation as a response-driven generation process rather than an interaction-driven process that jointly considers evolving user states and context.

3 Methodology

3.1 Task Definition

Given a multimodal dialogue history $H_{1:t-1}$ and current streaming user visual V_t and acoustic A_t input, continuous full-duplex emotional interaction aims to enable an agent to generate appropriate multimodal feedback during both listening and responding periods. Formally, the interaction process is defined as $(H_{1:t-1}, V_t, A_t) \rightarrow R_t$, where the response R_t depends on the agent's interaction state:

$$R_t = \begin{cases} F_t, & \text{listener backchannel,} \\ (Y_t, S_t, F_t), & \text{response generation,} \end{cases}$$

where F_t, Y_t, and S_t denote 3D facial animation, textual, and speech response, respectively. Although the interaction still involves transitions between listening and responding states, it differs from conventional turn-taking paradigms where the agent remains silent during user speaking periods. Instead, it continuously provides affective feedback while maintaining perception of the user's ongoing state.

3.2 Data Construction Pipeline

To support continuous full-duplex emotional interaction, **ContinuEI** is constructed to provide temporally aligned multimodal contexts and responses. The construction pipeline is based on American psychology TV Series "In Treatment" and its corresponding speech annotations (Chu et al., 2025), and consist of four stages.

Visual Identification Given a raw video episode, we first identify the visual speaker in each frame and extract speaker-consistent clips. Specifically, with a small number of manual annotations, InsightFace (Guo et al., 2021) is used to detect facial bounding boxes and associate faces across frames.

The identified regions are then used to separate each episode into clean speaker-specific video clips for subsequent training target construction.

Training Target Construction To construct high-fidelity 3D facial motion ground truth, we combine audio-driven face generator and vision-driven face reconstructor. For the audio modality, speaker-specific vocal signals are extracted through vocal separation[2] (Rouard et al., 2023) based on existing speech timestamps. We then employ NVIDIA Audio2Face-3D v3[3] (NVIDIA et al., 2025) to generate speech-driven neutral lip motion represented by BlendShapes (BS). For visual facial behaviors, we crop therapist-specific clips according to the detected bounding boxes and utilize MediaPipe[4] to recover vision-driven facial motions. To alleviate motion jitter caused by large pose variations, we further apply Savitzky-Golay filtering to preserve facial dynamics while smoothing reconstruction noise. However, it is still hard to reconstruct the lip motions from vision. Therefore, the audio-driven lip motions and vision-driven facial expressions are resampled to the same temporal resolution and integrated through kinematic modulation, producing high-fidelity ARKit BlendShape sequences. The effectiveness of the constructed facial motions is further validated through human evaluation. Details of modulation and validation are released in Appendix. The resulting ARKit BS representation provides semantically interpretable facial controls and are widely adopted in real-time avatar systems due to their compatibility with diverse embodiments.

Duplex Modality Temporal Alignment Different from conventional dialogue datasets that organize interactions as isolated user-system turns, **ContinuEI** preserves the temporal correspondence between users' ongoing affective cues and agents' multimodal responses. Specifically, we align user visual and speech inputs, system speech, and facial motion sequences along a unified timeline, enabling the modeling of dynamic transitions between user states and system behaviors during continuous interaction. To prevent information leakage, we perform strict modality isolation during alignment. For segments where users and therapists appear simultaneously, the therapist region is masked from the user visual input according to the detected facial bounding boxes, ensuring that the model only accesses user-side observations when generating responses. Through this duplex temporal alignment, **ContinuEI** provides synchronized multimodal contexts and responses for studying continuous emotional interaction.

Table 1: Statistics of proposed dataset **ContinuEI**.

Split (ratio)	Time (hours)	Dialogues	Sentences (responder ratio %)
train (98%)	29.620	1043	12003 (47.430%)
test (2%)	0.620	29	300 (47.667%)
total (100%)	30.240	1072	12303 (47.436%)

Streaming Simulation Finally, we simulate online interaction by chunking the above sentences. The final statistics of **ContinuEI** are as shown in Table 1.

3.3 Model Architecture

As illustrated in Fig.2, our framework **FuSE** enables continuous perception and emotional expression in multimodal emotional interaction, by integrating Qwen3-Omni (Xu et al., 2025) with an additional Actor module.

Facial Motion Latent Representation

Directly predicting high-dimensional facial motion sequences has been proven challenging due to the complex temporal and spatial correlations among facial movements. Therefore, following CodeTalker (Xing et al., 2023), we introduce a similar transformer-based VQ-VAE (van den Oord et al., 2018) as facial motion tokenizer to encode raw blendshape sequences into a compact latent space. The VQ-VAE consists of an encoder E, a decoder D, and a codebook $\mathcal{Z} = \{\mathbf{z}_j \in \mathbb{R}^C\}_{j=1}^N$ containing N facial motion primitives $\mathbf{z}_j$. The input sequence of 52 blendshape coefficients $\mathbf{x} \in \mathbb{R}^{T\times 52}$ is encoded into a continuous latent representation $\hat{\mathbf{z}} = E(\mathbf{x}) \in \mathbb{R}^{T\times G\times C}$, where T is the number of frames, G is the number of latent groups per frame, and C is the code dimension.

[2]We use demucs with mdx extra model as vocal separator.

[3]We use NVIDIA Audio2Face-3D v3 due to its strong speech-facial alignment capability.

[4]MediaPipe is an open-source facial reconstruction framework supporting ARKit BlendShape estimation. `https://developers.google.com/mediapipe`

The latent representation is then quantized by nearest-neighbor lookup:

$$\mathbf{z}_q = \text{quantize}(\hat{\mathbf{z}}) = \mathbf{z}_{k^*}, \quad k^* = \arg\min_{\mathbf{z}_k \in \mathcal{Z}} \|\hat{\mathbf{z}} - \mathbf{z}_k\|_2 . \tag{1}$$

Finally, the decoder reconstructs the input blendshape sequence from the quantized latent representation $\hat{\mathbf{x}} = D(\mathbf{z}_q) = D(\text{quantize}(E(\mathbf{x})))$.

Dual-Stream Conditional Modeling Inspired by LPM (Zeng et al., 2026), **FuSE** introduces a dual-stream conditional modeling mechanism to jointly capture user affective states and system response dynamics. Specifically, the *Actor* receives two temporally aligned conditions, including user multi-modal middel-hidden H_u from *Thinker* and system speech hidden H_a from *Talker*:

$$\begin{cases} H_u = \text{Thinker}_{<\ell/2}(V_u, A_u), \\ H_a = \text{Adapter}(\text{Talker}_{<\ell'-1}(H_u, T)), \end{cases} \tag{2}$$

where $\text{Adapter}(\cdot)$ is a one-layer transformer, V_u, A_u, and T denote user visual, audio input, and generated text response, respectively. The reason of using the former $\ell^{'} - 1$ layers from *Talker* will be expanded in the later training section. These two conditions are temporally aligned with the facial latent representation $\mathbf{z}^l \in \mathbb{R}^{T \times G \times h}$, and alternately integrated into cross-attention layers:

$$\tilde{\mathbf{z}}^l = \begin{cases} \text{CrossAttn}(\mathbf{z}^l, U_{proj}(H_u)) + \mathbf{z}^l, & l \in \text{even}, \\ \text{CrossAttn}(\mathbf{z}^l, T_{proj}(H_a)) + \mathbf{z}^l, & l \in \text{odd}. \end{cases} \tag{3}$$

where h and l is the hidden dimension and layer of *Actor*.

Spatial-Temporal Facial Modeling

Facial expressions exhibit complex dependencies across both temporal evolution and coordinated facial movements. Therefore, we adopt factorized spatial-temporal self-attention over facial motion latents:

$$\mathbf{z}_t^l = \text{Attn}_{time}(\mathbf{z}^l + \mathbf{P}_t), \quad \mathbf{z}_s^l = \text{Attn}_{space}(\mathbf{z}_t^l + \mathbf{P}_s) \tag{4}$$

where P_t and P_s denote temporal and spatial positional embeddings, respectively. This factorized design enables the *Actor* to model continuous expression trajectories while maintaining coordinated facial movements with reduced computational cost.

Real-time Facial Motion Generation

To support low-latency emotional interaction, **FuSE** adopt a chunk-wise streaming generation manner. Under a fixed chunk size, we introduce two temporal conditions for better management. The history length len_h indicates the available preceding context, which is automatically calculated during running. The commit length len_c specifies the portion of generated motion that has been confirmed for output. These conditions assist the model to handle first short packet and the subsequent long packets. Finally the *Actor* progressively generates facial motion chunks at 25 Hz, twice of the *Talker* speech token rate.

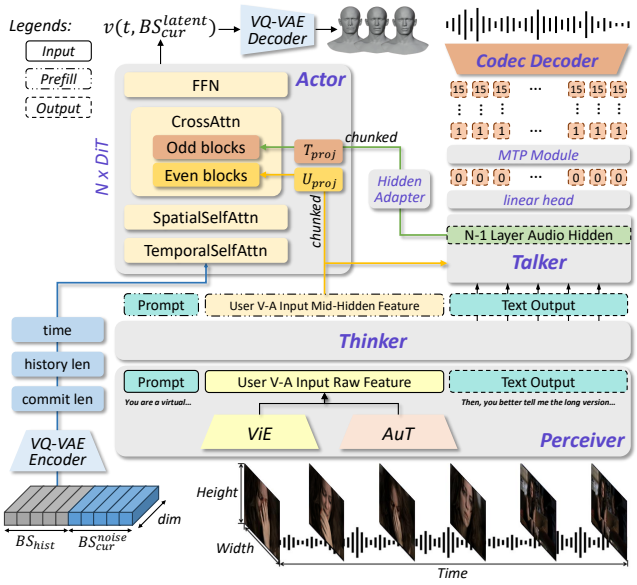


Figure 2: Overview of the proposed framework **FuSE**, which extends Qwen3-Omni with an *Actor* module to enable continuous full-duplex emotional interaction.

3.4 Training Strategy

Following the training strategy of Qwen3-Omni (Xu et al., 2025), we adopt a module-independent training

paradigm for **FuSE**, where the Thinker, Talker, and Actor are optimized with their corresponding objectives.

Thinker provides the shared multimodal hidden representation H_u for both the *Talker* and *Actor*. Following the original Qwen3-Omni design, H_u is extracted from the intermediate layer ($\ell/2$) of the *Thinker* and serves as the interface for downstream modules. Therefore, to preserve the hidden representation distribution and maintain compatibility with independently optimized downstream modules, we freeze $\text{Thinker}_{<\ell/2}$ and only fine-tune the upper $\ell/2$ layers with the standard autoregressive language modeling objective: $\mathcal{L}_{\text{Thinker}} = -\sum_i \log p_\theta(y_i \mid y_{<i}, H_u)$.

Talker takes the shared user representation H_u and the generated text response from the *Thinker* as inputs, and autoregressively predicts speech codec tokens. Since the original Qwen3-Omni implementation does not provide an explicit codec tokenizer, we replace the original speech generation modules with Qwen3TTS, which adopts a similar architecture. Given target speech, we use the tokenizer encoder to extract codec tokens $\mathbf{c} = \{c^i\}_{i=1}^{K-1}$, where c^0 denotes the main code and the remaining ones represent residual codes. The *Talker* is optimized with both main-code and residual-code prediction cross-entropy objectives: $\mathcal{L}_{\text{Talker}} = \text{CE}(\hat{c}^0, c^0) + \frac{1}{K-1}\sum_{k=1}^{K-1} \text{CE}(\hat{c}^k, c^k)$, where K is the codec size. To provide continuous system acoustic hidden representations for the *Actor* during training, we train a separate Qwen3TTS-based acoustic reconstructor to extract acoustic representations from target speech. And a lightweight one-layer adapter is trained with Mean Square Error (MSE) loss to map the $\ell - 1$ layer hidden states of the *Talker* into this acoustic representation space.

Actor is optimized with a flow-matching objective. Given a facial latent data sample $x_1 \sim \mathcal{D}$ and noise sample $x_0 \sim \mathcal{N}(0, \mathbb{I})$, the intermediate state can be defined as $x_t = (1-t)x_0 + tx_1$, and the model learns the velocity field: $\mathcal{L}_{FM} = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2\right]$.

4 Experiments and Results

4.1 Datasets and Models

We train and evaluate our models on **ContinuEI**. All models are trained using the LoRA (Hu et al., 2021) on two NVIDIA H200 80G GPUs. More details of the implementation are provided in the Appendix. For cascaded baselines, we use Gemini3.5Flash (Google DeepMind, 2026) plus Qwen3TTS-12Hz-0.6B (Hu et al., 2026), and Qwen3-Omni (A3B+A0.3B) (Xu et al., 2025) for Omni-to-Speech, and Emotalk (Peng et al., 2023) and Audio2Face-3D-v3 (NVIDIA et al., 2025) for Speech-to-Face. We further conduct module-level evaluation on public benchmarks. For text-to-speech, we use the English zero-shot subset of Seed-TTS-Eval (Anastassiou et al., 2024), which contains 1,088 prompt-target cases. For speech-to-face, we use the English test split from BEAT (Liu et al., 2022), an open-source dataset which contains Arkit Blendshape ground-truth at 25 FPS. The BEAT test set consists of 400 samples from 30 speakers, averaged 84.3 seconds each sample.

4.2 Evaluation Metrics

Since **FuSE** aims to enable continuous multimodal emotional interaction, we evaluate the generated responses from four aspects: verbal response quality, speech quality, facial expression quality, and interaction efficiency. **Verbal Response Quality**: We use GPT-5.5 to evaluate the generated textual responses from three perspectives: empathy (Emp.), contextual coherence (Coh.), and emotional consistency (Emo.). **Speech Quality**[5]: For generated speech, we adopt word error rate (WER) to evaluate intelligibility, speaker similarity (SS) to measure speaker identity preservation, and Emotional Accuracy (EmoACC) to assess acoustic emotional consistency with reference audio. **Facial Expression Quality**: For generated 3D facial motions, we use Lip Vertex Error (LVE) to measure lip motion accuracy, Facial Dynamics Distance (FDD) to evaluate the temporal consistency and

[5]WER is computed using Whisper Large-v3 (`openai/whisper-large-v3`); SS is computed as the cosine similarity between speaker embeddings extracted by ERes2NetV2-W24S4EP4 (`iic/speech_eres2netv2w24s4ep4_sv_zh-cn_16k-common`); and EmoACC measures agreement between utterance-level emotion labels predicted for generated and reference audio by emotion2vec+ Large (`iic/emotion2vec_plus_large`).

Table 2: Results of end-to-end continous emotional interaction on **ContinuEI**. † denotes NON-streaming models, where "Latency" is reported for whole sentences. WER, CEC, SLC, and NER are all calculated based on predicted text responses and speeches. Best results are in **bold**, while the second one are underlined.

(a) Text Qualitative Evaluation (Metrics-Models)

Metrics	Gemini3.5Flash	Qwen3Omni	Ours
Emp.↑	**2.05**	1.90	2.03
Coh.↑	2.61	2.49	**3.03**
Emo.↑	2.22	2.17	**2.41**
Latency(ms)↓	13,729.75	472.91	**430.03**

(b) Speech Qualitative Evaluation (Metrics-Models).

Metrics	Gemini3.5Flash + Qwen3TTS	Qwen3Omni	Ours
WER↓	**1.42%**	2.63%	1.61%
SS↑	0.374	0.224	**0.377**
EmoACC↑	**0.302**	0.221	0.235
Latency(ms)↓	230.09	219.06	**193.55**
Speech Naturalness↑	**4.14**	3.82	3.86

(c) Facial Motion Qualitative Evaluation (Models-Metrics).

MLLM Base Model	BS Model	Our Test (responder part)					Our Test (listener part)				
		Latency (ms)↓	Total Latency (ms)↓	Facial Naturalness↑	CEC↑	SLC↑	LVE ($\times 10^{-4}$mm)↓	FDD ($\times 10^{-5}$mm)↓	Latency (ms)↓	Facial Naturalness↑	NER↑
Gemini3.5Flash + Qwen3TTS	A2F-3D-v3	226.40	14,186.24	**3.75**	**3.64**	**4.03**	N/A	N/A	N/A	N/A	N/A
	EmoTalk†	343.51	18,153.26	3.10	3.12	2.87	N/A	N/A	N/A	N/A	N/A
Qwen3Omni	A2F-3D-v3	216.39	908.36	3.51	3.59	3.84	N/A	N/A	N/A	N/A	N/A
	EmoTalk†	**141.94**	12,179.76	3.06	2.98	3.03	N/A	N/A	N/A	N/A	N/A
Ours (Actor: fat300M, steps=20)		282.04	**905.62**	3.41	3.29	3.26	**3.94**	**3.87**	**287.99**	**3.45**	**3.59**

naturalness of facial movements, and Root Mean Square Error (RMSE) of BS coefficients. **Interaction Efficiency**: We record "Latency" as the first-packet latency for each module, while "Total Latency" as the elapsed time from the end of user input to the generation of the first BS frame. **Human Evaluation**: We recruit graduate students to evaluate the naturalness of generated speech and facial expressions, cross-modal emotional consistency (CEC), speech-lip consistency (SLC), and non-verbal emotional resonance (NER). Specifically, NER measures whether the generated facial motion preserves the emotional intensity conveyed in the user input video.

4.3 Main Results

Text Response Quality. As shown in Table 2(a), existing MLLMs still show limited response quality in the end-to-end video-chat setting. Gemini3.5Flash achieves competitive empathy but incurs high latency. In contrast, our model obtains the best contextual coherence and emotional expression scores, while maintaining the lowest text response latency. We attribute the limited performance of Gemini3.5Flash partly to the absence of user transcripts in our task setting, which aligns real-time interactions without transcripts as direct input. Consistent with this explanation, providing Gemini3.5Flash with user text improves its Emp., Coh., and Emo. scores to **2.81**, **3.85**, and **3.09**, respectively.

Speech Generation Quality. Table 2(b) shows that our *Talker* consistently outperforms Qwen3Omni Talker and remains competitive with Qwen3TTS, which receives one-shot reference audio and text from the training set. Our *Talker* achieves relatively better speaker similarity and lower first packet latency while maintaining comparable WER and speech naturalness. These results demonstrate improved identity preservation and streaming efficiency without compromising speech quality.

Facial Motion Generation Quality. Because predicted responses generally differ from the ground-truth utterances, reference-based facial metrics are not applicable to responder motion and are omitted. As shown in Table 2(c), our 300-million-parameter *Actor* achieves higher responder facial naturalness than the 600-million-parameter EmoTalk. A2F-3D-v3 nevertheless obtains the best responder-side human evaluation scores, indicating an advantage of its dedicated two-stage design, which explicitly decomposes audio-to-emotion and audio-to-face modeling. For listener motion, our *Actor* achieves competitive LVE and FDD while generating natural non-verbal feedback, serving as a baseline for future researches.

Latency Analysis. Cascaded systems incur relatively high latency because their modules operate sequentially. The delay increases further for non-streaming facial animation models such as EmoTalk, which require complete synthesized speech before generating facial motion. In contrast, our system achieves the lowest end-to-end latency among all evaluated methods. Notably, in practical online

Table 3: Text-to-speech module-independent evaluation on our test set and Seed-TTS-Eval. Best results are in **bold**, while the second one are underlined.

Model	Ours				Seed-TTS-Eval	
	WER↓	SS↑	EmoACC↑	Latency(ms)↓	WER↓	Latency(ms)↓
Qwen3-TTS (with ref)	**1.24%**	**0.427**	0.434	232.51	**1.83%**	232.21
Qwen3-Omni (Talker)	4.24%	0.279	0.196	220.63	2.32%	216.27
Our *Talker*	3.31%	0.423	**0.455**	**190.23**	2.14%	**177.65**

Table 4: Speech-to-face module-independent evaluation on our test set and BEAT-test. Best results are in bold, while the second one are underlined. † denotes NON-streaming models For fat300M *Actor*, we use 20 steps for diffusion.

Model	Our Test (responder part)				BEAT-test			
	LVE ($\times 10^{-4}$mm)↓	FDD ($\times 10^{-5}$mm)↓	RMSE↓	Latency (ms)↓	LVE ($\times 10^{-4}$mm)↓	FDD ($\times 10^{-5}$mm)↓	RMSE↓	Latency (ms)↓
A2F-3D-v3	3.36	**5.52**	0.211	235.18	2.72	**2.45**	0.181	**281.06**
EmoTalk†	4.25	5.53	0.171	198.17	2.24	2.90	0.143	561.97
UniTalker-D3†	**3.05**	6.50	0.158	**108.91**	**1.62**	4.20	**0.121**	416.37
Ours (fat300M)	3.76	5.69	**0.155**	282.04	2.19	4.52	0.208	320.35

deployment, the user-perceived response gap can be nearly eliminated. During responder inference, a parallel listener stream can continuously process incoming user input and condition the *Actor* to generate real-time non-verbal feedback. Once the first *Talker* token is generated, the system switches from listener feedback to responder generation. This strategy maintains continuous feedback while preparing the response, although the parallel listener stream requires additional GPU memory.

Table 5: Ablation results on *Actor* module under the testset of **ContinuEI**. N/A denotes Not Applicable.

Model	Our Test (responder part)				Our Test (listener part)			
	LVE ($\times 10^{-4}$mm)↓	FDD ($\times 10^{-5}$mm)↓	RMSE↓	Latency (ms)↓	LVE ($\times 10^{-4}$mm)↓	FDD ($\times 10^{-5}$mm)↓	RMSE↓	Latency (ms)↓
50M	3.79	5.76	0.153	691.14	3.69	3.86	0.179	750.18
50M w/o Spatial SA	3.85	6.01	0.155	504.18	3.73	3.99	0.173	491.75
50M w/o Temporal SA	4.11	5.43	0.170	470.71	4.5	6.12	0.204	466.66
50M w/o Talker CA	N/A	N/A	N/A	N/A	3.23	4.32	0.173	744.71
50M w/o User CA	3.66	6.01	0.152	376.56	N/A	N/A	N/A	N/A
deep300M	3.90	5.62	0.157	736.13	4.34	4.00	0.179	723.32
1B	4.27	5.39	0.170	964.87	4.74	3.94	0.192	736.13
fat300M step=1	3.34	6.99	0.149	19.92	3.24	4.81	0.150	24.33
fat300M step=5	3.49	6.18	0.148	101.70	3.63	4.12	0.169	103.16
fat300M step=10	3.64	5.76	0.152	246.29	3.80	3.92	0.177	248.78
fat300M step=20	3.76	5.69	0.155	282.04	3.94	3.87	0.181	287.99

4.4 Module-Level Evaluation

Text-to-Speech. We extend Qwen3TTS to condition on user multimodal representations, allowing a unified *Talker* to generate acoustic codec tokens and hidden states for facial generation. As shown in Table 3, our *Talker* outperforms the original Qwen3-Omni Talker under matched input-output conditions and performs comparably to Qwen3TTS with one-shot reference audio. It also achieves higher *EmoACC* and lower latency, while its Seed-TTS Eval results support cross-dataset generalization.

Speech-to-Face. As shown in Table 4, our speech-to-face module achieves a favorable accuracy-efficiency trade-off across both datasets. On our test set, it obtains the best RMSE of 0.155 while maintaining competitive LVE and FDD compared with existing methods. On BEAT-test, since the speech become much longer, our model achieves strong LVE performance and substantially lower latency than non-streaming model, demonstrating better efficiency under cross-dataset evaluation. These results suggest that the proposed module generalizes reasonably well. However, its FDD and RMSE indicate remaining limitations in fine-grained motion modeling under domain shift.

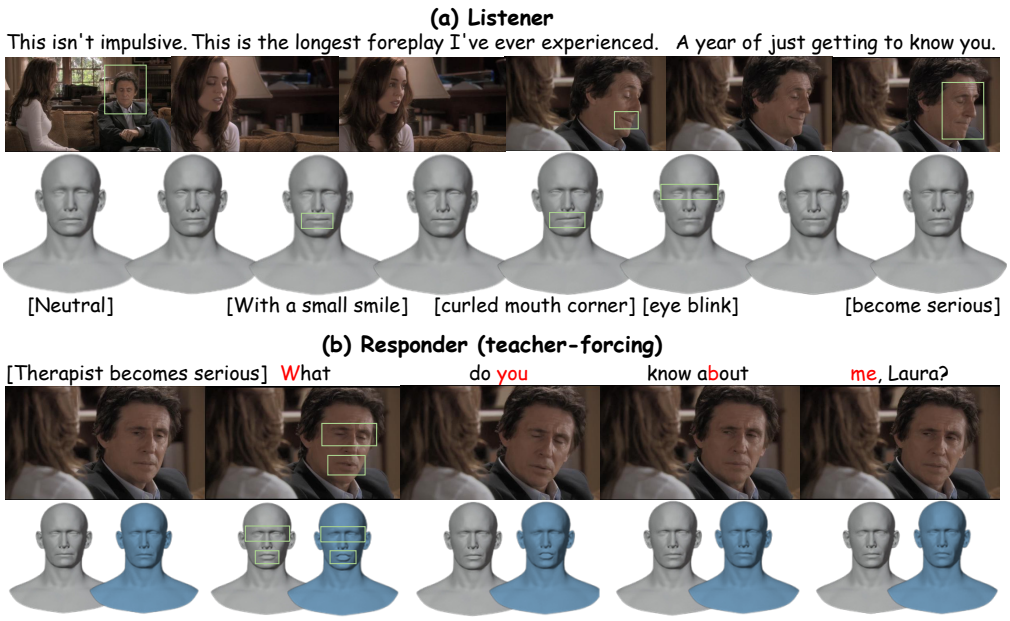


Figure 3: Case study on **FuSE** during continuous emotional interaction. The blue faces are targets for the responder sentence.

Ablation Study. As shown in Table 5, the ablation study verifies the effectiveness of both model design and diffusion sampling strategy in the *Actor* module. **Module Ablation.** Removing Spatial Self-Attention (SA) or Talker Cross-Attention (CA) generally degrades LVE and RMSE, while removing User CA leads to invalid listener prediction, indicating that cross-attention from listener context is essential for listener-side motion generation. Temporal SA improves temporal consistency, as removing it increases LVE and RMSE, especially on the listener part. **Diffusion Sampling Ablation.** Compared with larger variants, fat300M provides a better efficiency-accuracy trade-off: with only one diffusion step, it achieves the lowest latency and strong RMSE on both responder and listener parts, while increasing the number of steps improves FDD but also introduces higher latency. Overall, the results show that the proposed attention components are necessary for stable dyadic facial motion modeling, and fat300M with few-step sampling is the most practical configuration for efficient streaming inference. The details parameters for four scale are illustrated in appendix.

Case Study. As demonstrated in Figure 3, **FuSE** generates facial behaviors that are both speech-aligned and context-aware. In the listener case, it produces natural non-verbal reactions that are consistent with the speaker's utterance, such as smiles, mouth-corner movements, blinks, and serious expressions without direct speech input. This suggests that the model captures both context and acoustic cues from the user hidden. In the responder case, the generated expressions follow both the spoken content and the interlocutor's emotional state (frown in confusion), with facial emphasis appearing around key words. This demonstrates the model's ability to capture subtle dyadic facial dynamics beyond simple audio-to-motion mapping, enabling continuous emotional interaction.

5 Conclusion

We formulate multimodal continuous full-duplex emotional interaction as a task in which agents perceive evolving user states while concurrently generating verbal and non-verbal feedback. To support this task, we introduce **ContinuEI**, a dataset of temporally aligned user affective cues and multimodal agent responses, and **FuSE**, an end-to-end streaming framework for concurrent perception and response generation. Experiments support the feasibility of this setting, with **FuSE** achieving competitive interaction naturalness and first-packet latency relative to cascaded baselines. Together, they provide the data and an end-to-end reference implementation for future research on continuous multimodal emotional interaction.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, translation assistance, and minor writing refinement. We did not use generative AI tools to generate the core scientific ideas, experimental conclusions, or unsupported claims. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. **ContinuEI** is constructed from the publicly released MESC Chu et al. (2025) resources and follows the license and citation requirements of the original data, including the Apache-2.0 license. We use the data only for research on continuous emotional interaction modeling. Our released dataset will include the additional visual speaker annotations, 3D facial motions, and the construction code, but will not override the rights or terms of the original source materials. Details are provided in appendix.

REPRODUCIBILITY STATEMENT

The main paper describes the task definition, ContinuEI construction and temporal-alignment pipeline, FuSE architecture, training objectives, evaluation metrics, baseline settings, and latency protocol. The appendix provides additional details on data preprocessing and filtering, dataset splits, implementation settings, model and optimization hyperparameters, evaluation prompts, human-evaluation procedures, computing infrastructure, and additional experiments. We document the model versions, prompts, decoding settings, and query protocols used for closed-source evaluation models. Our data and code will be released upon publication.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024. URL https://arxiv.org/abs/2406.02430.

Karl Johan Åström and Lars Rundqwist. Integrator windup and how to avoid it. In *Proceedings of the 1989 American Control Conference*, pp. 1693–1698, Pittsburgh, PA, USA, 1989. doi: 10.23919/ACC.1989.4790464.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 27:8276–8287, 2025. doi: 10.1109/TMM.2025.3604951.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, pp. 10101–10111, 2019.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, 2024. URL https://arxiv.org/abs/2410.00037.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Google DeepMind. Gemini 3.5 flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026. Accessed: 2026-07-27.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

Shan He, Haonan He, Shuo Yang, Xiaoyan Wu, Pengcheng Xia, Bing Yin, Cong Liu, Lirong Dai, and Chang Xu. Speech4mesh: Speech-assisted monocular 3d facial reconstruction for speech-driven 3d facial animation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14146–14156, 2023. doi: 10.1109/ICCV51070.2023.01305.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960. doi: 10.1115/1.3662552. URL http://dx.doi.org/10.1115/1.3662552.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis, 2022. URL https://arxiv.org/abs/2203.05297.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer, 2024. URL https://arxiv.org/abs/2402.05712.

NVIDIA, :, Chaeyeon Chung, Ilya Fedorov, Michael Huang, Aleksey Karmanov, Dmitry Korobchenko, Roger Ribera, and Yeongho Seol. Audio2face-3d: Audio-driven realistic facial animation for digital avatars, 2025. URL https://arxiv.org/abs/2508.16401.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation, 2023. URL https://arxiv.org/abs/2303.11089.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1173–1182, October 2021.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1081–1115, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.55. URL https://aclanthology.org/2025.acl-long.55/.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12780–12790, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Yifan Yang, Zhi Cen, Sida Peng, Xiangwei Chen, Yifu Deng, Xinyu Zhu, Fan Jia, Xiaowei Zhou, and Hujun Bao. Streamingtalker: Audio-driven 3d facial animation with autoregressive diffusion model. *arXiv preprint arXiv:2511.14223*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, Jeremy Pi, Leo Li, Mingyi Shi, Shawn Wang, Sheng Bi, Steven Tang, Thorn Hang, Tobey Guo, Vincent Li, Xin Tong, Yikang Li, Yuchen Sun, Yue Zhao, Yuhan Lu, Yuwei Li, Zane Zhang, Zeshi Yang, and Zi Ye. Lpm 1.0: Video-based character performance model, 2026. URL https://arxiv.org/abs/2604.07823.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark, 2025. URL https://arxiv.org/abs/2502.04976.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13395–13412, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.723. URL https://aclanthology.org/2024.acl-long.723/.

A Ethics

Data Use and Release. **ContinuEI** is constructed from the publicly released MESC Chu et al. (2025) resources and follows the license and citation requirements of the original data, including the Apache-2.0 license. We use the data only for research on continuous emotional interaction modeling. Our released dataset will include the additional visual speaker annotations, 3D facial motions, and

the construction code, but will not override the rights or terms of the original source materials. Our data and code will be released upon publication.

Human Evaluation. Since the predicted text responses in end-to-end evaluation usually differ from the ground-truth responder utterances, reference-based facial motion metrics are not applicable to the responder part. Therefore, we recruit five human experts with at least a master degree, to evaluate the generated speech and facial expressions on a scale of 0 to 5. Specifically, evaluators are clearly informed of the task and scoring criteria before annotation:

- **Speech Naturalness**: Evaluators listened only to the generated speech and assessed whether the pronunciation was clear, the timbre was stable, and the intonation, pauses, and emotional delivery sounded natural. They also considered mechanical artifacts, discontinuities, clipping, noise, and other obvious signs of synthesis.
- **Facial Naturalness**: Evaluators viewed only the generated facial animation and assessed whether the facial expressions, eye blinks, head motion, and mouth movements resembled those of a real speaker. They also considered perceptual unnaturalness, rigidity, jitter, mesh penetration, abrupt motion jumps, abnormal eye closure, and facial drift.
- **Cross-modal Emotional Consistency (CEC)**: Evaluators jointly considered the generated speech and facial animation, and assessed whether the vocal emotion, facial expression, intensity, and temporal rhythm conveyed a consistent emotional state and attitude. If the speech carried weak emotional cues, a sample could still receive a high score when the facial animation was emotionally compatible with the audio.
- **Speech-Lip Consistency (SLC)**: Evaluators assessed whether mouth opening and closing, bilabial closures, vowel-related mouth aperture, and articulation timing were synchronized with the speech. They penalized early, delayed, or incorrect mouth movements.
- **Non-verbal Emotional Resonance (NER)**: Evaluators compared the generated facial motions with the source video clip to assess whether they preserved the emotional intensity conveyed by the client. Samples were penalized when the facial behaviors were emotionally flat, exaggerated, mismatched with the therapist's real reaction if applicable, or insensitive to changes in the client's expression.
- **Rating scale.** All criteria were rated on a 0–5 scale:
  - **5:** Highly natural or consistent output with almost no perceptible issues.
  - **4:** Generally good quality with only minor artifacts that do not substantially affect perception.
  - **3:** Acceptable quality, but with noticeable unnaturalness or asynchrony.
  - **2:** Clear issues that affect intelligibility or viewing comfort.
  - **1:** Unacceptable output with severe distortion, misalignment, or perceptual inconsistency.
  - **0:** Completely corrupted facial animation or audio.

All samples are anonymized and presented in a blind randomized order, without revealing the model identity. We report only averaged scores and do not collect personal or sensitive information from annotators beyond what is required for annotation management.

Responsible Use. The system is intended for research on embodied emotional interaction, not for clinical diagnosis, treatment, or replacement of professional mental-health services. Because generated responses may be inappropriate in high-risk settings, deployment should include human oversight, safety filtering, and clear disclosure that users are interacting with an AI system.

B Details of Data Construction

B.1 Visual Identification

We construct role-aware visual annotations for the visible *Client* and *Therapist* in each episode. First, we leverage YOLO-v26[6] to detect human bounding boxes in each frame, with a confidence threshold of 0.6. We then manually annotate three bounding boxes for each person appearing in every episode to ensure accurate role localization. For identity assignment, each detected human bounding box is processed by InsightFace[7] to extract face regions and identity embeddings. We

[6] YOLO-v26 is the latest YOLO model with improved detection accuracy and efficiency.

[7] InsightFace is framework for face analysis. We employ its state-of-the-art `buffalo_l` model for face detection and feature extraction

compute the cosine similarity between the extracted face embeddings and those stored in the identity label bank. To ensure annotation reliability, an identity is assigned only when the similarity score exceeds 0.6; otherwise, the identity is labeled as `Unknown` to prevent unreliable assignments from affecting subsequent processing. Notably, the identity label bank is maintained separately for each episode, where only the therapist identity is shared across episodes, as each episode typically contains interactions between one therapist and a fixed client in "In Treatment" TV series. This design further reduces potential identity ambiguity and improves matching robustness. We also manually inspected 100 randomly sampled visual identity assignments and found no incorrect assignments, demonstrating the reliability of this annotation pipeline. For subsequent processing, we extract therapist-consistent clips by retaining only frames containing the therapist identity and masking the facial regions of irrelevant individuals.

B.2 Facial Target Construction

For each valid therapist video clip, we construct the target facial motion as a 52-dimensional ARKit BlendShape sequence $B^T = \{B_t^T\}_{t=1}^T$, where $B_t^T \in [0,1]^{52}$. Inspired by speech-assisted monocular 3D facial reconstruction He et al. (2023), we use vision to anchor identity and expression baseline, and use speech to compensate for lip dynamics that are weakly constrained in monocular video.

Let $M_{t,j}$ denote the MediaPipe-estimated blendshape coefficient at video frame t and channel j, and let $A_{t,j}$ denote the Audio2Face-estimated NVIDIA et al. (2025) coefficient down-sampled to the same frame rate. MediaPipe is run in video mode with face detection and presence thresholds of 0.5, producing 52 ARKit coefficients per valid frame. Audio2Face is driven only by corresponding therapist vocal-only speech clips, where non-therapist intervals are muted before inference. Rather than directly averaging the two modalities, we decompose the final motion into a visual base and an audio-induced residual $B_{t,j}^T = F_{t,j} = M_{t,j} + D_{t,j}$. This decomposition reflects the different error profiles of the two estimators: MediaPipe better preserves poses on the camera plane (x-O-y), including brows, eyes, and non-speech expressions rather than z-axis related poses. While Audio2Face is capable of providing decomposed natural phoneme-synchronous mouth dynamics. Therefore, Audio2Face is used only as a relative motion cue on mouth-related channels. If there is no applicable therapist speech in current clip or the channel is outside the mouth-set, we directly keep the visual estimate, namely $D_{t,j} = 0$.

Motivated by recursive filtering Kalman (1960) and saturation-aware residual updates for bounded dynamical systems Åström & Rundqwist (1989), we update the audio-induced residual with a state-dependent saturation gate. For mouth-related channels, we first compute the Audio2Face temporal increment $\Delta A_{t,j} = A_{t,j} - A_{t-1,j}$, and normalize its scale by the clip-level motion ratio

$$\rho_j = \frac{\mathrm{Std}(M_{:,j})}{\mathrm{Std}(A_{:,j}) + \epsilon}, \tag{5}$$

where $\epsilon = 10^{-5}$ avoids numerical instability for nearly static channels. We then apply a saturation gate based on the current visual base and the accumulated residual:

$$g_{t,j} = \begin{cases} \mathrm{clip}\left(1 - (M_{t,j} + D_{t-1,j}), 0, 1\right), & \Delta A_{t,j} > 0, \\ 1, & \Delta A_{t,j} \le 0. \end{cases} \tag{6}$$

Here $D_{t-1,j}$ represents the audio-induced residual accumulated before the current update. Including it in the gate makes the saturation test depend on the actual pre-update activation $M_{t,j} + D_{t-1,j}$, rather than on the visual estimate alone. The residual is updated as

$$D_{t,j} = g_{t,j} D_{t-1,j} + \alpha \rho_j g_{t,j} \Delta A_{t,j}, \qquad D_{0,j} = 0, \tag{7}$$

where α is the only global modulation-strength parameter and is fixed for all clips. Finally, the target blendshape is

$$B_{t,j}^T = \begin{cases} \mathrm{clip}(M_{t,j} + D_{t,j}, 0, 1), & j \in \mathcal{L}, \\ M_{t,j}, & j \notin \mathcal{L}. \end{cases} \tag{8}$$

Qualitative analysis: This update preserves visual anchoring while allowing speech-driven lip modulation. When $M_{t,j} + D_{t-1,j}$ is small and $\Delta A_{t,j} > 0$, the gate satisfies $g_{t,j} \approx 1$, and the update becomes

$$D_{t,j} \approx D_{t-1,j} + \alpha \rho_j \Delta A_{t,j}, \tag{9}$$

Table 6: The results of human evaluation on target BS compared with source video. UFC and LFC denotes Upper Face Consistency and Lower Face Consistency, respectively.

Method	UFC	LFC	Avg.
Ours	**3.23**	**4.02**	**3.63**
MediaPipe	3.23	2.83	3.03
Audio2Face	1.03	3.55	2.32

allowing Audio2Face to contribute positive mouth motion. When $M_{t,j}+D_{t-1,j}$ is close to the upper bound and $\Delta A_{t,j} > 0$, $g_{t,j} \approx 0$, and thus

$$D_{t,j} \approx 0, \tag{10}$$

which suppresses further positive audio-driven activation and removes accumulated residual bias. For negative audio increments, $g_{t,j} = 1$, so

$$D_{t,j} = D_{t-1,j} + \alpha\rho_j\Delta A_{t,j}, \tag{11}$$

allowing the mouth motion to relax without being blocked by the saturation gate. Therefore, the modulation remains active for unsaturated or closing motions, while over-amplification is naturally limited when the current facial activation is already large.

Quantitative analysis: We further conduct human evaluation on the proposed BS target on 300 samples, compared with Mediapipe and Audio2Face-v3. The results in Table 6 demonstrate that this method is consistent with source video and effective.

C Implementation Details

C.1 Training Settings

Following the training strategy of Qwen3-Omni Xu et al. (2025), we adopt a module-independent training paradigm for **FuSE**, where the Thinker, Talker, and Actor are optimized with their corresponding objectives. Following the main stream fine-tuning methods, all modules are optimized with AdamW Loshchilov & Hutter (2019) using $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a warmup ratio of 0.03, followed by cosine schedule, where the maximum and minimum learning rates are set to 1×10^{-4} and 1×10^{-7}, respectively.

For the *Thinker* and *Talker*, we apply LoRA Hu et al. (2021) to the attention projections `q/k/v/o` and the feed-forward projections `gate_up/down`, with rank 16 and alpha 32, and conduct parameter-efficient fine-tuning. Specifically, we primarily train these two modules on responder sentences, and adopt a global effective batch size of 8 for 2 epochs. Notably, *Thinker* is initialized with Qwen3-Omni, while *Talker* is initialized with Qwen3TTS-12Hz-0.6B. Meanwhile, when LoRA fine-tuning *Talker*, its MTP module is trained together with the same settings, which improves residual acoustic code modeling. While the codec head is kept frozen to avoid substantially changing the original speech generation capability.

For the *Actor* and its VQ-VAE, we train them from scratch with a global batch size of 64, where *Actor* for 2 epochs and VQ-VAE for 100 epochs. The VQ-VAE is a Transformer-based BlendShape52 (BS52) tokenizer, where the encoder and decoder share the same model settings: a hidden size of 512, 6 layers, and 8 attention heads. The encoder takes BS52 coefficients as input and encode into a codebook with size of $N = 512$. Each frame is represented by $G = 8$ latent groups, where each group has a latent dimension of $C = 64$. The *Actor* supports a 100-frame blendshape window, corresponding to 4 seconds at 25 FPS, where 50 frames are used as history. To further support arbitrary commit lengths during streaming inference, we uniformly sample the committed length len_c from 1 to 50 during training.

In addition, we also train a speech reconstructor based on the same settings and objective as the *Talker*. Its task is to reconstruct the reference audio and provide the corresponding speech hidden for *Actor*. In this way, we train this end-to-end model **FuSE** in a module-independent manner.

Table 7: Details of parameter settings for four different scales.

Scale Name	Hidden Size	Layers	Heads	Head Dim.	FFN Size	Params.(M)
50M	512	10	8	64	1536	49.701
fat300M	2048	4	32	64	5120	304.412
deep300M	1024	15	16	64	3584	305.578
1B	2048	15	32	64	4096	1,026.347

C.2 Inference Settings during Evaluation

For text response generation, we apply nucleus sampling to the *Thinker* with temperature 0.7, top-k 50, and top-p 0.9. These sampling hyperparameters are only used for the *Thinker*, and are not applicable to the *Talker* and *Actor*, which generate speech codes and facial motions conditioned on the generated response.

For facial motion inference, we adopt different packet strategies for responder and listener generation. For the responder, we generate an initial packet of 10 blendshape frames, followed by subsequent committed packets, which improves the stability of talking-face motion. For the listener, we use 10-frame packets throughout the interaction, corresponding to 400 ms at 25 FPS, enabling more immediate non-verbal feedback.

We use the following system prompt for Qwen3-Omni during evaluation:

System Prompt for Qwen3-Omni

You are a virtual voice assistant with no gender or age. You are communicating with the user. In user messages, "I/me/my/we/our" refer to the user and "you/your" refer to the assistant. In your replies, address the user as "you/your" and yourself as "I/me/my"; never mirror the user's pronouns–always shift perspective. Keep original pronouns only in direct quotes; if a reference is unclear, ask a brief clarifying question. Interact with users using short (no more than 50 words), brief, straightforward language, maintaining a natural tone. Never use formal phrasing, mechanical expressions, bullet points, or overly structured language. Your output must consist only of the spoken content you want the user to hear. Do not include any descriptions of actions, emotions, sounds, or voice changes. Do not use asterisks, brackets, parentheses, or any other symbols to indicate tone or actions. You must answer users' audio or text questions, and do not directly describe the video content.

D Evaluation Details

We use `GPT-5.5` as a strict automatic judge for text evaluation. For each sample, the judge receives only the pure-text dialogue context, the ground-truth response, and the candidate response, and outputs integer scores from 0 to 5 for empathy, coherence, and emotion consistency.

System Prompt for LLM-based Text Evaluation

You are a strict evaluation judge for embodied empathetic dialogue. Evaluate the candidate response using only the provided pure-text dialogue context, ground-truth response, and candidate response.
Use exactly three metrics:

1. **Empathy**: How well the candidate recognizes, validates, and appropriately responds to the user's emotional state implied by the context.
2. **Coherence**: How well the candidate stays on topic with the context, avoids contradictions, and continues the conversation naturally.
3. **Emotion_consistency**: How well the candidate matches the emotional tone, attitude, and emotional direction expressed by the ground-truth response. Do not require lexical overlap; judge whether the candidate conveys a compatible emotion.

Scoring scale for every metric:
0 = missing, empty, non-response, or impossible to judge from the provided text.

1 = very poor; the candidate fails the criterion or is clearly harmful, incoherent, off-topic, or emotionally incompatible.
2 = poor; the candidate weakly addresses the criterion with major issues.
3 = fair; the candidate is acceptable but generic, incomplete, or partly inconsistent.
4 = good; the candidate mostly satisfies the criterion with only minor issues.
5 = excellent; the candidate strongly satisfies the criterion in a specific, natural, and context-aware way.
Metric-specific rubric:

- empathy
  0–1: ignores or invalidates the user's feelings; 2: acknowledges emotion superficially; 3: responds supportively but generically; 4: validates the feeling and offers relevant support; 5: shows nuanced understanding and responds warmly without overclaiming.
- coherence
  0–1: unrelated, contradictory, or nonsensical; 2: loosely related but misses important context; 3: generally follows the topic with some drift; 4: consistent and natural; 5: tightly grounded in the context and advances the conversation smoothly.
- emotion_consistency
  0–1: opposite or absent emotional tone compared with GT; 2: weakly compatible with clear mismatch; 3: partially compatible; 4: mostly matches GT emotion; 5: closely matches GT emotional direction and intensity.

Return only a valid JSON object with this exact schema:

```
{
  "empathy": integer from 0 to 5,
  "coherence": integer from 0 to 5,
  "emotion_consistency": integer from 0 to 5,
  "rationale": "brief English explanation"
}
```

Do not include markdown, extra keys, or any text outside the JSON object.

The user prompt instantiates the three text fields as `Dialogue context`, `Ground-truth response`, and `Candidate response`. Emp. measures whether the response recognizes and appropriately addresses the user's emotional state; Coh. measures topical consistency and natural continuation; Emo. measures whether the candidate follows the emotional tone and direction of the ground-truth response without requiring lexical overlap.