AuEmoChat: Authentic Emotion Understanding and Rendering for Conversational Speech Synthesis

Zhenqi Jia
jiazhenqi7@163.com
College of Computer Science
Inner Mongolia University
Hohhot, China

Yuan Zhao
zy404nf@163.com
College of Computer Science
Inner Mongolia University
Hohhot, China

Aruukhan
aruukhanres@gmail.com
College of Computer Science
Inner Mongolia University
Hohhot, China

Rui Liu*
imucslr@imu.edu.cn
College of Computer Science
Inner Mongolia University
Hohhot, China

Haizhou Li
haizhouli@cuhk.edu.cn
SRIBD, School of Artificial Intelligence
The Chinese University of Hong Kong
Shenzhen, China

Abstract

Conversational Speech Synthesis (CSS) aims to synthesize speech with human-like emotional expression and contextual consistency in user-agent interactions. Existing CSS methods struggle to render authentic human emotions due to limited predefined emotion label spaces (e.g., seven emotion categories), while redundant multimodal tokens in multi-turn dialogue history interfere with context understanding. To address these issues, we propose **AuEmoChat**, a CSS framework for authentic emotion understanding and rendering. First, we develop *AuEmoCodec*, which learns a discrete authentic emotion token space from large-scale emotional speech via finite scalar quantization, enabling a more authentic emotion representation than limited basic emotion categories. Furthermore, we propose *AuEmoToMe*, an authentic-emotion-guided token merging algorithm that merges redundant tokens in multimodal dialogue history while preserving emotion-relevant context. We integrate it into an autoregressive text-speech model to predict the target authentic emotion token and speech tokens. Finally, we propose *Authentic Emotion Flow Matching*, which renders speech by jointly conditioning on merged dialogue context, target authentic emotion, and acoustic priors. Extensive experiments on the NCSSD-EmCap dataset demonstrate that AuEmoChat outperforms state-of-the-art CSS baselines and generates more expressive and authentic emotional speech. The code and speech demos will be available at: https://github.com/anonymous-css/AuEmoChat.

CCS Concepts

• **Information systems** → **Multimedia content creation**; **Sentiment analysis**; • **Human-centered computing** → **Auditory feedback**.

Keywords

Conversational Speech Synthesis, User-Agent Interactions, Authentic Emotion, Token Merging, Flow Matching

ACM Reference Format:
Zhenqi Jia, Yuan Zhao, Aruukhan, Rui Liu, and Haizhou Li. 2026. AuEmoChat: Authentic Emotion Understanding and Rendering for Conversational Speech Synthesis . In *Proceedings of Proceedings of the 34th ACM International Conference on Multimedia (MM '26).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author.

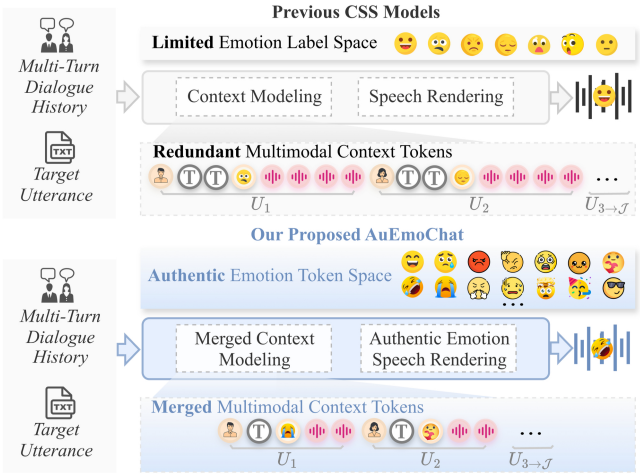

Figure 1: Previous CSS models rely on a limited emotion label space and directly model redundant multimodal context tokens, resulting in limited emotional speech expression. In contrast, AuEmoChat introduces an authentic emotion token space and models the context based on merged tokens, thereby generating more authentic emotional speech.

1 Introduction

Conversational Speech Synthesis (CSS) aims to leverage multimodal dialogue history to synthesize target speech with contextually appropriate affective prosody in User-Agent Interactions (UAI) [16, 17, 30]. In UAI systems, the ability of the agent to generate speech that is appropriate for the dialogue context and conveys human-like emotional expression is vital for enhancing user experience. As UAI becomes increasingly prevalent, CSS has become a crucial component of intelligent interactive systems [35, 42, 53] and plays an important role in embodied intelligence applications such as virtual assistants, voice agents, intelligent robots, smart home devices, and in-vehicle infotainment systems.

Previous CSS studies [8, 16–20, 25–27, 30, 31, 33, 47] mainly rely on a set of predefined emotion labels (including anger, disgust, fear, happiness, sadness, surprise, and neutral) derived from Ekman's basic emotion theory [12], and directly model the multi-turn dialogue history for context modeling. For instance, ECSS [30] encodes multimodal dialogue history into a heterogeneous graph to improve the emotional expressiveness of the synthesized target speech. GPT-Talker [31] converts multimodal information [14, 36, 39] in the dialogue history into discrete token sequences containing semantic and style information, and utilizes an autoregressive model to predict target acoustic tokens for synthesizing emotional speech. Recently, Chain-Talker [17] further generates empathetic captions based on these predefined emotion labels, adopting a three-stage framework that includes emotion understanding, semantic understanding, and empathy rendering to enhance the empathetic expression of the target speech.

Despite the progress of these advanced CSS methods in speech quality and expressiveness, they still face the following limitations, as illustrated in Fig. 1 (top): **1) Limited Emotion Label Space**: In real-world dialogues, emotional expressions are richer and broader [44, 52]. For example, emotions related to "happiness" include excitement, pleasure, and tearful joy. Such a broad emotional space is more capable of reflecting humans' authentic emotions (AuEmo). However, conventional predefined emotion labels cannot distinguish these richer and more diverse emotional states, limiting a model's ability to learn authentic emotional expressions. **2) Redundant Multimodal Context Tokens**: As dialogue turns increase, multimodal token sequences become increasingly long, inevitably introducing a large amount of redundant information [17, 31]. These redundant tokens introduce significant noise into contextual modeling, severely interfering with emotion understanding and high-quality speech generation for the target utterance [49].

To address these issues, we propose **AuEmoChat**, a CSS framework for authentic emotion understanding and rendering. Specifically, we first develop *AuEmoCodec*, which learns a discrete authentic emotion token space from large-scale emotional speech. AuEmoCodec quantizes emotional speech into an AuEmo token in an emotion codebook via Finite Scalar Quantization (FSQ) [37]. The AuEmo token is then used to reconstruct the perceived emotion scores of the source speech for seven basic emotion axes, enabling the model to capture more authentic and fine-grained emotional states. Second, we design an AuEmo-guided token merging (*AuEmoToMe*) algorithm that merges redundant text and speech tokens in multimodal dialogue history. This design reduces the interference caused by redundant tokens during emotion modeling while preserving emotion-relevant contextual information. As a result, the model can better infer the target AuEmo token and speech tokens for the target utterance. Finally, we propose *Authentic Emotion Flow Matching*, which generates emotionally expressive speech conditioned on the merged dialogue context and the predicted AuEmo token. In addition, this mechanism incorporates AuEmo classifier guidance during the flow matching process, enabling the synthesized speech to better align with the target authentic emotion. In summary, the main contributions of this work are as follows:

1) *We propose AuEmoChat*, the first CSS framework dedicated to authentic emotion understanding and rendering.
2) *We develop AuEmoCodec*, which leverages finite scalar quantization to learn an authentic emotion token space from large-scale emotional speech, thereby providing more authentic emotional representations.
3) *We design AuEmoToMe*, an authentic-emotion-guided token merging algorithm that merges redundant multimodal context tokens while preserving emotion-relevant context. Furthermore, *we propose Authentic Emotion Flow Matching* to improve context-consistent emotional speech rendering.
4) Extensive experiments on the NCSSD-EmCap dataset demonstrate that *AuEmoChat significantly outperforms advanced baselines* in both emotional expressiveness and speech quality.

2 Related Works

2.1 Speech Neural Codec

Speech neural codecs convert continuous speech waveforms into discrete tokens, providing the foundation for token-based speech modeling and large language model (LLM)-based audio processing [29, 45]. SpeechTokenizer [50] uses an encoder-decoder architecture with residual vector quantization (RVQ) to model semantic and acoustic information in speech jointly. SemantiCodec [29] adopts a dual-encoder architecture to extract ultra-low-bitrate and semantically rich audio tokens for speech, general sounds, and music. FACodec [22] decomposes speech into content, prosody, timbre, and acoustic-detail subspaces through factorized vector quantization (FVQ), enabling separate generation of different speech attributes. TF-Codec [21] integrates latent-domain predictive coding and learnable time-frequency compression into VQ for low-latency speech coding. CosyVoice [10] extracts semantically rich speech tokens with an Automatic Speech Recognition (ASR)-based self-supervised method, while CosyVoice2 [11] further improves token representation by introducing FSQ. Our AuEmoCodec has the following key distinctions: 1) It quantizes emotional speech into fixed AuEmo tokens rather than semantic or acoustic-related tokens. 2) Unlike conventional codecs that rely on ASR supervision to learn semantic information or reconstruct the original speech waveform, AuEmoCodec is trained to reconstruct quantized AuEmo tokens into a richer emotion perception space by modeling perceived scores across seven basic emotion axes, allowing the tokens to capture authentic human-like emotion.

2.2 Emotion Label Space

The emotion label space is commonly used to quantify emotional representations in speech and serves as a control condition for emotional speech synthesis [18, 30]. Traditional CSS methods learn emotions based on Ekman's basic emotion theory [12], which categorizes emotions into six basic types: anger, disgust, fear, happiness, sadness, and surprise. Emotions that cannot be clearly categorized are usually labeled as neutral. This type of model aligns better with people's intuitive understanding of emotions in daily life. However, its label space is limited and cannot fully cover the rich and subtle emotional states in real dialogues. Plutchik [41] points out that humans can express about 34,000 distinct emotions, which further indicates that traditional discrete models with limited categories cannot fully represent the authentic emotion space. To address the

limitation of label space, Lian et al. [28] introduce open-vocabulary emotion recognition, which expands emotion categories through open-vocabulary emotion labels. However, although it increases the number of categories, it also introduces many semantically similar or synonymous labels, failing to effectively disentangle the emotion representation space and thus increasing modeling difficulty. To construct a more authentic emotion token space, we use AuEmoCodec to learn AuEmo tokens from large-scale emotional speech. Specifically: 1) We build a discrete emotion codebook, where each token represents a distinct authentic emotional state, providing a finer-grained representation than traditional emotion categories. 2) Instead of relying on predefined emotion labels, AuEmoCodec learns the AuEmo token representation directly from large-scale emotional speech.

2.3 Token Merging

Token merging improves model efficiency and performance by combining redundant tokens in multimodal sequences [2]. For instance, ToMe [2] gradually merges similar tokens within the Transformer, enhancing inference speed while largely preserving performance. Token Merger [13] identifies meta tokens representing meaningful image cues and merges similar tokens adaptively, better preserving semantic context and generalizing across vision tasks. FastAST [1] applies token merging to audio spectrograms, improving inference speed while maintaining recognition accuracy. In multimodal dialogue history, speech tokens typically have high bitrate, and long conversations further increase sequence length, making emotional information highly redundant. Redundant tokens dilute model attention to key emotional cues, hindering accurate emotion understanding. Inspired by these works, we propose AuEmoToMe in AuEmoChat: 1) It is the first attempt to introduce token merging into AR-based CSS for modeling multimodal dialogue context. 2) It computes intra-modal similarity among text tokens and speech tokens in the multimodal dialogue history and merges redundant tokens. 3) To preserve emotional information, AuEmoToMe adopts an AuEmo-guided strategy that prioritizes merging redundant tokens while minimally affecting emotion-related key tokens.

3 Task Definition

A dialogue can be defined as a sequence of alternating interactions between the user and the agent, denoted as $\{U_1, U_2, \ldots, U_{\mathcal{J}}, U_{\mathcal{N}}\}$, where $\{U_1, U_2, \ldots, U_{\mathcal{J}}\}$ denotes the dialogue history, and $U_{\mathcal{N}}$ denotes the target utterance to be processed by the agent. For each historical utterance U_i, it contains multiple modalities $\{\mathrm{P}_i, \mathrm{T}_i, \mathrm{S}_i\}$, which represent the speaker, text, and speech of the i-th turn, respectively. For the target utterance $U_{\mathcal{N}}$, the speaker $\mathrm{P}_{\mathcal{N}}$ and text $\mathrm{T}_{\mathcal{N}}$ are given, and the goal is to generate the corresponding speech $\mathrm{S}_{\mathcal{N}}$. CSS based on authentic emotion understanding and rendering needs to address the following points: **1)** How to more accurately represent the authentic emotion of each utterance in the dialogue. **2)** How to effectively infer the authentic emotion token and speech token of the target utterance from a long multimodal dialogue history. **3)** How to synthesize speech with authentic emotion expression that is consistent with the current dialogue context.

4 Methodology: AuEmoChat

Fig. 2 (left) illustrates the overall architecture of **AuEmoChat**. The **Multimodal Dialogue Tokenization** module extracts text, speech, speaker, and AuEmo tokens from the input dialogue. The **AuEmoToMe-based Authentic Emotion Understanding** module incorporates the proposed AuEmoToMe into the LLM to merge redundant text and speech tokens in the dialogue history, based on which the LLM predicts the AuEmo token and speech tokens of the target utterance. The **Merged Context-Aware Authentic Emotion Rendering** module conditions on the merged history tokens, the predicted AuEmo token, and the speech tokens to synthesize authentic emotional speech that is consistent with the context through an authentic emotion flow matching mechanism. In addition, Fig. 2 (right) illustrates the training process of *AuEmoCodec*, which will be described in detail at the end of this section.

4.1 Multimodal Dialogue Tokenization

In this section, we introduce how multimodal dialogue data are processed to extract the corresponding multimodal tokens. The implementation details are as follows.

Text Tokenizer. We adopt a BPE-based Text Tokenizer [15] to tokenize the dialogue text. Specifically, as shown in Eq. 1, the input dialogue text sequence $\mathrm{T}_{1\to\mathcal{N}}$ is processed by the Text Tokenizer to obtain the corresponding text token embeddings:

$$f^{T}_{1\to\mathcal{N},1\to max_t} = \text{Text Tokenizer}(\mathrm{T}_{1\to\mathcal{N}}) \tag{1}$$

where "$1 \to \mathcal{N}$" denotes the dialogue turn indices, and "$1 \to max_t$" denotes the text token indices within each utterance.

Speaker Embedding. We use Speaker Embedding to encode the discrete identities of different speakers. Specifically, the speaker sequence $\mathrm{P}_{1\to\mathcal{N}}$ of the current dialogue is processed by the Speaker Embedding to obtain the corresponding speaker token embeddings:

$$f^{P}_{1\to\mathcal{N}} = \text{Speaker Embedding}(\mathrm{P}_{1\to\mathcal{N}}) \tag{2}$$

Speech Tokenizer. We adopt the Speech Tokenizer proposed in CosyVoice2 [11] to discretize the speech signals in the dialogue history. Specifically, the dialogue history speech sequence $\mathrm{S}_{1\to\mathcal{J}}$ is processed by the Speech Tokenizer to obtain the corresponding speech token embeddings:

$$f^{S}_{1\to\mathcal{J},1\to max_s} = \text{Speech Tokenizer}(\mathrm{S}_{1\to\mathcal{J}}) \tag{3}$$

where "$1 \to max_s$" denotes the speech token indices within each utterance.

AuEmo Tokenizer. To represent authentic emotional expressions in dialogue, we obtain an AuEmo Tokenizer by training AuEmoCodec, as shown in Fig. 2 (right). AuEmoCodec is trained on large-scale emotional speech data to learn a discrete authentic emotion token space via FSQ. We construct an emotion codebook of size 1000, from which 750 tokens are activated after training and used as AuEmo tokens. The detailed training procedure of *AuEmoCodec* is described in Sec. 4.4. The dialogue history speech sequence $\mathrm{S}_{1\to\mathcal{J}}$ is then processed by the AuEmo Tokenizer to obtain the corresponding AuEmo token embeddings:

$$f^{E}_{1\to\mathcal{J}} = \text{AuEmo Tokenizer}(\mathrm{S}_{1\to\mathcal{J}}) \tag{4}$$

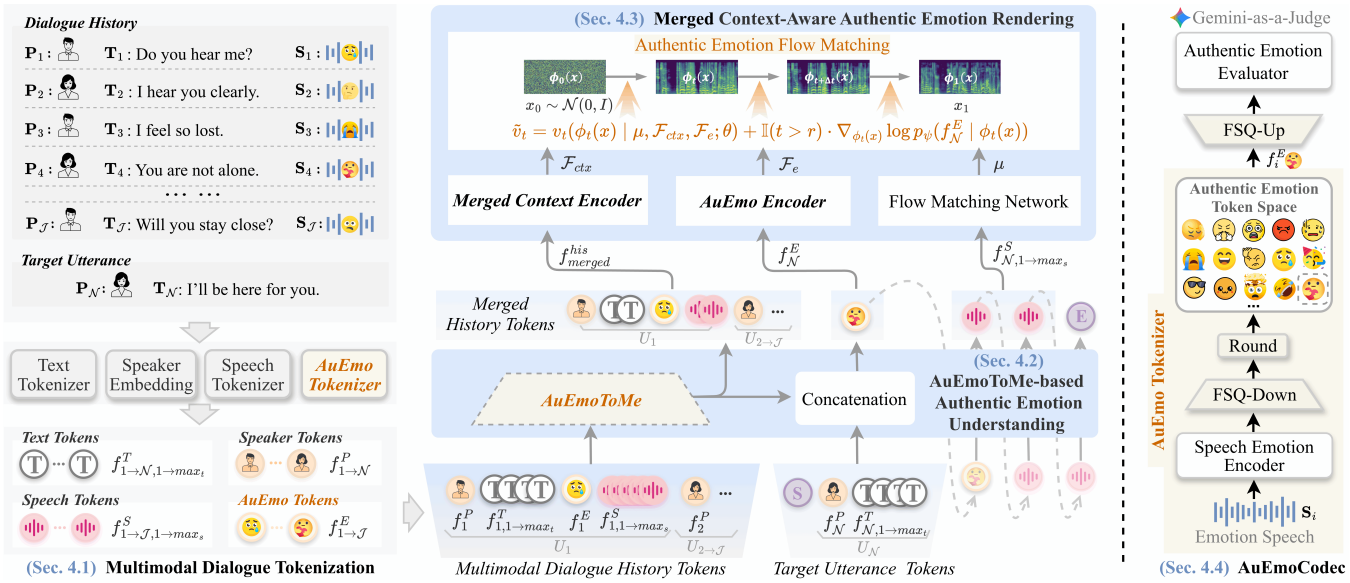

Figure 2: The left side illustrates the overall framework of the proposed AuEmoChat, which includes: Multimodal Dialogue Tokenization, AuEmoToMe-based Authentic Emotion Understanding, and Merged Context-Aware Authentic Emotion Rendering. The right side illustrates the overall framework of the proposed AuEmoCodec.

4.2 AuEmoToMe-based Authentic Emotion Understanding

In AR-based CSS, multimodal dialogue history is typically organized as a unified token sequence for contextual modeling. As dialogue turns increase, text and speech tokens accumulate rapidly, resulting in a long multimodal input sequence with many redundant tokens. To address this issue, we design the AuEmoToMe-based Authentic Emotion Understanding module, which merges redundant emotion-related text and speech tokens in the dialogue history and predicts the target AuEmo and speech tokens from the merged history representation. The module consists of three execution steps:

Multimodal Dialogue Token Sequence Construction. Following the order of dialogue turns, the dialogue history is constructed as the following multimodal dialogue history token sequence:

$$U_{1\to\mathcal{J}} = \{f_1^P, f_{1,1\to max_t}^T, f_1^E, f_{1,1\to max_s}^S, \ldots, f_{\mathcal{J}}^P, f_{\mathcal{J},1\to max_t}^T, f_{\mathcal{J}}^E, f_{\mathcal{J},1\to max_s}^S\} \tag{5}$$

where f_i^P, $f_{i,1\to max_t}^T$, f_i^E, and $f_{i,1\to max_s}^S$ denote the speaker token, text tokens, AuEmo token, and speech tokens of the i-th utterance.

For the target utterance, we only use the speaker token and text tokens as conditional inputs to construct the target utterance token sequence:

$$U_{\mathcal{N}} = \{\text{Ⓢ}, f_{\mathcal{N}}^P, f_{\mathcal{N},1\to max_t}^T\} \tag{6}$$

where Ⓢ denotes the start token of the target utterance.

Authentic Emotion-driven Token Merging (AuEmoToMe). To reduce the interference of redundant information in multi-turn multimodal dialogue history on target emotion modeling, we design AuEmoToMe based on ToMe [2]. It computes cosine similarity between tokens to identify redundant text and speech tokens, and merges them into a more compact merged history tokens sequence. To preserve authentic emotional information during compression, AuEmoToMe employs an AuEmo-guided merging strategy, where the AuEmo token of each historical utterance acts as an emotion anchor, guiding the weighted aggregation of matched text and speech redundant tokens separately. Thus, AuEmoToMe preserves more emotion-relevant information while merging redundant tokens.

Specifically, as shown in Algorithm 1, AuEmoToMe takes the dialogue history token sequence $U_{1\to\mathcal{J}}$ as input and performs token merging for each historical utterance. For each utterance U_i, its text tokens $f_{i,1\to max_t}^T$ and speech tokens $f_{i,1\to max_s}^S$ are separately used to construct token sequences X_i^M, where $M \in \{T, S\}$ denotes text or speech. The AuEmo token f_i^E serves as an emotion anchor. Given the merging ratio ρ, X_i^M is partitioned into a source set A^M and a destination set B^M. Cosine similarities between tokens in A^M and B^M are computed to select the top-k matched pairs for merging. The selected pairs are then weighted by their similarities to the AuEmo anchor and aggregated accordingly. Finally, the merged tokens are restored to their original temporal order, and the merged tokens of all utterances are concatenated to form the merged dialogue history tokens f_{merged}^{his}.

Target AuEmo Token and Speech Token Inference. The merged history tokens are concatenated with the target utterance token sequence:

$$\{f_{merged}^{his}, \text{Ⓢ}, f_{\mathcal{N}}^P, f_{\mathcal{N},1\to max_t}^T\} \tag{7}$$

The model first predicts the AuEmo token of the target utterance $f_{\mathcal{N}}^E$, and then generates the speech token sequence $f_{\mathcal{N},1\to max_s}^S$, which is terminated by an end token Ⓔ. Therefore, the target output sequence can be defined as:

$$\mathcal{Y}_{\mathcal{N}} = \{f_{\mathcal{N}}^E, f_{\mathcal{N},1\to max_s}^S, \text{Ⓔ}\} \tag{8}$$

Algorithm 1: AuEmoToMe

Input: Dialogue history tokens $U_{1\to\mathcal{J}}$, merge ratio ρ
Output: Merged history tokens f^{his}_{merged}
Initialize $f^{his}_{merged} \leftarrow \emptyset$;
for *each utterance* U_i *in* $U_{1\to\mathcal{J}}$ **do**
1. Extract text tokens $f^{T}_{i,1\to max_t}$, speech tokens $f^{S}_{i,1\to max_s}$, and AuEmo token f^{E}_{i}.
2. Construct token sequence X^M_i, where $M \in \{T, S\}$ denotes text or speech.
3. Partition X^M_i into source set A^M and destination set B^M, where $|B^M| = \lfloor |X^M_i|(1-\rho) \rfloor$ and $A^M = X^M_i \setminus B^M$.
4. Compute cosine similarity between tokens in A^M and tokens in B^M.
5. Match each token in A^M to its similar token in B^M.
6. Select the top-k pairs with the highest similarity.
for *each selected pair* $(a_u, b_v) \in (A^M, B^M)$ **do**
Compute cosine similarities $s_a = \mathrm{sim}(a_u, f^E_i)$ and $s_b = \mathrm{sim}(b_v, f^E_i)$.
Normalize them into fusion weights: $w_a = \frac{\exp(s_a)}{\exp(s_a)+\exp(s_b)}$, $w_b = \frac{\exp(s_b)}{\exp(s_a)+\exp(s_b)}$.
Update token representation: $b_v \leftarrow w_a a_u + w_b b_v$.
Update U_i with the merged tokens X^M_i.
Append the updated utterance U_i to f^{his}_{merged}.

The generation process follows a conditional autoregressive formulation:

$$p(\mathcal{Y}_{\mathcal{N}} \mid f^{his}_{merged}, \textcircled{S}, f^{P}_{\mathcal{N}}, f^{T}_{\mathcal{N},1\to max_t}; \Theta) \tag{9}$$

where Θ denotes the model parameters.

To explicitly model the authentic emotion understanding process, the prediction of the target AuEmo token is formulated as:

$$p(f^{E}_{\mathcal{N}} \mid f^{his}_{merged}, \textcircled{S}, f^{P}_{\mathcal{N}}, f^{T}_{\mathcal{N},1\to max_t}; \Theta) \tag{10}$$

and the speech token sequence is then generated conditioned on the predicted emotion token:

$$p(f^{S}_{\mathcal{N},1\to max_s} \mid f^{his}_{merged}, \textcircled{S}, f^{P}_{\mathcal{N}}, f^{T}_{\mathcal{N},1\to max_t}, f^{E}_{\mathcal{N}}; \Theta) \tag{11}$$

4.3 Merged Context-Aware Authentic Emotion Rendering

This module aims to generate conversational speech with authentic emotional expression by jointly conditioning on the merged history tokens f^{his}_{merged}, the target AuEmo token $f^{E}_{\mathcal{N}}$, and the target speech token sequence $f^{S}_{\mathcal{N},1\to max_s}$.

Merged Context Encoder. To obtain a context condition for Flow Matching, we feed f^{his}_{merged} into the Merged Context Encoder. This encoder uses a bidirectional GRU [5] to model the merged history tokens and capture the global dialogue context. The resulting context condition can be formulated as:

$$\mathcal{F}_{ctx} = \text{Merged Context Encoder}(f^{his}_{merged}) \tag{12}$$

AuEmo Encoder. To obtain an authentic emotion condition for Flow Matching, we feed the predicted AuEmo token $f^{E}_{\mathcal{N}}$ into the AuEmo Encoder. This encoder projects the AuEmo token into the conditioning space of the Flow Matching model. The resulting authentic emotion condition can be formulated as:

$$\mathcal{F}_{e} = \text{AuEmo Encoder}(f^{E}_{\mathcal{N}}) \tag{13}$$

Authentic Emotion Flow Matching. Given the predicted target speech token sequence $f^{S}_{\mathcal{N},1\to max_s}$, we first extract its acoustic prior representation, denoted as μ. The context condition $\mathcal{F}_{ctx}$ and the target emotion condition $\mathcal{F}_e$ are then injected into the Flow Matching network together with μ, such that the generation process is jointly constrained by the target speech prior, the merged dialogue context, and the target authentic emotion.

Following conditional Flow Matching, let x_1 denote the ground-truth mel-spectrogram of the target utterance. Let $x_0 \sim \mathcal{N}(0, I)$ denote the Gaussian noise. We define the interpolation path between x_0 and x_1 as:

$$\phi_t(x) = \big(1 - (1-\sigma_{min})t\big)x_0 + t x_1, \quad t \in [0,1] \tag{14}$$

where $\sigma_{\min} = 10^{-6}$, following CosyVoice2 [11]. The corresponding target vector field is:

$$u_t(\phi_t(x) \mid x_1) = x_1 - (1-\sigma_{min})x_0 \tag{15}$$

Conditioned on μ, $\mathcal{F}_{ctx}$, and $\mathcal{F}_e$, the Flow Matching network predicts the vector field:

$$v_t(\phi_t(x) \mid \mu, \mathcal{F}_{ctx}, \mathcal{F}_e; \theta) \tag{16}$$

where θ denotes the model parameters. The main objective is to make the predicted vector field match the target transport direction from noise to the real mel-spectrogram. Therefore, the Flow Matching loss ($\mathcal{L}_{fm}$) is defined as:

$$\mathcal{L}_{fm} = \mathbb{E}_{x_0,x_1,t}\left[\|v_t(\phi_t(x) \mid \mu, \mathcal{F}_{ctx}, \mathcal{F}_e; \theta) - u_t(\phi_t(x) \mid x_1)\|_2^2\right] \tag{17}$$

Although the $\mathcal{L}_{fm}$ learns the acoustic transport path, it does not explicitly constrain the intermediate mel states to preserve the target authentic emotion. To address this issue, we further introduce an AuEmo classifier guidance mechanism. Specifically, an auxiliary AuEmo classifier ψ is applied to the intermediate state $\phi_t(x)$ to estimate the probability that the current mel state matches the target AuEmo token $f^{E}_{\mathcal{N}}$. Based on this classifier, the guided vector field can be written as:

$$\tilde{v}_t = v_t(\phi_t(x) \mid \mu, \mathcal{F}_{ctx}, \mathcal{F}_e; \theta) + \mathbb{I}(t > r) \cdot \nabla_{\phi_t(x)} \log p_\psi\big(f^{E}_{\mathcal{N}} \mid \phi_t(x)\big) \tag{18}$$

where $\mathbb{I}(\cdot)$ is the indicator function, r is a activation threshold set to 0.7, and ψ denotes the pre-trained AuEmoCodec.

The purpose is to guide the intermediate mel state toward a direction that is more consistent with the target authentic emotion. In this way, the optimization trajectory is guided not only by the acoustic transport objective, but also by the target authentic emotion. Notably, we activate this guidance only when $t > r$, because the intermediate state is still close to Gaussian noise when t is small and does not yet contain stable emotion-related spectro-temporal patterns. Applying emotion guidance too early may introduce noisy gradients and weaken the learning of the basic transport path. When $t > r$, the intermediate mel state becomes more structured, and the emotion classifier can provide more reliable guidance.

During inference, the model starts from Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ and solves the following ordinary differential equation:

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x) \mid \mu, \mathcal{F}_{ctx}, \mathcal{F}_e; \theta), \quad t \in [0, 1] \tag{19}$$

The final output x_1 is the generated mel-spectrogram, which is then converted into a waveform by the HiFi-GAN vocoder [23].

4.4 Training Process of AuEmoCodec

AuEmoCodec aims to learn discrete AuEmo tokens from large-scale emotional speech data that are more discriminative and more consistent with real human emotional expression. Specifically, given an emotional speech segment S_i, we first use a Speech Emotion Encoder to extract its emotion representation. We then map this representation into a low-rank quantization space through FSQ-Down and discretize it with a bounded rounding operation. In our implementation, the quantization levels are set to levels = [8, 5, 5, 5]. The resulting discrete code is defined as the AuEmo token:

$$f_i^E = \text{ROUND}(\text{FSQ-Down}(\text{Speech Emotion Encoder}(S_i))) \tag{20}$$

To enable AuEmoCodec to focus exclusively on emotion representation learning, we do not train it to reconstruct the original waveform. Instead, motivated by the fact that authentic human emotion is often expressed as a complex mixture of multiple affective states [44, 52], we use the perceived scores of speech sample S_i over seven basic emotion axes as the reconstruction target. Specifically, the quantized AuEmo token f_i^E is first mapped back to a high-dimensional representation through FSQ-Up, from which the model reconstructs the multi-axis perceived emotion scores, denoted as P_i^{ae}. To obtain supervision targets, we use Gemini-2.5-Flash [6] to annotate the perceived scores of input speech along each emotion axis, denoted as GT_i^{ae}. The prompt used for Gemini-2.5-Flash, together with representative examples of the resulting perceived scores over the seven emotion axes, is provided in Appendix A. AuEmoCodec is then trained by minimizing the mean squared error (MSE) between P_i^{ae} and GT_i^{ae}, which encourages the learned AuEmo tokens to capture emotional nuances and thereby establish a more effective authentic emotion token space for CSS.

5 Experiments

This section presents the experimental setup, baseline and ablation models, and the evaluation metrics used to assess speech quality and emotional expressiveness. Additional experimental details are provided in Appendix B.

5.1 Experimental Setup

We use the open-source NCSSD-EmCap [17] dataset to evaluate the effectiveness of AuEmoChat. This dataset integrates three high-quality dialogue speech datasets: DailyTalk [25], NCSSD [31], and MultiDialog [40]. Specifically, NCSSD-EmCap contains approximately 384 hours of speech data. It comprises 18,580 dialogues with a total of 245,984 utterances, averaging 13.2 turns per dialogue, and covers 25 speakers. Among them, 124,599 utterances are from male speakers and 121,385 utterances are from female speakers. The dataset is divided into training, validation, and test sets with a ratio of 8:1:1, and the same split is used for training both AuEmoCodec and AuEmoChat. For training, AuEmoChat uses 4 NVIDIA A100 GPUs with a batch size of 4 and 8 gradient accumulation steps.

5.2 Baseline and Ablation Models

Baseline Models. To evaluate the effectiveness of AuEmoChat, we compare it with state-of-the-art (SOTA) emotional CSS systems, including **1)** ***BaseCSS*** [16, 25], **2)** ***ECSS*** [30], **3)** ***GPT-Talker*** [31], and **4)** ***Chain-Talker*** [17]. For fair comparison, all baseline models are configured to use the authentic emotion token space.

Ablation Models. To assess the contribution of each component in AuEmoChat, we conduct three groups of ablation studies. **1) Ablations on AuEmo Tokenizer:** We replace AuEmo Tokenizer with three alternative emotion representations to examine the role of the authentic emotion token space in supporting emotion understanding and rendering: ***Abl.1 w/ LimEmo***, which uses a traditional limited emotion label space. ***Abl.2 w/ OVEmo***, which uses an open-vocabulary (OV) emotion label space. ***Abl.3 w/ EmoCap***, which uses emotion captions. **2) Ablations on AuEmoToMe:** We remove AuEmoToMe or its AuEmo-guided merging strategy to investigate its importance for target-utterance emotion understanding and speech token prediction: ***Abl.4 w/o AuEmoToMe*** removes the entire AuEmoToMe module. ***Abl.5 w/o AuEmo-guided Strategy*** removes the AuEmo-guided token merging strategy. **3) Ablations on Authentic Emotion Flow Matching:** We remove different conditioning components to analyze the roles of merged dialogue context and authentic emotion constraints in context-consistent emotional speech rendering: ***Abl.6 w/o FM-AuEmo*** removes the AuEmo condition. ***Abl.7 w/o FM-ACG*** removes AuEmo classifier guidance (ACG). ***Abl.8 w/o FM-Context*** removes merged context condition.

5.3 Evaluation Metrics

We use the following metrics to evaluate each model's performance:

Subjective Metrics: Naturalness-DMOS (**N-DMOS**) [32, 33, 43] assesses the naturalness and overall quality of synthesized speech. Emotion-DMOS (**E-DMOS**) [17, 30] measures the emotional expressiveness and consistency with ground truth. A total of 30 trained, English-proficient evaluators participated in the subjective evaluation. Each evaluator reviewed the dialogue context and listened to each synthesized sample at least three times before rating its naturalness and emotional consistency.

Objective Metrics: We employ Word Error Rate (**WER**) [38], Mel Cepstral Distortion (**MCD**) [3, 24], and Speaker Similarity (**SpkSIM**) [7, 51] to evaluate the quality of the synthesized speech. For emotion expressiveness, Emotion Accuracy (**EmoACC**) evaluates performance under a seven-category basic emotion space using emotion2vec [34], while Authentic Emotion Accuracy (**AuEmoACC**) measures consistency in the authentic emotion token space using the trained AuEmoCodec.

AuEmoCodec Analysis Metrics: To validate the training strategy and hyperparameter design of AuEmoCodec, we employ the following metrics: 1) Tolerance at m axes (**Tol@m**) evaluates the reconstruction accuracy of multi-axis perceived emotion scores while allowing at most m mismatched axes. 2) Strict Emotion Axis Alignment (**S-EAA**) measures the reconstruction accuracy of perceived emotion scores across all emotion axes. 3) Dominant Emotion Axis

Table 1: Subjective (95% confidence interval [48]) and objective results with different comparative models. Blue indicates the best performance.

Methods	N-DMOS (↑)	E-DMOS (↑)	WER (↓)	MCD (↓)	SpkSIM (↑)	EmoACC (↑)	AuEmoACC (↑)
BaseCSS [25] (*ICASSP 2023*)	3.431 ± 0.023	3.299 ± 0.024	27.28	9.359	70.75	45.62	19.12
ECSS [30] (*AAAI 2024*)	3.675 ± 0.022	3.509 ± 0.022	25.10	8.568	72.51	49.85	22.31
GPT-Talker [31] (*MM 2024*)	3.812 ± 0.023	3.654 ± 0.022	20.49	8.061	77.68	57.08	23.45
Chain-Talker [17] (*ACL 2025*)	3.955 ± 0.025	3.756 ± 0.018	15.07	7.686	77.56	57.16	24.12
AuEmoChat (Ours)	**4.171 ± 0.026**	**3.979 ± 0.021**	**9.14**	**6.847**	**78.03**	**61.04**	**28.71**

Alignment (**D-EAA**) evaluates the accuracy of the reconstructed dominant emotion axis. 4) **Usage** measures codebook utilization.

6 Results and Discussion

In this section, we comprehensively evaluate AuEmoChat through baseline comparisons, component ablations, analysis of the AuEmoCodec architecture, and analysis of the context token merging rate. Additional experimental results are provided in Appendix C.

6.1 AuEmoChat *vs.* Baseline Models

Table 1 presents the subjective and objective evaluation results comparing AuEmoChat with SOTA CSS baselines. Overall, AuEmoChat achieves the best performance across all metrics, demonstrating its superiority in both speech quality and authentic emotion modeling. ***For subjective evaluations***, AuEmoChat achieves the highest scores in both N-DMOS (4.171) and E-DMOS (3.979), surpassing the strongest baseline Chain-Talker by 0.216 and 0.223, respectively. This indicates that AuEmoChat generates more natural speech with more accurate and context-consistent emotional expression. ***For objective evaluations***, AuEmoChat consistently outperforms all baselines in both speech quality and emotion expressiveness. Compared with the strongest baseline Chain-Talker, AuEmoChat reduces WER from 15.07 to 9.14 and MCD from 7.686 to 6.847, indicating improved pronunciation accuracy and acoustic fidelity. It also achieves the highest speaker similarity (78.03), showing better preservation of speaker identity. In terms of emotional expression, AuEmoChat improves EmoACC from 57.16 to 61.04 and AuEmoACC from 24.12 to 28.71. The improvement in AuEmoACC further demonstrates the effectiveness of the proposed AuEmoChat in capturing authentic emotional states beyond basic emotion labels. Overall, these results confirm that AuEmoChat effectively addresses the limitations of existing CSS systems by improving authentic emotion understanding, reducing redundant context interference, and generating more expressive and context-consistent speech.

6.2 Ablation Results of Key Components

Table 2 presents the ablation results obtained by removing or replacing key components in AuEmoChat. Overall, all ablation variants perform worse than the full AuEmoChat model across both speech quality and emotion expressiveness metrics, confirming the necessity of each proposed component. ***In Abl.1-Abl.3***, replacing AuEmo Tokenizer with alternative emotion representations leads to clear degradation in emotional expressiveness. This indicates that the learned authentic emotion token space is more effective than limited emotion labels, open-vocabulary emotion labels, and emotion captions, enabling the model to better capture emotional states in conversational speech. ***In Abl.4-Abl.5***, removing AuEmoToMe or its AuEmo-guided merging strategy degrades both speech quality and emotional expressiveness. This suggests that redundant dialogue tokens hinder effective context modeling, whereas AuEmoToMe improves target emotion understanding and speech token prediction by reducing irrelevant information. ***In Abl.6-Abl.8***, removing the AuEmo condition, AuEmo classifier guidance, or merged dialogue context degrades both speech quality and emotional expressiveness. This demonstrates that jointly modeling authentic emotion and merged dialogue context is essential for generating context-consistent and emotionally expressive speech. Notably, Abl.1, which uses traditional limited emotion labels, achieves the worst performance on both E-DMOS and AuEmoACC, highlighting the importance of modeling an authentic emotion. In Abl.8, removing merged dialogue context in flow matching significantly degrades speech quality, with WER, MCD, and SpkSIM reaching the worst values, indicating that contextual information plays a critical role in speech generation.

6.3 Analysis of AuEmoCodec Architecture

In this section, we analyze the key design choices of AuEmoCodec from three aspects: quantization paradigms, codebook capacity, and the judge model, as shown in Table 3. ***For quantization paradigms***, we evaluate Vector Quantization (VQ) [45], Residual VQ (RVQ), Finite Scalar Quantization (FSQ) [37], Residual FSQ (R-FSQ), and Grouped FSQ (G-FSQ). FSQ achieves the best performance on Tol@1, S-EAA, and D-EAA, and shows the most stable overall performance. Although its codebook usage is slightly lower than R-FSQ, it achieves stronger authentic emotion modeling, indicating that FSQ provides a better balance between representation capacity and efficiency. ***For codebook capacity***, as the codebook size increases from 200 to 1000, the performance on Tol@1, S-EAA, and D-EAA consistently improves, indicating that a larger codebook enhances the discrimination ability and representation precision of authentic emotions. However, as the codebook size increases further, codebook usage decreases and performance degrades, suggesting that an excessively large codebook reduces effective utilization and impacts authentic emotion modeling performance. Overall, a codebook size of 1000 achieves the best balance between performance and usage. ***For the judge model selection***, we evaluate Qwen2-Audio [4], Kimi-Audio [9], Qwen3-Omni-Captioner [46], and Gemini-2.5-Flash [6]. Gemini-2.5-Flash achieves the best performance on Tol@1, S-EAA, and D-EAA, demonstrating stronger

Table 2: Subjective (95% confidence interval) and objective results with different ablation models. Blue indicates the best performance, and underlined indicates the worst performance.

Methods	N-DMOS (↑)	E-DMOS (↑)	WER (↓)	MCD (↓)	SpkSIM (↑)	EmoACC (↑)	AuEmoACC (↑)
Ablations on AuEmo Tokenizer							
Abl.1 w/ LimEmo	3.983 ± 0.020	3.736 ± 0.025	9.69	6.921	77.59	58.46	20.38
Abl.2 w/ OVEmo	3.889 ± 0.024	3.859 ± 0.022	9.74	6.971	77.46	59.19	24.57
Abl.3 w/ EmoCap	3.979 ± 0.022	3.765 ± 0.021	9.43	6.936	77.46	58.54	24.15
Ablations on AuEmoToMe							
Abl.4 w/o AuEmoToMe	3.817 ± 0.027	3.750 ± 0.024	9.55	7.274	77.84	57.97	25.72
Abl.5 w/o AuEmo-guided Strategy	3.948 ± 0.022	3.815 ± 0.027	9.37	6.905	77.53	59.81	27.02
Ablations on Authentic Emotion Flow Matching							
Abl.6 w/o FM-AuEmo	3.982 ± 0.023	3.774 ± 0.024	9.15	6.865	77.18	58.33	23.88
Abl.7 w/o FM-ACG	4.008 ± 0.023	3.820 ± 0.022	9.27	6.989	77.71	59.84	25.47
Abl.8 w/o FM-Context	3.863 ± 0.024	3.873 ± 0.019	13.63	7.610	72.61	60.17	26.31
AuEmoChat (Ours)	**4.171 ± 0.026**	**3.979 ± 0.021**	**9.14**	**6.847**	**78.03**	**61.04**	**28.71**

Table 3: Analysis results of key design choices in AuEmoCodec. Blue indicates the best performance, and pink indicates the second-best performance.

Methods	Tol@1	Tol@2	S-EAA	D-EAA	Usage
Comparative Study of Quantization Paradigms					
VQ	35.34	62.55	10.85	47.46	55.10%
RVQ	54.39	81.07	18.53	71.25	61.10%
R-FSQ	53.97	81.69	16.25	66.15	87.30%
G-FSQ	54.46	**82.24**	18.30	70.81	72.10%
FSQ	**55.04**	81.17	**18.61**	**71.77**	75.00%
Comparative Study of Codebook Capacity					
Codebook Size (200)	45.82	79.52	10.27	46.44	76.00%
Codebook Size (600)	46.62	79.75	10.01	46.96	83.16%
Codebook Size (1000)	**55.04**	81.17	**18.61**	**71.77**	75.00%
Codebook Size (1400)	52.26	81.00	17.21	70.50	67.28%
Codebook Size (2000)	52.82	80.99	17.73	70.59	52.10%
Codebook Size (4000)	54.28	**81.21**	18.30	70.30	41.42%
Comparative Study of Different LLM-as-a-Judge Models					
Qwen2-Audio	39.99	74.12	15.17	54.44	73.40%
Kimi-Audio	43.25	78.54	17.25	64.07	74.10%
Qwen3-Omni-Captioner	43.04	**81.51**	17.88	68.51	**75.90%**
Gemini-2.5-Flash	**55.04**	81.17	**18.61**	**71.77**	75.00%

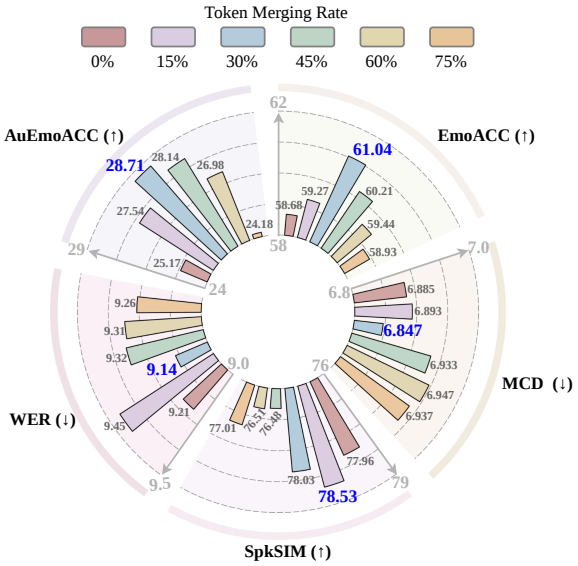

Figure 3: Analysis results of different token merging rates on speech quality and emotion expressiveness.

authentic emotion perception ability. In contrast, Qwen3-Omni-Captioner performs better on Tol@2 and usage. Overall, Gemini-2.5-Flash provides more accurate emotion supervision, which helps learn high-quality emotion representations. Based on the above analysis, we adopt FSQ as the quantization method, set the codebook size to 1000, and use Gemini-2.5-Flash as the judge model.

6.4 Analysis of Context Token Merging Rate

To analyze the impact of the token merging rate in AuEmoToMe, we conduct experiments with merging ratios of 0%, 15%, 30%, 45%, 60%, and 75%. As shown in Fig. 3, we report the performance on AuEmoACC, EmoACC, WER, MCD, and SpkSIM, where the five sectors correspond to different evaluation metrics. Overall, a token merging rate of 30% achieves the best overall performance. Under this setting, the model attains the best results on AuEmoACC, EmoACC, WER, and MCD, while maintaining near-best performance on SpkSIM. ***For emotion metrics***, as the merging ratio increases from 0% to 30%, both AuEmoACC and EmoACC consistently improve. This indicates that moderate token merging effectively reduces the interference of redundant tokens and enhances the model's ability to capture the authentic emotion of the target utterance. When the merging ratio is further increased, the performance degrades, suggesting that excessive merging impairs emotional expressiveness. ***For speech quality metrics***, WER and MCD achieve the best performance at 30%, indicating that appropriate token merging alleviates the interference of redundant context in speech modeling, thereby improving pronunciation accuracy. At higher merging ratios, the performance degrades, further confirming the negative impact of information loss on speech generation. Finally, we adopt

a token merging rate of 30% during inference to achieve the best balance between emotional expressiveness and speech quality.

7 Conclusion and Future Work

In this work, we propose AuEmoChat, a novel CSS framework for authentic emotion understanding and rendering. AuEmoChat introduces AuEmoCodec to learn an authentic emotion token space, AuEmoToMe to merge redundant multimodal context tokens while preserving emotion context information, and Authentic Emotion Flow Matching to generate context-consistent authentic emotional speech. To the best of our knowledge, AuEmoChat is the first CSS system explicitly devoted to authentic emotion modeling. We hope this work can inspire further research on more human-like emotional speech synthesis for conversational agents. In future work, we will extend AuEmoChat to multilingual settings, scale up AuEmoCodec training, and further improve the interpretability of authentic emotion for more comprehensive emotional CSS. More detailed limitations and future works are provided in Appendix D.

References

[1] Swarup Ranjan Behera, Abhishek Dhiman, Karthik Gowda, and Aalekhya Satya Narayani. 2024. Fastast: Accelerating audio spectrogram transformer via token merging and cross-model knowledge distillation. *arXiv preprint arXiv:2406.07676* (2024).

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461* (2022).

[3] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022. V2C: Visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21242–21251.

[4] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759* (2024).

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[7] Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton Van Den Hengel, Jian Yang, and Qingming Huang. 2025. Emodubber: Towards high quality and emotion controllable movie dubbing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 15863–15873.

[8] Yayue Deng, Jinlong Xue, Yukang Jia, Qifei Li, Yichen Han, Fengping Wang, Yingming Gao, Dengfeng Ke, and Ya Li. 2024. Concss: Contrastive-based Context Comprehension for Dialogue-Appropriate Prosody in Conversational Speech Synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10706–10710.

[9] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425* (2025).

[10] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407* (2024).

[11] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117* (2024).

[12] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[13] Zhanzhou Feng and Shiliang Zhang. 2023. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing* 32 (2023), 4156–4169.

[14] Rong Fu, Ziming Wang, Chunlei Meng, Jiaxuan Lu, Jiekai Wu, Kangan Qian, Hao Zhang, and Simon Fong. 2026. Missing-by-Design: Certifiable Modality Deletion for Revocable Multimodal Sentiment Analysis. *arXiv preprint arXiv:2602.16144* (2026).

[15] Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal* 12, 2 (1994), 23–38.

[16] Haohan Guo, Shaofei Zhang, Frank K Soong, Lei He, and Lei Xie. 2021. Conversational end-to-end tts for voice agents. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 403–409.

[17] Yifan Hu, Rui Liu, Yi Ren, Xiang Yin, and Haizhou Li. 2025. Chain-Talker: Chain Understanding and Rendering for Empathetic Conversational Speech Synthesis. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 1988–2003. doi:10.18653/v1/2025.findings-acl.101

[18] Yifan Hu, Rui Liu, Yi Ren, Xiang Yin, and Haizhou Li. 2025. UniTalker: Conversational Speech-Visual Synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10248–10257.

[19] Zhenqi Jia and Rui Liu. 2025. Intra-and inter-modal context interaction modeling for conversational speech synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[20] Zhenqi Jia, Rui Liu, Berrak Sisman, and Haizhou Li. 2025. Multimodal Fine-grained Context Interaction Graph Modeling for Conversational Speech Synthesis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 8863–8869.

[21] Xue Jiang, Xiulian Peng, Huaying Xue, Yuan Zhang, and Yan Lu. 2023. Latent-domain predictive neural speech coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2111–2123.

[22] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100* (2024).

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* 33 (2020), 17022–17033.

[24] Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, Vol. 1. IEEE, 125–128.

[25] Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[26] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. 2022. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7917–7921.

[27] Jingbei Li, Yi Meng, Xixin Wu, Zhiyong Wu, Jia Jia, Helen Meng, Qiao Tian, Yuping Wang, and Yuxuan Wang. 2022. Inferring speaking styles from multi-modal conversational context by multi-scale relational graph convolutional networks. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5811–5820.

[28] Zheng Lian, Haiyang Sun, Licai Sun, Lan Chen, Haoyu Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. 2024. Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark. (2024).

[29] Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. 2024. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing* 18, 8 (2024), 1448–1461.

[30] Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18698–18706.

[31] Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4187–4196.

[32] Rui Liu, Zhenqi Jia, Feilong Bao, and Haizhou Li. 2025. Retrieval-Augmented Dialogue Knowledge Aggregation for expressive conversational speech synthesis. *Information Fusion* (2025), 102948.

[33] Rui Liu, Zhenqi Jia, Jie Yang, Yifan Hu, and Haizhou Li. 2024. Emphasis Rendering for Conversational Text-to-Speech with Multi-modal Multi-scale Context Modeling. *arXiv preprint arXiv:2410.09524* (2024).

[34] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*. 15747–15760.

[35] Michael McTear. 2022. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

[36] Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. 2026. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8791–8800.

[37] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*

(2023).
[38] Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.. In *Interspeech*. 2765–2768.
[39] Jinzhong Ning, Yuanyuan Sun, Bo Xu, Zhihao Yang, Ling Luo, and Hongfei Lin. 2024. Breaking the Boundaries: A Unified Framework for Chinese Named Entity Recognition Across Text and Speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 1250–1260.
[40] Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16334–16348.
[41] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.
[42] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.
[43] Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (2016), 213–227.
[44] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis. In *Interspeech 2023*. 12–16. doi:10.21437/Interspeech.2023-1317
[45] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
[46] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765* (2025).
[47] Jinlong Xue, Yayue Deng, Fengping Wang, Ya Li, Yingming Gao, Jianhua Tao, Jianqing Sun, and Jiaen Liang. 2023. M 2-ctts: End-to-end multi-scale multi-modal conversational text-to-speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[48] Yusuke Yasuda and Tomoki Toda. 2023. Analysis of Mean Opinion Scores in Subjective Evaluation of Synthetic Speech Based on Tail Probabilities. In *Proc. INTERSPEECH 2023*. 5491–5495. doi:10.21437/Interspeech.2023-1285
[49] Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2025. Enhancing multimodal large language models complex reason via similarity computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10203–10211.
[50] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692* (2023).
[51] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. 2024. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM international conference on multimedia*. 7523–7532.
[52] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li. 2022. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing* 14, 4 (2022), 3120–3134.
[53] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009