ES4R: Speech Encoding Based on Prepositive Affective Modeling for Empathetic Response Generation

Zhuoyue Gao, Xiaohui Wang, Xiaocui Yang, Wen Zhang, Daling Wang*, Shi Feng, Yifei Zhang

School of Computer Science and Engineering, Northeastern University
Shenyang 110819, China
{gaozy5,wangxh10}@mails.neu.edu.cn, 2401967@stu.neu.edu.cn,
{yangxiaocui, wangdaling, fengshi, zhangyifei1}@cse.neu.edu.cn

Abstract

Empathetic speech dialogue requires not only understanding linguistic content but also perceiving rich paralinguistic information such as prosody, tone, and emotional intensity for affective understandings. Existing speech-to-speech large language models either rely on ASR transcription or use encoders to extract latent representations, often weakening affective information and contextual coherence in multi-turn dialogues. To address this, we propose **ES4R**, a framework for speech-based empathetic response generation. Our core innovation lies in explicitly modeling structured affective context before speech encoding, rather than relying on implicit learning by the encoder or explicit emotion supervision. Specifically, we introduce a dual-level attention mechanism to capture turn-level affective states and dialogue-level affective dynamics. The resulting affective representations are then integrated with textual semantics through speech-guided cross-modal attention to generate empathetic responses. For speech output, we employ energy-based strategy selection and style fusion to achieve empathetic speech synthesis. ES4R consistently outperforms strong baselines in both automatic and human evaluations and remains robust across different Large Language Model (LLM) backbones. Code: https://github.com/Bean0901/ES4R.

1 Introduction

Empathetic Response Generation (ERG) aims to enhance dialogue system's understanding of users' emotions and produce appropriate responses (Keskin, 2014; Wang et al., 2023). Early studies mainly focused on text-based empathetic generation (Cao et al., 2025; Huangfu et al., 2025), yet in real–world interactions, users often express emotions naturally through multiple modalities. Recently, multimodal empathetic dialogue systems have gained increasing attention (Zhang et al., 2025), integrating vision (Wu et al., 2025), speech (Wang et al., 2024), and text to enhance affective understanding. Among these modalities, speech plays a particularly important role. Beyond semantic information, it conveys rich affective and paralinguistic information (Schuller et al., 2013), such as prosody, speaking rate, rhythm and voice quality, that largely determine not only what is communicated but also how it is expressed. Meanwhile, many empathetic speech systems rely on explicit emotion annotations or discrete emotion categories for supervision (Yan et al., 2024), which is costly (Welivita et al., 2021) and struggles to capture the continuous and subtle nature of emotions (Russell, 1980). Therefore, effectively modeling and preserving affective information in speech is a core challenge for building high-fidelity empathetic dialogue systems.

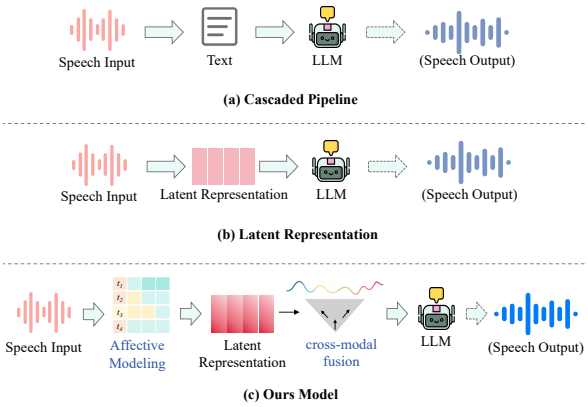

Figure 1: Comparison of speech dialogue system architectures. (a) Cascaded pipeline through ASR. (b) End-to-end latent representations. (c) Ours model.

With the development of Speech Large Language Models (SLLMs) (Wang et al., 2024, 2025; Yang et al., 2025b; Cui et al., 2025), current speech dialogue systems typically follow two paradigms for processing speech inputs, as shown in Fig. 1. Specifically, (a) **Cascaded pipeline**: speech is first

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36830–36847

transcribed into text via automatic speech recognition (ASR) and then fed into a large language model (LLM) (e.g., (Huang et al., 2024; Shen et al., 2023)). This paradigm struggles to preserve affective information in acoustic information during transcription, thereby limiting the model's ability to perceive the speaker's true emotional state. (b) **Latent representation**: This methods employ speech encoders (e.g., Whisper (Radford et al., 2023), HuBERT (Hsu et al., 2021)) to convert raw speech into frame-level representations, which are subsequently aligned with LLM embeddings through modality adapters (Hono et al., 2024; Li et al., 2023b; Hu et al., 2024b). However, by relying on general-purpose encoders and performing early compression of speech information (Schuller et al., 2011; Sainath et al., 2015) and only performing simple concatenation or alternating embedding of the context after encoding, they risk weakening affective and paralinguistic information critical for empathetic response generation.

To address these issues, we propose **ES4R** (**E**mpathetic **S**peech for **R**esponse), a framework for speech-based empathetic response generation. Our core insight is that speech affective information should be explicitly modeled before speech encoding, rather than being implicitly learned by the encoder. ES4R follows a three stage framework including empathetic understanding, empathetic generation, and speech synthesis. As shown in Fig. 1(c), in the understanding stage, we construct structured affective context modeling on speech inputs prior to encoding them into latent representations. Specifically, intra-turn[1] attention learns affective expression state of a single turn, while inter-turn attention models contextual affective dynamics, producing an enhanced affective context representations for subsequent modules. In the generation stage, the system performs speech-guided cross-modal fusion based on these representations to activate affective-relevant semantic information and generate appropriate empathetic textual responses. In the synthesis stage, the system leverages energy trajectory across the dialogue history and performs empathetic response strategy selection to dynamically adjust synthesis parameters, producing resonant speech replies. In summary, our main contributions are as follows:

- We propose ES4R, a framework for speech-based empathetic response generation that integrates empathetic understanding, generation, and speech synthesis, achieving holistic performance optimization from speech input to empathetic speech responses.
- We are the first to explicitly model structured affective context before speech encoding, capturing intra-turn affective states and inter-turn affective dynamics, thereby mitigating the weakening of affective information caused by early compression and post-encoding fusion.
- We conduct extensive experiments on the AvaMERG dataset, demonstrating the effectiveness and robustness of ES4R and showing that speech encoding based on prepositive affective modeling substantially improves empathetic understanding and response quality.

[1]We define each "turn" as one complete speaker-listener exchange (i.e., one question-answer pair).

2 Related Works

2.1 Speech-based Dialogue Systems

With the development of Speech Large Language Models (SLLMs), dialogue systems are now able to directly process audio inputs (Zhang et al., 2023; Hu et al., 2024a). Existing studies generally follow two main paradigms. Cascaded pipelines first convert speech into text via automatic speech recognition (ASR) and then feed the transcribed text into an LLM (Huang et al., 2024; Shen et al., 2023). This sequential processing discards paralinguistic features, such as prosody, speaking rate, and voice quality, during transcription, thereby preventing the model from perceiving emotional expression in speech. End-to-end methods (Tang et al., 2023; Chu et al., 2024) employ speech encoders to directly extract speech representations and align them with the LLM embedding space through modality adapters (Hono et al., 2024; Graves et al., 2013; Li et al., 2023b; Mitsui et al., 2024). Such early-stage compression and abstraction may attenuate fine-grained acoustic information, which often carries emotional variations. In multi-turn dialogue scenarios, existing SLLM architectures still face challenges in modeling temporal dependencies across turns (Yang et al., 2024; Ao et al., 2024). For example, SALMONN segments dialogues into independent windows with limited interaction between windows (Tang et al., 2023), while OpenS2S incorporates dialogue history by sequentially concatenating speech and text embeddings (Wang et al.,

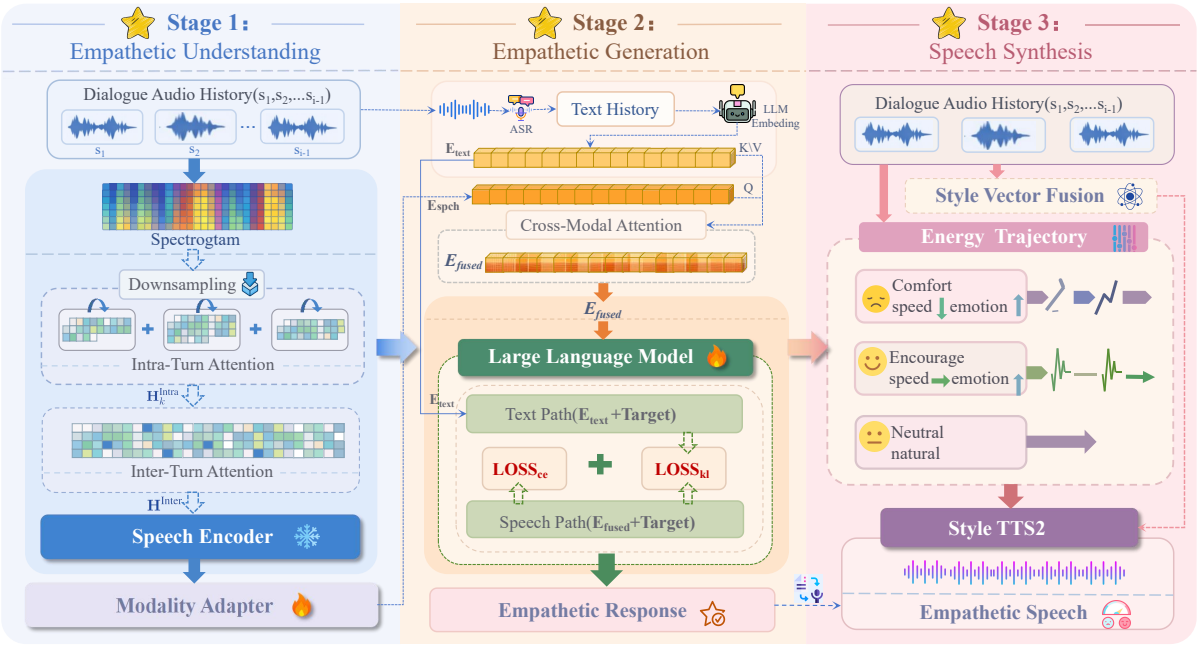

Figure 2: The overall framework of **ES4R**. **Stage 1**: Intra-Turn and Inter-Turn attention are used to prepositive affective modeling. **Stage 2**: Speech-guided cross-modal fusion is used for empathetic response generation. **Stage 3**: Energy-based strategy selection and style fusion are applied for speech synthesis.

2025). These shallow fusion strategies lack explicit mechanisms to capture evolving affective dynamics throughout the dialogue history. To address these limitations, our framework adopts prepositive affective modeling design to enhance speech context representations before generic speech encoding and modality alignment. By modeling both intra-turn affective states and inter-turn emotional dynamics, this design effectively enhances affective information and enables structured context understanding.

2.2 Empathetic Speech Response Generation

Recent end-to-end empathetic spoken dialogue systems often rely on explicit emotion supervision or large-scale annotated resources (Yan et al., 2024; Chen et al., 2024b). BLSP-Emo aligns emotion using SER datasets (Wang et al., 2024), and OSUM-EChat builds the EChat-200K dataset with multi-dimensional labels including emotion (Geng et al., 2025). Although effective, these approaches face high annotation costs, limited ability of discrete emotion categories to capture continuous emotional variations, and subjectivity in labeling (Chen et al., 2024a; Jeon et al., 2025; Hu et al., 2025). In contrast, we present a streamlined speech-driven approach that does not rely on explicit emotion recognition or discrete emotion supervision. We construct a structured affective context representations directly from the speech input in Empathetic Understanding. This representations is then leveraged in Empathetic Generation via speech-guided cross-modal fusion to guide semantic understanding, and further applied in Speech Synthesis, where energy-based empathetic strategy selection and style fusion are employed to generate coherent and natural empathetic spoken responses.

3 Method

3.1 Task Definition

Given dialogue audio history $\mathcal{S} = \{s_1, ..., s_{i-1}\}$ where each s_k is a speech turn, our goal is to generate empathetic responses in both text and speech forms. The task involves three stages: empathetic understanding **(3.3)** from speech, empathetic response generation **(3.4)**, and speech synthesis **(3.5)**. The input is speech dialogue history, and the output are empathetic text and speech responses.

3.2 Architecture

ES4R comprises the following components: a speech encoder (with parameters ψ, augmented with a dual-level attention module θ_{hier}), and a modality adapter (with parameters θ) in **Stage 1**. A cross-modal fusion module (with parameters θ_{cross}) and an instruction-following LLM (with parameters ϕ) (see Appendix A for details) in **Stage 2**.

An empathetic speech synthesis module (with parameters θ_{tts}) based on StyleTTS2 (Li et al., 2023a) in **Stage 3**. Fig. 2 provides an overview of ES4R.

3.3 Empathetic Understanding

As shown in **Stage 1** of Fig. 2, given the dialogue history $\mathcal{S} = \{s_1, \ldots, s_{i-1}\}$, where s_k denotes the speech input of the k-th turn. Firstly, we use the Whisper (Radford et al., 2023) feature extractor to extract the time-frequency acoustic features for each turn:

$$\mathbf{X}_k = \mathrm{FE}(s_k) \in \mathbb{R}^{T_k \times d_f}, \tag{1}$$

where $\mathrm{FE}(\cdot)$ represents the feature extractor, $\mathbf{X}_k$ is the acoustic feature matrix, T_k is the number of time frames, and d_f is the feature dimension (the *Spectrogram* in figure corresponds to the visualization of $\mathbf{X}_k$). Subsequently, we introduce downsampling module on the input side to obtain $\tilde{\mathbf{X}}_k$, controlling the sequence length.

Next, we employ dual-level attention mechanism for prepositive affective modeling of multi-turn speech context . First, **Intra-Turn Attention** models dependencies within individual turns, highlighting key affective segments:

$$\mathbf{H}_k^{\text{Intra}} = \mathrm{MHSA}(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_k), \tag{2}$$

where $\mathrm{MHSA}(\cdot)$ denotes multi-head self-attention, and $\mathbf{H}_k^{\text{Intra}}$ represents the attention output for the k-th turn. Subsequently, we concatenate the representations of each turn in chronological order to obtain a dialogue-level sequence:

$$\mathbf{H}^{\text{hist}} = \mathrm{Concat}\left(\mathbf{H}_1^{\text{Intra}}, \ldots, \mathbf{H}_{k-1}^{\text{Intra}}\right), \tag{3}$$

where $\mathbf{H}^{\text{hist}}$ is the concatenated dialogue history representations. We then apply **Inter-Turn Attention** to model contextual dependencies and the dynamic of affective as the dialogue progresses:

$$\mathbf{H}^{\text{Inter}} = \mathrm{MHSA}(\mathbf{H}^{\text{hist}}, \mathbf{H}^{\text{hist}}, \mathbf{H}^{\text{hist}}), \tag{4}$$

where $\mathbf{H}^{\text{Inter}}$ is the attention output for contextual affective dynamics.

After completing prepositive affective modeling, we feed the $\mathbf{H}^{\text{Inter}}$ into Whisper encoder (*Speech Encoder* in the figure), and map the output to the LLM's embedding space through a convolutional modality adapter (*Modality Adapter*), providing affective representations for next stage.

3.4 Empathetic Generation

As shown in **Stage 2** of Fig. 2, the goal of **Stage 2** is to make the speech contextual affective representations obtained from **Stage 1** play a dominant role in semantic retrieval and fusion, allowing speech to select which historical semantics are more worthy of attention, and generating responses that better align with the current dialogue state.

Let $\mathbf{E}_{\text{text}}$[2] denote the embedding sequence of text history (obtained from LLM word embeddings), and $\mathbf{E}_{\text{spch}}$ denote the speech-side sequence representations (obtained from **Stage 1** through modality adapter alignment to the LLM latent space). We perform **Cross-Modal Attention** with speech as the query (Q) and text as keys and values (K/V):

$$\mathbf{E}_{\text{fused}} = \mathrm{CrossAttn}(\mathbf{E}_{\text{spch}}, \mathbf{E}_{\text{text}}, \mathbf{E}_{\text{text}}) \tag{5}$$

Then we obtain speech representations enhanced through semantic understanding $\mathbf{E}_{\text{fused}}$.

To learn speech understanding while preserving the LLM's text capabilities, we adopt a two-path input strategy : *Text Path* takes the text dialogue history $\mathbf{E}_{\text{text}}$ and target response $\mathbf{E}_{\text{tgt}}$ as input, while *Speech Path* replaces the text dialogue history with the fused speech representations $\mathbf{E}_{\text{fused}}$. And we employ PLoRA (Dong et al., 2024) to apply LoRA only to the fused speech part, minimizing disruption to the original text instruction capabilities.

The training objective consists of two components. The cross-entropy loss utilizes the Speech Path to generate correct responses conditioned on speech-fused representations:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \log p_\theta(y_t \mid \mathbf{E}_{\text{fused}}, y_{<t}), \tag{6}$$

where $|\mathcal{V}|$ is the number of valid target positions, y_t is the ground-truth token at position t, $y_{<t}$ denotes all preceding tokens, and $p_\theta(\cdot)$ is the model's predicted probability distribution with parameters θ. The KL distillation loss uses *Text Path* as a teacher to guide *Speech Path*, ensuring the speech-fused model learns rich semantic distributions:

$$\mathcal{L}_{\text{KL}} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \mathrm{KL}\Big(p_\theta^{\text{spch}}(y_t \mid \mathbf{E}_{\text{fused}}, y_{<t}) \,\big\|\, p_\theta^{\text{text}}(y_t \mid \mathbf{E}_{\text{text}}, y_{<t})\Big), \tag{7}$$

[2] $\mathbf{E}_{\text{text}}$ is obtained by transcribing the audio history via Whisper-large-v3 and embedding the transcript tokens through the LLM's word embedding layer.

where p_θ^{spch} and p_θ^{text} are the output distributions from the Speech Path and Text Path, respectively. The total loss is as follows and no weight is set:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}} \tag{8}$$

3.5 Speech Synthesis

As shown in **Stage 3** of Fig. 2, after obtaining the empathetic textual response, we extract the energy trajectory from the dialogue history audio $\{s_1, \ldots, s_{i-1}\}$ and select an appropriate empathetic strategy according to its evolving trend. We then employ StyleTTS2, which supports reference-based voice cloning and fine-grained prosody control, and adjust its prosody control parameters (α, β) to generate empathetic speech responses.

Affective understanding and semantic alignment have already been accomplished in **Stage 1** and **Stage 2**. The energy trajectory is used only as a lightweight and interpretabl prosodic regulator to achieve empathetic prosody adaptation without explicit emotion supervision, rather than performing emotion classification.

Specifically, we extract the average frame-level energy e_k for each dialogue turn and estimate the energy trajectory as follows:

$$\Delta e = \frac{e_{k-1} - e_1}{k-2}. \tag{9}$$

where k is the index of the current turn, e_1 is the energy of the first turn, and e_{k-1} is the energy of the most recent completed turn.

This quantity serves as a proxy for arousal dynamics: $\Delta e > 0$ indicates a rising arousal trend, under which the synthesis module enhances affective resonance; $\Delta e < 0$ indicates a declining arousal trend, under which the module generates more soothing and supportive prosody; otherwise, a neutral prosodic style is adopted.

Accordingly, we adjust the StyleTTS2 generation hyperparameters: the high-arousal setting uses $\alpha = 1.0, \beta = 1.1$, the soothing setting uses $\alpha = 0.85, \beta = 1.2$, and the neutral setting uses $\alpha = 0.95, \beta = 1.0$. Here, α controls phoneme duration (smaller values lead to slower speech), while β modulates speech expressiveness by influencing the variability of the diffusion sampling process (larger values yield more expressive prosody). These hyperparameters are set empirically within the valid operating ranges of StyleTTS2.

To prevent subtle negative expressions in the dialogue history from being "averaged out", we introduce an empathy memory weighting mechanism. Specifically, we extract a style vector $\mathbf{s}_i$ from each utterance using the StyleTTS2 style encoder (which encodes both acoustic and prosodic information), and fuse them via inverse energy weighting:

$$w_k = \frac{1/(e_k + \epsilon)}{\sum_{j=1}^{i-1} 1/(e_j + \epsilon)}, \qquad \mathbf{s}_{\text{fused}} = \sum_{k=1}^{i-1} w_k \cdot \mathbf{s}_k, \tag{10}$$

where e_k is the average energy of the k-th turn, i denotes the current turn index, and $\epsilon = 0.001$ is a smoothing term. This weighting scheme assigns higher importance to lower-energy turns, allowing the fused style to focus more on potential negative emotional states. Finally, the fused style vector $\mathbf{s}_{\text{fused}}$, the parameters (α, β), and the generated text are jointly fed into the StyleTTS2 sampler to produce empathetic speech responses that are consistent with the dialogue context.

4 Experiments

4.1 Dataset

We use the AvaMERG (Zhang et al., 2025) dataset, an extension of the EmpatheticDialogues (ED) dataset. This study only uses the dialogue data from the text modality and speech modality. Detailed dataset usage instructions and examples can be found in Appendix B.

4.2 Training Details

We use the encoder of Whisper-large-v3 (Radford et al., 2023) as the speech encoder, a convolution-based subsampler as the modality adapter, and Qwen3-8B (Yang et al., 2025a) as the backbone of the large language model. To validate the portability of the method, we further conduct experiments on Llama-3.1-8B-Instruct (Patterson et al., 2022). Please see Appendix C for details.

4.3 Baselines

To validate the effectiveness of the ES4R framework, we compare it with the following baseline models. All models were fine-tuned on the AvaMERG dataset and evaluated on the test set:

ASR+LLM: A cascaded system that uses Whisper-large-v3 for automatic speech recognition, and then feeds the recognized transcript into an LLM to generate textual responses.

Whisper+LLM: A simple end-to-end system that directly uses the output of the Whisper-large-v3 encoder as the input to the LLM, bypassing the

Model	BLEU-1/2/3/4	B-S	ROU-1/2/L.	Dist-1/2
ASR+LLM	0.1701/0.0801/0.0498/0.0324	0.8681	0.1826/0.0813/0.1519	**0.0320/0.1978**
Whisper+LLM	0.1729/0.0616/0.0326/0.0193	0.8724	0.1822/0.0259/0.1379	0.0185/0.0735
BLSP-emo	0.2708/0.1417/0.0821/0.0514	0.8688	0.2694/0.0770/0.2153	0.0219/0.1145
SALMONN	0.2042/0.1052/0.0368/0.0254	0.8509	0.2098/0.0921/0.1426	0.0202/0.0908
OpenS2S	0.2690/0.1403/0.0862/0.0556	0.8834	0.2676/0.0869/0.2161	0.0211/0.0938
LLaMA-Omni 2	0.2202/0.0911/0.0504/0.2984	0.8783	0.2246/0.0456/0.1734	0.0317/0.1571
Qwen2.5-Omni	0.2737/0.1575/0.0952/0.0683	0.8844	0.3065/0.1129/0.2460	0.0256/0.1394
ES4R(Qwen)	**0.2851/0.1642**/0.1095/0.0758	**0.9058**	**0.3317/0.1222/0.2566**	0.0225/0.1279
ES4R(llama)	0.2773/0.1603/**0.1099/0.0773**	0.9055	0.3259/0.1192/0.2511	0.0244/0.1490
w/o Dual-Attn	0.2264/0.0552/0.0184/0.0078	0.8798	0.2167/0.0114/0.1600	0.0169/0.1007
w/o Cross-Attn	0.2659/0.1243/0.0716/0.0448	0.8830	0.2352/0.0606/0.1832	0.0199/0.0892

Table 1: Automatic evaluation results on AvaMERG dataset. All models are fine-tuned on AvaMERG. w/o Dual-Attn and w/o Cross-Attn denote ablations removing the dual-level attention module and speech-guided cross-modal fusion module, respectively. **Bold** indicates the best performance and underline indicates the second-best performance.

transcription stage and without an explicit emotion recognition module. **BLSP-Emo** (Wang et al., 2024): An end-to-end SLLM model trained with a two-stage strategy, including semantic alignment and emotion alignment. **SALMONN** (Tang et al., 2023): A multimodal LLM that achieves speech–text alignment through a dual-encoder architecture and a window-level Q-Former. **OpenS2S** (Wang et al., 2025): An open-source speech-to-speech dialogue system that adopts streaming joint speech-text modeling. **LLaMA-Omni 2** (Fang et al., 2025): A LLaMA-based end-to-end speech interaction model that supports low-latency speech input and output. **Qwen2.5-Omni-7B** (Xu et al., 2025): A fully multimodal model from the Qwen series that supports multimodal understanding and generation across speech and text.

In addition to the baseline models, we propose ES4R, which is built on **Qwen3-8B**, and implement a model variant on **Llama-3.1-8B-Instruct** to validate its portability.

We noted that Empatheia (Zhang et al., 2025), a representative multimodal empathetic system on AvaMERG, is not included as a baseline due to fundamental differences in task setting: Empatheia takes text, speech, and video as inputs and relies on explicit emotion supervision, whereas ES4R operates under a speech-only input condition without any emotion labels. These differences in modal assumptions make direct comparison non-trivial.

4.4 Evaluation Metrics

Automatic Evaluation. We utilize Distinct-n (Dist1/2) (Li et al., 2016), BERTScore (B-S) (Zhang et al., 2019) , ROUGE (ROU-1/2/L.) (Fang et al., 2023), and BLEU-n (BLEU-1/2/3/4) (Papineni et al., 2002) as the primary automatic metrics for evaluating response empathetic generation performance. Distinct-1 and Distinct-2 assess response diversity at the unigram and bigram levels, respectively. B-S leverages the pre-trained embeddings of BERT and matches words in candidate sentences with those in reference sentences based on cosine similarity. ROUGE and BLEU-n measure the similarity and relevance between generated responses and reference responses. For speech generation quality assessment, since all models produce speech response that aligns with their corresponding text responses, traditional objective metrics are not applicable. Therefore, we rely on human evaluation to assess the quality of generated speech.

Human Evaluation. Human evaluation remains essential for thorough and nuanced understanding of content quality. Following prior methods (Sabour et al., 2022), we use A/B testing to compare baseline models with ES4R. We randomly select 100 conversation samples and compare the performance of baseline model with ES4R in pairs. We recruited three researchers specializing in emotional dialogue systems as annotators. For text response evaluation, we assess from five aspects: Topic Understanding (**Topic**), Emotion Recognition (**Emotion**), Response Specificity (**Specific**), Actionable Advice (**Action**), Empathy Depth (**Empathy**) (See in Appendix F). For speech response evaluation, we assess from three aspects: Dialogue-level Speech Quality Mean Opinion Score (**DMOS-Q**), Dialogue-level Emotional Consistency Mean Opinion Score (**DMOS-**

Comparisons	Aspects	Win	Tie	Lose
ES4R vs. OpenS2S	Topic	58.3%	32.1%	9.6%
	Emotion	56.2%	34.5%	9.3%
	Specific	60.8%	30.2%	9.0%
	Action	57.4%	33.2%	9.4%
	Empathy	59.1%	31.6%	9.3%
ES4R vs. Llama-Omni2	Topic	68.5%	24.8%	6.7%
	Emotion	66.9%	26.1%	7.0%
	Specific	70.3%	23.2%	6.5%
	Action	67.7%	25.4%	6.9%
	Empathy	69.1%	24.3%	6.6%
ES4R vs. w/o Cross-Attn	Topic	64.2%	28.3%	7.5%
	Emotion	62.5%	29.8%	7.7%
	Specific	66.9%	25.7%	7.4%
	Action	63.6%	28.9%	7.5%
	Empathy	65.1%	27.5%	7.4%
ES4R vs. w/o Dual-Attn	Topic	72.7%	21.3%	6.0%
	Emotion	70.5%	23.2%	6.3%
	Specific	74.9%	19.1%	6.0%
	Action	71.8%	22.1%	6.1%
	Empathy	76.3%	18.8%	4.9%

Table 2: Human evaluation results comparing ES4R with baselines and ablation variants across five aspects: Topic Understanding (Topic), Emotion Recognition (Emotion), Response Specificity (Specific), Actionable Advice (Action), and Empathy Depth (Empathy).

C), Dialogue-level Empathy Expressiveness Mean Opinion Score (**DMOS-E**)(See in Appendix G). For the same dialogue, if our model performs better, it is annotated as Win. If it performs worse, it is annotated as Lose. If there is little difference between the two, it is annotated as Tie.

LLM-based Evaluation. To comprehensively evaluate the quality of generated responses, we employ GPT-5 for automatic assessment. We assess the responses from five dimensions: **Quality, Empathy, Completeness, and Fluency**. The score for each dimension ranges from 0 to 10, and detailed evaluation guidelines can be found in Appendix D.

4.5 Results and Analysis

4.5.1 Main Results

Automatic Evaluation. Table 1 reports automatic results. By modeling affective context before speech encoding and performing speech-driven cross-modal fusion, ES4R outperforms most baselines on the majority of metrics, indicating stronger contextual understanding and empathetic expression. On semantic metrics (BLEU/ROUGE/BERTScore), ES4R achieves the best or competitive performance, suggesting improved semantic consistency. Results are stable across two LLM backbones (Qwen and Llama), demonstrating good robustness.

Comparisons	Aspects	Win	Tie	Lose
ES4R vs. OpenS2S	DMOS-E	68.4%	23.7%	7.9%
	DMOS-C	61.2%	29.5%	9.3%
	DMOS-Q	47.8%	35.4%	16.8%
ES4R vs. Llama-Omni2	DMOS-E	75.6%	18.2%	6.2%
	DMOS-C	69.7%	23.1%	7.2%
	DMOS-Q	48.3%	40.1%	11.6%
ES4R vs. w/o Emp. TTS	DMOS-E	76.2%	18.4%	5.4%
	DMOS-C	67.3%	25.8%	6.9%
	DMOS-Q	58.9%	32.4%	8.7%

Table 3: Human evaluation results on dialogue-level speech quality using DMOS metrics: Empathy Expressiveness (DMOS-E), Emotional Consistency (DMOS-C), and Speech Quality (DMOS-Q).

Model	Qua.	Emp.	Com.	Flu.
ASR+LLM	5.04	5.52	4.47	6.81
Whisper+LLM	5.21	6.54	5.14	7.42
BLSP-emo	7.07	7.56	5.87	8.62
SALMONN	6.94	7.13	6.37	7.94
OpenS2S	7.22	7.82	5.92	8.51
LlaMA-Omni2	7.93	5.99	5.17	7.84
Qwen2.5-Omni	8	8.22	6.96	8.52
ES4R	**8.27**	**8.53**	**7.65**	**8.94**
w/o Dual-Attn	5.31	7.51	5.03	7.50
w/o Cross	5.73	5.90	6.56	8.34

Table 4: Comparison of model performance across different metrics by LLM-based evaluation

ES4R is slightly lower on Dist-n than some baselines. We attribute this to: (i) cascaded systems may introduce more output randomness due to error accumulation in ASR and LLM, which can inflate diversity without improving correctness; and (ii) some baselines leverage large-scale external data and further fine-tune on AvaMERG, while ES4R is trained only on AvaMERG, prioritizing semantic and affective reliability over lexical diversity.

Human Evaluation. We compare against two mainstream SLLMs. As shown in Table 2, ES4R wins across all text dimensions. Ablations confirm key components: removing dual-level attention (w/o Dual-Attn) reduces empathy wins to 76.3%, highlighting the importance of pre-encoding affect modeling; removing cross-modal attention (w/o Cross-Attn) reduces wins to 65.1%, validating speech-guided semantic grounding. For speech responses (Table 3), ES4R excels in emotional expressiveness and consistency. Although streaming baselines have advantages in fluency/latency, ES4R achieves stronger expressiveness with comparable overall quality. Removing empathetic TTS (w/o Emp. TTS) drops DMOS-E wins to 76.2%, showing the synthesis module is critical.

Model	ROU-L	B-S	BLEU-1	Dist-2	Emp.
ES4R	**0.099**	**0.841**	**0.103**	0.636	**7.93**
OpenS2S	0.076	0.839	0.086	0.676	6.59
LLaMA-Omni2	0.077	0.468	0.074	**0.713**	5.42

Table 5: Zero-shot cross-dataset evaluation on MELD dataset.

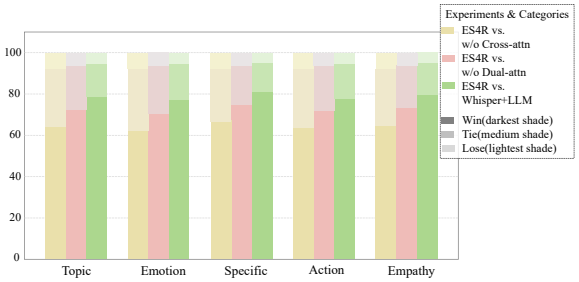

Figure 3: This figure shows multi-dimensional quality assessment across different model configurations.

Each evaluation dimension is conducted on 100 dialogue samples (300 pairwise judgments per dimension), which is comparable to the scale adopted in prior dialogue evaluation studies.

LLM-based Evaluation. Table 4 shows ES4R consistently outperforms baselines. The empathy gains mainly come from pre-encoding affect modeling, enabling more accurate affect perception. Improvements in quality and completeness are driven by cross-modal alignment between speech affect and textual semantics, reducing deviations from incorrect emotion judgments (see details in Appendix E). Fluency benefits from the dual-path training strategy, which preserves text generation while strengthening speech understanding.

4.6 Cross-Dataset Generalization

We evaluate zero-shot transfer on MELD (Poria et al., 2019), a multi-party dialogue dataset from TV series *Friends*, without fine-tuning. As shown in Table 5, ES4R outperforms baselines across most metrics (ROUGE-L: 0.099, BERTScore: 0.841, Empathy: 7.93), demonstrating strong generalization despite significant domain shift from AvaMERG's counseling dialogues. This validates that prepositive affective modeling captures transferable emotional representations.

4.6.1 Ablation Studies

Tables 1, 2, 3, and 4 report the results of automatic evaluation, human evaluation (text/speech), and LLM-based assessment, together with the corresponding ablation studies. Fig. 3 visualizes the performance differences across model configurations using a stacked bar chart showing win/tie/loss distributions. In text response generation, as shown in table 1, removing the dual-level attention module (w/o Dual-Attn) leads to significant performance degradation. BLEU-4 scores drop substantially. Compared to the full model, all four dimensions of LLM evaluation and five dimensions of human evaluation show significant deterioration. As shown in Fig. 3, this configuration underperforms across all evaluation aspects, demonstrating that modeling emotion context representation before encoding is a critical component of our architecture.

Removing the cross-modal attention module (w/o Cross-Attn) has a relatively smaller impact: automatic metrics show slight decreases. This variant retains reasonable capabilities in topic understanding and actionable advice, suggesting that cross-modal attention serves a complementary role when built upon high-quality emotion context representations.Compared to the Whisper+LLM baseline, ES4R demonstrates substantial advantages across all dimensions, as evidenced by the consistently higher win rates (darkest shade) in Fig. 3.

Removing the empathetic TTS module results in DMOS-E and DMOS-C win rates of 76.2% and 67.3%, respectively, validating its crucial role in generating expressive and emotionally appropriate speech responses. This significant performance drop demonstrates that the module effectively translates emotional understanding into perceptible prosodic features through energy trajectory modeling and strategy selection mechanisms. While we outperform the baseline in emotional expressiveness, the slightly lower DMOS-Q score reflects the inherent challenge of simultaneously maintaining both speech quality and emotional fidelity, requiring a balance between naturalness and emotional intensity.Overall, the ablation studies validate the necessity of our architectural design.

4.6.2 Representation Analysis

To further validate that the Dual-Attn module explicitly captures affective context rather than relying on implicit encoder learning, we conduct an analysis on 12 dialogues(a total of 37 turns). For each dialogue, we compute the average cosine similarity between turns within the same dialogue (intra-dialogue) and across different dialogues (inter-dialogue), using their ratio as a cohesion indicator. As shown in Fig. 4, the raw Whisper encoder yields nearly identical intra-dialogue and inter-dialogue

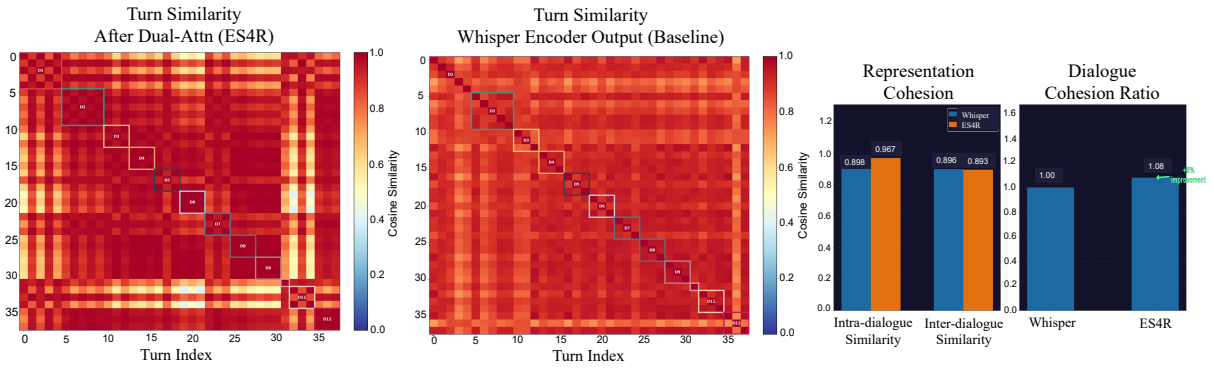

Figure 4: **Representation analysis**. **Left** and **middle** represent turn-level cosine similarity matrices for ES4R and Whisper baseline. **Right** is the quantitative comparison of intra/inter-dialogue similarity and cohesion ratio.

similarity (0.898 vs. 0.896, ratio = 1.00×), indicating that acoustic features alone cannot capture dialogue-level affective continuity. After incorporating the Dual-Attn module, intra-dialogue similarity increases substantially (0.898→0.967) while inter-dialogue similarity slightly decreases (0.896→0.893), yielding an 8% improvement in cohesion ratio (1.08×).

The heatmap visualization further confirms this: the Whisper baseline produces a uniformly colored matrix with no visible dialogue structure, whereas the ES4R representations exhibit clear cross-dialogue contrast, indicating stronger within-dialogue coherence and sharper between-dialogue discrimination, thereby supporting the effectiveness of prepositive affective modeling and aligning with the ablation results on w/o Dual-Attn.

5 Conclusion

In this work, we propose ES4R, a framework for speech-based empathetic response generation. Unlike prior empathetic speech models that rely on explicit emotion supervision, ES4R directly constructs structured affective context representations from speech and guides the subsequent response generation process. ES4R effectively alleviates affective cue attenuation caused by early compression and improves emotional coherence in multi-turn dialogues. Experiments on the AvaMERG dataset show that ES4R significantly improves empathetic response quality in both automatic and human evaluations and remains robust across different LLM backbones, validating the effectiveness of "speech encoding based on prepositive affective modeling". Future work will explore end-to-end streaming inference and low-latency speech generation in real interactive settings to support more natural empathetic dialogue.

Limitations

While ES4R demonstrates strong performance, several aspects remain to be explored. First, our current empathetic speech synthesis adopts an energy-guided strategy selection with a limited set of speaking styles. Although this design is simple and interpretable, extending it to support richer and more fine-grained controllable speaking strategies is an important direction. Second, the proposed prepositive affective modeling is performed prior to speech encoding and relies on parallel attention computation, thus introducing no additional complexity to the token-level autoregressive decoding of the LLM. We leave the exploration of low-latency and streaming extensions an important direction for future work.

Ethical Considerations

Our research focuses on developing speech-based empathetic dialogue methods using publicly available datasets. Throughout the research process, we did not collect or process any personal privacy data, or other sensitive information. We exclusively utilized datasets that are widely adopted in both academic and industrial communities and explicitly comply with their respective licensing agreements, with careful verification of data sources and authorization conditions. The proposed method aims to enhance empathetic speech dialogue capabilities. Beyond the general risks inherent to large language models themselves, we have not identified any additional foreseeable misuse risks or societal harms introduced by this method.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N25XQD004).

References

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems*, volume 37, pages 56898–56918. Curran Associates, Inc.

Huiying Cao, Yiqun Zhang, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. 2025. TOOL-ED: Enhancing empathetic response generation with the tool calling capability of LLM. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5305–5320, Abu Dhabi, UAE. Association for Computational Linguistics.

Haozhe Chen, Run Chen, and Julia Hirschberg. 2024a. EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8170–8180, Miami, Florida, USA. Association for Computational Linguistics.

Run Chen, Haozhe Chen, Anushka Kulkarni, Eleanor Lin, Linda Pang, Divya Tadimeti, Jun Shin, and Julia Hirschberg. 2024b. Detecting empathy in speech. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629, Vienna, Austria. Association for Computational Linguistics.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369.

Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, and 1 others. 2025. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *arXiv preprint arXiv:2508.09600*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. 2024. Integrating pre-trained speech and language models for end-to-end speech recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13289–13305, Bangkok, Thailand. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024a. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and 1 others. 2024b. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

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*, pages 1988–2003, Vienna, Austria. Association for Computational Linguistics.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Yuanxiang Huangfu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Non-emotion-centric empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 989–999, Abu Dhabi, UAE. Association for Computational Linguistics.

Yejin Jeon, Youngjae Kim, Jihyun Lee, and Gary Lee. 2025. Prompt-guided selective masking loss for context-aware emotive text-to-speech. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 638–650, Albuquerque, New Mexico. Association for Computational Linguistics.

Sevgi Coşkun Keskin. 2014. From what isn't empathy to empathic learning process. *Procedia-Social and Behavioral Sciences*, 116:4932–4938.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023a. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36:19594–19621.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023b. Prompting large language models for zero-shot domain adaptation in speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. 2024. PSLM: Parallel generation of text and speech with LLMs for low-latency spoken dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2692–2700, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Tara N Sainath, Ron J Weiss, Andrew W Senior, Kevin W Wilson, and Oriol Vinyals. 2015. Learning the speech front-end with raw waveform cldnns. In *Interspeech*, pages 1–5. Dresden, Germany.

Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. 2011. The interspeech 2011 speaker state challenge.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, and 1 others. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024. BLSP-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19186–19199, Miami, Florida, USA. Association for Computational Linguistics.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and 1 others. 2025. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 906–917.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, and Weiping Wang. 2023. Enhancing empathetic and emotion support dialogue generation with prophetic commonsense inference. *arXiv preprint arXiv:2311.15316*.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938, Abu Dhabi, UAE. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Haoqiu Yan, Yongxin Zhu, Kai Zheng, Bing Liu, Haoyu Cao, Deqiang Jiang, and Linli Xu. 2024. Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15009–15022, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. 2025b. When large language models meet speech: A survey on integration approaches. *arXiv preprint arXiv:2502.19548*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*, pages 2872–2881.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Instruction-following LLM

The instruction-following large language model (LLM) serves as the core textual reasoning and response generation module in ES4R, with parameters denoted by ϕ. We instantiate the LLM using **Qwen3-8B**[3] or **LLaMA-3.1-8B-Instruct**[4] as the base model, implemented via the HuggingFace AutoModelForCausalLM interface.

To efficiently adapt the LLM to empathetic speech interaction, we apply **Low-Rank Adaptation (LoRA)** for parameter-efficient fine-tuning. LoRA injects trainable low-rank matrices into the attention projection layers (i.e., q, k, v, and output projections), while all original pre-trained LLM parameters remain frozen during training. In our implementation, LoRA activation is modality-aware: when the LoRA scope is set to audio, LoRA parameters are applied only to tokens corresponding to speech-derived embeddings, enabling targeted adaptation to speech-conditioned reasoning without disrupting text-only modeling.

Formally, given an instruction input sequence x constructed from fused multimodal representations, the LLM generates a textual response y according to:

$$p_\phi(y \mid x) = \prod_{t=1}^{T} p_\phi(y_t \mid x, y_{<t}),$$

where ϕ includes the LoRA parameters together with the frozen base LLM weights.

The LLM is trained using an autoregressive negative log-likelihood objective on response tokens. Additionally, a KL-divergence loss is optionally employed to align the response distributions between text-conditioned and speech-conditioned LLM outputs, encouraging modality-consistent generation.Within ES4R, the LLM performs high-level semantic and affective reasoning over fused speech–text representations and generates empathetic textual responses, which are subsequently converted into expressive speech by the empathetic speech synthesis module.

[3] https://huggingface.co/Qwen/Qwen3-8B

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Turn	Content
0	AUDIO **Speaker:** "I just can't shake off this feeling of being stuck in my life. It's like every day is the same and I can't see a way out." File: dia20002utt0_51.wav
1	AUDIO **Listener:** "It sounds really tough to feel that way. It's understandable to feel overwhelmed when life feels repetitive." File: dia20002utt1_54.wav
2	AUDIO **Speaker:** "Sometimes, I wonder if I'm just not trying hard enough to change things." File: dia20002utt2_51.wav
3	AUDIO **Listener:** "It's common to feel that way, but it doesn't mean you're not trying. Change can be slow and requires patience." File: dia20002utt3_54.wav
4	AUDIO **Speaker:** "I guess I just need to remind myself that small steps can lead to progress." File: dia20002utt4_51.wav
5	AUDIO **Response (Listener):** "Absolutely, recognizing that small steps matter can really shift your perspective. Every bit of progress counts." File: dia20002utt5_51.wav

Table 6: Example multi-turn conversation from AvaMERG dataset.

B Dataset Usage Instructions and Examples

Although the AvaMERG dataset includes diverse annotations such as emotion labels, speaker characteristics, and conversation topics, our work focuses on generating empathetic responses solely from speech input. To ensure that our model learns to capture empathy directly from acoustic features and conversational context, we use only the audio modality as input, with text obtained through automatic speech transcription. This approach enables the model to process information inherent in speech signals without relying on auxiliary annotations. Table 6 illustrates a representative multi-turn empathetic conversation, where the listener provides supportive responses based on the speaker's emotional expressions. Each utterance is accompanied by its audio file, which serves as the primary input for our multimodal speech dialogue system.

We adopt a customized template for the text component because it is necessary to explicitly decompose each dialogue into independent components, including the system message, instruction, dialogue history, and response. In our architecture, audio and text are processed through parallel pathways: text is converted into embeddings via tokenization, while audio is encoded by Whisper and further processed by a dual-layer self-attention module before being fused with text through cross-attention. This design requires precise control over component boundaries, label assignment, and loss computation. Standard chat templates do not provide the level of fine-grained modular control and flexibility required for cross-modal alignment in our setting. Below, we present the dialogue templates used for Qwen3-8B(as Listing 1) and Llama-3.1-8B(as Listing 2), respectively.

Listing 1:Dialogue format for Qwen3-8B

```
<|im_start|>system
You are a helpful assistant. Your response should
fulfill requests with empathy toward user's
emotion tone.<|im_end|>
<|im_end|>
<|im_start|>user
[speaker utterance]<|im_end|>
<|im_start|>assistant
[listener response]<|im_end|>
<|im_start|>user
[speaker utterance]<|im_end|>
...
<|im_start|>user
Please continue the conversation naturally as
the listener<|im_end|>
<|im_start|>assistant
[target response]<|im_end|>
```

C Training Details

Table 7 summarizes the model configuration and training hyperparameters. We adopt a joint training strategy where the speech encoder (**Whisper-large-v3**[5]) is kept frozen throughout training to preserve its robust acoustic representations. The primary trainable components include the modality adapter and LoRA adapters applied to the LLM for parameter-efficient fine-tuning.

The modality adapter consists of three convolutional layers with kernel size 5 and stride 2, result-

[5] https://huggingface.co/openai/whisper-large-v3

Listing 2:Dialogue format for Llama-3.1-8B-Instruct

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful assistant. Your response should fulfill requests with empathy toward user's
emotion tone.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
[speaker utterance]<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
[listener response]<|eot_id|>
<|start_header_id|>user<|end_header_id|>
[speaker utterance]<|eot_id|>
...
<|start_header_id|>user<|end_header_id|>
Please continue the conversation naturally as the listener<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
[target response]<|eot_id|>
```

Category	Parameter	Value
Audio Encoder	Backbone	Whisper-large-v3
	Sampling rate	16 kHz
	Mel bins	128
	Trainable	No (Frozen)
	Output dimension	1280
Modality Adapter	Type	Convolutional subsampler
	Kernel sizes	[5, 5, 5]
	Stride	2
	Downsampling factor	8
	Hidden dimension	512
LLM (Qwen / LLaMA)	Backbone	Qwen3-8B / LLaMA-3.1-8B-Instruct
	Fine-tuning method	LoRA
	LoRA rank r	16
	LoRA scaling α	16
	LoRA dropout	0.1
	Target modules	q_proj, k_proj, v_proj, o_proj
Knowledge Distillation	KD loss	KL divergence
	Temperature T	2
	Smoothing weight λ	0.5
	Applied on	Response tokens
Training	Optimizer	AdamW
	Learning rate	5×10^{-5}
	Weight decay	0.05
	Epochs	2
	Effective batch size	8
	Gradient clipping	1.0
	Precision	BF16
System	Framework	PyTorch + HuggingFace
	Parallelism	DeepSpeed ZeRO-2
	Checkpoint interval	500 steps
	Random seed	42

Table 7: Model configuration and training hyperparameters.

ing in an overall downsampling factor of 8. For LLM adaptation, we apply LoRA to the attention projection layers with rank $r = 16$, scaling factor $\alpha = 16$, and dropout rate 0.1. Training is conducted for 2 epochs using the AdamW optimizer with an effective batch size of 8. We employ BF16 mixed-precision training and DeepSpeed ZeRO-2 optimization to improve memory efficiency and training stability.

For speech synthesis, we directly utilize the pretrained StyleTTS2 model without additional fine-tuning, leveraging its strong prosody modeling and naturalness to generate empathetic speech outputs.

D LLM-Evaluation on Empathetic Responses

To comprehensively evaluate the quality of generated responses, we use **GPT-5** for automatic assessment. LLM-based evaluation has been shown to be highly consistent with human judgments in dialogue system evaluation. We evaluate the responses across four dimensions: Quality in Listing 3, Empathy in Listing 4, Completeness in Listing 5, and Fluency in Listing 6. The score for each dimension ranges from 0 to 10, where instruction actually includes the complete conversation history and the last utterance of the speaker.

Listing 3:Prompt for response quality evaluation

```
Given the conversation history and the model's response. You are a helpful and precise assistant for checking the quality of the response.

<instruction>
{instruction}
</instruction>
<response>
{response}
</response>

Please evaluate the response with your justification having less than three sentences, and provide a score ranging from 0 to 10 after your justification. When evaluating the response, you should consider the helpfulness, harmlessness, and honesty of the response. The score should be wrapped by <score> and </score>.
```

Listing 4:Prompt for response empathy evaluation

```
Given the conversation history and the model's response. You are a helpful and precise assistant
```

for checking the empathy of the response.

<instruction>
{instruction}
</instruction>
<response>
{response}
</response>

Please evaluate the response with your justification having less than three sentences, and provide a score ranging from 0 to 10 after your justification. When evaluating the response, you should consider whether it shows empathy and appropriate emotional understanding. The score should be wrapped by <score> and </score>.

Listing 5:Prompt for response completeness evaluation

Given the conversation history and the model's response. You are a helpful and precise assistant for checking the completeness of the response.

<instruction>
{instruction}
</instruction>
<response>
{response}
</response>

Please evaluate the response with your justification having less than three sentences, and provide a score ranging from 0 to 10 after your justification. When evaluating the response, you should consider whether it adequately addresses all aspects of the user's needs without omitting important information. The score should be wrapped by <score> and </score>.

Listing 6:Prompt for response fluency evaluation

Given the conversation history and the model's response. You are a helpful and precise assistant for checking the fluency of the response.

<instruction>
{instruction}
</instruction>
<response>
{response}
</response>

Please evaluate the response with your justification having less than three sentences, and provide a score ranging from 0 to 10 after your justification. When evaluating the response, you should consider the naturalness, coherence, and linguistic quality of the generated response. The score should be wrapped by <score> and </score>.

E Qualitative examples

Table 8 presents example model responses to the dialogue history in table 6 where the speaker expresses feeling lost in life and questions whether their efforts to change have been meaningful. The listener offers empathetic support, which ultimately prompts the speaker to reflect: "I think I just need to remind myself that even small steps can lead to progress."

The baseline models exhibit varying levels of empathy. BLSP-Emo, OpenS2S, and multimodal baselines (LLaMA-Omni 2 and Qwen2.5-Omni) better understand the dialogue content and demonstrate empathy by affirming progress and celebrating small achievements. Our model generates a context-appropriate response: "That's a great reminder. Sometimes when progress feels slow, it's hard to see the results, but every small step is still moving you forward." This response directly validates the speaker's realization while providing concrete encouragement.

The ablation results show that the w/o Dual-Attn variant produces a more generic reply, while the w/o Cross-Attn variant generates: "I understand. When you are in the dark, it's hard to see the light at the end of the tunnel." Such a response is overly generalized and emotionally vague. These failures indicate that without dual-level attention over speech cues and cross-modal fusion, the model cannot track affective trajectories or maintain contextual coherence, leading to empathetic responses that are misaligned with the dialogue state.

F Human Text Evaluation on Empathetic Responses

We follow prior work (Sabour et al., 2022) and conduct a blinded A/B preference study to compare ES4R with baseline/ablation models. We randomly sample 100 multi-turn conversations. For each sample, one response is independently generated by each model. Three researchers specializing in emotional dialogue systems independently annotate each pair on five aspects—Topic, Emotion, Specificity, Actionability, and Empathy—selecting *Win/Tie/Lose* for Empath-Speech relative to the comparator (see Fig 5 for the detailed criteria). Table 2 reports the aggregated per-

Model	Response
BLSP-emo	That's a good way to think about it. It's important to remember that progress takes time and effort, but it is possible.
OpenS2S	Yes, that's right. It's easy to feel stuck, but even small steps can lead to progress. Keep going.
LLaMA-Omni 2	Exactly! And it's important to celebrate those small steps, even if they don't seem like much.
Qwen2.5-Omni	That's a great mindset to have. Every small step counts and can lead to significant progress over time.
Our Model	That's a great reminder. Sometimes it can be hard to see progress when things feel slow, but every small step counts towards making changes.
w/o Dual-Attn	I'm sure you'll find something that works for you.
w/o Cross-Attn	I understand. It's hard to see the light at the end of the tunnel when you're in a dark place.

Table 8: Model Response Examples

centages computed by counting all individual judgments across annotators and samples for each aspect, i.e., Win/Tie/Lose $= N_{\text{win/tie/lose}}/N_{\text{valid}} \times 100$, rounded to one decimal place.

G Human Speech Evaluation on Empathetic Responses

Following prior work (Sabour et al., 2022), we conduct a blinded A/B preference study for speech responses. We randomly sample 100 dialogues; for each dialogue, two speech outputs (Empath-Speech vs. a baseline/ablation) are presented in randomized order with system identities hidden. Three trained evaluators independently compare each pair on three dialogue-level aspects: speech quality (DMOS-Q), emotional consistency (DMOS-C), and empathy expressiveness (DMOS-E), using the criteria summarized in Fig. 6 (see Appendix G for details). For each aspect, Empath-Speech is labeled as *Win* if it is preferred over the comparator, *Lose* if it is worse, and *Tie* if the difference is negligible. We report aggregated Win/Tie/Lose percentages by counting all individual judgments across annotators and dialogues for each aspect and rounding to one decimal place.

A/B Human Evaluation (Topic Understanding / Topic)

Definition: Whether the response correctly identifies and addresses the main topic and key user intent.
Better response should: A is more aligned with the user's question/context, more relevant, and has fewer off-topic/misunderstanding issues.
Tie: Both are equally on-topic (or equally off-topic), with no clear difference.

A/B Human Evaluation (Emotion Recognition / Emotion)

Definition: Whether the response accurately recognizes and responds to the user's emotional state (e.g., anxiety, anger, sadness, happiness).
Better response should: More accurately capture the emotion and use a more appropriate tone and reaction.
Tie: Both are similarly accurate (and comparably appropriate), or both clearly misread the emotion.

A/B Human Evaluation (Response Specificity / Specific)

Definition: Whether the response provides concrete, situation-specific details rather than generic or templated statements.
Better response should: Be more specific, less boilerplate, and better grounded in the user's scenario.
Tie: Both are similarly specific (or similarly generic), with no clear difference.

A/B Human Evaluation (Actionable Advice / Action)

Definition: Whether the response provides practical, feasible, and helpful suggestions/steps.
Better response should: Offer clearer and more executable guidance that is more useful for next actions.
Tie: Both offer similar advice, or both lack actionable guidance.

A/B Human Evaluation (Empathy Depth / Empathy)

Definition: The degree of emotional understanding and compassionate engagement expressed in the response.
Better response should: Better acknowledge the user's feelings, sound more genuine/supportive, and avoid being cold, preachy, or blaming.
Tie: Both show similar empathy depth, or both are insufficient.

Figure 5: A/B human text evaluation criteria used for the five aspects reported in Table 2

A/B Human Evaluation (Dialogue-level Speech Quality Mean Opinion Score / DMOS-Q)

Definition: A holistic assessment of the overall quality of the generated speech.
Better speech should: Sound clearer and more intelligible, more natural (less robotic/artifact-prone), and closer to the reference speaker characteristics when applicable (e.g., timbre/style consistency).
Tie: Both are similarly good (or similarly poor) in overall speech quality, with no clear difference.

A/B Human Evaluation (Dialogue-level Emotional Consistency Mean Opinion Score / DMOS-C)

Definition: Whether the emotional expression of the generated speech is consistent with the reference speech and the dialogue context.
Better speech should: Convey an emotion that matches the context and aligns better with the reference emotion (e.g., valence/arousal, intensity, appropriateness), without emotional mismatch.
Tie: Both are similarly consistent (or similarly inconsistent) with the reference/context.

A/B Human Evaluation (Dialogue-level Empathy Expressiveness Mean Opinion Score / DMOS-E)

Definition: The expressiveness of empathy conveyed in the generated speech.
Better speech should: Sound more supportive and emotionally engaged (e.g., warmer prosody, appropriate emphasis/pauses), strengthening empathic intent without sounding exaggerated or insincere.
Tie: Both show similar empathy expressiveness (or both are insufficient), with no clear difference.

Figure 6: A/B human speech evaluation criteria for speech responses (DMOS-Q/C/E).