MEC^2-TT: Multimodal Emotion Consistency Correction and Trajectory Tracking for Empathetic Dialogue Generation

Anonymous Author(s)
Submission Id: 4096

Abstract

Multimodal empathetic dialogue generation aims to produce emotionally nuanced and compassionate responses. However, existing multimodal empathetic response generation (MERG) methods primarily focus on the current query, overlooking the emotional dynamics across dialogue history. Moreover, they lack effective mechanisms to handle cross-modal emotional inconsistencies that frequently arise in real-world interactions. To address these limitations, we propose **MEC^2-TT**, a novel framework that integrates **M**ultimodal **E**motion **C**onsistency **C**orrection and **T**rajectory **T**racking, enabling more coherent and context-aware empathetic responses. Specifically, an Emotion Consistency Correction module is first introduced to mitigate cross-modal inconsistencies via cross-modal alignment. Afterwards, we design an Emotion Trajectory Tracking module to capture the temporal evolution of user emotions throughout the dialogue. Finally, we leverage large language models with structured reasoning to enhance empathetic response generation with improved empathetic understanding. Comprehensive experiments on the AvaMERG and MELD benchmarks demonstrate that our method outperforms existing state-of-the-art approaches, particularly in terms of emotional response accuracy.

CCS Concepts

• **Human-centered computing** → **Interaction techniques**.

Keywords

Dialogue generation, Empathetic response, Multimodal, Large language model

1 Introduction

Empathetic dialogue generation, which aims to produce responses that accurately reflect and appropriately address users' emotional states, has emerged as an important research direction in affective computing [25]. Its ability to generate emotionally aware and contextually coherent responses holds significant promise for real-world applications, including mental health support, social companionship, and human-computer interaction.

Early work in empathetic response generation (ERG) [19, 28] primarily focused on textual inputs, leveraging natural language understanding to infer emotional cues from conversational context. However, human emotional expression is inherently multimodal. People naturally convey their feelings through language, vocal tone, and facial expressions, and these signals form a richer and more accurate picture of their true emotional state than text alone [38]. This has motivated the emergence of multimodal empathetic response generation (MERG), which integrates textual, acoustic, and visual signals to achieve more comprehensive emotion understanding and more empathetic response generation. Fig 1 illustrates the typical pipeline of a MERG system, where emotional cues are extracted from video, audio, and text modalities, then integrated into the dialogue model to facilitate emotion understanding and reasoning, ultimately producing responses with empathetic awareness.

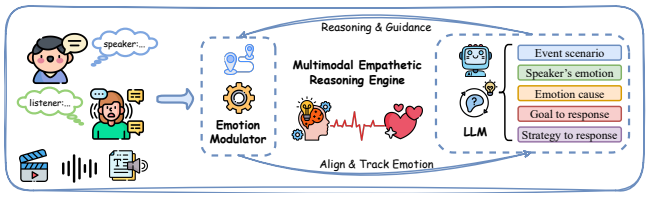

Figure 1: Illustration of the Multimodal Empathetic Reasoning process, comprising an Emotion Modulator for aligning and tracking user emotions from multimodal signals, and an LLM-driven reasoning module for generating multi-step empathetic guidance toward the final response.

Despite significant progress in MERG, three critical challenges remain unresolved. First, emotion representations extracted from each modality inherently carry modality-specific biases, as each modality captures only a partial and potentially skewed view of the user's emotional state from its own perspective [11]. Directly feeding these biased representations into the large language models (LLMs) hinders their ability to form a coherent understanding of the user's emotion, ultimately degrading the quality of generated responses. Second, existing approaches predominantly condition response generation on instantaneous emotional states, overlooking the temporal dynamics that unfold across dialogue history. However, a user's emotional trajectory provides critical context that a single-turn snapshot fails to capture [9]. For instance, a user experiencing persistent sadness requires a different empathetic strategy than one who abruptly shifts from joy to distress, even if their current surface states appear similar. Neglecting such temporal evolution risks producing responses that are locally plausible but globally incoherent. Third, existing approaches typically rely on end-to-end generation paradigms that directly map multimodal representations to responses [3, 31, 32, 40]. This black-box process lacks an explicit, intermediate reasoning stage. Without structurally breaking down the underlying causes of user emotions and strategizing the appropriate empathetic reaction step-by-step, models often fail to translate rich multimodal cues into deep, cognitive empathy, resulting in generic or superficial replies.

To address these challenges, we propose MEC^2-TT, a novel framework that integrates multimodal emotion consistency correction and emotion trajectory tracking, enabling more coherent and context-aware empathetic responses. The proposed MEC^2-TT comprises three dedicated components. The first is an Emotion Consistency Correction (ECC) module, which rectifies inter-modal discrepancies to produce reliable emotion representations. The

second is an Emotion Trajectory Tracking (ETT) module, which captures the temporal evolution of user affect. The third is an Empathetic Chain of Thought (ECoT) mechanism, which guides the large language model to explicitly reason over emotional states, underlying causes, and response strategies.

Specifically, the Emotion Consistency Correction module mitigates inherent cross-modal inconsistencies by dynamically aligning and refining modality-specific representations. Instead of simply concatenating multimodal features, this module explicitly models the distributional discrepancies among modalities and minimizes these biases, yielding a unified emotion-consistent representation. Furthermore, the Emotion Trajectory Tracking module is designed to capture the temporal dynamics of user emotions. By modeling the sequence of refined multimodal features from previous dialogue turns, it constructs a temporal representation that reflects both short-term emotional fluctuations and long-term emotional trends throughout the conversation, providing richer contextual cues for response generation. Finally, we introduce the ECoT mechanism to integrate the corrected emotion representations and trajectory information into a multimodal large language model (MLLM). Instead of directly decoding a response, the model first unpacks the user's emotional state and its underlying triggers. This intermediate reasoning step naturally dictates the appropriate empathetic strategy. This ensures the final generated responses are not only locally plausible but also globally coherent with the emotional flow of the dialogue.

In summary, the main contributions of this work are as follows:

- We propose a novel MERG framework equipped with an Emotion Consistency Correction module that explicitly models and mitigates cross-modal emotional discrepancies, enabling a more reliable perception of users' emotional cues.
- We introduce a dual-level contextual understanding approach by combining Emotion Trajectory Tracking with an ECoT mechanism. This allows the model to capture the global temporal evolution of emotions and perform explicit reasoning for deeper empathetic response generation.
- Extensive experiments conducted on large-scale multimodal benchmarks, including AvaMERG and MELD, demonstrate the superiority of the proposed framework, with notable improvements in both emotion understanding accuracy and the empathetic quality of generated responses.

2 Related work

Empathetic dialogue generation is a critical component in human–computer interaction, with its research scope evolving from text-based semantic understanding [18, 19, 25, 28] to multimodal emotional fusion [32, 35, 42], aiming to enable more natural and emotionally intelligent conversational experiences.

2.1 Text-Only Empathetic Dialogue Generation

Empathetic dialogue generation has traditionally been studied within the textual modality, where early approaches focused on modeling semantic relevance and emotional cues from user utterances [18]. Early approaches to empathetic dialogue generation primarily relied on recurrent neural architectures, including RNNs, GRUs, and LSTMs, to model sequential dependencies in dialogue [4, 8, 20]. Rashkin et al. [26] propose EMPATHETICDIALOGUES, a novel dataset of 25k conversations grounded in emotional situations. This progress laid a solid foundation for the rapid development of empathetic dialogue generation. Li et al. [14] propose an emotional cross-attention mechanism to learn the emotional dependencies from the emotional context graph. Fu et al. [7] propose a novel emotion correlation enhanced empathetic dialogue generation framework, which comprehensively realizes emotion correlation learning, utilization, and supervising. With the rapid advancement of large language models, their ability to generate coherent and empathetic dialogue has been substantially improved [23]. Furthermore, Chain of Thought reasoning further enhanced the ability of large language models to perform fine-grained emotional reasoning [15]. However, visual and audio modalities also contain rich emotional information, which text-only approaches fail to fully exploit in empathetic dialogue generation.

2.2 Multimodal Empathetic Dialogue Generation

The rapid advancement of multimodal large language models has provided momentum for progress in this field [2, 13, 24, 34], particularly demonstrating significant potential in integrating textual, acoustic, and visual information. Zhang et al. [39] present a large-scale high-quality benchmark dataset, AvaMERG, which extends traditional text-based ERG by incorporating authentic human speech audio and dynamic talking-face avatar videos. The dataset encompasses diverse avatar profiles and broadly covers various topics of real-world scenarios. E3RG [17] achieves natural, emotionally rich, and identity-consistent responses without additional training by integrating advanced expressive speech and video generative models. EmpathyEar [6] provides users with responses that achieve a deeper emotional resonance, closely emulating human-like empathy. Hazarika et al. [9] uses a multimodal approach comprising audio, visual, and textual features with gated recurrent units to model past utterances of each speaker into memories. Existing approaches have largely overlooked the heterogeneity of multimodal emotional features. Motivated by this observation, we design an emotion consistency correction module to mitigate cross-modal inconsistencies and propose an emotion trajectory tracking module to capture the evolution of user emotional states.

3 Methodology

In this section, we detail our MEC^2-TT framework for multimodal empathetic response generation. We describe the design and implementation of its three key components: the Emotion Consistency Correction module, the Emotion Trajectory Tracking module, and the Empathetic Chain of Thought mechanism.

3.1 Task definition

Let a dialogue history of T turns be denoted as:

$$\mathcal{H} = [(u_1, r_1), (u_2, r_2), \ldots, (u_{T-1}, r_{T-1}), u_T], \tag{1}$$

where u_t and r_t denote the user utterance and system response at turn t, respectively. Each user utterance $u_t = \{u_t^v, u_t^a, u_t^l\}$ consists

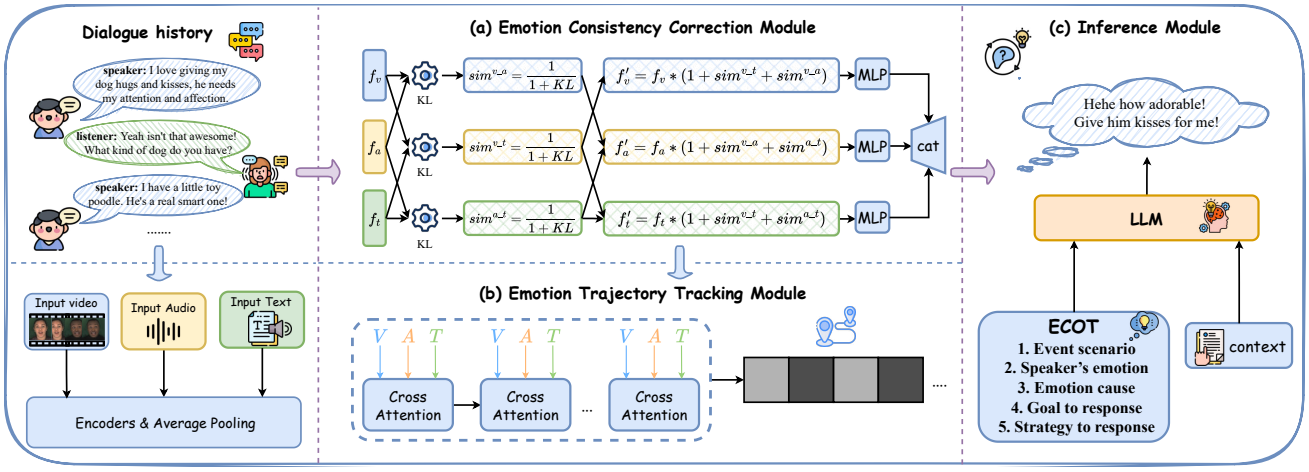

Figure 2: The overall architecture of our method. (a) The Emotion Consistency Correction Module aligns multimodal features via similarity modulation. (b) The Emotion Trajectory Tracking Module captures emotional dynamics across the dialogue history. (c) The Inference Module leverages an LLM with Empathetic Chain-of-Thought (ECOT) to generate the final response.

of visual, acoustic, and linguistic inputs, while the system response r_t is purely textual.

Given the dialogue history $\mathcal{H}$, MERG aims to generate a target response r_T that is emotionally aligned with the user's state and exhibits empathetic behavior. The generation process is formulated as:

$$r_T = argmax_r P(r_T|\mathcal{H};\theta), \tag{2}$$

where $P(r_T|\mathcal{H};\theta)$ is the conditional probability distribution over responses defined by the MERG model, and θ denotes the trainable parameters.

3.2 Framework Overview

As illustrated in Fig. 2, our MEC^2-TT framework is designed to capture user emotions and generate empathetic responses. The framework is built upon a pretrained large language model (e.g., Qwen2.5-7B [37]) as the backbone, which is augmented with three specialized modules, i.e., the Emotion Consistency Correction Module (ECC, Fig. 2 (a)), the Emotion Trajectory Tracking Module (ETT, Fig. 2 (b)), and the ECoT-Augmented inference module (Fig. 2 (c)).

Concretely, given a raw multimodal user utterance u_t from the dialogue history H, we first use pre-trained modality-specific encoders to extract initial visual, acoustic, and textual representations, denoted as X_t^v, X_t^a, and X_t^l. These features are then fed into the ECC Module, which aligns and refines them to mitigate inherent modality-specific biases and cross-modal inconsistencies. This module produces a unified emotion representation E_t for each turn.

Next, focusing on the current turn T, the ETT Module processes the sequence of unified features $E_1, \ldots, E_T$ to construct a trajectory-aware representation V_T that encodes global emotional variations up to that point.

Finally, to bridge the gap between multimodal perception and response generation, we introduce the ECoT mechanism. Rather than directly mapping features to outputs, ECoT provides structured, text-based reasoning steps that guide the model to explicitly interpret the unified representation E_T and the historical trajectory V_T, infer underlying emotional causes, and formulate empathetic response strategies to generate the final textual response r_T. The detailed formulations are introduced in the following subsections.

3.3 Multimodal Feature Extraction

Given the raw multimodal input signals $u_t = \{u_t^v, u_t^a, u_t^l\}$ at turn t, we first employ pre-trained modality-specific encoders to extract their corresponding initial feature representations. Specifically, we utilize pretrained EmotiEffLib [29] to process the video frames, Wav2Vec [1] for the audio signals, and pretrained EmoBERTa [12] to encode the textual utterances. This extraction yields the raw sequence-level representations $X_t^v \in \mathbb{R}^{d_v}$, $X_t^a \in \mathbb{R}^{d_a}$, and $X_t^l \in \mathbb{R}^{d_l}$, where d_v, d_a, and d_l denote the intrinsic feature dimensions of the respective pre-trained models.

3.4 Emotion Consistency Correction

We first project the initial visual, acoustic, and textual representations into a unified d-dimensional latent space via modality-specific linear transformations:

$$F_t^m = X_t^m W_m + b_m, \quad m \in \{v, a, l\}, \tag{3}$$

where X_t^m denotes the raw extracted feature for modality m at turn t, W_m is a learnable weight matrix, and $F_t^m \in \mathbb{R}^d$ serves as the base emotion representation.

Unlike general feature misalignment, cross-modal emotional inconsistencies arise from discrepancies in semantic probability distributions across modalities. For instance, visual cues might strongly indicate a specific emotion, while acoustic signals yield an ambiguous or even conflicting distribution. Standard geometric metrics, such as Euclidean or cosine distance, merely measure vector proximity and fail to capture these probabilistic semantic gaps. To address this, we adopt an information-theoretic perspective. Specifically, we quantify the relative entropy, i.e., the expected information difference, between the emotional states conveyed by different modalities. The Kullback-Leibler (KL) divergence naturally

satisfies this requirement, providing a rigorous measure of how one modality's emotional distribution diverges from another [5, 10].

To operationalize this information-theoretic measurement, we first map the latent representations F_t^m into emotion probability distributions $P_t^m \in \mathbb{R}^C$ over C latent affective factors. This is achieved via a shared projection layer followed by a Softmax operation:

$$P_t^m = \text{Softmax}(W_p F_t^m). \tag{4}$$

Subsequently, we compute the pairwise KL divergence between the distributions of modality i and modality j to quantify their affective semantic gap:

$$D_{\text{KL}}(P_t^i \parallel P_t^j) = \sum_{c=1}^{C} P_t^i(c) \log \left(\frac{P_t^i(c)}{P_t^j(c)} \right), \tag{5}$$

where $i, j \in \{v, a, l\}$ and $i \neq j$. A larger divergence indicates a severe inconsistency between the affective states conveyed by the two modalities, whereas a smaller value suggests emotional alignment.

To dynamically calibrate the modality-specific features based on these discrepancies, we transform the computed KL divergence into a bounded inter-modal similarity score $S_t^{i,j} \in (0, 1]$:

$$S_t^{i,j} = \frac{1}{1 + D_{\text{KL}}(P_t^i \parallel P_t^j)}. \tag{6}$$

This similarity score acts as an adaptive gating mechanism. It assigns higher confidence to modalities that share consistent emotional semantics while penalizing divergent ones. We then perform emotion consistency correction through a residual-like feature recalibration process. For a given modality m, its corrected representation $\tilde{F}_t^m$ is obtained by modulating the original feature F_t^m with the aggregated similarity scores from the other modalities:

$$\tilde{F}_t^m = \text{MLP}\left(F_t^m \odot \left(1 + \sum_{n \in \{v,a,l\} \setminus \{m\}} S_t^{m,n} \right) \right), \tag{7}$$

where $\odot$ denotes element-wise multiplication, and $\text{MLP}(\cdot)$ is a lightweight neural network comprising linear layers and ReLU activations to ensure deeper cross-modal semantic alignment.

Finally, the refined, bias-corrected features from all three modalities are concatenated to form the robust multimodal emotion representation $E_t \in \mathbb{R}^{3d}$ for the current utterance:

$$E_t = [\tilde{F}_t^v; \tilde{F}_t^a; \tilde{F}_t^l]. \tag{8}$$

This unified and unbiased representation E_t provides a trustworthy foundation for tracking the emotional trajectory in the subsequent dialogue history.

3.5 Emotion Trajectory Tracking

Human emotions in conversations are not isolated snapshots; they continuously evolve, forming an emotional trajectory. Accurately tracking the user's emotional transitions and historical affective patterns is a fundamental prerequisite for generating contextually coherent and empathetic responses. Therefore, building upon the bias-corrected multimodal affective representation E_t for each utterance, we propose an Emotion Trajectory Tracking module. This module captures the temporal dynamics and historical dependencies of the user's emotional states across the dialogue history.

For a dialogue history up to the current turn T, we first collect the sequence of corrected emotion representations $\mathcal{E}_{\leq T} = [E_1, E_2, \ldots, E_T]$. Because standard attention mechanisms are position-unaware, directly attending to these features would destroy the temporal flow of the conversation. To explicitly model the chronological trajectory, we inject trainable temporal positional encodings $PE(\cdot)$ into the representations:

$$U_t = E_t + PE(t), \quad \forall t \in \{1, 2, \ldots, T\}, \tag{9}$$

where U_t represents the time-aware emotion embedding for the t-th turn, $PE(t) \in \mathbb{R}^{3d}$ is a learnable positional embedding mapped from the discrete turn index i, allowing the model to explicitly capture the chronological order of the affective states.

To explicitly model the dynamic evolution of the user's emotions, we feed the time-aware sequence $\mathcal{U}_{<T} = [U_1, \ldots, U_{T-1}]$ into a Multi-Head Self-Attention (MHSA) block [30]. This operation constructs a cohesive emotional timeline, allowing the model to capture how feelings transition, accumulate, or resolve over time:

$$Z_{<T} = \text{MHSA}(\mathcal{U}_{<T}, \mathcal{U}_{<T}, \mathcal{U}_{<T}), \tag{10}$$

where the inputs correspond to the Query, Key, and Value, respectively. The output sequence $Z_{<T}$ encapsulates the contextualized emotional history, effectively establishing the continuous background trajectory of the dialogue before the current turn T. Finally, to understand how the user's current emotional state E_T is derived from this historical trajectory, we conceptualize a Cross-Attention mechanism as an active trajectory tracker. We use the current turn's time-aware representation U_T as the Query to scan and attend over the historical timeline $Z_{<T}$, which serves as Keys and Values:

$$\begin{aligned} Q &= U_T W_q, \\ K &= Z_{<T} W_k, \\ V &= Z_{<T} W_v, \\ V_T &= \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \right) V, \end{aligned} \tag{11}$$

where W_q, W_k, W_v are learnable projection matrices, and d_k is the scaling factor. Rather than simply appending historical features, this tracking operation uses the current emotion as an active probe to selectively retrieve the most relevant historical shifts that contextually support the present state.

The resulting vector V_T acts as our robust, trajectory-grounded emotional representation. By using the instantaneous emotion to selectively aggregate its historical derivation, V_T captures the complete affective journey of the user. This unified representation provides the necessary dynamic context for the subsequent Empathetic Response generation.

3.6 Empathetic Response Generation

While the trajectory-aware representation V_T effectively captures the deep multimodal context of the dialogue history, LLMs struggle to translate these dense, continuous vectors into explicitly empathetic responses without structured guidance. Direct end-to-end decoding risks bypassing the complex reasoning processes required for true empathy. To bridge this reasoning gap, we propose the Empathetic Chain of Thought (ECoT) mechanism, which is incorporated into the Qwen2.5-7B [37] backbone to decompose the empathetic reasoning process into an interpretable reasoning pipeline.

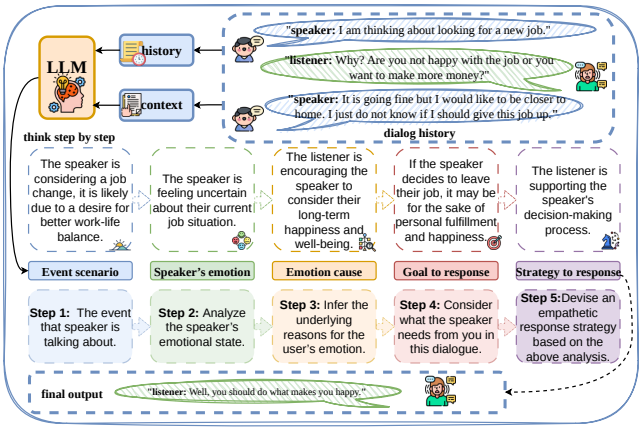

Figure 3: ECoT of our method. Based on the previously inferred context X, the model follows five reasoning steps to form the latent chain C_{ECoT} and generate empathetic responses aligned with the speaker's emotional state and conversational context.

Instead of treating the reasoning chain as the final objective, we utilize it as a structured prerequisite to guide the response generation, mirroring human cognitive empathy. Specifically, conditioned on the comprehensive context X, the model is instructed to sequentially deduce five intermediate cognitive components before formulating its reply: the event scenario (c_{event}), the speaker's emotion (c_{emo}), the emotion cause (c_{cause}), the implicit conversational goal (c_{goal}), and the corresponding empathetic strategy ($c_{strategy}$). We denote this latent reasoning chain C_{ECoT} as:

$$C_{ECoT} = [c_{event}, c_{emo}, c_{cause}, c_{goal}, c_{strategy}]. \tag{12}$$

To make this reasoning procedure clear and interpretable, we design a structured prompt template that explicitly guides the model through each intermediate step. Fig. 3 illustrates an example of the reasoning chain employed in our method. To inject the multimodal trajectory into the LLM, the representation V_T is projected into the textual embedding space of Qwen2.5-7B via a lightweight linear adapter, functioning as a continuous soft prompt P_V. Let I denote the task instruction and H_{text} denote the textual dialogue history. The unified multimodal context X is constructed by concatenating these inputs at the sequence level: $X = [I; H_{text}; P_V]$.

Based on this comprehensive context, the LLM operates in a two-stage auto-regressive generation paradigm. First, the model generates the intermediate reasoning steps C_{ECoT} token by token. The probability of generating the entire reasoning chain is formulated as:

$$P(C_{ECoT}|X) = \prod_{j=1}^{|C_{ECoT}|} P(s_j|X, s_{<j}; \theta), \tag{13}$$

where s_j is the j-th token of C_{ECoT}, and θ denotes the trainable parameters of the model.

Subsequently, conditioned on both the multimodal context X and the explicitly generated reasoning chain C_{ECoT}, the model infers the final empathetic response r_T:

$$P(r_T|X, C_{ECoT}) = \prod_{k=1}^{|r_T|} P(y_k|X, C_{ECoT}, y_{<k}; \theta), \tag{14}$$

where y_k is the k-th token of the response r_T. By explicitly conditioning the generation of r_T on C_{ECoT}, the model translates its internal affective analysis into a highly empathetic and contextually grounded textual reply.

To align the model with this structured reasoning paradigm, we employ a full fine-tuning strategy. The model is trained to auto-regressively maximize the likelihood of both the intermediate reasoning steps and the final textual response. Formally, the training objective is defined by the standard negative log-likelihood loss over the concatenated sequence:

$$\mathcal{L} = -\sum_{j=1}^{|C_{ECoT}|} \log P(s_j|X, s_{<j}; \theta) - \sum_{k=1}^{|r_T|} \log P(y_k|X, C_{ECoT}, y_{<k}; \theta), \tag{15}$$

where θ represents the trainable parameters of the LLM and the multimodal projector. s_j is the j-th token of the generated ECoT reasoning chain C_{ECoT}, and y_k is the k-th token of the target response r_T. This joint optimization ensures that the model learns to effectively leverage the multimodal trajectory V_T to perform rigorous, step-by-step empathetic reasoning, ultimately culminating in high-quality, contextually coherent responses.

4 Experiments

4.1 Datasets

We evaluate our method on two widely adopted datasets, AvaMERG and MELD, which serve as standard benchmarks for multimodal emotion recognition and generation. Given the limited availability of datasets supporting the multimodal emotion response generation (MERG) task, these two datasets have become de facto choices in the community, providing a solid basis for evaluation. The details of each dataset are introduced below.

AvaMERG[38] is built upon the existing text-based ERG benchmark EMPATHETIC DIALOGUE. They augment the dataset to include multimodal signals and annotations. For each utterance in the dialogue, they provide authentic human-reading speech and dynamic talking-face avatar videos (2D facial modeling) that both correspond to the intended emotion. It contains 33,048 annotated dialogues with 152,021 multimodal utterances.

MELD[22] is a widely used multimodal benchmark consisting of over 1,400 dialogues and around 13,000 utterances extracted from the TV series Friends, with each utterance annotated using eight emotion labels and three sentiment categories.

For the two aforementioned datasets, we adopted the same processing method: we organized the utterances into incrementally constructed dialogues, where each sample contains a growing history of previous turns. Specifically, the AvaMERG dataset includes ECoT annotations, which are directly leveraged in our approach.

4.2 Baselines

To demonstrate the effectiveness of our proposed method, we compare it against several representative baselines. Given the limited

Table 1: Evaluation results on AvaMERG and MELD datasets. Bold: best; Underline: second best.

Dataset	Model	Objective Evaluation					Human Evaluation		
		BLEU-1	ROUGE-1	BERTScore-f	BERTScore-p	BERTScore-r	Empathy	Relevance	Fluency
AvaMERG	E3RG [17]	6.01	14.76	86.35	86.73	85.99	3.18	3.09	3.92
	Empathyear [6]	7.63	16.75	81.42	78.55	84.54	3.45	3.63	**4.11**
	Qwen3-8B [36]	11.25	22.33	87.26	86.30	88.26	3.64	3.07	3.85
	Ours	**20.97**	**31.23**	**88.95**	**89.03**	**88.90**	**3.79**	**3.63**	3.95
MELD	E3RG [17]	6.26	14.91	86.07	85.01	86.61	2.79	3.46	**3.75**
	Empathyear [6]	2.83	8.32	79.20	77.12	81.44	2.81	3.26	3.35
	Qwen3-8B [36]	3.31	9.72	83.79	83.76	83.89	2.61	**3.52**	3.45
	PA-MERG† [32]	-	-	84.94	84.72	85.16	-	-	-
	ERGM† [32]	-	-	84.61	83.95	85.28	-	-	-
	CASE† [44]	-	-	82.73	81.59	83.91	-	-	-
	EmpSOA† [43]	-	-	83.68	82.71	84.68	-	-	-
	Ours	**14.30**	**24.13**	**88.33**	**89.10**	**87.68**	**2.84**	3.48	3.52

† Results taken directly from the original paper due to unavailable source code.

number of existing works in this field, we select the following models for evaluation: E3RG [17] decomposed the MERG task into multimodal empathy understanding, memory retrieval, and response generation, leveraging expressive speech and video generation, while we only evaluated its textual outputs. EmpathyEar [6] supported multimodal inputs and generated responses with talking avatars by combining large language models with multimodal encoders and generators; similarly, only its textual responses were considered. Qwen3-8B [36] served as a strong text-only baseline with a flexible thinking budget mechanism for adaptive inference. PA-MERG [32] incorporated multimodal information and modeled the interactions among personality, emotion, and context to generate empathetic responses. ERGM [33] focused on achieving affective and cognitive empathy in multimodal settings, utilizing guidance text to enhance emotional understanding without external databases. CASE [44] builds upon a commonsense cognition graph and an emotional concept graph and then aligns the user's cognition and affection at both the coarse and fine-grained levels. EmpSOA [43] clearly maintain, modulate, and inject the self-other aware information into the process of empathetic response generation.

4.3 Implementation Details

We implement our proposed framework using PyTorch and adopt Qwen2.5-7B [37] as the foundational large language model. To facilitate trajectory tracking, we structure the raw conversations into incremental dialogue sequences, where each training sample contains a progressively growing history of previous turns. All experiments are conducted on a server equipped with 4 NVIDIA RTX A6000 GPUs. To optimize memory footprint and computational efficiency, we leverage the DeepSpeed [27] library and perform both training and inference in FP16 half-precision. The model is trained for 3 epochs using the AdamW optimizer with an initial learning rate of 1e-5. During training, we set the per-device batch size to 2 and apply a gradient accumulation step of 3 to maintain a stable effective batch size.

4.4 Objective Evaluation

Following prior work, we evaluate our model using BLEU-1 [21] and ROUGE-1 [16] to measure surface-level lexical overlap, alongside BERTScore [41] to assess deep semantic equivalence between the generated and reference responses.

As shown in Table 1, our proposed method consistently achieves the best performance across all metrics on both the AvaMERG and MELD datasets. In terms of lexical and structural fidelity, our model significantly outperforms all baselines, yielding the highest BLEU-1 (20.97 and 14.30) and ROUGE-1 (31.23 and 24.13) scores, indicating superior fluency and content alignment. Furthermore, it demonstrates unparalleled semantic equivalence, achieving peak BERTScore-F1 values of 88.95 and 88.33. These comprehensive improvements highlight the architectural advantages of our framework. Compared to specialized baselines like E3RG and EmpathyEar, which rely on static voting or predefined reasoning, our method explicitly models cross-modal heterogeneity and temporal emotional dynamics. Meanwhile, although the general-purpose LLM Qwen3-8B shows competitive text generation, its lack of multimodal grounding severely limits its effectiveness (e.g., scoring only 11.25 / 3.31 on BLEU-1). Ultimately, these gains stem from the synergistic integration of the emotion consistency correction, emotion trajectory tracking, and ECoT mechanisms, which collectively ensure deep semantic coherence and natural linguistic formulation.

4.5 Human Evaluation

To assess the generated responses from a comprehensive human-centric perspective, especially in terms of empathic expression ability, we conduct human evaluations focusing on the multifaceted aspects of empathy and conversational quality. We measure the model performance through two distinct evaluation schemas: absolute scoring and relative comparison.

For the absolute scoring, we employ a 5-point Likert scale (1 = very poor, 5 = excellent) across three primary dimensions. These include *Empathy* to measure the model's ability to understand the

Table 2: A/B testing results of our model against baselines.

Comparisons	Win (%)	Lose (%)	Tie (%)
Ours vs. E3RG [17]	**57.5**	17.5	25.0
Ours vs. Empathyear [6]	**62.5**	14.0	23.5
Ours vs. Qwen3-8B [36]	**53.5**	26.0	20.5

user's emotional state and provide appropriate support, *Relevance* to evaluate contextual coherence and logical alignment with the dialogue history, and *Fluency* to assess the naturalness and grammatical correctness of the generated text. For the relative comparison, we conduct pairwise A/B testing where annotators directly choose the preferred response ("Win", "Lose", or "Tie") between our method and the given baselines. To ensure a fair and rigorous comparison, we randomly sample 200 dialogue context-response pairs. Five independent annotators with graduate-level backgrounds are recruited to conduct both evaluations in a strictly blind manner. All model identities are anonymized, and responses are randomly shuffled to eliminate subjective bias.

As shown in Table 1, our model achieves the highest or competitive scores in *Empathy*, *Relevance*, and *Fluency* across both datasets. It is particularly noteworthy that our model demonstrates a clear advantage in *Empathy*, achieving **3.79** on AvaMERG and **2.84** on MELD. This success can be directly attributed to the joint optimization of our proposed modules. Specifically, the emotion consistency correction resolves cross-modal inconsistencies, enabling the model to accurately grasp the user's true emotional state to provide appropriate support (*Empathy*). Building on this, the emotion trajectory tracking models the historical emotional dynamics, which guarantees tight contextual coherence and logical alignment with the dialogue history (*Relevance*). Ultimately, the Empathetic Chain of Thought (ECoT) mechanism explicitly structures the response strategy, translating these complex emotional understandings into linguistically natural and grammatically correct text (*Fluency*).

The A/B testing results (Table 2) confirm our model's superiority in human-perceived quality with win rates exceeding 50% across all baselines. The advantage over specialized models like Empathyear (62.5% win) and E3RG (57.5% win) validates our multimodal emotion consistency correction and trajectory tracking mechanisms. Furthermore, defeating Qwen3-8B (53.5% win vs. 26.0% lose) highlights the necessity of our ECoT. While Qwen3-8B's built-in CoT often prioritizes objective problem-solving, ECoT reasons through emotional causes and goals. This empathy-driven reasoning mechanism ensures responses are deeply attuned to user feelings, achieving stronger alignment with human empathetic preferences.

4.6 Ablation study

Table 3 presents the ablation results of our framework on the AvaMERG dataset, evaluating the contributions of the ECC module, the ETT module, and the ECoT mechanism. Overall, the full model consistently achieved the best performance across all evaluation metrics, demonstrating the effectiveness of jointly modeling multimodal emotion consistency, temporal emotional dynamics, and structured reasoning.

Table 3: Ablation Study Results Comparison.

ECC	ETT	ECoT	BLEU-1	BLEU-2	ROUGE-1	ROUGE-L	BERTScore
Effect of ECC, ETT & ECoT							
✗	✓	✓	18.37	12.15	28.83	25.53	88.23
✓	✗	✓	18.34	12.19	28.22	24.98	88.26
✓	✓	✗	9.86	6.17	15.73	13.41	85.72
Text-only Finetune							
✗	✗	✗	8.78	5.54	13.73	11.62	85.10
Full Method							
✓	✓	✓	**20.97**	**14.89**	**31.23**	**28.29**	**88.96**

Removing either the ETT or ECC module led to noticeable performance degradation across all metrics. Specifically, when removing ETT, BLEU-1 dropped from 20.97 to 18.34, and ROUGE-1 decreased from 31.23 to 28.22. Similarly, removing ECC led to comparable performance declines, indicating that both modules contributed significantly to performance. These results suggested that the ETT module played a crucial role in modeling the temporal evolution of user emotions, thereby improving contextual coherence, while the ECC module improved cross-modal consistency by mitigating emotional discrepancies between modalities. The comparable performance drops further indicated that these two modules were complementary and equally important.

Removing the ECoT mechanism resulted in a significant performance drop, far more pronounced than removing the multimodal modules. For instance, BLEU-1 sharply decreased to 9.86, and BLEU-2 to 6.17, representing nearly a 50% reduction compared to the full model. This substantial degradation indicated that the structured reasoning process introduced by ECoT played a crucial role in guiding the model to generate coherent and appropriate responses.

The performance further deteriorated to the lowest bound when only standard fine-tuning was applied. This underscores that standard end-to-end training fails to effectively activate the LLM's empathetic capabilities. It proves the absolute necessity of our integrated approach: the multimodal modules (ECC and ETT) provide a reliable sensory foundation, while the ECoT mechanism offers the indispensable cognitive framework to translate those signals into appropriate empathetic support.

4.7 Case study

Fig. 4 presents a qualitative case study of our approach in empathetic dialogue generation. As illustrated, MEC^2-TT effectively mitigates cross-modal emotional inconsistencies via emotion consistency correction, which serves as a crucial foundation for grasping the user's true affective state. Building upon this accurate understanding, our emotion trajectory tracking captures the dynamic evolution of user emotions across the conversation. Consequently, our model is empowered to generate contextually appropriate and deeply empathetic responses, successfully avoiding the pitfalls of superficial textual cues, particularly in scenarios of emotional camouflage.

In Case 1, the user experiences a significant workplace setback after a month of hard work and is dismissed. The user's text superficially conveys a positive and rationalizing attitude by claiming everything is fine. However, the audio modality captures a trembling voice, and the video modality shows a forced smile with red eyes,

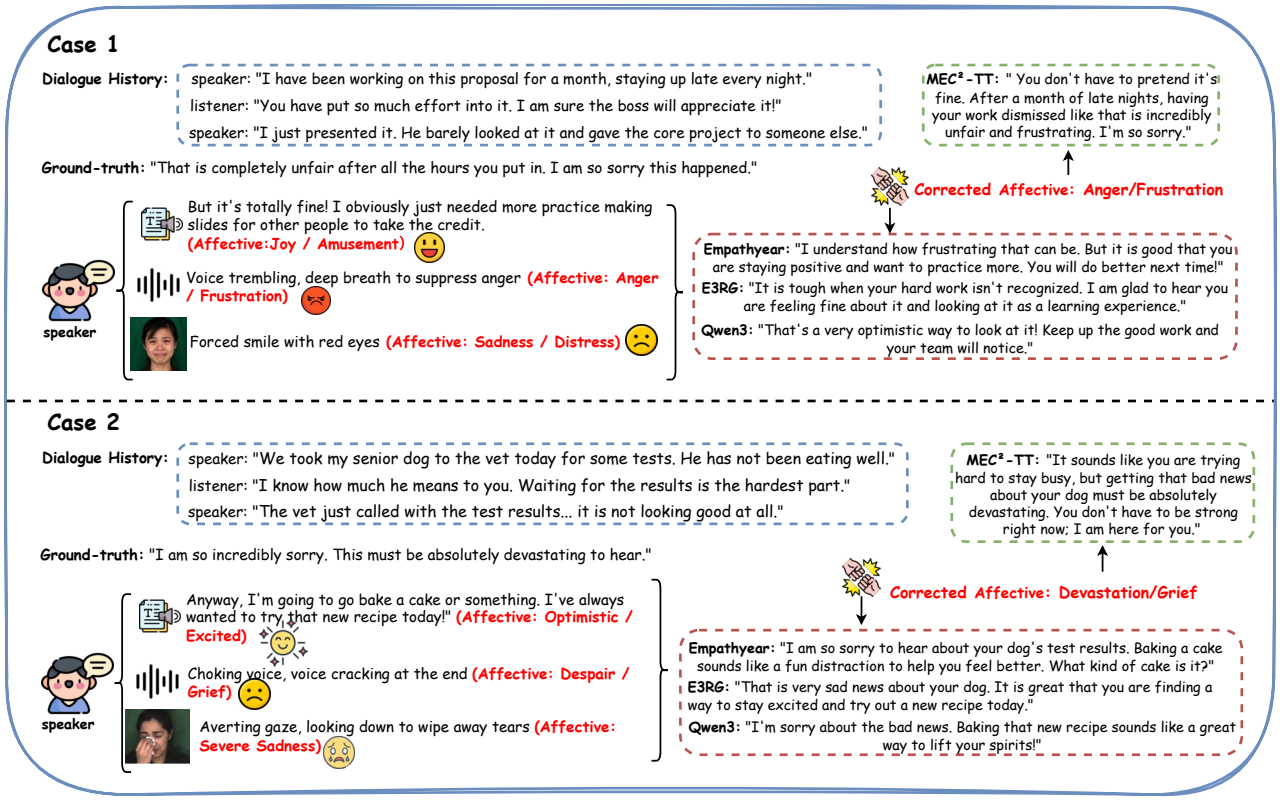

Figure 4: Case studies comparing our MEC^2-TT with baselines. Our model successfully infers the true underlying emotions despite superficial textual camouflage to generate highly empathetic responses.

jointly revealing suppressed anger and distress. An appropriate system response must see through this sarcastic camouflage. Baselines such as Empathyear, E3RG, and Qwen3 are entirely misled by the surface textual sentiment. They generate responses exhibiting toxic positivity by inappropriately encouraging the user to practice more and stay positive, completely invalidating the actual frustration. In contrast, MEC^2-TT successfully leverages cross-modal cues to correct the affective state to anger and frustration. By grounding its response in the dialogue history, it acknowledges the unfairness of the situation and provides genuine empathy.

Similarly, Case 2 demonstrates a scenario where the user receives devastating news about their senior dog but attempts to force a topic change by enthusiastically discussing a baking recipe. Despite this optimistic textual camouflage, the underlying severe sadness is betrayed by a choking voice and an averted gaze. The baselines over-rely on the superficial text and fall into the trap of emotional avoidance. By focusing entirely on the baking distraction, they generate responses that lack true empathetic depth. Conversely, MEC^2-TT leverages its consistency correction and emotion trajectory tracking to recognize this sudden shift as a vulnerable coping mechanism. It accurately identifies the true state of devastation and grief, delivering a deeply supportive response that validates the user's hidden pain and permits them to grieve rather than forcing a strong facade. This proves our framework's remarkable capability to provide genuine psychological support even when human expressions are highly contradictory.

Overall, these case studies demonstrate that MEC^2-TT consistently outperforms baseline models in terms of multimodal emotion correction, deep empathy tracking, and response coherence when facing complex human emotional expressions.

5 Conclusion

In this work, we proposed a novel framework for multimodal empathetic response generation integrating emotion consistency correction, emotion trajectory tracking, and an ECoT reasoning mechanism. The first module mitigated cross-modal affective inconsistencies to produce refined multimodal representations. The second captured the evolution of user emotions across dialogue turns, enabling the model to understand both transient and persistent emotional patterns. Guided by ECoT, the LLM reasoned over the multimodal emotional context to generate semantically accurate and emotionally aligned responses. Experiments on AvaMERG and MELD demonstrated consistent improvements, achieving state-of-the-art performance. These results highlight the importance of jointly modeling emotion correction, temporal dynamics, and structured reasoning for empathetic dialogue generation.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[2] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv* abs/2310.09478 (2023). https://api.semanticscholar.org/CorpusID:264146906

[3] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. arXiv:2406.11161 [cs.AI] https://arxiv.org/abs/2406.11161

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1724–1734. doi:10.3115/v1/D14-1179

[5] Jiequan Cui, Beier Zhu, Qingshan Xu, Zhuotao Tian, Xiaojuan Qi, Bei Yu, Hanwang Zhang, and Richang Hong. 2025. Generalized Kullback-Leibler Divergence Loss. arXiv:2503.08038 [cs.LG] https://arxiv.org/abs/2503.08038

[6] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot. arXiv:2406.15177 [cs.MM] https://arxiv.org/abs/2406.15177

[7] Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-CORE: Emotion Correlation Enhanced Empathetic Dialogue Generation. arXiv:2311.15016 [cs.CL] https://arxiv.org/abs/2311.15016

[8] Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

[9] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 2122–2132. doi:10.18653/v1/N18-1193

[10] Shuyi Ji, Zizhao Zhang, Shihui Ying, Liejun Wang, Xibin Zhao, and Yue Gao. 2022. Kullback–Leibler Divergence Metric Learning. *IEEE Transactions on Cybernetics* 52, 4 (2022), 2047–2058. doi:10.1109/TCYB.2020.3008248

[11] Juyeon Kim, Juyoung Hong, and Yukyung Choi. 2024. Causal Inference for Modality Debiasing in Multimodal Emotion Recognition. *Applied Sciences* 14, 23 (2024). doi:10.3390/app142311397

[12] Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. arXiv:2108.12009 [cs.CL]

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

[14] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2021. Knowledge Bridging for Empathetic Dialogue Generation. arXiv:2009.09708 [cs.CL] https://arxiv.org/abs/2009.09708

[15] Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought. arXiv:2401.06836 [cs.CL] https://arxiv.org/abs/2401.06836

[16] Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 605–612. doi:10.3115/1218955.1219032

[17] Ronghao Lin, Shuai Shen, Weipeng Hu, Qiaolin He, Aolin Xiong, Li Huang, Haifeng Hu, and Yap peng Tan. 2025. E3RG: Building Explicit Emotion-driven Empathetic Response Generation System with Multimodal Large Language Model. arXiv:2508.12854 [cs.AI] https://arxiv.org/abs/2508.12854

[18] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70. doi:10.1016/j.inffus.2020.06.011

[19] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8968–8979. doi:10.18653/v1/2020.emnlp-main.721

[20] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech 2010*. 1045–1048. doi:10.21437/Interspeech.2010-343

[21] 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*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[22] 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. arXiv:1810.02508 [cs.CL] https://arxiv.org/abs/1810.02508

[23] Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2024. Harnessing the Power of Large Language Models for Empathetic Response Generation: Empirical Investigations and Improvements. arXiv:2310.05140 [cs.CL] https://arxiv.org/abs/2310.05140

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231591445

[25] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. arXiv:1811.00207 [cs.CL] https://arxiv.org/abs/1811.00207

[26] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 5370–5381.

[27] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3505–3506. doi:10.1145/3394486.3406703

[28] Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-Aware Empathetic Response Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (Jun. 2022), 11229–11237. doi:10.1609/aaai.v36i10.21373

[29] Andrey Savchenko. 2023. Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing Correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 30119–30129. https://proceedings.mlr.press/v202/savchenko23a.html

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

[31] Yunxiao Wang, Meng Liu, Wenqi Liu, Kaiyu Jiang, Bin Wen, Fan Yang, Tingting Gao, Guorui Zhou, and Liqiang Nie. 2025. COMPEER: Controllable Empathetic Reinforcement Reasoning for Emotional Support Conversation. arXiv:2508.09521 [cs.CL] https://arxiv.org/abs/2508.09521

[32] 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*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 8925–8938. https://aclanthology.org/2025.coling-main.598/

[33] Jiaqiang Wu, Shangfei Wang, Yanan Chang, and Zhouan Zhu. 2025. Empathetic Response Generation Through Multi-Modality. *IEEE Transactions on Affective Computing* 16, 4 (2025), 3614–3623. doi:10.1109/TAFFC.2025.3599869

[34] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*. 53366–53397.

[35] 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)*. 15009–15022.

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui,

Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

[37] Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2024. Qwen2.5 Technical Report. *ArXiv* abs/2412.15115 (2024). https://api.semanticscholar.org/CorpusID:274859421

[38] 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. *arXiv preprint arXiv:2502.04976* (2025).

[39] 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. arXiv:2502.04976 [cs.MM] https://arxiv.org/abs/2502.04976

[40] Hanqing Zhang, Si Sun, and Dawei Song. 2026. Emotion-Aware LLM Adaptation for Empathetic Dialogue Generation. *IEEE Transactions on Audio, Speech and Language Processing* 34 (2026), 125–137. doi:10.1109/TASLPRO.2025.3642537

[41] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[42] Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. STICKERCONV: Generating Multimodal Empathetic Responses from Scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7707–7733. doi:10.18653/v1/2024.acl-long.417

[43] Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don't Lose Yourself! Empathetic Response Generation via Explicit Self-Other Awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13331–13344. doi:10.18653/v1/2023.findings-acl.843

[44] Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8223–8237. doi:10.18653/v1/2023.acl-long.457