MERIA: Empathetic Response Generation via Parallel Disentanglement and Uncertainty-Gated Fusion

Chenhao Dang*
dangchenhao@std.uestc.edu.cn
China Electronics Technology Group Corporation 15th Research Institute
Beijing, China

Zeyuan Zhu
2458039431@qq.com
China Electronics Technology Group Corporation 15th Research Institute
Beijing, China

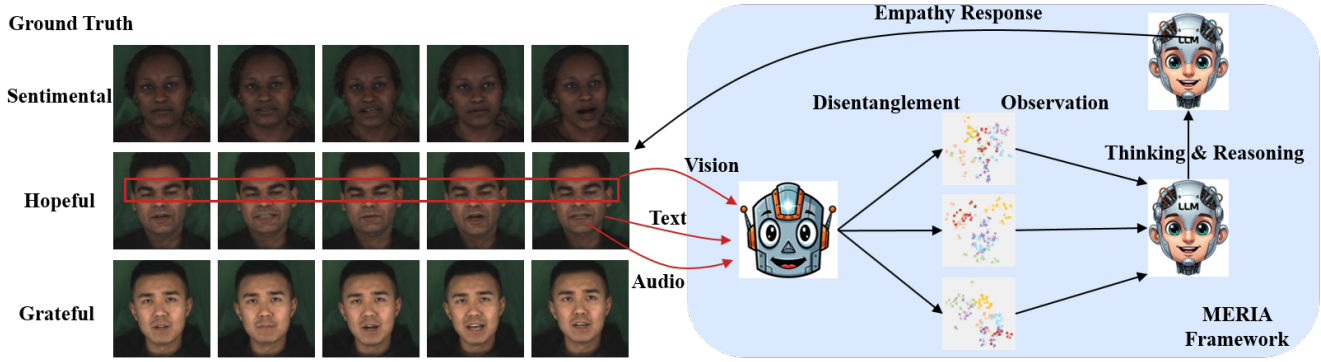


Figure 1: The Multimodal Empathetic Reasoning and Inconsistency-Aware (MERIA) framework addresses the issue of inconsistency in emotional representations across modalities in multimodal empathetic dialogues. The three video frames on the left, sourced from the AvaMERG dataset, exhibit a discrepancy between the visual emotion and the textual ground truth. On the right is an overview of the MERIA framework for multimodal empathetic response generation.

Abstract

A critical challenge in advancing human-like conversational AI systems is enabling models to understand and respond to user emotions contextually, a task known as Multimodal Empathetic Response Generation (MERG). While prevailing multimodal models attempt to resolve cross-modal emotional discrepancies via concatenation or cross-attention, their simplistic fusion mechanisms often fail to account for the nuanced and contradictory nature of human emotions. Consequently, the resulting feature representations suffer from these unresolved internal conflicts, limiting their effectiveness. In this paper, we propose a novel Multimodal Empathetic Reasoning and Inconsistency-Aware (MERIA) framework. MERIA introduces a multimodal disentanglement encoder based on β-VAE and extends the AvaMERG dataset with multimodal chains of empathy (M-CoE). Our framework outperforms existing methods, achieving the best human evaluation scores in the empathetic text response generation task of the MERG 2025 Challenge. Our code is available at https://github.com/DANG-ai/MERIA-MERG.

*Corresponding author of this paper.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762031

CCS Concepts

• **Information systems** → **Multimedia information systems**; • **Computing methodologies** → **Natural language generation**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Empathetic Response Generation, Multimodal Large Language Model, Modal Disentanglement, Affective Computing

ACM Reference Format:
Chenhao Dang and Zeyuan Zhu. 2025. MERIA: Empathetic Response Generation via Parallel Disentanglement and Uncertainty-Gated Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3762031

1 Introduction

In recent years, the emergence of Large Language Models (LLMs) has endowed machines with unprecedented intelligence, bringing us one step closer to achieving Artificial General Intelligence (AGI) [19]. However, AGI not only requires human-level logic and linguistic abilities but also necessitates emotional understanding and empathetic capabilities comparable to those of humans [10]. In this context, emotional understanding and empathy in human-computer interaction have become paramount [5], driving the development of the Empathetic Response Generation (ERG) task [19], which aims to enable machines to respond to users' emotional needs with

sensitivity and compassion. The expression of human emotions is inherently multimodal, encompassing vocal tone, facial expressions, and body language, often extending beyond the scope of text-based communication [14]. Therefore, expanding ERG to the multimodal domain, resulting in Multimodal Empathetic Response Generation (MERG), involves a shift from language comprehension to embodied human state perception and presents two core challenges: signal inconsistency and shallow reasoning [19].

Cross-modal emotional inconsistency poses a significant challenge. However, current multimodal models typically address emotional incongruence across modalities through simple fusion strategies [14], such as concatenating embeddings or employing cross-modal attention. We argue that these naive alignment approaches are inadequate. Given the complex and often paradoxical nature of human social emotions [11], forcing a simplistic fusion can introduce internal conflicts, leading to the extraction of degraded or even contradictory feature representations. For example, the text "That's great" may be accompanied by a sarcastic tone and a helpless expression. The AvaMERG dataset also demonstrates instances of emotional inconsistency, as shown in Figure 1, where facial expressions in the video do not fully match the sentiment conveyed by the text. This inconsistency is not merely noise, but rather a socially informative signal.

Additionally, AvaMERG [19] introduces the concept of a Chain of Empathy (CoE), which defines a reasoning path from event → emotion → cause → goal → response. However, the limitation of this model lies in its reliance solely on text, neglecting non-textual information such as facial expressions and vocal tone.

To address these challenges, we propose a Multimodal Empathetic Reasoning and Inconsistency-Aware (MERIA) framework, designed for deep, inconsistency-aware empathetic response generation. The core idea of MERIA is to avoid futile fusion at the source of signal conflicts. Instead, we first decouple the signals through a sophisticated disentangling mechanism and then perform intelligent, selective fusion and reasoning.

Our approach begins by extracting unimodal representations from raw input signals (text, audio, video) via pre-trained encoders. Central to our framework is the Parallel Disentangled Representation (PDR) module, an architecture inspired by parallel β-VAE [16], which we employ to decompose the entangled inputs into three orthogonal semantic subspaces. This module is trained independently using a dedicated semi-supervised loss suite.

Subsequently, we introduce an Uncertainty-Gated Fusion (UGF) mechanism. The UGF leverages the quantified uncertainty of each modality's private information to derive gating weights. These weights, in turn, dynamically modulate the contributions of the shared emotional representations during the fusion process.

To endow our model with sophisticated empathetic reasoning, we curated the Multimodal Chain of Empathy (M-CoE) dataset, a novel, manually annotated extension of the AvaMERG corpus. The structured annotation schema for M-CoE is adapted from the principles of CoE [19]. We then leverage this dataset to fine-tune a M-CoE LLM, which is tasked with generating explicit reasoning chains from the disentangled multimodal embeddings and dialogue context. In the final stage, a Response LLM is fine-tuned on a composite of these generated M-CoEs and the original dialogue text, enabling it to produce contextually and emotionally appropriate responses.

The main contributions are summarized as follows:

- To overcome the performance degradation caused by internal conflicts when simplistic fusion methods are confronted with incongruent multimodal emotional signals, we propose MERIA, a novel framework for empathetic response generation. The core PDR module effectively decouples shared emotions from modality-specific information and identity features using β-VAE.
- We extend the concept of M-CoE and provide multimodal annotated data, expanding empathetic reasoning from the pure text domain to a multimodal domain based on audiovisual evidence. Our annotated dataset will be made publicly available.
- Extensive experiments on the challenging AvaMERG benchmark demonstrate that MERIA significantly outperforms all baseline models in generating high-quality, deeply empathetic text responses and achieves the highest human evaluation score in the MERG 2025 Challenge.

2 Related Works

Multimodal Empathetic Response Generation. Human emotional expression is inherently multimodal, with non-verbal cues such as tone of voice and facial expressions playing a crucial role in empathetic communication. As a result, recent research has increasingly focused on multimodal empathetic response generation (MERG). The introduction of the AvaMERG benchmark [19] marked a significant advancement in this field, providing rich conversational data that includes text, speech, and facial video, and spurring the development of baseline systems such as Empatheia, which leverage multimodal large language models (MLLMs) [19]. Although initial successes have been achieved, existing MERG models share a fundamental limitation: they typically rely on simplistic fusion strategies (e.g., feature concatenation or cross-modal attention mechanisms) to integrate multimodal information [15]. While these approaches are effective when emotional signals across modalities align, they fail to perform well in scenarios involving emotional incongruence, a common occurrence in real-world interactions.

Multimodal Representation Disentanglement. To address the challenges posed by emotional signal inconsistencies across modalities, representation disentanglement offers a promising solution. This technique aims to decompose multimodal data into modality-invariant and modality-specific representations, allowing for more precise modeling of the source and role of various information types [8]. In recent years, multimodal disentanglement techniques have demonstrated significant modeling capabilities across various domains. For example, the CAMD model [20] successfully disentangled image-text information in crowdfunding projects into shared "commonality" and independent "specificity" features, utilizing an enhanced network to balance these representations and improve prediction accuracy.

Emotion-Related Fine-Tuning of Multimodal Large Language Models. The rapid development of multimodal large language models (MLLMs) has opened new avenues for achieving more advanced

artificial intelligence, particularly for fine-grained tasks such as emotion recognition. Advanced models, such as Emotion-LLaMA [3], have made significant strides in multimodal emotion recognition and reasoning tasks by integrating audio, visual, and text inputs through instruction tuning. Additionally, data-centric fine-tuning research, such as SoulChat [2], has demonstrated remarkable performance improvements by fine-tuning LLMs on a large-scale, multi-turn empathetic dialogue dataset containing over 2 million samples.

3 Method

To tackle the challenges of fine-grained emotional understanding, cross-modal inconsistency, and deep empathetic reasoning, we introduce MERIA—the Multimodal Empathetic Reasoning and Inconsistency-Aware framework. As depicted in Figure 2, MERIA consists of four major components: (1) a Parallel Disentangled Representation (PDR) module that concurrently extracts shared emotional cues, modality-private information, and speaker identity features from multimodal inputs; (2) an Uncertainty-Gated Fusion (UGF) module that adaptively integrates shared emotional representations by accounting for modality uncertainty; (3) a Multimodal Chain of Empathy (M-CoE) module, where a large language model (LLM) processes the fused multimodal signals and complete dialogue context to construct a structured M-CoE representation; and (4) a Response LLM, which leverages the M-CoE and the dialogue history to generate an empathetic textual response along with the predicted emotional state.

The training process of the MERIA framework is divided into three stages. In the first stage, the PDR module is trained using semi-supervised learning to disentangle multimodal inputs. In the second stage, a Large Language Model (LLM) is fine-tuned via LoRA using manually annotated M-CoE representations, the output signals from the trained PDR module, and the AvaMERG training set, resulting in the M-CoE LLM. In the third stage, the final Response LLM is obtained by further fine-tuning another LLM using the M-CoE outputs generated by the M-CoE LLM along with the AvaMERG training data. During inference, all model weights are frozen, and the data flow follows the pipeline illustrated in Figure 2.

3.1 Parallel Disentangled Representation (PDR) Module

The core component of our framework is the PDR module, which is designed to disentangle raw multimodal signals into three semantically distinct and independent subspaces: shared emotion, modality-private information, and speaker identity. Intuitively, we argue that text modality alone is insufficient to capture gender and age-specific features, and as such, we do not include an identity disentangled representation for the text modality. For each of the three modalities, the PDR module adopts a parallel architecture of two or three specialized β-Variational Autoencoders (β-VAE) [16], resulting in a total of eight encoders.

To handle the diverse nature of multimodal data, we employ specialized pre-processing and feature extraction pipelines for each modality.

Textual Modality: For each speaker sentence, we employ a pre-trained RoBERTa model [4] to obtain emotion-relevant contextualized embeddings X_T that capture nuanced affective semantics.

Audio Modality: We process the raw audio waveform using a pre-trained Emotion2Vec model [9]. The resulting embedding X_A captures not only emotion-relevant features but also rich prosodic and intonational cues, providing a comprehensive representation of the speaker's affective state.

Visual Modality: We utilize the pre-trained LibreFace model [1] to extract visual features X_V from video frames. LibreFace is a multi-task learning framework that produces a comprehensive set of facial representations aligned with our objectives: it encodes facial actions via Action Unit (AU) detection, captures expressive states through emotion classification, and models subtle muscle dynamics and intensity variations underlying micro-expressions via AU intensity estimation.

As illustrated in the PDR module in Figure 2, each component denoted by E represents a pre-trained β-VAE encoder. Given an input X, the encoder produces the parameters of the latent variable distribution, specifically the mean vector μ and the logarithm of the variance vector $\log \sigma^2$, which together define the conditional posterior $q(Z|X)$. Subsequently, the latent representation Z is obtained via the reparameterization trick, i.e., $Z = \mu + \sigma \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

To ensure that the eight parallel encoders learn their designated representations, we jointly train them using a unified, multi-objective loss function as shown in Figure 3. For each modality $\bar{M} \in \{\text{audio, vision}\}$, the input $X^{\bar{M}}$ is encoded and reparameterized into three latent representations: shared emotion $Z_S^{\bar{M}}$, modality-private information $Z_P^{\bar{M}}$, and identity $Z_I^{\bar{M}}$.

Given that speaker utterances in the textual modality typically lack explicit gender or age-related attributes, we omit the identity representation for text. Thus, only Z_S^T and Z_P^T are derived for the text modality. For each modality $M \in \{\text{audio, vision, text}\}$, a decoder reconstructs the input $\hat{X}^M$ from the concatenation of the available latent vectors. The total loss for each modality is defined as a weighted sum of five components:

$$\mathcal{L}_{\text{PDR}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}} + \alpha \mathcal{L}_{\text{align}} + \gamma \mathcal{L}_{\text{ortho}} + \delta \mathcal{L}_{\text{identity}} \quad (1)$$

1. Reconstruction Loss ($\mathcal{L}_{\text{recon}}$): Ensures information integrity by minimizing the Mean Squared Error (MSE) between the original and reconstructed feature vectors.

$$\mathcal{L}_{\text{recon}} = \sum_M \mathbb{E}[\|X^M - \hat{X}^M\|^2] \quad (2)$$

2. KL Divergence Loss ($\mathcal{L}_{\text{KL}}$): A standard component of β-VAEs [16] that regularizes the latent space by encouraging the learned posterior distribution $q(Z|X_M)$ to be close to a standard normal prior $p(Z)$. This is applied to all three latent spaces.

3. Cross-Modal Alignment Loss ($\mathcal{L}_{\text{align}}$): This loss exclusively shapes the *shared emotion* space. It enforces consistency by minimizing the distance between the shared emotion vectors from different modalities within the same dialogue turn.

$$\mathcal{L}_{\text{align}} = \sum_{M,N \in \{\text{text, audio, video}\}, M \neq N} \|Z_S^M - Z_S^N\|^2 \quad (3)$$

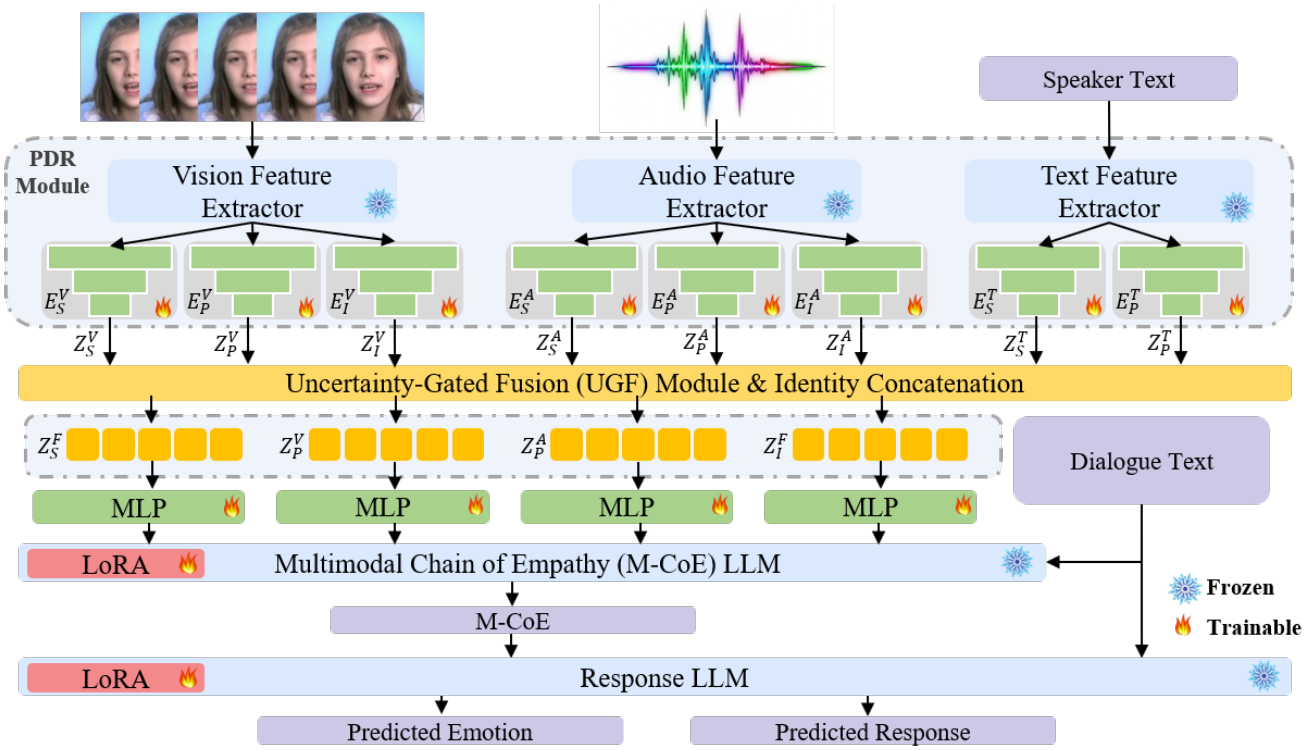


Figure 2: The overall architecture of our proposed MERIA framework. The superscript in the upper right corner denotes the modality: V for Vision, A for Audio, T for Text, and F for Fusion. The subscript in the lower right corner indicates the type of information: S for shared emotion, P for modality-private information, and I for identity. Multimodal inputs are first processed by the Parallel Disentangled Representation (PDR) module to disentangle them into shared emotion, modality-private information, and identity representations in parallel. The UGF module then processes these representations to produce a fused emotion vector and a conflict signal, respectively. The M-CoE LLM integrates all multimodal signals and the complete dialogue history into a structured M-CoE representation. Finally, the Response LLM utilizes the M-CoE and the dialogue history to generate an empathetic response along with the predicted emotion.

4. Orthogonality Loss ($\mathcal{L}_{\text{ortho}}$): To enforce independence between the three subspaces within a single modality, we impose an orthogonality constraint. This loss penalizes correlations between the latent vectors by minimizing the squared dot product.

$$\mathcal{L}_{\text{ortho}} = (Z_S^M \cdot Z_P^M)^2 + (Z_S^M \cdot Z_I^M)^2 + (Z_P^M \cdot Z_I^M)^2 \tag{4}$$

5. Hybrid Identity Loss ($\mathcal{L}_{\text{identity}}$): To fully leverage the annotated identity information (e.g., gender, age) while also capturing unique speaker characteristics, we propose a hybrid loss for the identity space. It combines a classification loss with a metric learning loss. An MLP classifier is attached to the identity encoder to predict the explicit labels, supervised by a Cross-Entropy loss ($\mathcal{L}_{\text{cls}}$). Simultaneously, a Triplet Loss ($\mathcal{L}_{\text{triplet}}$) ensures that utterances from the same speaker are closer than those from different speakers [12].

$$\mathcal{L}_{\text{identity}} = \delta_1 \mathcal{L}_{\text{triplet}} + \delta_2 \mathcal{L}_{\text{cls}} \tag{5}$$

3.2 Uncertainty-Gated Fusion (UGF) Module

Fusing multimodal information naively can be detrimental, especially in the presence of conflicting or noisy signals. The UGF module addresses this by performing an intelligent, uncertainty-aware fusion of the shared emotion representations. Crucially, the gating signal is derived from the uncertainty of the *private* latent space, based on the intuition that a modality is more reliable if it confidently captures a unique, private emotional cue [18].

For each modality M, we first quantify its certainty T^M as the inverse of the variance $\sigma_P^{2(M)}$ of its private latent distribution, aggregated to a scalar.

$$T^M = \frac{1}{\mathbb{E}(\sigma_P^{2(M)}) + \epsilon} \tag{6}$$

where ϵ is a small constant for numerical stability. These certainty scores are then normalized via a Softmax function to produce the final gating weights w_m.

$$[w^V, w^A, w^T] = Softmax([T^V, T^A, T^T]) \tag{7}$$

The final fused shared emotion representation Z_S^F is the weighted sum of the mean vectors Z_S^M from the shared emotion spaces.

$$Z_S^F = \sum_M w^M \cdot Z_S^M \tag{8}$$

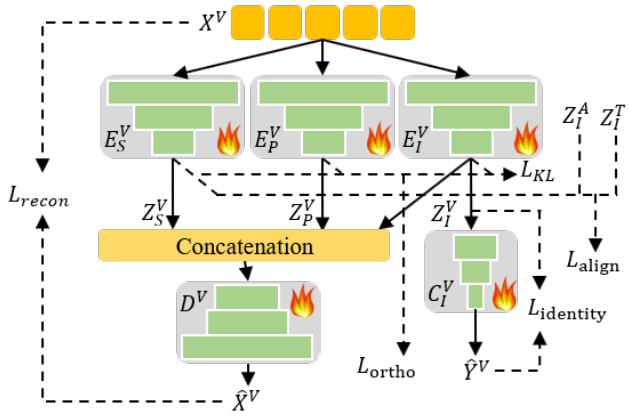


Figure 3: Illustration of the training process of the PDR module using the visual β-VAE encoders as an example. The outputs of the three encoders are concatenated and passed through a decoder D^V to reconstruct the original visual features. An additional MLP classifier C_I^V is attached to the identity encoder E_I^V to predict avatar profile attributes. The audio β-VAE encoders follow the same architecture, while the text β-VAE encoders exclude the identity branch E_I and its associated classifier C_I. All β-VAE encoders are jointly optimized using a unified, multi-objective loss function.

In addition, the identity latent embeddings from the audio and visual modalities, Z_I^A and Z_I^V, are further incorporated in this module. These are concatenated to form a unified identity representation Z_I^F. Similarly, the modality-private information latent embeddings, Z_P^A and Z_P^V, are directly passed to the subsequent module for downstream processing.

3.3 LLM Fine-Tuning

The M-CoE LLM, built upon the Qwen3-32B backbone [13], functions as the framework's core reasoning engine. Its primary role is to integrate the disentangled multimodal representations with the dialogue context to generate a structured M-CoE instance.

To bridge the modality gap between continuous vector embeddings and the LLM's discrete text input, we employ a soft prompting mechanism. As shown in Figure 2, we use lightweight, MLP-based projection networks to map each multimodal representation into a fixed-length sequence of virtual token embeddings. These virtual tokens are then prepended to the tokenized dialogue history, forming a unified input sequence that the LLM processes end-to-end.

The *M-CoE LLM* is trained after all upstream modules have been pre-trained and their parameters frozen. To facilitate supervised fine-tuning, we first construct a high-quality M-CoE dataset. Specifically, we leverage the AvaMERG training set and annotate it with step-by-step multimodal reasoning chains using GPT-4o [7] and Qwen2.5-Omni [17] as assistants. Each annotation follows our predefined M-CoE template: (visual observation, emotion, confidence), (audio observation, emotion, confidence)and a Chain-of-Empathy (CoE) from AvaMERG [19].

We fine-tune the Qwen3-32B model using the generated dataset. The model's task is to generate the reasoning chain given the dialogue history and the multimodal soft prompts. To make this process computationally efficient, we use Low-Rank Adaptation (LoRA) [6]. We freeze the entire LLM backbone and only train the small LoRA adapter matrices injected into the attention layers, along with the parameters of our soft prompt projection networks.

Finally, we also adopt Qwen3-32B as the backbone of our response LLM. The model takes as input the structured M-CoE representation and the corresponding dialogue text. Similar to the M-CoE LLM, we fine-tune the Response LLM using LoRA, with training samples derived from the AvaMERG dataset.

4 Experiments

4.1 Settings

All training and experiments were conducted on a computing cluster equipped with an Intel® Xeon® Platinum 8468 CPU and 8 NVIDIA H800 GPUs, each with 80 GB of memory. For visual feature extraction, we utilize the implementation from the LibreFace library [1]. For audio features, we adopt the Emotion2Vec+ large model [9], while for text, we employ the sentiment roberta large english model [4]. LoRA fine-tuning of the Qwen3-32B [13] backbone is performed using the LLaMA-Factory framework [21]. For all fine-tuning experiments, we adopt a consistent training configuration of 4 epochs. The LoRA module is configured with a rank r of 16 and a scaling factor α of 32. The learning rate is set to 0.0002, while all other hyperparameters follow their default settings.

All experiments are conducted on the AvaMERG dataset [19]. Given that audio and visual variations across different turns within a single dialogue are typically minor, we construct one M-CoE representation per dialogue. This results in a total of 24,696 M-CoE training samples.

In addition, we conduct ablation studies to evaluate the contribution of each module. Since the official ground truth for the AvaMERG test set is not publicly available, we sample 3,427 dialogues from the training set as the ablation test split. The remaining dialogues are used for training to maintain a consistent ratio between training and evaluation data.

4.2 Main Results

Table 1: Performance comparison of top-3 teams and the baseline. Accuracy is reported in percentage. Dist-1 and Dist-2 denote diversity metrics, while Human Evaluation reflects overall human preference. All metrics are the higher, the better.

Team	Acc. (%)	Dist-1	Dist-2	Human Eval.
Baseline [19]	48.51	2.69	14.76	-
AI4AI	41.79	**4.7940**	**31.7850**	4.1
It's MyGO!!!!!	**68.42**	3.9296	23.6446	3.8
DZ (Our)	52.75	3.0628	20.1742	**4.2**

Table 1 presents a comparative analysis of the top-performing teams and the official baseline across four dimensions: Accuracy,

Dist-1, Dist-2, and Human Evaluation. While the team It's MyGO!!!!! achieved the highest classification accuracy (68.42%), and AI4AI led in lexical diversity (Dist-1: 4.7940; Dist-2: 31.7850), our team (DZ) demonstrates a well-balanced performance across metrics and, critically, obtains the highest Human Evaluation score (4.2)—a strong indicator of real-world effectiveness in empathetic response generation.

This result highlights a key insight: accuracy and diversity metrics alone do not fully capture the quality of empathetic dialogues. Although DZ's accuracy (52.75%) is modest compared to the top-scoring team, the superior human preference suggests that our system is better aligned with human expectations of emotionally aware, contextually appropriate responses. This is further supported by competitive Dist-1 (3.0628) and Dist-2 (20.1742) scores, indicating our model's ability to produce varied and non-repetitive responses without sacrificing coherence or empathy.

The strong human-centric performance of team DZ validates the design philosophy of our framework, which prioritizes deep emotional reasoning, modality-awareness, and structured representation over raw classification scores. It underscores the importance of designing empathetic AI systems that optimize for subjective human experience, rather than relying solely on rigid quantitative metrics.

4.3 Ablation Experiment Results

Table 2: Ablation study on the MERIA framework components. "w/o" denotes "without".

Model	Acc. (%)	Dist-1	Dist-2
MERIA (Full Model)	**58.45**	**3.2914**	**21.0298**
w/o PDR (β-VAE)	43.16	2.7084	15.0855
w/o UGF	52.70	3.1592	20.4531
w/o M-CoE LLM	55.32	3.2816	20.9890

We conducted an ablation study to validate the contribution of each key component in MERIA, with results presented in Table 2.

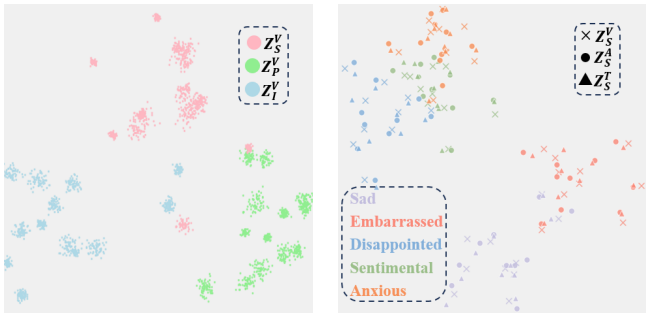
The removal of the PDR module, implemented via a β-VAE, causes the most significant performance degradation across all metrics. This confirms that disentangling multimodal signals into distinct semantic subspaces is fundamental for understanding nuanced emotional cues.

Excluding the Uncertainty-Gated Fusion (UGF) module also impairs performance, underscoring its importance in adaptively weighing and integrating information from different modalities to improve fusion quality.

Finally, omitting the M-CoE LLM primarily affects accuracy while having a minimal impact on diversity metrics. This suggests its main contribution is to enhance the generation of structured and logically coherent empathetic reasoning chains, leading to more contextually appropriate responses.

4.4 Qualitative Analysis of Disentanglement

To validate the effectiveness of our Parallel Disentangled Representation (PDR) module, we visualize the learned latent embeddings using t-SNE in Figure 4. Figure 4a illustrates the successful separation of semantic factors within a single modality. The clear spatial distinction between shared emotion (Z_S^V), private (Z_P^V), and modality-invariant (Z_I^V) representations confirms that our module can effectively isolate these distinct components. Figure 4b demonstrates robust alignment of shared emotion embeddings across different modalities. Within each fine-grained emotion cluster, the embeddings from video (×), audio (•), and text (△) are tightly intermingled. This shows that our PDR module, guided by the alignment loss, learns a modality-invariant representation of emotion, successfully factoring out modality-specific characteristics.



(a) (b)

Figure 4: T-SNE visualization of the latent space embeddings. (a) T-SNE visualization of the latent space embeddings derived from the visual β-VAE encoders; (b) T-SNE visualization of shared emotion embeddings for five hard-to-classify fine-grained emotions.

5 Conclusion

In this paper, we proposed MERIA, a novel framework for multimodal empathetic text generation. Due to human instincts, emotional inconsistencies often arise between different modalities, especially when audiovisual data is involved. Moreover, existing open-source datasets lack comprehensive multimodal chains of empathy, which limits the application of large multimodal models in empathetic scenarios. To address these issues, we (Team DZ) introduced a multimodal disentanglement encoder based on β-VAE and extended the chain of empathy (CoE) in the AvaMERG dataset to include multimodal aspects. Our proposed MERIA achieved the best human evaluation scores in empathetic text response generation task in the MERG 2025 Challenge. Additionally, we conducted ablation studies and visualized latent embeddings to assess the contribution of each module, further validating the disentanglement capability of the β-VAE encoders. These findings demonstrate the effectiveness of MERIA in enhancing multimodal empathy generation, paving the way for more advanced empathetic models in future applications.

References

[1] Di Chang and Yin. 2024. LibreFace: An Open-Source Toolkit for Deep Facial Expression Analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI, USA) *(WACV '24)*. IEEE/CVF, Piscataway, NJ, USA, 8205–8215. https://openaccess.thecvf.com/content/WACV2024/html/Chang_LibreFace_An_Open-Source_Toolkit_for_Deep_Facial_Expression_Analysis_WACV_2024_paper.html

[2] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1170–1183. doi:10.18653/v1/2023.findings-emnlp.83

[3] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* 37 (2024), 110805–110853.

[4] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* 40, 1 (2023), 75–87. doi:10.1016/j.ijresmar.2022.05.005

[5] Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. 2023. *Affective Conversational Agents: Understanding Expectations and Personal Influences.* arXiv:2310.12459 doi:10.48550/arXiv.2310.12459

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. *GPT-4o System Card.* arXiv:2410.21276 doi:10.48550/arXiv.2410.21276

[8] Jiayang Li, Xovee Xu, Yili Li, Ting Zhong, Kunpeng Zhang, and Fan Zhou. 2025. Commonality Augmented Disentanglement for Multimodal Crowdfunding Success Prediction. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Piscataway, NJ, USA, 1–5. doi:10.1109/ICASSP49660.2025.10889564

[9] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15747–15760. doi:10.18653/v1/2024.findings-acl.931

[10] Jingbo Meng, Renwen Zhang, Jiaqi Qin, Yu-Jen Lee, and Yi-Chieh Lee. 2025. AI-mediated social support: the prospect of human–AI collaboration. *Journal of Computer-Mediated Communication* 30, 4 (07 2025), zmaf013. arXiv:https://academic.oup.com/jcmc/article-pdf/30/4/zmaf013/63752662/zmaf013.pdf doi:10.1093/jcmc/zmaf013

[11] Zhibang Quan, Tao Sun, Mengli Su, and Jishu Wei. 2022. Multimodal Sentiment Analysis Based on Cross-Modal Attention and Gated Cyclic Hierarchical Fusion Networks. *Computational Intelligence and Neuroscience* 2022, 1 (2022), 4767437.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, Piscataway, NJ, USA, 815–823. doi:10.1109/CVPR.2015.7298682

[13] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[14] Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. *Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects.* arXiv:2505.20511 doi:10.48550/arXiv.2505.20511

[15] 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/

[16] Baao Xie, Qiuyu Chen, Yunnan Wang, Zequn Zhang, Xin Jin, and Wenjun Zeng. 2024. Graph-based unsupervised disentangled representation learning via multimodal large language models. *Advances in Neural Information Processing Systems* 37 (2024), 103101–103130.

[17] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. *Qwen2.5-Omni Technical Report.* arXiv:2503.20215 doi:10.48550/arXiv.2503.20215

[18] Yangshuyi Xu, Lin Zhang, and Xiang Shen. 2023. Multi-modal adaptive gated mechanism for visual question answering. *Plos one* 18, 6 (2023), e0287557.

[19] 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* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 2872–2881. doi:10.1145/3696410.3714739

[20] Qing Zhang, Jing Zhang, Xiangdong Su, Yonghe Wang, Feilong Bao, and Guanglai Gao. 2025. Domain disentanglement and fusion based on hyperbolic neural networks for zero-shot sketch-based image retrieval. *Information Processing & Management* 62, 1 (2025), 103963.

[21] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (Bangkok, Thailand) *(ACL 2024)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 400–410.