Towards Multimodal Empathetic Response Generation: A Rich Text-Speech-Vision Avatar-based Benchmark

Han Zhang
School of Electronic Engineering, Xidian University
Xi'an, Shaanxi, China
22021110280@stu.xidian.edu.cn

Zixiang Meng
School of Cyber Science and Engineering, Wuhan University
Wuhan, China
zixiangmeng@whu.edu.cn

Meng Luo
National University of Singapore
Singapore, Singapore
mluo@u.nus.edu

Hong Han
School of Electronic Engineering, Xidian University
Xi'an, Shaanxi, China
hanh@mail.xidian.edu.cn

Lizi Liao
Singapore Management University
Singapore, Singapore
lzliao@smu.edu.sg

Erik Cambria
Nanyang Technological University
Singapore, Singapore
cambria@ntu.edu.sg

Hao Fei*
National University of Singapore
Singapore, Singapore
haofei37@nus.edu.sg

Abstract

Empathetic Response Generation (ERG) is one of the key tasks of the affective computing area, which aims to produce emotionally nuanced and compassionate responses to user's queries. However, existing ERG research is predominantly confined to the singleton text modality, limiting its effectiveness since human emotions are inherently conveyed through multiple modalities. To combat this, we introduce an avatar-based Multimodal ERG (MERG) task, entailing rich text, speech, and facial vision information. We first present a large-scale high-quality benchmark dataset, **AvaMERG**, which extends traditional text ERG by incorporating authentic human speech audio and dynamic talking-face avatar videos, encompassing a diverse range of avatar profiles and broadly covering various topics of real-world scenarios. Further, we deliberately tailor a system, named **Empatheia**, for MERG. Built upon a Multimodal Large Language Model (MLLM) with multimodal encoder, speech and avatar generators, Empatheia performs end-to-end MERG, with Chain-of-Empathetic reasoning mechanism integrated for enhanced empathy understanding and reasoning. Finally, we devise a list of empathetic-enhanced tuning strategies, strengthening the capabilities of emotional accuracy and content, avatar-profile consistency across modalities. Experimental results on AvaMERG data demonstrate that Empatheia consistently shows superior performance than baseline methods on both textual ERG and MERG. All data and code are open at https://AvaMERG.github.io/.

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, Avatar Generation, Affective Computing

ACM Reference Format:
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 Web Conference 2025 (WWW '25), April 28–May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714739

1 Introduction

In recent years, the advent of Large Language Models (LLMs) [8–10, 20, 45, 52] has endowed machines with unprecedented levels of intelligence, bringing us closer to the realization of Artificial General Intelligence (AGI). However, the true essence of AGI extends beyond merely achieving human-level intelligent abilities; it must also encompass emotional understanding and empathetic capabilities comparable to those of humans. For instance, during human-machine interactions, it is crucial for machines to comprehend human emotions and intentions [12, 17, 28, 29, 53]. This necessity has driven the development of Empathetic Response Generation (ERG) [36], a task aimed at enabling machines to produce emotionally nuanced and compassionate responses to user queries, thereby facilitating emotion-aware conversations. Over the past decade, ERG has garnered significant research attention [31, 38, 49].

*Hao Fei is the corresponding author.

WWW '25, April 28–May 2, 2025, Sydney, NSW, Australia.

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714739

Figure 1: A snippet of avatar-based Multimodal Empathetic Response Generation (MERG) with rich multimodal signals: text (dialogue), audio (acoustic speech) and vision (dynamic talking-head avatar).

Due to its ability to support emotional interactions with humans, ERG has been applied in various practical scenarios, such as psychological therapy and elderly companionship dialogue systems.

However, current ERG research might encounter significant challenges due to its confinement to a singleton textual modality as task definition. It is worthwhile to reflect on how humans naturally express emotions; in many cases, the subtleties of emotions are more effectively and comprehensively conveyed through non-textual modalities. Specifically, in dynamic visual contexts, subtle facial expressions and body movements can communicate richer emotions and intentions. Simultaneously, in the auditory domain, variations in speech intonation and pitch can also convey emotional states that text alone cannot express. Figure 1 demonstrates a multimodal empathetic dialogue process. Existing text-based ERG tasks are restricted to providing users with mere textual responses, which lack enough warmth and emotional resonance inherent in human interactions, thereby falling short of achieving adequate empathetic effects. Furthermore, from the user's perspective, there is a desire to express emotions directly through speech or talking-facial video rather than being confined to text-based queries. In practical applications, numerous ERG scenarios require the ability to accept multimodal signal inputs and generate empathetic responses in multimodalities, such as in psychological therapy, companion robots, and electronic personal assistants. Unfortunately, there has yet to be any research on avatar-based Multimodal Empathetic Response Generation (MERG) within the community.

To bridge this gap, in this paper we present an **Avatar-based Multimodal Empathetic Response Generation** benchmark dataset (namely, **AvaMERG**). Building upon existing text-based ERG benchmark [36], we further augment the dataset to include multimodal signals and annotations. Specifically, for each utterance in the dialogue, we provide 1) authentic human-reading speech and 2) dynamic talking-face avatar videos (2D facial modeling) that both correspond to the intended emotion. AvaMERG features a wide variety of avatar profiles and covers broad common topics of real-world scenarios, including multiple age groups, genders, vocal tones, intonations, and appearances, thereby effectively simulating a diverse range of multimodal empathetic dialogue scenarios in realistic environments. We maintain the high quality of annotations through meticulous manual verification, guaranteeing the emotional accuracy and consistency of both the avatars' speech and video. Finally, we compile 33,048 annotated dialogues with 152,021 multimodal utterances, establishing a foundation for MERG research.

A direct approach to generating multimodal empathetic responses can be first producing the textual part of the response using existing text-based ERG models (e.g., high-performing LLMs), and then through a pipeline paradigm to invoke external well-trained speech generator and talking-head generator (e.g., diffusion-based models) to generate the corresponding multimodal content. However, there can be several non-trivial issues and inherent challenges. **First**, ensuring the emotional accuracy across the text, audio, and video is the most fundamental capability. **Second**, it is essential to maintain synchronization and consistency among the three modalities in terms of content, emotion, and style. Pipeline models often suffer from inadequate interaction between different modules, making it difficult to guarantee consistency. For example, the generated speech may convey the emotion of a happy girl, while the corresponding avatar depicts a crying boy. **Third**, the discrete approach (where LLMs invoke external audio and video generators) can largely lead to the quality decrease of the generated content due to error propagation.

To achieve high-quality MERG, we thus propose a novel Multimodal LLM, termed **Empatheia**. Architecturally, we employ a multimodal encoder to feed all input signals into the central LLM for comprehension and reasoning. We then utilize StyleTTS2 [22] as the speech generation module and DreamTalk [30] as the Talking Face Generation module. By using continuous embeddings as the medium for message passing, we connect the LLM to the frontend encoders and backend cross-modal generation modules, resulting in a full end-to-end system. Next, we optimize Empatheia by implementing a series of tuning strategies. We first devise a *Chain-of-Empathetic Inference* to assist the LLM to reason step-by-step, from understanding the emotion to identifying the underlying rationale and intent, and ultimately determining how to respond to the user's input. Then, we introduce *Content Consistency Learning*, which encourages the LLM to guide the two backend modules to produce speech and talking-face avatar videos that align with the empathetic textual content. Further, we propose a *Style-aware Alignment and Consistency Learning* mechanism to accurately identify the style signals transmitted by the central LLM, and ensure consistency in the style of both speech and video avatars, including emotion and profile. Finally, we perform overall MERG tuning to achieve overall high-quality multimodal empathetic responses.

We conduct experiments on the AvaMERG dataset, where the results demonstrate that our Empatheia system generates both textual and multimodal empathetic responses of higher quality compared to baseline models. In-depth analyses further reveal the underlying rationales for our model's advancements. Overall, this work pioneers the research of MERG, contributing a benchmark dataset and a strong-performing end-to-end MERG model, laying a solid foundation for future exploration in multimodal empathetic response generation.

2 Related Work

ERG [33, 34] is one of the crucial tasks within the field of affective computing, which aims at enabling dialogue models to produce responses imbued with empathy during human-machine conversations. Due to its significant practical applications, ERG has attracted substantial and sustained prior research attention [11, 24, 54]. Existing studies have developed various methods to enhance the performance of ERG systems [2, 13, 38, 50].

Yet current ERG approaches can be limited to a single text modality, which significantly restricts their effectiveness. In real-world dialogue scenarios, multiple modalities are often involved. As previously emphasized, multimodal information is crucial for generating more empathetic responses. Therefore, this paper tries to pioneer the research of Multimodal Empathetic Response Generation (MERG) by presenting a novel benchmark. It is also noteworthy that several recent related works have also touched upon multimodal ERG [48, 51].

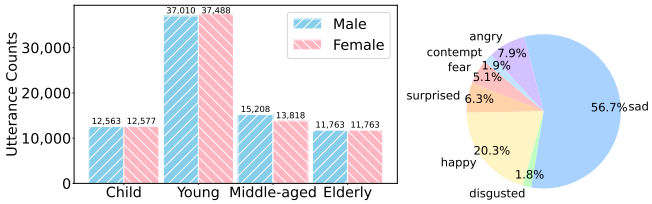
However, we emphasize that these studies do not fully address or cover all the modalities most relevant to empathy. Intuitively, both audio (capturing variations in a person's tone) and visual (capturing facial expressions) modalities can be important, and need to be simultaneously addressed. Moreover, it is insufficient to rely solely on emoticon-type visual features. Effective ERG that closely aligns with real-world application scenarios should present authentic facial visual signals.

Unlike existing text-based ERG models and methods, achieving multimodal emotional understanding and generating multimodal signals requires the utilization of multimodal-related technologies. First, our approach is related to research on Multimodal Large Language Models (MLLMs), with our system being based on a backbone MLLM. Various MLLMs, such as LLaVA [25], MiniGPT-4 [55], have been investigated and widely validated for their strong semantic understanding capabilities. However, most MLLMs are limited to multimodal information comprehension yet do not support the flexible generation of diverse modal content beyond text [1, 19, 39], such as audio and visual outputs. Although there are a few MLLMs that support the generation of various modal signals, such as NExT-GPT [46] and Unified-IO 2 [27], these models, unfortunately, are only capable of understanding and generating signals in general scenarios. They lack sufficient capabilities in emotion detection and emotional content generation. In other words, these MLLMs are unable to generate emotionally expressive speech or talking-face avatars. Therefore, we consider developing a novel MLLM for MERG, which is able to accurately generate emotionally charged speech and talking-face avatar videos. Additionally, we design a series of emotion-enhancement training strategies to ensure that our MLLM possesses highly-performing MERG capabilities.

3 AvaMERG Benchmark

3.1 Task Definition of MERG

Given a multimodal dialogue $\hat{D}=(Q_i|D_{<i})$, where Q_i denotes the current i-th round multimodal user query input, and $D_{<i}$ represents the dialogue history, MERG task is to produce a contextually appropriate and empathetic multimodal response R_i for Q_i, with each utterance (i.e., Q_i and R_i) consisting of three content-synchronized modalities: text t_i, speech audio s_i, and talking-face video v_i, i.e., $Q_i/R_i=(t_i^{q/r}, s_i^{q/r}, v_i^{q/r})$. This results in $D_i=\{(Q_1, R_1), \ldots, (Q_i, R_i)\}$, a total of i round of a multimodal dialogue, includes the user query Q_i and model response R_i. The task requires maintaining coherence and emotional congruence across these modalities to ensure that the generated response R_i well aligns with the emotional cues in user input and also context.



(a) Distribution of the utterance. **(b) Eemotion distribution.**

Figure 2: Visualized statistics of AvaMERG dataset.

Table 1: Statistics of AvaMERG dataset.

	Item	Stats
Dialogue	#Train Set	24,696
	#Valid Set	4,373
	#Test Set	3,979
	#Total	33,048
	Avg. Words Per Utterance	14.68
	Avg. Utterance Per Dialigue	4.6
Modality	Utterance Text	152,021
	Speech Audio	152,021
	Talking-head Video	152,021
	Avg. Length (Sec) Per Aud/Vid	5.67
Avatar	Child (Male/Female)	3/3
	Young (Male/Female)	25/17
	Middle-aged (Male/Female)	4/4
	Elderly (Male/Female)	5/4
	Tone (Emphatic/Mild/Gentle)	14/38/13
	Race	5
Emotion	Text/Multimodal	32/7
Topic&Scenario		10

3.2 Dataset Construction

We construct our *Ava-MERG* dataset by augmenting the existing pure-text ERG dataset, *Empathetic Dialogue* (ED) [36], where the textual empathetic response $t_i \in R_i$ with the query's corresponding emotion categories. First, we consider enriching the data with the identity information for both participants in the dialogue, including ages, genders, and also tone, such that MERG models can learn the correct avatar profile for both audio and video.

As the OpenAI GPT-4[1] has been validated for its remarkable performance in context understanding and thus extensively employed for data generation [28?], here we also adopt GPT-4 for our annotation. We define four age periods (*child*, *young*, *middle-aged*, *elderly*), binary genders (*male*, *female*), and three vocal tones (*emphatic*, *mild*, *gentle*). We ask GPT-4 to determine the above labels for each utterance in ED. Since the data in the raw ED is ill-balanced, e.g., most of the dialogues occurred between young or middle-aged participants, we further employ GPT-4 to produce more dialogue of ERG with above meta-information. Also, GPT-4 will detect the dialogue topics. Human annotators with 3-person cross-checking

[1] https://openai.com/index/gpt-4/, June, 2024

are recruited here to carefully check if the dialogue content, the meta-profile, and the topics are correct and of high quality. This led to the textual part of our AvaMERG data.

Next, we create the multimodal part of the information. First, we recruit a big number of English-speaking volunteers of the above different ages, genders, and vocal characteristics, and also different races (i.e., Asian, Caucasian, African, Latino, Indian). Then, we assign and group different pairs of two participants according to the profile determined in the AvaMERG dialogue. Next, we let these annotators carefully read the utterance text, with the correct emotional performance, including the tone, pitch, timbre and micro-facial expressions, where we then record their vocal speeches and talking-head videos. After the recordings, we recruit another group of well-trained annotators to evaluate each dialogue for content accuracy and emotional accuracy with same 3-person cross-checking. We ask each annotator to check: 1) whether the speech and video content match the content in textual utterance; 2) whether the speech and video style (including age, gender, tone, emotion) are consistent. Only the instance will be accepted where all three annotators vote for approval. This results in the final AvaMERG dataset.

3.3 Dataset Highlight

The data statistics are detailed in Table 1 and Figure 2. Here we summarize the data characteristics that are key to MERG. Due to the space limitation, we show the complete data description and statistics in Appendix §??.

Large Scale and High Quality. AvaMERG comprises a total of 33,048 dialogues with 152,021 utterances, which is large-scale enough to uncover the immense potential of the task. Also the construction undergoes a rigorous manual checking involving both textual and multimodal content verification, ensuring its high quality.

Multimodal Dialogue. Dialogues in AvaMERG cover three modalities: text, speech, and avatar video, which overcome the limitation of single-modality in existing textual ERG benchmarks.

Avatar Profile Diversity. The avatars encompass 4 distinct age groups, with each represented by male and female in 3 different vocal tones. Also avatars come from different races. This rich diversity of avatar profiles ensures the robustness of the MERG.

Emotion Diversity. AvaMERG includes 7 commonly occurred emotions: *sad*, *disgusted*, *surprised*, *contempt*, *happy*, *fear*, and *angry*.

Broad Topic Coverage. AvaMERG covers 10 primary common topics of real empathetic dialogue, along with hundreds of specific subtopics, fully covering the wide range of potential real-world applications for ERG.

4 Empatheia: MERG System

Figure 3 illustrates the overall architecture of our Empatheia system. Overall, Empatheia consists of three main blocks: multimodal encoding layer, LLM-based core reasoning layer, and multimodal generation layer.

4.1 Multimodal Encoder

To perceive the multimodal dialogue inputs, we employ the HuBERT [14] and CLIP ViT-L/14@336px [35] as the speech encoder and avatar video encoder. Essentially, the latent representations

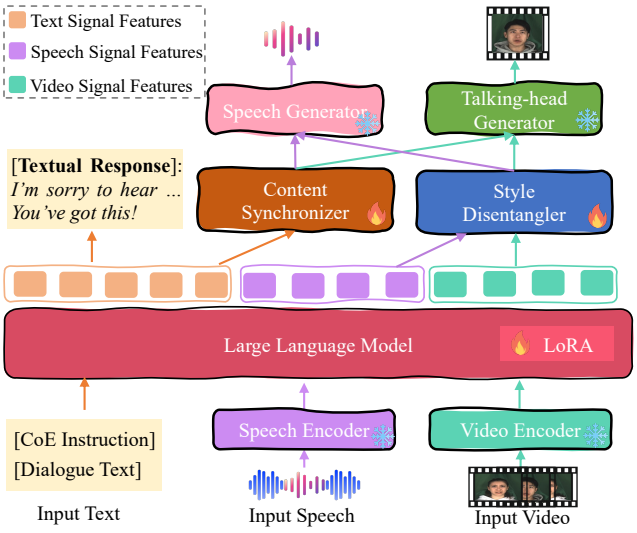


Figure 3: Architecture of our Empatheia MLLM for MERG.

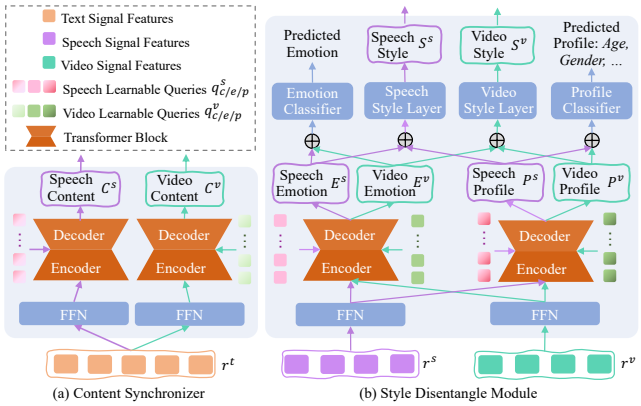
of synchronous text, speech, and talking face video should convey consistent semantics, meaning that ideally, their embeddings are aligned. We thus align the speech and avatar encoders' representation into the LLM's language semantic space via projections.

4.2 LLM-based Core Reasoner

LLM Backbone. The LLM serves as the "brain" of our system, responsible for understanding multimodal signals, reasoning about appropriate empathetic responses, and sending signals for multimodal generation. Given that Vicuna [3] is widely adopted as a baseline for MLLMs [6, 23] and demonstrates superior performance, we select it as our backbone LLM. After encoding the input multimodal dialogue $\hat{D}$, LLM is expected to output the representations of 1) text tokens r_i^t, 2) speech signal tokens r_i^s, and 3) video signal tokens r_i^v. Here r_i^s and r_i^v entail rich emotion and style features, which all will be used for controlling the follow-up modules.

Chain-of-Empathy Reasoning. Empathy is an advanced human capability that is challenging to interpret, and individuals often engage in several steps of contemplation before responding as listeners. Inspired by Chain-of-Thought [7, 47], we design a Chain-of-Empathy (CoE) reasoning mechanism. Specifically, we guide the LLM to think through the following progressive steps to gradually derive the final empathetic responses more accurately and more interpretably.

> • **CoE Instruction**:
> You are an empathetic conversational agent. Your goal is to understand the user's emotions and intentions, and respond or comfort them with appropriate language that helps them feel understood and cared for. Avoid rushing into your response; instead, carefully consider each step before replying by following these steps, one by one:
> ► **Step-1.** *Event scenario.* Reflect on the event scenarios that arise from the ongoing dialogue.
> ► **Step-2.** *User's emotion.* Analyze both the implicit and explicit emotions conveyed by the user.
> ► **Step-3.** *Emotion cause.* Infer the underlying reasons for the user's emotions.



(a) Content Synchronizer (b) Style Disentangle Module

Figure 4: Illustration of the Content Synchronizer and Style Disentangle modules.

- **Step-4.** *Goal to response.* Determine the goal of your response in this particular instance, such as alleviating anxiety, offering reassurance, or expressing understanding.
- **Step-5.** *Generating empathetic response.* Formulate a response that addresses the user's emotions and situation, ensuring it reflects the reasoning from the previous steps. The output should be purely focused on providing a thoughtful and empathetic reply.

These steps simulate the thought process that humans typically engage in. In the following §5.1 we expand the training of the CoE reasoning on our system.

4.3 Multimodal Generation

Multimodal Generator Backbones. Following the signal features (r_i^t, r_i^s, r_i^v) from LLM, the backbone speech generator and talking-head generator will produce the non-textual contents, respectively. To ensure high-quality multimodal generation, we employ the current state-of-the-art StyleTTS2 [22] and DreamTalk [30], respectively. Note that these generators are well-trained before integrating into our system. However, directly generating speeches and dynamic avatars would largely lead to the issues of inconsistency of both content and style. That is, two aspects of consistency are required: 1) **Consistency of content,** both the speech should be synchronized with the talking-head video, both of which should be further aligned with the textual response; 2) **Stylistic Coherence,** the style within text/speech/vision, including both the emotion and profile (age, gender, tone, appearance), should be kept consistent. For natural and accurate MERG, maintaining synchronized content and style across modalities is crucial.

For these purposes, we further design two modules before the two generators: content synchronizer and style disentangler.

Content Synchronizer. The content synchronizer (CS) aims to ensure that the speech and vision generators receive the correct response content information. As shown in Figure 4(a), the module is essentially a Transformer-based [42] variational auto-encoder (VAE) [16]. mainly consists of two transformer blocks, which CS encodes the r^t into latent representation z_c, from which the decoder reconstructs the content of speech C^s and vision C^v.

$$z_c^{s/v} = \text{Enc}^{\text{CS}}(\text{FFN}(r^t), q_c^{s/v}), \tag{1}$$

$$C^{s/v} = \text{Dec}^{\text{CS}}(\text{FFN}(z_c^{s/v}), q_c^{s/v}), \tag{2}$$

where q_c^s and q_c^v represent learnable content query features for two modalities, which are fed into the decoder along with the output from the encoder. C^s guides the speech generator to produce speech that correctly delivers the response text, while C^v guides the talking-head generator to generate accurate mouth movements reflecting the response text.

Style Disentangler. Style features (including emotions and profiles) can be subtly different in speech module and vision module. The style disentangler (SD) module thus aims to disentangle the style features from the LLM-output r_i^s and r_i^v, for two modules, respectively. As shown in Figure 4(b), similar to CS module, SD also uses VAE blocks to disentangle the emotion and profile representations for speech and video:

$$z_e^{s/v} = \text{Enc}^{\text{SD}}(\text{FFN}(r^s), q_e^{s/v}), \tag{3}$$

$$E^{s/v} = \text{Dec}^{\text{SD}}(\text{FFN}(z_e^{s/v}), q_e^{s/v}), \tag{4}$$

$$z_p^{s/v} = \text{Enc}^{\text{SD}}(\text{FFN}(r^s), q_p^{s/v}), \tag{5}$$

$$P^{s/v} = \text{Dec}^{\text{SD}}(\text{FFN}(z_p^{s/v}), q_p^{s/v}), \tag{6}$$

where $E^{s/v}$ are the disentangled emotion features. $P^{s/v}$ are the corresponding profile features. $q_e^{s/v}$ and $q_p^{s/v}$ denote the learnable query features. Then, we fuse the $E^{s/v}$ and $P^{s/v}$ by a speech/video style layer, and obtain the final speech/video style feature:

$$S^{s/v} = E^{s/v} \oplus P^{s/v}, \tag{7}$$

which will be passed to two generators separately. To further regulate the successful extraction of emotional and profile-aware features, we also fuse the emotion feature E^s and E^v into E, and the profile feature P^s and P^v into P. Then we use an emotion classifier and a set of profile classifiers to predict the labels of emotion, avatar's age, gender, and tone.

5 Empathetic-enhanced Training Strategy

With the above Empatheia model architecture, we now empower it with effective MERG capability via a series of training strategies.

5.1 Chain-of-Empathy Training

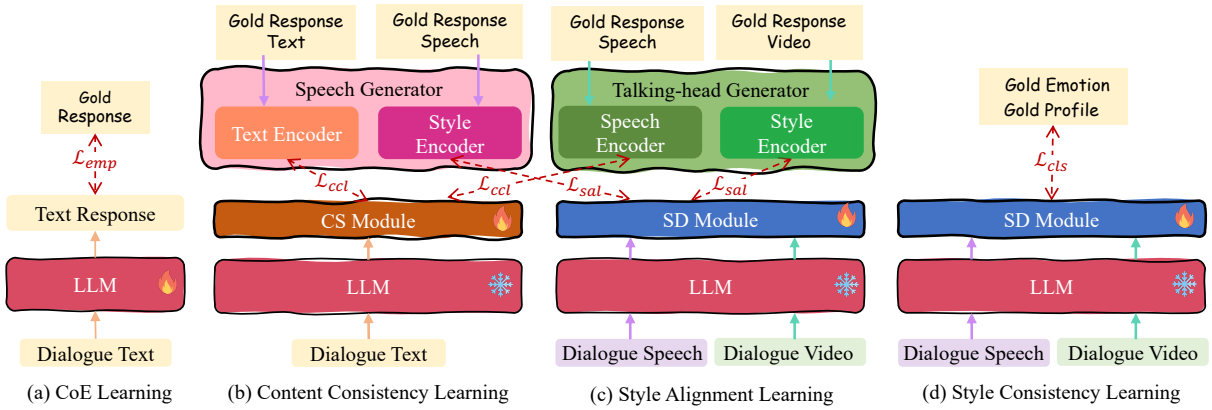
For the first stage, to teach Empatheia to learn how to perform CoE, we perform supervised fine-tuning. For this training, we annotate a set of CoE labels based on a subset of the Ava-MERG training data. Then, as shown in Figure 5(a), this training only updates the core LLM part for text generation, with Lora [15] technique.

$$\mathcal{L}_{emp} = -\sum_{i=1}^{N} \log P(x_i | x_1, \cdots, x_{i-1}), \tag{8}$$

where x_i denotes the output token of the LLM at i-th time step. Upon completion of training, the LLM is capable of not only generating empathetic responses but also providing a comprehensive CoE reasoning process.

5.2 Content Consistency Learning

The aim of the second training stage is to encourage the content signals output by CS module to guide the multimodal generator in producing content-consistent speech and video. This requires aligning the content representations of both sides. Therefore, as



(a) CoE Learning (b) Content Consistency Learning (c) Style Alignment Learning (d) Style Consistency Learning

Figure 5: Illustrations of the proposed training strategies.

shown in Figure 5(b), we minimize the Euclidean distance between C^s and the text embedding $\hat{C}^s$ encoded by the text encoder in the speech generator, as well as the distance between C^v and the audio embedding $\hat{C}^v$ encoded by the audio encoder in the talking-head generator:

$$\mathcal{L}_{ccl} = \|C^s - \hat{C}^s\|_2^2 + \|C^v - \hat{C}^v\|_2^2 . \tag{9}$$

Since the input text for the speech generator and the input audio for the video generator are well paired, the CS module naturally produces consistent multimodal content signal features after training. In this stage, we keep the LLM frozen to prevent it from forgetting the empathetic response capability.

5.3 Style Alignment and Consistency Learning

Style Alignment Learning. For the third stage, on the one hand, we aim to align the style features, ensuring that the multimodal generators accurately interpret the style signals provided by the SD module. As illustrated in Figure 5(c), we minimize the Euclidean distance between S^s (Equation 7) and the audio style features $\hat{S}^s$ encoded by the style encoder in the speech generator, as well as between S^v and the video style features $\hat{S}^v$:

$$\mathcal{L}_{sal} = \|S^s - \hat{S}^s\|_2^2 + \|S^v - \hat{S}^v\|_2^2 . \tag{10}$$

Style Consistency Learning. On the other hand, the target style features are not only exclusively composed of the predefined emotion and profile features, but also include additional modality-specific representations. For example, video style features may depict facial variations under specific emotional states. To further ensure style consistency across modalities, we constrain the SD to disentangle pure emotion and profile representations. We here introduce two classification losses for emotion and profile prediction:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i}^{N}\Big(\sum_{c}^{M_e} y_{i,c}\log(p_{i,c}) + \sum_{p}^{M_p}\sum_{c}^{p} y_{i,c}\log(p_{i,c})\Big), \tag{11}$$

where M_e represents the number of emotion categories, and M_p is the set of categories for gender, age, and tone. In this stage, we also fix the LLM to prevent loss of previously acquired capabilities. In summary, the total loss for the third stage is:

$$\mathcal{L}_{sac} = \mathcal{L}_{sal} + \mathcal{L}_{cls} . \tag{12}$$

5.4 Overall MERG Tuning

The previous training steps effectively decompose the MERG task into sub-processes of separate capabilities. To enhance the overall performance of MERG, comprehensive end-to-end fine-tuning is necessary. In this stage, we integrate all previous training processes, and jointly fine-tune the LLM, CS, and SD modules. The overall loss can be denoted as:

$$\mathcal{L}_{oal} = \mathcal{L}_{emp} + \alpha\mathcal{L}_{ccl} + \beta\mathcal{L}_{sac} . \tag{13}$$

By jointly optimizing the components, we aim to improve the consistency and accuracy of the generated speech and video outputs, while maintaining the empathetic dialogue capabilities learned in earlier stages. Furthermore, this unified fine-tuning stage allows the model to leverage cross-modal interactions more effectively, resulting in a more robust and coherent multimodal generation system tailored to the MERG task.

6 Experiment

6.1 Settings

Baseline. In our preliminary experiment, to identify the most suitable backbone LLM, we compare Flan-T5 XXL [4], ChatGLM3-6B [40], and Vicuna-7B [3]. Besides MERG, we also compare the text ERG performance with existing models, including KEMP [21], CEM [38] and CASE [54], where we evaluate our Empatheia using only text queries for generating textual responses only. Since no prior work addresses the MERG task, for the speech and video generation, we develop a pipeline-based baseline, where the LLM only outputs the invocation commands for the two backend multimodal generators, without feature embedding passing and end-to-end joint training. It first generates response text from the LLM, then passes the text into StyleTTS2 [22] to synthesize speech, and then processes the speech using DreamTalk [30] to generate the corresponding talking-head video.

Evaluation Metrics. For the text ERG task, we employ three evaluation metrics: Emotion Accuracy (Acc), and Distinct metrics (Dist-1 and Dist-2) [18]. For speech generation, we use the 5-scale Mean Opinion Score (MOS) [43] and Similarity MOS (SMOS) [26]. For talking head generation, we adopt the Cumulative Probability of Blur Detection (CPBD) [32], Structural Similarity Index Measure (SSIM) [44] and SyncNet confidence score (Sync_{cf}) [5].

Table 2: Comparisons of textual ERG on AvaMERG data. ↑: the higher the better; ↓: the lower the better.

Model	Acc ↑	Dis-1 ↑	Dis-2 ↑
KEMP [21]	35.87	0.41	1.78
CEM [38]	37.32	0.50	2.07
CASE [54]	40.96	0.54	2.14
Empatheia	**48.51**	**2.69**	**14.76**
w/o CoE	46.62	2.49	12.77
w/o SPC&VID	45.89	2.43	12.56

We also consider human evaluations. For textual ERG, we employ 4 human evaluation metrics: Empathy (Emp.), Coherence (Coh.), Informativity (Inf.), and Fluency (Flu.). For MERG, we newly define 6 metrics: Speech Content Accuracy (SCA), Video Content Accuracy (VCA), Speech Style Accuracy (SSA), Video Style Accuracy (VSA), Multimodal Content Consistency (MCC), and Multimodal Style Consistency (MSC).

Implementation Details. We fine-tune our model using LoRA [15] and DeepSpeed [37] techniques on a single 80GB A100 GPU. Each Transformer block comprises four encoder-decoder modules in CS and SD modules. To minimize training time and costs, we utilize BF16 precision and gradient accumulation. Also, we pre-extract content and style features for each speech and audio sample in the training set. Due to the space limitation, we leave more experimental settings in Appendix §??.

6.2 Automatic Evaluation Results

First, we compare the performance of different methods on textual ERG in Table 2, where we find that the Empatheia model performs the best. When we remove the speech and talking-face video information, a decline in performance is observed (though it still outperforms the baseline), indicating that multimodal information aids in better empathetic understanding. Also, removing the CoE strategy has the greatest impact on the response text, reflecting the importance of CoE. Next, we examine the performance of MERG in multimodal content generation, where we present the results of speech generation and avatar generation in Table 3 and Table **??**, respectively. It is evident that our Empatheia model consistently outperforms the pipeline system across all metrics for both speech and avatar video generation. We also analyze the model's ablation results. Firstly, when using different LLMs as backbones, we observe that Vicuna achieves better performance compared to ChatGLM3 and Flan-T5, so our subsequent evaluations are based on Vicuna. Then, when we remove the CS and SD modules individually, we observe a degradation in results, demonstrating the importance of both modules. Finally, we evaluate the impact of different learning strategies, where each causes varying degrees of performance decline, thus validating their effectiveness.

6.3 Human Evaluation Results

Since emotions represent a form of high-level human information, the above automatic evaluation metrics might be insufficient for assessing empathy-related capacities. Thus, we further present the results of human evaluations on textual ERG and MERG in Table 4 and Table 5. It is evident that Empatheia system significantly outperforms the baselines. Also, the model ablation results exhibit trends similar to those observed in the automatic evaluations. As seen, multimodal information contributes to enhanced empathetic understanding and generation. The effectiveness of the CoE mechanism is further confirmed. Moreover, the proposed CS and SD modules, along with various sophisticated training strategies, influence the overall system performance consistently, again revealing their efficacy and importance.

Table 3: Performance of MERG on AvaMERG for speech and talking-head avatar generation.

Model	Speech		Talking-head Avatar		
	MOS ↑	SMOS ↑	CPBD ↑	SSIM ↑	$Sync_{cf}$ ↑
Ground-Truth	4.35	4.81	0.20	1	3.93
Pipeline	3.88	3.97	0.08	0.43	1.95
Empatheia (ChatGLM3)	3.99	4.08	0.14	0.45	2.41
Empatheia (Flan-T5)	4.07	4.09	0.14	0.46	2.26
Empatheia (Vicuna)	**4.16**	**4.33**	**0.15**	**0.49**	**2.76**
w/o CS	3.90	4.07	0.08	0.44	2.21
w/o SD	3.83	4.10	0.11	0.41	2.16
w/o $\mathcal{L}_{emp} + \mathcal{L}_{ccl} + \mathcal{L}_{sac}$	3.90	4.11	0.10	0.33	2.14
w/o $\mathcal{L}_{ccl}$	4.04	4.25	0.13	0.45	2.36
w/o $\mathcal{L}_{sac}$	4.10	4.29	0.11	0.41	2.45

Table 4: Human evaluation on textual ERG.

Model	Emp. ↑	Coh. ↑	Inf. ↑	Flu. ↑
KEMP [21]	2.97	3.11	2.80	4.13
CEM [38]	3.18	3.17	3.15	4.39
CASE [54]	3.03	3.21	3.14	4.31
Empatheia	**4.33**	**4.02**	**3.95**	**4.67**
w/o SPC&VID	4.12	3.98	3.67	4.49
w/o CoE	4.03	3.77	3.49	4.35

Table 5: Human evaluation on MERG.

Model	SCA ↑	VCA ↑	SEA ↑	VEA ↑	MCC ↑	MSC↑
Pipeline	3.23	3.28	3.75	3.62	3.10	3.19
Empatheia	**3.92**	**3.85**	**4.39**	**4.46**	**3.98**	**3.91**
w/o CS	3.46	3.34	3.78	3.63	3.29	3.30
w/o SD	3.55	3.53	3.84	3.77	3.45	3.55
w/o $\mathcal{L}_{emp} + \mathcal{L}_{ccl} + \mathcal{L}_{sac}$	3.33	3.47	3.92	3.78	3.51	3.70
w/o $\mathcal{L}_{ccl}$	3.67	3.50	4.14	4.25	3.74	3.79
w/o $\mathcal{L}_{sac}$	3.88	3.82	3.99	4.04	3.81	3.74

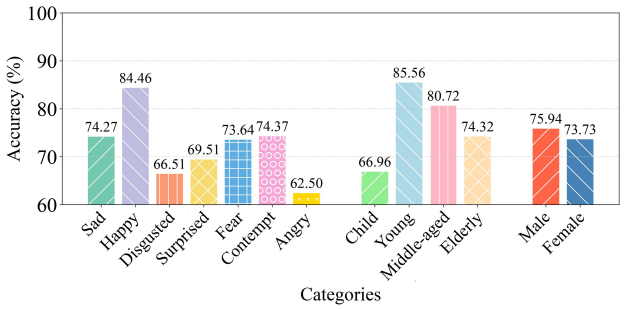


Figure 6: Results on various emotions, ages, and genders.

6.4 Analyses and Discussions

We now conduct more in-depth analyses of several key aspects of Empatheia, offering further insights for better understanding.

Q1. How does Empatheia perform across different emotions, genders, and age groups? Emotion prediction accuracy serves as an indirect measure of the model's capacity for empathetic understanding. We first study the emotion accuracy of Empatheia under

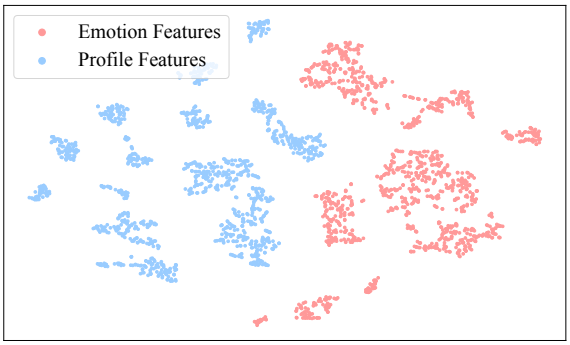


Figure 7: T-SNE visualization of emotion and profile features.

varying emotions, genders, and age groups. As shown in Figure 6, Empatheia is most sensitive to *sad* emotions.

In terms of gender, we observe that the model performs slightly better for males compared to females, which might be attributed to the higher number of male avatars compared to female avatars in the training set. Regarding age groups, Empatheia's accuracy in recognizing children's emotions is relatively low, potentially because children's facial expressions are more dynamic, or their emotional expression patterns differ significantly from adults.

Q2. Has SD module successfully disentangled emotion and profile features? While previous ablation experiments have validated the efficacy of the SD module, it remains uncertain whether it has fully achieved the intended goal of separating emotion and profile features. To explore this, we present the t-SNE [41] visualization on the fused multimodal emotion representations in Figure 7, where we select 500 samples with varying emotions from AvaMERG.

As shown, the results indicate that SD module significantly increases the separation between different emotion categories while clustering the representations of the same emotion. Similarly, the patterns on profile features confirm that SD has successfully disentangled the non-emotion avatar features.

6.5 Qualitative Case Study

Finally, we present two case studies to further demonstrate the specific multimodal empathetic generation capabilities of Empatheia, as illustrated in Figure 8, where we compare the outputs of the Pipeline baseline (without CoE). In the first instance, the user's text does not exhibit an explicit emotional inclination. However, the accompanying sad speech and facial expressions suggest that the user may feel sentimental about "*meeting a friend from middle school*". The Pipeline model, lacking the integration of the CoE strategy, generates an unempathetic response. Also, due to the absence of a style synchronization mechanism, there are inconsistencies in the emotions conveyed between the video and audio components. In contrast, our Empatheia system not only produces high-quality empathetic response content but also ensures that the speech and talking avatar exhibit correct and consistent emotional expressions. Similarly, in the second example, the Pipeline system erroneously interprets the user's emotion, mistakenly assuming that the user is happy about securing second place, whereas Empatheia accurately identifies the user's true emotional state through comprehensive multimodal understanding. Furthermore, the Pipeline incorrectly

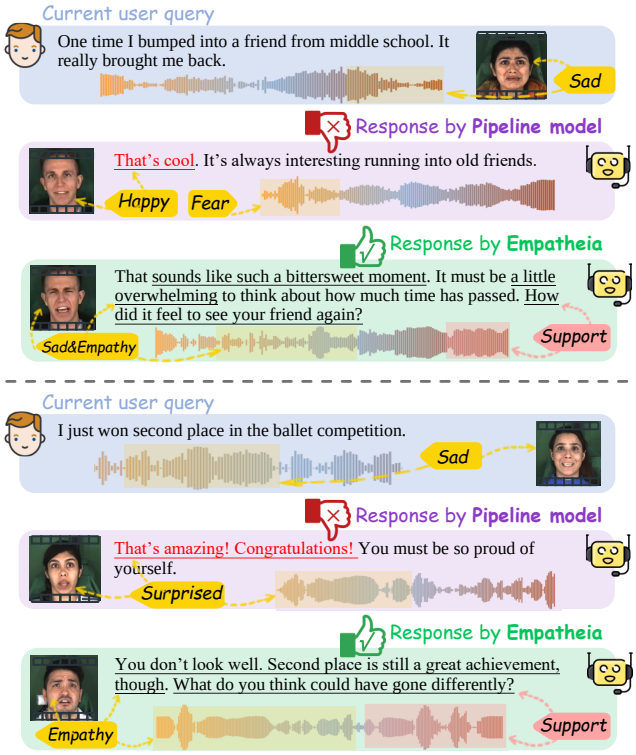


Figure 8: Qualitative results of two testing instances.

assigns the avatar's identity, presenting a male voice paired with a female avatar. On the contrary, our Empatheia shows outstanding capability in correctly handling the avatar profile consistency challenge. In Appendix §?? we showcase more instances for more sufficient case studies.

7 Conclusion

In this paper, we pioneer a novel task of avatar-based MERG. We first introduce AvaMERG, a large-scale high-quality benchmark dataset for MERG, which extends traditional text-based ERG by integrating authentic human speech audio and dynamic talking-face avatar videos. AvaMERG encompasses a diverse range of avatar profiles and covers various real-world scenarios, providing a robust foundation for multimodal empathetic dialogue research. Further, we present Empatheia, a benchmark system tailored for MERG. Based on a backbone LLM as the core reasoner, Empatheia leverages a multimodal encoder, speech generator, and talking-face avatar generator, forming an end-to-end system. We further enhance Empatheia with a Chain-of-Empathetic reasoning mechanism, and implement a series of empathetic-enhanced tuning strategies, including content consistency learning and style-aware alignment and consistency learning, to ensure emotional accuracy and content/profile consistency across modalities. Experimental results demonstrate that Empatheia consistently outperforms baseline methods in both textual ERG and MERG tasks, highlighting the efficacy of our approach.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[2] Changyu Chen, Yanran Li, Chen Wei, Jianwei Cui, Bin Wang, and Rui Yan. 2024. Empathetic Response Generation with Relation-aware Commonsense Knowledge. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 87–95.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[5] Joon Son Chung and Andrew Zisserman. 2017. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. 251–263.

[6] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499* (2023).

[7] Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 1171–1182.

[8] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. *Proceedings of the Advances in neural information processing systems*.

[9] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[10] Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From Multimodal LLM to Human-level AI: Modality, Instruction, Reasoning, Efficiency and Beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*. 1–8.

[11] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot. *arXiv preprint arXiv:2406.15177* (2024).

[12] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 7692–7699.

[13] Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*. 807–819.

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

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[16] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, et al. 2022. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. *arXiv preprint arXiv:2211.05705* (2022).

[18] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* (2015).

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. 19730–19742.

[20] Jian Li and Weiheng Lu. 2024. A Survey on Benchmarks of Multimodal Large Language Models. *arXiv preprint arXiv:2408.08632* (2024).

[21] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*. 10993–11001.

[22] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems* 36 (2024).

[23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[24] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687* (2019).

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[26] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262* (2018).

[27] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26439–26455.

[28] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. 2024. PanoSent: A Panoptic Sextuple Extraction Benchmark for Multimodal Conversational Aspect-based Sentiment Analysis. *arXiv preprint arXiv:2408.09481* (2024).

[29] Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. NUS-Emo at SemEval-2024 Task 3: Instruction-Tuning LLM for Multimodal Emotion-Cause Analysis in Conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. 1599–1606.

[30] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767* (2023).

[31] 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. *arXiv preprint arXiv:2010.01454* (2020).

[32] Niranjan D Narvekar and Lina J Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing* 20, 9 (2011), 2678–2683.

[33] Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140* (2023).

[34] Aravind Sesagiri Raamkumar and Yinping Yang. 2022. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing* (2022), 2722–2739.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. 8748–8763.

[36] Hannah Rashkin. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

[37] 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*. 3505–3506.

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

[39] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).

[40] GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv e-prints* (2024), arXiv–2406.

[41] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[42] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[43] Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer speech & language* 19, 1 (2005), 55–83.

[44] Zhou Wang and Alan C Bovik. 2002. A universal image quality index. *IEEE signal processing letters* 9, 3 (2002), 81–84.

[45] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Towards Semantic Equivalence of Tokenization in Multimodal LLM. *arXiv preprint arXiv:2406.05127* (2024).

[46] 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.

[47] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. *arXiv preprint arXiv:2405.18357* (2024).

[48] 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.

[49] Zhou Yang, Zhaochun Ren, Yufeng Wang, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Yunbing Wu, Yisong Su, Sibo Ju, and Xiangwen Liao. 2024. Exploiting emotion-semantic correlations for empathetic response generation. *arXiv preprint arXiv:2402.17437* (2024).

[50] Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing Empathetic Response Generation by Augmenting LLMs with Small-scale Empathetic Models. *arXiv preprint arXiv:2402.11801* (2024).

[51] Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, 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)*. 7707–7733.

[52] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[53] Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2023. ECQED: emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969* (2023).

[54] Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845* (2022).

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).