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

Anonymous Author(s)*

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. Overall, this work is expected to pioneer the MERG research by introducing a novel benchmark and an end-to-end model, laying a solid foundation for future advancements in multimodal empathetic response generation. All data and code (https://is.gd/iCVhv0) will be open.

1 Introduction

In recent years, the advent of Large Language Models (LLMs) [16, 46] 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

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).

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 [24, 25]. This necessity has driven the development of Empathetic Response Generation (ERG) [32], a task aimed at enabling machines to produce emotionally nuanced and compassionate responses to user queries, thereby facilitating emotion-aware open conversations. Over the past decade, ERG has garnered significant research attention [27, 34, 43]. 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 [32], 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 [18] as the speech generation module and DreamTalk [26] 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 [29, 30] 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 [8, 20, 47]. Existing studies have developed various methods to enhance the performance of ERG systems. For instance, some research focuses on detecting emotional signals from given dialogue contexts to generate emotionally consistent responses [9, 44], while others integrate external commonsense knowledge to deepen the understanding of emotions and improve empathetic generation [2, 34]. 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. For example, Yan et al. (2024) [42] go beyond text-based ERG by proposing an ERG system that incorporates acoustic information to enhance emotional understanding. Similarly, Zhang et al. (2024) [45] focus on integrating visual information, such as emoticons and stickers, into ERG to facilitate better empathetic understanding and generation in dialogues. 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 (i.e., talking-face avatar) to maximize the use of visual information.

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 [21], MiniGPT-4 [48], have been investigated and widely validated for their strong semantic understanding capabilities. However, most MLLMs are limited to
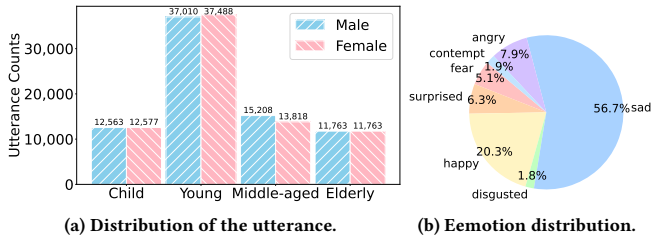
(a) Distribution of the utterance.   (b) Eemotion distribution.

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

multimodal information comprehension yet do not support the flexible generation of diverse modal content beyond text [1, 15, 35], such as audio and visual outputs. Although there are a few MLLMs that support the generation of various modal signals, such as NExT-GPT [41] and Unified-IO 2 [23], 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

This section will details the task definition, and then demonstrate the data construction.

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 $D_i=\{(Q_1, R_1), \dots, (Q_i, R_i)\}$,

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.

3.2 Dataset Construction

We construct our *Ava-MERG* dataset by augmenting the existing pure-text ERG dataset, *Empathetic Dialogue* (ED) [32], 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 Ope-nAI GPT-4[1] has been validated for its remarkable performance in context understanding and thus extensively employed for data generation [24, 41], 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 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 §C.

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

---

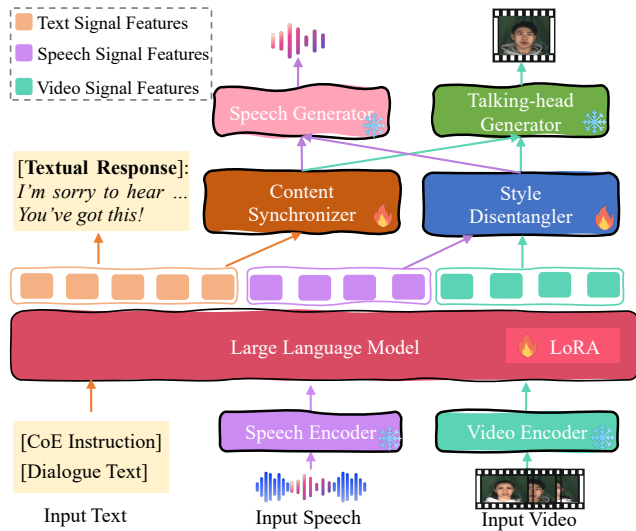[1]https://openai.com/index/gpt-4/, June, 2024

Figure 3: Architecture of our Empatheia MLLM for MERG.

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 [10] and CLIP ViT-L/14@336px [31] as the speech encoder and avatar video encoder. Essentially, the latent representations 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 [4] is widely adopted as a

baseline for MLLMs [7, 19] 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 this, 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.

▶ **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 [18] and DreamTalk [26], 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 example, if the text expresses anger and the speech is delivered in the voice of an angry middle-aged man, the face of the avatar should also reflect the expression and appearance of the man. For these purposes, we further design two modules before the two generators: content synchronizer and style disentangler.
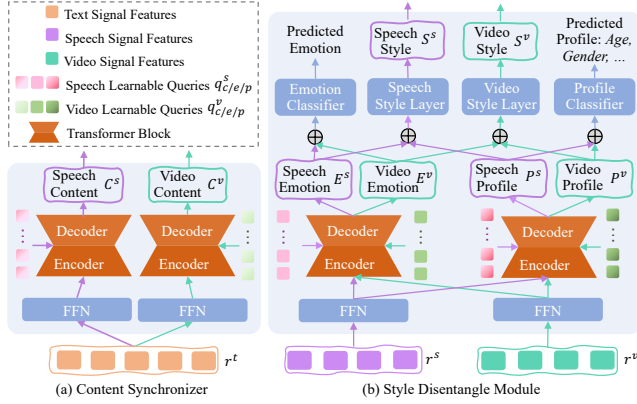
Figure 4: Illustration of the Content Synchronizer and Style Disentangle modules.

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 [38] variational auto-encoder (VAE) [12]. 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.

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

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

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:

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

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

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

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

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:

$$S^{s/v} = E^{s/v} \oplus P^{s/v}, \qquad (7)$$

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 [11] technique.

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

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 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:

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

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 $\hat{S^v}$:

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

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
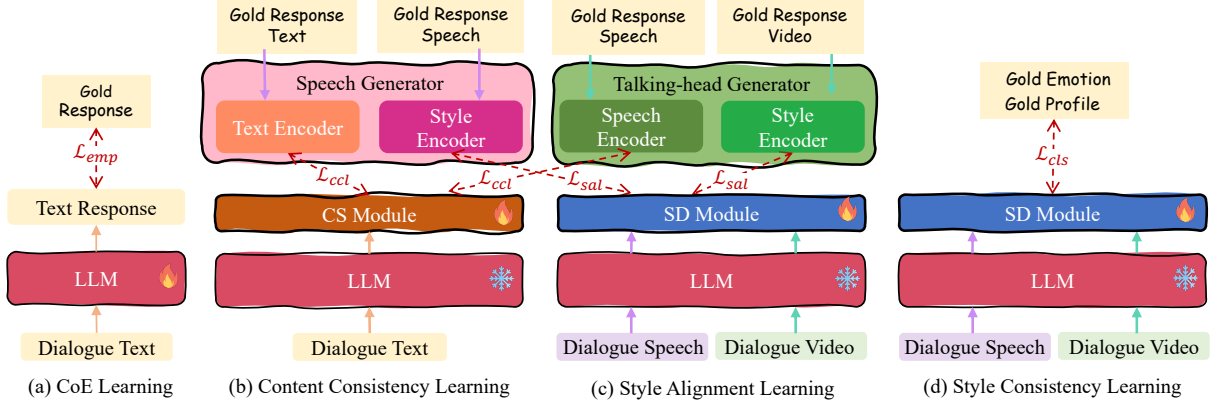
Figure 5: Illustrations of the proposed training strategies.

disentangle pure emotion and profile representations. We here introduce two classification losses for emotion and profile prediction:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_i^N \left(\sum_c^{M_e} y_{i,c} \log(p_{i,c}) + \sum_p^{M_p} \sum_c^p y_{i,c} \log(p_{i,c}) \right), \quad (11)$$

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:

$$\mathcal{L}_{sac} = \mathcal{L}_{sal} + \mathcal{L}_{cls} . \quad (12)$$

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:

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

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 [5], ChatGLM3-6B [36], and Vicuna-7B [4]. Besides MERG, we also compare the text ERG performance with existing models, including KEMP [17], CEM [34] and CASE [47], 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 [18] to synthesize speech, and then processes the speech using DreamTalk [26] 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) [14]. For speech generation, we use the 5-scale Mean Opinion Score (MOS) [39] and Similarity MOS (SMOS) [22], along with the objective Mel Cepstral Distortion (MCD) score [13]. For talking head generation, we adopt the Structural Similarity Index Measure (SSIM) [40], Cumulative Probability of Blur Detection (CPBD) [28], SyncNet confidence score ($Sync_{cf}$) [6], Landmark Distance around the mouth area (M-LMD) [3], and Landmark Distance on the full face (F-LMD).

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 [11] and DeepSpeed [33] 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 §E.

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 4,

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

Model	Acc ↑	Dis-1 ↑	Dis-2 ↑
KEMP [17]	35.87	0.41	1.78
CEM [34]	37.32	0.50	2.07
CASE [47]	40.96	0.54	2.14
Empatheia	**45.51**	**2.73**	**11.75**
w/o SPC&VID	44.19	2.33	10.56
w/o CoE	42.42	2.19	9.37

Table 3: Speech generation results of MERG on AvaMERG.

Model	MOS ↑	SMOS ↑	MCD ↓
Ground-Truth	4.35	-	-
Pipeline	3.88	3.97	3.89
Empatheia (ChatGLM3)	3.99	4.38	3.66
Empatheia (Flan-T5)	4.07	4.49	3.52
Empatheia (Vicuna)	**4.16**	**4.53**	**3.27**
w/o CS	3.90	4.07	3.78
w/o SD	3.83	4.10	3.69
w/o $\mathcal{L}_{emp} + \mathcal{L}_{ccl} + \mathcal{L}_{sac}$	3.90	4.21	3.79
w/o $\mathcal{L}_{ccl}$	4.04	4.45	3.47
w/o $\mathcal{L}_{sac}$	4.10	4.49	3.42

Table 4: Talking-head avatar generation performance of MERG on AvaMERG data.

Model	SSIM ↑	CPBD ↑	F-LMD ↓	M-LMD ↓	$Sync_{cf}$ ↑
Ground-Truth	1	0.20	-	-	-
Pipeline	0.50	0.08	5.98	5.53	1.95
Empatheia (ChatGLM3)	0.72	0.14	3.54	3.83	2.41
Empatheia (Flan-T5)	0.73	0.14	3.04	3.71	2.26
Empatheia (Vicuna)	**0.77**	**0.15**	**2.72**	**3.52**	**2.76**
w/o CS	0.56	0.08	4.88	4.88	2.21
w/o SD	0.63	0.11	4.78	4.96	2.16
w/o $\mathcal{L}_{emp} + \mathcal{L}_{ccl} + \mathcal{L}_{sac}$	0.63	0.10	4.79	5.21	2.14
w/o $\mathcal{L}_{ccl}$	0.72	0.13	4.33	5.79	2.36
w/o $\mathcal{L}_{sac}$	0.69	0.11	3.92	4.45	2.45

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 5 and Table 6. 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 5: Human evaluation on textual ERG.

Model	Emp. ↑	Coh. ↑	Inf. ↑	Flu. ↑
KEMP [17]	2.97	3.11	2.80	4.13
CEM [34]	3.18	3.17	3.15	4.39
CASE [47]	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 6: 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
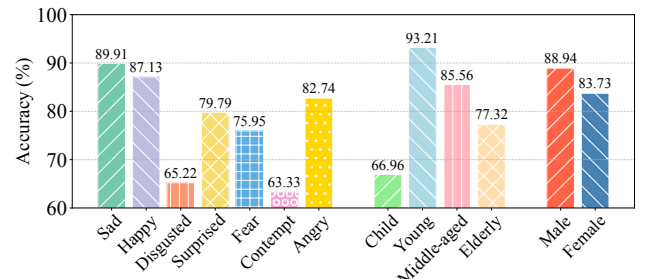
Figure 6: Results on various emotions, ages, and genders.

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 varying emotions, genders, and age groups. As shown in Figure 6, Empatheia is most sensitive to *sad* and *happy* emotions. In terms of gender, we observe that the model performs slightly better for females compared to males, which might be attributed to the tendency of males to suppress their emotional expressions. 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. How is CoE beneficial for empathetic response generation? CoE works by breaking down the overall hard problem into semantically coherent and easier sub-problems. To gain a deeper understanding of how each step of the CoE reasoning process contributes to the ERG, we conduct ablation experiments on CoE, by progressively using more leveled steps of CoE prompts. The results, as illustrated in Figure 7, show that as Empatheia advances through the steps of CoE, the model's empathetic performance improves significantly, where the capacity becomes stronger with each level of complexity, leading to more accurate and contextually appropriate responses. This indicates that each individual step in CoE
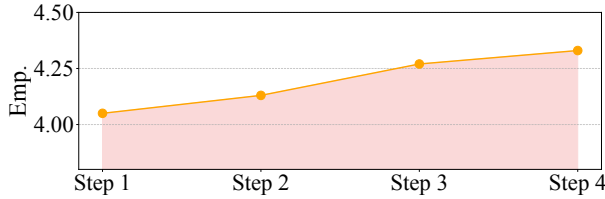
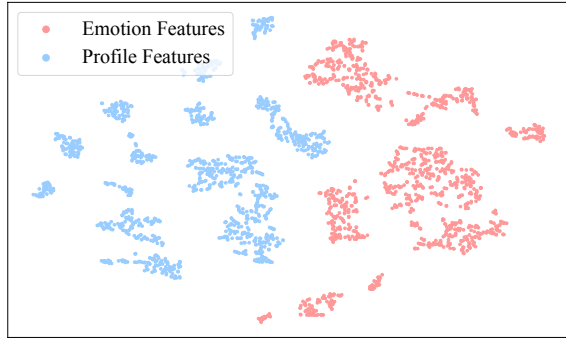Figure 7: In-Depth Analysis of the Effectiveness of CoE.

Figure 8: T-SNE visualization of emotion and profile features.

plays a vital role, enhancing the model's ability to comprehensively analyze and understand user's emotion, intention and empathy. By systematically breaking down the CoE into distinct phases, we highlight that the model's

Q3. 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 [37] visualization on the fused multimodal emotion representations in Figure 8, 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 9, 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
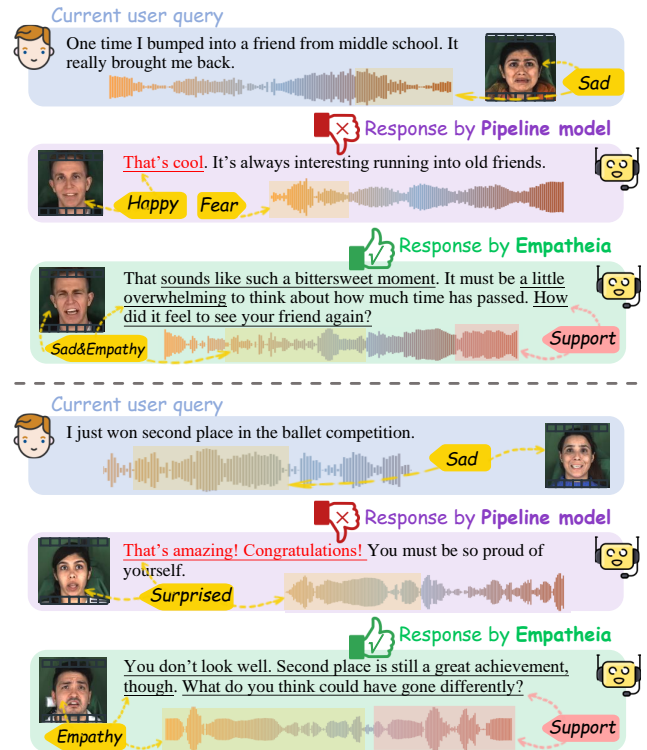
Figure 9: Qualitative results of two testing instances.

happy about securing second place, whereas Empatheia accurately identifies the user's true emotional state through comprehensive multimodal understanding. Furthermore, the Pipeline incorrectly 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 §F.3 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] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7832–7841.

[4] 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/

[5] 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.

[6] 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.

[7] 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).

[8] 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).

[9] 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.

[10] 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.

[11] 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).

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

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

[14] 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).

[15] 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.

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

[17] 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.

[18] 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).

[19] 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).

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

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

[22] 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).

[23] 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.

[24] 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).

[25] 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.

[26] 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).

[27] 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).

[28] 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.

[29] 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).

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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).

[36] 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.

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

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

[39] 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.

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

[41] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* (2023).

[42] 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.

[43] 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).

[44] 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).

[45] Yiqun Yang, 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.

[46] 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).

[47] 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).

[48] 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).

A   Ethic Considerations

We can identify the following potential ethical considerations for our work:

Privacy and Data Protection. Empatheia relies on multimodal inputs, including text, voice, and video, which contain highly sensitive personal data. It is essential to ensure that all data collected and processed by the system adheres to strict privacy and data protection regulations, such as the GDPR or CCPA. The system must implement strong encryption techniques for storage and transmission, while also ensuring user data is anonymized where possible. Users should have full control over their data, including the ability to delete their inputs and outputs from the system. Regular audits of data handling and retention practices should be conducted to maintain compliance with privacy standards.

Bias and Fairness. Empatheia's ability to generate empathetic responses relies on large language models and multimodal data, which can inherit biases from the training datasets. It is crucial to ensure that the system does not perpetuate harmful stereotypes or exhibit bias based on gender, race, age, or any other demographic characteristics. Diverse and inclusive training data, as well as regular testing for fairness, should be a priority. Additionally, measures should be taken to mitigate biases, such as using techniques like debiasing algorithms, and continuously refining the dataset to minimize any unintentional discrimination.

Emotional Manipulation and User Vulnerability. Since Empatheia is designed to interact empathetically with users, it may encounter individuals in emotionally vulnerable states. The system must avoid exploiting this vulnerability or manipulating emotions in harmful ways. Safeguards should be in place to ensure that the chatbot's responses are supportive but do not give inappropriate advice or encourage dependency. Ethical guidelines should be established to prevent the misuse of the chatbot, and users should be made aware that it is a machine-generated response system and not a substitute for professional psychological help. Where appropriate, the system could be designed to refer users to human professionals in cases of serious emotional distress.

Autonomy and Transparency. The nature of Empatheia's multimodal empathetic responses might blur the lines between human and machine interaction. It is essential to maintain transparency about the system's limitations and make users fully aware that they are interacting with an AI. Users should also have the autonomy to make informed decisions about using the system and be provided with clear options to opt-out or disengage at any time. Regular disclosures about the system's AI-driven nature, its data collection practices, and its purpose should be communicated transparently.

Potential for Misuse. As with any open-source system, there is a risk of Empatheia being misused in ways that could harm individuals or communities. Bad actors might leverage the system's empathetic capabilities for malicious purposes, such as manipulating others through emotion-driven content or creating deepfakes for deceptive purposes. To mitigate this, the development of Empatheia should include security measures to prevent exploitation, such as limiting the use of avatars and ensuring that any generated content is watermarked or traceable. The open-source release

should come with strict usage guidelines and community oversight to ensure responsible use of the system.

Long-Term Psychological Effects. The long-term effects of interacting with an empathetic AI system like Empatheia on human users should be carefully considered. While the system aims to foster deeper emotional connections, there is a risk that users may become overly reliant on AI for emotional support, potentially leading to social isolation or reduced human empathy. Further research should be conducted to assess the psychological impact of prolonged use of such systems, and regular evaluations should be made to ensure that the system enhances human emotional well-being rather than detracting from it.

B Future Work with AvaMERG

In this paper, we present a comprehensive exploration of multimodal empathetic response generation. We believe this work lays the foundation for future advancements in the field of multimodal sentiment analysis and empathetic interaction. From our practice, several promising directions for future research can be identified.

Exploring Higher Performance of MLLMs and Efficient Training Methods. Future work can investigate the performance of various MLLMs in the generation of empathetic responses, particularly their advantages and limitations when processing multimodal inputs. Currently, we utilize state-of-the-art speech and avatar generators; however, their performance remains limited. Therefore, it is essential to enhance the quality of multimodal generation. Also, more efficient training methods, such as transfer learning, few-shot learning, or self-supervised learning, can be explored to improve the training efficiency and performance of these models. Through systematic experimental comparisons, the aim is to identify best practices that enhance the quality and responsiveness of empathetic response generation.

Developing Multidimensional Evaluation Methods. Currently, the evaluation of the multimodal generation component of MERG relies solely on human evaluations, which introduces significant uncertainty. Future research should aim to establish multidimensional evaluation methods to comprehensively and automatically assess the effectiveness and quality of multimodal empathetic responses. This can be achieved by combining automated evaluations with human assessments. Specifically, deep learning-based evaluation models can be developed to automatically analyze the semantic consistency of generated responses, the accuracy of emotional conveyance, and the synergistic effects of multimodal inputs. Additionally, emotional analysis tools and semantic understanding techniques should be utilized to conduct detailed emotional depth analyses of the generated responses. Furthermore, the research should explore how to assess cross-modal correlations, such as evaluating the consistency between text, audio, and video, to further enhance the comprehensiveness and accuracy of the evaluation.

Enhancing the Model's Contextual Understanding. Future research can focus on improving the model's understanding of conversational context, particularly in retaining and utilizing historical information during long dialogues. Consideration could be given to incorporating more complex memory mechanisms or contextual attention mechanisms to enhance the model's contextual awareness.

Exploring Cross-cultural Expressions of Empathy. Future work can investigate how to effectively generate empathetic responses across different cultural contexts. The research could focus on analyzing the impact of cultural differences on emotional expression and communication styles, adjusting the model based on these findings to better accommodate users from diverse cultural backgrounds.

Improving Dataset Diversity and Quality. Future work can focus on collecting and constructing larger-scale, more diverse multimodal datasets to encompass a wider range of emotional expressions and conversational scenarios. By enhancing the representativeness of the dataset, the model's generalization ability and robustness in diverse emotional interaction contexts can be further improved.

C More Details of Datasets

C.1 Extended Details of Data Constructions

C.1.1 Dialogue Enriching.

Augmenting the Empathetic Dialogue (ED) Dataset. We begin by augmenting the existing pure-text Empathetic Dialogue (ED) dataset to construct our AvaMERG dataset. The ED dataset consists of dialogues aimed at empathetic response generation (ERG) but lacks multimodal and identity-specific information essential for Multimodal Empathetic Response Generation (MERG). To address this, we first enrich each textual empathetic response $t_i \in R_i$ with corresponding emotion chain, thereby constructing an emotional chain of thought (CoT).

Leveraging the advanced contextual understanding capabilities of OpenAI's GPT-4, we annotate each utterance in the ED dataset with emotion chain. We define an emotion CoT: *emotion* → *emotion_cause* → *goal_to_response*. GPT-4 assigns an appropriate emotion chain to each utterance based on the dialogue context.

An example of our prompt template:

As an expert in empathetic dialogue analysis, your task is to analyze the emotional dynamics and intentions behind a conversation between two participants: a 'speaker' and a 'listener.' The goal is to identify the 'Emotion Cause' for the speaker and the 'Goal to Response' for the listener's final reply.

Task Overview: 1. **Emotion Cause**: Based on the conversation context, sentiment, and dialogue history, analyze and identify the underlying emotional cause or trigger for the speaker. 2. **Goal to Response**: Analyze the last response from the listener and identify the intended goal behind that response. The goal should relate to how the listener is attempting to address the speaker's emotional state.

Input JSON Field Descriptions:
- **doc_id**: dialogue['doc_id']
- **sentiment**: dialogue['sentiment']

```
    - context: dialogue['context']
    - dialogue:
    for utterance in dialogue['dialogue']:
    user_content += f" - utterance['role']:
    utterance['utterance']"
```

Expected Output: Please provide the following in **concise** JSON format: - **Event Scenario**: A **short description** summarizing the main context or situation of the dialogue (e.g., 'The speaker is expressing fear after experiencing something unsettling'). - **Emotion Cause**: A **brief** explanation of the **specific event** or experience that triggers the speaker's emotion (e.g., 'Elevator game brings horror experience'). - **Goal to Response**: A **concise** goal that describes the **specific emotional state** the listener is attempting to address (e.g., 'Alleviating fear').

Example Output:

```json
{
  "doc_id": "<doc_id>",
  "event_scenario": "The speaker is expressing
fear after experiencing something unsettling",
  "emotion_cause": "Elevator game brings horror
        experience",
  "goal_to_response": "Alleviating fear"
}
```

To ensure the accuracy of the emotion annotations, we implement a validation step where human annotators review the GPT-4 assigned emotions. This process involves cross-referencing the emotion labels with the dialogue content to verify consistency and appropriateness. Any discrepancies are resolved through discussion among annotators, ensuring high-quality emotion annotations.

Enriching Identity Information. To enable MERG models to generate appropriate avatar profiles for both audio and video modalities, we further annotate each utterance with identity information for both participants in the dialogue. This includes:

- **Age**: We define four age periods—*child* (0-15 years), *young* (16-34 years), *middle-aged* (35-59 years), and *elderly* (60+ years).
- **Gender**: Binary genders—*male* and *female*.
- **Timbre**: Three vocal timbres—*low*, *mid*, and *high*.

GPT-4 is utilized to determine the above labels for each utterance, ensuring that the dialogue reflects realistic interactions between participants with diverse profiles. The identity annotations are critical for training models to generate contextually appropriate and personalized empathetic responses in multiple modalities.

Data Balancing and Expansion. Observing that the raw ED dataset is imbalanced (e.g., most dialogues involve young or middle-aged participants), we employ GPT-4 to generate additional dialogues that include underrepresented age groups and genders, as well as a balanced distribution of timbres. GPT-4 also detects and labels the dialogue topics, covering 10 primary common topics: [Social Issues and Moral Dilemmas, Achievements and Self-Realization, Support and Comfort, Emotions and Feelings, Disappointments and Expectations, Life Events, Interpersonal Relationships, Health and

Well-being, Uncertainty About the Future, Personal Struggles and Challenges].

To enhance diversity, GPT-4 is instructed to generate dialogues that based on various races, cultural backgrounds, and socio-economic statuses, reflecting a realistic and inclusive range of human experiences. This process results in a more balanced and representative dataset.

Enriching Identity for Dialogue Generation

You are an AI language model tasked with generating a dialogue between two participants, incorporating detailed identity information and topic annotations.

Requirements:

- **Dialogue Structure:** The dialogue contains 3 turns, with alternating participants (max 6 utterances). Each turn includes the full dialogue history and the listener's empathetic response.
- **Identity Information:** Profiles for both participants should include: **Age** (child: 0-15, young: 16-34, middle-aged: 35-59, elderly: 60+); **Gender** (male, female); **Timbre** (low, mid, high). Dialogue content should reflect these identities.
- **Emotion and Empathy Chain:** Each turn should include the speaker's emotion (constant across turns) and evolving fields for **event scenario**, **emotion cause**, and **goal to response**.
- **Data Balancing:** Ensure representation of underrepresented age groups, genders, and timbres.
- **Topics:** The conversation should revolve around one of these topics: Social Issues and Moral Dilemmas; Achievements and Self-Realization; Support and Comfort; Emotions and Feelings; Disappointments and Expectations; Life Events; Interpersonal Relationships; Health and Well-being; Uncertainty About the Future; Personal Struggles and Challenges.

JSON Format Example:

```json
{"conversation_id": "string",
 "speaker_profile": { "age": "string", "gender":
 "string", "timbre": "string" },
 "listener_profile": { "age": "string", "gender"
 : "string", "timbre": "string" },
 "topic": "string",
 "turns": [{"turn_id": "string", "context":
 "string",  "dialogue_history": [
  { "index": int, "role": "string", "utterance"
   : "string" }],
 "response": "string",
 "chain_of_empathy": {
 "speaker_emotion": "string",
 "event_scenario": "string",
 "emotion_cause": "string",
 "goal_to_response": "string"}
 ]
}
```

Human Annotation and Cross-Checking. To ensure the quality

and accuracy of the augmented dataset, we recruit human annotators for rigorous manual checking. Each dialogue undergoes a 3-person cross-checking process where annotators verify:

- **Content Accuracy**: Whether the dialogue content is coherent, contextually appropriate, and free from biases or offensive language.
- **Meta-Profile Consistency**: Whether the assigned identity information and emotion labels are accurate and consistent with the dialogue content.
- **Topic Relevance**: Whether the dialogue topics are correctly identified and relevant.

Annotators receive detailed guidelines and training to ensure consistency in their evaluations. Discrepancies among annotators are discussed and resolved collectively. Only dialogues that receive unanimous approval from all three annotators are included in the AvaMERG dataset.

C.1.2 Audio & Video Recording.

Volunteer Recruitment. We recruit a large and diverse group of English-speaking volunteers representing different ages, genders, vocal characteristics, and races (Asian, Caucasian, African, Latino, Indian). Recruitment is conducted through community outreach, social media, and collaboration with institutions to ensure diversity. All volunteers provide informed consent and are compensated for their participation. We standardized the filming environment requirements. Due to the need for diversity, 30% of volunteers with certain characteristics could not be found offline, so they communicated with us online and submitted their recorded results. Similarly, they adhered to consistent environment requirements during their recording.

Participant Pairing and Assignment. Volunteers are paired and grouped according to the profiles determined in the annotated AvaMERG dialogues. Each pair corresponds to the identity profiles of the dialogue participants, ensuring that the multimodal data accurately reflects the textual annotations. Care is taken to match volunteers to profiles that they can authentically portray, enhancing the realism of the dataset.

Recording Sessions. During the recording sessions, volunteers perform the dialogues by carefully reading the utterance text. They are provided with context about the dialogue, including the emotional state and background of the characters. Instructions are given to exhibit the correct emotional performance, paying close attention to:

- **Vocal Attributes:** Tone, pitch, and timbre corresponding to the annotated vocal timbre and emotion.
- **Facial Expressions:** Micro-facial expressions that align with the emotional content, captured using high-resolution cameras.

Professional recording equipment is used to ensure high-quality audio and video data. Sessions are supervised by directors who provide guidance to volunteers to achieve the desired performances.

Post-processing. Recorded data undergoes post-processing to enhance quality. This includes noise reduction in audio files, color correction in videos, and synchronization of audio and video streams.

Metadata is added to files to link them with the corresponding textual annotations and identity profiles. After processing, considering storage space and the input requirements of our model, we standardized the format to meet both our storage capacity and the model's needs, while ensuring the audio and audiovisual content remained unchanged.

C.1.3 Manual Annotation Verification.

Verification Process. To ensure the highest quality of our dataset, we implement a thorough manual annotation verification process following the post-processing stage.

Each recorded dialogue, along with its corresponding annotations and multimodal content, undergoes a comprehensive review by at least two independent annotators. The verification process focuses on:

- **Content Alignment:** Checking that the spoken words and visual expressions in the recordings accurately match the textual utterances and annotated emotion chains.
- **Profile Consistency:** Ensuring that the age, gender, timbre, and emotional expressions portrayed by the volunteers align with the assigned profiles.
- **Technical Quality**: Verifying that the audio and video recordings meet the required technical standards, including clarity, resolution, and synchronization.

Quality Metrics Calculation. We calculate the Cohen's Kappa Score to measure the agreement between annotators. Achieving a score of 0.78 indicates a high level of consistency and reliability in the annotation process. Any instances where annotators disagree or identify potential issues are reviewed collectively, and problematic data is either corrected or discarded.

Finalization. Only dialogues that pass the manual verification process with unanimous approval are included in the final AvaMERG dataset. This rigorous quality control ensures that the dataset is both reliable and suitable for training and evaluating MERG models.

C.2 Detailed Data Highlights

Here, we extend the content of Dataset Construction from the main article to provide a more comprehensive introduction to all the highlights of our AvaMERG dataset. AvaMERG boasts a diverse array of avatar profiles. This diversity ensures that models trained on AvaMERG can generalize across various demographic profiles, promoting inclusivity and reducing biases. AvaMERG contains a total of **33,048** dialogue samples and **152,021** dialogue utterances, providing ample usable data for the field. AvaMERG contains the following notable strengths:

A reasonable and well-balanced tone distribution. In terms of Tone shown in Figure 10, Gentle tones were the most common (**42.43%**, **64,504**), followed by Emphatic (**31.28%**, **47,553**) and Mild (**26.29%**, **39,964**) tones. Regarding Tone, listeners were mostly Gentle (**67.73%**, **51,481**), followed by Mild (**26.05%**, **19,801**) and Emphatic (**6.22%**, **4,728**), tones. Speakers were mostly Emphatic (**56.34%**, **42,825**), followed by Mild (**26.53%**, **20,163**) and Gentle

Table 7: A snippet of an annotated data instance based on provided dialogue data.

KEY	VALUE
Conversation-ID	01797
Speaker Profile	Age: young, Gender: male, Tone: mild, ID: 35
Listener Profile	Age: young, Gender: female, Tone: emphatic, ID: 20
Topic	Personal Struggles and Challenges
Dialogue	1. Speaker: When I left the bathroom in high school once I had toilet paper stuck to my shoe. 2. Listener: I bet it was very embarrassing? 3. Speaker: Yeah it sure was, you know how mean teenagers can be. It's like they have no empathy or think about what if it was them. 4. Listener: It's alright, we've all been there many times.
Speaker Emotion	Embarrassed
Event Scenario	The speaker experienced embarrassment after accidentally leaving the bathroom with toilet paper stuck to their shoe.
Emotion Cause	Embarrassing public incident involving toilet paper.
Goal of Response	Validate the speaker's embarrassment, providing reassurance and comfort.

Table 8: Detailed information of dialogue topic.

Topic	Keywords
Social Issues and Moral Dilemmas	income inequality, climate change activism, ethical consumerism, human rights violations, systemic racism, gender discrimination, animal welfare concerns, mental health stigma, refugee crisis, digital privacy issues
Achievements and Self-Realization	bar exam, college graduation ceremony, soccer match, kitchen cooking test, financial success, project completion, professional certification, art exhibition, personal fitness milestone, community service award
Support and Comfort	dealing with burnout, feeling overwhelmed at work, losing a pet, struggling with loneliness, facing a tough decision, breakup with partner, family argument, lossing a job
Emotions and Feelings	obtaining driver's license, reckless driving behavior, forgotten debt, unexpected compliment, cherished memory, disappointing news, long-awaited vacation, stressful deadline
Disappointments and Expectations	missed opportunity, unfulfilled promise, low performance review, delayed project completion, unmet goals, failed partnership, incomplete application, rejected proposal, unsatisfactory results, broken commitment
Life Events	flight costs, unexpected surprise, enjoyable day trip, receiving unexpected flowers, mother's death, job interview, significant life transition, family gathering, grocery shopping
Interpersonal Relationships	marriage relationship, emotional distress, providing support, misbehaving with friends, overwhelming academic pressure, unexpected friendship formation, trust building, mutual Respect, forgiveness and reconciliation, social connection, shared experiences
Health and Well-being	mental resilience, physical fitness routine, healthy eating habits, stress management techniques, mindfulness practice, regular health check-ups, emotional balance, quality sleep, work-life harmony, social connection
Uncertainty About the Future	career instability, financial insecurity, unpredictable life events, fear of change, unfulfilled dreams
Personal Struggles and Challenges	financial stress, unexpected romantic encounter, feeling lonely, career uncertainty, self-doubt , mental health struggles, parental pressure
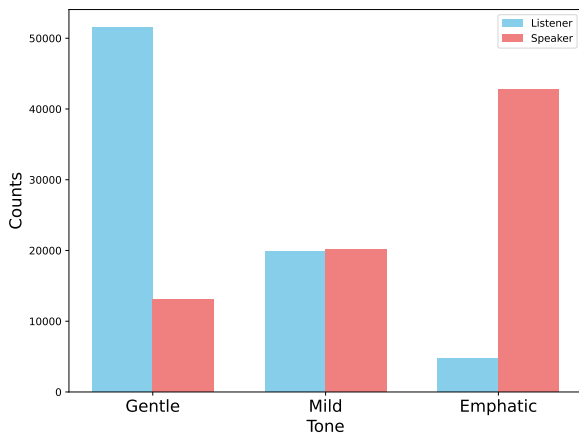
Figure 10: The distribution of tone.

(**17.13%**, **13,023**) tones. AvaMERG provides a comprehensive distribution of tones between speakers and listeners, promoting balanced

dialogue interactions, and its tone distribution is designed to align with the tones that different speaker identities would adopt in real-world scenarios.

A diverse and balanced age-gender distribution for real-world applicability. AvaMERG sets itself apart with a well-balanced distribution across both age groups and gender, as illustrated in the figure 2. The dataset includes avatars representing different life stages: Child, Young, Middle-aged, and Elderly, with a nearly equal gender representation within each category. This ensures that the models trained on AvaMERG can generalize effectively to a wide variety of demographic profiles, promoting inclusivity and reducing bias in real-world dialogue systems. For children, there is an almost equal split between male (**12,563**) and female (**12,577**) utterances. Among the young adult group, the dataset includes **37,010** male utterances and **37,488** female utterances. In the middle-aged category, there are **15,208** male utterances and **13,818** female

utterances. Lastly, the elderly group shows an equal number of utterances for both genders, with **11,763** male and **11,763** female utterances. This diverse age and gender representation ensures that dialogue systems trained on AvaMERG can perform robustly across various demographic profiles, enhancing fairness and inclusivity in real-world applications.

Detailed and rich emotional design for real-world applicability. We adopt 32 fine-grained textual emotions and 7 coarse-grained multimodal emotions, and map them appropriately, as shown in Figure 11. As shown in the emotional distribution in Figure 2, AvaMERG covers a wide range of emotional expressions. Among the 33,048 dialogues, the most prevalent emotion is sadness, accounting for **56.7%** of the samples, which reflects the empathetic nature of many conversations, especially in scenarios requiring emotional support or assistance. This is followed by happiness (**20.3%**), which captures the dialogues involving positive reinforcement or joyful interactions. Other emotions such as anger (**7.9%**), contempt (**7.9%**), and surprise (**6.3%**) are also represented in reasonable proportions, ensuring that the dataset includes not only empathetic responses but also situations where the user expresses negative or unexpected emotions. Fear and disgust appear less frequently, with **5.1%** and **1.8%** respectively, but still provide valuable instances for training models that can handle a full spectrum of emotional states. This rich emotional diversity ensures that models trained on AvaMERG are capable of understanding and generating appropriate responses to a broad range of emotional expressions, improving the model's ability to handle real-world interactions where emotions play a critical role.

A comprehensive topic design. The AvaMERG dataset showcases a meticulously designed mapping between various real-world topics and the emotional responses they elicit, providing an invaluable resource for training empathetic response models. As shown in the heatmap Figure 12, topics such as "Achievements and Self-Realization", "Disappointments and Expectations", "Health and Well-being", and "Personal Struggles and Challenges" elicit a wide range of emotional responses, including anger, happiness, and sadness. For instance, "Health and Well-being" is associated with a significant number of utterances reflecting fear (**1,158**), sadness (**2,214**), and disgust (**810**), which mirrors real-world conversations where health concerns often evoke complex emotions. Similarly, "Achievements and Self-Realization" is more frequently linked to positive emotions, with **3,214** utterances expressing happiness, demonstrating the natural alignment between positive life events and joyful emotions. In contrast, "Social Issues and Moral Dilemmas" and "Uncertainty About the Future" are characterized by more negative emotions, such as fear and contempt, reflecting the inherent challenges and anxieties that arise when discussing social complexities or future uncertainties. By covering a wide spectrum of topics and emotions, the dataset enables the development of empathetic models capable of understanding and responding to the nuanced emotional underpinnings of different conversations.
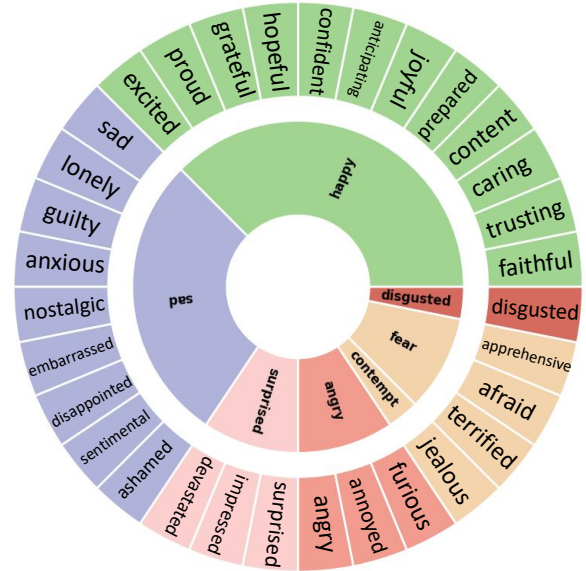
Figure 11: The mapping of fine-grained textual emotions to coarse-grained multimodal emotions.
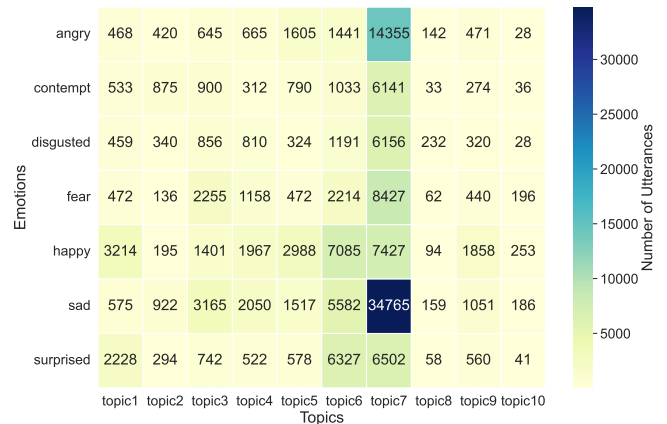
Figure 12: Emotion-topic heatmap for dialogue utterances. Topics 1 to 10 represent achievements and self-realization, disappointments and expectations, emotions and feelings, health and well-being, interpersonal relationships, life events, personal struggles and challenges, social issues and moral dilemmas, support and comfort, uncertainty about the future, respectively.

D More Details of Methods

In this part, we provide an extension to our Empatheia system, including the input construction format, the specifics of the generator, and the training details.

D.1 Empatheia Text Input

Empatheia is designed with four complementary training stages, each of which employs a distinct input structure. For the CoE learning and Content Consistency Learning (CCL) stages, the input format is as follows:

An Input Example for the CoE and CCL

Input:
Provide an empathetic response based on the given dialogue context below. Don't rush to give the response, thinking step by step.
Dialogue Context:
{
Speaker:I paid all my bills today, I feel great!
Assistant:Every little accomplishment counts! What bills did you have to pay?
Listener:Rent and electricity. We've been struggling financially so it's such a relief to pay bills.
}

Target:
Firstly, the event scenario of this conversation is: Paying off overdue rent and electricity bills amid financial struggles.
Secondly, the emotion of the speaker is: content
Thirdly, the emotion cause is: Relief from the burden of financial stress after successfully paying bills.
Fourthly, the goal to response is:Providing support and validation for the speaker's sense of accomplishment.
Finally, the response is: Sorry to hear that. Well, at least this will be a weight off your shoulders.

For different modalities of dialogue input, we standardize them into feature vectors, which can then be fed into the LLM for multimodal information comprehension. For the stages of style consistency learning, our input consists solely of multimodal speech and video. This setup encourages the model to learn how to generate multimodal responses with consistent emotion and style. Specifically, the input format to the LLM is structured as follows:

An Input Example for Style Learning

Input:
Provide an multimodal empathetic response based on the given dialogue context below.
Dialogue Context:
{
Speaker:<Aud> <Vid>.
Listener:<Aud> <Vid>.
Speaker:<Aud> <Vid>.
}

Target:
<AUD1><AUD2><AUD3><AUD4><AUD5><AUD6>
<AUD7><AUD8><AUD9><AUD10><AUD11><AUD12>
<AUD13><AUD14><AUD15><AUD16>
<VID1><VID2><VID3><VID4><VID5><VID6><VID7>
<VID8><VID9><VID10><VID11><VID12><VID13>
<VID13><VID14><VID15><VID16>

Here, <Aud> and <Vid> are special placeholders, and before being input into the LLM, the token embeddings at these positions will be replaced with the corresponding audio and video features. <AUDi> and <VIDi> are used as multimodal generation signals.

Combining the first two phases, the final overall training input consists of three modalities. We define the input format for the LLM in this phase as follows:

An Input Example for Overall Training

Input:
Provide a multimodal empathetic response based on the given dialogue context below. Don't rush to give the response, thinking step by step.
Dialogue Context:
{
Speaker:I paid all my bills today, I feel great!<Aud> <Vid>.
Listener:Every little accomplishment counts! What bills did you have to pay?<Aud><Vid>.
Speaker:Rent and electricity. We've been struggling financially so it's such a relief to pay bills.<Aud> <Vid>.
}

Target:
Firstly, the event scenario of this conversation is: Paying off overdue rent and electricity bills amid financial struggles.
Secondly, the emotion of the speaker is: content
Thirdly, the emotion cause is: Relief from the burden of financial stress after successfully paying bills.
Fourthly, the goal to response is:Providing support and validation for the speaker's sense of accomplishment.
Finally, the response is: Sorry to hear that. Well, at least this will be a weight off your shoulders.
<AUD1>...AUD16> <VID1>...<VID16>

D.2 Technical Details of Two Generators

Next, we will introduce the details of the speech generator and video generator we used, as well as how they receive the features passed from the CS and SD modules.

D.2.1 Speech Generator.

For our system's speech generation, we employ the state-of-the-art TTS model, StyleTTS2. Leveraging advanced diffusion models and adversarial training, StyleTTS2 is able to produce speech with emotions that are more authentic and natural than earlier TTS models, which is critical for our application. The input to StyleTTS2 includes the target empathic response text, denoted as t, and an optional reference mel-spectrogram, denoted as x. StyleTTS requires the input of the text script to be converted into speech, along with a reference audio containing emotional cues to serve as the style template. The modules responsible for reconstructing x in StyleTTS2 are listed as follows.

Text Encoder. The acoustic text encoder E_{aco} encodes the input phonemes into hidden representations:

$$\boldsymbol{h}_{\text{text}} = E_{aco}(\boldsymbol{t}) \tag{14}$$

Style Encoder. The style encoder E_{ref} encodes the input reference mel-spectrogram x into a styled vector:

$$\boldsymbol{ref_s} = E_{ref}(\boldsymbol{x}) \tag{15}$$

where ref_s encapsulates style information such as timbre, emotion, and other stylistic characteristics present in the reference audio.

Text Aligner. The text aligner A utilizes a dot product to extract aligned phoneme representations:

$$h_{algn} = h_{text} \cdot a_{pred} \quad (16)$$

from the input speech x and phonemes t. Here, $a_p red$ denotes the duration prediction, which is computed by

$$h_{bert} = BERT(t), \quad (17)$$

$$a_{pred} = E_{dur}(h_{bert}) \quad (18)$$

where BERT is pre-trained on extensive corpora of Wikipedia articles as a prosodic text encoder. D_{dur} is the duration predictor.

Duration Predictor. The duration predictor S predicts the duration of the reconstructed phonemes by:

$$d_{pred} = S(h_{text}, s) \quad (19)$$

Prosody Predictor. The prosody predictor P predicts the pitch and energy of the reconstructed phonemes by:

$$\hat{p}_x, \hat{n}_x = P(h_{text}, s) \quad (20)$$

Finally, the reconstructed speech is obtained by a speech decoder G:

$$\hat{x} = G(h_{algn}, ref_s, \hat{p}_x, \hat{n}_x) \quad (21)$$

D.2.2 Talking Head Generator.

DreamTalk is a sophisticated framework for generating expressive talking heads, utilizing diffusion models to deliver high-quality performance while minimizing reliance on expensive style references. The framework is comprised of a denoising network, a style-sensitive lip expert, and a style predictor. The denoising network employs diffusion models to create audio-driven facial movements that reflect the speaking style indicated by a reference video. The style-sensitive lip expert guarantees accurate lip synchronization and dynamic facial expressions, while the style predictor derives personalized speaking styles directly from the audio input.

Here, we will focus on introducing the denoising network E_θ, which learns to denoise the noisy motion m to obtain the predicted motion $m^*(0)$ under the conditions of audio window A_w and reference video R:

$$m^*(0) = E_\theta(A_w, R, m, t) \quad (22)$$

where t represents the time step, and E_θ consists of two encoders: the audio encoder E_{aud} and the style encoder E_{sty}. We denote the style code obtained from the style encoder as s. During inference, based on the style code s, DreamTalk employs the DDPM sampling algorithm to generate predicted facial motions.The generated facial motions are subsequently rendered into videos by the PIRenderer.

D.3 Training Details

Empatheia comprises four complementary training stages, each playing a crucial role in the overall process. In the following sections, we will provide a more detailed overview of the specific training details for each stage.

► **Training step1: CoE Learning Stage**

▽ Training Data: As shown in D.1, the CoE training data consists of dialogue context and targets. The dialogue context includes both the dialogue history and the user's current query input, while the target represents the CoE reasoning process and the empathetic response.

▽Training Objective: In this phase, we aim for the LLM to learn how to engage in step-by-step reasoning based on CoE to generate high-quality empathetic responses, effectively completing the ERG task. This stage lays the groundwork for subsequent content consistency learning.

▽Training Method: During this phase, we employ LoRA and Negative Log-Likelihood (NLL) loss to fine-tune Vicuna, training it to engage in empathetic reasoning by calculating the loss associated with the target component.

► **Training step2: Content Consistency Learning**

▽ Training Data: The training data for this phase not only includes the dialogue context from the first phase but also incorporates the pre-extracted audio and video content representations corresponding to each response text. Specifically, for speech, we input the response text into the two text encoders, E_{aco} and BERT, in StyleTTS2. The resulting embeddings, h_{text} and h_{bert}, are concatenated to produce the gold content representation for speech. For video, we utilize the audio encoder E_{aud} in DreamTalk to obtain the gold content representation for video.

▽Training Objective: The training objective for this phase is to align the speech and video representations output by the content synthesizer with the gold content representation. This alignment enables the CS to learn to produce outputs that are both accurate and exhibit consistent content signal features.

▽Training Method: In this phase, we freeze the parameters of the LLM and only fine-tune the parameters of the CS module. We calculate the L2 loss between the predicted content representation and the ground truth content representation.

► **Training step3: Style Aligning and Consistency Learning**

▽ Training Data: The input data format for this phase is illustrated in D.1, which uses multimodal special token placeholders in a dialogue format. The target training data consists of the audio and video style representations. The gold audio style representation is pre-extracted through the E_{ref} encoder of StyleTTS2, which processes the gold speech, while the gold video style representation is pre-extracted using DreamTalk's E_{sty} encoder, which processes the gold video. To ensure style consistency, in this stage, we also apply supervised constraints to the SD module using emotion and profile labels.

▽Training Objective:The goal of this phase is to align the speech and video style representations output by the SD module with the gold style representations, allowing the SD module to learn to produce accurate and consistent style signals.

▽Training Method: During training, we use ImageBind and a mapping layer to extract audio and video features from the multimodal dialogue, which are then used to replace the special token

placeholders. As in the previous phase, we freeze the parameters of the LLM and train the SD module using L2 loss and classification loss.

▶**Training step4: Overall Training**

▽ Training Data: As illustrated in D.1, the training data format for this stage encompasses all three modalities, integrating the training data from previous steps. The target data also includes all pre-extracted multimodal content and style representations.

▽Training Objective: This stage constitutes a comprehensive fine-tuning process aimed at equipping the model with the full capability of multimodal empathetic responses. It aims to provide not only sufficiently empathetic textual responses but also synchronized, content- and style-accurate multimodal responses.

▽Training Method: In this stage, we employ LoRA for fine-tuning the LLM while simultaneously updating the CS and SD modules. The overall loss is calculated as the weighted sum of the losses from the previous three stages.

E   Extended Experiment Settings

In this section, we provide more detailed experimental settings, encompassing hyperparameters, the construction of the pipeline, and the specifics of the evaluation.

E.1   Hyper-Parameter Settings

The basic hyperparameter settings for our training process are shown in Table 9.

Table 9: Some Basic Hyper-Parameter Settings

Basic Hyper-Parameters	Value
α	0.2
β	0.3
lora_r	16
lora_alpha	32
Num of Generate Tokens	16
Num of layers in Transformer Block	4

In order to reduce the usage of GPU memory and accelerate the training speed, we utilized Deepspeed to train our model, with some of the parameter settings presented in the Table 10.

Table 10: Some Deepspeed Hyper-Parameter Settings

Deepspeed Hyper-Parameters	Value
fp16	True
lr	5e-5
weight decay	0.001
train batch size	32
train micro batch size per gpu	4
gradient accumulation steps	8
zero optimization	stage2

E.2   Pipeline Baseline Implementation Details

We utilized the same LLM backbone as Empatheia to construct our pipeline, adopting StyleTTS2 and DreamTalk as multimodal generators. Unlike Empatheia, however, our pipeline solely relies on explicit meta-response texts to propagate generated instructional signals, whereas Empatheia employs implicit instructional features for signal transmission. For the training of our pipeline, we exclusively utilized text data and additionally introduced profile attributes. Consequently, after training, we can guide the subsequent

generation of speech and talking face videos based on the response text, emotion, and profile information outputted by the model.

E.3   Evaluation Details

Here, we detail how we conduct the evaluation for the various tasks, including the textual ERG task, speech generation task, and talking head generation task on the AvaMERG dataset.

E.3.1   Text ERG Evaluation Metrics.

For the text ERG task, we employ three evaluation metrics: Emotion Accuracy (Acc), and Distinct metrics (Dist-1 and Dist-2). These metrics are designed to evaluate both the emotional correctness of the generated responses and their lexical diversity.

Emotion Accuracy (Acc). Acc measures the percentage of correctly predicted emotions in the generated responses. The correct emotion is defined as the exact match with the ground truth emotion label. Acc is computed as follows:

$$Acc = \frac{\#\text{correct emotions}}{\#\text{total emotions}} \quad (23)$$

where 'correct emotions' represents the total number of responses where the predicted emotion matches the gold label.

Distinct-1 (Dis-1). This evaluates the diversity of unigrams (single words) in the generated responses. It is defined as the ratio of unique unigrams to the total number of unigrams in the generated text:

$$Dis\text{-}1 = \frac{\#\text{unique unigrams}}{\#\text{total unigrams}} \quad (24)$$

A higher Distinct-1 score indicates more lexical diversity and reduces the likelihood of repetitive responses.

Distinct-2 (Dis-2). This is analogous to Distinct-1, but it measures the diversity of bigrams (two consecutive words). It is computed as follows:

$$Dis\text{-}2 = \frac{\#\text{unique bigrams}}{\#\text{total bigrams}} \quad (25)$$

A higher Distinct-2 score indicates that the model is generating more contextually diverse phrases.

Human Evaluation of Textual ERG. For human evaluation, we randomly selected 200 dialogues from the test dataset. Taking into account both the cost of human labor and the reliability of the results, we chose competitive models from the last year as representative baselines. Given the dialogue context and the responses generated by these models, we engaged three annotators to score the responses using a majority voting system. They rated each response on a scale from 1 to 5 (1: not at all, 3: adequate, 5: excellent) based on four key criteria: Empathy, Coherence, Informativity, and Fluency. Specifically, these criteria are:

1) **Empathy** (Emp): whether the response demonstrates an understanding of the user's emotions and experiences, and responds appropriately.
2) **Coherence** (Coh): whether the response is logically consistent and contextually relevant.
3) **Informativity** (Inf): whether the response provides useful and meaningful information.
4) **Fluency** (Flu): whether the response is grammatically well-formed and easy to read.

E.3.2 Speech Generation Evaluation Metrics.

For the speech generation component of MERG, we use subjective and objective metrics to assess the quality and emotional expressiveness of the generated speech.

Mean Opinion Score (MOS). MOS is a subjective evaluation metric where human evaluators rate the naturalness of the generated speech on a scale of 1 to 5, with 5 indicating highly natural speech. MOS is computed as:

$$MOS = \frac{\sum \text{ratings}}{\#\text{evaluators}} \tag{26}$$

Similarity MOS (SMOS). SMOS measures how similar the generated speech is to a reference speech sample in terms of emotional tone and expressiveness. Like MOS, it is rated on a 5-point scale by human evaluators.

$$SMOS = \frac{\sum \text{similarity ratings}}{\#\text{evaluators}} \tag{27}$$

Mel Cepstral Distortion (MCD). MCD is an objective metric that measures the distance between the generated speech and the reference speech. Lower values of MCD indicate higher similarity between the two speech samples. It is computed using the logarithmic differences between the mel-cepstral coefficients (MFCC) of the generated and reference audio:

$$MCD = \frac{10}{\log 10} \sqrt{2 \sum_{i=1}^{d} (c_i^{\text{gen}} - c_i^{\text{ref}})^2} \tag{28}$$

where c_i^{gen} and c_i^{ref} are the mel-cepstral coefficients of the generated and reference speech, respectively.

E.3.3 Talking Head Avatar Generation Evaluation Metrics.

For evaluating the quality of the generated talking head avatars, we employ a combination of perceptual and geometric metrics to measure the visual fidelity and synchronization of the avatar's lip movements with the speech.

Structural Similarity Index Measure (SSIM). SSIM assesses the visual similarity between the generated avatar video frames and the ground truth video frames. It measures the perceived quality in terms of luminance, contrast, and structure:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{29}$$

where μ_x and μ_y are the means of the generated and reference images, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance of the two images. C_1 and C_2 are constants to stabilize the division.

Cumulative Probability of Blur Detection (CPBD). CPBD quantifies the perceptual sharpness of the generated video frames by estimating the probability that an observer would detect blurring based on edge width analysis. Higher CPBD values indicate sharper images and fewer perceived blurs. The CPBD score is calculated by first analyzing the cumulative distribution of edge widths in an image. For each edge in the image, the probability that the edge is perceived as blurred is computed. The overall CPBD score is the average of these probabilities over all edges. The formula can be expressed as:

$$CPBD = \frac{1}{N} \sum_{i=1}^{N} P(e_i) \tag{30}$$

where $P(e_i)$ is the probability of blur detection for edge e_i, and N is the total number of detected edges in the image. This metric gives a perceptual estimate of how likely it is for human viewers to notice blur across the video frames.

SyncNet Confidence Score (Sync_{cf}). Sync_{cf} is used to evaluate the synchronization between the generated speech and the avatar's lip movements. It measures the alignment between the visual lip movements and the audio, with higher scores indicating better synchronization:

$$Sync_{cf} = \frac{\text{aligned speech and lip movements}}{\text{total frames}} \tag{31}$$

Full-Face Landmark Distance (F-LMD). F-LMD measures the Euclidean distance between facial landmarks on the entire face in the generated video and the ground truth reference. Lower values indicate better alignment between the generated and real facial movements:

$$F\text{-}LMD = \frac{1}{N} \sum_{i=1}^{N} \|L_{gen}(i) - L_{ref}(i)\| \tag{32}$$

where $L_{gen}(i)$ and $L_{ref}(i)$ represent the facial landmarks of the generated and reference avatars, respectively, and N is the total number of frames.

Mouth Landmark Distance (M-LMD). M-LMD focuses on the landmarks around the mouth region, which are crucial for evaluating lip synchronization. Like F-LMD, lower values indicate better alignment of mouth movements with speech.

$$M\text{-}LMD = \frac{1}{N} \sum_{i=1}^{N} \|M_{gen}(i) - M_{ref}(i)\| \tag{33}$$

E.3.4 Human Evaluation of MERG.

To better evaluate the performance of models on the MERG task, we have newly defined six human evaluation metrics, specifically:

- **Speech Content Accuracy (SCA):** Assesses whether the content in the generated audio is complete and consistent with the response text.
- **Video Content Accuracy (VCA):** Evaluates whether the face in the generated video accurately and fluently reads out the response text.
- **Speech Style Accuracy (SSA):** Determines the accuracy of the emotion conveyed in the generated speech and whether the voice matches the intended character profile.
- **Video Style Accuracy (VSA):** Assesses the accuracy of the emotion expressed by the generated avatar and whether the avatar's appearance and behavior match its intended profile.
- **Multimodal Content Consistency (MCC):** This is a comprehensive comparison that evaluates the consistency of content across the three modalities (speech, video, and text).
- **Multimodal Style Consistency (MSC):** Provides an overall evaluation of the consistency of style across the three modalities.

During the testing process, we deliberately engaged three experienced evaluators to ensure the comprehensiveness and accuracy
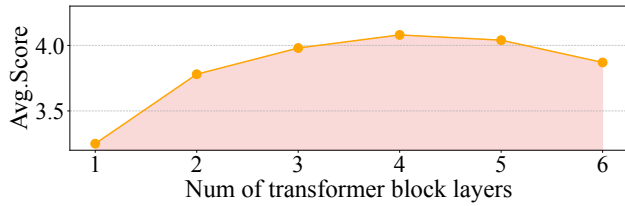
Figure 13: Impact of the transformer layer number.
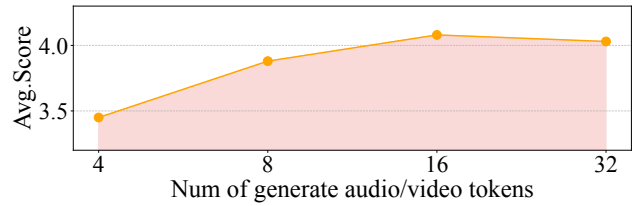
Figure 14: Impact of special token numbers.

Figure 15: Impact of loss weights.

of the evaluation. The task of these three evaluators was to conduct meticulous assessments of 200 test cases based on the aforementioned six aspects, grading them from 1 to 5 according to the model's performance (where 1 represents extremely poor performance and 5 represents excellence). Prior to the testing, we organized a detailed training session to clarify the scoring criteria, grading rules, and handling methods for potential special cases, thereby further ensuring the reliability of the scores.

F   More Experiments and Analysis

In this part, we present additional experimental results and analyses to further demonstrate the performance of Empatheia.

F.1   Impact of Hyperparameter Settings

Here we study the impact of various hyperparameters on model performance, including the number of transformer blocks, the number of special signal tokens used for multimodal generation, and the loss weights alpha and beta.

Figure 13 illustrates the model's performance when varying the number of transformer blocks in CS and SD modules, specifically evaluating the model's output on multiple metrics. The "Avg. Score" reflects the average performance across six carefully crafted manual evaluation metrics, which together capture various aspects of the model's ability to generate multimodal empathetic responses. The results show that the transformer block reaches its peak performance at the 4 layers. Deeper layers can, to some extent, enhance the model's learning ability by capturing more complex interactions between patterns. However, further increasing the number of transformer layers did not lead to significant improvement. Therefore, Empatheia sets the number layers of transformer blocks to 4. Figure 14 illustrates the impact of different numbers of audiovisual special tokens on model performance. As shown, the model's performance peaks when the number of tokens reaches 16.

We scale the content and style learning losses using α and β to bring them closer to the magnitude of the Vicuna's inherent loss, aiming to achieve balanced training. The experiments on loss hyperparameters are visualized in a heatmap in Figure 15, the model performs better when α and β are relatively balanced, indicating that a more balanced loss contributes to the model's convergence to an optimal solution.

F.2   Impact of Training Data Amounts

Figure 16 presents the performance of Empatheia and the pipeline under varying proportions of training data. We incrementally increased the proportion of the training dataset from 0%, 20%, 50%, to 100%. It is observable that both the Pipeline and Empatheia exhibit enhanced performance as the volume of data increases. Notably,
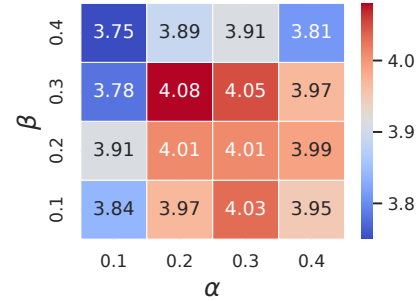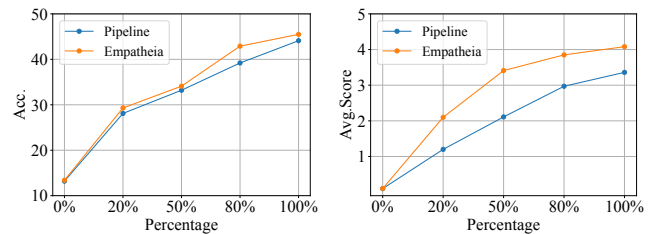
Figure 16: Performance comparison under different quantities of training data.

due to a lack of understanding of multimodality and a synchronous module for multimodal generation, the average score growth rate of the Pipeline in terms of accuracy and consistency in multimodal generation lags significantly behind Empatheia. This once again demonstrates the robust potential of Empatheia.

F.3   More Case Study

We further provide several examples to demonstrate the performance differences between our model and pipeline.

MERG Performance Comparisons. As illustrated in Figure 17, Figure 18, Figure 19 and Figure 20, our model exhibits more consistent and empathetic multimodal responses, indicating Empatheia's deeper understanding of multimodal contexts. This is attributed to our multi-stage training approach, as well as the advanced content synchronizer and style deconstructor employed.

CoE Qualitative Results. We also present several examples to visualize the reasoning process of CoE. As shown in Figure 21, Figure 22, Figure 23 and Figure 24, CoE can significantly enhance the model's empathy capabilities and the accuracy of perceiving users' emotions, the underlying intention, and finally the correct response.

Dialogue history

A little while ago, my friend went on another beach vacation. It was her 3rd one for the year. My family and I hadn't been on vacation for several years, due to issues beyond our control, so I begrudged her trip a bit.

Being jealous of another person is a perfectly fine response. Despite this, you never know what is going on behind the scenes, she may have taken out huge loans just to fund her trips.

Current user query   *Sad face, Sad voice*

Yes, true, I know they tend to let bills go for things like that, which I don't do.

Sad face, Sad voice   **Gold Response**

I bet you would go on many vacations if you did the same as her. Instead, you are being a responsible person and doing what is best for your future.

sad face, angry voice   Response by **Pipeline model**

Well, vacations aren't everything. There are plenty of other things to focus on instead. Everyone has their own priorities.

sad face, sad voice   Response by **Empatheia**

It's understandable that you feel that way. It can be tough when you see others enjoying things that are out of reach for you right now.

Figure 17: Qualitative results A of MERG.

Dialogue history

When I was pregnant, I felt ill until it was over. The doctors were worried that something was wrong with the baby.

Oh no! I'm not sure how all that works, but I thought that was fairly normal for a pregnant woman.

Current user query   *happy face, happy voice*

Usually the sickness stops within the first few months, so the fact that I fell ill so much was worrisome. Despite it all, I had a healthy, bouncing baby boy. I never have felt more grateful in my life.

happy face, happy voice   **Gold Response**

Well, I'm glad for you! And for him! A new life is exciting.

fear face, fear voice   Response by **Pipeline model**

Well, it sounds like things worked out in the end, so that's good. I've heard pregnancy can be tough for some people.

happy face, happy voice   Response by **Empatheia**

That sounds like such a difficult time, with all the uncertainty and worry throughout your pregnancy. It must have been so stressful, but I'm so glad to hear that everything turned out well and you have a healthy baby boy.

Figure 18: Qualitative results B of MERG.

Dialogue history

Lately, I find myself struggling to keep up with daily tasks. It feels like I'm losing my grip.

It's understandable to feel that way, especially with so much change happening. Do you want to talk about what's been most challenging for you?

Current user query

fear face, fear voice

That sounds really tough. It's hard when your body doesn't cooperate like it used to. Have you found any little things that help you cope?

fear face, fear voice **Gold Response**

That sounds really tough. It's important to take care of yourself during challenging times. Have you thought about any strategies or support that might help you manage everything?

disgusted face, fear voice **Response by Pipeline model**

Honestly, it sounds like you're just not trying hard enough. Everyone has to deal with feeling tired sometimes, but they still get things done.

fear face, fear voice **Response by Empatheia**

It's completely normal to struggle when things become overwhelming. It must be exhausting to deal with such changes. If you'd like to share more about what's been weighing on you, I'm here to listen.

Figure 19: Qualitative results C of MERG.

Dialogue history

One time, I smelled an odd burning smell coming from under the hood of my car.

That's strange. What did you discover under there?

Current user query

disgusted face, disgusted voice

Well, when I went to lift the engine cover, I noticed a dead, rotting squirrel on top of the engine block! It was so repulsive.

disgusted face, disgusted voice **Gold Response**

Oh my! That is repulsive indeed! I would've taken it home for dinner. Waste not want not!

fear face, fear voice **Response by Pipeline model**

That must have been such a shock! I can't imagine how alarming it was to find something like that in your car.

disgusted face, disgusted voice **Response by Empatheia**

Oh no, that sounds absolutely disgusting! Finding a rotting squirrel on your engine block must have been so gross. I can't blame you for being repulsed by it — what an awful thing to deal with!

Figure 20: Qualitative results D of MERG.

Dialogue history

I went to a heavy metal music festival in Las Vegas a few years ago! I had fun, but to be honest it was a bit of an emotional roller coaster.

Oh, I don't know why concerts make me feel a bit sentimental.

Current user query

I think it's because everybody can relate when it comes to live music. My girlfriend broke up with me the second day, but we had to stay in the same room.

CoE reasoning in Empatheia

1. Event Scenario: A breakup occurred during a heavy metal festival in Las Vegas while sharing a room with the ex-girlfriend.

2. User Emotion: Lonely

3. Emotion Cause: The emotional turmoil was caused by the breakup during an otherwise enjoyable music festival, combined with the discomfort of being stuck in the same room with the ex-girlfriend.

4.Goal to Response: Acknowledge the conflicting emotions, validate the difficulty of the situation, and offer empathy for the discomfort and emotional roller coaster the speaker went through.

Response by Empatheia

It sounds like that must have been a really tough experience for you, having to go through such an emotional breakup in the middle of what was supposed to be a fun time. I'm really sorry you had to go through that.

Figure 21: Qualitative results A of CoE mechanism.

Dialogue history

I am so glad that I decided to pack a kit for my car in case of bad winter weather.

I bet you are, that's some pretty good foresight right there! Did you have to use it this past winter?

Current user query

I did, two days after packing it we had a blizzard and my tire went flat out in the middle of it. I would have been very unhappy had I not packed it.

CoE reasoning in Empatheia

1. Event Scenario: An emergency kit was packed for winter weather, which proved essential during a blizzard.

2. User Emotion: Prepared

3. Emotion Cause: Relief and appreciation come from the foresight of packing the kit, which turned out to be essential when their tire went flat in the middle of a blizzard.

4.Goal to Response: Validate the speaker's sense of relief and preparedness, acknowledging the stressful situation they avoided.

Response by Empatheia

Wow, it's a good thing you were so prepared! That must have been such a stressful situation, but having the kit with you really saved the day. It's amazing how a little foresight can make such a big difference.

Figure 22: Qualitative results B of CoE mechanism.

Dialogue history

A week before my cruise this summer, I was counting down the days with bated breath. I thought I would actually DIE before I got to go on it.

It's the worst waiting for vacations! You want the time before to go fast but then it to go slow during. Never works out that way, unfortunately.

Current user query

Yes! Completely the opposite of what you want. The universe can be funny like that. But I had an absolute blast!

CoE reasoning in Empatheia

1. Event Scenario: Eagerly counting down the days, anxiety built for the upcoming summer cruise.

2. User Emotion: Excited

3. Emotion Cause: The anticipation and anxiety surrounding the upcoming trip.

4. Goal to Response: To provide empathy and acknowledge the speaker's excitement and enjoyment of the trip.

Response by Empatheia

I totally understand that feeling of anticipation! Waiting for a vacation can feel so long, but it's all worth it when you finally get to enjoy the trip. It sounds like you had an amazing time on the cruise, which is fantastic!

Figure 23: Qualitative results C of CoE mechanism.

Dialogue history

I feel so bad when I think back to times that I got angry and yelled at my kids for no real good reason -- just a combination of the day's events and feeling irritated, really.

We get carried away with our emotions sometimes. Did you explain it to them?

Current user query

Yes! That I feel good about! Making a point to use it as an opportunity to talk. I've definitely used it when it comes to the way others might act -- emphasizing that it's not about *them*, it's really about the person and where they're at.

CoE reasoning in Empatheia

1. Event Scenario: Reflecting on moments of losing temper with children without justification.

2. User Emotion: Guilty

3. Emotion Cause: Irritation and frustration accumulated throughout the day led to these outbursts.

4. Goal to Response: Offer empathy and validation for the speaker's feelings, acknowledging the challenge of parenting and the importance of communication.

Response by Empatheia

Wow, it's a good thing you were so prepared! That must have been such a stressful situation, but having the kit with you really saved the day. It's amazing how a little foresight can make such a big difference.
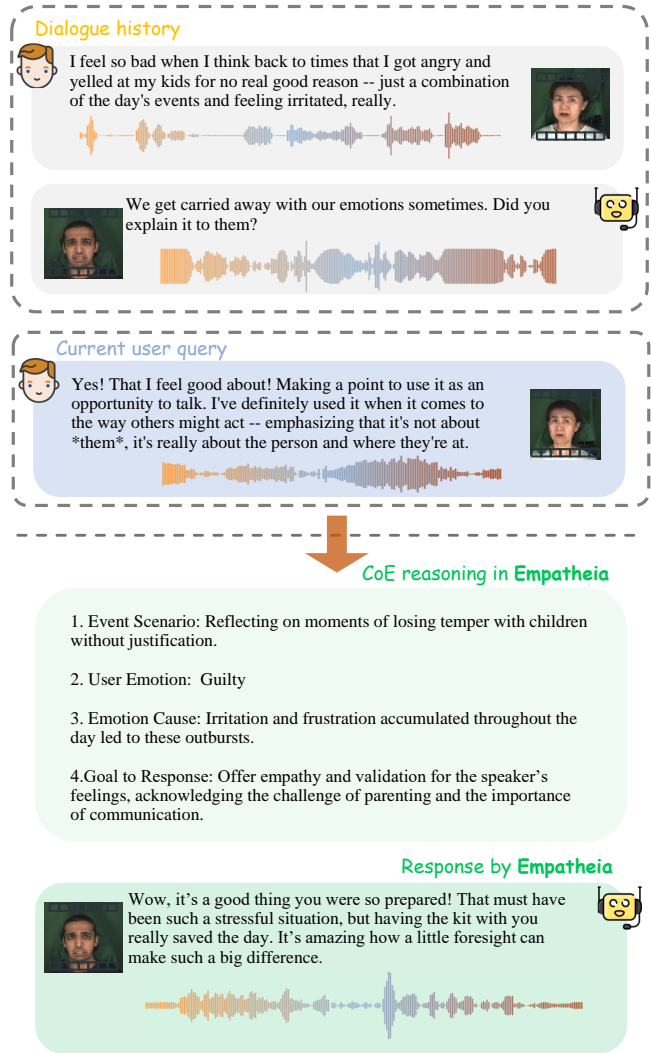
Figure 24: Qualitative results D of CoE mechanism.