Speak From Heart: An Emotion-Guided LLM-Based Multimodal Method for Emotional Dialogue Generation

Chenxiao Liu
lcxxx@mail.ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

Zheyong Xie
xiezheyong@mail.ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

Sirui Zhao
sirui@mail.ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

Jin Zhou
zhou229103@mail.ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

Tong Xu*
tongxu@ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

Minglei Li
liminglei29@huawei.com
Huawei Cloud
Shenzhen, China

Enhong Chen
cheneh@ustc.edu.cn
University Of Science And Technology Of China
Hefei, China

ABSTRACT

Recent advancements in Large Language Models (LLMs) have greatly enhanced the generation capabilities of dialogue systems. However, progress on emotional expression during dialogues might be still limited, especially when capturing and processing the multimodal cues for emotional expression. Therefore, it is urgent to fully adapt the multimodal understanding ability and transferability of LLMs to enhance the emotional-oriented multimodal processing capabilities. To that end, in this paper, we propose a novel Emotion-Guided Multimodal Dialogue model based on LLM, termed ELMD. Specifically, to enhance the emotional expression ability of LLMs, our ELMD customizes an emotional retrieval module, which mainly provides appropriate response demonstration for LLM in understanding emotional context. Subsequently, a two-stage training strategy is proposed, founded on previous demonstration support, to support uncovering nuanced emotions behind multimodal information and constructing natural responses. Comprehensive experiments demonstrate the effectiveness and superiority of ELMD.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Discourse, dialogue and pragmatics**.

*Corresponding Author.

ICMR '24, June 10–14, 2024, Phuket, Thailand.

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658104

KEYWORDS

Large Language Models; Emotional expression; Multimodal cues; Emotional retrieval module; Dialogue systems

ACM Reference Format:
Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024. Speak From Heart: An Emotion-Guided LLM-Based Multimodal Method for Emotional Dialogue Generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3652583.3658104

1 INTRODUCTION

Emotional expression, as an essential element in human communication, plays a crucial role in improving the naturalness and user experience of dialogue. Correspondingly, it raises a significant challenge for dialogue systems to effectively perceive and respond the emotional factors during dialogues. For example, as shown in Figure 1, in order to generate an appropriate response, the dialogue system should understand the man's anger and express apologies, otherwise, it could further enrage the man and completely ruin this conversation. Therefore, large efforts have been made to achieve better emotional expression for dialogue systems [21]. Traditional efforts usually rely on pre-defined emotion categories to generate targeted responses [31, 40], or attempt to combine visual cues (e.g., actions and facial expressions) to promote the multimodal emotional comprehension [5, 29, 38]. Unfortunately, they still suffer the limited capabilities to understand and generate long text, as well as the lack of ability to deeply understand visual information. Therefore, more comprehensive solutions for emotional dialogue systems are still urgently required.

Recently, the rapid development of large language models (LLMs) unlocks new possibilities in the field of sentiment analysis. Along this line, InstructERC [12] made the initial attempt to correlate

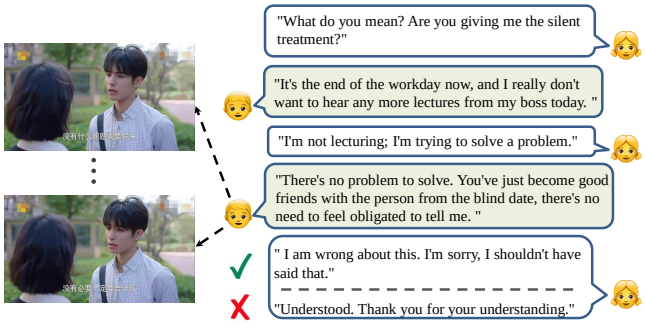

Figure 1: Example of Emotional Dialogue.

sentiment with LLMs via enhancing the emotion recognition in dialogue. However, relying solely on semantic information in the retrieval module may lead to emotional biases, and InstructERC could not be applied in multimodal scenarios. To that end, some other researches, e.g., [1, 16, 42], have attempted to incorporate visual information into MLLMs for better emotion recognition. Although these methods have achieved good results, they may still face difficulties in visual understanding, especially the implicit emotional cue, which severely hinders their ability to produce responses that closely mimic the nuanced and natural emotional expressions exhibited by humans.

To address these problems, in this paper, we propose a novel Emotion-Guided LLM-based Multimodal Dialogue (ELMD) framework, to enhance the emotional recognition and expression ability of LLM. The ELMD framework mainly consists of three parts, namely emotional retrieval module (ERM), Response Emotion Prediction (REP), and Emotion-Enhanced Response Generation (EERG). Specifically, in the ERM stage, we first adopt a contrastive learning loss to train the emotion representation component and self-separated component, for obtaining the fine-grained emotional representations. In this way, better retrieval will be achieved to construct response demonstration. Then, in the REP stage, we guide the learning process of visual features by predicting emotion labels from generated responses. In this way, the model will be significantly enhanced to uncover emotional cues within the scene context. Finally, we conduct additional training for the entire model during the EERG stage, which aims to harmonize information from diverse modalities, fine-tune the interplay between various components of the model, and refine the ability to generate natural responses with nuanced emotions. Generally, the technical contribution of this paper could be summarized as follows:

- To the best of our knowledge, we are the first LLM-based emotional multimodal dialogue framework proposed in Chinese corpus, aiming to make full use of multimodal cues to achieve an emotion-rich and human-like dialogue system.
- We propose an emotional retrieval module and next emotion prediction training method, which is used to supplement the emotional content, provide guidance for LLM from the perspective of both textual and visual cues, and enhance the emotional dialogue generation ability.
- Our proposed method has achieved impressive results in multiple real-world datasets, and the results also prove that emotional information supplement enhances the model's ability to generate emotionally rich textual content that is suitable for the current scene.

2 RELATED WORK

2.1 Emotion Recognition in Conversation

In the realm of conversational human-machine interaction, the initial and pivotal step lies in discerning user emotions, paving the way for more nuanced and intelligent communication. Existing ERC models can be classified into three categories. The first category is recurrent-based methods [7, 9, 10, 25]. These approaches typically involve the design of various levels of encoders and recurrent neural units to extract emotional features. For instance, Majumder et al. [25] employs three GRUs to model the speaker, context, and emotion aspects. The second category focuses on graph-based methods, as highlighted by previous research studies such as [8, 30]. These methods leverage nodes and edges to represent the intricate roles and conversational dynamics within dialogues. Notably, a recent study by [13] proposes an innovative approach called S+PAGE, which employs a two-stream conversational Transformer architecture. This architecture effectively captures features from both the current conversation and the broader context. The third category of emotion recognition methods centers around LLM-based techniques, which harness the context-aware capabilities of Large Language Models trained on extensive corpora. These models aim to directly generate emotion labels by capturing emotional information from the dialogue context.

2.2 Emotional & Empathetic Conversation

Building upon the ability to recognize user emotions, effectively conveying emotions is paramount in fostering meaningful conversations. Emotional chatting aims to generate responses with predefined emotions [40]. Zhou and Wang [41] propose an Emotional Chatting Machine (ECM) to generate emotional responses. In addition to explicitly providing emotion labels, on certain social platforms, emojis sometimes convey hidden emotional information [38, 40]. Zhou et al. [40] adopts the attention-based SEQ2SEQ model to extract the concealed emotional information from emojis, thereby aiding in the generation of emotional responses. Compared to emotional chatting, which merely generates response with specific emotion, empathetic conversation aims to generate dialogues with empathetic responses [18, 23, 24, 39]. It not only attends to the emotional expressions but also emphasizes understanding the user's sentiments, responding in a compassionate, empathetic, and supportive manner. For instance, Zheng et al. [39] proposes a multi-factor hierarchical framework called CoMAE for empathetic response generation, which models three factors: communication mechanism, dialog act, and emotion.

2.3 Multimodal Dialouge Systems

While text carries rich information, human communication often involves multiple modalities, encompassing aspects beyond textual content, such as voice, facial expressions, and body language.

Numerous studies delve into the integration of multimodal information into conversational systems [14, 20, 26, 32, 35]. Nie et al. [26] propose the MAGIC model, which determines the type of response by understanding the intent within a given multimodal context. Then it employs RNNs to generate responses. Recently, MLLMs have emerged as a new research focus, leveraging the powerful text generation capabilities of large language models to function as the brain for executing multimodal tasks[1, 3, 16, 36, 42]. Existing MLLM models employ two main types of approaches to incorporate visual information. The first kind of approach, exemplified by Flamingo [1], utilizes a frozen visual encoder and employs a gated cross-modal attention mechanism for alignment between text and images. The second type of approach, represented by models like Mingpt4 and LLava [16, 42], leverages Q-Former to reduce the sequence length of visual features and projects visual features onto the same dimension as text features through a linear layer.

3 METHODOLOGY

3.1 Task Definition

To describe multimodal emotional dialogue generation intuitively, we provide a formal definition of this task. For a given multimodal dialogue data, we assume the current dialogue utterance is represented as u and the corresponding dialogue history is denoted as $\mathcal{H} = \{u_1, u_2, ..., u_n\}$. Additionally, the visual information of the current user's scene is represented as $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, which contains m frames. Ultimately, our goal is to learn a model $P(r|u, \mathcal{H}, \mathcal{V}; \theta)$ capable of producing responses r that align with the emotional nuances inherent in the present scenario.

3.2 Overview

To achieve the objective we stated before, we propose a novel Emotion-Guided LLM-Based Multimodal Dialogue Method, which mainly consists of three parts:

(1) Emotional Retrieval Module (ERM): This module is designed to fetch an emotionally relevant pair of u' and r', where u' contains semantically and emotionally relevant content that aligns with current utterance u. The r' is extracted as supplementary information to integrate into the demonstration.
(2) Response Emotion Prediction (REP): This stage is designed to bridge the significant gap between visual and emotional information. It establishes the correlation between visual context and emotional outcomes in response through explicit emotional guidance.
(3) Emotion-Enhanced Response Generation (EERG): This stage integrates context, visual information, and supplementary content to enhance the emotional response generation ability of the target model.

Then we will provide a detailed description of these components in following sections, step-by-step.

3.3 Emotional Retrieval Module

Inspired by [6, 17], we find that providing task-relevant data as a demonstration when using LLM can effectively enhances the model's expressiveness. Therefore, we customize an Emotional Retrieval Module to retrieve response content r' associated with utterance u' similar to u, serving as a demonstration to supplement the model with necessary exemplar information. However, traditional semantic representations contain only semantic information, lacking emotional relevance. This deficiency of emotion can lead to biases in the target demonstration. Moreover, the original representations lacks fine-grained use of information. To address these limitations, as shown in Figure 2, we jointly train the Emotional Retrieval Module with two components: an emotion representation training component (ERT) and a self-separated component (SSC).

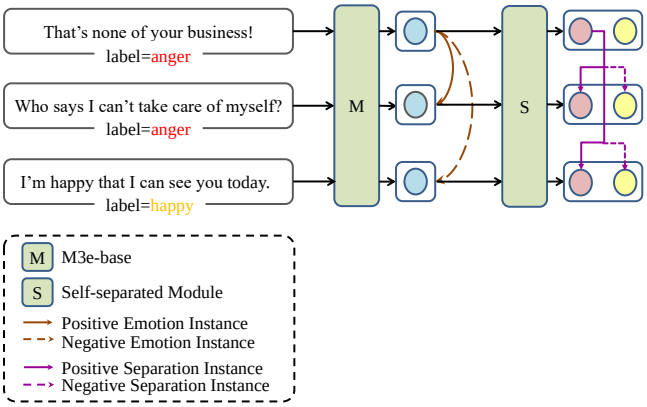

Figure 2: The architecture overview of the Emotional Retrieval Module.

Specifically, for a given user utterance u, we initially employ M3e-base [34] to obtain its text embedding $e_u \in \mathbb{R}^{1\times d}$ as a representation of its semantic content.

$$e_u = M3e\text{-}base(u) \tag{1}$$

Subsequently, for the emotion representation training component, we adopt a contrastive learning approach based on emotion labels. Specifically, we treat utterances with the same emotion as positive samples and those with different emotions as negative samples. More formally, for a given representation e_u from a sample, a more formal expression is provided as follows:

$$\mathcal{L}_e = -log\frac{exp(sim\left(e_u, e_u^p\right)/\tau)}{\sum_{k=1}^{n_e} exp(sim\left(e_u, e_u^k\right)/\tau)} \tag{2}$$

Here, e_u^p represents the representation of the utterance for positive samples, τ is the temperature hyperparameter, $sim\left(e_u, e_u^p\right)$ is the cosine similarity, and n_e is the sum of positive and negative samples for the current sample.

Based on the training of emotional enhancement representations, we further enhance representations through a self-separated module motivated by Tang et al. [33]. Specifically, we utilize the self-separated module constructed by a linear layer to progressively partition the input feature e_u into a stack of s vectors, yielding vectors $E_s = \{h_1, h_2, \ldots, h_s\}$, seeing in Algorithm 1. Each vector $h_i \in \mathbb{R}^{1\times d}$ is composed of distinct information.

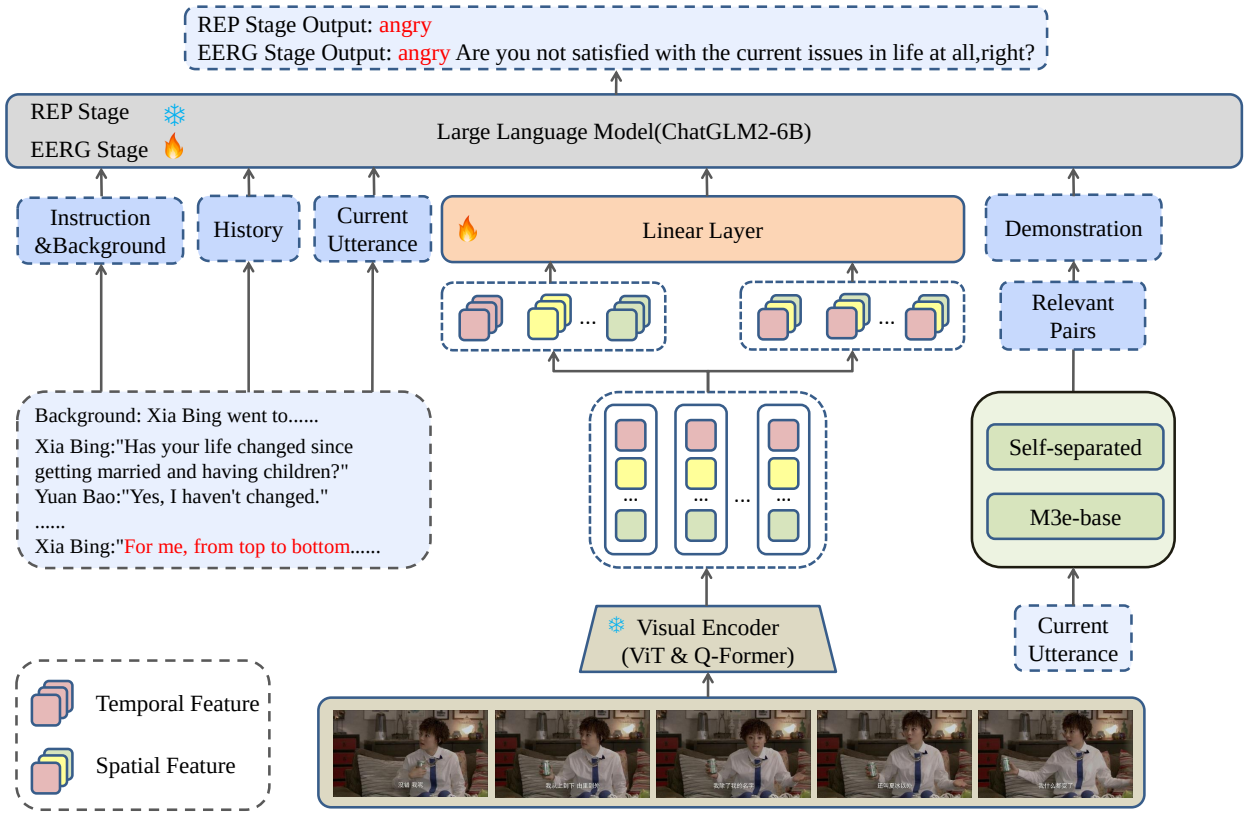

Figure 3: The architecture overview of the LLM-Based Emotion-Guided Multimodal Dialogue Method.

Algorithm 1 Self-Separation

Input: $e_u \in \mathbb{R}^{1\times d}$:the vector representation of original utterance;
s: hyper-parameter, the self-separation coefficient
d: the dimension of the hidden state;
Output: $E_s \in \mathbb{R}^{s\times d}$: the vector representations of emotional and semantic information after processing. In this context, it is the form of a set;

Initialize E_s
Set $g \leftarrow$the integer of d/s;
for $i = 1$ to s do **do**
  Initialize augment vector
  $c_i \leftarrow (0, 0, \ldots, 0)_{1\times d}$
  Set $c_i[(i-1)\times g + 1 : i \times g] \leftarrow (1, 1, \ldots, 1)_{1\times g}$;
  $E_s[i, :] \leftarrow MLP(e_u + c_i; c_i)$;
end for
return E_s

To let the sparse semantic component automatically partition the e_u, we conduct unsupervised training on the sparse semantic components concurrently with emotion enhancement training. Specifically, for two different samples, we choose one sample $E_s = \{h_1, h_2, \ldots, h_s\}$ as the basic one, and select representations at the same positions from another sample $E_s^{'} = \{h_1^{'}, h_2^{'}, \ldots, h_s^{'}\}$ as positive samples and representations at different positions as negative samples. Through this way, we can construct s sets of different positive-negative sample pairs. The training objective is as follows:

$$\mathcal{L}_c = \sum_{j=1}^{s} -log\frac{exp(sim\left(h_j, h_j^{'}\right)/\tau)}{\sum_{k=1}^{s} exp(sim\left(h_j, h_k^{'}\right)/\tau)} \tag{3}$$

As constructing one emotional positive-negative sample pair introduces $n_e + 1$ dialogue utterance, we simultaneously build $n_s = \frac{(n_e+1)n_e}{2}$ positive-negative sample pairs for training the self-separated module. In the example provided in Figure 2, n_e is 2. Therefore, our ultimate training objective can be described as:

$$\mathcal{L}_r = \lambda\mathcal{L}_e + \frac{1}{sn_s}\sum_{i=1}^{n_s}\mathcal{L}_c \tag{4}$$

where λ is a hyperparameter used to adjust the weight of the emotion-contrastive loss in the joint training loss.

After training completion, for each u from a dialogue, we obtain the corresponding representation E_s through this module. Then, we select a $u^{'}$ with the maximum cosine similarity to E_s and return its response $r^{'}$ with a special instruction as demonstration $\mathcal{R}$.

3.4 Response Emotion Prediction

In human communication, individuals often convey emotions through facial expressions and body movements. To capture the emotion behind these visual actions, we employ a visual encoder to extract features from images [11]. However, these high-level representations frequently lack a direct incorporation of emotional information. To bridge the gap between visual features and emotional information, we have designed a response emotion prediction training strategy, as illustrated in Figure 3.

More specifically, we employ the target response emotion to guide the model in capturing visual cues that are intricately linked to emotions. Firstly, we employ ViT and Qformer as our visual encoders with initialized weights sourced from VisualGLM, and ChatGLM2-6B[1] serves as the LLM. To process video input, we adopt the approach from Video-Chatgpt [22] to obtain features in both temporal and spatial dimensions. More formally, given a video sample $\mathcal{V}$ with m frames, we pass it through the visual encoder to obtain $X \in \mathbb{R}^{m \times o \times q}$.

$$X = VisualEncoder(\mathcal{V}) \in \mathbb{R}^{m \times o \times q} \tag{5}$$

We then perform pooling on X in both temporal and spatial dimensions, yielding temporal feature $V_t \in \mathbb{R}^{o \times q}$ and spatial feature $V_s \in \mathbb{R}^{m \times q}$. Concatenating these temporal and spatial features results in the video-level feature $V_c \in \mathbb{R}^{(m+o) \times q}$.

$$V_t = \text{mean}_{\text{dim}=0}(X) \in \mathbb{R}^{o \times q} \tag{6}$$

$$V_s = \text{mean}_{\text{dim}=1}(X) \in \mathbb{R}^{m \times q} \tag{7}$$

$$V_c = [V_t, V_s] \in \mathbb{R}^{(m+o) \times q} \tag{8}$$

Subsequently, a linear layer is applied to project it into the word embedding space of the LLM input, yielding the fixed-length video token E_V.

$$E_V = Linear(V_c) \tag{9}$$

After obtaining the general visual feature, we need to align it with the response emotion through the LLM. Following the general practice of previous work [12], we build a prompt template for LLM input as shown in Figure4, including six parts. Specifically, the instruction $\mathcal{I}$ and background $\mathcal{B}$ provide the basic guidance and the background information of dialogue, which employs the summaries of the preceding episodes of dialogue segments[2]. Then, we use the template to combine visual features as model input for response emotion prediction training. It is worth noting that we freeze the visual encoder and LLM and train only the linear layer. Therefore, the training objective is:

$$\mathcal{L}_V = -\sum_{i=1}^{N} \log p\left(e_i | \mathcal{I}, \mathcal{B}, \mathcal{H}, u, E_v, \mathcal{R}; \theta\right) \tag{10}$$

Here, e_i is the token for the emotion label, θ represents the parameters of the model. N denote the size of training dataset.

[1]https://github.com/THUDM/ChatGLM2-6B
[2]https://baike.baidu.com/

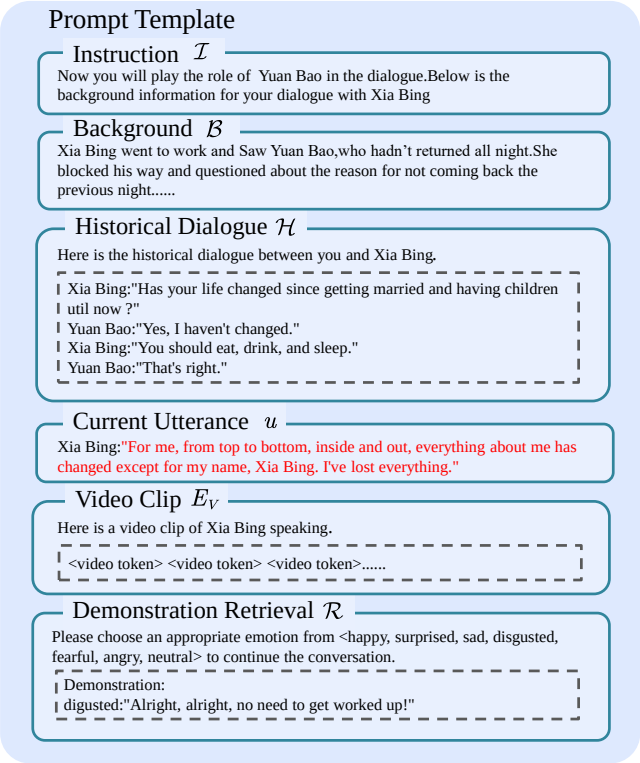

Figure 4: The Schematic of Prompt Template

3.5 Emotion-Enhanced Response Generation

After the training in the REP phase, the Linear layer has acquired a certain ability to extract emotional information from visual features. To harmonize information from diverse modalities, we further conduct response generation training with information acquired previously. In detail, we employ the ptuning[19] to fine-tune the LLM and optimize the visual linear layer at the same time. Here we obtain the same input as mentioned in Section 3.4 as model input. Therefore, our final training objective is:

$$\mathcal{L} = -\sum_{i=1}^{N} \log p\left(\epsilon_i | \mathcal{I}, \mathcal{B}, \mathcal{H}, u, E_v, \mathcal{R}; \theta\right) \tag{11}$$

where the ϵ_i represents the tokens for the emotion and response.

Table 1: Partition information of multimodal emotional dialogue datasets.

Dataset	Train	Valid	Test
M3ED	6071	758	758
MMED	5468	683	683

Table 2: The automatic evaluation results on M3ED and MMED datasets with selected metrics. The best results are in bold.

	Model	BLEU-2	BLEU-4	METEOR	ROUGE-1	ROUGE-L
M3ED	Livebot	0.7041	0.070	5.237	13.46	10.42
	DialoGPT	2.1513	0.282	5.726	12.29	9.137
	VisualGLM	1.822	0.421	6.559	13.180	9.984
	ours	**6.723**	**1.552**	**9.845**	**17.014**	**12.609**
MMED	Livebot	0.054	0.004	3.657	9.189	7.558
	DialoGPT	0.163	0.018	5.206	10.440	8.021
	VisualGLM	0.620	0.120	6.794	12.384	9.053
	ours	**5.525**	**1.129**	**8.154**	**13.820**	**9.998**

4 EXPERIMENTAL SETUP

4.1 Datasets

We use two Chinese datasets to verify our method, here we introduce the detail of two datasets.

M3ED [37] is a Chinese dataset comprising diverse multimodal emotional dialogues, extracted from various TV series. In the M3ED dataset, a long individual utterance in one turn may be split into multiple sentences, and each sentence carry one of the seven emotion labels (happy, surprise, sad, disgust, anger, fear, neutral). To meet the requirements of dialogue tasks, we merge multiple sentences from the same individual, determining the predominant emotion among them as the label for the combined utterance. The dataset is then split into training set, validation set, and test sets in an $8:1:1$ ratio.

MMED is a multimodal emotion dataset that we autonomously collect. The dialogues within it are also sourced from various TV series. The dataset format and preprocessing are consistent with M3ED.

We summarize key statistical data for two datasets in Table 1.

4.2 Baselines

To better compare model capabilities, we introduce three models as baselines, including:

- **DialoGPT** [28] is an extension of the GPT-2 model designed for generative tasks. We opt for the Chinese version of the DialoGPT model[3] and conduct training exclusively using dialogue text from two datasets.
- **Livebot** [20] is a model designed for generating live comments. It utilizes a unified Transformer model to capture the dependency information between comments and videos. The model consists of three parts: a video encoder, a text encoder, and a comment decoder.
- **VisualGLM** is a multimodal large language model [4], built upon ChatGLM-6b [4], and pre-trained on a large-scale dataset containing both Chinese and English text-image pairs. Since VisualGLM takes images as input, we randomly sample a frame from the input video as model input. We freeze the parameters of the video encoder and employ p-tuning's fine-tuning [19] on VisualGLM.

[3]https://github.com/yangjianxin1/GPT2-chitchat
[4]https://github.com/THUDM/VisualGLM-6B

Table 3: Ablation experiments results on M3ED

Model	BLEU-4	METEOR	ROUGE-L
ours	**1.552**	**9.845**	**12.609**
w/o ERM	1.161	8.788	12.272
w/o ERT	1.085	8.514	12.314
w/o SSC	1.342	9.183	12.477
w/o V	1.378	9.324	12.563
w/o REP	1.139	9.841	12.331

4.3 Implementation Details

We opt for ChatGLM2-6B as our language model and initialize our visual encoder with the weights from VisualGLM's[4] ViT and QFormer. Additionally, we utilize the m3e-base model[5] as the text embedding model for our retrieval module. During the training process of the Emotional Retrieval Module, we set the self-separation coefficient to 4, the learning rate to 5e-5, and the batch size to 32. In contrastive learning loss, the temperature coefficient is set to 0.5. In the stage of response emotion prediction, we set the batch size to 16 and the learning rate to 2e-5. For the stage of emotion-enhanced response generation, we continue with a batch size of 16 but increase the learning rate to 2e-2. All training optimizers use AdamW. We conduct training on four A100 40G GPUS, and the total training time takes approximately 15 hours. In the text generation phase, we employ nucleus(top-p) sampling with a p-value of 0.7 and a temperature set to 0.95 and use the nltk [6] and chinese-rouge [7] to calculate our metrics.

4.4 Evaluations

In order to obtain accurate and comprehensive performance comparisons, we use both automatic and human evaluations.

Automatic Evaluation: In our experiments, We adopt BLEU-2 and BLEU-4 [27], METEOR [2], and ROUGE-1 and ROUGE-L [15] as the automatic evaluation metrics to evaluate the performance of our model.

Human Evaluation: In order to better compare with the three baseline methods, We randomly sample 100 examples from the test set for all models. Then we hire two well-educated volunteer

[5]https://huggingface.co/moka-ai/m3e-base
[6]https://github.com/nltk/nltk
[7]https://github.com/Isaac-JL-Chen/rouge_chinese

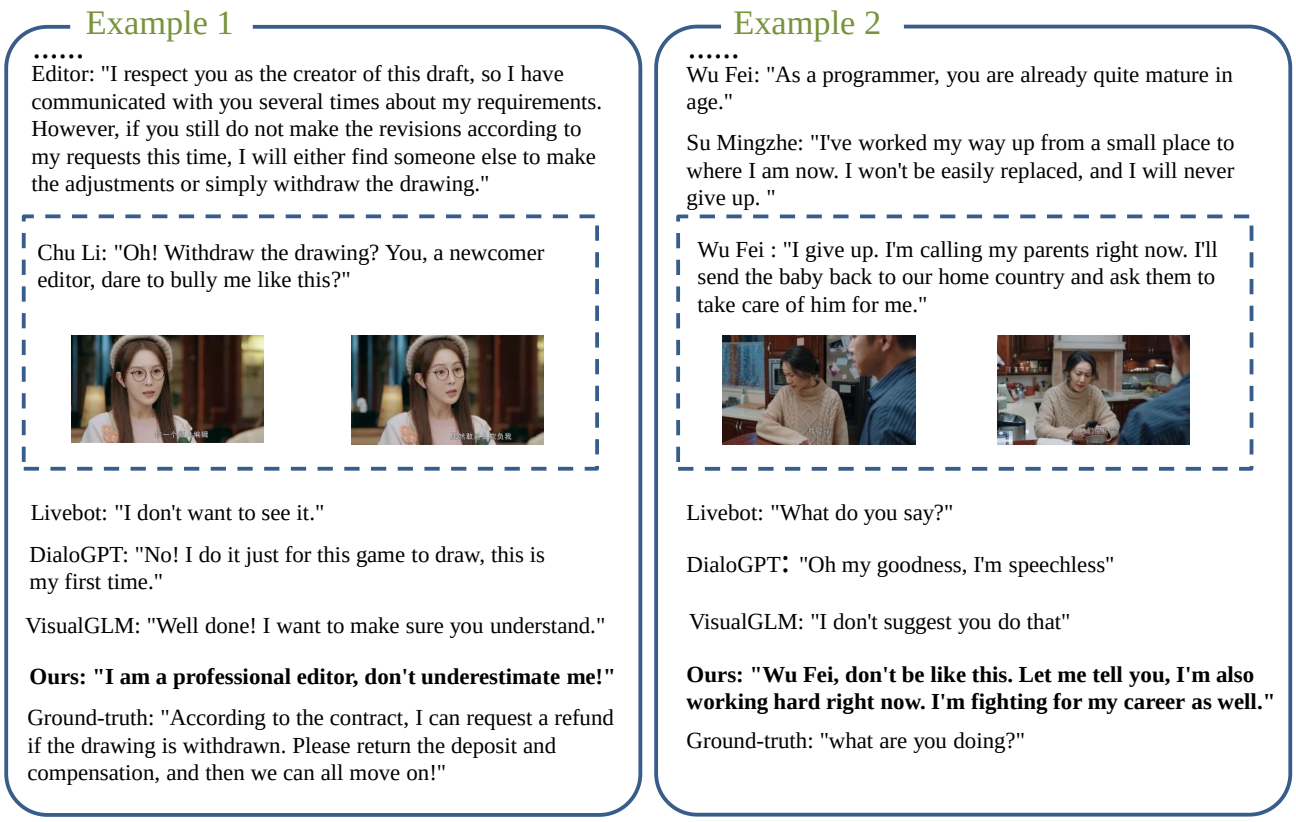

Figure 5: A case study. Example 1 occurred when Chu Li did not heed the editor's advice to make revisions to the artwork, leading to a dispute between them. In Example 2, the context is that Su Mingzhe, who worked as a programmer, lost his job, and Wu Fei is suggesting that he consider a career change.

Table 4: The result of human evaluation on two datasets.

	Model	R1	R2
M3ED	Livebot	1.545	1.56
	DiaoGPT	1.765	1.775
	VisualGLM	3.42	3.28
	ours	**4.13**	**4.055**
	Ground-Truth	4.84	4.875
MMED	Livebot	1.105	1.155
	DiaoGPT	1.57	1.74
	VisualGLM	3.105	2.935
	ours	**3.895**	**3.495**
	Ground-Truth	4.735	4.640

annotators to score each model based on two metrics: Contextual Relevance (R1) and Emotional Relevance (R2). R1 represents the degree of relevance between the generated response and the context, while R2 indicates the level of emotional alignment between the generated content and the current context. Scores for each metric range from 1 to 5. Additionally, we report Cohen's Kappa[8] to indicate the consistency among annotators.

[8]https://en.wikipedia.org/wiki/Cohen%27s_kappa

5 EXPERIMENTAL RESULTS

5.1 Automatic Evaluation Results

Table 2 presents the automatic evaluation results of all models on two datasets based on selected metrics. Our model surpasses all baselines across all metrics on both datasets. In comparison with DialoGPT without visual information, our model achieves superior performance. This improvement is likely attributed to the enhanced semantic and emotional depth provided by visual content, including character behavior and expression. Consequently, adept utilization of visual cues empowers our model to generate text with a more comprehensive understanding of the scene.

Unlike DialoGPT, both Livebot and VisualGLM-6B incorporate visual information as an additional input. However, a substantial performance gap persists between these models and our proposed method. Specifically, when compared with Livebot, the integration of a large language model affords our proposed model significant advantages in language generation, while the inclusion of visual information with appropriate alignment does not hinder its generation ability. As for VisualGLM-6B, our proposed method can extract richer event and emotional information from dialogue scenes through continuous frames, thereby enabling the generation of more refined responses.

5.2 Human Evaluation Results

The results of the human evaluation for both datasets are presented in Table 4. Cohen's Kappa was computed for the annotators on each dataset, yielding a coefficient of 0.41 for M3ED and 0.38 for MMED, indicative of a moderate level of agreement between the two annotators. Across both the M3ED and MMED datasets, VisualGLM and our model, leveraging LLM, demonstrate significant superiority over livebot and DialogGPT in the R1 and R2 metrics. This underscores the importance of integrating LLM in multimodal emotion dialogue tasks, given its robust text generation capabilities which lay a solid groundwork for seamless conversations.

Further comparison between VisualGLM and our model reveals a notable enhancement achieved by our model. This highlights the efficacy of our training framework in enriching emotional expression. The human evaluation corroborates these findings, demonstrating the superior performance of our model. Overall, the congruence between human and automatic evaluations reinforces the excellence of our model in generating emotionally resonant responses.

5.3 Ablation Study

To assess the influence of various components on our model's performance, we performed an ablation study using the M3ED dataset, observing comparable patterns as those found in the MMED dataset. The outcomes of this study are detailed in Table 3.

Firstly, the removal of the Emotion Retrieval Module (ERM) led to a decrease in performance across all metrics, underscoring its critical role in providing emotional cues during the two-stage training and impacting the final outputs. Next, we eliminated the Emotion Representation Training component (ERT) and the Self-Separated Component (SSC) from the retrieval module sequentially. The subsequent decline in performance underlined the essential efficacy and necessity of these components. Notably, the exclusion of ERT resulted in a more significant drop in metrics than removing the ERM, emphasizing that relying purely on semantic representations for retrieval can introduce noise and underscores the importance of integrating emotional information through ERT to refine the representation. Further, the elimination of visual input (V) led to reduced performance, revealing the value of body language and facial expressions in videos as they convey additional emotional cues that are not present in text alone. Finally, the removal of the Response Emotion Prediction (REP) component caused a decline in performance across various metrics, confirming the efficacy of REP in distilling emotional content from abstract visual features and minimizing noise in the visual data during the second training stage.

5.4 Analysis of Self-separation Coefficient

In the Emotional Retrieval Module, the self-separation coefficient plays a crucial role in partitioning the embedding. To examine its influence on model performance more closely, we experimented with different self-separation coefficients and presented the effects in Figure 6. Examination reveals that the coefficient's increment causes fluctuations in BLEU-4 scores, while METEOR and ROUGE-L generally exhibit an upward trend. The findings suggest that a minimal self-separation coefficient hinders performance, whereas an excessively large coefficient impedes retrieval efficiency with minimal metric enhancement. Thus, selecting an optimal coefficient value is imperative, contingent upon the particular context.

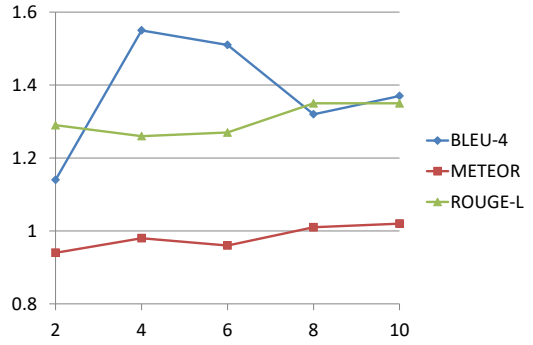

Figure 6: Experiments with different self-separation coefficient in Emotional Retrieval Module on M3ED. For ease of viewing, Mentor and ROUGE-L are divided by a factor of 10.

5.5 Case Study

To further substantiate the effectiveness of our model, we conducted a case study, and the results are presented in Table 5. In Example 1, when Chun Li expresses a condescending and threatening tone, our model responds assertively, expressing dissatisfaction with Chun Li's behavior. In Example 2, Wu Fei expresses anger and disappointment towards Su Mingzhe, who is unemployed and unwilling to change careers. This prompts Wu Fei to consider having their child return to their home country. Our model initially conveys surprise at Wu Fei's actions and then emphasizes the determination to pursue a career as a programmer. In comparison to the ambiguous responses from other models, our model generates responses with clear emotions and accurately expresses its intentions. Both examples illustrate how our model aligns seamlessly with the context in terms of both emotion and content.

6 CONCLUSION

In this work, we develop an LLM-Based Emotion-Guided Multimodal Dialogue Method. This approach maximizes the potent textual expression capabilities of LLM while leveraging the limited dataset of multimodal emotional dialogues. Specifically, we have engineered an emotional retrieval module and a response emotion prediction training process to discern latent emotional cues embedded within multimodal information. Simultaneously, Our approach incorporates six distinct input components, encompassing both emotional and natural expressions, for training the model. Notably, our method exhibits great performance across two real-world datasets, offering an effective solution to generate fluent and emotionally nuanced responses guided by emotional cues.

7 ACKNOWLEDGEMENTS

This work was supported in part by the grants from National Natural Science Foundation of China (No.62222213, U22B2059, 62072423)

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Yunkai Chen, Qimeng Wang, Shiwei Wu, Yan Gao, Tong Xu, and Yao Hu. 2024. TOMGPT: Reliable Text-Only Training Approach for Cost-Effective Multi-modal Large Language Model. *ACM Transactions on Knowledge Discovery from Data* (2024).

[4] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 320–335. https://doi.org/10.18653/v1/2022.acl-long.26

[5] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1555–1566.

[6] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

[7] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2470–2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

[8] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 154–164. https://doi.org/10.18653/v1/D19-1015

[9] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7042–7052. https://doi.org/10.18653/v1/2021.acl-long.547

[10] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 397–406. https://doi.org/10.18653/v1/N19-1037

[11] Dacher Keltner, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior* 43 (2019), 133–160.

[12] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework. arXiv:2309.11911 [cs.CL]

[13] Chen Liang, Jing Xu, Yangkun Lin, Chong Yang, and Yongliang Wang. 2022. S+PAGE: A Speaker and Position-Aware Graph Neural Network Model for Emotion Recognition in Conversation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 148–157. https://aclanthology.org/2022.aacl-main.12

[14] Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A Visual Experience Powered Conversational Agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5596–5611. https://doi.org/10.18653/v1/2021.acl-long.435

[15] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 605–612.

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.

[17] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[18] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *ACL*.

[19] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. *CoRR* abs/2110.07602 (2021). arXiv:2110.07602 https://arxiv.org/abs/2110.07602

[20] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. Livebot: Generating live video comments based on visual and textual contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6810–6817.

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

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424* (2023).

[23] Avinash Madasu, Mauajama Firdaus, and Asif Ekbal. 2023. A Unified Framework for Emotion Identification and Generation in Dialogues. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Elisa Bassignana, Matthias Lindemann, and Alban Petit (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 73–78. https://doi.org/10.18653/v1/2023.eacl-srw.7

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

[25] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6818–6825.

[26] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM international conference on multimedia*. 1098–1106.

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

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[29] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 5370–5381. https://doi.org/10.18653/v1/P19-1534

[30] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907* (2021).

[31] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3685–3695.

[32] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2854–2866. https://doi.org/10.18653/v1/2022.acl-long.204

[33] Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing Personalized Dialogue Generation with Contrastive Latent Variables: Combining Sparse and Dense Persona. *arXiv preprint arXiv:2305.11482* (2023).

[34] He sicheng Wang Yuxin, Sun Qingxuan. 2023. *M3E: Moka Massive Mixed Embedding Model.*

[35] Zheyong Xie, Weidong He, Tong Xu, Shiwei Wu, Chen Zhu, Ping Yang, and Enhong Chen. 2023. Comprehending the Gossips: Meme Explanation in Time-Sync Video Comment via Multimodal Cues. *ACM Transactions on Asian and*

Low-Resource Language Information Processing 22, 8 (2023), 1–17.

[36] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549* (2023).

[37] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5699–5710. https://doi.org/10.18653/v1/2022.acl-long.391

[38] Sirui Zhao, Hongyu Jiang, Hanqing Tao, Rui Zha, Kun Zhang, Tong Xu, and Enhong Chen. 2023. PEDM: A Multi-Task Learning Model for Persona-Aware Emoji-Embedded Dialogue Generation. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 132 (feb 2023), 21 pages. https://doi.org/10.1145/3571819

[39] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. *arXiv preprint arXiv:2105.08316* (2021).

[40] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[41] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1128–1137. https://doi.org/10.18653/v1/P18-1104

[42] 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).