Context-Aware Emotion Classification: Using Image & Audio Captioning, and context-based utterance history

YoonSik Park HoSik Hwang
Hanyang University
yoons1595@hanyang.ac.kr audtla22@hanyang.ac.kr

Abstract

Emotion Recognition in Conversion (ERC) predicts the current speaker's emotions through utterance. However, a single utterance can be interpreted differently depending on the context. Recent studies use various models to solve this problem, but it is not easy to predict the speaker's emotions according to the situation. To alleviate this problem, we propose a method to predict emotion by providing LLM with an optimal caption for the speaking context. To achieve this, we generate information about visual and auditory contexts through captioning, and conduct history prediction by accumulating previous utterances relevant to each context. Furthermore, to overcome the limitations of non-deterministic LLM, ERC is performed through refining prediction that converts the output into target emotion. Through this, various modalities could be unified into a single modality called text, and through this, good predictive performance could be obtained without conflict between modalities. And zero-shot classification could be performed using the pretrain model without additional learning.

1 Introduction

Emotion Recognition in Conversation (ERC) predicts the current speaker's emotions through utterance. In order to accurately predict this emotion, various information such as the context of the vision/audio/current utterance is required. However, solving this problem is a very important task in ERC because visual information (video) and auditory information (audio) have modality different from utterance. To solve this problem, we proceed with emotion prediction based on the situation at the time of utterance using the MELD dataset with three modalities. The MELD dataset has a total of three modalities: text, video, and audio. Many previous studies have been conducted on text. Papers such as CKERC[1], InstructERC[2], TelME[3], and SPCL-CL-ERC[4] tried to perform emotion classification by focusing on the text of the MELD dataset. Recently, as a lot of advanced LLM have emerged, the process is focusing on emotion classification using LLM. There is a paper called HCAM[5] that utilizes all three modalities of the MELD dataset. It aims to achieve effective emotion classification through joint representation space using cross attention. In another paper[6], while using all three modalities, the face plays the most important role in predicting emotion, so they focused on this to create a good representation space by grasping the facial expression well. In previous studies, the performance was the best in papers using text. As mentioned earlier, the fundamental problem of the MELD dataset is that it is quite difficult to predict emotions with one sentence in text. This also occurs in other modalities, such as the fact that the speaker's face does not come out in the video or it is difficult to grasp who is the speaker, and in audio, voice files with a lot of noise that are not suitable for ERC can be exemplified. To solve this problem, we decided to use the following approach. Since the text-based approach performed the best, we decided to unify text, audio, and video into one modality called text. First, in the video, we tried to obtain information on the facial expression of a person corresponding to a speaker through captioning. In audio, we tried to obtain captioning related to the speaker's emotions. For example, we tried to obtain captioning related to emotions such as increasing the voice, explaining the crying sound,

and describing the worked up voice. Through this, we tried to obtain good emotion classification performance by unifying three modalities into one and creating one prompt. For captioning, in video, fine-tuned architecture was used using a vision encoder-decoder-based coco dataset, and since it was a burden on gpu to use the entire video, the last frame (image) of video was used. The reason for using the last frame is that the person's emotions often appear on the face at the end of the speech. In audio captioning, an EnCLAP[7] model with SOTA performance was used. And for the text, the maximum utterances that can be used in one dialog were included in the text data. In the InstructERC[2] paper, it was decided how much to refer to the previous sentence when trying to predict the emotion of a specific utterance in one dialog through a hyperparameter. However, since this approach through hyperparameter accumulates a fixed size of utterances, it is possible to reduce the prediction performance, including unnecessary sentences. We advanced this by creating an adaptive window size to effectively utilize accumulated previous utterance sets for each sentence, using the LLama3 model. The contributions of this work can be summarized as follows:

• By unifying modalities, we reduced the effort required to obtain a good representation space.

• Instead of using a fixed hyperparameter for window size, we dynamically selected the amount of previous context for each utterance, resulting in better prediction performance.

• Utilizing pre-trained models enabled zero-shot classification without additional training.

2 Related Work

2.1 Emotion Recognition in conversation

ERC is a highly influential field in the domain of emotion analysis. ERC can be broadly categorized into two tasks: text-based[1-4] and multimodal [5] methods. Text-based models have traditionally focused on the relationship between speakers and context, but recent trends favor the use of pretrained language models (LLM) for prediction, which currently exhibit the best performance. Research has also explored multimodal approaches aiming to achieve a robust representation space through cross attention. However, due to the superior performance observed with text-based methods, there has been relatively less research in multimodal approaches, and their performance tends to slightly lag behind that of text-based methods.

2.2 Image Captioning & Audio Captioning

Recent research trends involve leveraging various models based on transformer-based pretrained models to produce research outcomes. There is also active research on encoder-decoder architectures tailored for generation tasks such as captioning. Training through the alignment of language and vision is being utilized for downstream tasks like captioning. The captioning research on face expression[8] has not been studied very much in recent years except for a few studies. In audio captioning[7,9], various studies have been conducted by using a pretrain audio encoder and a language pretrain model as a decoder. Models such as gpt[10] and bart[11] are actively being developed as decoders.

3 Proposed Method

Our ovarall pipeline can be seen in Figure 1. First of all, we will discuss captioning and history prediction, the process of converting other modalities into the text based modality. Next, we will introduce the process of making a prompt for prediction by combining these. Finally, we will show the process of emotion prediction and refine prediction.

3.1 Multimodal to unimodal by captioning

Three different modalities of video, audio, and utterance were converted into one text based. For video and audio, captioning is used, and utterance proceeded with history prediction to create an optional accumulated utterance set for better prediction.

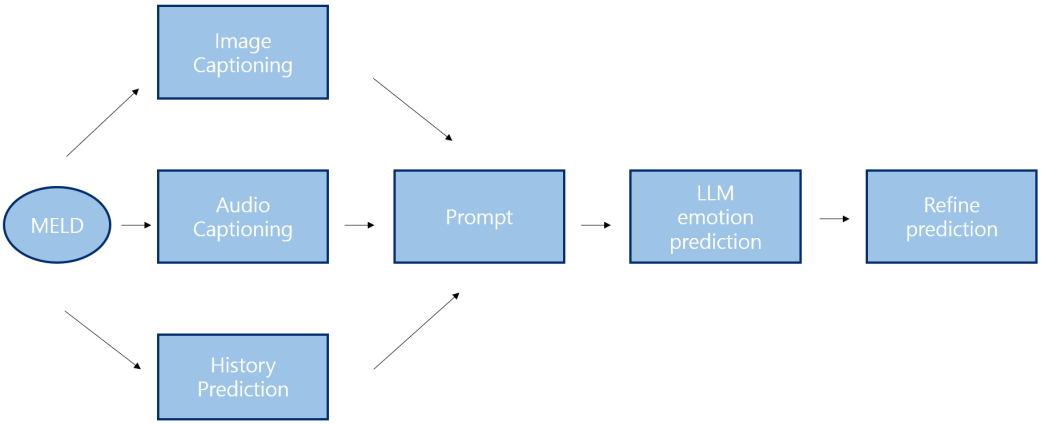

Figure 1: Pipeline

3.1.1 Video

The video dataset of the MELD consists of short images of about 3 to 10 seconds. What is needed for Emotion prediction is information about the speaker, especially detailed information related to facial expressions. This information mainly appeared at the end of the speech, which is the second half of the video. Therefore, we decided to extract the last frame of the video and caption this image. For this, a bit-swin-base-224-gpt2-image-captioning model was used. This model has a Swin transformer as an encoder and gpt2 as a decoder and is fine-tuned through the coco dataset. Since fine-tuned is performed through the coco dataset, it has the advantage of captioning information about the object well. The result of image captioning is as follows.

Figure 2: Image captioning - Last frame and image captioning result

3.1.2 Audio

The audio dataset of the MELD dataset similarly consists of short voice files of 3-10 seconds. En-CLAP[7], which is currently state-of-the-art in audio captioning, was used. Through this, information on the current situation in the voice environment could be converted into text. The result of audio captioning is as follows.

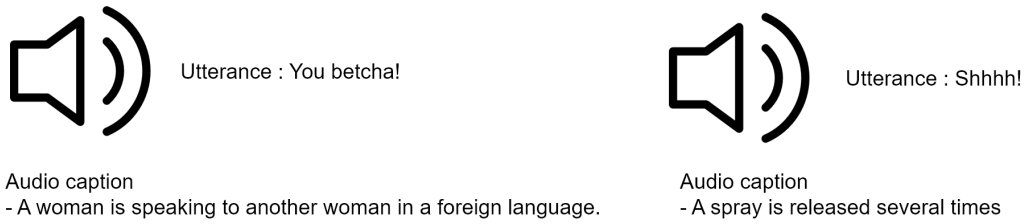

Figure 3: Audio captioning - utterance and audio captioning result

3.2 History Prediction

As mentioned earlier in the introduction, a single utterance does not contain the context of the conversation. In order to caption the context, reference to the previous conversation is required. The naive approach for reference to the previous conversation is to create a utterance set that accumulates all previous conversations. However, this method has the disadvantage of excessively increasing the length of the input for one prediction. If the length of the input is excessively increased, the number of tokens increases and unnecessary sentences may be included in the prediction. More specifically, There is a moment when the context is switched in one dialog, and conversations before that context are not very helpful information, not related to the main context for current prediction. Therefore, simply accumulating the previous conversation has the disadvantage of greatly increasing the length of the input without considering the context. In order to solve the previous naive approach, history prediction was conducted to predict the part related to the current utterance during the previous conversation using LLM. we used the llama-8B model, as it demonstrated the best performance among the models available in our GPU. As shown in the figure 4, we create an accumulated utterance set for all previous utterances in the same dialogue. In addition, we use llama3 to predict the most optimal starting point for emotion prediction of the final utterance within the accumulated set and make the optimal utterance set for prediction.

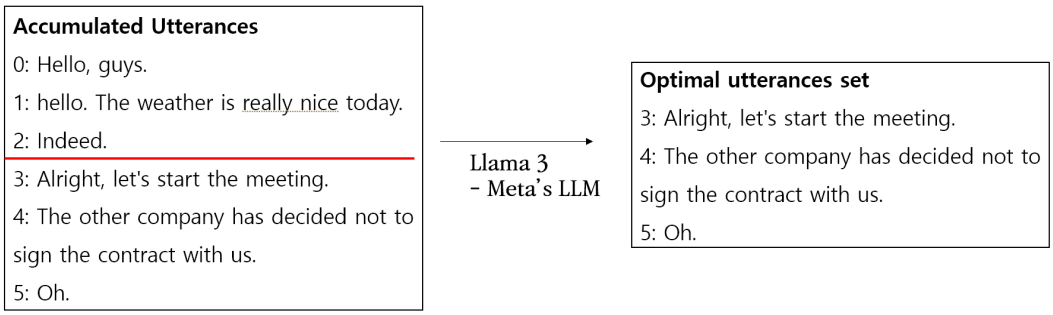

Figure 4: history prediction - Naive accumulated utterances to Optimal utterances set

3.3 LLM Emotion Prediction

Emotion prediction was conducted among eight classes. In the case of Prompts, it will be described in detail in case 1, and based on case 1, the prompt of other cases were modified and proceeded. The prompt of cases 2-8 can be found in appendix. 1) History prediction + Image caption + Audio caption 2) History prediction + Image caption 3) History prediction + audio caption 4) History prediction 5) Accumulated utterances with hyperparameter 6) Only Utterance 7) Only Utterance + simple prompt 8) Only utterances + simple prompt + prompt engineering First, we measured the performance of our pipeline, which includes history prediction, image and audio captioning. Next, we tried case 2

and 3 to determine which one has a more major influence between image and audio captioning. In addition, we examined the performance when using only history prediction. we create an utterance set through naive accumulation with a hyperparameter and evaluate the performance to determine whether history prediction truly enhances performance or not. Furthermore, we used only simple utterance and compared the performance with using a simple prompt without prompt engineering. Finally, in case 8, we compared the performance using prompt engineering to that of using a simple prompt.

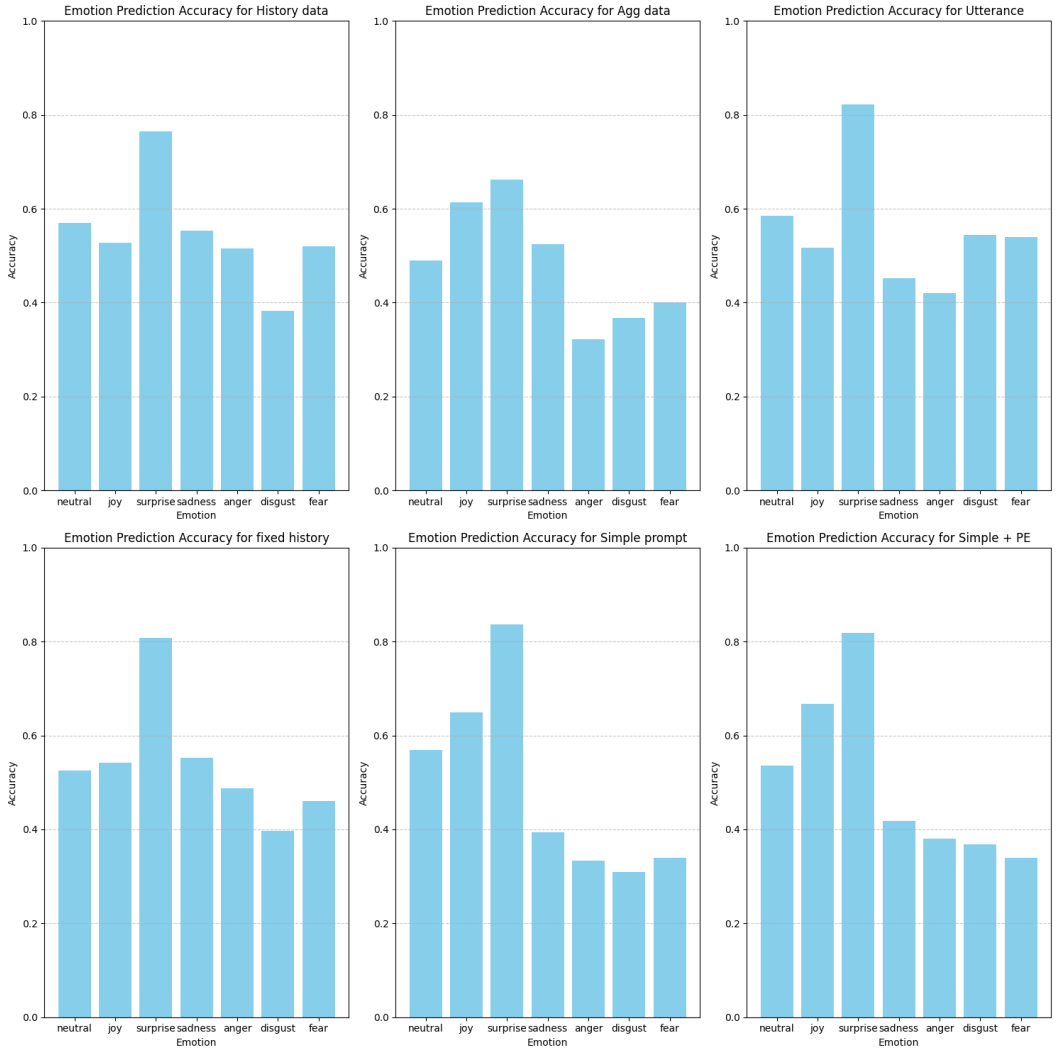

Figure 5: When accuracy was measured using GPT-4o, significant differences were observed in the prediction accuracy for each emotion. "Agg data" refers to the combined input of text, video, and audio captioning into the LLM. "Fixed history" denotes setting the history hyperparameter to 3. "Simple prompt" indicates a simplified version of a longer prompt, while "Simple prompt + PE" refers to applying slight prompt engineering to the simple prompt.

4 Results

We used the test set of the MELD dataset to check the performance of the emotion classification. In the last emotion prediction process, llama3 and gpt-4o were used. In the results of Llama3, 'history + captions' approach involves creating a single prompt by combining the optimal utterance set derived from history prediction with video and audio captioning, subsequently conducting emotion prediction. 'Utterance' predicts using only single utterance, and 'history' also predicts using the

utterance set through history prediction. 'History + audio' is the combination of history and audio captioning, and 'history + video' is the combination of history and video captioning. Similarly, in gpt 4o result, 'History + Captions', 'Utterance', and 'History' proceeded in the same way as described above. Additionally, 'History (prev3)' was conducted by accumulating the previous three utterances for emotion prediction. Moreover, simple prompt reduced the existing prompt by about 30%, consequently reduced the token length. Finally, simple prompt + PE is a prompt engineering that gives role to simple prompt.

1) In emotion prediction, it was confirmed that the performance of gpt-4o was better in all cases than llama3. 2) Both llama3 and gpt-4o performed best when only utterance and history were used. Through this, it was judged that the background information of audio and video caption did not give meaningful information to emotion classification. 3) Through the results of history prediction and history (prev3), it was confirmed that considering the context in history prediction, rather than simply accumulating a fixed number of utterances, led to improved performance. 4) From the results of simple utterance and simple utterance + PE, it was confirmed that prompt engineering did not have a significant effect. This is likely because prompt engineering is less impactful in simple emotion classification tasks compared to tasks that require extracting long sentences using LLM. 5) As can be seen from Figure 6, the ratio of the output converted through the refine output process was large when the prompt was long (history + captions). It was judged that this was the result of increasing input token. The overall performance summarized in a table can be found in Table 1.

prediction	
<emotion: neutral>	754
<emotion: joy>	544
<emotion: surprise>	483
<emotion: sadness>	261
<emotion: anger>	158
<emotion: disgust>	71
<emotion: fear>	70
<emotion: frustration>	45
<emotion : neutral>	39
<emotion : joy>	24
<emotion : surprise>	23
<emotion: curiosity>	21
<emotion: confusion>	13
<emotion: annoyance>	12
<emotion: concern>	11
<emotion : anger>	7
<emotion : fear>	6
<emotion: sarcasm>	6
<emotion: determination>	6
<emotion: encouragement>	6
<emotion: embarrassment>	6
<emotion : sadness>	5
<emotion : disgust>	3
<emotion: hope>	3
...	
<emotion: neutrality>	1
<emotion: frustration	1
<emotion: neutrality	1
<emotion: self-deprecating humor>	1

(a) History + captions

prediction	
<emotion: neutral>	852
<emotion: surprise>	618
<emotion: joy>	344
<emotion: sadness>	256
<emotion: anger>	214
<emotion: fear>	82
<emotion: disgust>	68
<emotion: frustration>	40
<emotion : neutral>	22
<emotion : surprise>	19
<emotion : joy>	11
<emotion: concern>	9
<emotion: curiosity>	8
<emotion: embarrassment>	6
<emotion: encouragement>	5
<emotion : fear>	5
<emotion: confusion>	4
<emotion: urgency>	4
<emotion: sarcasm>	4
<emotion: annoyance>	4
<emotion: hope>	3
<emotion : anger>	3
<emotion: determination>	3
<emotion: relief>	2
...	
<emotion: envy>	1
<emotion: guilt>	1
<emotion : annoyance>	1
<emotion : concern>	1

(b) Simple prompt

Figure 6: The relationship between misprediction and input length
(a) history + captions (b) simple prompt

Table 1: Main results & Performance Comparison

LLama3	Performance (%)					
	History + Captions	Utterance	History	History + Audio	History + Video	-
	30.11%	32.87%	32.49%	29.77%	29.89%	
GPT-4o	History + Captions	Utterance	History	History (prev 3)	Simple Prompt	Simple Prompt + PE*
	51.99%	56.69%	**56.97%**	55.06%	55.4%	54.83%

5 Discussion

Our approach addressed the problem by converting multimodal data into unimodal text. Using an adaptive historical window, referred to as history prediction, we accumulated previous effective sentences for utterance emotion, creating an optimal set. We unified the modalities of audio and video through captioning, and ultimately unified text, video, and audio into a single modality. Using LLM prompt engineering, we proposed a method to obtain classification results. Secondly, using pretrained models enabled us to achieve robustness against noise in audio and video. Training data is characterized by various noise sources, as indicated in Fig 7, with an imbalance in class distribution. In addition, since it is difficult to distinguish who is the speaker in the video, there is a problem related to the matching of the speaker and the person on the current screen. In audio, various voices (for example, the audience's cheers and the actor's utterances) are mixed, so it is difficult to focus on who, and this effect could be reduced to some extent through the pretrain model.

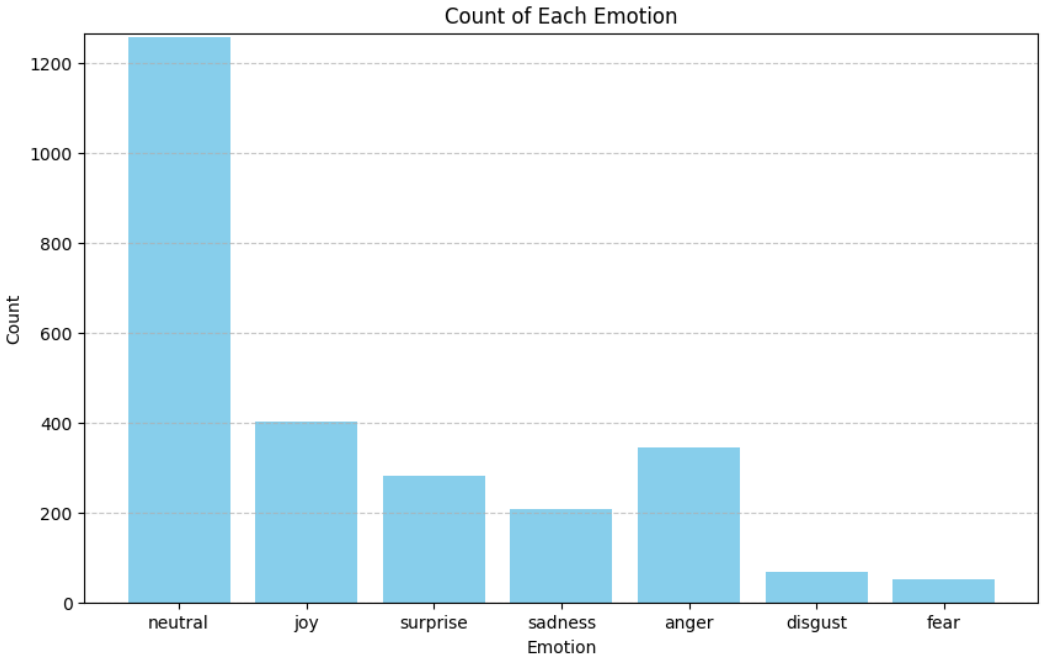

Figure 7: Label distribution of MELD test dataset

6 Conclusion and Future Work

Our research achieved good performance by converting multimodal data to unimodal text through captioning, thereby compensating for the poor utilization of vision and audio data. By using the pre-trained model, good classification performance was achieved through LLM without additional training. Furthermore, we proposed adaptive historical window which consider whole context by LLM (i.e. history prediction). However, there is still significant room for improvement. First of all, audio captioning and image captioning have not been able to generate meaningful captions that are helpful for prediction. Image described information about the background situation (e.g., two men standing in the kitchen), which is quite irrelevant to emotion classification. In audio captioning, extracting captions such as audience laughter and other background noises was insufficient for achieving good prediction performance. As an improvement, better performance could be achieved if video captions focus on the speaker's face and describe the emotion, or if audio captions describe the speaker's voice (e.g., tone, accent). Second, the performance may vary depending on how prompts are given to the LLM. The ablation study will be required to find a better prompt. Third, fine-tuning the LLM model for emotion prediction could reliably constrain the output format and enhance predictive performance in ERC. Fourth, due to GPU limitations, we used a lightweight model such as the LlaMA3-8B to extract history. If we utilize a more powerful LLM, such as GPT-4o, we can achieve

better history prediction, thereby ensuring improved ERC performance. Finally, due to GPU capacity, captioning for the video modality was performed using only the last frame of the image. However, if facial expression captioning is performed using the entire video, there will be room for performance improvement. In summary, Our approach is meaningful in that it aims to achieve better performance by unifying modality to text, thereby eliminating the difficulty of forming a joint representation space for multimodal data. In addition, ERC was conducted by creating an optional accumulated output set to predict the emotions of the utterances through history prediction. Furthermore, it is expected that stronger performance will be achieved through the various approaches mentioned in future works.

7 Appendix

7.1 Prompt for history prediction

```
1.↵
You are a great psychologist who finds the best starting point for the
conversation to predict the feelings of the last utter speaker in the
following conversation.↵
↵
2. Utterance format is as follows. ex) {Speaker's name}: {Utterance}↵
3. Each Utterance is separated by $$$. for example, {Speaker's name1}:
{Utterance1}$$${Speaker's name2}: {Utterance2}$$${Speaker's name3} :
{Utterance3}. Since there have been two $$$ before, the third utterance is
Utterance3. If there have been three $$$ Utterances before, the fourth
utterance.↵
4. What I give you is all the previous conversations. You need to find the
optimal starting point for the last speaker's emotion prediction.↵
5. Your output is the most appropriate starting point, and you only need to
print that number. For example, if you thought the i-th output was
appropriate, your output must be only number(I   ).'''},↵
```

Figure 8: Prompt for history prediction: 1. Role Assignment 2. Description of Utterance Format 3. Description of Accumulated Utterances 4. Detailed Explanation for History Prediction 5. Output Format Constraints

7.2 LLM emotion prediction

History prediction + Image caption + Audio caption Generate optimal utterances based on image captions, audio captions, and historical predictions. Use prompt to predict the emotions. **History Prediction + Image Caption** Emotion prediction was conducted using accumulated utterances from history prediction and image captions. **History Prediction + Audio Caption** Emotion prediction was conducted using history prediction and audio captions. **History Prediction** Emotion prediction was conducted using history prediction alone. **Utterance** Emotion prediction was conducted using a single utterance. **Accumulated Utterances with Hyperparameter** Without considering context through history prediction, prediction was conducted by accumulating the preceding w utterances using a hyperparameter w. **Single Utterance** Emotion prediction was conducted using a single utterance. **Single Utterance + Simple Prompt** Prediction was conducted using a simplified prompt reduced by 32% compared to the previous single utterance approach. **Single Utterance + Simple Prompt + PE** Prediction was conducted using the simplified prompt with the addition of a role.

7.3 Refine prediction

The MELD dataset requires predicting one of seven emotions. Therefore, the prompt was constrained to specify the emotion type and output format. However, due to the non-deterministic nature of the LLM, there were instances of misprediction where these constraints were not adhered to. We addressed these mispredictions as follows. First, we converted both the mispredicted emotion and the

```
'''you are a psychologist with great ability to predict the feelings of the
last utterance of the following drama conversation. ↵
↵
    1. I will give you the drama dialogue and 2 helpful situation information.
In other words, I will let you know 3 things. ↵
    a) The dialogue (utterances)↵
    b) Description of the visual information (frame) of the drama↵
    c) Description of the audio information (audio) of the drama↵
↵
    2. Using this information, you must judge the emotion of the last
utterance of the conversation.↵
↵
    3. In the conversation, each utterance is separated by $$$. You just must
predict the emotion of the last utterance.↵
    For example, {Speaker's name1}: {Utterance1}$$${Speaker's
name2}:{Utterance2}$$${Speaker's name3}:{Utterance3}. => You must predict the
emotion of utterance3.↵
↵

    4. The last utterance is speaking with one of the following seven
emotions. Never predict another emotion except for these seven sentiments. ↵
    <neutral, anger, surprise, joy, disgust, sadness, fear>↵
    You must predict the last utter speaker's emotion among seven emotions.↵
↵
    5. Your output format should be next. ↵
    <emotion : {your_prediction}>↵
↵
    You can't talk about the basis of judgment, the situation, only to keep
that form all the time.↵
↵
     For example, if you judged that the emotion of the last utterance was a
specific sentiment1, your output should be <emotion: sentiment1}>. '''↵

↵

f"(a) the dialogue: {row['Accumulated_Utterance_by_prediction']} \n (b)
Description of the visual information : {row['image caption']} \n (c)
Description of the audio information :  {row['audio caption']↵
```

Figure 9: Example of emotion prediction

```
4. The last utterance is speaking with one of the following seven emotions.
Never predict another emotion except for these seven sentiments. ↵
    <neutral, anger, surprise, joy, disgust, sadness, fear>↵
    You must predict the last utter speaker's emotion among seven emotions.↵
↵
    5. Your output format should be next. ↵
    <emotion: {your_prediction}>↵
↵
    You can't talk about the basis of judgment, the situation, only to keep
that form all the time.↵
```

Figure 10: Example of prompt when misprediction occurs

target emotion into vectors using spaCy. Then, we calculated the cosine similarity between these vectors. Finally, we converted the mispredicted emotion to the target emotion with the highest cosine similarity value.

References

[1] Fu, Y. CKERC: Joint Large Language Models with Commonsense Knowledge for Emotion Recognition in Conversation. arXiv preprint arXiv:2403.07260. (2024)

[2] Lei, Shanglin, et al. "Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework." arXiv preprint arXiv:2309.11911 (2023).

[3] Yun, Taeyang, et al. "TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation." arXiv preprint arXiv:2401.12987 (2024).

[4] Song, Xiaohui, et al. "Supervised prototypical contrastive learning for emotion recognition in conversation." arXiv preprint arXiv:2210.08713 (2022).

[5] Dutta, Soumya, and Sriram Ganapathy. "HCAM–Hierarchical Cross Attention Model for Multimodal Emotion Recognition." arXiv preprint arXiv:2304.06910 (2023).

[6] Halawa, Marah, et al. "Multi-Task Multi-Modal Self-Supervised Learning for Facial Expression Recognition." arXiv preprint arXiv:2404.10904 (2024).

[7] Kim, Jaeyeon, et al. "EnCLAP: Combining Neural Audio Codec and Audio-Text Joint Embedding for Automated Audio Captioning." arXiv preprint arXiv:2401.17690 (2024).

[8] Mohamad Nezami, Omid, et al. "Face-cap: Image captioning using facial expression analysis." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. Springer International Publishing, 2019.

[9] Shin, Wooseok, et al. "Rethinking Transfer and Auxiliary Learning for Improving Audio Captioning Transformer." Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Vol. 2023. 2023.

[10] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[11] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).