Supplementary Material: Leveraging CLIP Encoder for Multimodal Emotion Recognition (Appendix)

Yehun Song
Agency for Defense Development
kidswave@add.re.kr

Sunyoung Cho
Sookmyung Women's University
sycho22@sookmyung.ac.kr

In this supplementary material, we present additional experimental results in Sec. 1 to show the performance of different query settings for sentiment and emotion recognition. Examples of qualitative results are provided in Sec. 2, and a pseudo-code for the inference process of our MER-CLIP is provided in Sec. 3.

1. Additional results based on different emotion queries

In our main paper, we use the term '*Emotion*' for the emotion query in the emotion recognition task and '*Sentiment*' for the sentiment analysis task, selecting terminology that aligns with the respective task names. To demonstrate the diverse range of words associated with emotions, we conduct experiments to compare the performance of synonyms for '*Emotion*' and '*Sentiment*'.

Table 1 and Table 2 present comparisons of the emotion recognition and the sentiment analysis tasks, respectively, using various terms for emotion queries: '*Emotion*', '*Sentiment*', '*Feeling*', '*Impression*', '*Mood*', and '*Sensation*'. The term '*(no word)*' indicates the exclusive use of learnable prompts for the emotion query. The results in Table 1 indicate that '*Emotion*' achieves the highest micro-F1 score, '*Feeling*' achieves the highest recall score, and '*Sensation*' achieves the highest accuracy/precision scores. Interestingly, '*(no word)*' shows the comparable results across all metrics. The results in Table 2 indicate that '*Impression*' and '*Mood* yield higher performance than other words, while the results in Table 3 show that '*Sensation*' achieves the highest performance. Note that '*(no word)*' also shows the comparable results across all metrics on both CMU-MOSEI and CMU-MOSI datasets. These results indicate that the semantic information conveyed by the emotion queries impacts the performance of MER-CLIP. We can also observe that using only learnable parameters (*e.g.*, '*(no word)*') without prior knowledge yields promising results, suggesting that '*(no word)*' can be applied to arbitrary tasks where determining the optimal query setting is challenging.

Table 1. Performance comparison of words for emotion query on multimodal emotion recognition on CMU-MOSEI dataset.

Words	Accuracy(%)	Precision(%)	Recall(%)	Micro-F1(%)
Emotion	49.3	53.1	63.4	**57.8**
Sentiment	49.3	55.1	59.5	57.2
Feeling	49.3	52.3	**64.3**	57.7
Impression	48.9	54.7	59.0	56.8
Mood	49.2	53.3	61.9	57.3
Sensation	**49.6**	**55.5**	59.6	57.5
(no word)	49.1	52.7	62.8	57.3

Table 2. Performance comparison of words for emotion query on multimodal sentiment analysis on CMU-MOSEI dataset.

Words	ACC_2(%)	F1(%)
Emotion	85.2	85.1
Sentiment	85.3	85.1
Feeling	85.4	85.2
Impression	**85.5**	**85.5**
Mood	**85.5**	85.3
Sensation	85.0	84.8
(no word)	85.0	85.0

Table 3. Performance comparison of words for emotion query on multimodal sentiment analysis on CMU-MOSI dataset.

Words	ACC_2(%)	F1(%)
Emotion	85.1	85.0
Sentiment	84.0	84.0
Feeling	83.5	83.4
Impression	84.5	84.3
Mood	83.9	83.7
Sensation	**85.7**	**85.5**
(no word)	84.0	84.0

2. Qualitative results

We show qualitative results for three examples, each from CMU-MOSEI and CMU-MOSI datasets, respectively.

Fig. 1 shows results on CMU-MOSEI dataset. The first row is the file ID, followed by three rows of input modality data. The fifth row indicates the ground-truths for emotion recognition and sentiment analysis, and the last row shows the prediction results. In the prediction results, probabilities exceeding the threshold (0.6) for emotion recognition are highlighted in bold to demonstrate the multi-label prediction capability, while higher scores for sentiment analysis are also highlighted in bold. The prediction results of sentiment analysis are processed with the softmax function to clarify the results. We can observe that our method accurately predicts labels for both emotion recognition and sentiment analysis.

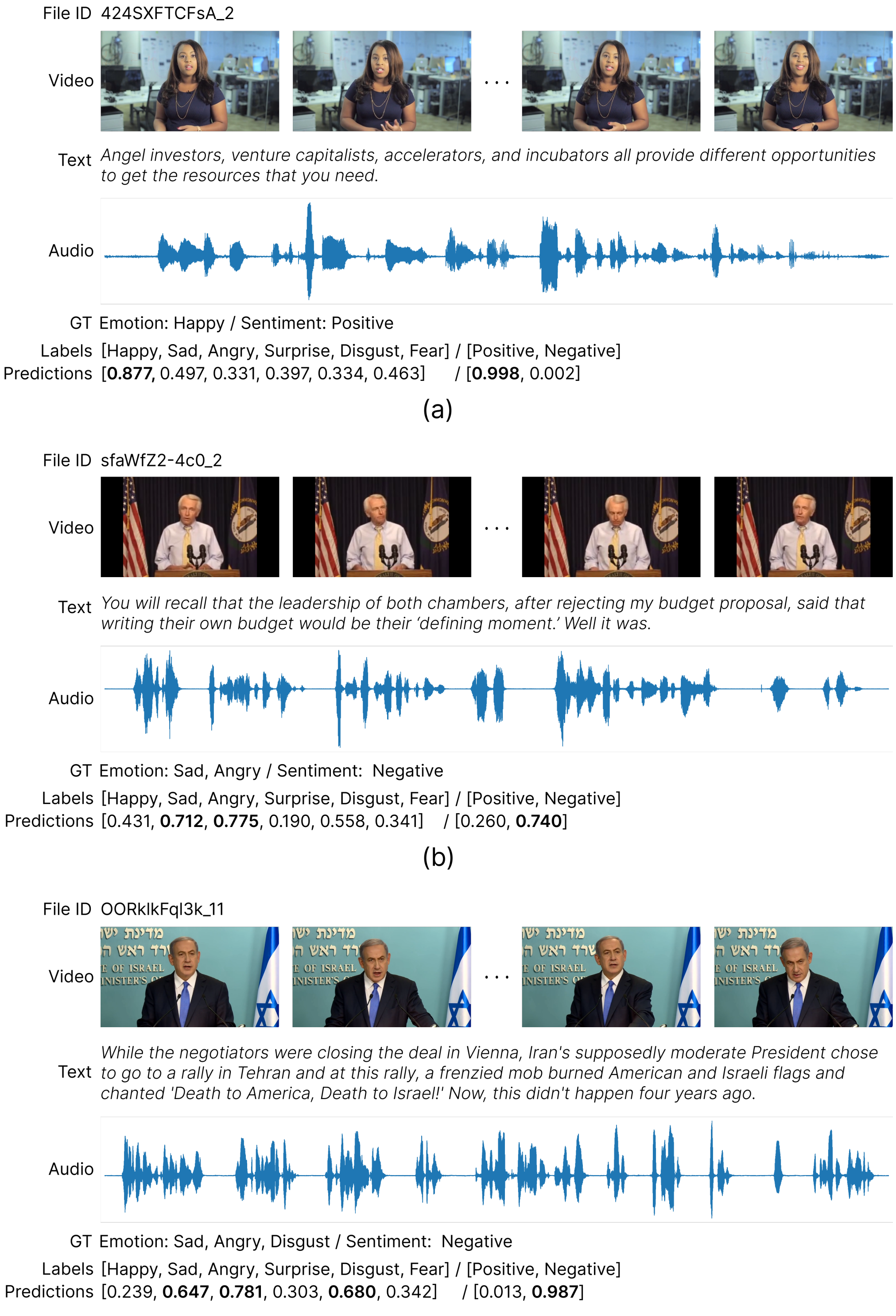
Fig. 2 shows results on CMU-MOSI dataset. Since CMU-MOSI has only sentiment labels, we denoted the results of its predictions on sentiment analysis. The higher score is highlighted in bold, and we can also observe that our MER-CLIP accurately predicts sentiment labels for all three examples in the CMU-MOSI dataset.

3. Algorithms

Algorithm 1 presents the pseudo-code for the inference procedure of our emotion recognition and sentiment analysis tasks.

Algorithm 1 Inference Process for Emotion Recognition and Sentiment Analysis

```
Require: Visual feature $X^V$, Audio feature $X^A$, Language feature $X^L$, Emotion query embedding $Z^E$, Label embedding $Z^L$
Ensure: Predicted class label $\hat{Y}$
Put $X^L$, $X^V$, $X^A$ in a list $M = [X^L, X^V, X^A]$ as the predetermined LVA order.
Put $M$ as key, value, and $Z^E$ as query in CMD and get the final output $Z^{[3]}$.
After processing CMD, get cosine similarity $sim(\cdot) = Z^{[3]} \cdot Z^L$.
if emotion recognition then
    Apply standard normalization and sigmoid function to $sim(\cdot)$.
    Transform logits to a vector by converting values greater than the threshold (0.6) to 1 and all others to 0.
else if sentiment analysis then
    Multiply $sim(\cdot)$ with the learnable logit scale initialized with $exp(log(1/0.07))$.
    Transform logits into a vector by setting the class with the larger logit to 1 and the other class to 0.
end if
return $\hat{Y}$ {Return the predicted class label}
```


File ID 424SXFTCFsA_2
Video
Text *Angel investors, venture capitalists, accelerators, and incubators all provide different opportunities to get the resources that you need.*
Audio
GT Emotion: Happy / Sentiment: Positive
Labels [Happy, Sad, Angry, Surprise, Disgust, Fear] / [Positive, Negative]
Predictions [**0.877**, 0.497, 0.331, 0.397, 0.334, 0.463] / [**0.998**, 0.002]

(a)

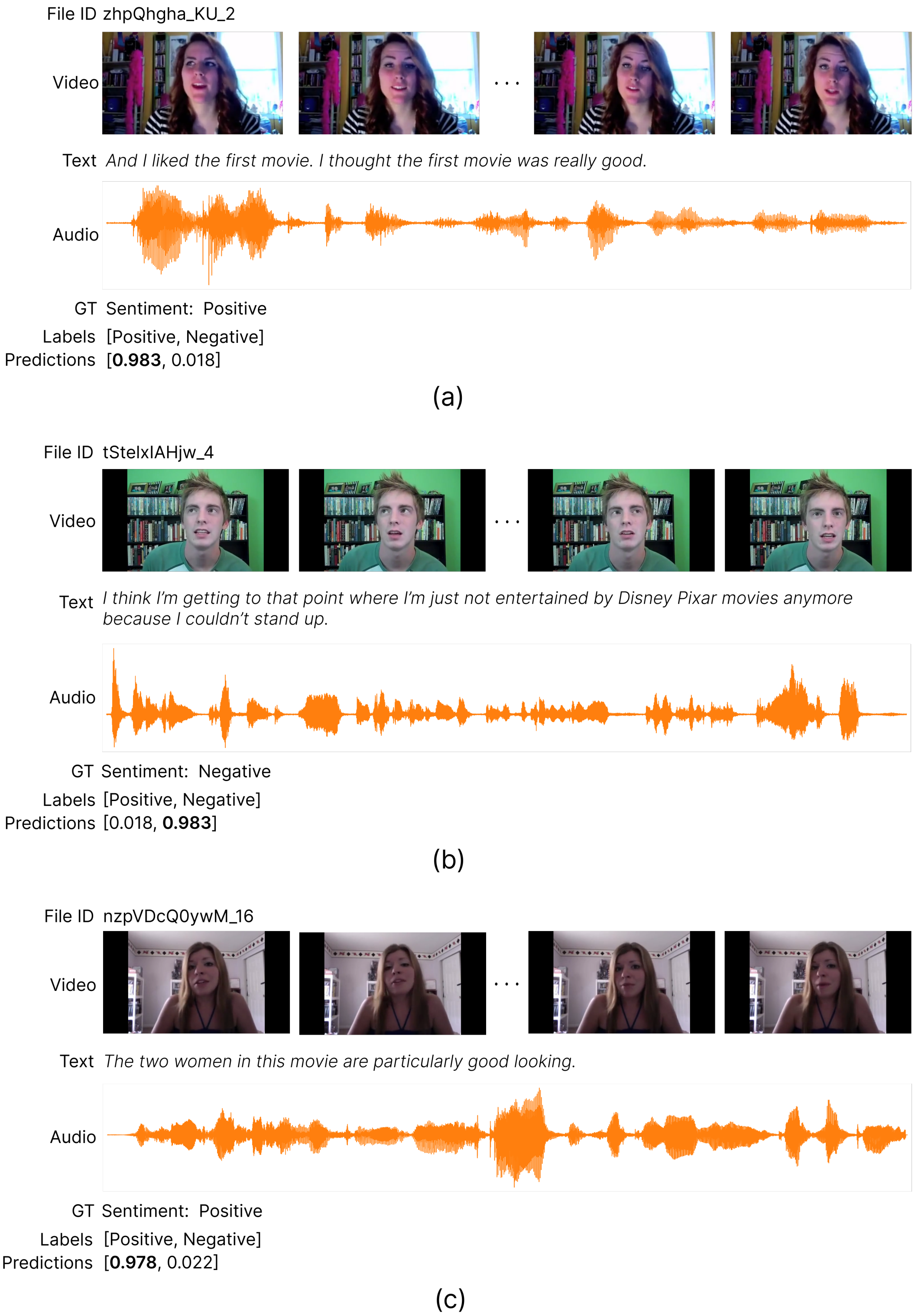
File ID sfaWfZ2-4c0_2
Video
Text *You will recall that the leadership of both chambers, after rejecting my budget proposal, said that writing their own budget would be their 'defining moment.' Well it was.*
Audio
GT Emotion: Sad, Angry / Sentiment: Negative
Labels [Happy, Sad, Angry, Surprise, Disgust, Fear] / [Positive, Negative]
Predictions [0.431, **0.712**, **0.775**, 0.190, 0.558, 0.341] / [0.260, **0.740**]

(b)

File ID OORklkFql3k_11
Video
Text *While the negotiators were closing the deal in Vienna, Iran's supposedly moderate President chose to go to a rally in Tehran and at this rally, a frenzied mob burned American and Israeli flags and chanted 'Death to America, Death to Israel!' Now, this didn't happen four years ago.*
Audio
GT Emotion: Sad, Angry, Disgust / Sentiment: Negative
Labels [Happy, Sad, Angry, Surprise, Disgust, Fear] / [Positive, Negative]
Predictions [0.239, **0.647**, **0.781**, 0.303, **0.680**, 0.342] / [0.013, **0.987**]

(c)

Figure 1. Qualitative results on CMU-MOSEI dataset.



File ID zhpQhgha_KU_2

Text *And I liked the first movie. I thought the first movie was really good.*

GT Sentiment: Positive

Labels [Positive, Negative]

Predictions [**0.983**, 0.018]

(a)

File ID tStelxIAHjw_4

Text *I think I'm getting to that point where I'm just not entertained by Disney Pixar movies anymore because I couldn't stand up.*

GT Sentiment: Negative

Labels [Positive, Negative]

Predictions [0.018, **0.983**]

(b)

File ID nzpVDcQ0ywM_16

Text *The two women in this movie are particularly good looking.*

GT Sentiment: Positive

Labels [Positive, Negative]

Predictions [**0.978**, 0.022]

(c)

Figure 2. Qualitative results on CMU-MOSI dataset.