

Supplementary Materials of Observe before Generate: Emotion-Cause aware Video Caption for Multimodal Emotion Cause Generation in Conversations

Anonymous Authors

1 ANNOTATION SAMPLE

We store the text data of our ECGF dataset in JSON files. The attached file "ECGF_samples.json" contains annotation samples of 20 conversations, where 111 emotion utterances are annotated with abstractive causes.

```
{
  "Conversation_id": 2,
  "Utterances": [
    {
      "utterance_id": 1,
      "text": "I do not want to be single, okay? I just... I just... I just... I just wanna be married again!",
      "video": "dia2utt1.mp4",
      "speaker": "Ross",
      "emotion": "sadness",
      "cause": "Ross doesn't want to be single, he just wants to be married again."
    },
    {
      "utterance_id": 2,
      "text": "And I just want a million dollars!",
      "video": "dia2utt2.mp4",
      "speaker": "Chandler",
      "emotion": "neutral"
    },
    {
      "utterance_id": 3,
      "text": "Rachel?!",
      "video": "dia2utt3.mp4",
      "speaker": "Monica",
      "emotion": "surprise",
      "cause": "Rachel shows up at the cafe, dressed like a bride."
    }
  ]
},
```

Figure 1: An annotated conversation in our ECGF dataset.

2 DETAILS OF OUR FRAMEWORK

Figure 4 illustrates the model architecture of our proposed framework ObG. Both ECCap and CGM are encoder-decoder models that take multimodal inputs and output text sequences. Detailed descriptions are provided in Section 5 of the main paper.

3 INSTRUCTION TEMPLATES FOR LLMs

To investigate the capability of Large Language Models (LLMs) in emotion cause generation, we utilize GPT-3.5 and Gemini under a few-shot setting to make inferences on the test set of our ECGF dataset, with the instruction templates shown in Figure 2 and Figure 3. Specifically, we concatenate the task prompt, three annotated samples as demonstrations, and the input sample to be tested together and feed them to the model, instructing it to generate the emotion cause for the given emotion in the input sample. GPT-3.5 and Gemini take only text input, to limit the input length, we retain only the conversation context within a window range of [-5,2], i.e., between the five utterances preceding and the two succeeding the

target emotion utterance (as clues to the cause typically appear near the emotion utterance [1, 2]). For Gemini-Pro, we add one keyframe from the video of each utterance to the input and use a smaller window range of [-3,0] to meet its maximum input limitation of 16 images. To conserve space, some context is omitted in the figures.

| |
|---|
| Task Prompt Given a conversation with multiple utterances, each including a speaker and the text, please write a sentence of no more than 40 words to describe what triggered the given emotion based on the context. |
| Annotation Demonstrations ... U8. Joey: Where do you think, Zelda? U9. Rachel: You found my book?! U10. Joey: Yeah I did! U11. Rachel: Joey, what... what are you doing going into my bedroom?! ... In U11, Rachel shows anger because: <i>Joey went into Rachel's room and found her book.</i> |
| Input Sample U1. You got the clothes clean. Now that is the important part. U2. Oh, I guess. Except everything looks like jammies now. U3. Whoa, I am sorry. Excuse me. We had this cart. In U3, Rachel shows anger because: |

Figure 2: The instruction template for GPT-3.5 / Gemini to generate emotion causes in a few-shot learning setting.

| |
|---|
| Task Prompt Given a conversation with multiple utterances, each including a speaker, the text, and a key frame from the video, please write a sentence of no more than 40 words to describe what triggered the given emotion based on the multimodal context. |
| Annotation Demonstrations  U1. Rachel: Barry? U2. Barry: C'mon in. U3. Rachel: Are you sure? ... In U3, Rachel shows surprise because: <i>Barry invites Rachel to come in while he has a patient.</i> |
| Input Sample  U3. Joey: What the matter? U4. Rachel: Nothing. U5. Joey: What is it? Hey! ... In U5, Joy shows sadness because: |

Figure 3: The instruction template for Gemini-Pro to generate emotion causes in a few-shot learning setting.

To highlight the benefits of the emotion-cause aware video captions generated by our trained model ECCap, we additionally leverage the image captioning capability of Gemini-Pro to generate plain

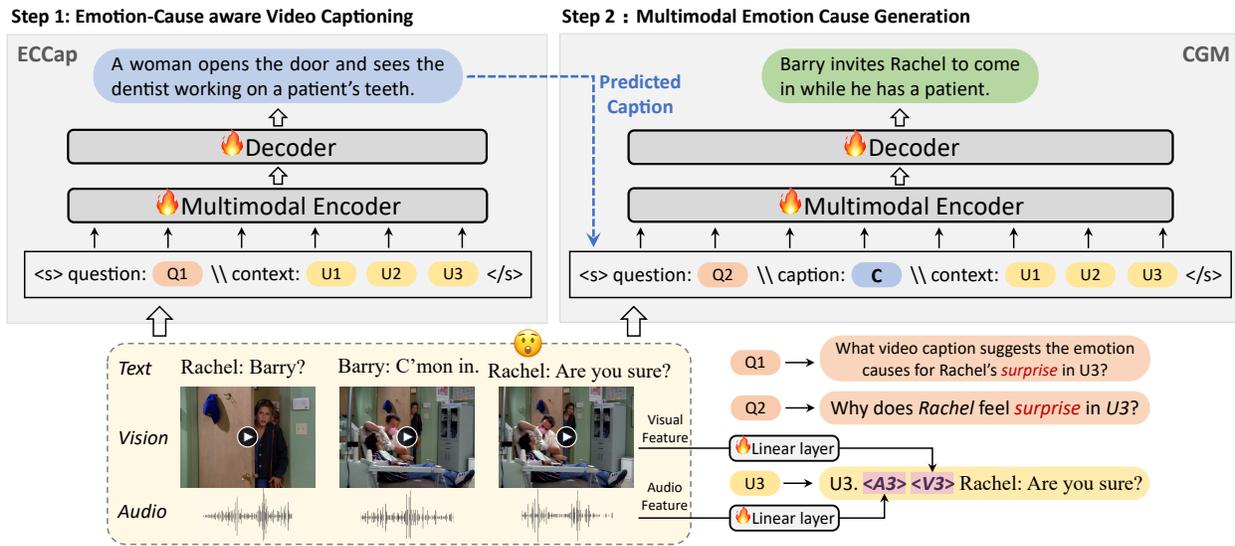


Figure 4: The model architecture of our framework ObG.

video captions and then use them to assist in emotion cause generation (the experimental results are presented in Table 5 of the main paper). Figure 5 illustrates the instruction template. For each sample, we input one keyframe from each utterance within the window range of [-3,0] surrounding the emotion utterance, i.e., four images in total, and instruct the model to output a video caption.

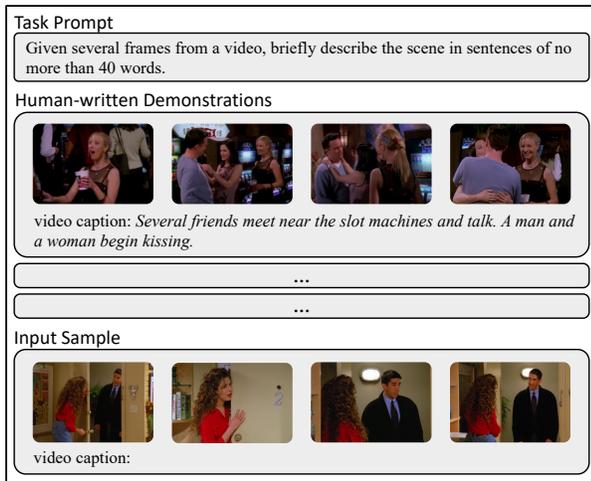


Figure 5: The instruction template for Gemini-Pro to generate plain captions in a few-shot learning setting.

As mentioned in Section 6.5 of the main paper, we also explore the effect of utterance-level captions (generating a video caption for each utterance, different from generating a conversation-level caption for the entire context), and conduct comparative experiments. Specifically, for each utterance within the window range of [-5, 2] around the emotion utterance, we input all the keyframes

(no more than 4 frames) from its video into Gemini-Pro to generate an emotion-cause aware video caption. The instruction template is shown in Figure 6. After training ECCap to obtain the utterance-level emotion-cause aware video captions, we concatenate each utterance with its caption and feed them into CGM for emotion cause generation, with the input template as follows: “question: Why does [Rachel] feel [surprise] in U[3]? context: U1. <A1> <V1> Rachel: Barry? [caption: ...] U2....”.

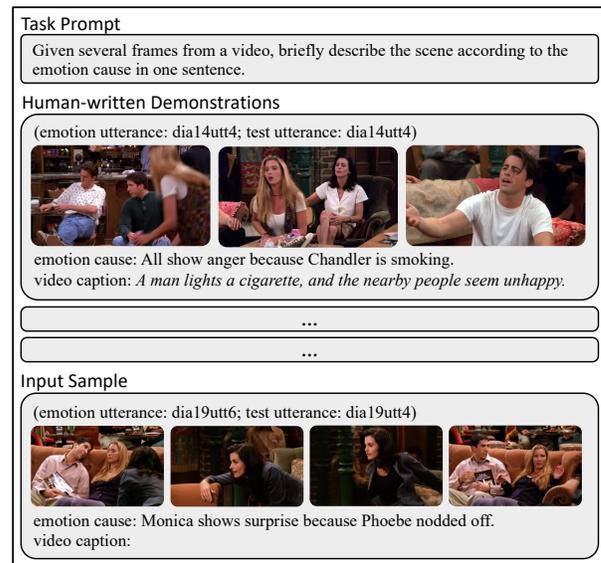


Figure 6: The instruction template for Gemini-Pro to generate utterance-level emotion-cause aware video captions in a few-shot learning setting.

REFERENCES

[1] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations.

[2] Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal Emotion-Cause Pair Extraction in Conversations. *IEEE Transactions on Affective Computing* 14, 3 (2023), 1832–1844.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348