Facial Affective Behavior Analysis with Instruction Tuning

Supplementary Material

This is the supplementary material of Facial Affective Behavior Analysis with Instruction Tuning.

A Limitations and negative effects

A.1 Limitations

This work also has its limitations. Firstly, we haven't tried other face-prior feature extractors except for landmark features. We leave this as an exploring research direction for future work. Secondly, our annotations from the training set also include noise induced by GPT-4V's hallucinations [15, 16] (see Appendix C.3), which may introduce some bias for models. We think our benchmark can be regarded as the learning from noisy descriptions task. Thirdly, the metric we proposed for evaluating the expression recognition ability may not reflect the nuanced expressions, which could be further improved.

A.2 Negative effects

There also exist some potential negative effects of our EmoLA. Firstly, *privacy issues*. Facial affective behavior analysis (FABA) may infringe on users' privacy, especially when deploying EmoLA on public spaces or systems. Individuals' faces may be unconsciously captured and recorded, leading to privacy concerns. Secondly, *misdirections and misjudgments*. Our EmoLA is not entirely accurate and may produce misjudged emotions. This can lead to misunderstanding or misdirection issues, especially in applications like security or judicial systems. Thirdly, *technological abuse*. Our method may be abused to suppress dissent or monitor political activities, thereby leading to social hazard or freedom restrictions.

B Background of instruction tuning

According to [18], instruction tuning refers to "the process of further training LLMs on a dataset consisting of (instruction, output) pairs in a supervised fashion, which bridges the gap between the next-word prediction objective of LLMs and the users' objective of having LLMs adhere to human instructions." The LLMs are typically trained on large general language corpora, which may differ from the users' objectives. As a result, to make the LLMs follow users' instructions, instruction tuning is proposed. For instance, InstructGPT [13]/

ChatGPT [12], FLAN-T5 [4], OPT-IML [6] are tuned with instruction-following data to enable their counterparts GPT-3 [1], T5 [14], OPT [19] have better generalization and few-shot abilities. Inspired by the success of instruction tuning for LLMs, LLaVA [9] attempts to extend this technique to the multimodal space, by introducing an MLP connector to map the visual tokens to language token space. Following that, other methods [2, 3, 17, 22] also adopt this mechanism on multiple downstream tasks and achieve remarkable results.

C Annotation details

C.1 Instructions in FABA-Instruct

We tried different instructions in our FABA-Instruct datasets. Specifically, we adopt 100 carefully designed instructions for emotion and action unit (AU) recognition tasks, respectively. Some of these instructions for emotion and AU are delineated in Fig. 2a and Fig. 2b, respectively. As shown in these examples, these instructions are all with natural language format.

C.2 AU types in FABA-Instruct

As mentioned in the paper, we select 12 AUs of FABA-Instruct for evaluation, *i.e.*, AU1, AU2, AU4, AU5, AU6, AU10, AU12, AU17, AU24, AU25, AU26, AU43. Also, there exists in total of 19AUs in the training annotations, and the meaning of these AUs are given in Tab. 1.

C.3 The accuracy estimation of training annotations

There also exists some noise in the training set due to the hallucinations in GPT-4V. To estimate the label accuracy in training annotations, we randomly sample 200 samples from each task in FABA-Instruct, and manually re-annotate these samples. After that, we can roughly estimate the accuracy or F1 score of training annotations according to these manual annotations. Specifically, for the emotion task, we calculate accuracy by classifying the text into 7 classes, which has been introduced in the main content. For the AU task, we evaluate all the AUs using the F1 score.

For emotion annotations in FABA-Instruct, as shown in Tab. 2, the accuracy of training annotations is about 91%. For AU annotations, the average F1 is 76.1% for all the AUs (see Tab. 1). From the estimation results, it can be observed that although there are some noisy labels in both two tasks, the recognition performance of GPT-4V on two tasks is still high. Therefore, it's reasonable to use these annotations for further research. Our FABA-Bench can not only be utilized for FABA tasks but also be regarded as the learning from noisy annotations task. Some examples are be found in Fig. 1.

Table 1: The meaning of AUs in FABA-Instruct.

AUs	Meaning
AU1	inner brow raiser
AU2	outer brow raiser
AU4	brow lowerer
AU5	upper lid raiser
AU6	cheek raiser
AU7	lid tightener
AU9	nose wrinkler
AU10	upper lip raiser
AU12	lip corner puller
AU14	dimpler
AU15	lip corner depressor
AU17	chin raiser
AU20	lip stretcher
AU23	lip tightener
AU24	lip pressor
AU25	lips part
AU26	jaw drop
AU27	mouth stretch
AU43	eyes closed

D Experiments

D.1 Details about the traditional FABA datasets

We utilize three traditional AU datasets (BP4D, DISFA, GFT) and one traditional emotion datasets (RAF-DB) for better evaluating the performance of our EmoLA compared with the previous SOTA methods. Furthermore, we also sampled around 20,000 face images from AffectNet to construct our FABA-Instruct dataset.

Emotion datasets. *AffectNet* [11] is a large-scale in-the-wild emotion database with more than 1M face images crawled from internet. We randomly sampled 20,000 images from this database for dataset construction. RAF-DB [7] is an facial expression dataset with 29,672 in-the-wild face images, which is annotated by 40 annotators with 7 single-label and two-tab compond emotion categories. In this paper, we mainly compare with other methods on the single-label subset.

AU datasets. *BP4D* [20] is a spontaneous facial AU dataset with 328 videos from 41 subjects (23 females and 18 males). There are in total of 140,000 frames with 12 AUs (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24). *DISFA* [10] consists of 26 subjects (12 females and 14 males) and 130,000 frames with the AU intensities (from 0 to 5) annotations. Following [21], an AU with the intensity equal or greater than 2 is considered to be activated. 8 (1, 2, 4, 6, 9, 12, 25, 26) of the 12 AUs are utilized for evaluation. *GFT* [5] contains 96 participants from 32 three-person groups. 10 AUs (1, 2, 4, 6, 10, 12, 14, 15, 23, 24) are selected

GT: Angry

The person in the image appears to be engaging in a conversation or reacting to something; their eyes are open wide, and the eyebrows are raised, which might suggest **surprise or intense interest**. However, without more context, it's challenging to precisely pinpoint the emotion as facial cues can be interpreted differently depending on the situation.

GT: sadness

The individual in the image appears to have a somewhat furrowed brow and a tightness around the mouth, which might suggest feelings of **anger, frustration, or intensity**. There's also a possibility that the person is in the midst of speaking or reacting to something, as people's facial expressions can change rapidly depending on the context. Without additional context, it is hard to determine the exact emotion with certainty.

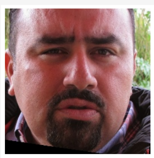
GT: Neutral

In the image, the facial expression appears to be a **mix of emotions**; it's not a clear-cut display of a single emotion. The person seems to be giving a somewhat intense gaze, which could be interpreted as **focused, determined, or possibly pensive**. The lips are slightly parted, which might suggest a moment of **contemplation or surprise**. However, without additional context, it's not definitive. Facial expressions can be complex and open to interpretation.

GT: Angry

The person in the image appears to be showing an emotion that could be interpreted as **shock, surprise, or fear**. This is suggested by the wide-open eyes and mouth agape, combined with the slightly raised eyebrows. Such facial expressions often indicate a strong emotional reaction to something unexpected or alarming.

(a) Examples of wrong emotion annotations.



GT: AU4, AU7, AU23

The facial action coding system (FACS) is a tool for categorizing human facial movements by their appearance on the face. Based on the image provided, I can identify several facial action units (AUs) that might be in use by the person in the photo:
1. AU4: Brow Lowerer - The person's eyebrows appear to be drawn together and slightly downwards, which could be indicative of this action unit.
2. **AU9: Nose Wrinkler** - There seems to be a slight wrinkling of the nose, suggesting this action unit might be in play.
3. **AU15: Lip Corner Depressor** - The corners of the mouth appear to be pulled downward, which is consistent with this AU.
4. AU23: Lip Tightener - The lips seem to be tight and pursed, which could hint at the activation of this action unit.
5. **AU24: Lip Pressor** - If the lips are indeed pressed tightly against each other, this action unit could also be in use.
These AUs may combine to convey emotions such as confusion, skepticism, or mild disapproval. Please note that deducing the exact emotional state or intent solely from facial cues without additional context might not be accurate, as facial expressions can be complex and subject to individual interpretation.

GT: AU1, AU5, AU25

In the image, the person seems to display a combination of facial expressions. The Facial Action Coding System (FACS) can be applied to describe the individual action units (AUs):
- AU1+AU2: Raised eyebrows, indicated by the visible lines across the forehead and the position of the eyebrows above the neutral position.
- **AU4: Brow Lowerer**, with the eyebrows drawn together, which often indicates confusion or concentration.
- AU5: Upper eyelid raise, though this is not very prominent here.
- **AU6: Cheek Raiser**, which is associated with the orbicularis oculi muscle contracting and might be part of a squint.
- **AU7: Lid Tightener** may also be present to some degree, suggested by the tension around the eyes.
- **AU12: Lip Corner Puller**, as indicated by the corners of the lips being drawn back slightly, usually associated with smiling or a positive emotion. However, the intensity seems low as the expression does not look like a strong smile.
- AU25+**AU26**: Lips part slightly, which could imply a speech beginning or a slight surprise.
Please note that without the complete context of the facial expression and additional movement data, this analysis is speculative based on static visual cues. Facial expressions are dynamic, and more accurate interpretations often require motion and context.

(b) Examples of wrong AU annotations.

Fig. 1: Wrong annotation examples on FABA-Instruct.

considering the challenges brought by head motion and occlusion. There are in total of 108000 training images and 24600 evaluating images in GFT.

D.2 Baseline details of FABA-Bench

As presented in the Experiment section, we reproduced four baselines which have similar arhitectures as our EmoLA on FABA-Bench, *i.e.*, MiniGPT4-V2 [2], mPLUG-Owl2 [17], Shikra [3], and LLaVA-1.5 [8]. LLaVA utilize a image encoder to obtain the visual tokens, and map the visual tokens to the language space through a linear layer. After that, the image tokens and language tokens are passed to an LLM decoder to generate descriptions. LLaVA undergoes a two-stage training process: initially, it exclusively trains the projector, followed

Table 2: Estimated accuracy of Emotion annotations in FABA-Instruct.

Total	Correct	Acc.
200	182	91.0

Table 3: Estimated F1 score of AU annotations in FABA-Instruct.

AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15
92.2	95.6	84.2	90.7	82.6	65.9	61.3	58.1	84.3	70.6
AU17	AU20	AU23	AU24	AU25	AU26	AU27	AU43	**Avg.**	
66.7	55.2	79.1	79.5	84.9	83.9	78.8	57.1	76.1	

Table 4: The multitask performance of EmoLA on FABA-Bench.

Methods	Emotion			AU		
	S_{re}	S_{ge}	S_{rege}	S_{re}	S_{ge}	S_{rege}
EmoLA (single task)	64.5	31.7	96.2	56.3	35.2	91.5
EmoLA (multi-task)	64.3	32.0	96.3	54.7	33.9	88.6

by a phase where only the LLM decoder is trained. LLaVA-1.5 enhances its performance by incorporating a two-layer MLP and adopting higher image resolutions. Shikra shares a similar architecture with LLaVA; however, it distinguishes itself by fine-tuning both the projector and the LLM decoder during its training phases. Similarly, MiniGPT4-V2, while architecturally similar to LLaVA, employs higher-resolution images to improve visual perception and aggregates every four neighboring visual tokens into a single token to optimize training efficiency. mPLUG-Owl2 introduced a visual abstractor module which aggregates the information from visual tokens by learnable queries. Furthermore, mPLUG-Owl2 leverarged a modality-adaptive module for facilitating multi-modal inputs to a shared semantic space for enabling modality collaboration.

D.3 Multi-task performance on FABA-Bench

We also perform experiments to evaluate the multi-task performance of EmoLA on our FABA-Bench. Specifically, we train our EmoLA using dataset from two tasks instead of the individual dataset. As shown in Tab. 4, the performance on Emotion of EmoLA under multi-task setting is almost the same to the single task setting. While for AU task, the performance of multi-task EmoLA drops com-

Table 5: The location of prior token.

Methods	Emotion S_{re}	Emotion S_{ge}	Emotion S_{rege}	AU S_{re}	AU S_{ge}	AU S_{rege}
EmoLA (prior token before visual tokens)	63.5	32.1	95.6	55.4	34.3	89.7
EmoLA (prior token after visual tokens)	64.5	31.7	96.2	56.3	35.2	91.5

pared to the single task version. We assume it's because the emotion recognition task will somehow affect the AU recognition performance by making model focus more on the general emotion recognition. Moreover, the multi-task is harder than the single task, which may also decrease the performance on AU task.

D.4 Location of prior token

We also investigate the location of prior token in EmoLA in Tab. 5. From the results, we can observe that if we put the prior token before the visual tokens the recognition performance on two tasks will decrease. We assume this will affect the visual tokens' representation due to the causal mask in the decoder. As a result, we put the prior token after the visual tokens.

E More generation cases

E.1 Successful cases

We present some successful cases of EmoLA's prediction on our FABA-Instruct dataset in Fig. 3 and Fig. 4, respectively.

E.2 Failed cases

We also present some failed cases of EmoLA's prediction on our FABA-Instruct dataset in Fig. 5a and Fig. 5b, respectively.

- "What is the emotion of this face?",
- "Tell me about the emotion on this face",
- "Label the emotion of this face?",
- "What emotions do you perceive in this face?",
- "How does the emotion on this face make you feel?",
- "Can you tell me the emotion displayed in this image?",
- "What emotional response does this face evoke?",
- "Can you determine the emotion portrayed in this image?",
- "Provide a detailed interpretation of the emotion on this face",
- "What can you say about the emotional expression in this image?",
- "Can you analyze the emotion depicted in this face?",
- "What emotions come to mind when you see this face?",
- "How do you perceive the emotional tone of this face?",
- "Can you provide insights into the emotions conveyed in this image?",
- "What emotions do you think this face is expressing?",
- "How does the emotion in this image resonate with you?",
- "Can you articulate the emotional context of this face?",
- "What is your interpretation of the emotions displayed by this person?",
- ....

(a) Examples of emotion instructions.

- "What is the action units of this face?"
- "List all the action units of this face."
- "What action units can be identified in this face?"
- "Please enumerate the action units observed in this particular face?"
- "Provide a breakdown of the action units present on this face?"
- "Identify the action units exhibited by this face."
- "List the action units that characterize this face."
- "Can you specify the action units found in this face?"
- "Enumerate the action units displayed by this face."
- "Describe the action units visible on this face."
- "What action units are evident in this face?"
- "List the action units observable in this face."
- "Give a detailed list of action units for this face."
- "What are the action units present in this face?"
- "List all the action units of this face."
- "Please enumerate the action units observed in this particular face?"
- "What specific facial movements correspond to action units of this image?"
- "Provide a detailed breakdown of the action units visible in this face."
- ....

(b) Examples of AU instructions.

Fig. 2: The instructions used in FABA-Instruct.

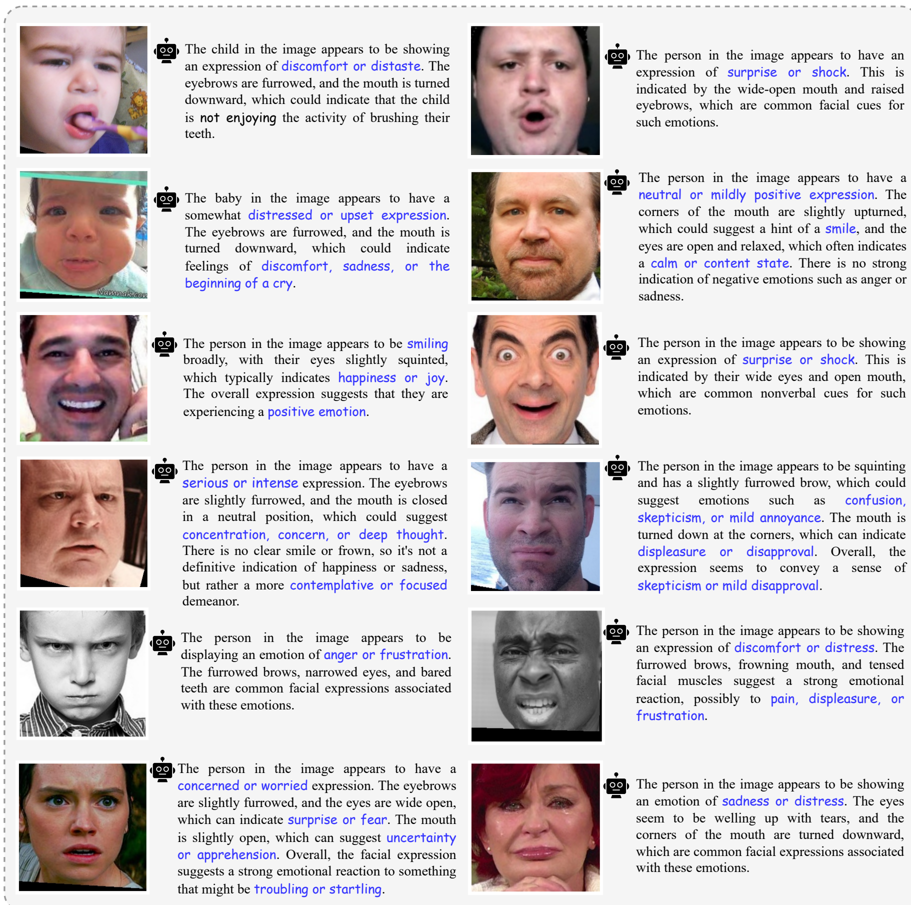


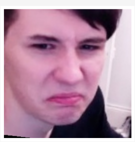
Fig. 3: Examples of EmoLA's successful prediction on emotion task.

The image shows a person with a facial expression that can be broken down into several Facial Action Coding System (FACS) Action Units (AUs). Here are the AUs that appear to be present:
1. AU4 (Brow Lowerer): There is a slight downward pull of the brows, which could indicate a frown or a concentration.
2. AU7 (Lid Tightener): The eyelids appear to be tightened, which can be associated with a squint or a focused gaze.
3. AU24 (Lip Pressor): The lips appear to be pressed together, which can be a sign of tension or determination.

The overall expression seems to convey a sense of seriousness or intensity, but without additional context, it's difficult to determine the exact emotional state or intent behind the expression.

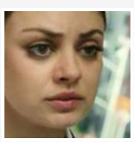
The image shows a person with a smiling expression. The Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face. Based on this system, we can describe the facial action units (AUs) present in this expression:
1. AU6 (Cheek Raiser): This is indicated by the appearance of crow's feet or wrinkles that form around the outer corners of the eyes, suggesting that the orbicularis oculi muscle is contracted to create a squinting effect.
2. AU12 (Lip Corner Puller): The corners of the lips are drawn back and up, which is characteristic of a smile. This action is primarily caused by the zygomatic major muscle.
3. AU25 (Lips Part): The lips are slightly parted, which can be a component of a relaxed or natural smile.

The combination of these AUs typically conveys a positive emotion, such as happiness or contentment.



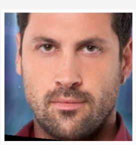
The image shows a person displaying a facial expression with several discernible features. Based on the Facial Action Coding System (FACS), we can identify several Action Units (AUs) that are active in this expression:
1. AU4: Brow Lowerer – The eyebrows are drawn together and downward, which is indicative of a frown or a look of concentration or anger.
2. AU9: Nose Wrinkler – There is a slight wrinkling of the nose, which can be associated with disgust or a negative reaction.
3. AU10: Upper Lip Raiser – The upper lip appears to be raised slightly, which can be part of an expression of disdain or contempt.
4. AU15: Lip Corner Depressor – The lip corners are pulled downwards, which can be a sign of sadness or displeasure.
5. AU17: Chin Raiser – The chin appears to be pushed upwards, which can be associated with a grimace.

The overall combination of these AUs suggests a facial expression that could be interpreted as one of disgust, disapproval, or displeasure. Facial expressions are complex and can be influenced by context, so this interpretation is based solely on the visible AUs without additional information.



Label: The image shows a person with a neutral to slightly negative facial expression. The Facial Action Coding System (FACS) can be used to describe the facial movements, which are called Action Units (AUs). Here are some possible AUs that might be present based on the visible features of the face:
- AU4: Brow Lowerer – There appears to be a slight downward pull of the brows, which could indicate a frown or a look of concentration or concern.
- AU7: Lid Tightener – The eyelids seem to be slightly tightened, which can be associated with a focused or intense gaze.
- AU24: Lip Pressor – The lips appear to be pressed together, which can be a sign of tension or determination.

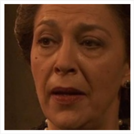
It's important to note that interpreting AUs can be subjective and context-dependent. The exact AUs and their intensity can vary depending on the individual and the situation.



The image shows a person with a neutral to slightly intense expression. The Facial Action Coding System (FACS) is a system to taxonomize human facial movements by their appearance on the face. Based on this image, here are some possible Action Units (AUs) that might be present:
- AU4: Brow Lowerer – There appears to be a slight downward pull of the brows, which could indicate concentration, confusion, or a slight frown.
- AU5: Upper Lid Raiser – The upper eyelids appear to be slightly raised, which can indicate alertness or widening of the eyes.
- AU7: Lid Tightener – The eyelids appear to be tightened, which can be associated with a focused or intense gaze.
- AU24: Lip Pressor – The lips appear to be pressed together, which can indicate tension or determination.

It's important to note that interpreting AUs from a static image without context can be challenging and may not fully capture the nuances of the person's expression. Facial expressions are complex and can be influenced by a wide range of factors, including individual differences, cultural expressions, and situational context.

Fig. 4: Examples of EmoLA's successful prediction on AU task.

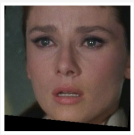


GT: Neutral

The person in the image appears to have a **somber or concerned** expression. The eyebrows are slightly furrowed, and the corners of the mouth are turned downward, which can indicate **sadness, worry, or contemplation**. The eyes seem to be focused intently, which can suggest **deep thought or concern**.

GT: Angry

The person in the image appears to have a **serious or intense** expression. The eyebrows are slightly furrowed, and the mouth is set in a straight line, which could suggest concentration, concern, or deep thought. There is no clear smile or frown, so it's not a definitive indication of happiness or sadness. The overall impression is one of **seriousness or contemplation**.



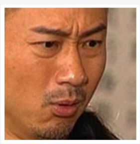
GT: Sadness

The person in the image appears to be showing an emotion of **concern or worry**. The eyebrows are slightly furrowed, and the eyes are wide, which can indicate a sense of anxiety or fear. The mouth is slightly open, which can suggest surprise or shock. Overall, the facial expression conveys a strong emotional response that could be associated with **stress or apprehension**.

GT: Happy

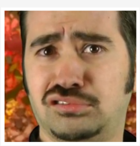
The person in the image appears to have a slightly open mouth and wide eyes, which could suggest **surprise, shock, or mild alarm**. The eyebrows are raised, which often indicates **surprise or confusion**. However, without additional context, it's difficult to determine the exact emotion with certainty.

(a) Examples of EmoLA's failed prediction on emotion task.



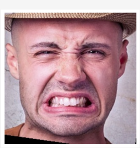
GT: AU4, AU9, AU17, AU23

The image shows a person displaying a facial expression with several distinct features. Based on the Facial Action Coding System (FACS), we can identify several Action Units (AUs) that are active in this expression:
1. AU4: Brow Lowerer – The eyebrows are drawn together and downward, which is indicative of a frown or a look of concentration or anger.
2. AU9: Nose Wrinkler – There is a slight wrinkling of the nose, which can be associated with disgust or a strong smell.
3. **AU10: Upper Lip Raiser** – The upper lip appears to be raised, which can be part of an expression of disdain or contempt.
4. AU17: Chin Raiser – The chin appears to be pushed upwards, which can be associated with a grimace or a display of anger or frustration.
5. AU23: Lip Tightener – The lips appear to be tightened or pressed together, which can be a sign of tension or determination.
The overall expression seems to convey a strong emotion, possibly anger, frustration, or disgust. Facial expressions are complex and can be interpreted in various ways depending on the context and other nonverbal cues.



GT: AU1, AU9, AU12, AU15, AU20, AU25

The image shows a person displaying a facial expression with several distinct features. Based on the Facial Action Coding System (FACS), we can identify several Action Units (AUs) that are active in this expression:
1. **AU4: Brow Lowerer** – The eyebrows are drawn together and downward, which is indicative of a frown or a look of concentration or concern.
2. AU9: Nose Wrinkler – There is a slight wrinkling of the nose, which can be associated with disgust or a reaction to an unpleasant smell.
3. **AU10: Upper Lip Raiser** – The upper lip appears to be raised slightly, which can be part of an expression of disgust or contempt.
4. AU12: Lip Corner Puller – The lip corners are pulled back, but this is not a typical smile; it could be part of a grimace or a grimacing smile.
5. AU15: Lip Corner Depressor – The lip corners are also slightly turned downward, which can indicate sadness or displeasure.
6. AU20: Lip Stretcher – The lips appear to be stretched horizontally, which can be associated with tension or grimacing.
7. AU25: Lips Part – The lips are parted, which can be a sign of surprise or the beginning of speaking.
The overall expression seems to convey a complex emotion, possibly a mix of disgust, concern, or discomfort. Facial expressions are complex and can represent a combination of different emotions, and the context in which the expression occurs is also important for accurate interpretation.



GT: AU4, AU9, AU10, AU23, AU25

The image shows a person displaying a facial expression with several distinct features. Based on the Facial Action Coding System (FACS), we can identify several Action Units (AUs) that are being activated:
1. AU4: Brow Lowerer – The eyebrows are drawn together and downward, which is indicative of a frown or a scowl.
2. AU9: Nose Wrinkler – There is a wrinkling of the nose, which can be associated with disgust or a strong negative emotion.
3. AU10: Upper Lip Raiser – The upper lip is raised, exposing the upper teeth, which can be part of an expression of anger or disgust.
4. **AU17: Chin Raiser** – The chin appears to be pushed upwards, which can be part of an expression of anger or determination.
5. AU23: Lip Tightener – The lips appear to be tightened and stretched horizontally, which can be associated with anger or grimacing.
6. AU25: Lips Part – The lips are parted, which can be a component of an expression of surprise or an intense emotion.
7. **AU26: Jaw Drop** – The jaw appears to be dropped, which can be part of an expression of surprise or shock, but in this context, it seems more related to the grimace.
The overall expression suggests a strong negative emotion, possibly anger, frustration, or disgust. Facial expressions are complex and can be interpreted in various ways depending on the context and the combination of different AUs.

(b) Examples of EmoLA's failed prediction on AU task.

Fig. 5: Failed cases of EmoLA on FABA-Instruct.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. vol. 33, pp. 1877–1901 (2020) 2
2. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023) 2, 4
3. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 2, 4
4. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 2
5. Girard, J.M., Chu, W.S., Jeni, L.A., Cohn, J.F.: Sayette group formation task (gft) spontaneous facial expression database. In: IEEE FG. pp. 581–588 (2017) 3
6. Iyer, S., Lin, X.V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P.S., et al.: Opt-iml: Scaling language model instruction meta-learning through the lens of generalization. arXiv preprint arXiv:2212.12017 (2022) 2
7. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593 (2017) 3
8. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 4
9. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Adv. Neural Inform. Process. Syst. **36** (2024) 2
10. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013) 3
11. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017) 3
12. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2**, 13 (2023) 2
13. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 27730–27744 (2022) 1
14. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR **21**(140), 1–67 (2020) 2
15. Tan, Z., Chen, T., Zhang, Z., Liu, H.: Sparsity-guided holistic explanation for llms with interpretable inference-time intervention. In: AAAI. pp. 21619–21627 (2024) 1
16. Tan, Z., Cheng, L., Wang, S., Bo, Y., Li, J., Liu, H.: Interpreting pretrained language models via concept bottlenecks. arXiv preprint arXiv:2311.05014 (2023) 1
17. Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023) 2, 4

18. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al.: Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023) 1
19. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022) 2
20. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. Ima. Vis. Comput. **32**(10), 692–706 (2014) 3
21. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: CVPR. pp. 3391–3399 (2016) 3
22. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2