

---

# EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language

---

Phoebe Chua<sup>1,2\*</sup> Cathy Mengying Fang<sup>1\*</sup> Takehiko Ohkawa<sup>3</sup> Raja Kushalnagar<sup>4</sup>  
Suranga Nanayakkara<sup>2</sup> Pattie Maes<sup>1</sup>

<sup>1</sup>MIT Media Lab   <sup>2</sup>National University of Singapore   <sup>3</sup>The University of Tokyo   <sup>4</sup>Gallaudet University

{phochua, catfang, pattie}@media.mit.edu  
{pchua, scn}@nus.edu.sg  
ohkawa-t@iis.u-tokyo.ac.jp  
raja.kushalnagar@gallaudet.edu

## Abstract

Unlike spoken languages where the use of prosodic features to convey emotion is well studied, indicators of emotion in sign language remain poorly understood, creating communication barriers in critical settings. Sign languages present unique challenges as facial expressions and hand movements simultaneously serve both grammatical and emotional functions. To address this gap, we introduce EmoSign, the first sign video dataset containing sentiment and emotion labels for 200 American Sign Language (ASL) videos. We also collect open-ended descriptions of emotion cues. Annotations were done by 3 Deaf ASL signers with professional interpretation experience. Alongside the annotations, we include baseline models for sentiment and emotion classification. This dataset not only addresses a critical gap in existing sign language research but also establishes a new benchmark for understanding model capabilities in multimodal emotion recognition for sign languages. The dataset is made available at <https://huggingface.co/datasets/catfang/emosign>.

## 1 Introduction

The emotional content of a speech comes not only from its linguistic content but also *how* it is spoken—from pitch and intonations to non-verbal expressions [12]. For sign language, however, emotional indicators are less understood and studied [15, 17, 30]. This ambiguity has practical negative consequences from misinterpretations of signers’ feelings to causing biases and prejudices in legal settings [42] and emergency departments [20]. The root of the ambiguity comes from the differences in the makeup of sign languages. Sign languages are comprised of five parameters: Handshape, Place of Articulation (where the sign is made), Movement (how the articulators move), Orientation (the hands’ relation towards the Place of Articulation), and Non-manual behaviors (what the body and face are doing). They function similarly to the cavities, articulators and features of spoken languages [45, 37]. The non-manual behaviors serve as linguistic (e.g., grammatical) markers in addition to conveying emotions. One category of such non-manual behaviors is facial expressions. For example, “puffing of one’s face” indicates intensity and “raised eyebrows” indicates a yes/no question [45]. At the same time, one’s facial expressions also convey emotions, which are applied on top of the linguistic meaning [30]. Other ways of expression emotions in sign languages include the hand movements, such as the tempo, rhythm, and size of the movements [39]. For example, the

---

\*These authors contributed equally to this work.

duration of sentences was shorter, the sign movements were more angular, and their movement paths shortened when expressing anger as compared to a neutral condition.

Emotion recognition is a common task in the machine learning community. Datasets and models have been constructed to recognize the sentiment and emotions based on text [19, 43], speech [33, 6], facial expressions [4], and a combinations of multiple input modalities [51, 7, 38]. Emerging multi-modal foundation models have demonstrated abilities to understand context of the videos [31]. Few have extended the capabilities of pre-trained models to understand affect via fine-tuning and instruction-tuning [24, 8] on new datasets and benchmarks [27]. Emotion recognition on sign languages, however, remain sparse. Sign language translation itself is already difficult. In addition, current approaches fail to recognize the complex linguistic and affective functions of facial expressions during signing.

To address this critical research gap, we introduce EmoSign, the first comprehensive dataset containing sentiment and emotion labels for American Sign Language (ASL) videos. The dataset includes 200 ASL video clips annotated by 3 Deaf ASL signers with professional interpretation experience, who provided: (1) Overall sentiment ratings on a 7-point scale, (2) Presence and intensity ratings for 10 distinct emotion categories (3) Detailed descriptions of specific emotion cues. Unlike existing sign language datasets that focus primarily on translation capabilities, EmoSign specifically targets the affective dimensions of signing. Our work contributes to both sign language accessibility and emotion recognition research by:

1. Providing the first dedicated dataset for studying emotional expression in ASL
2. Documenting descriptions of how emotions manifest through manual and non-manual components through the lens of native signers
3. Establishing baseline model performance for sentiment analysis and emotion classification tasks in sign language

The impact of our work extends beyond emotion recognition of sign languages. Studying how emotions can be recognized from a mix of non-manual and manual components allows us to study emotion differently. This also presents a new technical challenge to the AI/ML community with a challenging new benchmark for multimodal AI systems that tests their ability to distinguish between grammatical and affective functions of emotional cues in sign languages.

## 2 Related Work

### 2.1 Emotion recognition

Datasets for emotion recognition are typically annotated using either discrete or dimensional emotion frameworks. Discrete emotion labels are based on theories of basic emotion, which suggest that there are a limited number of emotion states (e.g., fear, happiness) associated with distinct expressions and physiological states [14]. In contrast, dimensional approaches quantify emotion along continuous axes, predominantly emotional arousal and valence [11]. Recent advances in large language models (LLMs) have expanded the scope of emotion recognition beyond the traditional paradigm of emotion label prediction. These models facilitate a more generative approach to emotion *understanding*, producing detailed, comprehensive descriptions of emotional states in natural language [27]. This shift has prompted the development of new datasets and metrics that accommodate rich natural language descriptions of emotions, allowing for greater nuance in emotion analysis.

Many approaches to recognizing human emotions have used facial expressions [23], speech or text sentiment. Increasingly, multimodal approaches that combine video, audio, text and image inputs are being explored as a way of improving model robustness in complex environments [26]. However, many existing multimodal models are still heavily reliant on [29], or biased towards [47], the language modality. To address this limitation, recent work has begun investigating novel sources of data that capture nonverbal social cues without relying on language, such as mime videos [25]. Sign language videos are also a rich source of expressive nonverbal data. They present unique challenges, but also opportunities, for multimodal emotion recognition. In hearing communities, many emotional expressions are universal in the sense that they can be reliably understood by people across many different cultures [10]. However, in sign languages, facial expressions and other non-manual components (such as mouth shapes and body language) often simultaneously serve grammatical

Dataset	Size	Signers	ASL Fluency	Source	Labels
YouTube-ASL [44]	984h	>2500	Mixed/unknown	Web	English captions
OpenASL [41]	288h	220	DHH and interpreters	Web	English captions
ASL STEM Wiki [49]	>300h	37	Interpreters	Lab	English sentences
ASLLRP [36]	2,651 utterances	19	ASL native signers		English and gloss captions, non-manual information
How2Sign [13]	79h	11	Interpreters	Lab	English captions
MS-ASL [22]	24h	222	Mixed/unknown	Web	English caption

Table 1: Overview of existing American Sign Language (ASL) datasets.

and emotional functions. As such, recognizing the emotion of sign videos is often challenging for non-signers [30].

## 2.2 Machine learning research on sign language

In the context of machine learning research, work on sign language has focused on sign language translation and production. Sign language translation (SLT) methods typically use either raw image data or skeletal representation of the signer’s pose as input. In terms of model architectures, Transformer-based architectures have been used for word-level sign language recognition based on 2D body pose sequence representations [5]. The translation capabilities of LLMs also appear to extend to sign languages. SignLLM [16] proposes a framework for transforming sign videos into language-like representations that can easily be passed to off-the-shelf LLMs. To do so, the sign videos are first processed into a sequences of discrete character-level sign tokens using a learned codebook. Then, the tokens are composed into word-level sign tokens to form a sentence. LLaVA-SLT [28] takes a slightly different approach. First, they pretrain an LLM using paired gloss-text data to learn the grammatical structure of sign language as well as its relationship with natural language. Next, they pretrain a visual encoder using about 400 hours of video-text pairs. Finally, a simple two-layer neural network is used to map the pretrained visual language embeddings into the LLM’s token embedding space and produce a translation. In parallel with SLT, research has also investigated sign language production (SLP), which has useful applications such as automatic sign language captioning. SLP typically involves converting text to gloss (a method of sign language transcription), mapping the gloss to pose, then rendering the pose into a video or avatar [16]. Despite significant advances in both SLT and SLP, challenges remain in developing systems that accurately capture and convey emotional nuance in sign language.

To the best of our knowledge, there does not exist a dataset that captures the emotion labels of sign languages. We hope to contribute to the expansion of current foundational models’ capabilities and set up a new benchmark that have specific societal impacts.

## 3 EmoSign: Multimodal Dataset of Sign Language with Emotion Labels

### 3.1 Dataset Collection and Pre-processing

Table 1 summarizes existing continuous signing ASL datasets. We excluded YouTube-ASL due to the uncertainty of the quality of the signing and captions in the dataset. For the remaining datasets, we sampled several hours of video from each and segmented the videos into sentence or utterance-long clips. We then used VADER [19] to calculate the text sentiment of each clip’s caption. Across the board, we found that a large majority of videos were associated with neutral or close-to-neutral text captions. Based on this preliminary review, we selected ASLLRP as the base dataset as it provides the most comprehensive and high-quality labels for each video and also contains videos with strong emotional intensity.

To address the limitations of existing ASL datasets, we collect additional emotion labels of sentiment, emotion and open-ended visual emotion cues using continuous signing videos from ASLLRP [36] as the base. The annotation process is costly due to the manual effort involved in annotating multiple emotion labels for each video, as well as the difficulty of recruiting ASL-native signers for the task. To manage costs, we selected a sample of 200 utterances from ASLLRP for annotation.

We first preselected a subset of ASLLRP’s continuous signing videos based on a manual inspection of the text content for emotional expressiveness. The videos were then segmented into utterances, which typically consist of a single sentence or phrase, based on the utterance start and end frames

label	alpha	label	alpha	label	alpha
Sentiment	0.738	surprise_neg	0.119	disgust	0.166
joy	0.699	worry	0.555	frustration	0.330
excited	0.552	sadness	0.333	anger	0.370
surprise_pos	0.381	fear	0.351	<b>average</b>	<b>0.593</b>

Table 2: Krippendorff’s Alpha scores of each label of the Dataset on a scale of -1 to 1, where 1 indicates unanimous agreement and -1 indicating systematic disagreement.

provided in the dataset. As before, we used VADER to calculate the text sentiment of each utterance. Despite the pre-selection for emotionally expressive videos, a large proportion of utterances were still associated with neutral or close-to-neutral text captions. To achieve a range of labels in the dataset, we selected 100 utterances with the most positive and most negative VADER scores for inclusion in the final sample.

### 3.2 Annotation Construction

This study was reviewed and approved by our Institutional Review Board, protocol number NUS-IRB-2024-1081. We recruited 3 ASL native signers through a third-party vendor to annotate the selected video clips (for more details about annotator recruitment, see Appendix A.1). Prior to beginning the annotations, the signers attended a training session with the researchers. The goal of the session was to walk through the annotation process, align on the expected responses for each annotation task, and clarify and questions they may have regarding the tasks. The annotation process took roughly x hours per individuals. The annotations were collected via Qualtrics (See Appendix ?? for details about the annotation interface). The interface was refined through pilot tests with individuals whose first language is ASL (excluded from the final team of annotators).

For each video, the signers completed three annotation tasks in the following order: (1) sentiment analysis, (2) emotion classification, and (3) free response description of attributions of emotions. For the general sentiment, the annotators were asked to determine the overall sentiment of the video from strongly negative to strongly positive on a linear scale from  $-3$  to  $+3$  [51]. For emotion classification, the annotators were asked to determine the level of presence of 10 emotions from “not present at all” to “extremely present” on a scale of 0 to 3. The set of emotions are “joy”, “excited”, “surprise (positive)”, “surprise (negative)”, “worry”, “sadness”, “fear”, “disgust”, “frustration”, and “anger”. The labels build on Ekman’s basic emotions, [14] and the circumplex model of affect [40], and was also informed by prior work on emotion datasets that expand on basic emotion categories for richer annotations [32].

After the first two tasks, the annotators were also asked to rate their level of confidence regarding their scores on a scale of 0-100 (0: not confident at all; 100-extremely confident). Finally, the annotators were asked to describe specific cues that led them to identify the emotions they chose. Example descriptions were given about the speed and scale of movement, head and body movement, facial expressions, and signs that were emphasized. These guided questions were derived based on prior literature and interview results [9] as well as our pilot tests. Note that we allow the annotators to skip any videos that they did not wish to annotate because of content or the quality of the videos.

### 3.3 Dataset Post-processing

Each label of the clip was labeled by minimally 1, maximally 3 annotators, given certain clips were skipped. For each label, we used the “majority vote” approach to find the most popular rating. In the case of a tie, we used the annotators’ self-reported confidence score and selected the label from the most confident annotation. Table 2 shows the Krippendorff’s alpha scores used to measure the agreement between the annotators of each label. Overall, the average score is 0.593, within the emotion categories, positive emotion labels had higher inter-annotator agreement than negative emotion labels.

### 3.4 Final Dataset Analysis

The final dataset includes 200 utterances with an average length of 4.8s per utterance and a total of about 16 minutes of video. Figure 1 shows the distribution of the duration of clips in the stimuli set.

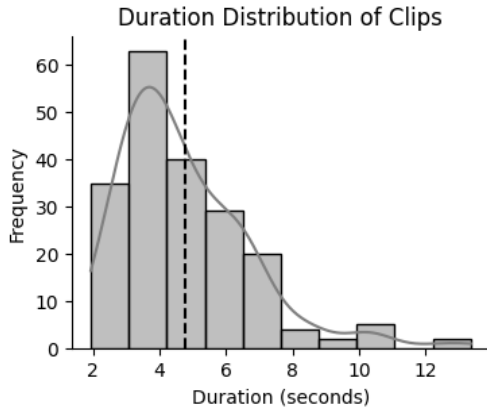


Figure 1: Duration distribution of the clips in the dataset. Dashed line indicates mean.

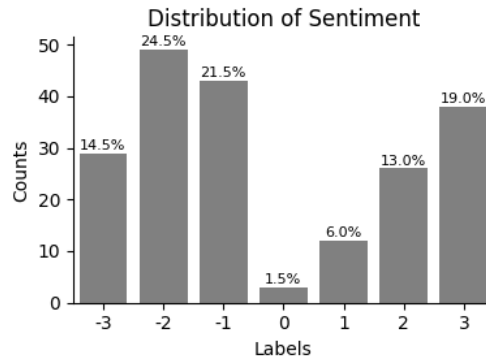


Figure 2: Distribution of sentiment labels. The labels correspond to the 7-point Likert scale where -3 is extremely negative, 0 is neutral, and 3 is extremely positive. Numbers above the bars indicate count.

The dataset includes 4 different signers, and primarily depicts scenarios from everyday life such as conversations about the weather, family members and medical checkups.

Figure 2 shows the distribution of the sentiment labels. There are relatively few clips with neutral sentiment, but this is expected, since we selected clips with captions that had salient positive or negative emotions based on VADER. Figure 3 shows the distributions of the emotion categories. We binarized the presence of each emotion and the detailed breakdown distribution can be found in the Appendix A.5.

Looking at the annotators’ response to attributions of emotional cues, we found some common themes which we elucidate in brief: (1) The non-manual markers are the primary cue for recognizing emotion in ASL. These includes *facial expressions* such as furrowed brows, pursed lips, and squinted eyes, *head movements* such as the head thrusts, tilting, and orientation changes that intensify the emotion, *mouth movements* such as “O” shape, tongue out, and puffed lips, and *body movement* such as shoulder raising and full-body tilting. (2) Signs were modified and emphasized for emotional expressions. Sign size (large/small), speed (fast/slow), repetition, and finger-spelling (sometimes for emphasis) are all noted as sign-based emotional markers. For stronger emotions (both positive and negative), signs are produced more broadly, quickly, or emphatically, very often in parallel with expressive non-manuals markers. (3) The role and context of the sentence is important to disambiguate emotions. Markers such as shifts in eye gaze, physical orientation, and changes in signing space are all used by signers to signal narrative switching or changes in perspective. The lack of context can cause uncertainty in emotion identification.

## 4 Benchmarks

### 4.1 Benchmark Tasks

We describe three benchmark tasks of increasing complexity that we carried out using the EmoSign dataset.

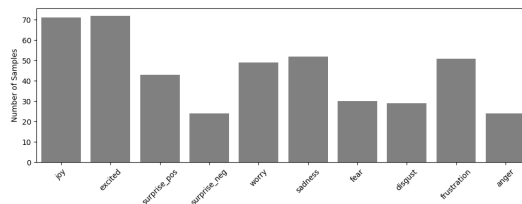


Figure 3: Distribution of emotion categories based on binarized presence across clips.

**Sentiment Analysis.** In this classification task, the goal is to predict the overall sentiment of a sign video on a 7-point Likert scale (-3 to 3), with a score of -3 corresponding to Strongly Negative and a score of +3 corresponding to Strongly Positive. In addition, we also formulated a coarse version of the task where the model is simply required to predict whether the sentiment is positive, neutral or negative. Following prior work [51, 50], we evaluate performance using accuracy (ACC) and weighted F1-score (WAF). We chose WAF as the primary metric and ACC as the secondary metric due to label imbalance.

**Single-label Emotion Classification.** Building on the sentiment analysis task, the goal of the emotion classification task is to predict the specific emotion that is most dominant or most present in a sign video. We first combined “joy” and “excited” into one category “happiness” due to their high co-occurrence within a single video (Jaccard similarity score of 0.81, Figure 5 in Appendix). If all of the emotion categories were labeled as “not present”, then a “neutral” label is assigned. We then separated the dataset into two sets for single-expression set and multi-expression set. The multi-expression set only contains classes that have more than 2 samples. The single expression set consists of 140 clips and the multiple expression set consist of 37 clips. Appendix A.5 shows the distributions of the single and multiple expression set respectively.

The model is required to select a single emotion label from a predefined set of ten possible labels (described in Section 3.2), as well as the label Neutral. Note that the model is not required to assess emotional intensity in this task. For each single-expression label, we present the accuracy and F1 score. Holistically, we chose weighted accuracy (wAcc) and weighted F1-score (wF1) as the evaluation metric based on prior approaches [21, 32].

**Emotion Cue Grounding.** The ability to accurately identify task-relevant temporal and spatial regions of a video, otherwise known as grounding, is a crucial component of video question-answering [34] systems and has a wide range of valuable applications ranging from human-computer interaction [46] to clinical settings [48]. Yet, there remains a significant performance gap between MLLMs and human annotators on visual grounding tasks [47].

The goal of the emotion cue grounding task is to identify video frames and spatial regions relevant to the sentiment analysis and emotion classification tasks described above.

## 4.2 Baselines

We selected several multimodal LLMs (MLLMs) that support video-language inputs to obtain baseline results on the EmoSign dataset. Model cards and qualitative examples of model outputs are provided in Appendix A.2. For all models and tasks, we conduct ablation studies to better understand the influence of individual modalities. Inference was conducted on a single 80GB NVIDIA Tesla A100 GPU.

We used GPT-4o to annotate the videos in the same way that a human annotator would have (See Appendix A.3 for prompts and sample outputs). Each video was sampled at 10 fps and coded in base64. We forced structured output where each API calls outputs responses to all three tasks.

We also evaluated the performance of AffectGPT [27], Qwen2.5-VL-7B-Instruct [2] and MiniGPT4 [1] on the benchmark tasks. AffectGPT is a multimodal LLM (MLLM) fine-tuned on various emotion recognition and understanding tasks. Qwen2.5-VL-7B-Instruct is a general-purpose MLLM optimized for instruction following, while MiniGPT4 is a vision-language model that uses an additional projection layer to align visual features with the language branch that displays strong performance on a range of benchmark tasks. In preliminary tests, we found that AffectGPT, Qwen2.5-VL-7B-Instruct and MiniGPT4 were unable to consistently produce clean output when prompted to respond to all three benchmark tasks at once. We adapted the prompts (detailed in Appendix A.4) to improve the structure and interpretability of model outputs and conducted inference on each task separately.

## 5 Results

In this section, we present the baseline results of the selected MLLMs against the ground truth provided by the ASL native signers. One of the authors (hearing, not fluent in ASL) also annotated each video using the same interface, providing an estimate of how a human without access to the

modality	model	sentiment (3-class)		sentiment (7-class)	
		wAcc	wF1	wAcc	wF1
video	MiniGPT4	34.68	40.00	14.46	13.03
	Qwen 2.5	27.34	16.47	10.26	2.44
	AffectGPT	33.33	0.04	14.29	0.04
	GPT-4o	40.72	24.43	19.81	5.97
video + caption	MiniGPT4	21.65	36.89	9.76	12.18
	Qwen 2.5	41.10	54.29	15.84	14.51
	AffectGPT	<b>56.18</b>	64.37	21.02	16.13
	GPT-4o	52.13	<b>76.72</b>	<b>22.89</b>	<b>26.35</b>
	Hearing person*	55.64	57.64	25.48	21.39

Table 3: Benchmark results of sentiment analysis.

sign video captions would understand the emotions in the video. The experimental results revealed that providing access to the video captions almost always improved model performance; below, we provide detailed error analysis.

### 5.1 Sentiment Analysis

When only provided with the sign videos, models exhibited diverse biases and behaviors. AffectGPT consistently output sentiment as Neutral, suggesting an almost-complete lack of ability to recognize emotions in sign videos from visual cues alone. GPT-4o and Qwen2.5-VL-7B-Instruct tend to skew towards predicting positive sentiment, while we did not observe a consistent pattern for MiniGPT4. Interestingly, the hearing annotator demonstrated a tendency to perceive neutral clips as negative and positive clips as neutral. These results align with previous study findings that hearing non-signers often misinterpret linguistic facial markers as indicators of negative emotion [30].

When also given access to the video captions, GPT-4o shows improved performance but still retains a slight positive bias, and rarely selects the Neutral label. Both AffectGPT and Qwen2.5-VL-7B-Instruct have a tendency to predict simply Positive, Neutral or Negative even in the 7-class sentiment task, suggesting limitations in their ability to perform fine-grained sentiment analysis. MiniGPT4 shows a notable positive bias and appears to struggle to effectively integrate text data into its reasoning.

While hearing annotators tend to lean towards negative interpretations in the absence of text captions for additional context, the models, in general, exhibit a slight neutral-to-positive bias. This discrepancy highlights the gap in human versus machine perception in the domain of sign language sentiment analysis. A possible reason for this gap is that many foundational models are pre-trained with an emphasis on being helpful, harmless and honest [3], leading to a neutral or positive bias to mitigate potentially harmful or incorrect assertions about a person’s emotion state, especially in ambiguous contexts. However, more research is required to fully understand these observed model behaviors.

### 5.2 Single-label Emotion Classification

In the context of single-label emotion classification, when only provided with the sign videos, the models demonstrated limited ability to identify emotions beyond very broad and common categories. GPT-4o almost always classified videos as displaying either happiness or frustration, suggesting that it falls back to common emotional descriptors without the presence of text as a contextual guide. Similarly, AffectGPT limited its predictions mostly to happiness, sadness, or neutral emotions, Qwen2.5-VL-7B-Instruct to happiness and neutral, and MiniGPT4 predominantly classified videos as happy.

Access to the video captions led to improved model performance, allowing for more accurate and nuanced emotional classification. Although GPT-4o still occasionally defaults to happiness and frustration, it shows enhanced capacity to distinguish emotions such as worry and disgust, and generally succeeded in identifying sentiment correctly. GPT-4o’s tendency to favor labels such as worry and fear were aligned with the emotion co-occurrence patterns observed in the ground truth labels in EmoSign, suggesting that these emotions are relatively close together in the language embedding space [18].

AffectGPT still retained its tendency to give neutral predictions, though less so than before. It occasionally confused frustration for happiness, but otherwise generally succeeded in identifying

modality	model	HP	SP(P)	SP(N)	WR	SD	FR	DG	FS	AG	NE	total	total
		Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	wAcc	wF1
video	MiniGPT4	69	20	14	0	0	0	0	0	0	27	13.01	22.02
	Qwen 2.5	35	0	<b>43</b>	0	0	0	0	5	<b>33</b>	27	14.39	18.53
	AffectGPT	11	0	0	7	<b>30</b>	0	0	5	0	<b>73</b>	12.62	11.03
	GPT-4o	35	0	0	7	0	0	20	53	0	0	11.50	20.76
video + caption	MiniGPT4	<b>89</b>	0	14	7	<b>30</b>	<b>43</b>	10	0	<b>33</b>	9	23.56	35.89
	Qwen 2.5	63	<b>40</b>	29	64	0	14	20	32	<b>33</b>	55	34.96	44.67
	AffectGPT	85	20	0	50	<b>30</b>	14	10	32	<b>33</b>	27	30.17	47.77
	GPT-4o	<b>89</b>	20	0	<b>79</b>	20	29	<b>50</b>	<b>74</b>	0	0	<b>35.97</b>	<b>55.09</b>

Table 4: Benchmark results of single expression emotion classification. HP: happiness; SP(P): surprise (positive); SP(N): surprise (negative); WR: worry; SD: sadness; FR: fear; DG: disgust; FS: frustration; AG: anger; NE: neutral.

overall sentiment correctly. Qwen2.5-VL-7B-Instruct showed improved performance on the emotion classification task, but developed a tendency to predict frustration. Like AffectGPT, it occasionally confused frustration for happiness. MiniGPT4 continued to display a bias towards labeling videos as happy, even for videos with a ground truth of negative emotions such as disgust. These persistent biases suggest the need for further model enhancements and fine-tuning on sign videos to improve their emotion recognition capabilities.

### 5.3 Reasoning analysis

To obtain a preliminary understanding of model abilities to perform emotion cue grounding, we conducted a manual inspection of several randomly selected videos alongside the ground truth and each model’s corresponding reasoning outputs.

Without captions, the inference outputs show that MiniGPT4 and GPT-4o are capable of identifying specific facial expressions within the sign videos and using them to reason about the signer’s emotions, suggesting that these models can capture and interpret visual nuance to some extent. In contrast, AffectGPT and Qwen2.5-VL-7B-Instruct were only able to provide generic descriptions such as "The signer’s facial expressions and body language do not convey strong emotions", or "The facial expressions are neutral, and the body language shows no particular direction or activity".

With the captions provided, the models appear to use the linguistic context to guide their reasoning over the visual inputs. Despite a general improvement in task performance, we observed several failure modes: models sometimes misinterpreted the sentiment of the text caption, or correctly understood the text sentiment but evaluated the visual inputs in ways that diverged from the Deaf annotators. Furthermore, models would claim to identify specific cues, such as hand gestures and posture, that suggested certain emotions. When checking these cues against the sign videos, we observed that the cues recognized by GPT-4o, Qwen2.5-VL-7B-Instruct and MiniGPT4 (e.g., a thumbs-up sign) were truly present in the video. However, there was a recurring sense that the models were attempting to construct explanations that were consistent with their judgment of the text sentiment, rather than independently recognizing emotions from visual cues.

GPT-4o frequently repeated statements such as "relaxed body language" and "generally positive sentiment", indicating a possible over-reliance on common language patterns without truly consulting the visual context. Qwen2.5-VL-7B-Instruct often highlighted the lack of audio and exhibited reluctance to make definitive statements about the signer’s emotion. Its reasoning sometimes demonstrated a lack of understanding about sign language as a concept, with outputs such as "the exact content of the sign language cannot be determined without audio." As in the sentiment and emotion classification tasks, AffectGPT’s reasoning displays a bias towards the neutral label, frequently repeating statements such as "neutral expression" and "lack of obvious body language cues." Like Qwen2.5-VL-7B-Instruct, it was also hesitant to make statements about the signer’s emotional state.

We observed a significant performance gap between the ground-truth labels and the MLLM predictions, especially when models are only given access to the visual modality. This gap underscores the difficulty of the benchmark tasks as well as the current limitations of MLLMs in comprehending the nonverbal emotion cues in sign language. These findings align with prior work indicating that MLLMs often struggle with visual understanding, and that strong performance on visual question-answering tasks are likely not due to genuine visual comprehension but rather a result of language shortcuts and spurious correlations with irrelevant visual information [47].

## 6 Limitations

In this paper, we propose a new dataset of sign videos with sentiment, emotion and open-ended emotion cue labels, as well as benchmark tasks on the dataset. We acknowledge several limitations of the present work.

While VADER [19] is an efficient way of performing sentiment analysis for large text datasets, its lack of sophistication in emotion understanding could have influenced the diversity and representativeness of selected videos. The sign videos around which we build EmoSign were derived from the ASLLRP continuous signing corpus [36], which does not include more complex contexts often present in real-world scenarios such as in-the-wild footage or videos featuring multiple speakers. Due to the high cost of manual annotations, we only selected 200 sign utterances for inclusion in the initial EmoSign dataset. As such, there are some restrictions in the range of emotions captured in the current dataset. Class imbalances pose an additional challenge due to the relatively small number of videos. To address many of these issues, we plan to expand the dataset size in future work. Further, we plan to expand the open-ended emotion cue annotations by incorporating spatial and temporal information, which could provide critical context for the emotion cue grounding task.

In terms of benchmarking, we evaluated a relatively small set of models, which may not capture the full range of capabilities and limitations across different model architectures and weights. We also did not evaluate model performance on multi-label emotion classification tasks [32]. A broader evaluation across more models and tasks could provide a more comprehensive picture of current models' abilities to recognize emotion in sign language videos. Importantly, none of the models we tested were specifically trained on American Sign Language (ASL). Future work could investigate fine-tuning models on ASL data to improve recognition of sign gestures, which could directly support more robust emotion recognition.

## 7 Conclusion

In this paper, we introduced EmoSign, the first multimodal dataset containing sentiment and emotion labels specifically for ASL videos. By providing annotations from Deaf native ASL signers on sentiment, emotion categories, and detailed descriptions of emotion cues, EmoSign addresses a critical gap in both sign language research and emotion recognition. Our benchmark evaluations of several state-of-the-art multimodal LLMs revealed significant limitations in their ability to recognize emotions in sign language videos, particularly when relying solely on visual input without text captions. We hope that EmoSign will encourage more research on the nuanced emotional expressions unique to sign languages as well as innovations in how multimodal models' ability to understand emotion-laden signed videos, leading to the creation of more emotionally expressive sign language interpretation.

## Acknowledgments and Disclosure of Funding

We would like to acknowledge Paul Liang for providing us with feedback and support, and students from Gallaudet University who assisted with pilot tests of the annotation interface.

## References

- [1] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

- [4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.
- [5] Matyáš Boháček and Marek Hruží. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191, 2022.
- [6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.
- [7] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171, 2017.
- [8] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*, 2024.
- [9] Phoebe Chua, Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. Perspectives on capturing emotional expressiveness in sign language, 2025. URL <https://arxiv.org/abs/2505.08072>.
- [10] Daniel Thomas Cordaro. *Universals and cultural variations in emotional expression*. University of California, Berkeley, 2014.
- [11] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38): E7900–E7909, 2017.
- [12] Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201, 1997.
- [13] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744, 2021.
- [14] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [15] Eeva A Elliott and Arthur M Jacobs. Facial expressions, emotions, and sign languages. *Frontiers in psychology*, 4:115, 2013.
- [16] Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. Signllm: Sign languages production large language models. *arXiv preprint arXiv:2405.10718*, 2024.
- [17] Jari K Hietanen, Jukka M Leppänen, and Ulla Lehtonen. Perception of emotions in the hand movement quality of finnish sign language. *Journal of nonverbal behavior*, 28:53–64, 2004.
- [18] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- [19] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.
- [20] Tyler G James, Kyle A Coady, Jeanne-Marie R Stacciarini, Michael M McKee, David G Phillips, David Maruca, and JeeWon Cheong. “they’re not willing to accommodate deaf patients”: communication experiences of deaf american sign language users in the emergency department. *Qualitative Health Research*, 32(1):48–63, 2022.

- [21] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.
- [22] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.
- [23] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019.
- [24] Deng Li, Xin Liu, Bohao Xing, Baiqiang Xia, Yuan Zong, Bihan Wen, and Heikki Kälviäinen. Eald-mlm: Emotion analysis in long-sequential and de-identity videos with multi-modal large language model. *arXiv preprint arXiv:2405.00574*, 2024.
- [25] Hengzhi Li, Megan Tjandrasuwita, Yi R Fung, Armando Solar-Lezama, and Paul Pu Liang. Mimeqa: Towards socially-intelligent nonverbal foundation models. *arXiv preprint arXiv:2502.16671*, 2025.
- [26] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614, 2023.
- [27] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.
- [28] Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juzhe Zhang, Xin Chen, Jingyi Yu, and Lan Xu. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*, 2024.
- [29] Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Hemm: Holistic evaluation of multimodal foundation models. *arXiv preprint arXiv:2407.03418*, 2024.
- [30] Hyunchul Lim, Minghan Gao, Franklin Mingzhe Li, Nam Anh Dang, Ianip Sit, Michelle M Olson, and Cheng Zhang. Exploring the impact of emotional voice integration in sign-to-speech translators for deaf-to-hearing communication. *arXiv preprint arXiv:2412.05738*, 2024.
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL <https://arxiv.org/abs/2310.03744>, 3(4):5, 2023.
- [32] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32, 2022.
- [33] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.
- [34] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024.
- [35] Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. New shared & interconnected asl resources: Signstream@ 3 software; dai 2 for web access to linguistically annotated video corpora; and a sign bank. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, Miyazaki, Language Resources and Evaluation Conference 2018*, 2018.

- [36] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp). *arXiv preprint arXiv:2201.07899*, 2022.
- [37] Roland Pfau, Markus Steinbach, and Bencie Woll. *Sign language*. De Gruyter Mouton, 2012.
- [38] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.
- [39] Judy S Reilly, Marina L McIntire, and Howie Seago. Affective prosody in american sign language. *Sign Language Studies*, 75(1):113–128, 1992.
- [40] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [41] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022.
- [42] Eiji Taira and Shizuka Itagaki. How hearing people understand the deaf and some legal implications of their misinterpretation of visual expressions. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 32:819–829, 2019.
- [43] Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. Roberta- lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525, 2022.
- [44] Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36:29029–29047, 2023.
- [45] Clayton Valli and Ceil Lucas. *Linguistics of American sign language: An introduction*. Gallaudet University Press, 2000.
- [46] Greg Wadley, Vassilis Kostakos, Peter Koval, Wally Smith, Sarah Webber, Anna Cox, James J Gross, Kristina Höök, Regan Mandryk, and Petr Slovák. The future of emotion in human-computer interaction. In *CHI Conference on human factors in computing systems extended abstracts*, pages 1–6, 2022.
- [47] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.
- [48] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150, 2012.
- [49] Kayo Yin, Chinmay Singh, Fyodor O Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex X Lu, and Danielle Bragg. Asl stem wiki: Dataset and benchmark for interpreting stem articles. *arXiv preprint arXiv:2411.05783*, 2024.
- [50] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.
- [51] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

## Neurips Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: See Section 1.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See Section 6.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, we provide the model cards in Appendix A.2 as well as the prompts used to generate the main experimental results in Appendix A.3 and A.4.

### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide access to the EmoSign dataset and code used for data processing. We also provide all prompts used to generate the main experimental results in Appendix A.3 and A.4.

### 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: We did not run any experiments fine-tuning or training new models.

### 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our baseline evaluations did not include factors that require error bars, confidence intervals or statistical significance tests.

### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4.2 specifies that all inference was conducted on a single 80GB NVIDIA Tesla A100 GPU.

### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: All annotators were paid fair wages to the best of our knowledge. The research was carried out after procedures were reviewed and approved by our IRB. As a sign language dataset, the videos inevitably contain personally identifiable information. We only release the video and utterance IDs of the clips included in the EmoSign dataset, following the access rules on the ASLLRP dataset which it is based on.

#### 10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss potential positive and negative societal impacts in Sections 1, 2 and 7.

#### 11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We release an extension of a subset of the ASLLRP dataset [36]. The utterances were created by ASL consultants so that they 1) cover a range of linguistic constructions (e.g., questions and conditional sentences), and 2) are natural for signers to produce [35]. There are no unsafe images or videos in the sign data, which are publicly available through a data access interface\*.

#### 12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The sign videos were derived from ASLLRP [35, 36]. The license, terms of use and url of the source data is provided at our huggingface link: <https://huggingface.co/datasets/catfang/emosign>.

#### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Documentation is provided in Section 3.2 and on Huggingface.

#### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The full text of instructions given to participants and screenshots of the annotation interface are available in Appendix A.1. We do not have the details of how much each individual annotator was paid for their work, as annotators were hired through a third-party vendor. However, we disclose the amount paid to the third-party vendor also in Appendix A.1

#### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

---

\*<https://www.bu.edu/asllrp/indexright.html>

Answer: [Yes]

Justification: We recruited human annotators to label sign videos for sentiment, emotion and free-response descriptions of emotion attributions. No risks outside of those normally encountered in everyday life were anticipated. The paper describes our IRB approval in 3.2.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe our use of LLMs for benchmarking in Sections 4.2 and 5 as well as in Appendix A.3 and A.4.

# A Appendix / supplemental material

## A.1 Data Collection Additional Details

We employed a total of 3 annotators through a third-party vendor, specifically a full-service sign language interpreting and captioning company specializing in American Sign Language (ASL). The annotators utilized the interface shown in Figure 4 for the annotation tasks. All three annotators completed all annotation tasks for all sign videos in the dataset. The vendor was paid \$400 per annotator.

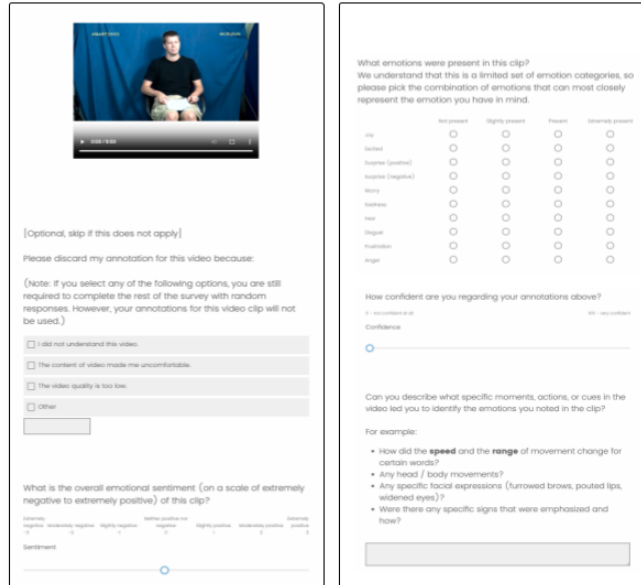


Figure 4: Annotation Interface

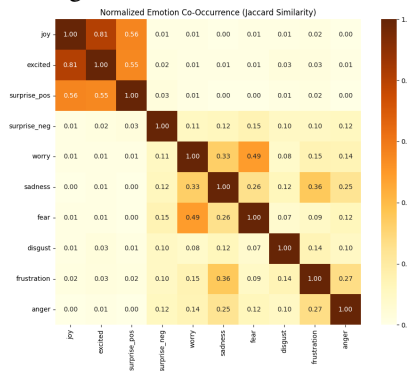


Figure 5: Jaccard Similarity of the original set of Emotion Labels.

## A.2 Model Cards and Outputs

Model	Link
Qwen2.5-VL-7B-Instruct [2]	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct/tree/main">https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct/tree/main</a>
Qwen3-8B *	<a href="https://huggingface.co/Qwen/Qwen3-8B">https://huggingface.co/Qwen/Qwen3-8B</a>
AffectGPT [27]	<a href="https://github.com/zeroQiaoba/AffectGPT">https://github.com/zeroQiaoba/AffectGPT</a>
MiniGPT4-video [1]	<a href="https://github.com/Vision-CAIR/MiniGPT4-video">https://github.com/Vision-CAIR/MiniGPT4-video</a>

Table 5: Model cards for MLLMs used in EmoSign baseline results.

## A.3 GPT-4o Details

The prompt template used for GPT-4o is as follows:

You are an expert in the field of emotions.

Please focus on facial expressions, body language, environmental cues, and events in the video and predict the emotional state of the character. Please ignore the character’s identity. We uniformly sample 10 frames per second from this clip. Please consider the temporal relationship between these frames.

The video involves a person signing a sentence in ASL. You have three tasks:

Task 1 - On a scale of extremely negative (-3) to extremely positive (+3), what is the overall affective sentiment of this clip? Output your classification as a number between -3 and 3.

Task 2 - Which of the following affective categories were present in this clip? You may choose multiple options. The scale is from 0 to 3, where 0 is not present and 3 is extremely present. The possible categories are: joy, excited, surprise (positive), surprise (negative), worry, sadness, fear, disgust, frustration, anger. Output your classification as a json with the name of the emotion and a number between 0 and 3.

Task 3 - Describe what specific moments, actions, or cues in the video that led you to your responses above? (e.g., the speed at which the person signed, specific facial expressions, content of the sign, etc). Output your 2-sentence response as a string.

You should provide a structured response in the form of a Json string.

A sample response is formatted as follows:

```
"filename": "00ADU7t7IWI_1",
"Sentiment": "2",
"joy": "3",
"excited": "2",
...
"anger": "0",
"QA": "The person is smiling and signing with energetic movements, indicating a positive and engaging message. The consistent smile and lively hand gestures suggest joy and excitement."
```

#### A.4 Other MLLM prompts for Sentiment and Emotion Recognition

The remaining MLLMs used for baseline evaluation struggled to follow the prompt template used for GPT-4o, which is relatively long and complex. As such, we created separate prompts for the sentiment analysis and single-label emotion recognition tasks presented in Section 5.. Several of the multimodal LLMs used in the analysis had a dedicated method for incorporating video captions, in which case we used it to pass the caption to the mode. Otherwise, we included the caption directly within the prompt.

**Sentiment Prompt:** You are an expert in the field of emotions. The video involves a person signing a sentence in ASL. Please focus only on facial expressions and body language of the signer in the video and try to recognize the emotional state of the signer. Your task is: on a 7 point Likert scale ranging from extremely negative to extremely positive, what is the overall affective sentiment of this clip? You must select an answer from this list: [Extremely Negative, Negative, Somewhat Negative, Neutral, Somewhat Positive, Positive, Extremely Positive].

**Single-label Emotion Recognition Prompt:** You are an expert in emotion analysis. Your task is to analyze the given video clip and select the SINGLE MOST DOMINANT emotion from the following list: [joy, excited, surprise (positive), surprise (negative), worry, sadness, fear, disgust, frustration, anger].

**\*\*Rules you MUST follow:\*\***

1. **\*\*You must choose ONLY ONE emotion\*\***—the one that is most prominent in the clip.
2. If multiple emotions seem present, select the strongest one.
3. If unsure, make your best judgment—do not leave the response blank.

**\*\*Penalties for not following instructions:\*\***

- If you output more than one emotion, you will fail the task.
- If you refuse to answer or leave it blank, you will fail the task.

**\*\*Output format:\*\*** [selected emotion]

Now, analyze the clip and provide the SINGLE emotion label:

**Sentiment extraction prompt:** Based on this text, what is the sentiment of the signer? The text is: [inference output]. Keep your answer brief without any explanation.

**Emotion label extraction prompt:** Based on this text, what are the signer’s emotions? The text is: [inference output]. Keep your answer brief without any explanation.

Figure 6 provides an example of a sign video and the associated MLLM outputs from our baseline experiments.



Figure 6: Sample frames from a sign video in the EmoSign dataset. The caption is: "If Mary gets home late, John will probably be upset."

**Qwen2.5-VL-7B-Instruct (with captions):** Based on the signer's hand gestures and body language, it appears that they are expressing a sense of concern or worry about the situation described in the video caption. The signer's facial expression seems to convey a sense of anticipation or anxiety, which aligns with the idea of being upset if someone arrives late. Therefore, the most appropriate emotional state for the signer in this context would be "worry".

**Qwen2.5-VL-7B-Instruct (without captions):** Based on the signer's hand gestures and body language, it appears that they are expressing a sense of frustration or anger. The signer seems to be making emphatic movements with their hands, which could indicate a strong emotional response. However, without audio context, it's challenging to determine the exact nature of the emotion.

**AffectGPT (with captions):** The visual cues from the facial expressions and body movements of the signer, along with the caption "If Mary comes back late, John must be upset," suggest that the signer might be conveying a sense of worry or anxiety. This emotional state could be related to the potential impact of Mary's lateness on John. Therefore, the primary emotions displayed by the signer in the video might include worry, sadness, or anxiety.

**AffectGPT (without captions):** Given the video description, the signer's facial expressions and body language do not reveal any obvious signs of strong emotion. The head is slightly tilted, possibly indicating interest or concentration in the signing action. The eyes are closed, possibly to focus or concentrate on the signing, and the mouth is open, possibly due to speech or to express emotion. Considering these clues, it can be inferred that the signer might be in a calm or neutral emotional state. However, without additional information about the context of the signing or the individual's background, it is not possible to accurately determine the character's emotional state.

**MiniGPT4-video (with captions):** Mostly, John appears to be upset or frustrated due to Mary delaying returning home. His eyebrows are furrowed together, his lips curled downwards (shown by the wrinkles above them), as well as a slight clench of this jaw muscles which indicates stress or irritation - all these signs suggest he feels either anger/frustration (disgust) at her lateness. Secondly one can see worry or anxiety expressed through brow wiggling while watching oneself being signing - there is an urge hidden behind those raised upper lids that could turn into full blown concern if things go wrong further...

**MiniGPT4-video (without captions):** The signer often exhibits a joyful expression during certain moments in the signature process. For example, when they finish signing an important sentence or document with satisfaction. They also show signs of happiness such as smiling wide and laughing occasionally which indicates positive emotions. There were instances wherein excitement is palpable from their energetic gestures particularly when discussing thrilling topics like travel experiences abroad or new technologies that arouse enthusiasm.

## A.5 Additional Descriptive Statistics of the Dataset and Benchmark Results

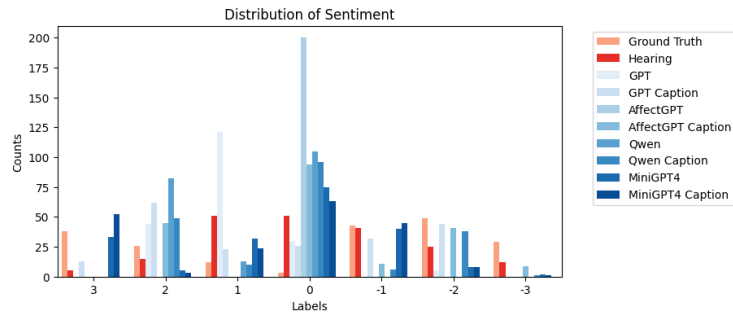


Figure 7: Distribution of Sentiment Analysis Results (7-class).

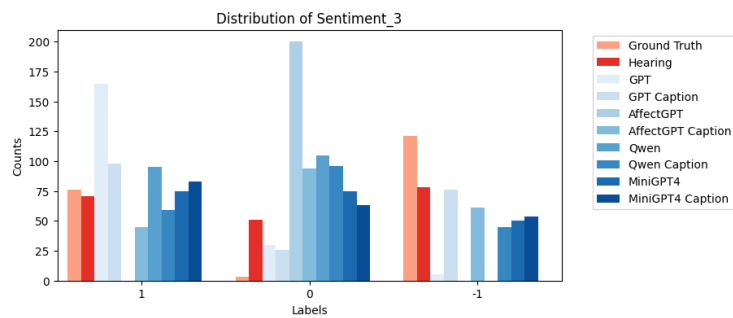


Figure 8: Distribution of Sentiment Analysis Results (3-class).

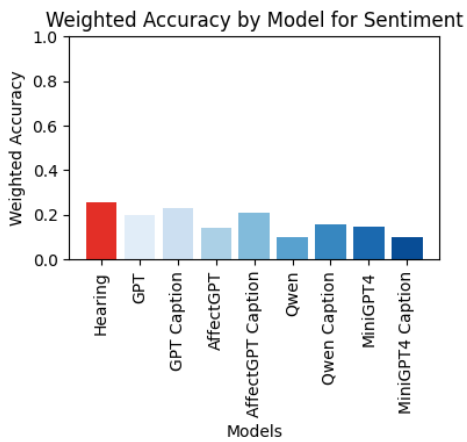


Figure 9: Weighted Accuracy of Models on Sentiment Analysis Task (7-class).

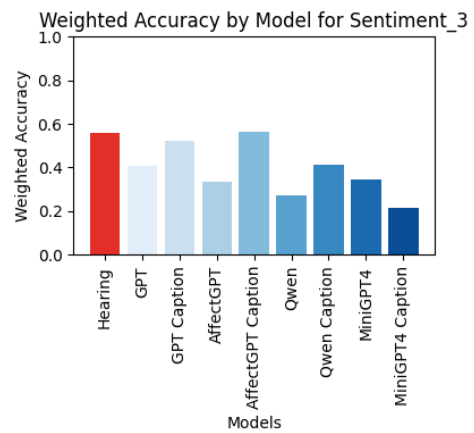


Figure 10: Weighted Accuracy of Models on Sentiment Analysis Task (3-class).

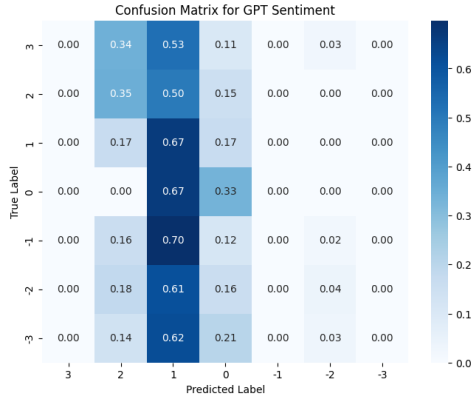


Figure 11: Confusion Matrix of GPT (without caption) output on Sentiment Analysis Task (7-class).

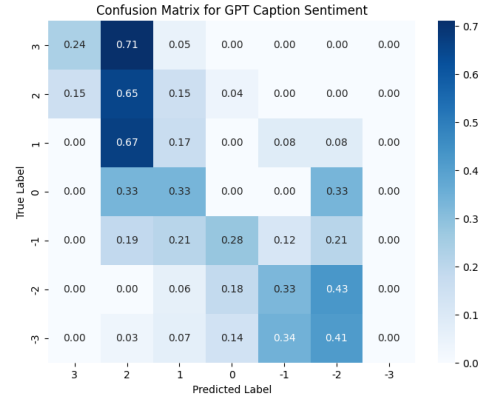


Figure 12: Confusion Matrix of GPT with caption output on Sentiment Analysis Task (7-class).

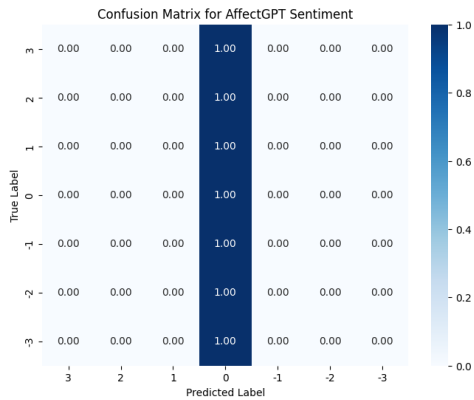


Figure 13: Confusion Matrix of AffectGPT (without caption) output on Sentiment Analysis Task (7-class).

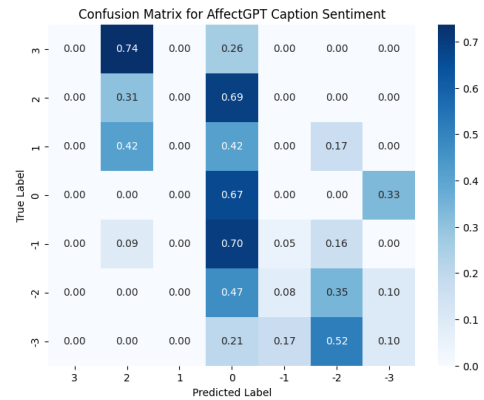


Figure 14: Confusion Matrix of AffectGPT with caption output on Sentiment Analysis Task (7-class).

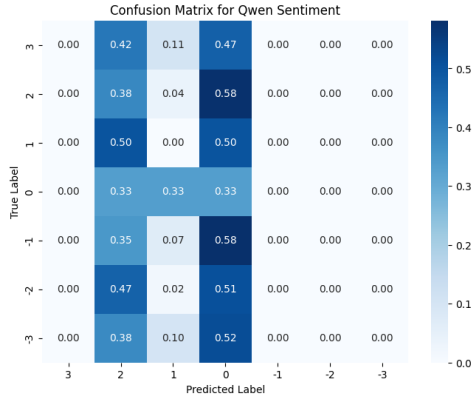


Figure 15: Confusion Matrix of Qwen-2.5 (without caption) output on Sentiment Analysis Task (7-class).

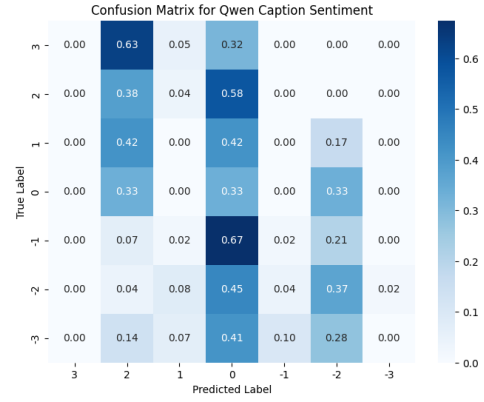


Figure 16: Confusion Matrix of Qwen-2.5 with caption output on Sentiment Analysis Task (7-class).

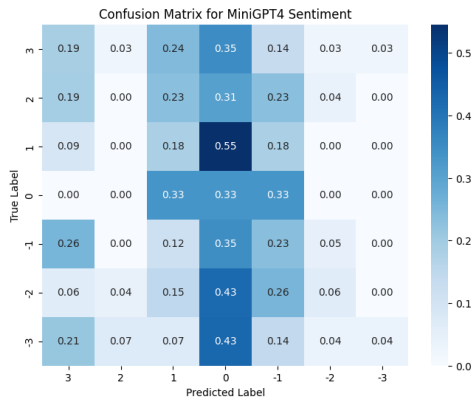


Figure 17: Confusion Matrix of MiniGPT4 (without caption) output on Sentiment Analysis Task (7-class).

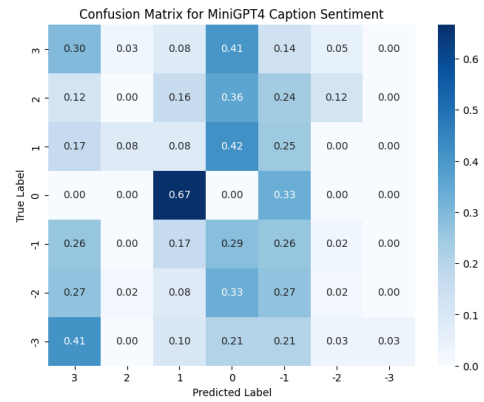


Figure 18: Confusion Matrix of MiniGPT4 with caption output on Sentiment Analysis Task (7-class).

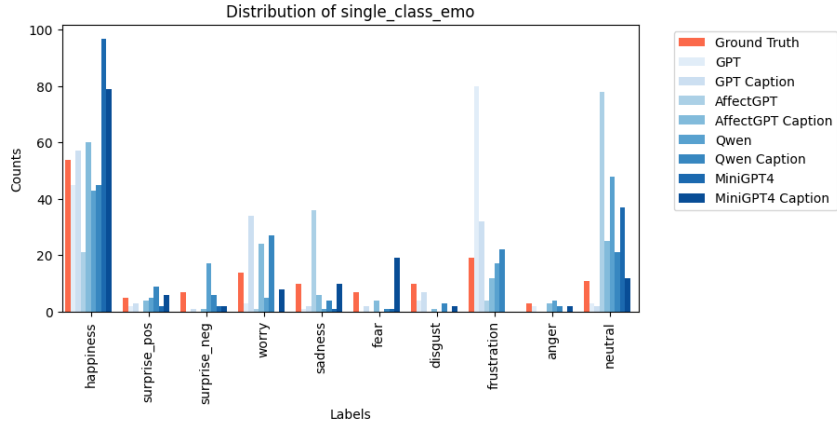


Figure 19: Distribution of Single Expression Emotion Classification Results.

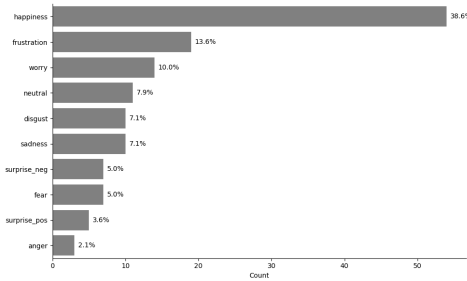


Figure 20: Distribution of emotion categories within the single expression set. Numbers above the bars indicate count.

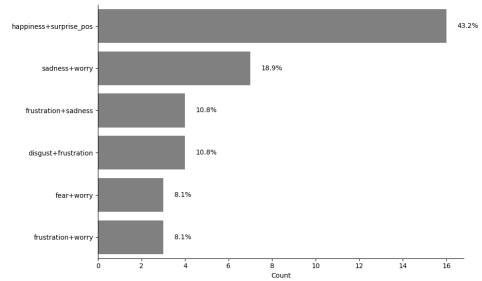


Figure 21: Distribution of emotion categories within the multi expression set. Numbers above the bars indicate count.

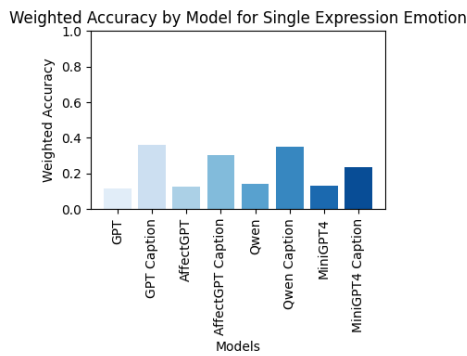


Figure 22: Weighted Accuracy of Models on Single Expression Emotion Classification Task.

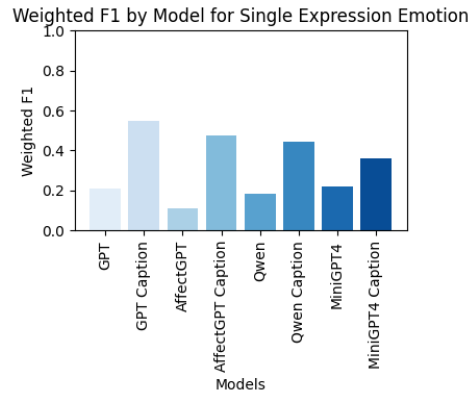


Figure 23: Weighted Accuracy of Models on Single Expression Emotion Classification Task.

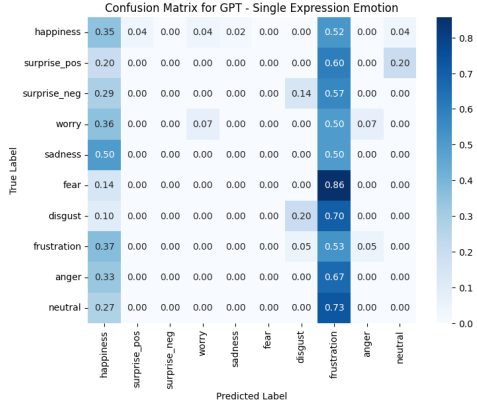


Figure 24: Confusion Matrix of GPT (without caption) output on Single Expression Emotion Classification Task.

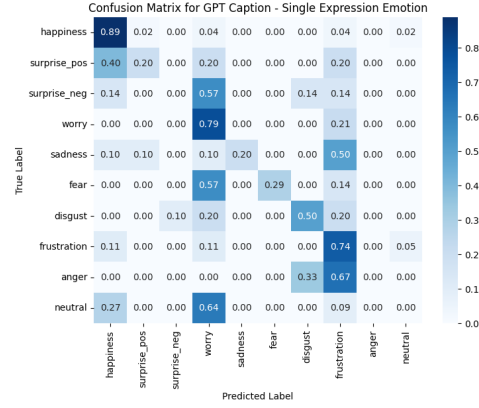


Figure 25: Confusion Matrix of GPT with caption output on Single Expression Emotion Classification Task.

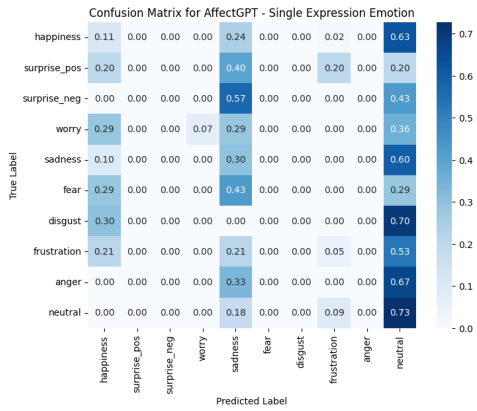


Figure 26: Confusion Matrix of AffectGPT (without caption) output on Single Expression Emotion Classification Task.

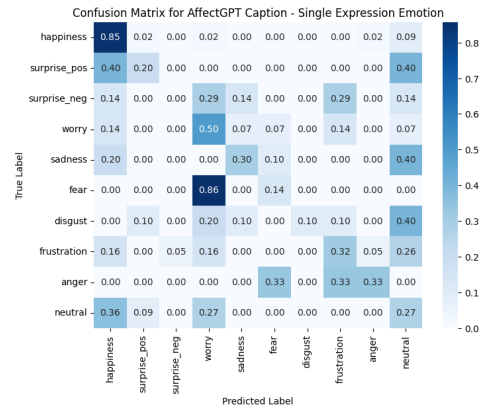


Figure 27: Confusion Matrix of AffectGPT with caption output on Single Expression Emotion Classification Task.

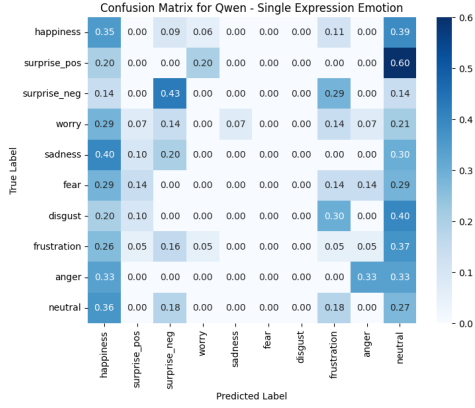


Figure 28: Confusion Matrix of Qwen-2.5 (without caption) output on Single Expression Emotion Classification Task.

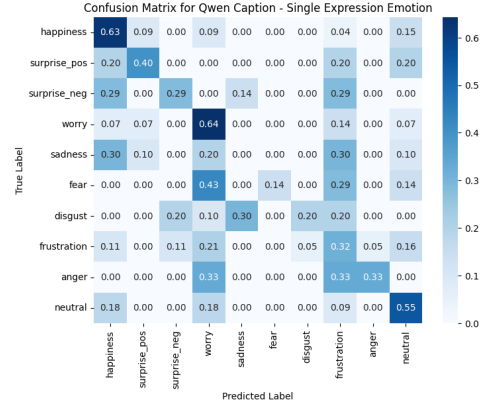


Figure 29: Confusion Matrix of Qwen-2.5 with caption output on Single Expression Emotion Classification Task.

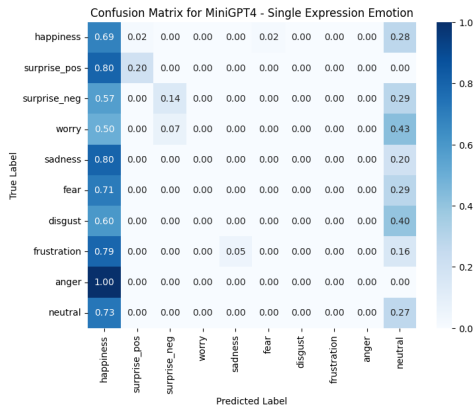


Figure 30: Confusion Matrix of MiniGPT4 (without caption) output on Single Expression Emotion Classification Task.

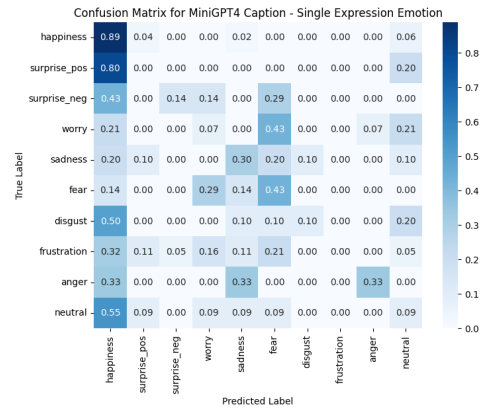


Figure 31: Confusion Matrix of MiniGPT4 with caption output on Single Expression Emotion Classification Task.