Emotion Recognition in Signers

Anonymous ACL submission

Abstract

Recognition of signers' emotions suffers from one theoretical challenge and one practical challenge, namely, the overlap between grammatical and affective facial expressions and the scarcity of data for model training. This paper addresses these two challenges in a cross-lingual setting using our eJSL dataset, a new benchmark dataset for emotion recognition in Japanese Sign Language signers, and BOBSL, a large British Sign Language dataset with subtitles. In eJSL, two signers expressed 78 distinct utterances with each of seven different emotional states, resulting in 1,092 video clips. We empirically demonstrate that 1) textual emotion recognition in spoken language mitigates data scarcity in sign language, 2) temporal segment selection has a significant impact, and 3) incorporating hand motion enhances emotion recognition in signers. Finally we establish a stronger baseline than spoken language LLMs (Qwen 2.5 and GPT-4o).

1 Introduction

Emotion recognition is a core topic not only in natural language processing (Yun et al., 2024) but also in affective computing and human-computer interaction (Zeng et al., 2009; El Ayadi et al., 2011), enabling more natural and empathetic systems. Such systems are equally or more important for social minorities. Recently, more light is shed on sign language (Long et al., 2024; Yin et al., 2024; Wang et al., 2025), however, automatic emotion recognition in signers has not been explored at all. To our best knowledge, the single contribution in this direction is the EmoSign dataset for American Sign Language (ASL) (Chua et al., 2025).

In this paper, we introduce eJSL[1], a new benchmark dataset for emotion recognition in Japanese Sign Language JSL). We asked two signers to express 78 distinct sentences with each of seven different emotional states, resulting in 1,092 video clips. Because human languages are highly context-dependent, any linguistic expression potentially can be expressed with any emotion. In this dataset, thus, the task is recognizing the emotions of signing signers rather than that of signed contents.

Here, the arising challenge is that emotion expressions in signers are further complicated because facial expressions convey both grammatical and affective information (Brentari, 1999; Wilbur, 2000). For example, eyebrow movement can signal a yes/no question (Pfau and Quer, 2010) or express surprise (Valli and Lucas, 2000), creating ambiguity for emotion recognition models trained on non-signers.

To address the challenge, we investigate three hypotheses: (1) caption-based weakly labeled data can support effective model fine-tuning, (2) selecting temporal segments less affected by grammatical expressions improve accuracy, and (3) hand gesture features enhance recognition beyond facial features alone. Experiments on multiple datasets validate these hypotheses and offer insights into understanding of emotional communication in signers.

2 Emotion Recognition and Sign Language

As discussed, a unique challenge in emotion recognition in signers lies in the overlap between grammatical facial expressions (GFEs) and affective facial expressions (AFEs). Unlike spoken language, sign language uses non-manual markers such as facial movements and head gestures to encode syntax. These signals often occur simultaneously with AFEs, making their separation critical for accurate understanding of communicative information.

To this end, Silva et al. (2020) annotated their corpus with facial Action Units (AUs) to encode GFEs. However, this corpus is not annotated in terms of emotion. Although there are many other sign language datasets (see Table 2 of (Albanie et al., 2021) for a not-exaustive but rich list of 30

[1] https://released-but-anonymized-for-reviewing/

datasets), none of them are with emotion annotation except for EmoSign and our eJSL. However, both of them are small-scale benchmark-oriented datasets. Thus the scarcity of data available for supervised training is another challenge.

In human multimodal communication, verbal and non-verbal information can be independent, even contradictory in sentiment. In such contradictory situations, facial information can be highly dominant more than verbal information (Mehrabian, 1971). Nevertheless, in usual situations, it is repeatedly observed that textual information is dominant by the multimodal spoken language emotion recognition literature (Li et al., 2023; Yun et al., 2024). Therefore, we can expect that, even in sign language, textual caption/subtitle data (i.e., translations in spoken language) are useful to induce the emotion state of original signers in existing corpora. In this paper, we explore this possibility.

3 Datasets

We use three sign language datasets: eJSL, EmoSign, and BOBSL. We use eJSL and EmoSign for evaluation and BOBSL for both evaluation and neural model training. Although all three datasets are in different sign languages, as our focus is paralinguistic (or even non-linguistic) and we are at a very early stage of research, we assume the impact of the differences is marginal.[2]

3.1 eJSL

The eJSL (emotional Japanese Sign Language) dataset is our original video corpus containing 78 distinct utterances. As illustrated in Figure 1, each utterance performed by one male and one female signer across the six Ekman's basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) (Ekman, 1992) and the neutral state, yielding 1,092 clips in total. The signers are native JSL signers who work as vocational deaf actors. The signers can also read and write fluently in Japanese as well as non-signers. Thus all instructions and utterances were textually presented in Japanese.

Each clip is a complete JSL utterance with a single intended emotion. The 78 utterances were adopted from a public transcript[3] with substantial modifications in consultation with a professional sign language interpreter so that signers have less difficulties in uttering (e.g., replacing proper names with pronouns, avoiding onomatopoeic words, etc.).

3.2 EmoSign

EmoSign (Chua et al., 2025) is an ASL dataset of 200 clips drawn from the ASLLRP corpus (Neidle et al., 2022), designed for affective analysis. We use its *Single Expression Set* of 140 clips, which are labeled with a single dominant emotion. It covers ten emotion categories (for mapping to our label set, see Appendix A) and serves for model comparison, as Chua et al. (2025) provide established baselines using vision-capable large language models.

3.3 BOBSL

The BOBSL dataset (Albanie et al., 2021) contains over 1,460 hours of British Sign Language video data from BBC programs by 39 sign language interpreters. Using the official subtitle and alignment data from BOBSL, we derive a dataset by applying a textual emotion recognition (TER) model to subtitles, producing large-scale weak labels in seven basic emotions according to the steps below.

First we extract two base subsets: **BOBSL-A** from automatically subtitle-alignned data (113,826 clips), and **BOBSL-M** manually subtitle-aligned data (34,046 clips).

A portion of BOBSL-M is held out (1438 clips) and manually annotated by two English-speaking non-signers based on subtitles, with a high-confidence overlap set (**BOBSL-M_C**, 930 clips for testing) showing moderate-to-substantial agreement on emotion labels (see Appendix B) .

Finally, we apply a pre-trained TER model[4] to BOBSL-A, as we identified the model works best according to our preliminary verification using BOBSL-M_C. We refer to the resulting emotion-annotated dataset as **BOBSL-A_TEA** for training.

4 Experiments

In this section, we validate our three hypotheses: (1) TER on subtitles mitigates the data scarcity issue in sign language, (2) selecting temporal segments less affected by grammatical expressions

[2]We recognize that this is a strong assumption, and interlingual and intercultural differences will naturally have an impact, but we believe these will only emerge as issues once more research has progressed and recognition performance has improved. *When and how cultural variations limit this assumption* is an important research question for future work.

[3]https://github.com/mmorise/ita-corpus/blob/main/emotion_transcript_utf8.txt

[4]https://huggingface.co/michellejieli/emotion_text_classifier

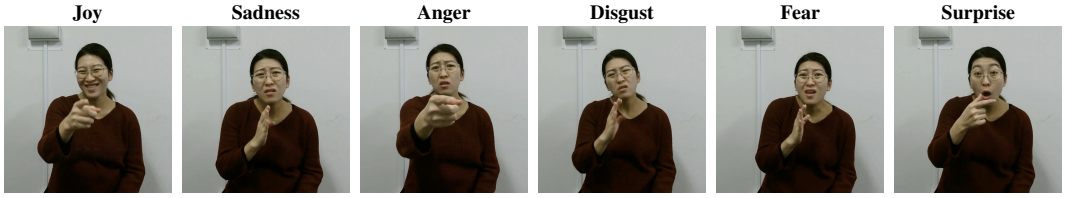

Figure 1: Examples of six different emotional expressions for the same utterance *"What? That's definitely a lie, right? Hurry and say it was a lie."* by a signer in the eJSL dataset.

improves accuracy, and (3) hand gesture features enhance recognition beyond facial features alone.

4.1 Emotion recognition models and metrics

Through our experiments, we adopt EMO-AffectNet (Ryumina et al., 2022) for video-based face emotion recognition (FER), with a minor extension to include hand gesture features. For our hand gesture extension, see Appendix C.

Ryumina et al. (2022) provide a comprehensive cross-corpus study covering eight emotion datasets. Their framework combines a ResNet-50 FER backbone, pretrained on VGGFace2, with temporal modeling modules using multiple data augmentation strategies and label balancing. As our primary baseline, we use their public model weights[5], which were optimized in such a way with non-signer data. Here after we refer to the plain Emo-AffectNet architecture as EAN and the hand gesture extended version as EANwH.

In accordance with the emotion recognition literature, we use weighted accuracy (wAcc) and macro F1 as performance metrics.

4.2 TER-based automatic data labeling

To validate the effectiveness of TER-based automated labeling on sign language datasets, we fine-tuned the baseline EAN model with the BOBSL-A_TEA datasets introduced in section 3.3.

As shown in Table 1 and Table 2, finetuning with BOBSL-A-TEA improved recognition performance significantly not only on BOBSL but also on eJSL, although the current overall performance is still quite low in comparison to that on non-signers. Nevertheless, the results support our hypothesis that TER-based weak labeling would mitigate the scarcity of sign language emotion recognition data.

[5] https://github.com/ElenaRyumina/EMO-AffectNetModel

Method	wAcc (%)	macro F1 (%)
EAN w/ non-signers data	15.54	12.12
EAN w/ BOBSL-A_TEA	**27.85**	**17.75**

Table 1: Performance of fine-tuning with TER-based labeling on BOBSL-M_C.

Method	wAcc (%)	macro F1 (%)
EAN w/ non-signers data	7.41	9.25
EAN w/ BOBSL-A-TEA	**15.11**	**12.11**

Table 2: Performance of fine-tuning with TER-based labeling on eJSL.

Method	wAcc (%)	macro F1 (%)
Full Clip Input	15.11	12.11
Random 2s Segment	15.20	12.29
Post-Signing 2s Segment	**23.17**	**19.26**

Table 3: Comparison of temporal segment selection strategies on eJSL.

4.3 Temporal segment selection

If GFEs really obscure affective cues, selecting non-signing temporal segments used for FER should improve the recognition performance. This has been theoretically expected but has not been verified quantitatively yet. Especially, by observation, post-signing segments seem to be emotionally salient, at least in acted eJSL.

Therefore, we compare the following three strategies: (1) using full clip, which is equivalent to the previous experiment settings for Table 1 and Table 2; (2) randomly selecting a 2-second segment in each clip; and (3) using the post-signing 2-second segment in each clip.[6]

As expected, the results shown in Table 3 confirm temporal segment selection of non-signing or emotionally salient segments is quite effective.

[6] We extracted the post-signing segment automatically using motion intensity, for we instructed signers to remain still for 3 seconds after signing each sentence.

Method	wAcc (%)	macro F1 (%)
EAN (full clip)	27.85	17.75
EANwH (full clip)	**32.72**	**20.03**

Table 4: Performance of EANwH on BOBSL-M_C.

Method	wAcc (%)	macro F1 (%)
EAN (full clip)	15.11	12.11
EAN (post 2s)	23.17	19.26
EANwH (full clip)	**24.63**	**21.09**

Table 5: Performance of EANwH on eJSL.

4.4 Incorporating hand motion

Hand motions are expected to serve cues for signing segments. Then, by incorporating hand features, a model would learn an effective way to attend only to non-signing moments. To confirm this possibility, we applied EANwH (see section 4.1), a hand-feature extended version of EAN.

As expected, the results shown in Table 4 and Table 5 confirm that incorporating hand features are effective both for BOBSL and eJSL. For eJSL, EANwH using full clips performs better than the post-signing segment selection with EAN.

4.5 Comparison to vision-capable LLMs

Finally, we compare our EANwH model to vision-capable LLMs (Qwen 2.5 and GPT-4o) using EmoSign (Chua et al., 2025). We applied the procedure presented in (Chua et al., 2025) and could reproduce mostly the same results with them.[7]

The results shown in Table 6 suggest that EANwH is better than the tested LLMs. Especially, EANwH has superior performance on Neutral. Appendix Table 8 shows that Neutral is the majority class in BOBSL-M_C, as is often the case in real-world data. Therefore, performance on the Neutral class may strongly influence users' perceived performance in practical applications.

Table 7 shows the evaluation results on eJSL with the same procedure. EANwH obtained the best results, consistently with the test on BOBSL-M_C shown in Table 6.

4.6 Discussion

Statistical tests on eJSL confirm that EAN w/ BOBSL-A-TEA is significantly better than EAN w/ non-signers ($p < 0.0001$, Table 2), and EANwH (full clip) is significantly better than EAN (full clip) ($p < 0.0001$, Table 5). However, the gain from EAN (post 2s) to EANwH (full clip) is not significant.

While we observed gains from the simplest baseline, the achieved overall performance is still very limited.[8] However, as EANwH, our hand motion-enhanced version, is also quite naive, there should be much room for technical improvements.

Utilization of existing resource will also enhance performance. While this paper utilized annotated data in a cross-lingual setting, use of more datasets from the same sign language in training will improve results.

Fundamentally both EAN and EANwH do not understand signed linguistic content in utterances. As discussed in section 2, linguistic content can serve strong emotion indicators in usual situations. Thus, integration with sign language understanding also must be be explored.

Model	Joy	Sad.	Ang.	Dis.	Fear	Sur.	Neu.	Total
Qwen2.5	39.18	4.26	**28.57**	0.00	0.00	17.65	10.17	14.26
GPT-4o	38.38	**27.27**	0.00	**28.57**	8.33	0.00	0.00	14.65
EANwH	30.99	16.67	26.67	8.33	**10.53**	0.00	**25.00**	**16.88**

Table 6: Per-class F1 and overall macro F1 scores of vision-capable LLMs and EANwH on EmoSign.

Model	Joy	Sad.	Ang.	Dis.	Fear	Sur.	Neu.	Total
Qwen2.5	20.91	**11.98**	2.53	12.10	9.57	1.27	19.84	11.17
GPT-4o	7.38	4.64	**15.93**	**23.79**	8.61	11.00	6.67	11.15
EANwH	**35.91**	10.64	15.55	14.29	**9.65**	**21.10**	**40.49**	**21.09**

Table 7: Per-class F1 and overall macro F1 scores of vision-capable LLMs and EANwH on eJSL.

5 Conclusion

To push forward the research on emotion recognition in signers, this paper introduced a new sign language benchmark dataset eJSL, in which two JSL signers acted seven emotions for 78 utterances.

With eJSL and other two datasets, i.e., BOBSL and EmoSign, we empirically demonstrated effectiveness of textual emotion recognition, temporal segment selection and hand motion. We hope our eJSL and findings contribute emotion recognition in signers and sign language.

[7] Only a few clips were differently classified from the results reported in their confusion matrices. We set the temperature parameter 0.

[8] We sampled 70 clips (10 per class) of one signer and asked the other signer to classify them. The achieved macro F1 score was 77.78, while a non-signer achieved 57.85 on the same 70 clips. The former would be the upper-bound and the latter would be the lower-bound for practical applications.

6 Limitations

As discussed in section 4.6, the current achieved performance of emotion recognition in signers is very limited. Therefore, the findings in this paper may not be applicable after the performance is significantly improved in future.

Our eJSL dataset contains only two JSL signers. The data may not be representative of the JSL community. In addition, the data are acted and may be different from real spontaneous data. (Note that, however, the current standard emotion recognition datasets in English (Busso et al., 2008; Poria et al., 2019) are also acted.)

eJSL marks the first step in research into emotion recognition associated with sign languages (especially JSL), and has a certain significance as a benchmark. However, since it only includes two signers, it will be essential to use it in conjunction with other datasets. Increasing the number of signers of eJSL itself is also an important future challenge.

The dataset may be used for purposes other than benchmarking, such as analysis by linguists, but should not be used to train models without careful consideration.

7 Ethical Considerations

The eJSL data collection was conducted in February 2025 after obtaining the signers' consents using the standard consent form of our institute. In accordance with the institutional ethics committee's ethical review requirements checklist, the data collection was exempt from full ethical review in advance. The interpreter and signers were appropriately compensated in accordance with the institutional regulations.

The eJSL dataset has been reviewed by two signatories prior to its release. It is available for loan on the condition that it is not redistributed and is used for research purposes only.

References

Albanie, S., Varol, G., Momeni, L., Afouras, T., Ma, X., Wang, Y., Chung, J. S., Bear, H., Hain, T., Cox, S., Buehler, P., and Zisserman, A. (2021). BBC-Oxford British sign language dataset. *arXiv preprint arXiv:2111.03635*.

Brentari, D. (1999). *A Prosodic Model of Sign Language Phonology*. The MIT Press.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Chua, P., Fang, C. M., Ohkawa, T., Kushalnagar, R., Nanayakkara, S., and Maes, P. (2025). EmoSign: A multimodal dataset for understanding emotions in american sign language.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Li, D., Wang, Y., Funakoshi, K., and Okumura, M. (2023). Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Long, Z., Liu, X., Qiao, J., and Li, Z. (2024). Sign language recognition based on facial expression and hand skeleton. In *Proceedings of the International Conference on Automation and Artificial Intelligence*. Southeast University. Available at: https://arxiv.org/abs/2407.02241.

Mehrabian, A. (1971). *Silent Messages*. Wadsworth, Belmont, CA.

Neidle, C., Opoku, A., and Metaxas, D. (2022). Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp).

Pfau, R. and Quer, J. (2010). *Nonmanuals: their grammatical and prosodic roles*, page 381–402. Cambridge Language Surveys. Cambridge University Press.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Ryumina, E., Dresvyanskiy, D., and Karpov, A. (2022). In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514:435–450.

Silva, E. P. d., Costa, P. D. P., Kumada, K. M. O., De Martino, J. M., and Florentino, G. A. (2020). Recognition of affective and grammatical facial expressions: A study for brazilian sign language. In *ECCV 2020 Workshops*, pages 218–236. Springer.

Valli, C. and Lucas, C. (2000). *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington, D.C.

Wang, Z., Li, D., Jiang, R., and Okumura, M. (2025). Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. *IEEE Access*, 13:5491–5506.

Wilbur, R. (2000). Phonological and prosodic layering of nonmanuals in american sign language. In *Sign Language & Linguistics*.

Yin, K., Regier, T., and Klein, D. (2024). American sign language handshapes reflect pressures for communicative efficiency. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15715–15724, Bangkok, Thailand. Association for Computational Linguistics.

Yun, T., Lim, H., Lee, J., and Song, M. (2024). Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95, Mexico City, Mexico. Association for Computational Linguistics.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

A Mapping from EmoSign to Ekman's Basic Emotions

We map EmoSign surpris_ pos and surprise_neg to surprise, worry to fear and frustration to sadness, based on semantic similarity (Russell, 1980). Table 8 shows this mapping and the original counts.

eJSL (Ekman)	EmoSign	Count
Joy	Happyness	54
Sadness	Sadness	10
	Frustration	19
Anger	Anger	3
Disgust	Disgust	10
Fear	Fear	7
	Worry	14
Surprise	Surprise_pos	5
	Surprise_neg	7
Neutral	Neutral	11

Table 8: Emotion distribution of the EmoSign single expression set (N=140) and mapping to Ekman's basic emotions.

B Manual Emotion Annotation on BOBSL subtitles

In the BOBSL-M subset, we manually annotated a selected subset of 1,438 clips for emotion labels using two independent annotators, both of whom labeled the same set of video segments (1 male and 1 female students who graduated universities in North America). Based on these annotations, we created two subsets: **BOBSL-M_A1** and **BOBSL-M_A2**, corresponding to the individual annotations from each annotator. The intersection of segments where both annotators provided consistent labels forms a high-confidence subset named **BOBSL-M_C** (Table 9).

Annotation Instructions The annotators were instructed as follows:

```
Task description: Use 7 emotion labels to annotate sentences in several text document. Each sentence can only correspond to exactly 1 emotion label. Use the annotation tool to label sentences.

Emotion category: Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise. For the "Neutral" label, it is used for the sentence that does not have an obvious emotion.
```

Emotion	M_A1	M_A2	M_C
Joy	59	251	48
Sadness	37	110	25
Anger	35	92	26
Disgust	19	55	10
Fear	21	33	5
Surprise	34	47	8
Neutral	1233	850	808
Total	1438	1438	930

Table 9: Number of instances per emotion of BOBSL-M subsets.

```
How to use the labeling tool: The tool shows the sentence to be annotated and its context, determine the emotion of the sentence to be annotated with its context. When labeling, each emotion maps to a key, just press a key to do the corresponding labeling: 'a': 'Anger', 'd': 'Disgust', 'f': 'Fear', 'j': 'Joy', 'n':'Neutral', 's': 'Sadness', 'u': 'Surprise.

Examples for each emotion category:
* The quoted sentences are from the internet.
* The unquoted sentences are from the dataset.
1. Anger:
"I can't believe you did that! How could you be so careless?"
What the fuck is wrong with you?!
2. Disgust:
"The way they treated those poor animals is revolting."
Oh, horrible.
3. Fear:
"I'm really scared about what might happen next. This is terrifying."
You must be a nightmare to live with.
4. Joy:
"I'm so happy! This is the best news I've heard all day!"
Everyone was cheering, clapping. 5. Neutral:
"I went to the store today and bought some groceries."
It's one of the oldest and grandest houses in Henrietta Street, built in 1743.
6. Sadness:
"I'm really feeling down today. Everything seems so hopeless."
I regret to inform you, Mr Keys, that Thomas was killed this morning, in Iraq, in the line of duty.
7. Surprise:
"Wow, I didn't see that coming! What a shock!"
```

`I just can't believe it, just out there, Daddy, is the inner city London.`

Agreement between M_A1 and M_A2 To evaluate the annotation consistency between the two annotators, we computed the Gwet's AC1 (Gwet, 2008) as a robust measure of inter-rater agreement. Compared to Cohen's Kappa (Cohen, 1960), AC1 is less sensitive to category imbalance and prevalence issues, making it more appropriate for our dataset, where the neutral class dominates the distribution (Feinstein and Cicchetti, 1990). The observed agreement (P_o) was 0.6467, and the expected agreement by chance (P_e) was 0.0762. Using the formula:

$$\text{AC1} = \frac{P_o - P_e}{1 - P_e}$$

we obtained a Gwet's AC1 value of 0.6176, indicating moderate to substantial agreement between annotators. This level of consistency supports the reliability of the overlapping subset **BOBSL-M_C**, which is used as a high-confidence evaluation set in our experiments.

C EMO-AffectNet (EAN) and EANwH

We extend EMO-AffectNet (Ryumina et al., 2022) to incorporate hand motion as shown Figure 2.

C.1 Feature Extraction

In our model, we extract modality-specific features from both facial images and hand skeletal data.

Facial Feature Extraction. For facial features, we adopt the same methodology as described in the large-scale visual cross-corpus study by Ryumina et al. (2022). Specifically, each face image, cropped to 224×224 resolution using MTCNN, is passed through a ResNet-50 backbone pretrained on VGGFace2. The output is a 512-dimensional embedding extracted from the global average pooling (GAP) layer before the final classification head. This representation captures rich identity-independent emotional features and has been shown to generalize well across datasets with varying demographics and acquisition conditions. The extracted features are stored frame-by-frame as a temporal sequence of fixed-length vectors for downstream sequence modeling.

Hand Feature Extraction. For hand features, we follow the approach proposed in the sign language recognition model by Long et al. (2024). From each frame, we obtain 21 hand keypoints per hand (totaling 42 keypoints) using the MediaPipe Hands pipeline. These keypoints are represented as 2D coordinates and normalized relative to the wrist joint to ensure translation invariance. We also apply coordinate transformation to align the hand pose into a canonical hand-centered coordinate system, as described in their work. This process effectively reduces spatial variance and emphasizes articulation differences across signs. The final hand representation for each frame is a 42×2 feature matrix, which is flattened and stored as part of the temporal input sequence.

C.2 Feature Synchronization and Fusion.

To effectively integrate facial and hand-derived features, we adopt an early fusion strategy at the frame level. For each frame in the video, the 512-dimensional facial feature vector extracted from the ResNet-50 backbone is concatenated with the flattened 84-dimensional hand skeleton vector (21 keypoints × 2D), resulting in a unified 596-dimensional feature vector. This frame-level concatenation preserves temporal alignment between the two modalities and enables the model to capture low-level interactions between facial expressions and hand gestures.

The sequence of fused multimodal vectors is then passed into a temporal modeling module, which captures the temporal dependencies and emotional dynamics across the video using two LSTM layers of 512 and 256 hidden units, resulting in about 300M parameters. This early fusion design allows for efficient joint modeling of modality-specific and cross-modal patterns without requiring complex attention-based alignment mechanisms. It also ensures robustness against partial modality noise, as both facial and skeletal information are encoded into a shared temporal embedding space from the outset.

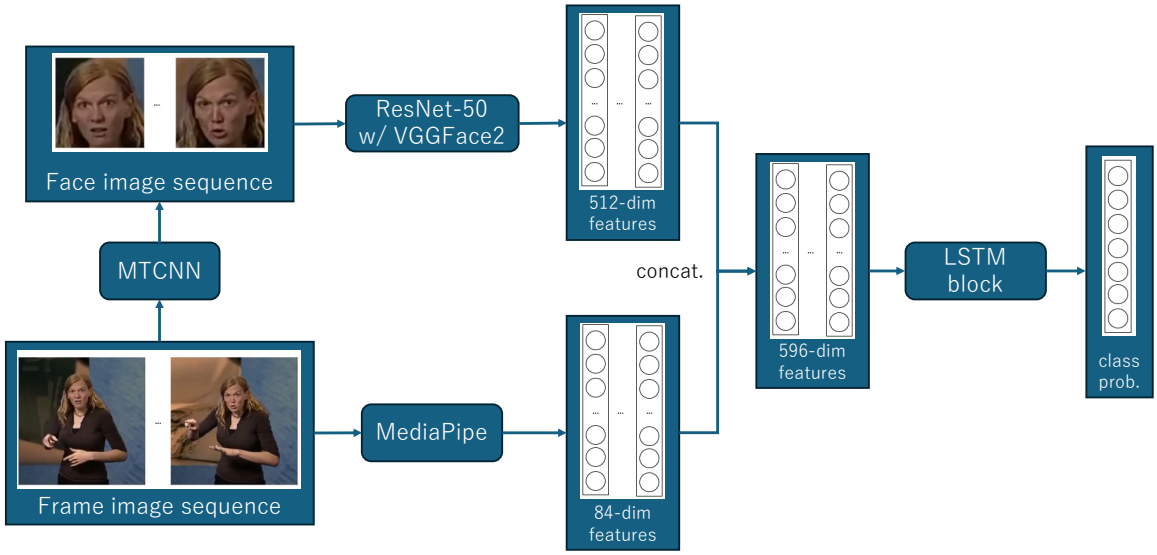

Figure 2: EANwH model architecture using both facial and hand features.