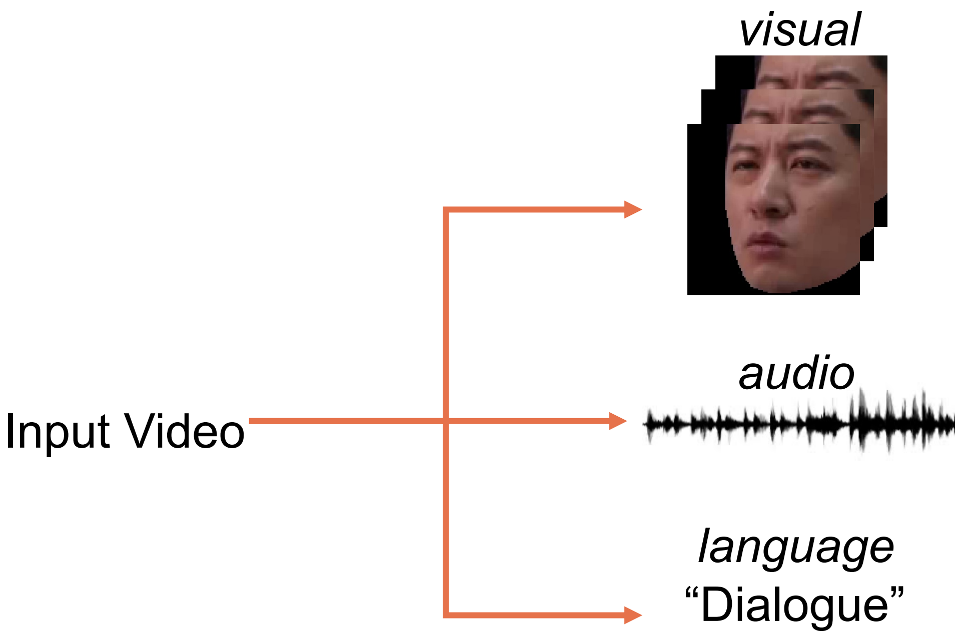
Semi-Supervised Multimodal Emotion Recognition with Expression MAE

Multimodal Intelligent Perception System Lab
MIPS-Lab

Motivation – Multimodal Learning

Multimodal Emotion Recognition Challenge 2023 (MER2023):

- **MER2023:** The objective of *MER2023* is to investigate emotion recognition using *audio, language, and visual signals*, thus enhancing the *robustness* of affective computing.
- **MER-SEMI:** This track provides large amounts of unlabeled video samples, encouraging participants to leverage *semi-supervised learning* to improve emotion recognition performance.



Partition	# of samples labeled	# of samples unlabeled	Duration
Train&Val	3373	0	03:45:47
MER-MULTI	411	0	00:28:09
MER-NOISE	412	0	00:26:23
MER-SEMI	834	73148	67:41:24

unbalanced data

Motivation – Semi-Supervised Learning

Class Imbalance:

➢ Classes in the training set are extremely imbalanced. Making some classes hard to learn by the model.

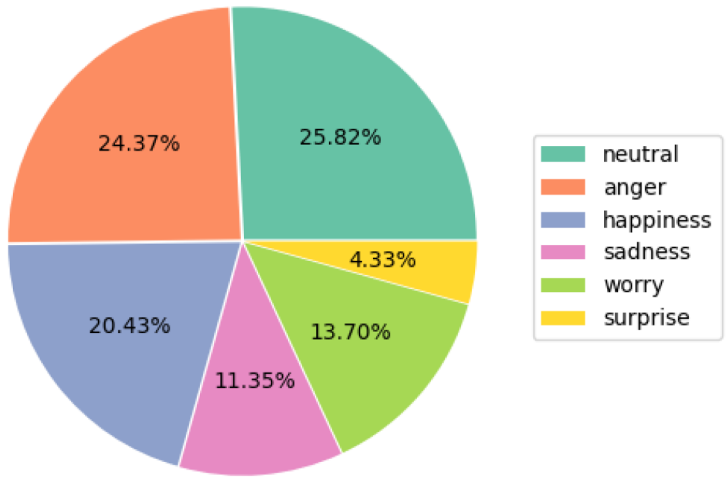
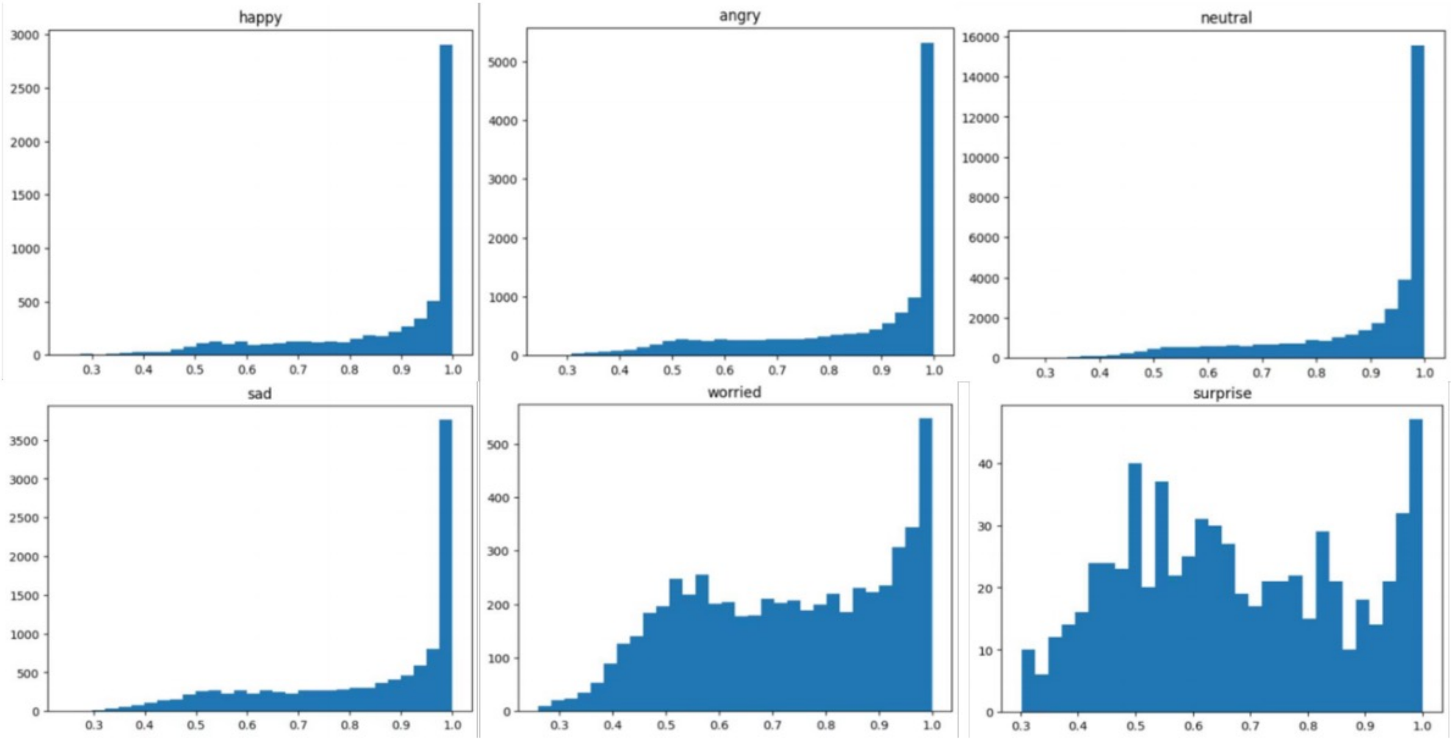


Figure 3: Distribution of discrete emotions (Train&Val)

Pre-experiment



Class of worried and surprise are hard to learn by the baseline model, with a lower confidence score.

Related Works

Multimodal Feature Extractor:

***Visual*:**

- 1. MAE: containing an Encoder-Decoder structure, are a type of ***self-supervised*** learners for computer vision. Once the encoder has been trained, it can be directly reused for downstream tasks.
- 2. VideoMAE: is designed to process video inputs and apply a tube masking strategy to prevent inadvertent feature information leakage, effectively deriving ***dynamic visual features***.

Language:

- 3. MacBERT: mitigates the gap between the pre-training and fine-tuning stages by masking a word with a similar word.

Audio:

- 4. HuBERT: introduces a self-supervised approach with an unsupervised clustering step, addressing problems in the acoustic field through masked prediction of hidden units.

Cross Modality:

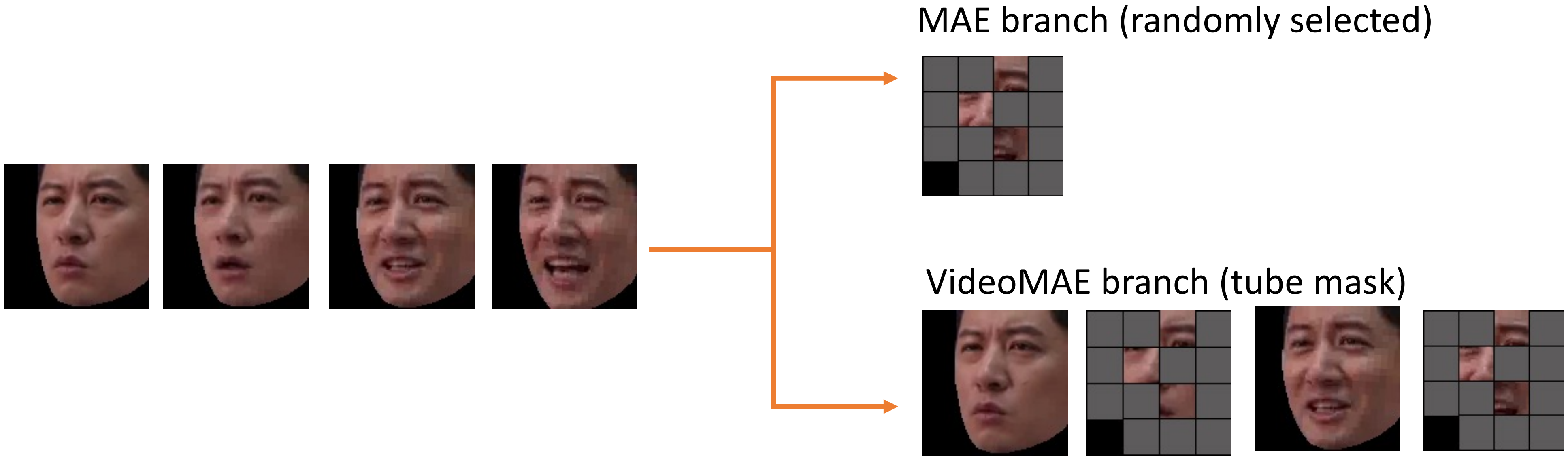
- 5: CLIP: is pre-trained on a large dataset model for matching images and text modalities.
- 6: Tacotron-Var: is a pre-trained speech synthesis model to integrate text and speech features.

Pipeline – Expression MAE

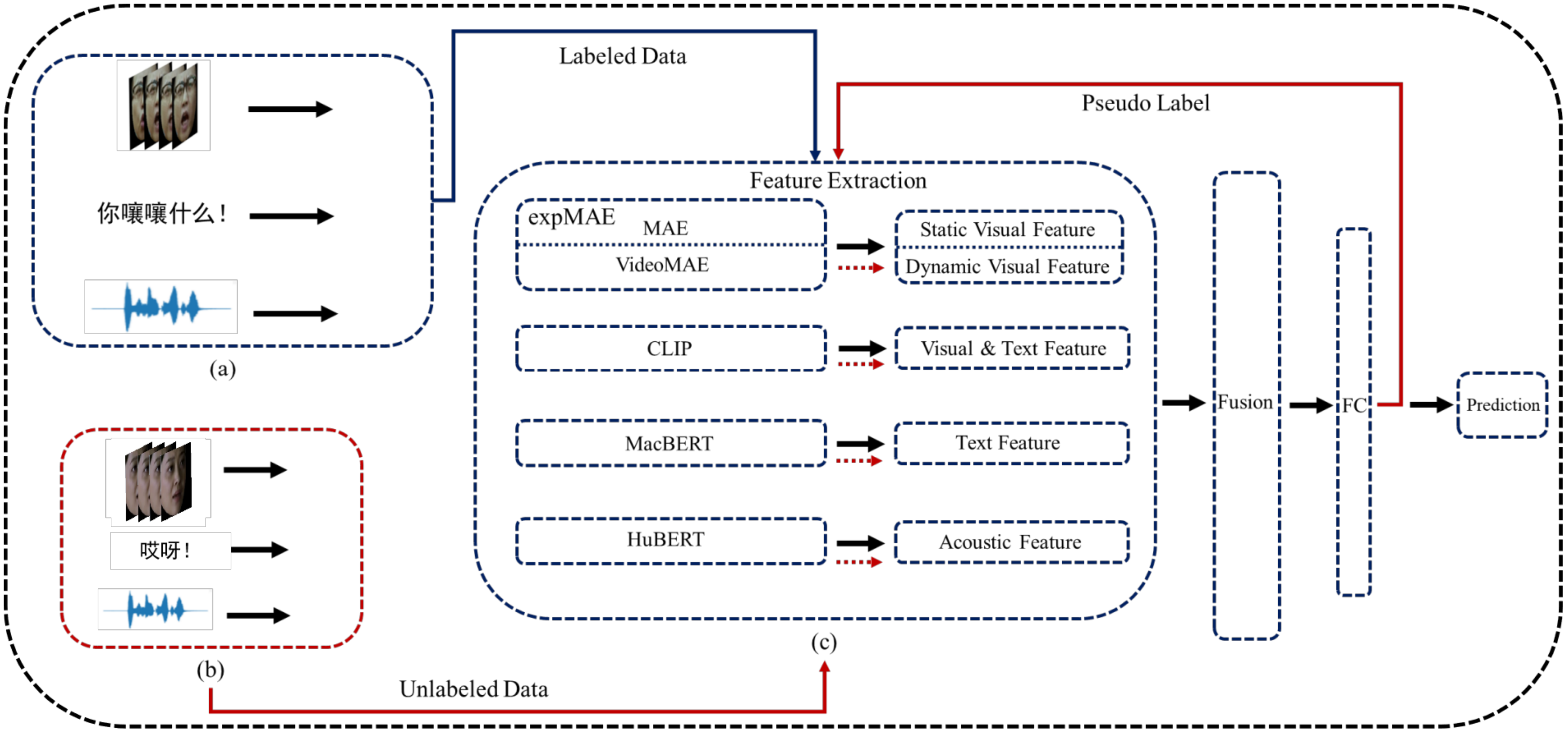
Expression MAE (expMAE):

- 1. The MAE model's limitations become apparent as it can only glean ***static features***, not accommodating changes in facial expressions during the progression of a video.
- 2. VideoMAE is designed to process video inputs and apply a tube masking strategy to prevent inadvertent feature information leakage, effectively deriving ***dynamic visual features***.

By combing MAE and VideoMAE, we introduce expression MAE (expMAE).



Pipeline - Overall



Baseline Experiments

Metric(e): emotion label
Metric(v): Valence
Metric: metric(e) – 0.25*metric(v)

Table 1: Unimodal results of the baseline.

Feature	Train&Val		
	$metric_e$ (↑)	$metric_v$ (↓)	$metric$ (↑)
Acoustic Modality			
HuBERT-base[7]	60.72	1.53	0.22
HuBERT-large[7]	**65.67**	**1.27**	**0.34**
Lexical Modality			
MacBERT-base[4]	40.96	2.42	-0.19
MacBERT-large[4]	**42.62**	**2.39**	**-0.17**
Visual Modality			
MANet-RAFDB[18]	57.48	1.38	0.23
DFER[17]	43.63	2.02	-0.06
MAE [6]	60.01	1.42	0.25
VideoMAE [14]	61.98	1.33	0.28
expMAE	**62.56**	**1.29**	**0.30**
Cross Modality			
Tacotron-Var [16]	44.01	2.44	-0.17
CLIP [11]	**60.99**	**1.26**	**0.29**

Pipeline – Fusion block

- 1. After the baseline experiments, we use the best unimodal result model for multimodal feature fusion.
- 2. we apply the factorized bilinear pooling (FBP) module to fuse each contextually related feature, generating the fused features

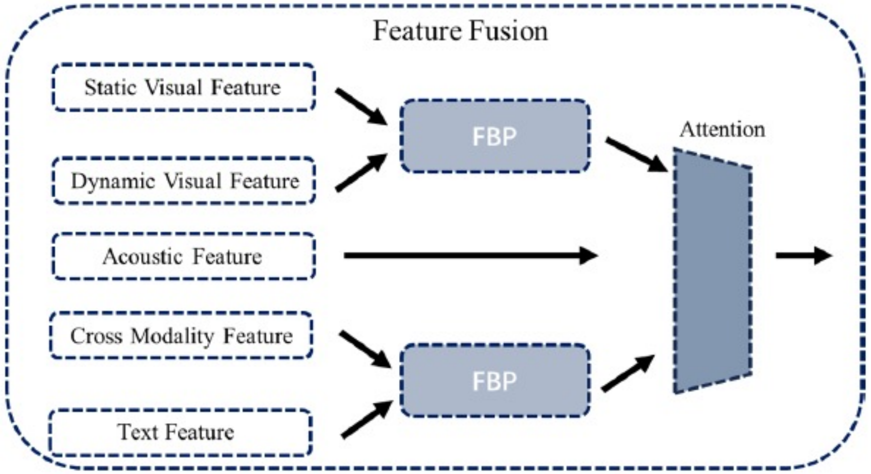


Figure 2: Illustration of the Fusion block in Figure 1.

Multimodal Experiments

A: Acoustic
L: Lexical
V: Visual
C: Cross modality
HL: HuBERT-large
ML: MacBERT-large
MR: MANet-RAFDB
T-Var: Tacotron-Var

A	L	V	C	Train&Val $metric_e$ (↑)	$metric_v$ (↓)	$metric$ (↑)
Bimodal Results						
HL	ML	—	—	67.02	1.16	0.38
HL	—	MR	—	72.92	0.86	0.52
HL	—	MAE	—	71.32	0.89	0.49
HL	—	VideoMAE	—	72.90	0.81	0.52
HL	—	expMAE	—	**73.40**	**0.78**	**0.53**
HL	—	—	CLIP	71.32	0.79	0.51
HL	—	—	T-Var	65.24	1.19	0.35
—	ML	MR	—	61.19	1.28	0.29
—	ML	MAE	—	63.66	1.32	0.30
—	ML	VideoMAE	—	66.70	1.12	0.39
—	ML	expMAE	—	**67.13**	**1.10**	**0.39**
—	ML	—	CLIP	64.14	1.12	0.35
—	ML	—	T-Var	53.64	2.19	-0.01
Trimodal Results						
HL	ML	MR	—	73.39	0.87	0.52
HL	ML	MAE	—	73.8	0.92	0.49
HL	ML	VideoMAE	—	72.42	0.78	0.53
HL	ML	expMAE	—	**74.52**	**0.76**	**0.55**
Multi Results						
HL	ML	MAE	T-Var	73.07	0.84	0.518
HL	ML	VideoMAE	T-Var	74.52	0.77	0.552
HL	ML	expMAE	T-Var	74.65	0.76	0.556
HL	ML	MAE	CLIP	72.65	0.89	0.502
HL	ML	VideoMAE	CLIP	74.78	0.74	0.561
HL	ML	expMAE	CLIP	**75.01**	**0.66**	**0.585**

. FBP is the best classifier in MER-SEMI

Table 3: Comparison of different classifiers.

Classifier	Train&Val $metric_e$ (↑)	$metric_v$ (↓)	$metric$ (↑)
SVM	61.70	—	—
Transformer (3 layers)	73.55	0.89	0.512
Transformer (6 layers)	72.14	0.88	0.499
naive attention	75.01	**0.66**	**0.585**
FBP	**75.53**	0.82	0.550

Semi-supervised Learning Strategy

To mitigate the impact of skewed class distribution on the classifier, we introduced two data augmentation techniques:

- 1. **Threshold-based reliable labeling:** samples with pseudo-label confidence exceeding 0.8 were filtered and added to the training set.

- 2. **Threshold-based weak reinforcement**: only added the unbalanced classes, since they are hard to learn by model.

Final Experiments

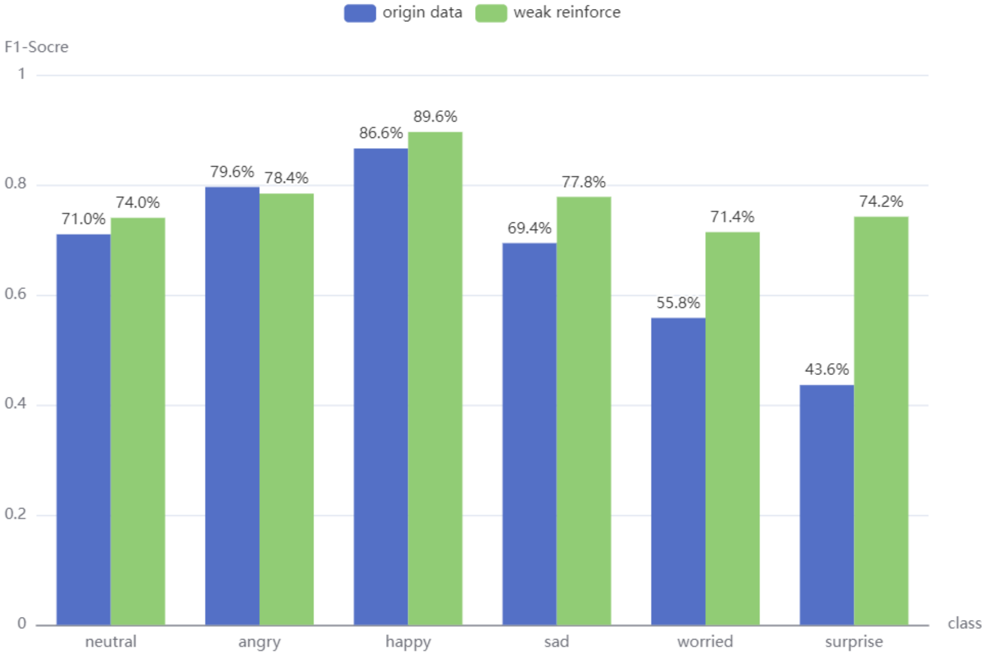
MIPS-Lab

By adding our semi-supervised strategy, we achieve the best results for our model, ranking 2 at the MER-SEMI challenge.

Table 4: Comparison of semi-supervised strategy, MER-SEMI shows the test result on the MER-SEMI dataset.

Augmentation	Train&Val			MER-SEMI
	$metric_e$ (↑)	$metric_v$ (↓)	$metric$ (↑)	$metric_e$ (↑)
baseline	75.53	0.82	0.550	0.8799
threshold (reliable label)	76.01	0.75	0.570	0.8759
threshold (weak reinforce)	**78.10**	**0.64**	**0.622**	**0.8855**

Performance after the semi-supervised strategy.



Future Works

We would explore two aspects for future direction. Firstly, a stronger semi-supervised training strategy may be utilized in the task, such as Multi-view Learning, Network Embedding.

Secondly, it may be interesting to finetune both the encoder (such as expMAE, MacBERT, HuBERT) and the classifier (such as the fusion module) together rather than only train on the decoder.