Supplementary Material for Multimodal Emotion Regression with Multi-Objective Optimization and VAD-Aware Audio Modeling for the 10th ABAW EMI Track

Anonymous CVPR submission

Paper ID 20

A. Additional Ablation Studies

A.1. Ablation on Temporal Enhancement Components

We further investigate whether more sophisticated temporal enhancement modules can improve performance on top of our pretrained multimodal feature pipeline. Starting from the concatenation-based fusion setting, we separately introduce three temporal components: Temporal Convolutional Network (TCN), temporal BiGRU, and attention pooling. The results are reported in Table 1.

The results show that *Baseline + TCN* achieves 0.466014, *Baseline + temporal BiGRU* achieves 0.463559, and *Baseline + Attention Pooling* achieves 0.461779. None of these variants surpass our final model. This observation suggests that, under our pretrained-feature regime, the bottleneck does not lie in insufficient temporal modeling. Instead, additional temporal enhancement may introduce redundant transformations and optimization interference, thereby degrading the effectiveness of the original pretrained representations.

A.2. Attempts on Textual Triplet Loss and Middle-layer Prediction

In addition, we examine two alternative design choices beyond the main framework. The first is *Ours + tri loss*, which follows prior work and introduces a text-guided triplet loss for the textual modality. As shown in Table 2, this variant yields 0.444169, substantially lower than the performance of *Ours* (0.478567). This result indicates that the triplet-based contrastive objective does not provide useful supervision for our continuous EMI regression task, and may even distort the optimization target.

The second variant is *Ours + mid repr*, motivated by recent studies suggesting that intermediate-layer representations may contain richer semantic information. In this setting, intermediate features are directly used for prediction. This variant achieves 0.472401, which is competitive but still below our final model. Therefore, while intermediate-layer prediction preserves part of the discriminative information, it does not lead to better performance than our default design.

Table 1. Ablation on temporal enhancement components on the validation set.

Method	Validation Score
Baseline + Temporal Enhancement	0.441066
Baseline + TCN	0.466014
Baseline + Temporal BiGRU	0.463559
Baseline + Attention Pooling	0.461779

Table 2. Attempts on textual triplet loss and middle-layer prediction on the validation set.

Method	Validation Score
Ours	0.478567
Ours + tri loss	0.444169
Ours + mid repr	0.472401

Overall, these supplementary experiments further confirm that our final framework provides the best trade-off between effectiveness and stability. The strongest performance is obtained without extra temporal enhancement, triplet supervision, or intermediate-layer prediction, supporting our conclusion that, given strong pretrained features, carefully designed fusion and optimization are more important than adding extra temporal or auxiliary modules.

B. Comparison with Other Participating Teams

We further compare our method with other participating teams, as shown in Table 3. Our method obtains a test score of 0.674, which is very close to *USTC-IAT-United* (0.680), demonstrating competitive performance on the final evalu-

Table 3. Comparison with the results of the 8th participating teams.

Team	Validation Set	Test Set
HCAI-VIS	0.716	0.719
USTC-IAT-United	0.513	0.680
Ours	0.479	0.674

ation set.

While our validation score (0.479) is lower than those of *HCAI-VIS* (0.716) and *USTC-IAT-United* (0.513), the much smaller gap on the test set indicates that our model maintains good transferability to unseen data. In particular, the difference between our method and *USTC-IAT-United* is only 0.006 on the test set, showing that our approach remains competitive under the final benchmark setting. Nevertheless, the performance gap to *HCAI-VIS* suggests that further improvements are still needed in terms of robustness and ranking consistency across different data splits.