

# Supplementary Materials: 2M-AF: A Strong Multi-Modality Framework For Human Action Quality Assessment with Self-supervised Representation Learning

Anonymous Authors

## 1 DETAILS OF THE EXPERIMENTS.

**Settings for each dataset.** We implement our 2M-AF using the PyTorch framework. One NVIDIA RTX 4090 GPU are used to accelerate training. Overall, for SME-GCN, the learning rate is  $1e^{-3}$ , and the temperature parameter  $\tau$  is set as 1. And for two modality streams, the numbers of the output channels are selected as  $C^r = 400$  and  $C^s = 256$ . Similar to other methods, we fix the weights of the pre-trained I3D model. The hyperparameters of 2M-AF include the number of clips  $n$  for I3D, the proportions of the loss functions  $\lambda_1$  and  $\lambda_2$ , as well as the training parameters such as learning rate (lr), and weight decay (wd). To pre-train the single I3D ConvNet, the dropout is randomly set from 0.4 to 0.7 (Especially, on the 'Sync. 10m' of the AQA-7 dataset, the dropout is set as 0.1).

**a. Settings in 'Compare with State-of-the-art'.** For different datasets, our model configurations are shown in Table 1. Additionally, for the variable-length MMFS-63 dataset, we process videos to 128 frames by filling in zeros for missed frames and randomly cropping the extra frames.

**b. Settings in 'Ablation study'.** In the ablation experiments, in addition to the aforementioned configuration, we also conducted experiments using all categories of AQA-7 data. The configuration we used was as follows: we set  $n = 10$ ,  $\lambda_1 = \lambda_2 = 0.5$ , learning rate as 0.0005 and weight decay is set as 0.5.

**c. The effects of  $k$ .** Our variance mask works well for  $k$  values 1-5, as evidenced by AQA-7 results in Table 2, and we set  $k=2$  for the best performance.

**Download links of the datasets.** For the AQA-7 and UNLV-diving datasets, the data can be obtained from [rtis.oit.unlv.edu/datasets.htm](http://rtis.oit.unlv.edu/datasets.htm). And for the MMFS-63 dataset, the data can be obtained from [github.com/dirReno/MMFS](https://github.com/dirReno/MMFS).

Table 1: Hyperparameters settings on three datasets.

Dataset	$n$	$\lambda_1$	$\lambda_2$	lr	wd
<b>AQA-7</b>					
Diving	10	0.5	0.5	0.0001	0.00001
Gym Vault	10	0.5	0.5	0.0001	0.00001
BigSki	10	0.5	0.5	0.0005	0.00001
BigSnow	10	0.5	0.5	0.0001	0.00001
Sync. 3m	10	0.5	0.5	0.0001	0.00001
Sync. 10m	10	0.6	0.4	0.0001	0.00001
UNLV-Diving	14	0.5	0.5	0.001	0.0005
MMFS-63	10	0.4	0.6	0.001	0.5

Table 2: Different setting of  $k$  on the AQA-7 dataset.

$k$	1	2	3	5	10
Acc	0.8840	0.8848	0.8825	0.8803	0.8770

## 2 THE CONFUSION MATRIXES OF PFM

To demonstrate the selection ability of PFM, we plotted the confusion matrices for the AQA-7 and UNLV-Diving datasets. We define the positive sample ('P') as the skeleton modality, while the negative sample ('N') is the RGB modality. As the results are shown in Figure 1, on the AQA-7 dataset, 90% sample is correctly selected (48% skeleton and 42% RGB). On the UNLV-Diving dataset, 96% data is correctly selected (7% skeleton and 89% RGB). The experimental results show that, whether in cases where the two modalities are similar (AQA-7) or have significant differences (UNLV-Diving), PFM accurately selects the more suitable modality for each sample.

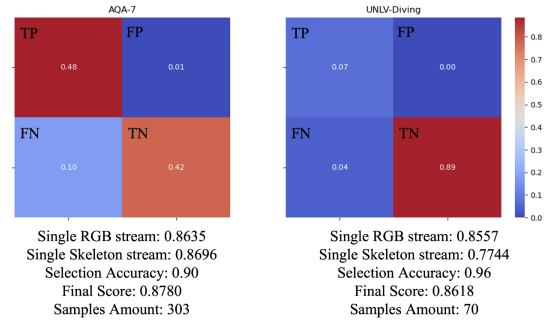


Figure 1: The plot of the confusion matrices of PFM, where 'final score' is the accuracy of the assessment. We adopt the original CTR-GCN+I3D+PFM configuration to show the full performance of PFM.