MCMoE: Completing Missing Modalities with Mixture of Experts for Incomplete Multimodal Action Quality Assessment

Huangbiao Xu[1,2], **Huanqi Wu**[1,2], **Xiao Ke**[1,2*], **Junyi Wu**[1,2], **Rui Xu**[1,2], **Jinglin Xu**[3]

[1]Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350116, China
[3]School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
kex@fzu.edu.cn,{huangbiaoxu.chn, wuhuanqi135, xurui.ryan.chn, xujinglinlove}@gmail.com, junyi.wu-1@outlook.com

Abstract

Multimodal Action Quality Assessment (AQA) has recently emerged as a promising paradigm. By leveraging complementary information across shared contextual cues, it enhances the discriminative evaluation of subtle intra-class variations in highly similar action sequences. However, partial modalities are frequently unavailable at the inference stage in reality. The absence of any modality often renders existing multimodal models inoperable. Furthermore, it triggers catastrophic performance degradation due to interruptions in cross-modal interactions. To address this issue, we propose a novel Missing Completion Framework with Mixture of Experts (MCMoE) that unifies unimodal and joint representation learning in single-stage training. Specifically, we propose an adaptive gated modality generator that dynamically fuses available information to reconstruct missing modalities. We then design modality experts to learn unimodal knowledge and dynamically mix the knowledge of all experts to extract cross-modal joint representations. With a mixture of experts, missing modalities are further refined and complemented. Finally, in the training phase, we mine the complete multimodal features and unimodal expert knowledge to guide modality generation and generation-based joint representation extraction. Extensive experiments demonstrate that our MCMoE achieves state-of-the-art results in both complete and incomplete multimodal learning on three public AQA benchmarks.

Code — https://github.com/XuHuangbiao/MCMoE
Extended version — https://arxiv.org/abs/2511.17397

Introduction

Action quality assessment (AQA) has gained attention for its objective evaluation of action execution proficiency, with wide applications in sports, rehabilitation (Ding, Xu, and Li 2023; Bruce et al. 2024), and skill determination (Xu et al. 2025b). Multimodal AQA (Xu et al. 2024a, 2025c; Zeng and Zheng 2024) has emerged as a promising paradigm beyond skeleton-based (Pan, Gao, and Zheng 2019; Bruce et al. 2024) and vision-based (Ke et al. 2024; Xu et al. 2024c,b; Xu, Yin, and Peng 2025) methods. By leveraging complementary information from temporally aligned modalities, it better discriminates subtle intra-class variations in highly similar actions through enriched contextual cues.

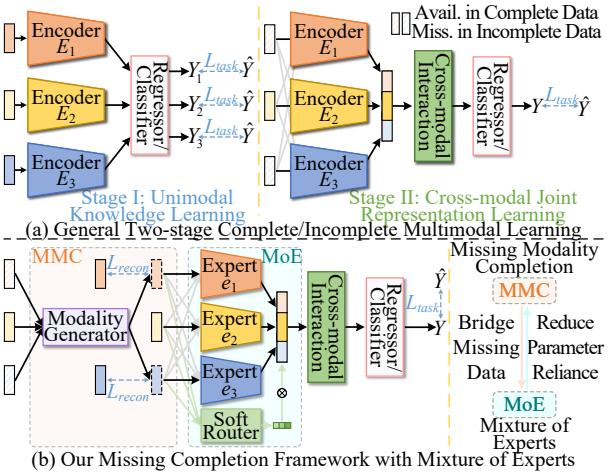


Figure 1: (a) Existing two-stage methods first learn unimodal features from complete multimodal data and then model cross-modal representations to address missing data, leading to higher training cost and complexity. (b) Our MCMoE unifies unimodal and joint representation learning in a single stage by exploiting the complementarity between modality completion and mixture of experts.

However, existing multimodal models often assume full modality availability during training and inference. Yet, real-world inference faces inevitable missing modalities due to sensor failures (Liu et al. 2021), environmental constraints (Wang et al. 2021), or privacy concerns (Jaiswal and Provost 2020). Prior works (Wei, Luo, and Luo 2023; Park et al. 2023) show that such incompleteness severely degrades performance by disrupting cross-modal interactions. Moreover, diverse AQA applications require distinct modalities (e.g., audio in sports (Xia et al. 2023), text in action feedback (Zhang et al. 2024a), and pose in rehabilitation (Bruce et al. 2024; Zhou et al. 2023a)), challenging the generalizability of fixed-architecture multimodal models. Therefore, a flexible framework for incomplete multimodal scenarios is crucial.

Prior solutions address incomplete modalities problems

*Corresponding author.

by modality completion or cross-modal joint representation learning. The former reconstructs missing data using available ones with complex generators like diffusion models (Wang, Li, and Cui 2023; Meng et al. 2024), variational autoencoders (Shi et al. 2019; Wu and Goodman 2018), and generative adversarial networks (Cai et al. 2018; Yoon, Jordon, and Schaar 2018), which incur high computational costs unsuitable for real-time scenes. In contrast, the latter extracts joint cross-modal features (Xu, Jiang, and Liang 2024; Park et al. 2023) at a lower cost. This joint learning is also widely used for multimodal learning (Zeng and Zheng 2024; Xu et al. 2024a) and has received more attention. However, state-of-the-art methods (Zeng and Zheng 2024; Xu, Jiang, and Liang 2024; Park et al. 2023) often adopt costly two-stage pipelines that sequentially learn unimodal and cross-modal representations (Fig. 1 (a)). This inevitably increases training and optimization costs. Thus, efficiently integrating specific and shared modality knowledge to compensate for missing modalities is of great significance.

While seeking efficient solutions, we identify an indispensable element: the capability to adaptively model correlations between available and missing modalities. Adaptive cross-modal fusion reduces reliance on high-fidelity reconstruction, minimizing dependence on heavyweight generators. Also, specific unimodal knowledge drives precise integration of reconstructed and available modalities. Thus, we naturally turn to the prevalent **M**ixture **o**f **E**xperts (**MoE**) (Li et al. 2025), which flexibly employs experts specialized in processing diverse modal inputs. Based on the benefits of MoE, unimodal experts facilitate accurate modality completion, while selective collaboration specializes in diverse incomplete combinations. By exploiting complementarity between **M**issing **M**odality **C**ompletion (**MMC**) and MoE, cross-modal knowledge dynamically refines features generated by MMC, enhancing incomplete multimodal learning.

To achieve this, we propose a novel **M**issing **C**ompletion Framework with **M**ixture **o**f **E**xperts (**MCMoE**), unifying unimodal and joint learning in single-stage training. Specifically, we adaptively generate missing modalities from all available ones to preserve modality-specific knowledge modeling despite incompleteness. Then, we learn unimodal experts for each modality and design a soft router to dynamically fuse semantics, compensating for generated features and reducing reliance on heavyweight generators. By aligning the complete modality-specific and generated cross-modal representations, the model is motivated to focus on both specific and shared knowledge modeling. As shown in Fig. 1 (b), our MCMoE leverages the complementarity between MMC and MoE, achieving a balanced learning of unimodal and joint representations within single-stage training.

To bridge cross-modal semantic gaps, we further design a shared temporal enhancement module that alleviates the difficulty of completing missing data from available ones. Then, we propose a novel **A**daptive **G**ated **M**odality **G**enerator (**AGMG**) that cost-effectively adapts to various incomplete combinations. AGMG dynamically processes existing modalities to iteratively complete missing ones and employs gating layers for selective fusion. Finally, a cross-modal fusion module integrates the modality features processed by MoE for quality assessment. Extensive experiments on three public AQA benchmarks (Rhythmic Gymnastics (Zeng et al. 2020), Fis-V (Xu et al. 2019), and FS1000 (Xia et al. 2023)) show that our MCMoE outperforms state-of-the-art methods in both complete and incomplete multimodal scenarios. Abundant ablations validate the contribution of each component. *To our knowledge, this is the first work to explore incomplete multimodal action quality assessment.* The main contributions are:

- We propose a novel missing completion framework with mixture of experts for incomplete multimodal action quality assessment, reducing the reliance on heavyweight generative architectures with unified unimodal and joint representation learning in single-stage training.
- We propose the adaptive gated modality generator to selectively complete missing modalities from available ones, and design shared temporal enhancement and cross-modal fusion modules to fill the semantic gaps between modalities and fuse cross-modal semantics.
- We conduct extensive experiments and ablation studies to reveal the complementarity between missing modality generation and mixture of experts and the state-of-the-art performance of our method in both complete and incomplete multimodal scenarios on three public benchmarks.

Related Work

Multimodal Action Quality Assessment

Multimodal learning (Lai et al. 2025; Cai et al. 2024, 2025; Wu et al. 2025a; Huang et al. 2025) has achieved notable success recently. Similarly, multimodal AQA aims to evaluate action execution by integrating complementary cues (e.g., RGB, flow, audio, text), offering finer semantic understanding than unimodal methods (Xu et al. 2022; Zhou et al. 2023b). Recent works (Xu et al. 2024a; Majeedi et al. 2024; Xia et al. 2023; Zeng and Zheng 2024) have explored multi-granular semantic alignment, temporal visual-audio fusion, and adaptive modality integration. However, they assume full modality availability during both training and inference, which rarely holds in real-world scenarios due to sensor failure (Liu et al. 2021), environment (Wang et al. 2021), privacy concerns (Jaiswal and Provost 2020) or upper-layer algorithm failure (Wu et al. 2025b, 2024a,b; Chen et al. 2025). Moreover, modality importance varies across domains, e.g., audio in sports (Xia et al. 2023), text in feedback (Zhang et al. 2024a), and gesture in rehabilitation (Bruce et al. 2024). This poses higher challenges to the modal adaptability of models. *Thus, we address this gap by introducing the first framework for incomplete multimodal AQA, adaptable to diverse missing-modality conditions.*

Incomplete Multimodal Learning

Incomplete multimodal learning has become important for real-world tasks such as emotion recognition, action recognition, and medical analysis (Xu, Jiang, and Liang 2024; Park et al. 2023; Shi et al. 2024). Existing solutions mainly rely on (1) modality completion via generative models (e.g., VAE (Shi et al. 2019), GAN (Yoon, Jordon, and Schaar

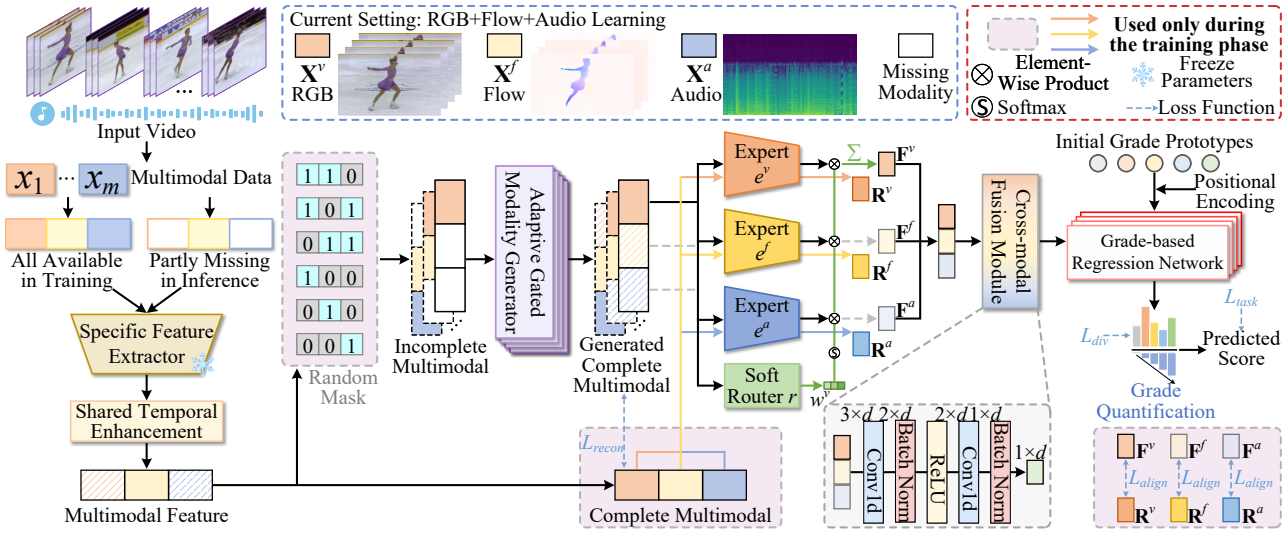


Figure 2: Overview of our missing completion framework with mixture of experts (MCMoE). Following the SOTA multimodal AQA method PAMFN, we use RGB, Flow, and Audio inputs. All modalities are visible during training, and the missing inputs during inference are zero-vector initialized. Frozen modality-specific extractors extract features, enhanced by a shared temporal enhancement module to bridge cross-modal gaps. Random masking simulates modality incompleteness during training and an adaptive gated modality generator completes missing representations. Then, unimodal experts and a soft router enable dynamic fusion, followed by cross-modal integration and grade-based regression for score prediction. (Best viewed in color.)

2018), Diffusion (Meng et al. 2024)) or (2) joint representation learning through cross-modal consistency (Lian et al. 2023; Park et al. 2023; Xu, Jiang, and Liang 2024). Yet the former depends on heavy generative models, and the latter often requires two-stage training, both increasing cost and complexity. *In contrast, our method exploits the complementarity between modality completion and mixture of experts to unify unimodal and joint learning in a single stage.*

Method

In this section, we detail our MCMoE, which combines modality completion with MoE to jointly learn unimodal and joint representations in single-stage training (Fig. 2).

Overview

Our framework accepts multimodal inputs. Its flexible architecture readily extends to arbitrary modality counts. Following mainstream incomplete multimodal research (Xu, Jiang, and Liang 2024; Lian et al. 2023; Woo et al. 2023) and the SOTA multimodal AQA method (Zeng and Zheng 2024), we focus on three modalities: RGB visual (v), optical flow (f), and audio (a). Following convention, videos are divided into T segments, which are fed into pre-trained modality-specific extractors to obtain features. These features are processed by a shared temporal enhancement module to bridge cross-modal semantic gaps, yielding a multimodal feature set $\mathbf{X}^M = \{\mathbf{X}^m | m \in \{v, f, a\}\}$. For each modality, the unimodal set is $\mathbf{X}^m = \{\boldsymbol{x}_t^m\}_{t=1}^T$. Under incomplete modalities, $\mathbf{X}^M$ partitions into available ($\mathbf{X}^{\tilde{M}} = \{\mathbf{X}^{\tilde{m}} | \tilde{m} \in \tilde{M}\}$) and missing ($\mathbf{X}^{\bar{M}} = \{\mathbf{X}^{\bar{m}} | \bar{m} \in \bar{M}\}$) subsets, where $M = \tilde{M} \cup \bar{M} = \{v, f, a\}$ and $\tilde{M} \cap \bar{M} = \emptyset$.

We propose an adaptive gated modality generator $\mathcal{G}$ to dynamically complete missing features $\mathbf{X}^{\bar{M}}$ using available features $\mathbf{X}^{\tilde{M}}$, optimized via reconstruction loss $\mathcal{L}_{recon}$. Generated features $\hat{\mathbf{X}}^{\bar{M}}$ combine with $\mathbf{X}^{\tilde{M}}$ to form a new complete feature set $\hat{\mathbf{X}}^M = \{\hat{\mathbf{X}}^m | m \in M\}$. We design a mixture-of-experts $\mathcal{E}^M = \{e^m | m \in M\}$ to dynamically extract both unimodal and joint representations: expert e^m mines modality-specific knowledge from $\hat{\mathbf{X}}^m$, while others assist in capturing cross-modal joint patterns. Then, a shared soft router r selectively fuses these features. Formally,

$$\hat{\mathbf{X}}^M = \mathbf{X}^{\tilde{M}} \cup \hat{\mathbf{X}}^{\bar{M}}, \hat{\mathbf{X}}^{\bar{M}} = \mathcal{G}\left(\left(\mathbf{X}^{\bar{M}}, \mathbf{X}^{\tilde{M}}\right) | \psi\right), \quad (1)$$

$$\mathbf{F}^m = \sum_{k \in \{v,f,a\}} r\left(\hat{\mathbf{X}}^m | \sigma\right) e^k\left(\hat{\mathbf{X}}^m | \tau^k\right), \quad (2)$$

where ψ, σ, and τ^k denote the learnable parameters of $\mathcal{G}$, r, and expert e^k of modality k. $\mathbf{F}^m$ is the feature of modality m that fuses modality-specific and shared information.

During training, we obtain the unimodal feature $\mathbf{R}^m = e^m(\hat{\mathbf{X}}^m | \tau^m)$ using the modality expert e^m. An alignment loss $\mathcal{L}_{align}$ between $\mathbf{R}^m$ and $\mathbf{F}^m$ motivates joint learning of unimodal and joint representations. Then, all $\mathbf{F}^M$ are concatenated and processed by the cross-modal fusion module $\mathcal{C}$ to capture complementary patterns, yielding multimodal features $\mathbf{H}$. Finally, a grade-based regression network $\mathcal{R}$ models performance quality patterns $\mathbf{P}^N$ from $\mathbf{H}$, which a fully connected layer regresses to rank weights $\boldsymbol{s}^N$. The final score s combines $\boldsymbol{s}^N$ with grade quantifications $\mathbf{G}^N$, where N is the number of grades. Formally,

$$\mathbf{H} = \mathcal{C}\left(\text{Concat}\left(\mathbf{F}^M\right) | \phi\right), \quad (3)$$

$$s = \sum_{n=1}^{N} \boldsymbol{s}^n \otimes \mathbf{G}^n, \mathbf{P}^N = \mathcal{R}\left(\mathbf{H}|\varphi\right), \tag{4}$$

where ϕ, φ are learnable parameters of $\mathcal{C}$ and $\mathcal{R}$, and $\otimes$ denotes element-wise product in a batch. A diversity loss $\mathcal{L}_{div}$ ensures grade patterns focus on distinct performance aspects, while task-specific loss $\mathcal{L}_{task}$ fits quality assessment. With score label $\hat{s}$, the final objective $\mathcal{J}$ is:

$$\begin{aligned} \min\ \mathcal{J} = \lambda_1 \mathcal{L}_{recon}(\hat{\mathbf{X}}^M, \mathbf{X}^M) + \lambda_2 \mathcal{L}_{align}(\mathbf{R}^M, \mathbf{F}^M) \\ + \lambda_3 \mathcal{L}_{div}(\mathbf{P}^N) + \lambda_4 \mathcal{L}_{task}(s, \hat{s}), \end{aligned} \tag{5}$$

where λ_1, λ_2, λ_3, and λ_4 are the balancing weights.

Feature Extraction

For fairness, we follow preprocessing pipelines from existing multimodal AQA works (Zeng and Zheng 2024; Xia et al. 2023). On Rhythmic Gymnastics and Fis-V, videos are split into T non-overlapping 32-frame segments $\mathbf{I}_T$. Within each segment, frozen pre-trained extractors—VST (Liu et al. 2022) for RGB, I3D (Carreira and Zisserman 2017) for flow, and AST (Gong, Chung, and Glass 2021) for audio—extract temporally aligned features $\mathbf{I}_T^v$, $\mathbf{I}_T^f$, and $\mathbf{I}_T^a$ of dimension 1024, 1024, and 768, respectively. For FS1000, segments are 5 seconds long with 3 seconds overlap, and TimeSformer (Bertasius, Wang, and Torresani 2021), I3D, and AST process non-overlapping 8-frame clips in each segment with dimensions 768, 1024, and 768. Formally,

$$\mathbf{I}_T^v = \mathcal{V}\left(\mathbf{I}_T\right), \mathbf{I}_T^f = \mathcal{F}\left(\mathbf{I}_T\right), \mathbf{I}_T^a = \mathcal{A}\left(\mathbf{I}_T\right), \tag{6}$$

where $\mathcal{V}$, $\mathcal{F}$, and $\mathcal{A}$ are the frozen specific feature extractors.

For cross-modal interactions, inputs are projected to a shared latent space of dimension d via modality-specific modules $\mathcal{P}^M$ with parameters ω^M. We then design a **S**hared **T**emporal **E**nhancement **M**odule (**STEM**) $\mathcal{T}$ to bridge semantic gaps. Implemented as stackable Transformer encoders (Vaswani et al. 2017), $\mathcal{T}$'s shared parameters ϑ capture cross-modal commonalities while enhancing modalities separately. This avoids direct multimodal feature interaction, facilitating subsequent unimodal learning. Formally,

$$\mathbf{X}^M = \mathcal{T}\left(\mathcal{P}^M\left(\mathbf{I}_T^M|\omega^M\right)|\vartheta\right). \tag{7}$$

Adaptive Gated Modality Generator

To complete the missing modalities, we propose a novel **A**daptive **G**ated **M**odality **G**enerator (**AGMG**), which adaptively generates absent features iteratively based on available ones (Fig. 3). Inspired by (Vaswani et al. 2017), AGMG first applies multi-head cross-attention: the concatenated available modality features $\mathbf{X}^{\tilde{M}}$ serve as *keys/values*, while zero-initialized missing features $\mathbf{X}^{\bar{M}}$ act as *queries*. Outputs become subsequent layer queries for iterative refinement over L layers (l denotes the current l-th layer). Thus, for the missing modality $\bar{m}$, $\mathbf{X}_Q^l = \mathbf{W}_Q^{\bar{m}} \mathbf{X}^{\bar{m}}$,

$$\mathbf{X}_K = \mathbf{W}_K^{\bar{m}} \text{Concat}\left(\mathbf{X}^{\tilde{M}}\right), \mathbf{X}_V = \mathbf{W}_V^{\bar{m}} \text{Concat}\left(\mathbf{X}^{\tilde{M}}\right), \tag{8}$$

$$\mathbf{X}_Q^{l+1} = \text{Softmax}\left(\mathbf{X}_Q^l (\mathbf{X}_K)^{\mathrm{T}} / \sqrt{d}\right) \mathbf{X}_V, \tag{9}$$

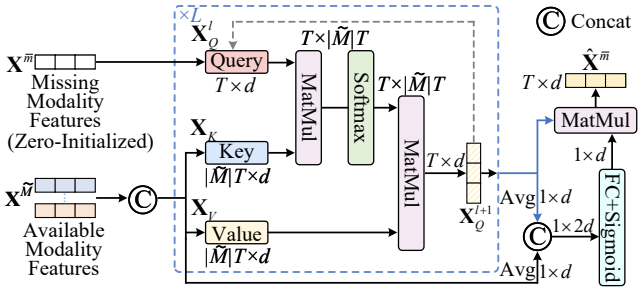


Figure 3: The illustration of our proposed Adaptive Gated Modality Generator (AGMG).

where $\sqrt{d}$ is a normalization factor, $\mathbf{W}_Q^{\bar{m}}$, $\mathbf{W}_K^{\bar{m}}$, and $\mathbf{W}_V^{\bar{m}}$ are modality-specific learnable weights. The final layer yields $\mathbf{X}_{new}^{\bar{m}}$. AGMG then employs gating layers g to dynamically weight fusions based on current input completeness, mitigating potential error propagation from imperfect generation:

$$\hat{\mathbf{X}}^{\bar{m}} = \mathbf{X}_{new}^{\bar{m}} \cdot g\left(\text{Avg}\left(\text{Concat}\left(\mathbf{X}^{\tilde{M}}\right)\right), \text{Avg}\left(\mathbf{X}_{new}^{\bar{m}}\right)\right). \tag{10}$$

In inference, missing modalities are zero-initialized without any information leakage, as in (Xu, Jiang, and Liang 2024; Woo et al. 2023). In training, we simulate incompleteness via random masking of complete modalities.

Complementarity between MMC and MoE

This work exploits the complementarity between **M**issing **M**odality **C**ompletion (**MMC**) and **M**ixture **o**f **E**xperts (**MoE**) to avoid heavy generative models and unify unimodal/joint learning in one stage. We detail this in two parts:

Benefits of MMC for MoE. Inspired by prior works (Xu, Jiang, and Liang 2024; Zhang et al. 2024b), we employ a MoE to dynamically mix expert knowledge for cross-modal representation in incomplete multimodal scenarios. However, missing modalities limit or mislead cross-modal semantic extraction, and existing two-stage solutions (Xu, Jiang, and Liang 2024; Park et al. 2023) increase training cost and optimization complexity. In contrast, MMC generates higher-confidence features instead of zero matrices based on available modalities. Guided by the reconstruction loss $\mathcal{L}_{recon}$ with real-modality supervision, these features carry reliable unimodal knowledge and cross-modal cues, benefiting both unimodal and joint learning for modality experts and enabling experts with fewer parameters.

Benefits of MoE for MMC. Existing MMC methods often rely on heavyweight generative models (Cai et al. 2018; Wang, Li, and Cui 2023; Meng et al. 2024) to complete missing modalities. In contrast, our MoE dynamically selects and fuses modality knowledge to handle generated features. It adaptively balances real available and generated features to prevent error propagation from imperfect generation. When processing missing modality $\bar{m}$, MoE further refines missing data via inter-modal correlations from available ones, boosting robustness and accuracy of joint representations. These advantages reduce reliance on high-fidelity generators in MMC and lower parameter costs.

Datasets	Methods	Year	Testing Condition (Spearman Correlation (↑)/Mean Square Error (↓))							
			$\{v, f\}$	$\{v, a\}$	$\{f, a\}$	$\{v\}$	$\{f\}$	$\{a\}$	Average	$\{v, f, a\}$
FS1000 (7-class)	♡MLP-Mixer*	2023	0.722/25.56	0.542/60.57	0.474/87.20	0.623/101.35	0.472/87.59	0.177/68.43	0.520/71.78	0.819/14.56
	♡PAMFN*	2024	0.727/59.96	0.644/62.80	0.561/56.36	0.713/92.10	0.486/117.63	0.145/62.13	0.571/75.16	0.855/13.02
	♣ActionMAE*	2023	0.775/24.66	0.766/64.13	0.556/26.51	0.761/50.64	0.462/**21.47**	0.458/41.66	0.651/38.18	0.809/17.96
	♠GCNet*	2023	0.730/25.56	0.740/23.86	0.507/24.97	0.696/26.67	0.447/31.27	0.442/39.40	0.610/28.62	0.764/21.82
	♠IMDer*	2023	0.760/22.34	0.745/28.46	0.573/24.86	0.724/35.99	0.424/22.92	0.488/32.56	0.636/27.86	0.788/25.95
	♠MoMKE*	2024	0.798/18.86	0.805/23.88	0.541/24.96	0.785/37.96	0.398/23.31	0.499/27.53	0.668/26.08	0.819/16.85
	♠SDR-GNN*	2025	0.789/17.50	0.785/25.08	0.564/22.29	0.749/28.47	0.504/29.96	0.477/25.46	0.665/24.79	0.817/15.91
	MCMoE (Ours)	-	**0.845/12.66**	**0.882/11.85**	**0.738/14.88**	**0.845/13.64**	**0.650**/22.47	**0.615/16.72**	**0.782/15.37**	**0.881/11.53**
Fis-V (2-class)	♡MLP-Mixer*	2023	0.732/30.34	0.651/46.70	0.572/26.96	0.618/48.46	0.546/27.18	0.325/67.25	0.586/41.15	0.772/13.97
	♡PAMFN	2024	0.801/33.49	0.661/54.34	0.622/110.50	0.644/84.93	0.616/110.42	0.141/86.16	0.610/79.97	0.822/15.33
	♣ActionMAE*	2023	0.704/33.54	0.678/27.61	0.575/25.55	0.616/40.07	0.484/24.87	0.486/29.29	0.597/30.16	0.698/17.34
	♠GCNet*	2023	0.738/19.86	0.656/21.32	0.594/21.72	0.667/20.87	0.602/19.61	0.455/34.77	0.626/23.03	0.698/16.93
	♠IMDer*	2023	0.748/15.19	0.658/22.66	0.568/23.99	0.675/25.45	0.618/26.38	0.405/31.96	0.622/24.27	0.703/17.02
	♠MoMKE*	2024	0.754/14.84	0.689/20.60	0.646/19.62	0.684/23.16	0.654/22.46	0.497/29.09	0.660/21.63	0.747/17.30
	♠SDR-GNN*	2025	0.752/14.99	0.680/20.62	0.619/20.89	0.689/20.26	0.648/21.50	0.479/32.13	0.651/21.73	0.733/16.45
	MCMoE (Ours)	-	**0.813/11.02**	**0.787/14.64**	**0.727/17.41**	**0.765/15.14**	**0.698/15.39**	**0.557/28.54**	**0.734/17.02**	**0.829/12.15**
RG (4-class)	♡MLP-Mixer*	2023	0.733/7.23	0.614/14.09	0.485/10.18	0.655/9.67	0.566/11.45	0.244/16.01	0.567/11.44	0.754/7.48
	♡PAMFN	2024	0.764/6.78	0.616/38.87	0.448/122.11	0.658/39.86	0.483/123.44	0.131/151.69	0.543/80.46	0.819/6.64
	♣ActionMAE*	2023	0.724/7.30	0.621/8.76	0.545/10.39	0.689/12.41	0.521/11.29	0.251/16.84	0.575/11.16	0.709/7.01
	♠GCNet*	2023	0.738/6.69	0.638/7.95	0.556/11.75	0.701/8.12	0.568/36.01	0.225/15.20	0.591/14.29	0.716/6.45
	♠IMDer*	2023	0.746/6.03	0.646/7.55	0.569/8.80	0.699/7.59	0.596/9.35	0.206/14.19	0.598/8.92	0.724/6.37
	♠MoMKE*	2024	0.762/5.69	0.656/7.97	0.629/9.39	0.693/8.42	0.621/10.08	0.264/13.25	0.623/9.13	0.747/6.18
	♠SDR-GNN*	2025	0.758/6.08	0.655/7.38	0.612/9.40	0.727/7.77	0.591/9.80	0.264/13.53	0.621/8.99	0.742/6.35
	MCMoE (Ours)	-	**0.822/5.33**	**0.781/5.83**	**0.699/8.15**	**0.767/6.25**	**0.662/8.59**	**0.278/13.20**	**0.697/7.89**	**0.842/4.85**

Table 1: Comparisons of performance on three benchmarks with incomplete modalities. v, f, and a refer to the RGB, flow, and audio modalities. "Average" denotes the average result of all six incomplete multimodal combinations. The **bold** / underline indicate the best / second-best results. * indicates our reimplementation. ♡, ♣, and ♠ mean the evaluated method sources for multimodal AQA, incomplete multimodal action recognition, and incomplete multimodal emotion recognition.

Building on the above complementarity, our MoE adopts a lightweight two-layer multilayer perceptron (MLP) for each expert. We aggregate all experts' outputs to derive both unimodal and joint representations for a given modality. For example, the unimodal ($\mathbf{F}_v^v$) and joint representations ($\mathbf{F}_f^v$ and $\mathbf{F}_a^v$) of visual modality v can be obtained as follows:

$$\mathbf{F}_m^v = e^m\left(\hat{\mathbf{X}}^v\right) = \text{MLP}_{\tau^m}\left(\hat{\mathbf{X}}^v\right), m \in \{v, f, a\}. \quad (11)$$

To handle diverse missing-modality scenarios, we employ a two-layer MLP soft router r to dynamically estimates the importance of unimodal and joint representations based on the input $\hat{\mathbf{X}}^v$. The final visual feature $\mathbf{F}^v$ is obtained by weighting based on the importance weights $\boldsymbol{w}_m^v$:

$$\left\{\boldsymbol{w}_v^v, \boldsymbol{w}_f^v, \boldsymbol{w}_a^v\right\} = r\left(\hat{\mathbf{X}}^v\right) = \text{Softmax}\left(\text{MLP}_\sigma\left(\hat{\mathbf{X}}^v\right)\right), \quad (12)$$

$$\mathbf{F}^v = \sum_{m\in\{v,f,a\}} \boldsymbol{w}_m^v \cdot \mathbf{F}_m^v. \quad (13)$$

The flow ($\mathbf{F}^f$) and audio ($\mathbf{F}^a$) features are obtained similarly. We also apply the alignment loss $\mathcal{L}_{align}$ to align $\mathbf{F}^m$ with true unimodal features $\mathbf{R}^m = e^m(\hat{\mathbf{X}}^m)$. This motivates both unimodal and joint learning within single-stage training, and also indirectly promotes AGMG to generate more faithful information. Finally, we design a **C**ross-modal **F**usion **M**odule (**CFM**) to capture inter-modal correlations and map features into a task-specific latent space. Our CFM can rely on a simple convolutional block (Fig. 2) to leverage the MoE's cross-modal semantics for efficient fusion.

Score Generation and Optimization

As shown in Eq. 3, we extract multimodal features $\mathbf{H}$ via CFM. For scoring, we employ state-of-the-art grade-based regression (Xu, Zeng, and Zheng 2022; Xu et al. 2025a; Liu et al. 2025). We initialize N learnable grade prototypes with sine–cosine positional encodings, then use a three-layer Transformer decoder to implement regressor $\mathcal{R}$ to aggregate $\mathbf{H}$ into grade patterns $\mathbf{P}^N$. The final score s is computed by Eq. 4, where the grade quantification $\mathbf{G}^n = \frac{n-1}{N-1}$.

As shown in Eq. 5, our method uses four losses: reconstruction loss $\mathcal{L}_{recon}$, alignment loss $\mathcal{L}_{align}$, diversity loss $\mathcal{L}_{div}$, and task-specific loss $\mathcal{L}_{task}$. Specifically, $\mathcal{L}_{recon}$ uses Mean Square Error (MSE) to train AGMG for high-fidelity features. $\mathcal{L}_{align}$ minimizes Kullback-Leibler (KL) divergence between $\mathbf{F}^M$ and $\mathbf{R}^M$ for unified unimodal and joint learning. $\mathcal{L}_{div}$ applies triplet loss to separate grade patterns. For quality assessment, $\mathcal{L}_{task}$ uses MSE to fit predictions to expert scores. Hence, Eq. 5 can be rewritten as:

$$\text{MSE}(y, \hat{y}) = \|y - \hat{y}\|^2, \quad (14)$$

$$\mathcal{L}_{recon}\left(\hat{\mathbf{X}}^M, \mathbf{X}^M\right) = \text{MSE}\left(\hat{\mathbf{X}}^M, \mathbf{X}^M\right), \mathcal{L}_{task}(s, \hat{s}) = \text{MSE}(s, \hat{s}), \quad (15)$$

$$\mathcal{L}_{align}(\mathbf{R}^M, \mathbf{F}^M) = \text{KL}\left(\mathbf{R}^M \parallel \mathbf{F}^M\right) = \textstyle\sum_t \mathbf{R}_t^M \log\left(\frac{\mathbf{R}_t^M}{\mathbf{F}_t^M}\right), \quad (16)$$

$$\mathcal{L}_{div}(\mathbf{P}^N) = \sum_n \left[\max\left(\text{sim}\left(\mathbf{P}^n, \mathbf{P}^i\right)\right) - \min\left(\text{sim}\left(\mathbf{P}^n, \mathbf{P}^i\right)\right) + \delta\right]_+, \quad (17)$$

where $i \neq n$, δ is a margin parameter, which is set to 1. The $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, $[\cdot]_+$ means $\max(0, \cdot)$.

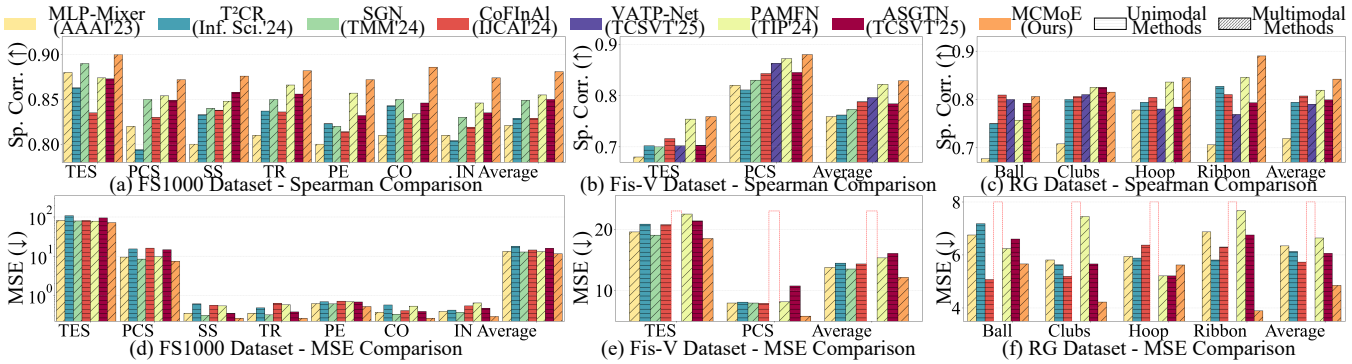


Figure 4: Comparisons of performance with complete modalities. * indicates our reimplementation based on the official code.

Experiments

Datasets and Metrics. We evaluate our method on three public AQA benchmarks: FS1000 (Xia et al. 2023), Fis-V (Xu et al. 2019), and Rhythmic Gymnastics (RG) (Zeng et al. 2020), which provide RGB, flow, and audio modalities. Following standard protocols (Xia et al. 2023; Du et al. 2024), we report Spearman's Rank Correlation (ρ) and Mean Square Error (MSE). MSE measures the numerical error (Eq. 14), while ρ assesses the rank agreement between predictions and ground truth.

Implementation Details. All experiments use an RTX 3090 GPU (PyTorch 1.12.0). We evaluate incomplete modalities using the common fixed-missing protocol (Xu, Jiang, and Liang 2024; Wang, Li, and Cui 2023), covering the full set $\{v, f, a\}$ and six subsets ($\{v, f\}$, $\{v, a\}$, $\{f, a\}$, $\{v\}$, $\{f\}$, $\{a\}$). For FS1000/Fis-V/RG, we randomly sample 95/124/68 continuous clips. The λ_1, λ_2, λ_3, and λ_4 in Eq. 5 are 1, 1, 1/0.5/1, and 10. The grade N is 4. We set a dropout of 0.3 to avoid over-fitting. The batch size is 32 and learning rate is 1e-4/2e-4/2e-4. We optimize with Adam (weight decay 1e-4) and cosine annealing (decay 0.01). For better convergence, we train models with different epochs as in (Xu, Zeng, and Zheng 2022; Zeng and Zheng 2024; Zhou et al. 2024). More details are in (Xu et al. 2025d).

Comparison with State-of-the-art

Incomplete Multimodal Scenarios. To evaluate our method under incomplete modalities, we test all modal combinations on three datasets (Tab. 1). Comparisons include SOTA methods from multimodal AQA (Xia et al. 2023; Zeng and Zheng 2024), incomplete multimodal action recognition (Woo et al. 2023), and incomplete multimodal emotion recognition (Lian et al. 2023; Wang, Li, and Cui 2023; Xu, Jiang, and Liang 2024; Fu et al. 2025). We extend the bimodal MLP-Mixer (Xia et al. 2023) to support trimodal inputs via secondary bimodal interactions. Our MCMoE outperforms all baselines on nearly all metrics and incomplete configurations. Averaged over six combinations, it improves SP. Corr./MSE by 17.1%/38.0%, 11.2%/21.3%, and 11.9%/11.5% on the three datasets. Existing SOTA multimodal AQA methods degrade notably with missing modalities, especially on MSE, likely due to disrupted cross-modal interactions. Incomplete multimodal baselines from other domains lack tailored modeling for action semantics and assessment patterns, resulting in poor AQA performance. Our MCMoE maintains strong performance under both complete and incomplete settings, with average gains of 2.23%/15.3% under full modalities. This highlights the benefit of jointly leveraging MMC and MoE to compensate for missing modality interference.

Methods	Year	1-stage	#Params	#FLOPs	Average	$\{v, f, a\}$
MLP-Mixer	2023	✓	14.32M	49.90G	0.52/71.8	0.82/14.6
PAMFN	2024	×	18.06M	2.56G	0.57/75.2	0.86/13.0
ActionMAE	2023	✓	14.05M	62.12G	0.65/38.2	0.81/18.0
GCNet	2023	✓	8.78M	1191.39G	0.61/28.6	0.76/21.8
IMDer	2023	×	7.97M	23.53G	0.61/27.9	0.79/26.0
MoMKE	2024	×	5.39M	2.60G	0.67/26.1	0.82/16.9
SDR-GNN	2025	✓	22.63M	24.35G	0.67/24.8	0.82/15.9
Ours	-	✓	**4.90M**	**1.34 G**	**0.78/15.4**	**0.88/11.5**

Table 2: Compare the computational costs with the SOTA.

Complete Modality Scenarios. Fig. 4 compares our method with unimodal(Ke et al. 2024; Zhou et al. 2024; Liu et al. 2025) and multimodal (Xia et al. 2023; Zeng and Zheng 2024; Gedamu et al. 2025; Du et al. 2024) SOTA AQA models under full-modality settings. Our method achieves the best or second-best results across all categories and consistently superior averages, with SP. Corr./MSE gains of 3.0%/9.7%, 0.9%/10.1%, and 2.8%/15.4% on the three datasets. This validates the effectiveness of MMC+MoE synergy, which enhances adaptive multimodal fusion and mitigates the impact of missing data. In addition, our single-stage learning avoids knowledge forgetting and complexity issues common in two-stage training.

As shown in Tab. 1 and Fig. 4, our MCMoE achieves balanced and state-of-the-art performance in both complete and incomplete scenarios. Moreover, Tab. 2 shows it offers a better performance–efficiency trade-off compared to SOTA incomplete multimodal methods on FS1000.

Ablation Study

To validate the effectiveness of our components, we build a **Baseline** that extracts and projects multimodal features, directly summed and fed into a grade-based regressor with $\mathcal{L}div$ and $\mathcal{L}task$. As shown in Tab. 3, performance im-

Settings	RG		Fis-V	
	Average	$\{v, f, a\}$	Average	$\{v, f, a\}$
Baseline	0.532/25.79	0.718/7.05	0.577/61.45	0.724/16.38
+ STEM	0.573/19.92	0.744/6.83	0.637/39.72	0.748/14.30
+ AGMG	0.647/9.45	0.779/6.04	0.678/21.36	0.773/13.26
+ MoE (w/o CFM)	0.675/7.91	0.817/5.30	0.701/18.07	0.790/13.10
+ CFM (Ours)	**0.697/7.89**	**0.842/4.85**	**0.734/17.02**	**0.829/12.15**
w/o AGMG	0.635/16.54	0.724/6.44	0.609/50.45	0.742/14.37
w/o MoE	0.658/7.97	0.777/5.48	0.684/19.20	0.767/13.32
w/o STEM	0.585/10.96	0.735/7.62	0.643/19.34	0.739/14.16
w/o $\mathcal{L}_{recon}$	0.617/9.30	0.771/5.50	0.699/18.45	0.813/12.47
w/o $\mathcal{L}_{align}$	0.648/8.56	0.797/5.16	0.708/19.25	0.790/12.59
w/o $\mathcal{L}_{div}$	0.624/8.13	0.791/4.93	0.693/17.54	0.795/12.29

Table 3: Ablation results on the RG and Fis-V. The top half adds our components in order, and the bottom half individually removes one. Results are shown by $\rho(\uparrow)$/MSE($\downarrow$).

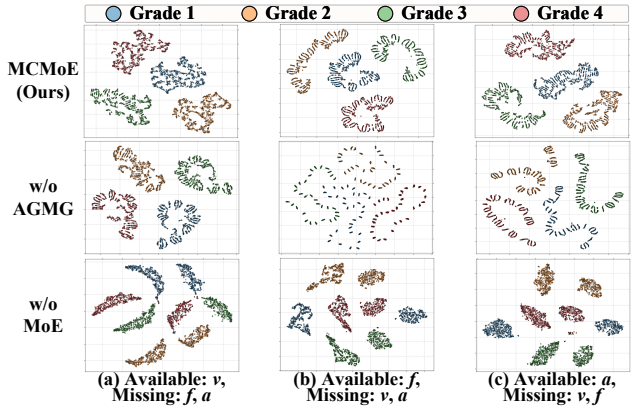


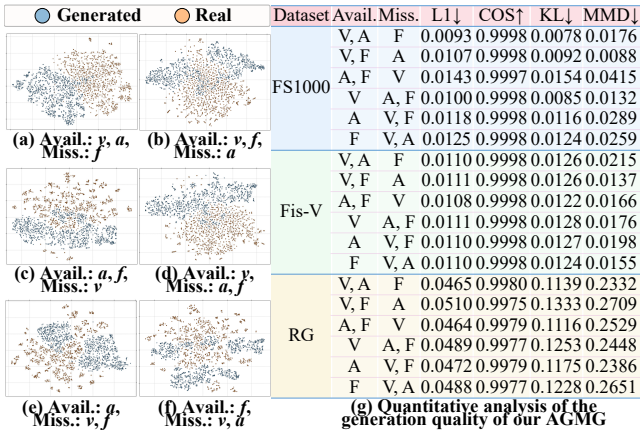
Figure 5: The t-SNE grade distributions in the three extreme unimodal scenes contrasting without AGMG and MoE.

proves as components are incrementally added. Notably, adding AGMG significantly boosts accuracy, showing that dynamic modality completion effectively bridges semantic gaps of missing data. Adding MoE further improves results, especially under incomplete settings, with average gains of 3.9%/15.9%. This is likely due to its adaptive fusion of unimodal and cross-modal knowledge. Removing any component from the full model leads to clear performance drops. Notably, STEM is critical for capturing temporal context and cross-modal semantics, essential in long video understanding (Xu, Zeng, and Zheng 2022; Zhou et al. 2024) and multimodal learning (Zeng and Zheng 2024; Woo et al. 2023).

We also ablate the loss terms. Beyond the core $\mathcal{L}_{task}$, removing $\mathcal{L}_{recon}$, $\mathcal{L}_{align}$, or $\mathcal{L}_{div}$ each harms performance. $\mathcal{L}_{recon}$ provides key supervision for reliable modality completion, avoiding generating misleading misinformation. $\mathcal{L}_{align}$ motivates MoE to jointly learn unimodal and cross-modal features within single-stage training. $\mathcal{L}_{div}$ enforces diversity across grade patterns, aiding accurate AQA.

Visualization Analysis

To visualize the complementary effects of our emphasized MMC and MoE, Fig. 5 presents t-SNE distributions of grade



Dataset	Avail.	Miss.	L1↓	COS↑	KL↓	MMD↓
FS1000	V, A	F	0.0093	0.9998	0.0078	0.0176
	V, F	A	0.0107	0.9998	0.0092	0.0088
	A, F	V	0.0143	0.9997	0.0154	0.0415
	V	A, F	0.0100	0.9998	0.0085	0.0132
	A	V, F	0.0118	0.9998	0.0116	0.0289
	F	V, A	0.0125	0.9998	0.0124	0.0259
Fis-V	V, A	F	0.0110	0.9998	0.0126	0.0215
	V, F	A	0.0111	0.9998	0.0126	0.0137
	A, F	V	0.0108	0.9998	0.0122	0.0166
	V	A, F	0.0111	0.9998	0.0128	0.0176
	A	V, F	0.0110	0.9998	0.0127	0.0198
	F	V, A	0.0110	0.9998	0.0124	0.0155
RG	V, A	F	0.0465	0.9980	0.1139	0.2332
	V, F	A	0.0510	0.9975	0.1333	0.2709
	A, F	V	0.0464	0.9979	0.1116	0.2529
	V	A, F	0.0489	0.9977	0.1253	0.2448
	A	V, F	0.0472	0.9979	0.1175	0.2386
	F	V, A	0.0488	0.9977	0.1228	0.2651

(g) Quantitative analysis of the generation quality of our AGMG

Figure 6: (a-f) t-SNE distributions of our generated and true features; (g) Generation quality analysis on four metrics.

patterns on FS1000. Each point is a grade feature. Our MCMoE effectively distinguishes action qualities, yielding compact clusters with clear inter-grade boundaries. In contrast, omitting AGMG or MoE leads to scattered distributions and weaker separability. Without AGMG's modality completion, MoE fails to maintain intra-grade consistency, resulting in dispersed features. Without MoE's dynamic cross-modal fusion, unimodal semantics cannot collaborate effectively, splitting quality grades into two clusters due to modality divergence. This visualization highlights the critical roles of MMC and MoE in quality space modeling.

Moreover, Fig. 6 qualitatively and quantitatively shows the quality of features generated by AGMG. The generated and real features are well mixed in the t-SNE space in all incomplete combinations on FS1000, indicating high semantic similarity that makes them hard to distinguish. We conduct a comprehensive quantitative analysis using four similarity metrics—L1 Distance, Cosine Similarity, KL Divergence, and Maximum Mean Discrepancy. Even on the smallest RG with limited multimodal training data, AGMG produces high-quality features. These results show that AGMG generates semantically aligned features at a low cost, effectively mitigating the limitations caused by missing data.

Conclusion

In this paper, we introduce MCMoE, a framework for incomplete multimodal action quality assessment that leverages the complementarity between Missing Modality Completion (MMC) and Mixture of Experts (MoE). MCMoE uses an adaptive gated modality generator to reconstruct missing modalities and a MoE architecture with unimodal experts and a soft router to fuse modality-specific and cross-modal information. This design mitigates the impact of missing data, reduces dependence on heavy generative models, and enables unified unimodal and joint representation learning in a single stage. As a result, MCMoE achieves superior performance on three public AQA benchmarks under both complete and incomplete settings, striking a balance between performance and cost.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the National Natural Science Foundation of China under Grant 61972097, U21A20472, 62522102, and 62373043, in part by the Major Scientific Research Project for Technology Promotes Police under Grant 2025YZ040003, 2024YZ040001, in part by the Natural Science Foundation of Fujian Province under Grant 2025J01536.

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, volume 139, 813–824.

Bruce, X.; Liu, Y.; Chan, K. C.; and Chen, C. W. 2024. EGCN++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE TPAMI*.

Cai, J.; Li, Q.; Shen, Y.; Pan, J.; and Liu, W. 2024. Efficient Semantic Segmentation for Compressed Video. In *ICRA*, 4266–4272.

Cai, J.; Su, J.; Li, Q.; Yang, W.; Wang, S.; Zhao, T.; He, S.; and Liu, W. 2025. Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+ X Semantic Segmentation. In *CVPR*, 10587–10598.

Cai, L.; Wang, Z.; Gao, H.; Shen, D.; and Ji, S. 2018. Deep adversarial learning for multi-modality missing data completion. In *ACM SIGKDD*, 1158–1166.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Chen, H.; Wu, S.; Wang, Z.; Yin, Y.; Jiao, Y.; Lyu, Y.; and Liu, Z. 2025. Causal-Inspired Multitask Learning for Video-Based Human Pose Estimation. In *AAAI*, 2052–2060.

Ding, X.; Xu, X.; and Li, X. 2023. SEDSkill: Surgical Events Driven Method for Skill Assessment from Thoracoscopic Surgical Videos. In *MICCAI*, 35–45.

Du, Z.; He, D.; Wang, X.; and Wang, Q. 2024. Learning Semantics-Guided Representations for Scoring Figure Skating. *IEEE TMM*, 26: 4987–4997.

Fu, F.; Ai, W.; Yang, F.; Shou, Y.; Meng, T.; and Li, K. 2025. SDR-GNN: spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *KBS*, 309: 112825.

Gedamu, K.; Ji, Y.; Yang, Y.; Shao, J.; and Shen, H. T. 2025. Visual-semantic Alignment Temporal Parsing for Action Quality Assessment. *IEEE TCSVT*, 35(3): 2436–2449.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Huang, C.; Su, Y.; Xu, H.; and Ke, X. 2025. Progressive Modality-Adaptive Interactive Network for Multi-Modality Image Fusion. In *IJCAI*, 1161–1169.

Jaiswal, M.; and Provost, E. M. 2020. Privacy enhanced multimodal neural representations for emotion recognition. In *AAAI*, 7985–7993.

Ke, X.; Xu, H.; Lin, X.; and Guo, W. 2024. Two-path target-aware contrastive regression for action quality assessment. *Inf. Sci.*, 664: 120347.

Lai, X.; Ke, X.; Xu, H.; Wu, S.; and Guo, W. 2025. MSP: Multimodal Self-Attention Prompt Learning. *IEEE TIP*, 34: 5978–5988.

Li, Y.; Niu, Y.; Xu, H.; Da, H.; Xu, R.; and Liu, W. 2025. IPCMoE: Integrating Perceptual Cues with Mixture-of-Experts for Joint Low-Light Image Enhancement and Deblurring. In *ACM MM*, 7644–7652.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE TPAMI*, 45(7): 8419–8432.

Liu, A.; Tan, Z.; Wan, J.; Liang, Y.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Face anti-spoofing via adversarial cross-modality translation. *IEEE TIFS*, 16: 2759–2772.

Liu, J.; Wang, H.; Zhou, W.; Stawarz, K.; Corcoran, P.; Chen, Y.; and Liu, H. 2025. Adaptive Spatiotemporal Graph Transformer Network for Action Quality Assessment. *IEEE TCSVT*, 1–1.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *CVPR*, 3202–3211.

Majeedi, A.; Gajjala, V. R.; GNVV, S. S. S. N.; and Li, Y. 2024. RICAˆ 2: Rubric-Informed, Calibrated Assessment of Actions. In *ECCV*, 143–161.

Meng, X.; Sun, K.; Xu, J.; He, X.; and Shen, D. 2024. Multi-modal modality-masked diffusion network for brain MRI synthesis with random modality missing. *IEEE TMI*, 43(7): 2587–2598.

Pan, J.-H.; Gao, J.; and Zheng, W.-S. 2019. Action assessment by joint relation graphs. In *ICCV*, 6331–6340.

Park, Y.; Woo, S.; Lee, S.; Nugroho, M. A.; and Kim, C. 2023. Cross-modal alignment and translation for missing modality action recognition. *CVIU*, 236: 103805.

Shi, J.; Shang, C.; Sun, Z.; Yu, L.; Yang, X.; and Yan, Z. 2024. PASSION: Towards Effective Incomplete Multi-Modal Medical Image Segmentation with Imbalanced Missing Rates. In *ACM MM*, 456–465.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*, 15692–15703.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, S.; Yang, D.; Zhai, P.; Chen, C.; and Zhang, L. 2021. Tsa-net: Tube self-attention network for action quality assessment. In *ACM MM*, 4902–4910.

Wang, Y.; Li, Y.; and Cui, Z. 2023. Incomplete multimodality-diffused emotion recognition. In *NeurIPS*, 17117–17128.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *CVPR*, 20039–20049.

Woo, S.; Lee, S.; Park, Y.; Nugroho, M. A.; and Kim, C. 2023. Towards good practices for missing modality robust action recognition. In *AAAI*, 2776–2784.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Chen, Y.; and Wu, Q. 2025a. Enhanced Visual-Semantic Interaction with Tailored Prompts for Pedestrian Attribute Recognition. In *CVPR*, 9570–9579.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 5580–5590.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *ACM MM*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE TDSC*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025b. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Xia, J.; Zhuge, M.; Geng, T.; Fan, S.; Wei, Y.; He, Z.; and Zheng, F. 2023. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *AAAI*, 2901–2909.

Xu, A.; Zeng, L.-A.; and Zheng, W.-S. 2022. Likert scoring with grade decoupling for long-term action assessment. In *CVPR*, 3232–3241.

Xu, C.; Fu, Y.; Zhang, B.; Chen, Z.; Jiang, Y.-G.; and Xue, X. 2019. Learning to score figure skating sport videos. *IEEE TCSVT*, 30(12): 4578–4590.

Xu, H.; Ke, X.; Li, Y.; Xu, R.; Wu, H.; Lin, X.; and Guo, W. 2024a. Vision-Language Action Knowledge Learning for Semantic-Aware Action Quality Assessment. In *ECCV*, 423–440.

Xu, H.; Ke, X.; Wu, H.; Xu, R.; Li, Y.; and Guo, W. 2025a. Language-Guided Audio-Visual Learning for Long-Term Sports Assessment. In *CVPR*, 23967–23977.

Xu, H.; Ke, X.; Wu, H.; Xu, R.; Li, Y.; Xu, P.; and Guo, W. 2025b. DanceFix: An Exploration in Group Dance Neatness Assessment Through Fixing Abnormal Challenges of Human Pose. In *AAAI*, 8869–8877.

Xu, H.; Wu, H.; Ke, X.; Li, Y.; Xu, R.; and Guo, W. 2025c. Quality-Guided Vision-Language Learning for Long-Term Action Quality Assessment. *IEEE Transactions on Multimedia*, 27: 7326–7339.

Xu, H.; Wu, H.; Ke, X.; Wu, J.; Xu, R.; and Xu, J. 2025d. MCMoE: Completing Missing Modalities with Mixture of Experts for Incomplete Multimodal Action Quality Assessment. *arXiv preprint arXiv:2511.17397*.

Xu, J.; Rao, Y.; Yu, X.; Chen, G.; Zhou, J.; and Lu, J. 2022. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, 2949–2958.

Xu, J.; Rao, Y.; Zhou, J.; and Lu, J. 2024b. Procedure-aware action quality assessment: Datasets and performance evaluation. *IJCV*, 132(12): 6069–6090.

Xu, J.; Yin, S.; and Peng, Y. 2025. Human-Centric Fine-Grained Action Quality Assessment. *IEEE TPAMI*, 47(8): 6242–6255.

Xu, J.; Yin, S.; Zhao, G.; Wang, Z.; and Peng, Y. 2024c. FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Action Quality Assessment. In *CVPR*, 14628–14637.

Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *ACM MM*, 438–446.

Yoon, J.; Jordon, J.; and Schaar, M. 2018. Gain: Missing data imputation using generative adversarial nets. In *ICML*, 5689–5698.

Zeng, L.; and Zheng, W. 2024. Multimodal Action Quality Assessment. *IEEE TIP*, 33: 1600–1613.

Zeng, L.-A.; Hong, F.-T.; Zheng, W.-S.; Yu, Q.-Z.; Zeng, W.; Wang, Y.-W.; and Lai, J.-H. 2020. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM MM*, 2526–2534.

Zhang, S.; Bai, S.; Chen, G.; Chen, L.; Lu, J.; Wang, J.; and Tang, Y. 2024a. Narrative Action Evaluation with Prompt-Guided Multimodal Interaction. In *CVPR*, 18430–18439.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2024b. Mixture of modality knowledge experts for robust multi-modal knowledge graph completion. *CoRR, abs/2405.16869*.

Zhou, K.; Cai, R.; Ma, Y.; Tan, Q.; Wang, X.; Li, J.; Shum, H. P. H.; Li, F. W. B.; Jin, S.; and Liang, X. 2023a. A Video-Based Augmented Reality System for Human-in-the-Loop Muscle Strength Assessment of Juvenile Dermatomyositis. *IEEE TVCG*, 29(5): 2456–2466.

Zhou, K.; Li, J.; Cai, R.; Wang, L.; Zhang, X.; and Liang, X. 2024. CoFInAl: Enhancing Action Quality Assessment with Coarse-to-Fine Instruction Alignment. In *IJCAI*, 1771–1779.

Zhou, K.; Ma, Y.; Shum, H. P.; and Liang, X. 2023b. Hierarchical graph convolutional networks for action quality assessment. *IEEE TCSVT*, 33(12): 7749–7763.