This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

BriMA: Bridged Modality Adaptation for Multi-Modal Continual Action Quality Assessment

Kanglei Zhou[1], Chang Li[1], Qingyi Pan[2], Liyuan Wang[1,†]
[1] Department of Psychological and Cognitive Sciences, [2] Department of Statistics and Data Science, Tsinghua University

Abstract

*Action Quality Assessment (AQA) aims to score how well an action is performed and is widely used in sports analysis, rehabilitation assessment, and human skill evaluation. Multi-modal AQA has recently achieved strong progress by leveraging complementary visual and kinematic cues, yet real-world deployments often suffer from non-stationary modality imbalance, where certain modalities become missing or intermittently available due to sensor failures or annotation gaps. Existing continual AQA methods overlook this issue and assume that all modalities remain complete and stable throughout training, which restricts their practicality. To address this challenge, we introduce **Bri**dged **M**odality **A**daptation (BriMA), an innovative approach to multi-modal continual AQA under modality-missing conditions. BriMA consists of a memory-guided bridging imputation module that reconstructs missing modalities using both task-agnostic and task-specific representations, and a modality-aware replay mechanism that prioritizes informative samples based on modality distortion and distribution drift. Experiments on three representative multi-modal AQA datasets (RG, Fis-V, and FS1000) show that BriMA consistently improves performance under different modality-missing conditions, achieving 6–8% higher correlation and 12–15% lower error on average. These results demonstrate a step toward robust multi-modal AQA systems under real-world deployment constraints. Our code is available at* `https://github.com/ZhouKanglei/BriMA`.

1. Introduction

Action Quality Assessment (AQA) [11, 12, 31, 36, 40, 64] aims to measure how well an action is performed, supporting applications in sports analysis [50, 51], rehabilitation assessment [3, 63], and skill evaluation [14, 27]. With the rise of multi-modal sensing, recent AQA models [43, 45, 46, 49] increasingly combine visual and other cues such as audio and motion flow, achieving notable progress in accuracy.

† Corresponding author: `liyuanwang@tsinghua.edu.cn`

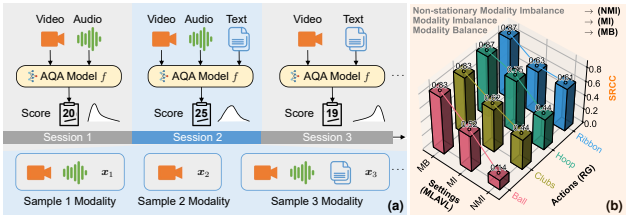


Figure 1. Our motivation: non-stationary modality imbalance **(a)** significantly challenges continual AQA performance **(b)**.

However, real-world AQA rarely operates under fully reliable sensing conditions [20, 42, 58]. Cameras drop frames, sensors degrade over time, annotations are often missing, etc., leading to **non-stationary modality imbalance**, where modality availability varies over time (see Fig. 1(a)). This phenomenon naturally arises in practical AQA workflows and poses a significant obstacle to stable scoring.

Despite its practical prevalence, our empirical study reveals that current AQA efforts fail under such changing modality availability. Multi-modal AQA methods [45, 46] assume fully observed inputs and therefore suffer noticeable failures once modalities become missing (supported by the performance decline in Fig. 1(b) and Appendix Fig. S1). Meanwhile, continual AQA approaches [6, 22, 66] mitigate task-level forgetting but also rely on complete and stable modalities during training, resulting in notable degradation under evolving modality imbalance (see the suboptimal performance of Fs-Aug [22] and MAGR [66] in Tabs. 1 to 3). These failures show that non-stationary modality imbalance is not a single missing-data issue. In AQA, modality changes reshape both the feature space and its temporal dynamics, forcing the model to cope with simultaneous input sparsity and shifting task distributions. This interaction destabilizes representation learning and amplifies forgetting, ultimately degrading scoring reliability.

The inherent complexity of AQA makes this problem fundamentally different from related challenges in recognition or general multi-modal learning [39, 41, 70]. AQA requires fine-grained score reasoning, where subtle temporal cues directly determine the score [45, 46]. This tight coupling between modalities and score semantics breaks the assumptions underlying standard completion techniques.

Simple imputation [42] introduces bias, retrieval-based reconstruction [19, 41] pulls mismatched contexts, and generative synthesis [9, 26] fails to preserve score-critical geometry under limited supervision. While these methods can restore missing modalities at the representation level, they distort the scoring manifold and undermine the ranking consistency essential for AQA. Consequently, directly applying these techniques to multi-modal continual AQA yields only suboptimal performance (see Tab. 5 and Fig. 5), highlighting the need for a tailored solution that explicitly respects the score-sensitive nature of this task.

To address these challenges, we propose **Bri**dged **M**odality **A**daptation (BriMA), an innovative approach to handling evolving modality imbalance and distribution shifts in multi-modal continual AQA. Our core idea is to construct a stable bridging space that aligns missing modalities with shared task structure and memory, enabling reliable modality completion and drift-aware adaptation that existing imputation, retrieval, and generative strategies cannot provide. BriMA is instantiated with two components. The first is a memory-guided bridging module that retrieves structurally aligned exemplars from past tasks and predicts only a minimal residual correction instead of full-feature synthesis, ensuring stable and score-faithful modality completion. The second is a modality-aware replay module that maintains a curated buffer of representative samples with reliable modalities and balanced score coverage, and dynamically prioritizes replay by modality distortion and score drift, effectively counteracting distribution shifts over time.

Experiments on three representative multi-modal AQA datasets (RG [57], Fis-V [30], and FS1000 [43]) show that BriMA consistently improves scoring accuracy and mitigates forgetting under heterogeneous and incomplete modality conditions. On average, it boosts rank correlation by 6.1%, 8.3%, and 1.4%; reduces error by 12.7%, 15.3%, and 6.4%; and lowers relative error by 13.9%, 14.1%, and 5.2% on the three datasets, establishing a strong baseline for multi-modal continual AQA under realistic scenarios.

Our contributions can be summarized as follows:

- We identify non-stationary modality imbalance in multi-modal continual AQA and show its practical significance.
- We propose BriMA, consisting of memory-guided bridging imputation and modality-aware replay to stabilize modality completion and mitigate distribution shifts.
- Extensive experiments on RG, Fis-V, and FS1000 show consistent gains in correlation and error, establishing a strong baseline for this realistic setting.

2. Related Work

Action Quality Assessment (AQA) [28–30, 47, 52, 65, 69] aims to automatically evaluate how well an action is performed and has been widely applied in sports scoring [1, 25], skill assessment [7, 27], and rehabilitation [3, 21, 63]. Existing efforts evolve from hand-crafted representations to deep video models based on CNNs [44], RNNs [16], 3D ConvNets [44, 62], and Transformers [2], as well as ranking and contrastive supervision [17]. Despite these advances, AQA remains challenging due to its fine-grained, score-sensitive nature, and large variations across action types [50, 51, 64]. Recent studies demonstrate that multi-modal cues such as pose [8, 13, 32], flow [56], audio [43, 46], or gaze [14] can significantly enhance AQA. Fusion and alignment strategies, including cross-attention, graph modeling [45, 48], and correlation-based alignment, have been proposed to leverage complementary information. However, real-world AQA data often exhibit modality imbalance, missing-modality cases, and inconsistent modality quality, whereas existing multi-modal AQA methods generally assume complete and synchronized inputs [64]. Although some studies [6, 22, 66] adopt continual learning techniques to mitigate task-wise non-stationarity, they are limited to single-modality settings and do not address the additional challenges introduced by modality heterogeneity with missing modalities in real-world scenarios.

Continual Learning (CL) [37–39] seeks to incrementally learn new tasks without forgetting previously learned knowledge. Prior work spans regularization-based [18, 59], replay-based [4, 34, 54], and architecture-based [35, 37] methods. Although effective on classification benchmarks, these methods struggle with the fine-grained regression, heterogeneous distributions, and modality inconsistencies inherent to AQA. Existing multi-modal CL methods [15, 24, 61] do not address modality imbalance issues, leaving multi-modal continual AQA largely unexplored.

Incomplete Multi-Modal Learning [42, 70] deals with scenarios where one or more modalities are missing during training or inference. Existing efforts typically rely on simple imputation strategies such as zero-filling, feature masking, or hallucination networks that reconstruct missing modalities from available ones. More advanced methods [19, 20, 53, 58] employ cross-modal alignment or shared latent spaces to mitigate modality gaps, yet these methods assume stationary data distributions and consistent cross-modal relationships. In contrast, our approach addresses incomplete multi-modal learning in a continual setting, where modality availability evolves across tasks and distribution shifts amplify modality imbalance.

3. Problem Definition and Notations

We define the multi-modal continual AQA setting with non-stationary modality imbalance and introduce the notations.

Multi-Modal AQA. Traditional multi-modal AQA methods [43, 45, 46] operate under a fully observed multi-modality setting. An action instance is represented by M heterogeneous modalities $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ with an associated score $y \in \mathbb{R}$. Each modality $\boldsymbol{x}^m$ is encoded as a feature vec-

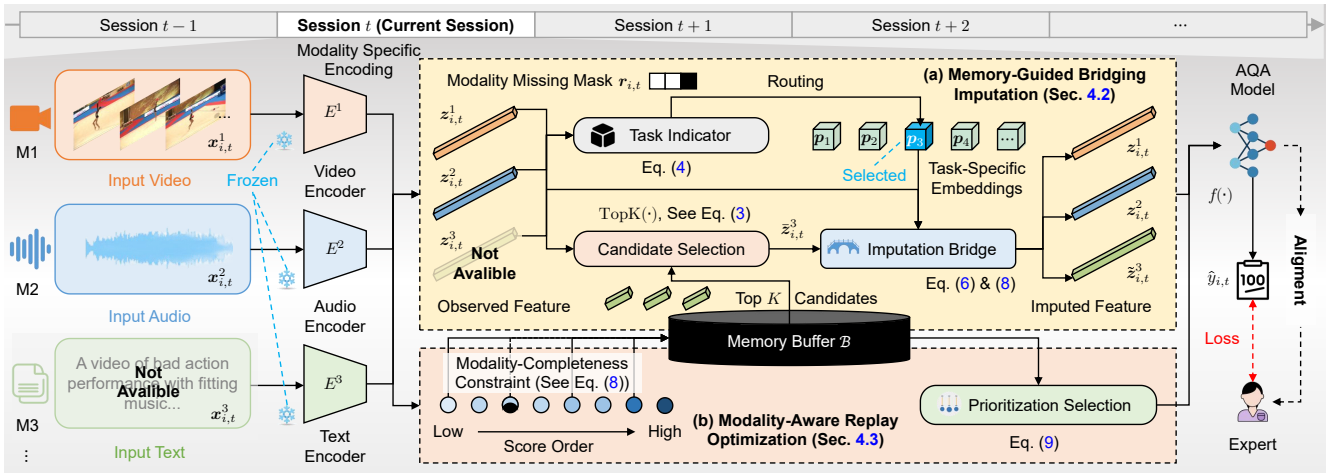


Figure 2. Overview of BriMA. At each session, incomplete multi-modal inputs are encoded and completed via memory-guided bridging imputation (see Sec. 4.2), then scored by the AQA model. The modality-aware replay optimization (see Sec. 4.3) selects and updates representative prototypes in the memory bank to maintain consistency in continual adaptation across tasks.

tor $\boldsymbol{z}^m = E^m(\boldsymbol{x}^m)$, and the AQA model f jointly performs feature fusion and score prediction as $\hat{y} = f(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^M)$.

Continual AQA. Existing continual AQA methods [6, 22, 66] consider a sequence of tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$, where each task $\mathcal{T}_t$ provides samples for score prediction. The model updates its parameters θ_f using only data from the current task, while being required to retain performance on all previously learned tasks. A standard formulation is

$$\min_{\theta_f} \; \mathcal{L}_{\text{score}} + \lambda_{\text{mem}} \, \mathcal{L}_{\text{mem}}, \tag{1}$$

where $\mathcal{L}_{\text{score}}$ is typically a mean-squared error loss on the current task, and $\mathcal{L}_{\text{mem}}$ (see Eq. (10)) leverages a memory buffer $\mathcal{B}_{t-1}$ to regularize updates and mitigate forgetting.

Non-Stationary Modality Imbalance. The above formulations assume that all modalities remain fully observed across tasks, which rarely holds in real-world AQA. For a sample (i,t) at task t, only a subset of modalities $\mathcal{O}_{i,t} \subseteq \{1, \ldots, M\}$ may be available, while the remaining $\mathcal{M}_{i,t} = \{1, \ldots, M\} \setminus \mathcal{O}_{i,t}$ are missing due to sensor failures or annotation gaps. When $m \in \mathcal{M}_{i,t}$, the modality input $\boldsymbol{x}^m_{i,t}$ and its feature $\boldsymbol{z}^m_{i,t}$ are absent, altering both the input structure and the underlying data distribution. As the missing pattern $\mathcal{M}_{i,t}$ evolves over tasks, the model must cope with simultaneous input sparsity and temporal distribution drift, posing a modality-driven CL challenge that existing multi-modal and continual AQA methods cannot handle.

Multi-Modal Continual AQA. We consider multi-modal continual AQA under non-stationary modality imbalance, where each task $\mathcal{T}_t$ provides samples $\mathcal{D}_t = \{(\{\boldsymbol{x}^m_{i,t}\}_{m \in \mathcal{O}_{i,t}}, y_{i,t})\}_{i=1}^{N_t}$ with only a subset of modalities $\mathcal{O}_{i,t}$ observed and the rest $\mathcal{M}_{i,t}$ missing. Observable modalities are encoded as $\boldsymbol{z}^m_{i,t} = E^m(\boldsymbol{x}^m_{i,t})$, while missing ones require reconstructed features $\tilde{\boldsymbol{z}}^m_{i,t}$. Compared with continual AQA, multi-modal continual AQA additionally learns both the scoring parameters θ_f and the reconstruction parameters θ_g, which are jointly optimized to produce accurate score predictions while maintaining cross-task stability. This leads to the optimization objective

$$\min_{\theta_f, \theta_g} \; \mathcal{L}_{\text{score}} + \lambda_{\text{mem}} \, \mathcal{L}_{\text{mem}} + \lambda_{\text{rec}} \, \mathcal{L}_{\text{rec}}, \tag{2}$$

where $\mathcal{L}_{\text{score}}$ and $\mathcal{L}_{\text{mem}}$ follow Eq. (1), and $\mathcal{L}_{\text{rec}}$ (see Eq. (7)) enforces feature-level reconstruction for missing modalities. Unlike raw-data imputation, we store and reconstruct lightweight feature embeddings for efficiency and privacy.

4. BriMA: Bridged Modality Adaptation

We propose **Bri**dged **M**odality **A**daptation (BriMA) for multi-modal continual AQA under practical modality imbalance, outlining its core idea and two main components.

4.1. Core Idea and Framework Overview

Challenges. Multi-modal continual AQA is challenging because evolving modality availability reshapes the feature space, while score prediction is highly sensitive to these shifts, leading to representation drift that interacts with continual forgetting. Existing strategies fall short: simple imputation [42] amplifies noise in score estimation, retrieval-based completion [19, 41] often pulls mismatched scoring contexts that disrupt representation alignment, and generative synthesis [9, 26] cannot maintain score fidelity under limited supervision common in AQA. These errors accumulate across tasks and interact with CL dynamics, making it difficult to jointly achieve accurate scoring in Eq. (2).

Core Idea. BriMA introduces a stable bridging space that anchors incomplete modalities to shared task structure and cross-task memory. By reconstructing missing features through exemplar-aligned residuals and reinforcing temporal consistency via modality-aware replay, BriMA produces

robust, drift-resistant scoring. Grounding completion in this structured space avoids the noise of imputation and the instability of generative synthesis, preventing error accumulation and maintaining reliable performance across tasks.

Framework Overview. The framework of BriMA is illustrated in Fig. 2. At session t, observed modalities produce features $\boldsymbol{z}_{i,t}^m = E^m(\boldsymbol{x}_{i,t}^m)$, while missing modalities $\mathcal{M}_{i,t}$ require reconstructed features $\tilde{\boldsymbol{z}}_{i,t}^m$. The Memory-guided Bridging Imputation (MBI) module (see Sec. 4.2) generates these reconstructed features by retrieving structurally aligned exemplars from the memory $\mathcal{B}_{t-1}$ and predicting a residual correction that anchors each modality to a stable, task-consistent bridging space. The scoring model f then performs multi-modal fusion and prediction jointly as $\hat{y}_{i,t} = f(\{\boldsymbol{z}_{i,t}^m\}_{m\in\mathcal{O}_{i,t}}, \{\tilde{\boldsymbol{z}}_{i,t}^m\}_{m\in\mathcal{M}_{i,t}})$. To maintain temporal stability, the Modality-aware Replay Optimization (MRO) module (see Sec. 4.3) selects informative samples from $\mathcal{B}_{t-1}$ based on modality distortion and score drift, replaying them to counteract distribution shift and forgetting. After learning on task t, representative and reliable samples are updated into the memory. A complete description of the training pipeline is given in **Appendix** Algorithm 1.

4.2. MBI: Memory-Guided Bridging Imputation

Design Idea. MBI targets the core difficulty of modality absence in multi-modal continual AQA, where direct imputation [42, 58] amplifies noise, retrieval-based filling [19, 41] retrieves mismatched scoring contexts, and generative synthesis [9, 26] is unreliable under limited supervision. To overcome these limitations, MBI constructs a task-consistent bridging space that anchors incomplete inputs to stable cross-task structure. It operates in three steps: (1) retrieve structurally aligned exemplars from memory, (2) estimate a residual correction rather than full-feature synthesis, and (3) combine task-specific cues to regularize reconstruction. This design yields noise-resistant and score-preserving modality completion suitable for continual adaptation.

Candidate Selection. For each missing modality $m \in \mathcal{M}_{i,t}$, MBI retrieves a set of K exemplar features $\{\boldsymbol{z}_{j,t'}^m\}_{j=1}^K$ from the memory buffer $\mathcal{B}_{t-1}$, where $t' < t$ indexes previous continual sessions. Retrieval is performed by computing the cosine similarity between the current observed representation $\boldsymbol{z}_{i,t}^{\mathcal{O}}$ and stored cross-modal embeddings:

$$s_{j,t'} = \frac{\langle \boldsymbol{z}_{i,t}^{\mathcal{O}},\, \boldsymbol{z}_{j,t'}^{\mathcal{O}} \rangle}{\|\boldsymbol{z}_{i,t}^{\mathcal{O}}\|\,\|\boldsymbol{z}_{j,t'}^{\mathcal{O}}\|}, \quad \{\boldsymbol{z}_{j,t'}^m\}_{j=1}^K = \mathrm{TopK}_j(s_{j,t'}). \tag{3}$$

This selects semantically aligned exemplars from earlier tasks, ensuring that modality reconstruction is anchored to contextually consistent cross-modal structure.

Task Indicator. We introduce a task indicator to specify which modalities are missing and to provide lightweight task-specific conditioning for the bridging network. In this study, we follow the standard assumption that the missing-modality set $\mathcal{M}_{i,t}$ is known during training, so no detector needs to be learned. Accordingly, we define a binary mask

$$\boldsymbol{r}_{i,t} \in \{0,1\}^M, \quad r_{i,t}^m = \mathbb{1}[m \in \mathcal{M}_{i,t}], \tag{4}$$

and use it to select a task-specific embedding $\boldsymbol{p}_t^m$ from a learnable pool $\mathcal{P} = \{\boldsymbol{p}_t^1, \ldots, \boldsymbol{p}_t^M\}$. The conditioning vector

$$\boldsymbol{c}_t^m = \mathrm{MLP}([\boldsymbol{p}_t^m]) \tag{5}$$

modulates the bridging network for modality m, ensuring that reconstruction remains consistent with the task domain under evolving modality imbalance.

Imputation Bridge. Given the retrieved exemplar set $\{\boldsymbol{z}_{j,t'}^m\}_{j=1}^K$ with similarity scores $\{s_{j,t'}\}$, we compute normalized exemplar weights $w_j = \mathrm{softmax}(s_{j,t'})$ and form an initial estimate $\bar{\boldsymbol{z}}_{i,t}^m = \sum_{j=1}^K w_j\, \boldsymbol{z}_{j,t'}^m$. Conditioned on $\boldsymbol{z}_{i,t}^{\mathcal{O}}$ and the task-specific vector $\boldsymbol{c}_t^m$, the bridging network learns a residual refinement:

$$\Delta\boldsymbol{z}_{i,t}^m = B_\Theta\big(\boldsymbol{z}_{i,t}^{\mathcal{O}},\, \bar{\boldsymbol{z}}_{i,t}^m,\, \boldsymbol{c}_t^m\big), \quad \tilde{\boldsymbol{z}}_{i,t}^m = \bar{\boldsymbol{z}}_{i,t}^m + \Delta\boldsymbol{z}_{i,t}^m. \tag{6}$$

The reconstruction loss enforces alignment between reconstructed and ground-truth features:

$$\mathcal{L}_{\text{rec}} = \big\|\tilde{\boldsymbol{z}}_{i,t}^m - \boldsymbol{z}_{i,t}^m\big\|_2^2, \tag{7}$$

where the weighted exemplar prior offers a well-justified starting point and the residual correction adapts to task-dependent modality shifts, yielding semantically consistent imputation under non-stationary modality imbalance.

4.3. MRO: Modality-Aware Replay Optimization

Design Idea. MRO is designed to address the combined challenges of catastrophic forgetting and evolving modality imbalance in continual AQA. Unlike uniform replay [22, 66], which indiscriminately reuses all stored samples and propagates noisy or distorted modalities, MRO performs reliability-aware replay by (1) maintaining a clean, modality-balanced memory, (2) selecting past samples based on modality distortion and score drift, and (3) prioritizing those most critical for stabilizing prediction under shifting modality availability. This strategy ensures that only informative exemplars contribute to updating the model, resulting in more stable cross-task scoring.

Sample Selection. After completing task $\mathcal{T}_t$, we construct a compact and reliable memory buffer $\mathcal{B}_t$ for future replay. All training samples are first ranked by their predicted scores $\hat{y}_{i,t}$ and partitioned into Q quantile bins. From each bin, we aim to uniformly select instances that satisfy the modality-completeness constraint:

$$\mathcal{B}_t = \Big\{(\boldsymbol{x}_{i,t}, y_{i,t}) \;\Big|\; i \in \mathcal{S}_t,\; \sum_{m=1}^{M} \mathbb{1}[m \in \mathcal{M}_{i,t}] = 0\Big\}, \tag{8}$$

Table 1. Performance comparison on the RG dataset. **Bold** values indicate the best results. Joint Training (JT) and Sequential Training (ST) denote the upper and lower bounds, while rehearsal-free and rehearsal-based methods represent different CL strategies. ↑: higher is better; ↓: lower is better. Average SRCC is computed using Fisher-z transformation.

Method	Publisher	SRCC (↑)					MSE (↓)					RL2 (↓)				
		Ball	Clubs	Hoop	Ribbon	Avg.	Ball	Clubs	Hoop	Ribbon	Avg.	Ball	Clubs	Hoop	Ribbon	Avg.
JT-MLAVL[46]	CVPR'25	0.826	0.829	0.871	0.866	0.849	5.57	4.20	4.11	3.99	4.47	N/A	N/A	N/A	N/A	N/A
Modality Missing Rate $\beta = 10\%$																
ST-MLAVL [46]	CVPR'25	0.480	0.575	0.659	0.664	0.599	14.34	9.19	7.26	8.98	9.94	3.829	4.311	2.734	3.358	3.558
SI [59]	ICML'17	0.481	0.597	0.605	0.685	0.597	14.61	8.70	8.63	8.74	10.17	3.903	4.083	3.249	3.271	3.626
EWC [18]	PNAS'17	0.524	0.594	0.597	0.692	0.605	13.35	9.41	9.16	9.11	10.26	3.565	4.415	3.448	3.409	3.709
LwF [23]	TPAMI'17	0.484	0.468	0.542	0.664	0.545	16.43	10.36	16.49	8.17	12.86	4.387	4.861	6.206	3.056	4.628
MER [34]	ICLR'19	0.594	0.773	0.704	0.786	0.722	**8.33**	6.06	8.22	5.00	6.77	**2.223**	2.843	3.093	1.870	2.473
DER++ [4]	NeurIPS'20	0.543	0.605	**0.727**	0.753	0.665	12.92	8.29	9.86	7.31	9.60	3.451	3.890	3.712	2.736	3.447
NC-FSCIL [54]	ICLR'23	0.457	0.620	0.664	0.731	0.628	14.10	9.06	7.70	9.40	10.06	3.766	4.248	2.899	3.515	3.607
SLCA [60]	ICCV'23	0.467	0.569	0.695	0.744	0.631	18.52	8.31	23.85	8.52	14.80	4.946	3.898	8.976	3.186	5.251
Fs-Aug [22]	TCSVT'24	0.463	0.585	0.703	0.703	0.623	14.89	8.46	12.56	8.51	11.10	3.976	3.970	4.728	3.182	3.964
MAGR [66]	ECCV'24	0.460	0.606	0.672	0.720	0.623	13.85	8.52	9.23	8.32	9.98	3.700	3.996	3.475	3.113	3.571
ASAL [67]	TVCG'25	0.552	0.629	0.693	0.678	0.641	13.37	10.02	9.30	8.35	10.26	3.571	4.699	3.501	3.123	3.724
BriMA (Ours)	–	**0.648**	**0.788**	0.710	**0.836**	**0.746**	8.65	**4.88**	**7.12**	**4.39**	**6.26**	2.310	**2.291**	**2.681**	**1.641**	**2.231**
Modality Missing Rate $\beta = 25\%$																
ST-MLAVL [46]	CVPR'25	0.406	0.576	0.549	0.674	0.558	16.89	8.42	9.39	9.15	10.96	4.512	3.949	3.536	3.423	3.855
SI [59]	ICML'17	0.409	0.589	0.584	0.649	0.564	17.18	8.65	**7.81**	9.33	10.74	4.587	4.057	**2.940**	3.489	3.769
EWC [18]	PNAS'17	0.379	0.580	0.638	0.652	0.571	16.85	9.13	8.91	9.60	11.12	4.499	4.284	3.354	3.592	3.932
LwF [23]	TPAMI'17	0.448	0.586	0.487	0.642	0.545	13.55	7.55	9.98	9.05	10.03	3.618	3.541	3.757	3.384	3.575
MER [34]	ICLR'19	0.527	0.766	0.519	0.761	0.660	**9.39**	7.10	13.36	5.29	8.79	**2.509**	3.329	5.028	1.980	3.212
DER++ [4]	NeurIPS'20	0.528	0.574	0.683	0.739	0.639	14.97	8.33	10.94	7.91	10.54	3.999	3.910	4.119	2.958	3.746
NC-FSCIL [54]	ICLR'23	0.422	0.608	0.615	0.721	0.601	15.04	8.85	9.87	9.07	10.71	4.016	4.150	3.717	3.394	3.819
SLCA [60]	ICCV'23	0.397	0.551	**0.720**	0.715	0.612	18.92	8.44	24.82	9.30	15.37	5.054	3.962	9.343	3.480	5.460
Fs-Aug [22]	TCSVT'24	0.472	0.565	0.607	0.699	0.592	17.04	8.56	16.65	9.23	12.87	4.551	4.016	6.268	3.453	4.572
MAGR [66]	ECCV'24	0.364	0.542	0.589	0.695	0.558	15.11	8.63	11.93	8.39	11.02	4.035	4.051	4.490	3.140	3.929
ASAL [67]	TVCG'25	0.505	0.609	0.676	0.682	0.623	14.36	10.15	10.31	8.24	10.77	3.835	4.762	3.880	3.084	3.890
BriMA (Ours)	–	**0.592**	**0.771**	0.522	**0.794**	**0.686**	9.43	**6.10**	10.11	**4.92**	**7.64**	2.519	**2.860**	3.806	**1.841**	**2.757**
Modality Missing Rate $\beta = 50\%$																
ST-MLAVL [46]	CVPR'25	0.142	0.436	0.444	0.613	0.422	23.42	8.67	12.21	10.08	13.59	6.255	4.067	4.595	3.770	4.672
SI [59]	ICML'17	0.163	0.441	0.463	0.622	0.436	23.39	8.69	**9.31**	10.30	12.92	6.248	4.078	**3.503**	3.854	4.421
EWC [18]	PNAS'17	0.046	0.410	0.370	0.528	0.349	25.13	8.95	12.72	12.47	14.82	6.712	4.201	4.786	4.664	5.091
LwF [23]	TPAMI'17	0.136	0.227	0.310	0.628	0.342	24.18	14.04	12.91	9.02	15.04	6.458	6.584	4.858	3.373	5.318
MER [34]	ICLR'19	0.499	**0.752**	0.389	0.662	0.594	10.55	5.31	15.15	8.69	9.92	2.818	2.493	5.701	3.249	3.565
DER++ [4]	NeurIPS'20	0.225	0.510	0.467	0.571	0.452	19.68	8.74	11.32	9.73	12.37	5.257	4.100	4.262	3.639	4.315
NC-FSCIL [54]	ICLR'23	0.226	0.561	0.499	0.625	0.490	19.37	9.44	12.41	11.24	13.11	5.172	4.428	4.669	4.205	4.619
SLCA [60]	ICCV'23	0.198	0.508	0.406	0.658	0.458	21.03	8.52	24.52	9.72	15.95	5.618	3.997	9.230	3.637	5.620
Fs-Aug [22]	TCSVT'24	0.216	0.508	0.535	0.651	0.492	20.09	8.95	17.78	9.84	14.17	5.366	4.199	6.692	3.680	4.984
MAGR [66]	ECCV'24	0.210	0.520	0.423	0.668	0.471	16.45	8.64	12.99	9.92	12.00	4.393	4.054	4.888	3.712	4.262
ASAL [67]	TVCG'25	0.305	0.543	0.544	0.531	0.486	17.30	10.23	10.90	10.34	12.19	4.620	4.801	4.104	3.868	4.348
BriMA (Ours)	–	**0.569**	0.739	**0.583**	**0.716**	**0.659**	**9.77**	**5.13**	10.85	**6.99**	**8.18**	**2.609**	**2.407**	4.085	**2.614**	**2.928**

where $\mathcal{S}_t$ denotes selected indices across bins. Suppose a bin contains no samples satisfying the constraint, we iteratively re-rank the unselected samples, re-partition them, and repeat the selection until the buffer is filled.

Replay Prioritization. When learning the next task $\mathcal{T}_{t+1}$, replay samples are prioritized according to their vulnerability to continual drift. We compute (i) the modality distortion $d_i = \|\tilde{\boldsymbol{z}}^m_{i,t} - \boldsymbol{z}^m_{i,t}\|_2^2$ and (ii) the score drift $\Delta y_i = \left|f_{\theta_f^{t+1}}(\{\boldsymbol{z}^m_{i,t}\}) - f_{\theta_f^t}(\{\boldsymbol{z}^m_{i,t}\})\right|$, where θ_f^t and θ_f^{t+1} denote the scoring-network parameters before and after learning $\mathcal{T}_{t+1}$. Each sample's priority is defined as

$$q_i = \alpha\, d_i + (1-\alpha)\, \Delta y_i, \tag{9}$$

where α controls the trade-off between modality discrepancy and scoring stability. Samples with larger q_i are replayed more frequently, enabling targeted correction for instances most susceptible to drift.

Score Consistency. To further stabilize continual adaptation, we regularize the scoring network against its previous snapshot on stored memory features. Given a replay exemplar with multi-modality embeddings $\{\boldsymbol{z}^m_{i,t}\}$, the loss is

$$\mathcal{L}_{\text{mem}} = \mathbb{E}_{i\in\mathcal{B}_t}\left[\left\|f_{\theta_f^{t+1}}(\{\boldsymbol{z}^m_{i,t}\}) - f_{\theta_f^{t}}(\{\boldsymbol{z}^m_{i,t}\})\right\|_2^2\right]. \tag{10}$$

This term penalizes temporal prediction drift and complements prioritized replay, yielding a more stable continual AQA process under evolving modality imbalance.

5. Experiments

In this section, we present the main results and analyses. Additional results and analyses are provided in **Appendix**.

Table 2. Performance comparison on the Fis-V dataset.

Method	Publisher	SRCC (↑)			MSE (↓)			RL2 (↓)		
		PCS	TES	Avg.	PCS	TES	Avg.	PCS	TES	Avg.
JT-MLAVL[46]	CVPR'25	0.863	0.766	0.823	7.17	19.44	13.31	N/A	N/A	N/A
Modality Missing Rate $\beta = 10\%$										
ST-MLAVL [46]	CVPR'25	0.648	0.566	0.609	17.03	33.31	25.17	2.721	2.795	2.758
SI [59]	ICML'17	0.630	0.556	0.594	16.81	28.77	22.79	2.685	2.414	2.550
EWC [18]	PNAS'17	0.584	0.551	0.568	18.36	33.43	25.89	2.933	2.805	2.869
LwF [23]	TPAMI'17	0.369	0.165	0.270	25.43	52.97	39.20	4.062	4.446	4.254
MER [34]	ICLR'19	0.624	0.416	0.528	19.86	43.13	31.50	3.173	3.620	3.396
DER++ [4]	NeurIPS'20	0.682	0.539	0.616	14.05	33.56	23.81	2.244	2.816	2.530
NC-FSCIL [54]	ICLR'23	0.702	0.540	0.628	15.34	30.08	22.71	2.451	2.524	2.487
SLCA [60]	ICCV'23	0.671	0.501	0.592	24.25	36.53	30.39	3.874	3.066	3.470
Fs-Aug [22]	TCSVT'24	0.664	0.514	0.594	14.65	36.34	25.50	2.341	3.049	2.695
MAGR [66]	ECCV'24	**0.717**	0.488	0.615	17.39	38.18	27.79	2.778	3.204	2.991
ASAL [67]	TVCG'25	0.695	0.457	0.589	14.53	39.12	26.83	2.322	3.283	2.802
BriMA (Ours)	–	0.705	**0.594**	**0.653**	**13.60**	**25.78**	**19.69**	**2.173**	**2.164**	**2.168**
Modality Missing Rate $\beta = 25\%$										
ST-MLAVL [46]	CVPR'25	0.622	0.572	0.598	17.98	30.11	24.05	2.873	2.527	2.700
SI [59]	ICML'17	0.594	0.544	0.569	18.30	31.37	24.83	2.924	2.632	2.778
EWC [18]	PNAS'17	0.625	0.573	0.599	17.69	30.90	24.29	2.826	2.593	2.709
LwF [23]	TPAMI'17	0.216	0.136	0.177	33.66	53.09	43.37	5.377	4.455	4.916
MER [34]	ICLR'19	0.545	0.459	0.503	21.57	38.43	30.00	3.446	3.225	3.336
DER++ [4]	NeurIPS'20	0.649	0.537	0.596	16.85	36.47	26.66	2.691	3.060	2.876
NC-FSCIL [54]	ICLR'23	0.617	0.576	0.597	18.58	31.96	25.27	2.968	2.682	2.825
SLCA [60]	ICCV'23	0.597	0.569	0.583	20.93	41.93	31.43	3.344	3.519	3.431
Fs-Aug [22]	TCSVT'24	0.649	0.551	0.602	21.61	35.30	28.45	3.451	2.962	3.207
MAGR [66]	ECCV'24	0.667	0.513	0.596	16.04	33.30	24.67	2.563	2.795	2.679
ASAL [67]	TVCG'25	0.664	0.519	0.596	20.24	36.20	28.22	3.233	3.037	3.135
BriMA (Ours)	–	**0.688**	**0.632**	**0.661**	**15.37**	**26.94**	**21.15**	**2.455**	**2.261**	**2.358**
Modality Missing Rate $\beta = 50\%$										
ST-MLAVL [46]	CVPR'25	0.386	0.569	0.483	23.41	29.33	26.37	3.740	2.461	3.100
SI [59]	ICML'17	0.493	0.551	0.523	21.82	31.65	26.73	3.486	2.656	3.071
EWC [18]	PNAS'17	0.442	0.479	0.461	31.39	48.39	39.89	5.014	4.061	4.537
LwF [23]	TPAMI'17	0.435	0.257	0.349	25.72	46.07	35.90	4.109	3.866	3.988
MER [34]	ICLR'19	0.494	0.351	0.425	26.59	43.90	35.25	4.248	3.684	3.966
DER++ [4]	NeurIPS'20	0.592	0.530	0.562	18.88	31.12	25.00	3.016	2.611	2.814
NC-FSCIL [54]	ICLR'23	0.630	0.544	0.588	17.58	37.10	27.34	2.808	3.114	2.961
SLCA [60]	ICCV'23	0.441	0.564	0.505	35.12	46.28	40.70	5.610	3.884	4.747
Fs-Aug [22]	TCSVT'24	0.602	0.488	0.547	24.22	41.61	32.92	3.869	3.492	3.681
MAGR [66]	ECCV'24	0.634	0.446	0.547	18.31	35.12	26.72	2.925	2.947	2.936
ASAL [67]	TVCG'25	0.407	0.494	0.451	37.05	38.71	37.88	5.918	3.249	4.583
BriMA (Ours)	–	**0.648**	**0.620**	**0.634**	**15.70**	**28.03**	**21.86**	**2.508**	**2.352**	**2.430**

5.1. Experimental Setting

Datasets. We evaluate BriMA on three representative multi-modal AQA benchmarks, considering three commonly available modalities: video, audio, and textual commentary. The **Rhythmic Gymnastics (RG)** dataset [57] contains 1,000 videos with a relatively small sample size but diverse action types (Ball, Clubs, Hoop, Ribbon), with 200 training and 50 evaluation samples per type. In contrast, the **Figure Skating Video (Fis-V)** dataset [30] offers a moderate-scale collection of 500 videos but includes only two score types: Total Element Score (TES) and Program Component Score (PCS). The **FS1000** dataset [43] is a large-scale benchmark with 1,000 training and 247 validation videos, providing seven component scores (TES, PCS, SS, TR, PE, CO, IN). Following prior works [45, 46], we train separate models for each action or score type and split each action or subscore into $T = 5$ sequential tasks.

Evaluation Metrics. Existing works adopt different evaluation metrics, making direct comparison difficult. To ensure fairness and comprehensiveness, we report all three existing metrics: Spearman's Rank Correlation Coefficient (SRCC), MSE, and RL2. SRCC measures the monotonic relationship between predicted scores $\hat{y}_i$ and ground-truth scores y_i based on their rank correlation, defined as

$$\text{SRCC} = \frac{\sum_{i=1}^{N}(r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sqrt{\sum_{i=1}^{N}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{N}(\hat{r}_i - \bar{\hat{r}})^2}}, \tag{11}$$

where r_i and $\hat{r}_i$ denote the rank values of the ground-truth and predicted scores, respectively. MSE quantifies the average squared prediction error between $\hat{y}_i$ and y_i, while RL2 represents the normalized mean squared error, defined as

$$\text{RL2} = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{(\max(\boldsymbol{y}) - \min(\boldsymbol{y}))^2} \times 100. \tag{12}$$

Table 3. Performance comparison on the FS1000 dataset. **Subscore** results are shown in Appendix Tab. S2.

Method	Publisher	SRCC (↑)			MSE (↓)			RL2 (↓)		
		10%	25%	50%	10%	25%	50%	10%	25%	50%
JT-MLAVL[46]	CVPR'25	**0.90**	N/A	N/A	10.39	N/A	N/A	N/A	N/A	N/A
ST-MLAVL [46]	CVPR'25	0.728	0.724	0.665	32.56	35.47	54.24	1.592	1.645	2.012
SI [59]	ICML'17	0.719	0.725	0.656	36.77	39.83	46.74	1.650	1.631	1.932
EWC [18]	PNAS'17	0.716	0.710	0.635	54.97	56.58	58.06	2.150	2.041	2.354
LwF [23]	TPAMI'17	0.699	0.626	0.583	35.19	34.67	45.76	1.726	2.271	2.625
MER [34]	ICLR'19	0.739	0.710	0.627	29.95	32.58	39.40	1.758	1.818	2.596
DER++ [4]	NeurIPS'20	0.755	0.733	0.668	31.23	41.45	46.17	1.513	1.623	2.003
NC-FSCIL [54]	ICLR'23	0.746	0.728	0.659	32.62	38.59	49.92	1.493	1.643	2.011
SLCA [60]	ICCV'23	0.740	0.719	0.675	53.75	58.41	59.01	2.109	2.133	2.036
Fs-Aug [22]	TCSVT'24	0.741	0.725	0.677	34.81	38.38	47.98	1.637	1.711	1.999
MAGR [66]	ECCV'24	0.730	0.711	0.658	32.75	39.53	42.62	1.609	1.681	2.067
ASAL [67]	TVCG'25	0.739	0.724	0.666	**29.90**	35.61	38.75	1.623	1.667	1.923
BriMA (Ours)	–	**0.756**	**0.740**	**0.698**	33.35	**31.35**	**35.29**	**1.441**	**1.572**	**1.823**

Implementation Details. All experiments are implemented in PyTorch and run on a single NVIDIA RTX 5090 GPU. All models are trained for 50 epochs using Adam optimizer with an initial learning rate of 3×10^{-4}, weight decay of 10^{-4}, and a cosine learning rate scheduler with decay rate 0.01. The batch size is set to 4 for RG and 8 for Fis-V and FS1000. The mini-batch size for replay sampling is 2 and 4 for the respective datasets. Following [66], we set $\lambda_{\text{mem}}, \lambda_{\text{rec}}$ to 1 and the memory buffer size to 50 for all rehearsal methods ($Q = 10$). The replay coefficient is fixed to $\alpha = 0.5$. To simulate non-stationary modality imbalance, we vary the missing-modality rate $\beta \in \{10\%, 25\%, 50\%\}$. For fair comparison, all methods use the same random seed to generate missing-modality patterns. Other implementation details are supplemented in the Appendix.

5.2. Comparisons with State-of-the-Arts

Assessment Performance. Across Tabs. 1 to 3, Sequential Training (ST) [46] suffers the largest drop from the Joint Training (JT) upper bound as β increases, confirming severe forgetting under evolving modality imbalance. Rehearsal-free methods (SI [59], EWC [18], LwF [23]) moderately alleviate forgetting but remain unstable when modalities are corrupted or missing, while rehearsal-based methods (MER [34], DER++ [4], NC-FSCIL [54], SLCA [60], Fs-Aug [22], MAGR [66], ASAL [68]) improve stability yet

Table 4. Efficiency and relative performance comparison on RG (average across all items). Results are reported relative to [46].

Method	SRCC Gain	MSE Drop	RL2 Drop	Params (M)	Training Time (h)	Inference (Sample/s)	Storage (MB)
ST-MLAVL [46]	0%	0%	0%	8.30	0.416	0.078	0.00
EWC [18]	+3.1%	+7.5%	+5.8%	8.30	0.597	0.078	0.00
NC-FSCIL [54]	+5.2%	+13.6%	+9.3%	8.30	1.025	0.078	32.25
MAGR [66]	+9.5%	+22.8%	+15.2%	8.40	1.154	0.080	32.25
ASAL [67]	+11.2%	+26.7%	+19.0%	8.30	0.876	0.084	32.25
BriMA (Ours)	**+14.6%**	**+33.4%**	**+23.1%**	8.80	1.417	0.118	32.25

Table 5. Ablation study on the RG dataset ($\beta = 10\%$).

ID	Setting	SRCC (↑)				
		Ball	Clubs	Hoop	Ribbon	Avg.
1	Ours	0.648	0.788	0.710	0.836	0.746
2	Ours w/o MBI	0.471	0.660	0.693	0.742	0.640
3	Ours w/o Bridge	0.640	0.727	0.656	0.796	0.730
4	Ours w/o Candidate	0.639	0.755	0.628	0.773	0.699
5	Ours w/o MRO	0.640	0.727	0.656	0.775	0.702
		MSE (↓)				
1	Ours	8.65	4.88	7.12	4.39	6.26
2	Ours w/o MBI	10.70	6.80	8.03	9.34	8.72
3	Ours w/o Bridge	7.39	6.82	8.34	5.19	6.94
4	Ours w/o Candidate	7.76	6.57	9.20	5.08	7.15
5	Ours w/o MRO	7.39	6.82	8.34	5.51	7.02
		RL2 (↓)				
1	Ours	2.310	2.291	2.681	1.641	2.231
2	Ours w/o MBI	2.857	3.190	3.024	3.495	3.142
3	Ours w/o Bridge	1.974	3.197	3.140	1.941	2.563
4	Ours w/o Candidate	2.072	3.083	3.461	1.900	2.629
5	Ours w/o MRO	1.974	3.197	3.140	2.060	2.593

fail to balance modality completion and cross-task retention. BriMA consistently achieves the best performance across all datasets and missing rates, surpassing the second-best baseline by 3.8%, 4.5%, and 5.2% in average SRCC on RG, Fis-V, and FS1000, respectively. As β rises from 10% to 50%, our relative improvement further expands by up to 7.1%, demonstrating that the proposed MBI and MRO jointly provide robust generalization and stable scoring under increasingly severe modality degradation.

Computational Efficiency. Due to its lightweight residual reconstruction design, BriMA maintains a favorable efficiency–performance trade-off. As shown in Tab. 4, BriMA achieves the largest SRCC improvement (+14.6%) and error reductions (MSE -33.4%, RL2 -23.1%) with only a modest increase in parameters (+0.10M), training time (+1.001 hours), and inference latency (+0.100 sample/s). These results show that the proposed mechanisms achieve substantial accuracy gains with limited computational overhead.

5.3. Ablation Study

Core Modules. As shown in Tab. 5, removing either MBI or MRO notably degrades performance, verifying that both modules are critical to handle evolving modality imbalance. Without MBI (Row 2), SRCC drops by 10.6% and

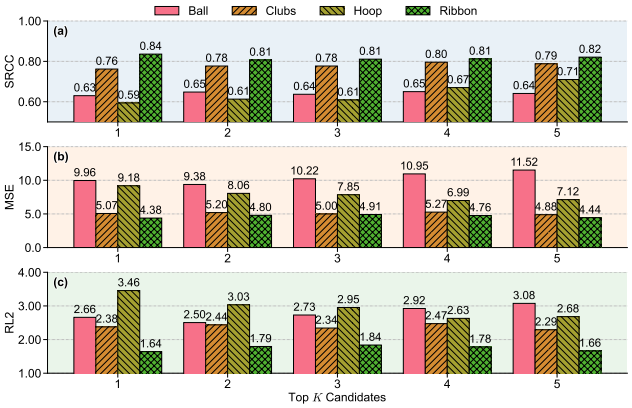


Figure 3. Results of different candidates.

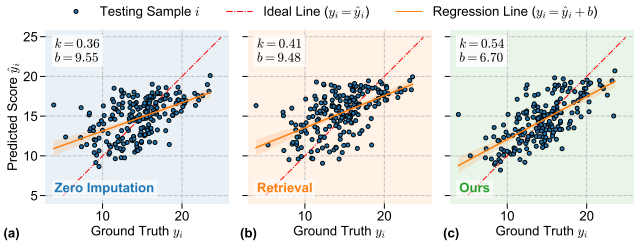


Figure 4. Results of different imputation strategies.

MSE rises by 39.3%, indicating the importance of stable bridging-based imputation. Excluding MRO (Row 5) yields weaker score consistency across tasks, reflected by a 16.2% increase in RL2. Further ablations within MBI reveal complementary roles of the bridge and candidate selection mechanisms. Removing the bridge (Row 3) leads to higher reconstruction bias and temporal instability, while omitting candidate selection (Row 4) causes uneven memory sampling and degraded score correlation. Both components contribute jointly to modality-consistent reconstruction, validating the necessity of the full MBI design.

Effect of Different Candidates. Fig. 3 analyzes the impact of the retrieval size K in the bridging imputation. When retrieval is disabled ($K = 0$, Row 4 in Tab. 5), SRCC notably drops and both MSE and RL2 increase, demonstrating that retrieved exemplars are crucial for stabilizing imputation. Introducing retrieval ($K > 0$) consistently improves all metrics, confirming that exemplar-guided residual correction mitigates reconstruction bias. Across actions, the optimal K varies slightly (e.g., K=2 for Ball, K=4 for Hoop, K=5 for Ribbon), but the differences remain marginal ($< 2\%$ SRCC), indicating that once retrieval is enabled, the generative bridge dominates reconstruction quality rather than the exact number of exemplars.

Effect of Different Imputation Strategies. Fig. 4 compares different imputation strategies on the RG dataset. The zero imputation baseline (see Fig. 4(a)) shows a weak correlation ($\hat{y}_i = 0.36y_i + 0.55$) and large variance, implying severe underfitting and bias accumulation. Retrieval-based

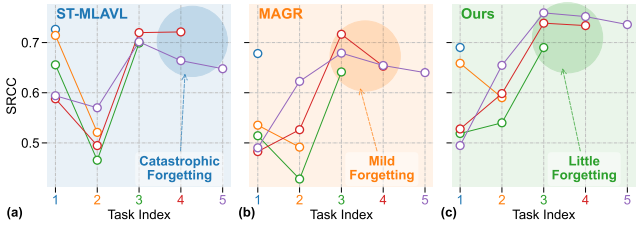


Figure 5. Task-wise SRCC performance comparison.

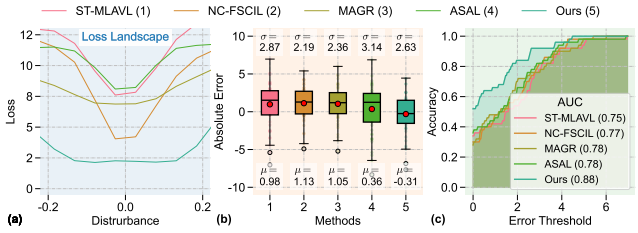


Figure 6. Performance comparisons: (a) loss landscape, (b) error box, (c) accuracy Area Under the Curve (AUC).

imputation (see Fig. 4(b)) achieves moderate improvement ($\hat{y}_i = 0.41y_i + 9.48$) by incorporating exemplar information but still suffers from inconsistent scaling and offset errors. In contrast, our method (see Fig. 4(c)) produces the strongest linearity ($\hat{y}_i = 0.54y_i + 0.70$) with minimal deviation, demonstrating that the proposed bridge-guided imputation effectively aligns reconstructed scores with ground truth while maintaining stable distribution across tasks.

5.4. Visualizations

Forgetting Mitigation. To examine whether forgetting is effectively mitigated, we visualize task-wise SRCC performance in Fig. 5 on RG. In Fig. 5(a), ST-MLAVL [46] suffers from severe forgetting, where after training on the fifth task, the performance on the fourth task drops noticeably compared to that after the fourth task training. With MAGR [66] (see Fig. 5(b)), forgetting is partially alleviated; the fourth-task performance remains stable between the fourth and fifth training stages, although mild degradation still appears on the third task. In contrast, our method (see Fig. 5(c)) shows slight forgetting, exhibiting a stable SRCC increase across tasks 3 and 4. This indicates that our method not only preserves prior task knowledge but also leverages new data to enhance earlier task performance.

Flat Minima. To analyze generalization stability, we perturb model weights ten times and average the resulting loss surfaces across tasks. Fig. 6(a) compares the loss landscapes of ST-MLAVL [46], NC-FSCIL [54], MAGR [66], ASAL [67], and our method. Our model exhibits a flatter and lower loss basin than the others, indicating better convergence stability and stronger generalization under CL.

Error Analysis. As shown in Fig. 6(b), our method achieves the lowest mean absolute error (-0.31) among all compared approaches, substantially outperforming ST-MLAVL [46] (0.98), NC-FSCIL [54] (1.13), MAGR [66] (0.84), and ASAL [67] (0.62). It also yields the most stable predictions with a standard deviation of 2.63, indicating strong robustness to task transitions. Fig. 6(c) further shows AUC improvements, where ours (0.88) clearly surpasses ST-MLAVL (0.75), NC-FSCIL (0.77), MAGR (0.78), and ASAL (0.78), confirming superior accuracy and generalization across continual AQA tasks.



Figure 7. Case study. See Appendix Fig. S2 for more cases.

Case Study. Fig. 7 presents a rhythmic gymnastics ball routine (Ball #022) used to evaluate the scoring accuracy of different methods. The left part shows key frames, while the right part compares predicted scores from ST-MLAVL [46], NC-FSCIL [54], MAGR [66], ASAL [67], and ours against the ground truth (16.70). ST-MLAVL and ASAL achieve close predictions (16.48 and 16.31), while NC-FSCIL and MAGR underestimate (14.29 and 14.19). Our method produces the most consistent visual–score alignment, demonstrating robust perception of fine-grained motions and superior temporal reasoning in complex routines.

6. Conclusion and Discussion

In this work, we present BriMA, a novel approach to multi-modal continual action quality assessment under non-stationary modality imbalance. BriMA integrates memory-guided bridging imputation and modality-aware replay optimization to jointly address missing-modality reconstruction and cross-task stability. Extensive experiments across multiple datasets demonstrate that BriMA consistently outperforms state-of-the-art methods, achieving superior correlation, lower errors, and stronger resistance to forgetting. We believe BriMA offers a general paradigm for building more reliable multi-modal systems, advancing robust perception and evaluation in real-world scenarios.

Limitations and Future Work. BriMA introduces moderate training overhead due to exemplar retrieval, residual bridging, and priority-based replay, achieving an efficiency–stability trade-off. Its pattern-level conditioning may face scalability challenges with larger modality sets. The framework does not explicitly model fine-grained temporal dynamics and assumes observable modality availability. Moreover, while the residual bridging design is conservative, handling weak modalities and unseen missing patterns at scale requires further investigation. Future work will explore more efficient retrieval, adaptive memory compression, lighter reconstruction modules, and temporal modeling to enhance scalability and robustness.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (NSFC, No. 62406160), the Beijing Natural Science Foundation (L247011), the Beijing Major Science and Technology Project (No. Z251100008425003), and the China Postdoctoral Science Foundation (No. 2025M781489).

References

[1] Kumar Ashutosh and Kristen Grauman. Learning skill-attributes for transferable assessment in video. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[2] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European Conference on Computer Vision*, pages 422–438. Springer, 2022. 2

[3] XB Bruce, Yan Liu, Keith CC Chan, and Chang Wen Chen. Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6471–6485, 2024. 1, 2

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. *Advances in Neural Information Processing Systems*, 33:15920–15930, 2020. 2, 5, 6, 3

[5] Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. *arXiv preprint arXiv:2410.10167*, 2024. 5

[6] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *WACV*, pages 42–52, 2024. 1, 2, 3

[7] Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer Nature Switzerland Cham, 2023. 2

[8] Linfeng Dong, Wei Wang, Yu Qiao, and Xiao Sun. Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment. *Advances in Neural Information Processing Systems*, 2025. 2

[9] Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 108–116, 2018. 2, 3, 4

[10] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14314–14324, 2025. 5

[11] Ruisheng Han, Kanglei Zhou, Amir Atapour-Abarghouei, Xiaohui Liang, and Hubert PH Shum. Finecausal: A causal-based framework for interpretable fine-grained action quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6018–6027, 2025. 1

[12] Ruisheng Han, Kanglei Zhou, Shuang Chen, Amir Atapour-Abarghouei, and Hubert PH Shum. Caflow: Enhancing long-term action quality assessment with causal counterfactual flow. *arXiv preprint arXiv:2511.21653*, 2025. 1

[13] Seiji Hirosawa, Takaaki Kato, Takayoshi Yamashita, and Yoshimitsu Aoki. Action quality assessment model using specialists' gaze location and kinematics data—focusing on evaluating figure skating jumps. *Sensors*, 23(22):9282, 2023. 2

[14] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1, 2

[15] Yukang Huo and Hao Tang. When continue learning meets multimodal large language model: A survey. *arXiv preprint arXiv:2503.01887*, 2025. 2

[16] Hiteshi Jain, Gaurav Harit, and Avinash Sharma. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2260–2273, 2020. 2

[17] Chung-In Joung, Seunghwan Byun, and Seungjun Baek. Contrastive learning for action assessment using graph convolutional networks with augmented virtual joints. *IEEE Access*, 2023. 2

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2, 5, 6, 7, 3

[19] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18035–18043, 2025. 2, 3, 4

[20] Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37: 28515–28536, 2024. 1, 2

[21] Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. *arXiv preprint arXiv:2406.08877*, 2024. 2

[22] Yuan-Ming Li, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Continual action assessment via task-consistent score-discriminative feature distribution modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2, 3, 4, 5, 6

[23] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 5, 6, 3

[24] Guoqiang Liang, Chuan Qin, De Cheng, Shizhou Zhang, and Yanning Zhang. Boosting multi-modal alignment: Geometric feature separation for class incremental learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1880–1889, 2025. 2

[25] Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. A figure skating jumping dataset for replay-guided action quality assessment. In *ACM International Conference on Multimedia*, pages 2437–2445, 2023. 2

[26] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2302–2310, 2021. 2, 3, 4

[27] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28952–28962, 2025. 1, 2

[28] Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2241–2244. IEEE, 2016. 2

[29] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019.

[30] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017. 2, 6, 1

[31] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *European Conference on Computer Vision*, pages 105–123. Springer, 2022. 1

[32] Mengshi Qi, Hao Ye, Jiaxuan Peng, and Huadong Ma. Action quality assessment via hierarchical pose-guided multi-stage contrastive regression. *arXiv preprint arXiv:2501.03674*, 2025. 2

[33] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23059–23068, 2025. 5

[34] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019. 2, 5, 6, 3

[35] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 2

[36] Jiahao Wang, Zhengyin Du, Annan Li, and Yunhong Wang. Assessing action quality via attentive spatio-temporal convolutional networks. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 3–16. Springer, 2020. 1

[37] Liyuan Wang, Xingxing Zhang, Qian Li, Mingtian Zhang, Hang Su, Jun Zhu, and Yi Zhong. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence*, 5(12):1356–1368, 2023. 2

[38] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[39] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Hang Su, and Jun Zhu. Hide-pet: continual learning via hierarchical decomposition of parameter-efficient tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2

[40] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *ACM International Conference on Multimedia*, pages 4902–4910, 2021. 1

[41] Xun Wang, Bingqing Ke, Xuanping Li, Fangyu Liu, Mingyu Zhang, Xiao Liang, and Qiushi Xiao. Modality-balanced embedding for video retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2578–2582, 2022. 1, 2, 3, 4

[42] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024. 1, 2, 3, 4

[43] Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *Association for the Advancement of Artificial Intelligence*, pages 2901–2909, 2023. 1, 2, 6

[44] Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran. S3d: Stacking segmental p3d for action quality assessment. In *IEEE International Conference on Image Processing*, pages 928–932. IEEE, 2018. 2

[45] Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *European Conference on Computer Vision*, 2024. 1, 2, 6

[46] Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23967–23977, 2025. 1, 2, 5, 6, 7, 8, 3

[47] Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, Peirong Xu, and Wenzhong Guo. Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8869–8877, 2025. 2

[48] Huangbiao Xu, Huanqi Wu, Xiao Ke, Yuezhou Li, Rui Xu, and Wenzhong Guo. Quality-guided vision-language learning for long-term action quality assessment. *IEEE Transactions on Multimedia*, 2025. 2

[49] Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. *arXiv preprint arXiv:2511.17397*, 2025. 1

[50] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2949–2958, 2022. 1, 2

[51] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14628–14637, 2024. 1, 2

[52] Jinglin Xu, Sibo Yin, and Yuxin Peng. Human-centric fine-grained action quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[53] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446, 2024. 2

[54] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023. 2, 5, 6, 7, 8, 3

[55] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 5

[56] Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 33: 1600–1613, 2024. 2

[57] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM International Conference on Multimedia*, pages 2526–2534, 2020. 2, 6, 1

[58] Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–433. Springer, 2024. 1, 2, 4

[59] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017. 2, 5, 6, 3

[60] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023. 5, 6, 3

[61] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[62] Haoyang Zhou, Teng Hou, and Jitao Li. Prior knowledge-guided hierarchical action quality assessment with 3d convolution and attention mechanism. In *Journal of Physics: Conference Series*, page 012027. IOP Publishing, 2023. 2

[63] Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li, Hubert PH Shum, Frederick WB Li, Song Jin, and Xiaohui Liang. A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2456–2466, 2023. 1, 2

[64] Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert P. H. Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *arXiv preprint arXiv:2412.11149*, 2024. 1, 2

[65] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *International Joint Conference on Artificial Intelligence*, pages 1771–1779, 2024. 2

[66] Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang. Magr: Manifold-aligned graph regularization for continual action quality assessment. In *European Conference on Computer Vision*, pages 375–392, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[67] Kanglei Zhou, Zikai Hao, Liyuan Wang, and Xiaohui Liang. Adaptive score alignment learning for continual perceptual quality assessment of 360-degree videos in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 5, 6, 7, 8, 2, 3

[68] Kanglei Zhou, Qingyi Pan, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Xiaohui Liang, and Liyuan Wang. Continual action quality assessment via adaptive manifold-aligned graph regularization. *arXiv preprint arXiv:2510.06842*, 2025. 6

[69] Kanglei Zhou, Hubert PH Shum, Frederick WB Li, Xinxing Zhang, and Xiaohui Liang. Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE Transactions on Image Processing*, 34:3718–3732, 2025. 2

[70] Qian Zhou, Ting Chen, Hua Zou, and Xuan Xiao. Uncertainty-aware incomplete multimodal fusion for few-shot central retinal artery occlusion classification. *Information Fusion*, 104:102200, 2024. 1, 2