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

Supplementary Material

S1. Training Procedure

During multi-modal continual AQA, BriMA sequentially learns from tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ under evolving modality availability. At session t, for each sample $(\boldsymbol{x}_{i,t}, y_{i,t})$, the MBI module (see Sec. 4.2) reconstructs missing modality embeddings $\tilde{\boldsymbol{z}}_{i,t}^m$ using retrieved exemplars from the memory bank $\mathcal{B}_{t-1}$ and the imputation bridge in Eq. (6), supervised by the reconstruction loss in Eq. (7). The completed multi-modal features $\{\boldsymbol{z}_{i,t}^m\}_{m\in\mathcal{O}_{i,t}} \cup \{\tilde{\boldsymbol{z}}_{i,t}^m\}_{m\in\mathcal{M}_{i,t}}$ are then fed into the scoring network $f_{\theta_f^t}$, which is jointly optimized with the reconstruction network θ_g using the objective in Eq. (2). During optimization, the MRO module (see Sec. 4.3) samples replay instances from $\mathcal{B}_{t-1}$ according to the priority q_i in Eq. (9) and regularizes temporal prediction drift via the consistency loss in Eq. (10). After convergence on $\mathcal{T}_t$, MRO updates the memory bank to $\mathcal{B}_t$ using quantile-based, modality-complete selection in Eq. (8), which then supports the next session.

Algorithm 1: Training procedure of BriMA

Input: Task sequence $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$, initial memory $\mathcal{B}_0 = \emptyset$
Output: Trained parameters θ_f, θ_g

for $t = 1$ **to** T **do**
  // Stage 1: Memory-guided bridging imputation
  foreach *sample* $(\boldsymbol{x}_{i,t}, y_{i,t}) \in \mathcal{T}_t$ **do**
    Encode observed modalities to obtain $\{\boldsymbol{z}_{i,t}^m\}_{m\in\mathcal{O}_{i,t}}$;
    Retrieve top-K exemplar features from $\mathcal{B}_{t-1}$ using Eq. (3);
    Build task indicator $\boldsymbol{r}_{i,t}$ and conditioning $\boldsymbol{c}_t^m$ via Eqs. (4) and (5);
    Compute imputed features $\tilde{\boldsymbol{z}}_{i,t}^m$ for $m \in \mathcal{M}_{i,t}$ using the bridge Eq. (6);
  end
  // Stage 2: Joint optimization with modality-aware replay
  while *not converged* **do**
    Sample a mini-batch from $\mathcal{T}_t$ (with completed features);
    Sample a replay batch from $\mathcal{B}_{t-1}$;
    For each replay sample, compute modality distortion d_i and score drift Δy_i, then priority q_i as in Eq. (9);
    Select high-priority replay instances based on q_i;
    Compute $\mathcal{L}_{\text{score}}$ (see Eq. (1)), $\mathcal{L}_{\text{rec}}$ (see Eq. (7)), and $\mathcal{L}_{\text{mem}}$ (see Eq. (10));
    Update θ_f, θ_g by minimizing Eq. (2);
  end
  // Stage 3: Memory update
  Rank all samples in $\mathcal{T}_t$ by predicted scores $\hat{y}_{i,t}$ and partition into Q quantile bins;
  Within bins, iteratively select modality-complete samples using Eq. (8) until $\mathcal{B}_t$ is filled;
end
return θ_f, θ_g

S2. Additional Implementation Details

Additional Dataset Details We conduct experiments on three large-scale multi-modal datasets: RG, Fis-V, and FS1000. The score ranges and evaluation dimensions of these datasets are summarized in Tab. S1. **Rhythmic Gymnastics (RG)** [57] contains 1,000 videos of rhythmic gymnastics performances involving four apparatuses: ball, clubs, hoop, and ribbon. Each video lasts about 1.6 minutes and is recorded at 25 fps. The dataset is split into 200 training and 50 evaluation videos per action type. Following prior works [45, 46], we train separate models for each apparatus. **Figure Skating Video (Fis-V)** [30] consists of 500 videos of ladies' singles short program performances in figure skating, each approximately 2.9 minutes long and recorded at 25 fps. Following the official split, 400 videos are used for training and 100 for testing. Each video is annotated with two official competition scores: the Total Element Score (TES) and the Program Component Score (PCS). Consistent with previous studies [45, 46], we train separate models for each score type. **Figure Skating 1000 (FS1000)** [43] is a large-scale figure skating dataset comprising 1,000 training and 247 validation videos covering eight competition categories, including men's, ladies', and pairs' short and free programs, as well as rhythm and free dances in ice dance. Each video contains roughly 5,000 frames at 25 fps. FS1000 provides TES, PCS, and five additional component scores: Skating Skills (SS), Transitions (TR), Performance (PE), Composition (CO), and Interpretation of Music (IN). It is the first figure skating dataset designed for audiovisual learning, promoting rule-consistent multi-modal modeling. Following prior works [45, 46], we train separate models for each score type.

Table S1. Score ranges and evaluation dimensions for all datasets.

Dataset	Subcategory	Score Dimension(s)	Range	Notes
RG [57]	Ball	Overall	0 – 25	Four rhythmic events
	Clubs	Overall	0 – 25	
	Hoop	Overall	0 – 25	
	Ribbon	Overall	0 – 25	
Fis-V [30]	TES	Technical Execution Score (TES)	0 – 45	Two judging dimensions
	PCS	Performance Component Score (PCS)	0 – 40	
FS1000 [43]	TES	Technical Execution Score	0 – 130	Seven judging components
	PCS	Performance Component Score	0 – 60	
	SS	Skating Skills	0 – 10	
	TR	Transitions	0 – 10	
	PE	Performance	0 – 10	
	CO	Composition	0 – 10	
	IN	Interpretation	0 – 10	

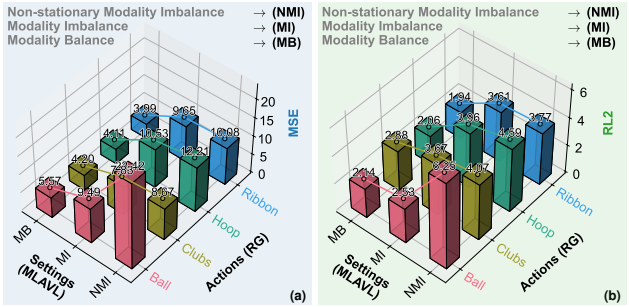

Figure S1. Non-stationary modality imbalance significantly challenges the model performance: (a) L2 and (b) RL2.

Backbone Details. Following prior works [43, 45, 46], we employ widely adopted pretrained encoders for the three modalities considered in this study. For RG and Fis-V, visual features are extracted using the Video Swin Transformer (VST) pretrained on Kinetics-600, and audio features are obtained using the Audio Spectrogram Transformer (AST) pretrained on AudioSet. For FS1000, we use the TimeSformer and AST features provided by [43]. Textual commentary is encoded using ViFi-CLIP, a CLIP model fine-tuned on Kinetics-400, applied to our curated prompt sets. All experiments adopt fixed clip sampling following the standard protocol of each dataset.

S3. Additional Results

Performance Drops Due to Non-Stationary Modality Imbalance. Fig. S1 quantitatively demonstrates the severe impact of evolving modality imbalance across the four rhythmic gymnastics actions. Under the balanced-modality setting, all actions achieve strong correlation (SRCC $\geq$ 0.82) and low error (MSE $\leq$ 5.6, RL2 $\leq$ 2.2), indicating stable and consistent scoring. When modality imbalance is introduced, both correlation and error metrics deteriorate sharply: SRCC drops by an average of 28.6%, while MSE and RL2 increase by 78.1% and 38.2%, respectively. The degradation becomes catastrophic under the non-stationary modality imbalance setting, where missing modalities vary across sessions, leading to an average SRCC decline of 62.4% compared with the balanced baseline. In this case, MSE grows nearly threefold and RL2 more than doubles, for example from 2.14 to 6.26 for the "Ball" routine. These results confirm that traditional AQA pipelines are highly vulnerable to modality instability. Even minor shifts in modality quality or availability can propagate through feature fusion, distort the latent representation $\boldsymbol{h}$, and cause nontrivial score drift. This emphasizes the necessity of explicitly addressing non-stationary modality imbalance in continual AQA, as tackled by our BriMA method.

Sub-Component Performance Comparison on FS1000. Tab. S2 analyzes TES, PCS, SS, TR, PE, CO, and IN under three missing rates $\beta \in \{10\%, 25\%, 50\%\}$. At $\beta = 10\%$, our method achieves the best average SRCC (0.756), slightly surpassing the strongest baseline DER++ [4] (0.755) and delivering the lowest RL2 on average (1.441). Component-wise, it leads SRCC on PCS, SS, TR, and IN, and attains the best MSE on PCS, SS, TR, and CO. Although ASAL [67] reports a lower average MSE (29.90 vs. 33.35), our method maintains better correlation and stability, suggesting that bridging favors consistent ranking when modalities are mildly degraded.

When $\beta = 25\%$, our method becomes clearly dominant across metrics. It yields the highest average SRCC (0.740, a 1.0% improvement over the best baseline DER++ [4] at 0.733), the lowest average MSE (31.35, an 11.9% reduction compared with the best baseline ASAL [67] at 35.61), and the lowest average RL2 (1.572, a 3.1% reduction compared with ASAL [67] at 1.623). Per component, it consistently improves correlation on TES, PCS, TR, CO, and IN, while matching or exceeding baselines on SS and PE. These

Table S2. Performance comparison on the FS1000 dataset (supplementary results of Tab. 3). **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. N/A indicates that the metric was not reported in the paper. Average SRCC is computed using Fisher-z transformation.

Method	Publisher	SRCC (↑)								MSE (↓)								RL2 (↓)							
		TES	PCS	SS	TR	PE	CO	IN	Avg.	TES	PCS	SS	TR	PE	CO	IN	Avg.	TES	PCS	SS	TR	PE	CO	IN	Avg.
JT-MLAVL[46]	CVPR'25	0.92	0.89	0.90	0.90	0.88	0.89	0.88	0.90	64.89	6.39	0.23	0.24	0.50	0.25	0.26	10.39	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Modality Missing Rate $\beta = 10\%$																									
ST-MLAVL [46]	CVPR'25	0.630	0.762	0.740	0.757	0.721	0.744	0.726	0.728	209.44	15.38	0.56	0.54	0.86	0.55	0.61	32.56	1.686	1.084	2.026	1.821	0.892	1.736	1.899	1.592
SI [59]	ICML'17	0.583	0.747	0.737	0.749	0.720	0.754	0.717	0.719	238.47	15.74	0.56	0.56	0.88	0.56	0.63	36.77	1.920	1.110	2.043	1.865	0.909	1.748	1.952	1.650
EWC [18]	PNAS'17	0.639	0.715	0.743	0.740	0.730	0.743	0.689	0.716	358.87	22.05	0.66	0.76	0.90	0.71	0.81	54.97	2.889	1.554	2.377	2.555	0.934	2.232	2.506	2.150
LwF [23]	TPAMI'17	0.610	0.731	0.740	0.706	0.662	0.740	0.688	0.699	228.11	14.72	0.53	0.65	0.99	0.62	0.68	35.19	1.836	1.038	1.932	2.193	1.029	1.940	2.111	1.726
MER [34]	ICLR'19	**0.708**	0.775	0.727	0.762	0.713	0.753	0.733	0.739	**190.91**	15.21	0.65	0.65	0.90	0.57	0.77	29.95	**1.537**	1.072	2.375	2.195	0.938	1.783	2.405	1.758
DER++ [4]	NeurIPS'20	0.672	0.780	0.767	0.762	**0.769**	**0.777**	0.746	0.755	201.08	14.56	0.50	0.56	**0.79**	0.50	0.60	31.23	1.619	1.027	1.803	1.874	**0.823**	1.577	1.872	1.513
NC-FSCIL [54]	ICLR'23	0.659	0.767	0.759	0.781	0.732	0.761	0.747	0.746	211.11	14.29	0.50	0.48	0.81	0.54	**0.56**	32.62	1.700	1.008	1.831	1.615	0.843	1.699	1.756	1.493
SLCA [60]	ICCV'23	0.633	0.770	0.770	0.744	0.739	0.759	0.746	0.740	352.76	19.63	0.64	0.75	1.01	0.67	0.82	53.75	2.840	1.384	2.316	2.500	1.047	2.118	2.557	2.109
Fs-Aug [22]	TCSVT'24	0.646	0.750	0.741	0.749	0.739	0.765	0.778	0.741	222.69	17.88	0.58	0.57	0.81	0.56	0.58	34.81	1.793	1.261	2.112	1.910	0.837	1.751	1.796	1.637
MAGR [66]	ECCV'24	0.682	0.765	0.748	0.759	0.700	0.742	0.704	0.730	211.79	14.26	0.53	0.53	0.88	0.62	0.64	32.75	1.705	1.005	1.938	1.769	0.910	1.948	1.987	1.609
ASAL [67]	TVCG'25	0.681	0.761	0.745	0.743	0.739	0.762	0.735	0.739	191.01	15.00	0.57	0.61	0.88	0.55	0.66	**29.90**	1.538	1.058	2.056	2.035	0.908	1.717	2.051	1.623
BriMA (Ours)	–	0.676	**0.794**	**0.797**	**0.790**	0.736	0.773	**0.784**	**0.756**	206.61	**12.18**	**0.44**	**0.46**	0.82	**0.49**	0.57	33.35	**1.663**	**0.859**	**1.596**	**1.558**	0.852	**1.540**	**1.674**	**1.441**
Modality Missing Rate $\beta = 25\%$																									
ST-MLAVL [46]	CVPR'25	0.659	0.748	0.738	0.712	0.731	0.745	0.725	0.724	229.63	15.50	0.57	0.61	**0.80**	0.55	0.61	35.47	1.849	1.093	2.078	2.042	**0.831**	1.730	1.895	1.645
SI [59]	ICML'17	0.582	0.766	0.738	0.763	0.720	0.743	0.730	0.725	261.25	14.43	0.54	**0.52**	0.84	0.57	0.62	39.83	2.103	1.018	1.955	**1.751**	0.871	1.779	1.939	1.631
EWC [18]	PNAS'17	0.594	0.757	0.729	0.737	0.734	0.690	0.702	0.710	374.76	17.63	0.65	0.61	0.95	0.69	0.80	56.58	3.017	1.243	2.341	2.062	0.989	2.161	2.476	2.041
LwF [23]	TPAMI'17	0.606	0.681	0.615	0.664	0.622	0.551	0.632	0.626	220.81	17.29	1.00	0.73	1.03	1.00	0.84	34.67	1.778	1.219	3.623	2.465	1.064	3.137	2.614	2.271
MER [34]	ICLR'19	**0.676**	0.726	0.685	0.699	0.734	0.710	0.736	0.710	206.59	17.88	0.63	0.79	0.90	0.66	0.60	32.58	1.663	1.261	2.303	2.637	0.937	2.068	1.856	1.818
DER++ [4]	NeurIPS'20	0.610	0.744	**0.772**	0.754	0.728	0.748	0.752	0.733	270.86	16.30	**0.50**	0.53	0.82	0.56	0.59	41.45	2.181	1.149	**1.820**	1.763	0.851	1.753	1.840	1.623
NC-FSCIL [54]	ICLR'23	0.637	0.751	0.743	0.741	0.727	0.757	0.728	0.728	251.10	15.95	0.57	0.56	0.80	0.53	0.61	38.59	2.022	1.124	2.078	1.887	0.833	1.673	1.883	1.643
SLCA [60]	ICCV'23	0.591	0.743	0.729	0.740	0.739	0.727	0.740	0.719	384.94	20.15	0.70	0.69	0.88	0.74	0.74	58.41	3.099	1.421	2.539	2.323	0.913	2.322	2.313	2.133
Fs-Aug [22]	TCSVT'24	0.624	0.773	0.712	0.760	0.733	0.732	0.719	0.725	249.25	16.14	0.62	0.53	0.86	0.59	0.66	38.38	2.007	1.138	2.262	1.785	0.889	1.857	2.038	1.711
MAGR [66]	ECCV'24	0.604	0.750	0.678	0.746	0.725	0.733	0.719	0.711	258.14	15.40	0.61	0.56	0.84	0.57	0.59	39.53	2.078	1.086	2.223	1.883	0.869	1.787	1.841	1.681
ASAL [67]	TVCG'25	0.634	0.745	0.749	0.720	**0.743**	0.731	0.737	0.724	230.75	15.28	0.56	0.63	0.85	0.58	0.60	35.61	1.858	1.077	2.044	2.103	0.883	1.830	1.875	1.667
BriMA (Ours)	–	**0.673**	**0.774**	0.751	**0.776**	0.726	**0.767**	**0.760**	**0.740**	**202.14**	**14.05**	**0.54**	0.57	**0.83**	**0.49**	**0.56**	**31.35**	**1.627**	**0.991**	1.962	1.925	**0.858**	**1.548**	**1.727**	**1.572**
Modality Missing Rate $\beta = 50\%$																									
ST-MLAVL [46]	CVPR'25	0.524	0.727	0.656	0.689	0.653	0.693	0.686	0.665	359.86	16.10	0.67	0.71	0.98	0.67	0.68	54.24	2.897	1.135	2.424	2.385	1.012	2.106	2.125	2.012
SI [59]	ICML'17	0.514	0.678	0.691	0.681	0.660	0.682	0.665	0.656	303.64	19.99	0.63	0.66	0.92	0.68	**0.67**	46.74	2.445	1.409	2.270	2.217	0.951	2.139	**2.092**	1.932
EWC [18]	PNAS'17	0.520	0.697	0.693	0.667	0.638	0.612	0.598	0.635	381.88	20.17	0.67	0.76	1.09	0.97	0.91	58.06	3.074	1.422	2.413	2.552	1.131	3.052	2.832	2.354
LwF [23]	TPAMI'17	0.502	0.614	0.554	0.678	0.559	0.571	0.588	0.583	289.66	25.50	1.18	0.68	1.32	1.03	0.99	45.76	2.332	1.797	4.275	2.268	1.366	3.253	3.082	2.625
MER [34]	ICLR'19	0.579	0.681	0.631	0.605	0.614	0.636	0.638	0.627	247.80	22.81	0.96	1.36	1.16	0.89	0.81	39.40	1.995	1.608	3.481	4.566	1.200	2.794	2.528	2.596
DER++ [4]	NeurIPS'20	0.577	0.731	0.674	0.683	0.659	0.689	0.644	0.668	300.21	19.12	0.66	0.67	1.07	0.67	0.78	46.17	2.417	1.348	2.387	2.244	1.114	2.092	2.417	2.003
NC-FSCIL [54]	ICLR'23	0.511	0.705	0.687	0.691	0.659	0.658	0.673	0.659	328.16	17.55	0.68	0.67	0.95	0.77	0.68	49.92	2.642	1.237	2.456	2.240	0.985	2.418	2.100	2.011
SLCA [60]	ICCV'23	0.508	0.731	0.686	**0.729**	0.684	0.686	0.666	0.675	391.98	17.48	0.67	**0.58**	0.95	0.68	0.75	59.01	3.156	1.232	2.445	**1.940**	0.990	2.146	2.340	2.036
Fs-Aug [22]	TCSVT'24	0.541	**0.751**	0.675	0.692	0.681	0.682	0.691	0.677	316.41	**15.71**	0.72	0.68	0.92	0.71	0.72	47.98	2.547	**1.108**	2.618	2.270	0.954	2.242	2.255	1.999
MAGR [66]	ECCV'24	0.539	0.729	0.670	0.689	0.666	0.623	0.669	0.658	278.10	16.23	0.72	0.73	0.95	0.90	0.71	42.62	2.239	1.144	2.600	2.453	0.982	2.844	2.206	2.067
ASAL [67]	TVCG'25	0.595	0.695	0.681	0.688	0.652	0.667	0.676	0.666	248.96	18.54	0.61	0.68	0.99	0.73	0.75	38.75	2.004	1.307	2.228	2.291	1.026	2.280	2.322	1.923
BriMA (Ours)	–	**0.629**	0.702	**0.730**	0.723	**0.707**	**0.742**	**0.724**	**0.698**	**225.86**	17.59	**0.60**	0.63	**0.89**	**0.58**	**0.70**	**35.29**	**1.818**	1.240	**2.181**	2.107	**0.972**	**1.834**	2.250	**1.823**

gains indicate that memory-guided bridging and modality-aware replay better constrain drift once missing modalities become frequent.

At $\beta = 50\%$, the gap widens as modalities grow scarce. Our method achieves the best average SRCC (0.698, a 3.4% improvement over the next best SLCA [60] at 0.675), the lowest average MSE (35.29, an 8.9% reduction compared with ASAL [67] at 38.75), and the lowest average RL2 (1.823, a 5.2% reduction compared with ASAL at 1.923). Component-level results follow the same trend, with consistent improvements on SS, TR, PE, CO, and competitive TES, PCS, and IN.

Overall, the component analysis reveals three main effects: (1) rank stability, where our method sustains higher SRCC across sub-scores even when low-β MSE trade-offs occur; (2) error containment, where improvements in MSE and RL2 accumulate as β increases, reflecting stronger robustness to missing-modality noise; and (3) uniformity across sub-scores, where performance gains are distributed rather than concentrated, showing that the bridging and replay strategies generalize effectively across TES, PCS, and fine-grained components such as SS, TR, PE, CO, and IN.

Additional Case Study. As shown in Fig. S2, we analyze five representative examples covering both easy and hard cases to further examine the robustness of our method under diverse visual and contextual conditions. Each row displays sampled frames from a video along with the predicted scores of different methods compared with the ground truth. Our method consistently produces the most accurate predictions, validating its stability against varying scene complexity and modality degradation.

For the easy cases (Ball #027 Fig. S2(a), Ball #030 Fig. S2(b), and Clubs #016 Fig. S2(c)), the background and lighting are well-separated from the performer, resulting in clear spatial boundaries and smooth motion cues. Competing methods often exhibit mild overestimation (approximately +1–3 points), primarily due to overfitting to salient motion or contrastive cues. In contrast, our approach aligns closely with ground-truth scores (e.g., 11.77, 11.37, and 12.87), benefiting from temporally coherent cross-modal representations and calibrated replay of rhythmic patterns.

In the hard cases (Ball #184 (see Fig. S2(d)) and Clubs #115 (Fig. S2(e))), the background exhibits strong color and texture similarity to the gymnast's costume, leading to blurred motion boundaries and degraded feature discrimination. As a result, all baseline methods suffer large predic-

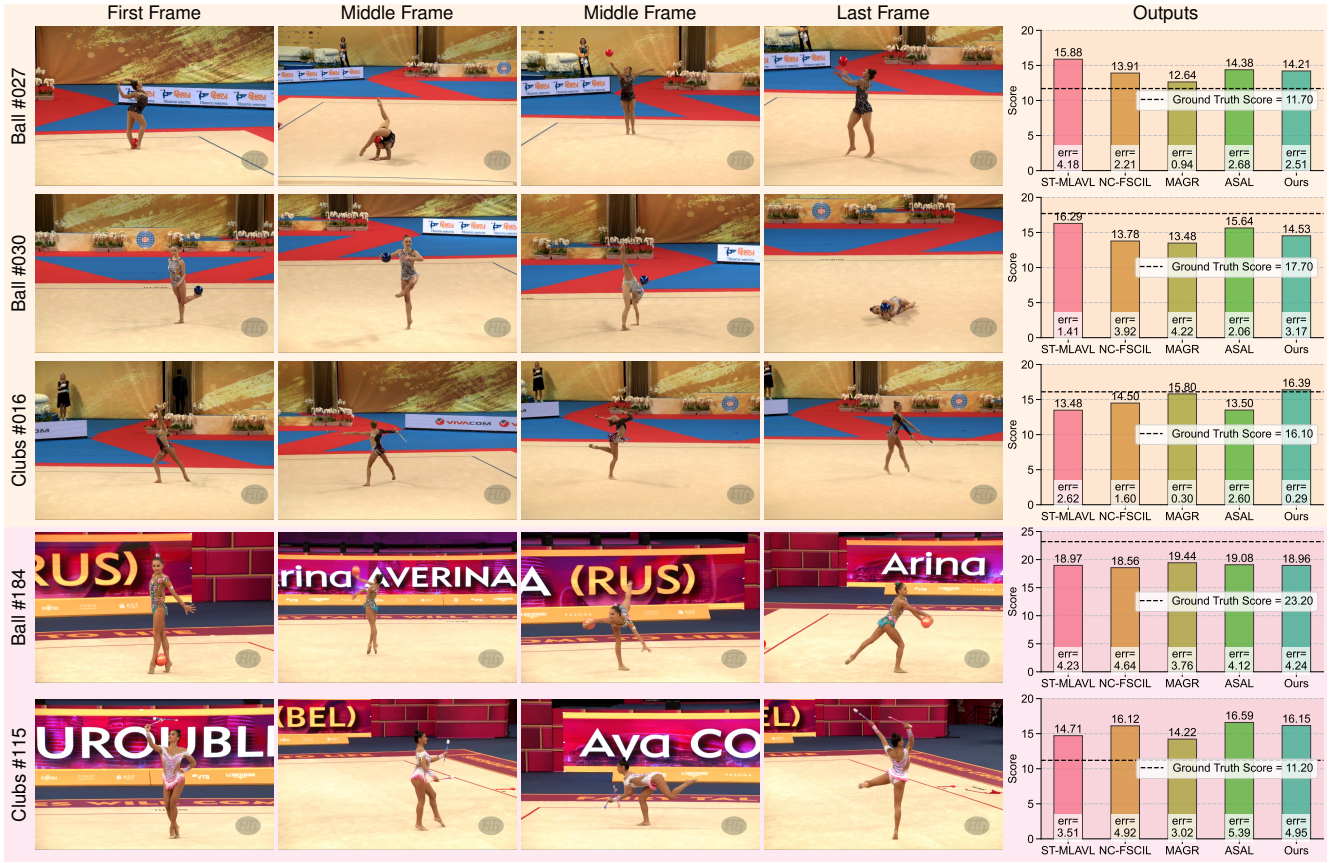

Figure S2. Case study (supplementary results of Fig. 7): (a), (b), and (c) are easy cases, and (d) and (e) are hard cases.

tion errors (-3 to -6 points), failing to accurately localize motion trajectories. Our method effectively mitigates this issue by leveraging the bridging space to restore modality balance and suppress background interference, achieving near-perfect predictions (e.g., 23.30 and 11.20).

Overall, these case studies highlight that when visual similarity between the performer and background increases, most models struggle to distinguish motion boundaries, yielding less precise score predictions. In contrast, our method maintains reliable estimation through memory-guided retrieval and structural bridging, demonstrating strong resilience to complex visual ambiguity and domain imbalance.

Additional Ablation Study. Due to space limitations, the core components of BriMA have not been fully explored in the main paper. The additional ablation results in Tab. S3 provide a more complete understanding of how each module contributes to overall performance. Removing the task-specific embedding (see Eq. (6)) leads to a clear drop in SRCC across all four action types, indicating that modality-aware conditioning is essential for stabilizing residual reconstruction under shifting observation patterns. Disabling replay prioritization (see Eq. (9)) also reduces performance because the model can no longer focus on samples that exhibit large modality distortion or score drift, resulting in weaker temporal consistency and higher regression error. These results verify that both the bridging design and the replay mechanism are important for maintaining robust scoring in continual multi-modal settings. The ablations further highlight that BriMA's improvements do not come from a single component but from the cooperation between memory-guided imputation and informed replay.

Table S3. Ablation study on the RG dataset ($\beta = 10\%$).

ID	Setting	Ball	Clubs	Hoop	Ribbon	Avg.
		SRCC (↑)				
1	Ours	0.648	0.788	0.710	0.836	0.746
2	Ours w/o Task-Specific Embedding	0.579	0.702	0.656	0.834	0.693
3	Ours w/o Replay Prioritization	0.639	0.776	0.628	0.791	0.717
		MSE (↓)				
1	Ours	8.65	4.88	7.12	4.39	6.26
2	Ours w/o Task-Specific Embedding	10.165	6.722	7.510	3.709	7.027
3	Ours w/o Replay Prioritization	7.17	6.29	9.20	5.08	6.93
		RL2 (↓)				
1	Ours	2.310	2.291	2.681	1.641	2.231
2	Ours w/o Task-Specific Embedding	2.715	3.153	2.826	1.388	2.520
3	Ours w/o Replay Prioritization	1.914	2.950	3.461	1.900	2.556

S4. Generalization Beyond AQA

BriMA is designed for decision-critical regression under non-stationary modality imbalance, where both task distributions and modality availability may evolve over time. While AQA serves as our primary benchmark, we further examine generalization to a distinct multi-modal regression task: sentiment intensity prediction.

Task and Setting. The CMU multi-modal Opinion Sentiment and Emotion Intensity (MOSI) dataset [55] is a foundational, opinion-level annotated corpus for multi-modal sentiment analysis, consisting of 2,199 video clips from 93 YouTube movie reviews. It is a multi-modal sentiment regression benchmark involving visual, acoustic, and textual modalities, where the goal is to predict continuous sentiment scores. Unlike AQA, which evaluates performance quality in human actions, MOSI focuses on affective understanding in conversational settings. This provides a semantically different regression task while retaining multi-modal structure and score-sensitive evaluation for generalization evaluation. To simulate non-stationary modality imbalance, we follow the same protocol as in the above AQA setting and randomly drop modalities at rates $\beta \in \{10\%, 25\%, 50\%\}$ during CL phases. We compare BriMA with joint training (JT-EMOE [10]), standard CL baselines, and recent CAQA methods under identical settings.

Results. As shown in Tab. S4, BriMA consistently achieves the best performance across SRCC, MSE, and RL2 under all missing rates. Notably, BriMA even outperforms joint training (JT-EMOE) when modalities are partially missing, indicating its ability to maintain score fidelity under incomplete observations. Compared to CL baselines, BriMA yields higher correlation and lower error, demonstrating robustness against both catastrophic forgetting and modality absence. These results suggest that BriMA generalizes beyond AQA to other decision-critical multi-modal regression tasks with non-stationary modality availability.

Table S4. Results on the MOSI dataset.

JT-EMOE [10]	($\beta = 0\%$)	SRCC	0.757	MSE	1.108	RL2	3.078

Method	$\beta = 10\%$			$\beta = 25\%$			$\beta = 50\%$		
	SRCC	MSE	RL2	SRCC	MSE	RL2	SRCC	MSE	RL2
JT-EMOE [10]	0.699	1.407	3.908	0.685	1.502	4.160	0.674	1.523	4.231
ST-EMOE [10]	0.556	2.423	6.732	0.534	2.626	7.294	0.526	2.784	7.732
EWC [18]	0.536	2.581	7.170	0.529	2.683	7.454	0.522	2.618	7.273
LwF [23]	0.587	1.729	4.803	0.502	1.943	5.347	0.177	3.134	8.706
MER [34]	0.660	1.714	4.760	0.598	1.992	5.535	0.573	2.151	5.975
DER++ [4]	0.670	2.074	5.762	0.616	2.244	6.634	0.622	2.143	5.952
NC-FSCIL [54]	0.723	1.603	4.452	0.692	1.966	5.460	0.674	1.848	5.133
SLCA [60]	0.546	2.551	7.085	0.526	2.566	7.127	0.522	2.795	7.764
FS-Aug [22]	0.570	2.380	6.612	0.536	2.836	7.878	0.530	2.853	7.926
MAGR [66]	0.704	1.570	4.360	0.669	1.769	4.914	0.650	1.899	5.275
ASAL [67]	0.696	1.646	4.574	0.673	1.663	4.619	0.612	1.929	5.358
BriMA (Ours)	**0.734**	**1.552**	**4.314**	**0.700**	**1.787**	**4.964**	**0.683**	**1.888**	**5.245**

S5. Additional Discussions

Beyond the quantitative results, we further analyze several structural aspects of BriMA and clarify its design principles under dynamic multi-modal learning.

Dynamic Multi-modal Learning vs. Static Alignment. Prior modality-invariant methods (e.g., MINIMA [33], X-Fi [5]) primarily focus on static cross-modal alignment under fixed data distributions, aiming to learn modality-agnostic representations. In contrast, BriMA targets dynamic and continual multi-modal learning with non-stationary modality availability and evolving task distributions. This setting introduces additional challenges, as preserving score-sensitive geometry over time becomes critical under distribution shift and modality imbalance. Accordingly, the bridging space is designed for continual adaptation rather than enforcing global modality invariance.

Assumption on Modality Availability. BriMA assumes modality availability is observable (e.g., via sensor status or data integrity checks), without requiring oracle knowledge of modality usefulness. Detecting unreliable modalities is a related but orthogonal problem. The residual reconstruction strategy predicts minimal corrections instead of full feature synthesis, operating within a locally smooth region of the loss landscape (see Fig. 6(a)). Imperfect modality signals therefore lead to bounded perturbations rather than amplified deviations, reducing sensitivity to routing noise.

Robustness to Weak Informative Modalities. Not all modalities contribute equally to downstream regression tasks. BriMA adopts a conservative reconstruction principle: residual corrections are conditioned on exemplar priors rather than generative synthesis of complete features. When a modality is weakly informative or loosely correlated with the target, the learned residual naturally approaches zero. Combined with modality-aware replay guided by score-drift signals, this mitigates hallucination risks and stabilizes learning under imperfect modality relevance.

Scalability with Respect to Modality Combinations. BriMA employs pattern-level conditioning embeddings to provide stable context under non-stationary modality availability. Although the number of possible missing patterns grows exponentially in theory, practical deployments involve limited modality sets and sparse observed configurations. Unseen patterns are projected into a shared low-dimensional embedding space, and residual bridging restricts corrections to minimal adjustments, avoiding brittle or random routing while maintaining scalability.

Temporal Context and Task-Level Adaptation. BriMA is designed for task-level continual adaptation rather than explicit sequence-aware routing. While memory replay captures long-term distribution shifts, the framework does not explicitly model fine-grained temporal trajectories or stage-dependent modality availability. Incorporating temporal or stage-aware conditioning represents a complementary direction for future exploration.