A Details of the AQA Network

Fig. S1 illustrates the network architecture of our AQA framework, employing the `I3D+MLP` paradigm. Following the previous work [16], we introduce the reparameterization technique [4] to ensure robust score regression.

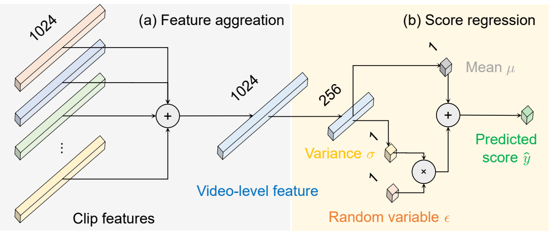

Fig. S1: The network architecture for score regression: (a) depicts feature aggregation and (b) delineates score regression.

Feature Aggregation. To address computational challenges in Action Quality Assessment (AQA), methods like [1, 7, 8, 13, 16] often opt to divide videos into clips. This process involves uniformly dividing the entire video sequence $\mathbf{x}$ into 10 clips $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{10}$. Each of these clips is then fed into the I3D backbone to extract clip features. The division helps manage the computational intensity associated with processing large video datasets, enabling more efficient computation and improved memory utilization.

These clip features are aggregated using the widely used average pooling method to obtain the whole video-level representation $\boldsymbol{h}$. Thus, the aforementioned process can be represented as:

$$\boldsymbol{h} = \texttt{AvgPool}(\texttt{i3d}(\mathbf{c}_i), \texttt{i3d}(\mathbf{c}_2), \cdots, \texttt{i3d}(\mathbf{c}_{10})), \tag{S1}$$

where $\texttt{AvgPool}(\cdot)$ denotes the average pooling function. While the simple process is effective, it has limitations, such as its coarse-grained temporal approach, which may not capture fine-grained action quality [3, 16]. It is noted that our main focus is on the integration of CL and AQA, and addressing these limitations is out of our work.

Score Regression. The video-level representation is then utilized in MLPs for the final score regression. The detailed illustration (see Fig. S1(b)) provides clarity on the sequential steps involved in the score regression process.

We employ a probabilistic layer to transform the video-level feature $\boldsymbol{h}$ into a random score variable y. The encoded score variable y follows a Gaussian distribution defined by:

$$p(y; \boldsymbol{h}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{h})}} \exp\left(-\frac{(y - \mu(\boldsymbol{h}))^2}{2\sigma^2(\boldsymbol{h})}\right), \tag{S2}$$

where μ and σ^2 are the mean and variance parameters with respect to the feature representation. These parameters quantify the quality and uncertainty of the action score, respectively. To sample from the distribution, we apply the reparameterization trick [4], which involves sampling from another random variable ϵ following the standard normal distribution $\mathcal{N}(0,1)$. In this way, the predicted score $\hat{y}$ can be calculated as:

$$\hat{y} = \mu(\boldsymbol{h}) + \epsilon \cdot \sigma(\boldsymbol{h}), \tag{S3}$$

where $\sigma(\cdot)$ represents the standard deviation. This ensures that the score distribution sampling process is differentiable, allowing feasible training of our score regression network.

B Additional Experimental Details

This section provides supplementary information on the datasets and implementation details used in our experiments.

B.1 Deatils of Datasets

MTL-AQA [8] serves as a robust resource for AQA research, offering a comprehensive collection of 16 distinct diving events. With a total of 1412 samples, it encompasses a wide range of scenarios, featuring both male and female athletes participating in single and double diving competitions across 3 m springboard and 10 m platform categories. Notably, MTL-AQA provides detailed annotations for action categories and commentary alongside AQA scores, enhancing the dataset's utility for comprehensive analysis. A total of 1059 samples were allocated for training purposes, while 353 samples were reserved for testing.

FineDiving [11] is a recently introduced large-scale fine-grained diving dataset comprising 3000 diving samples extracted from various prestigious events including the Olympics, World Cup, World Championships, and European Aquatics Championships. This dataset covers 52 action types, 29 sub-action types, and 23 difficulty degree types, providing rich annotations for detailed analysis. Additionally, FineDiving includes fine-grained annotations such as action type, sub-action type, coarse-grained and fine-grained temporal boundaries, and action scores in addition to AQA scores. A total of 75% of the samples were allocated for training purposes, while the remaining 25% were reserved for testing.

UNLV-Dive [9] comprises 300 training videos and 70 testing videos, selected from the (semi-) final of the 10-meter platform diving event in the 2012 London Olympics. The dataset's final scores range from 21.6 to 102.6, with execution scores falling within the [0, 30] range.

JDM-MSA [15] comprises 14 types of actions categorized into three groups: time-related actions, position-related actions, and uncertain actions. Time-related actions are evaluated based on completion duration, and position-related actions on body part movement magnitude and position, while uncertain actions require subjective judgments of difficulty. For this study, we focus on six challenging actions, prioritizing those requiring subjective assessment. Data collection efficiency was enhanced using an iPad, an iPhone, and two USB cameras. To ensure consistency, all samples were normalized to the same resolution and frame count.

B.2 Implementation Details

In the detailed architecture presented in Fig. S1, the video-level representation undergoes an initial encoding process, resulting in a 1024-dimensional vector. This vector then traverses a series of Multi-Layer Perceptrons (MLPs). The first MLP layer, with a dimensionality of 256, serves as the initial transformation. Following this, two additional MLP layers process the vector further, ultimately mapping it to mean and variance scores. These mean and variance parameters play a crucial role in shaping the final prediction. The MP module in MAGR follows a similar structure with a 2-layer MLP. The input dimension is set to 1024, the hidden size is 256, and the output size remains at 256. This design choice aims to capture the intricate relationships within the data manifold, enhancing the ability to align features with evolving data distributions across sessions.

C Additional Experiments

This section delves into additional computational analysis, online performance, the inclusion of difficulty degree, additional ablation study, the impact of memory size, and offers visualizations to further verify the effectiveness of our method.

C.1 Model Size and Computational Time Comparison

We conducted measurements on both model size (backbone + regressor) and offline training time (with the same hyper-parameter setting using 2 Nvidia RTX 3090 GPUs in distributed parallel computing on the same machine). The results are reported in Tab. S1. It can be seen that our method employs comparable model size and computational time while achieving much better performance than the most recent strong baselines. Specifically, the inclusion of a manifold projector slightly increases our model size compared to these baselines, while our training time remains competitive.

Table S1: Computational performance on the MTL-AQA dataset.

Method	Param. (M)	Training Time (h)	Offline Performance		
			ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)
SLCA [14]	13.62	2.27	0.7223	0.1852	0.1665
NC-FSCIL [12]	12.62	2.33	0.8426	0.1146	0.0718
Feature MER	12.62	2.22	0.7283	0.2255	0.0535
MAGR (Ours)	12.63	2.23	0.8979	0.0223	0.1914

C.2 Real-Time Performance

In our online continual learning setting, we evaluate the performance of models in real-time scenarios where data arrives sequentially and models need to learn and adapt continuously without revisiting previous data. Each method is trained and updated in an online manner, processing one data point or a small batch at a time (training 1 epoch for all models), simulating real-world conditions where retraining on the entire dataset is infeasible. The results in Tab. S2 highlight the superior online CL performance of MAGR, which achieves the highest average correlation (ρ_{avg}) across all datasets: MTL-AQA (0.5196), FineDiving (0.4641), UNLV-Dive (0.4202), and JDM-MSA (0.2029). These results demonstrate that MAGR's graph regularization effectively preserves the feature space structure, making it a strong contender for real-time action assessment.

Table S2: Online continual learning (ρ_{avg} is the main metric).

Method	MTL-AQA			FineDiving			UNLV-Dive			JDM-MSA		
	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)
SLCA [14]	0.4880	0.0430	-0.0282	0.3935	0.3360	0.2346	0.3119	0.1641	-0.3082	0.1726	0.0589	0.0382
NC-FSCIL [12]	0.4971	0.0291	-0.0463	0.3810	0.0079	**0.2518**	0.3136	**0.1282**	-0.4892	0.1540	**0.0355**	0.0378
Feature MER	0.3571	0.1444	**-0.0213**	0.1935	0.0998	0.1559	0.1308	0.2126	-0.4571	0.1699	0.0356	0.0382
MAGR (Ours)	**0.5196**	**0.0269**	-0.0337	**0.4641**	**0.0062**	0.2020	**0.4202**	0.1947	**-0.0499**	**0.2029**	0.0356	**0.0449**

C.3 Comparison with the Inclusion of Difficulty Degree

Previous research [1, 13, 16] has indicated that integrating difficulty degree labels can significantly enhance the performance of AQA models, particularly on the MTL-AQA dataset. In alignment with this observation, we conducted experiments incorporating difficulty degree labels in our evaluation, as depicted in Tab. S3. The setting is the same as the previous work [13]. The results reconfirm the beneficial impact of leveraging difficulty degree information on AQA performance. Notably, MAGR maintains its superiority over recent strong baselines, reaffirming its effectiveness in addressing AQA challenges.

Table S3: Results on the MTL-AQA dataset with the difficulty degree.

Method	Publisher	Memory	MTL-AQA		
			ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)
Joint Training	-	None	0.9587	-	-
Sequential FT	-	None	0.8684	0.1418	0.2282
EWC [5]	PNAS'17	None	0.8625	0.1267	0.1776
LwF [6]	TPAMI'17	None	0.7852	0.1501	0.0912
DER++ [2]	NeurIPS'20	Raw Data	0.9037	0.1230	0.3122
TOPIC [10]	CVPR'20	Raw Data	0.8782	0.1394	0.2304
SLCA [14]	ICCV'23	Feature	0.6885	0.2029	0.0958
NC-FSCIL [12]	ICLR'23	Feature	0.9034	0.0878	0.1456
MAGR (Ours)	-	Feature	**0.9237**	**0.0615**	**0.1944**

C.4 Ablation Study

The ablation results in Tabs. S4 and S5 on both the UNLV-Dive and JDM-MSA datasets demonstrate the vital role each component our MAGR model.

Ablation Study on UNLV-Dive. From the results in Tab. S4, it is evident that each component of our MAGR model plays a crucial role in its performance. Omitting the MP module leads to a 21% decline in ρ_{avg} and a marked 69% increase in ρ_{aft}, emphasizing its significance in managing feature deviation. The graph regularizers, both local (II-GR) and global (J-GR), when removed individually or together, induce notable reductions in ρ_{avg} and substantial increments in ρ_{aft}, emphasizing their essential role in regularizing the feature space. Lastly, without the OUS strategy, we observe a decrease in all performance metrics, underlining its importance in model robustness across sessions. The collective results underscore the intertwined significance of all the components in achieving the peak performance of MAGR on the UNLV-Dive dataset.

Table S4: Ablation results on the UNLV-Dive dataset.

Setting	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)
MAGR (Ours)	0.7668	0.0827	0.1227
w/o MP	$0.6026^{\downarrow 21\%}$	$0.1396^{\uparrow 69\%}$	$0.1075^{\downarrow 12\%}$
w/o II-GR	$0.7189^{\downarrow 6\%}$	$0.1726^{\uparrow 109\%}$	$0.1226^{\downarrow 0\%}$
w/o J-GR	$0.6549^{\downarrow 15\%}$	$0.1267^{\uparrow 53\%}$	$0.1466^{\uparrow 20\%}$
w/o IIJ-GR	$0.6261^{\downarrow 18\%}$	$0.2102^{\uparrow 154\%}$	$0.1442^{\uparrow 18\%}$
w/o OUS	$0.7356^{\downarrow 4\%}$	$0.0599^{\downarrow 28\%}$	$0.0867^{\downarrow 29\%}$

Ablation Study on JDM-MSA. The ablation study on the JDM-MSA dataset provides insight into the importance of the components in our MAGR model. The results demonstrate that each component significantly contributes to the model's performance. Omitting the MP module resulted in a 20% decrease in ρ_{avg} and a 10% decrease in ρ_{aft}, highlighting its role in addressing feature deviation. Similarly, removing the graph regularizers, including II-GR, J-GR, and IIJ-GR, led to notable reductions in ρ_{avg}, emphasizing their essential role in regularizing the feature space. The proposed OUS strategy proved to be crucial for maintaining model performance, with its removal resulting in performance declines across all metrics. These findings emphasize the critical nature of each component for achieving optimal AQA performance on the JDM-MSA dataset.

Table S5: Ablation results on the JDM-MSA dataset.

Setting	ρ_{avg} (↑)	ρ_{aft} (↓)	ρ_{fwt} (↑)
MAGR (Ours)	0.7166	0.1069	0.4957
w/o MP	$0.5725^{\downarrow 20\%}$	$0.1185^{\downarrow 10\%}$	$0.4942^{\downarrow 0\%}$
w/o II-GR	$0.6755^{\downarrow 6\%}$	$0.1962^{\uparrow 83\%}$	$0.4956^{\downarrow 0\%}$
w/o J-GR	$0.6066^{\downarrow 15\%}$	$0.0933^{\downarrow 13\%}$	$0.3953^{\downarrow 20\%}$
w/o IIJ-GR	$0.5792^{\downarrow 19\%}$	$0.1055^{\downarrow 10\%}$	$0.4085^{\downarrow 18\%}$
w/o OUS	$0.6880^{\downarrow 4\%}$	$0.1280^{\downarrow 16\%}$	$0.4945^{\downarrow 0\%}$

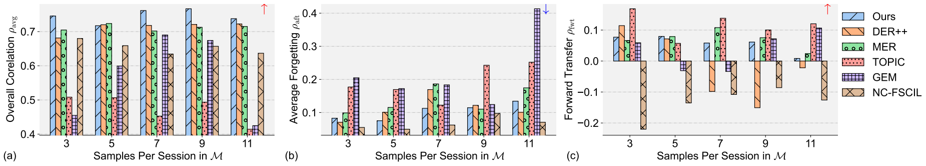

Fig. S2: Memory size comparisons with replay-based methods on UNLV-Dive.

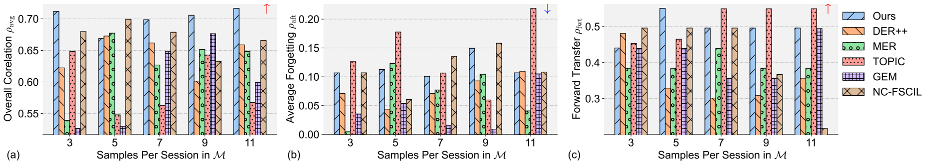

Fig. S3: Memory size comparisons with replay-based methods on JDM-MSA.

C.5 Impact on the Memory Size

The performance comparison of various replay methods on the UNLV-Dive and JDM-MSA datasets for different memory sizes is presented in Figs. S2 and S3, respectively. On the UNLV-Dive dataset, our method consistently outperforms other approaches, achieving the highest overall correlation (ρ_{avg}). When memory size is small, all methods experience performance degradation. However, as the memory size increases, our method exhibits superior resilience and maintains strong performance, while the performance of other methods tends to saturate or even decline. On the JDM-MSA dataset, similar trends are observed, with our method achieving the highest ρ_{avg} and demonstrating greater stability as memory size increases. These results highlight the effectiveness and robustness of our approach across varying memory sizes and datasets.

C.6 Visualization of Mitigating Catastrophic Forgetting

The visualization presented in Fig. S4 offers valuable insights into the effectiveness of MAGR in mitigating catastrophic forgetting on the UNLV-Dive dataset. The TSNE plots (Figs. S4(a) to S4(f)) showcase the distribution of feature representations learned by MAGR (top) and feature MER (bottom) across different sessions. It is evident that MAGR's MP and IIJ-GR contribute to maintaining the consistency and continuity of feature distributions over sequential sessions, thereby alleviating the adverse effects of catastrophic forgetting. Notably, the correlation plots (Figs. S4(g) and S4(h)) further emphasize MAGR's superior performance in preserving the correlation between predicted quality scores and ground truth labels, even amidst evolving feature distributions. This highlights MAGR's robustness and efficacy in addressing the challenges posed by non-stationary data distributions and underscores its potential to enhance CL and AQA research.

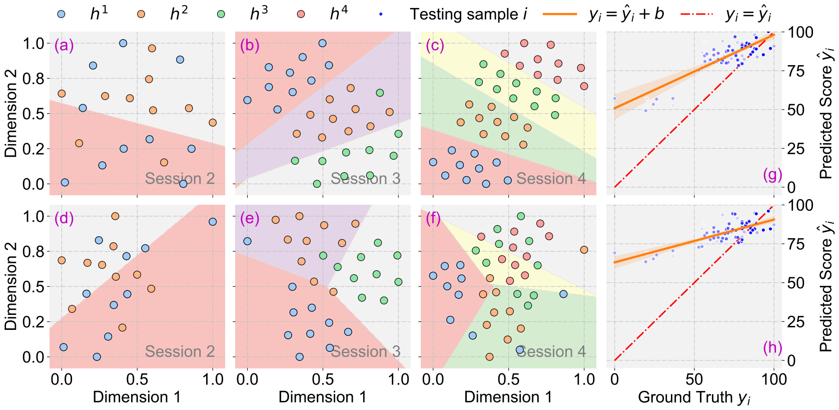

Fig. S4: Visualizations of feature distribution (a-f) and score correlation (g-h) on the UNLV-Dive dataset: MAGR (top) and Feature MER (bottom).

C.7 Visualization of Performance Changes

In Fig. S5, the performance changes (ρ_{avg}) of several recent strong baselines across different sessions on the MTL-AQA dataset are illustrated. Initially, all methods exhibit comparable performance. However, as sessions progress, our method consistently maintains a stable level, while other methods experience substantial fluctuations. This underscores the effectiveness of our approach in balancing learning plasticity and memory stability over sequential sessions.

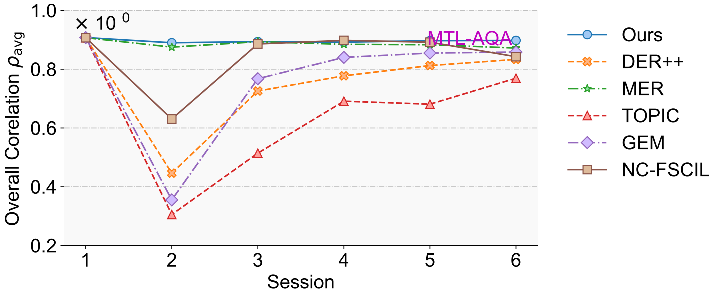

Fig. S5: Performance comparison across different sessions.

D Discussions

Larger Dataset Validation. We evaluated our method, MAGR, across four AQA datasets of varying scales, consistently achieving leading performance. Given the typically scarce training samples for AQA tasks, FineDiving is one of the largest available datasets. We plan to update our results with larger AQA datasets as they become available.

Overfitting with Smaller Datasets. We acknowledge the potential risk of overfitting with smaller datasets or fewer feature points, as noted. To address this, we conducted experiments with reduced training samples, as shown in Fig. 8(a) of our main paper. When reducing the number of training samples per session, the performance of several recent strong baselines, especially those with non-graph feature replay methods like SLCA [14] and NC-FSCIL [12], decreases significantly. In contrast, our method maintains relatively stable performance, resulting in a significant performance lead. This stability is attributed to the use of graph construction, which captures and retains relationships between feature points, thereby reducing the risk of overfitting.

Handling Difficult Cases of Actions and Qualities. We also analyzed the specific errors related to actions and qualities in our main paper. Fig. S5 evaluates the performance of specific actions learned in respective sessions. Our performance remains consistently high across different actions, whereas other strong baselines experience performance degradation in specific actions, particularly in session 2. Additionally, Fig. 9(g) and 9(h) evaluate the performance of specific qualities of each action. Compared to Feature MER in Fig. 9(h), our method, shown in Fig. 9(g), aligns much better with the ground truth, indicating fewer specific errors. While our method shows slight deviations within the low-score area due to limited training samples, the high-score area, which is densely sampled, is more accurate. This issue can be mitigated by collecting or re-sampling more training samples in the low-score area.

References

1. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: European Conference on Computer Vision. pp. 422–438. Springer (2022)
2. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in Neural Information Processing Systems **33**, 15920–15930 (2020)
3. Gedamu, K., Ji, Y., Yang, Y., Shao, J., Shen, H.T.: Fine-grained spatio-temporal parsing network for action quality assessment. IEEE Transactions on Image Processing (2023)
4. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691 (2019)
5. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114
6. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
7. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6331–6340 (2019)
8. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: 2019 IEEE Winter Conference on Applications of Computer Vision. pp. 1468–1476. IEEE (2019)

9. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28 (2017)
10. Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12183–12192 (2020)
11. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2949–2958 (2022)
12. Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., Tao, D.: Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In: ICLR (2023)
13. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7919–7928 (2021)
14. Zhang, G., Wang, L., Kang, G., Chen, L., Wei, Y.: Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. arXiv preprint arXiv:2303.05118 (2023)
15. Zhou, K., Cai, R., Ma, Y., Tan, Q., Wang, X., Li, J., Shum, H.P., Li, F.W., Jin, S., Liang, X.: 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)
16. Zhou, K., Ma, Y., Shum, H.P., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE Transactions on Circuits and Systems for Video Technology (2023)