Action Quality Assessment with Temporal Parsing Transformer: Supplementary Material

Yang Bai[1,2*†], Desen Zhou[2*], Songyang Zhang[3], Jian Wang[2], Errui Ding[2], Yu Guan[4], Yang Long[1], and Jingdong Wang[2‡]

[1] Department of Computer Science, Durham University
[2] Department of Computer Vision Technology (VIS), Baidu Inc.
[3] Shanghai AI Laboratory, [4] University of Warwick
{yang.bai,yang.long}@durham.ac.uk,{zhoudesen,wangjingdong}@baidu.com

In this supplementary material, we provide more results on AQA-7 dataset.

1 Datasets

MTL-AQA dataset is the largest dataset for AQA task. In MTL-AQA, 1412 fine-grained samples are collected from 16 different events with different views. The dataset mainly focuses on diving covering various categories. In this dataset, different annotations are available to support research on different tasks, including action quality assessment, action recognition, and comment generation. In addition, raw annotation of score and difficulty (DD) from the multiple judges is available. We split the dataset into 1059 training samples and 353 test data following the evaluation protocol suggested in paper [3].
AQA-7 dataset contains samples from seven different action categories, including gymnastic vaulting, big air skiing, big air snowboarding, synchronous diving - 3m springboard, synchronous diving - 10m platform, and trampoline. Following the setting in [2], we excluded the trampoline category with much longer videos than the other categories, resulting in 803 training videos and 303 testing videos.
JIGSAWS [1] is a surgical activities dataset that contains three tasks, namely Suturing (S), Needle Passing (NP), and Knot Tying (KT). To evaluate different features of videos, samples in the dataset are annotated with multiple scores, and the final score is the sum of all annotations. We applied four-fold cross-validation following [4] to align with previous work.

2 Additional Ablation Study

2.1 Ablation on different number of queries and layers

Number of queries We show the ablation study of the number of queries in Tab. 1a. We found that too many queries hurt performance. As a result, we choose query number to be 5.

* Equal contribution. † Work done when Yang Bai was a research intern at VIS, Baidu.
‡ Corresponding Author.

Table 1: More ablation studies on MTL-AQA dataset.

(a) Different number of queries.

Query number	Sp. Corr.	R-ℓ_2
3	0.9562	0.2583
5	**0.9607**	**0.2378**
7	0.9572	0.2337

(b) Different number of decoder layers.

Layer number	Sp. Corr.	R-ℓ_2
0 (baseline)	0.9498	0.2893
1	0.9563	0.2736
2	**0.9607**	0.2378
3	0.9594	**0.2303**

Number of decoder layers Our temporal parsing transformer has a multi-layer decoder structure, we show the ablation study of different decoder layers in Tab.1b. We found that 2-layer decoder achieves comparable performance compared with 3-layer decoder. We finally select 2-layer decoder for model simplicity.

2.2 Ablation study on AQA-7 dataset

In this subsection, we perform experiments to show the effectiveness of each designed components, i.e. temporal parsing transformer(TPT), ranking loss(L_{rank}) and sparsity loss($L_{sparsity}$) on AQA-7 dataset. The results are shown in Tab.2. We can observe that with only TPT component, the average performance of six categories is improved from 0.8229 Corr. to 0.8478 Corr. With our ranking loss and sparsity loss, the performance is further significantly improved from 0.8478 Corr. to 0.8715 Corr., showing the effectiveness of our temporally ordered supervision strategy.

Table 2: Ablation study of different components on AQA-7 dataset.

Sp. Corr	TPT	L_{rank}	$L_{spar.}$	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg.
Baseline	×	×	×	0.8597	0.7117	0.6625	0.6342	0.9336	0.9189	0.8229
	✓	×	×	0.8889	0.7837	0.6753	0.6722	0.9401	0.9279	0.8478
	✓	✓	×	0.8892	0.7999	**0.7367**	0.6722	0.9429	0.9440	0.8622
Ours	✓	✓	✓	**0.8969**	**0.8043**	0.7336	**0.6965**	**0.9456**	**0.9545**	**0.8715**
R-ℓ_2(×100)	TPT	L_{rank}	$L_{spar.}$	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg.
Baseline	×	×	×	0.70	2.18	4.03	4.08	0.78	2.40	2.36
	✓	×	×	0.76	1.73	4.13	3.50	0.47	1.64	2.04
	✓	✓	×	0.54	1.69	3.31	3.49	0.49	1.94	1.91
Ours	✓	✓	✓	**0.53**	**1.68**	**2.89**	**3.30**	**0.33**	**1.33**	**1.68**

2.3 Visualization results on AQA-7 dataset

We provide some visualization results of AQA-7 dataset in Fig.1 and Fig.2. In Fig.1, we visualize the frames with the highest attention responses in cross attention maps of the last decoder layer. Four representative categories are selected

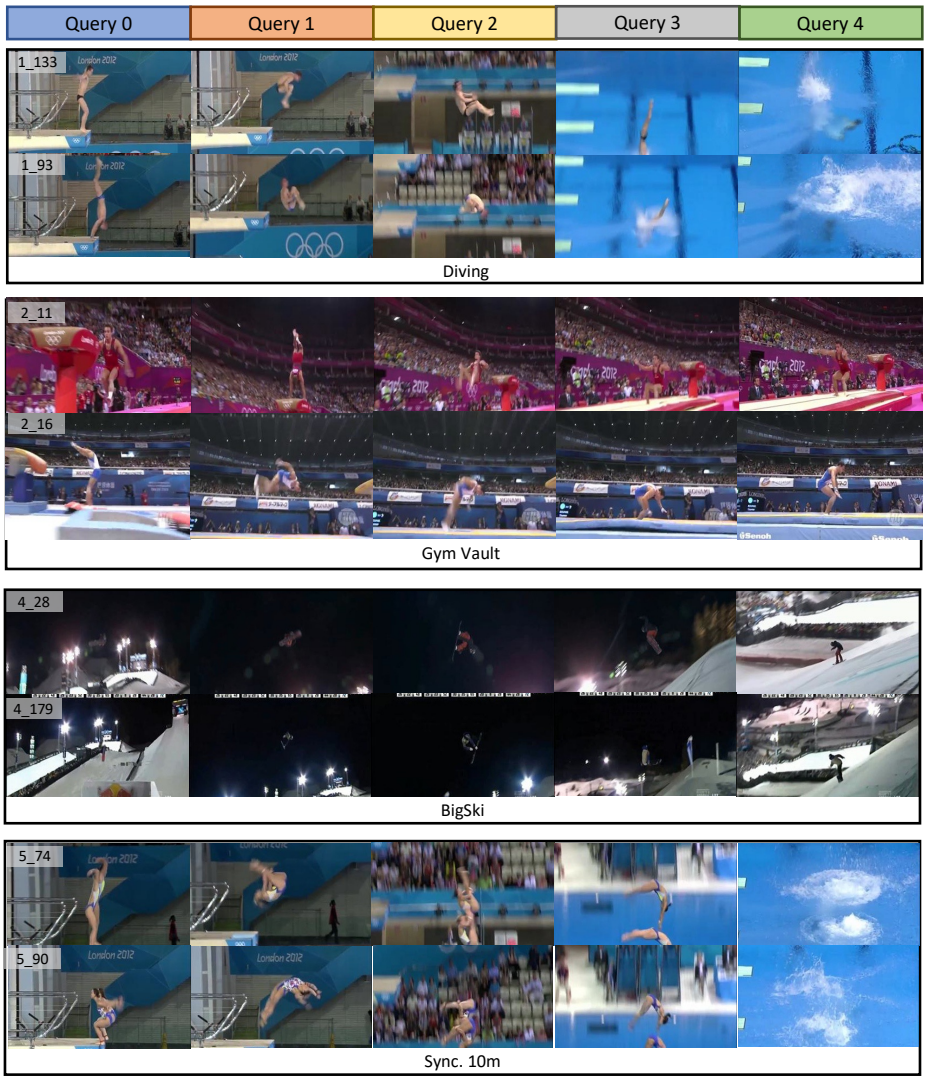

Fig. 1: Visualization of the frames with highest attention responses in decoder cross attention maps on AQA-7 dataset. Each row represents a test video from different representative categories (diving, gymnastic vault, big air snowboarding, synchronous diving - 10m platform), whose ID is shown in the left first frame. Different columns correspond to temporally ordered queries (note that different categories do not share same query embeddings). The above results show that our transformer is able to capture semantic temporal patterns with learned queries.

for showing, the other two categories are similar. Since each clip consists of

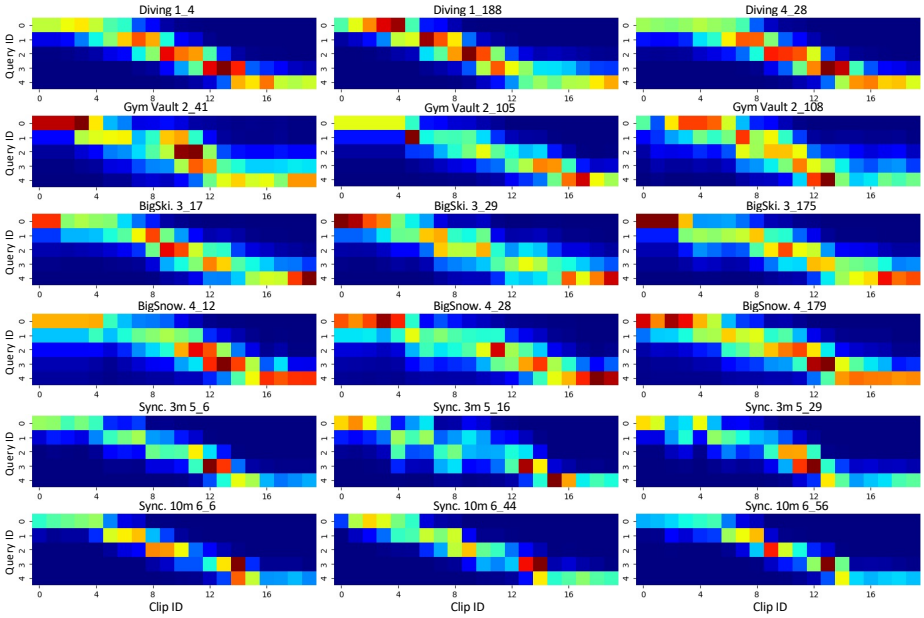

Fig. 2: Visualization of cross attention maps on video samples from the AQA-7 dataset covering all categories (diving, gymnastic vault, big air skiing, big air snowboarding, synchronous diving - 3m springboard and synchronous diving - 10m platform), where video IDs and category names are shown at the top of each attention map. In each subfigure, each row indicates one query, and each column indicates one clip. We can observe that the bright grids(with high attention responses) have a consistent temporal order due to ranking loss, and the attention maps are sparse due to our sparsity loss.

multiple frames, we select the middle frame of a clip as representative. We can observe that our transformer is capable of parsing a diving video into temporal patterns such as the take-off, the flight, and the entry with learned queries. Other categories are similar.

Fig.2, we visualize the cross attention maps on AQA-7 dataset for all categories. We can observe similar results from the MTL-AQA dataset that the attention responses have a consistent temporal order and are adaptive for different video samples, which demonstrates the effectiveness of our proposed ranking loss and sparsity loss. We also observe that the categories with more samples (diving category with 370 training samples) have more distinguishable cross attention responses of parts than categories with fewer samples (91 training samples from synchronous diving - 10m platform). We suppose that more training samples might be beneficial to learn atomic patterns.

References

1. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI workshop: M2cai. vol. 3, p. 3 (2014)
2. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1468–1476. IEEE (2019)
3. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 304–313 (2019)
4. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9839–9848 (2020)