Supplementary Material for "Likert Scoring with Grade Decoupling for Long-term Action Assessment"

Angchi Xu[1], Ling-An Zeng[2], Wei-Shi Zheng[1,3,4,*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]School of Artificial Intelligence, Sun Yat-sen University, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{xuangch, zenglan3}@mail2.sysu.edu.cn, wszheng@ieee.org

1. Additional Implementation Details

Label Normalization. An AQA dataset with N_v videos is generally annotated with non-negative real labels $\{\hat{y}^{(i)}\}_{i=1}^{N_v}$. Following [1], we introduce a normalizing constant ϵ to constraint the label interval to [0,1]:

$$y^{(i)} = \hat{y}^{(i)}/\epsilon, \tag{1}$$

where $\{y^{(i)}\}_{i=1}^{N_v} \in [0, 1]$ are used as labels for our training. The value of ϵ is related to the maximum score of the training set. Specifically, we set it as 25/45/40 for RG / Fis-V(TES) / Fis-V(PCS), respectively.

Feature Extraction. Since the AQA is a fine-grained action understanding task, which needs to capture the rapid informative events (spanning only a few frames) and examine the entire duration of the video [2], we adopt a dense-sampling strategy. Specifically, given a video with N frames and the number of frames per segment N_s (32 in our experiments), we select $N_{use} = \left\lfloor \frac{N}{N_s} \right\rfloor \times N_s$ frames to produce $\left\lfloor \frac{N}{N_s} \right\rfloor$ segments. We take consecutive frames in the middle of the video, namely, discard the first $\left\lfloor \frac{N - N_{use}}{2} \right\rfloor$ frames and the last $\left\lceil \frac{N - N_{use}}{2} \right\rceil$ frames.

2. More Visualizations

More Visualizations of Cross-attention Weights. Here we provide more visualizations of the cross-attention weights of different grade prototypes in GAD. As shown in Figures 1 to 3, the prototypes are able to capture video segments related to different grades.

More Visualizations of Response Intensities. In the paper, we have taken RG(Ball) as an example to illustrate how the response intensities change with the label scores. Here we provide the visualizations of other five classes in Figure 4. The same changing law as RG(Ball) can be observed in the figure, which demonstrates the robustness of our model.

References

[1] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2526–2534, 2020. 1

[2] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 1

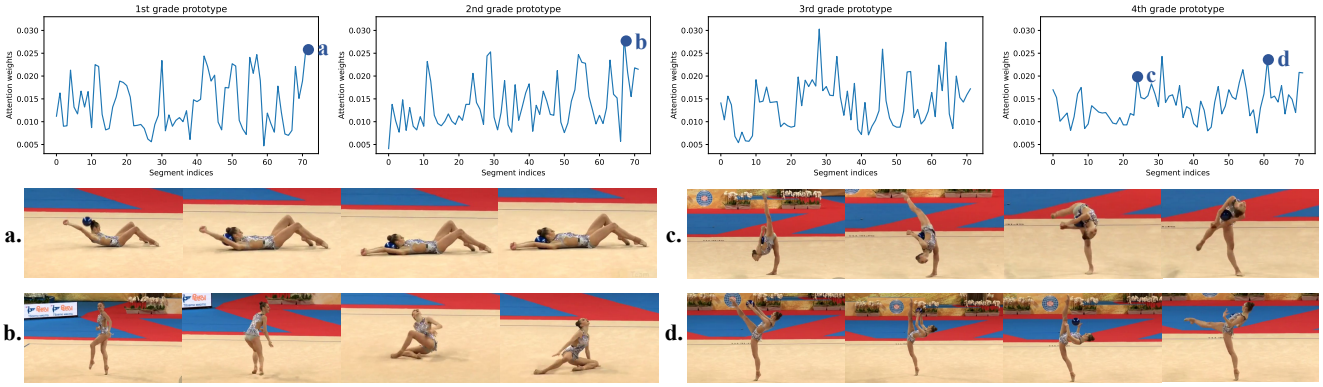

Figure 1. Visualization of cross-attention weights of each grade prototype in GAD. The sample is the *Ball_030* video in RG. The first row shows four weight curves of four prototypes on video segments. The next two rows are four video segments corresponding to four markers on the curves, *i.e.*, a, b, c and d.

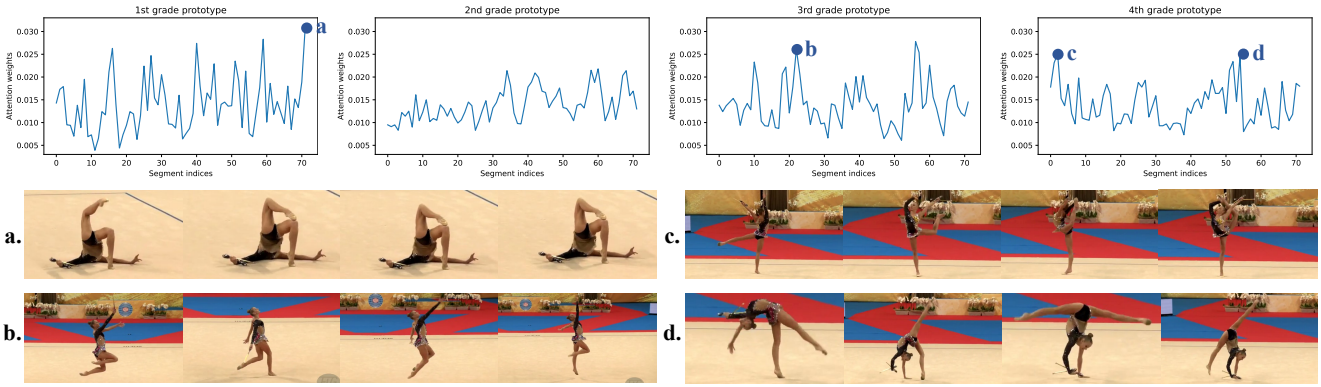

Figure 2. Visualization of cross-attention weights of each grade prototype in GAD. The sample is the *Clubs_016* video in RG. The first row shows four weight curves of four prototypes on video segments. The next two rows are four video segments corresponding to four markers on the curves, *i.e.*, a, b, c and d.

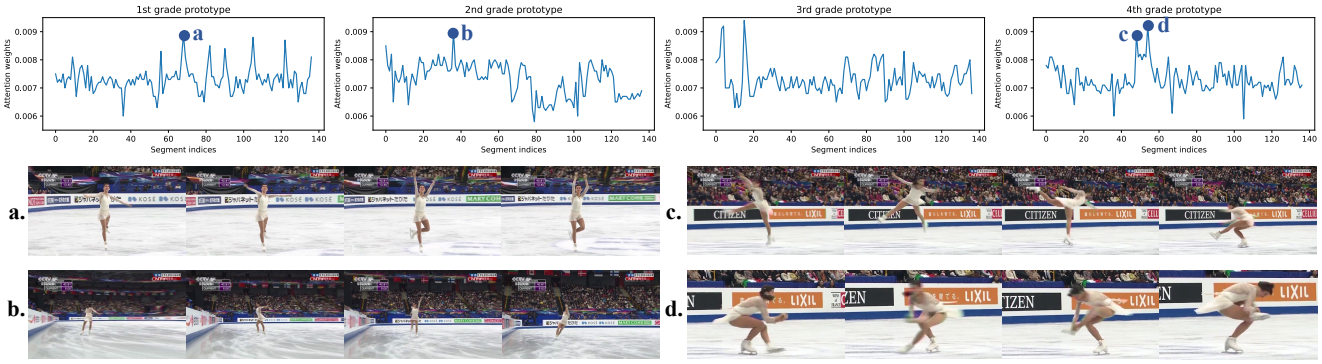

Figure 3. Visualization of cross-attention weights of each grade prototype in GAD. The sample is the *#230* video in Fis-V and the class is TES. The first row shows four weight curves of four prototypes on video segments. The next two rows are four video segments corresponding to four markers on the curves, *i.e.*, a, b, c and d.

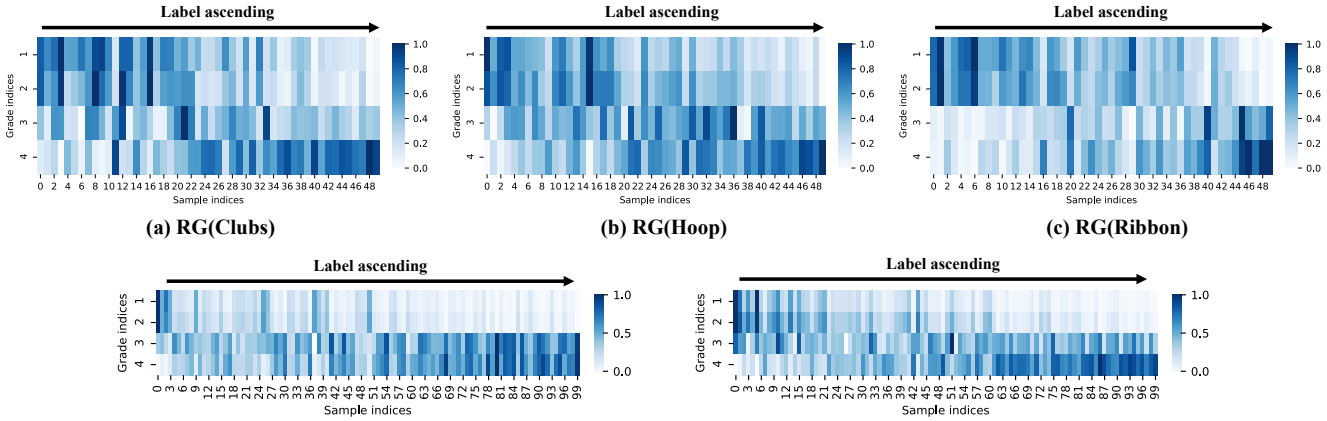

Figure 4. Visualization of response intensities at each grade of all video samples on the test sets of all classes except RG(Ball) (see the paper). In each sub-figure, each column represents a sample, and all samples are sorted in ascending order of label scores. To better observe the relative changes of intensities with the sample scores, we normalize each row, *i.e.*, all intensities of the same grade, by min-max scaling. Best viewed in color.