# FineGrade: A Rule-Consistent Scoring Framework for Fine-Grained Action Quality Assessment

Supplementary Material

## 1. More Annotation Details

We retain only videos covering the four annotated apparatus: Vault (VT), Floor Exercise (FX), Balance Beam (BB), and Uneven Bars (UB) and discard warm-ups, training, or clips without official annotations. For each valid competition, we fetch the official result book and manually verify day or stage correspondence (*e.g.*, qualification vs. final) to match scores with videos. After filtering, 173 competition videos remain for alignment. FineGym-AQA extracts all labels (Athlete, D/E/ND/T, Sub-code/DV) directly from official result books and FIG CoP [1], ensuring traceability. Figure 1 illustrates the detailed annotation process.

## 2. Feature Extraction

We follow the dense temporal sampling scheme of Xu et al. [2]. Given a video with $N$ frames, we set the segment length to $N_s = 32$ frames. We first compute the number of complete segments $K = \lfloor N/N_s \rfloor$ and keep $N_{\text{use}} = KN_s$ frames. To focus on the central part of the sequence, we discard $\lfloor (N - N_{\text{use}})/2 \rfloor$ frames from the beginning and drop the remaining extra frames at the end. The remaining $N_{\text{use}}$ frames are then split into $K$ non-overlapping segments of length $N_s$, and each segment is fed into the backbone network to obtain a length-$K$ feature sequence.

## 3. More Implementation Details

**FineGym-AQA.** Inputs are variable-length VST features and event IDs $e \in \{1, 2, 3, 4\}$ (VT, FX, BB, UB). The temporal encoder is a 2-layer Transformer ($d_{\text{model}} = 384$, 4 heads). A boundary head decodes segments using global parameters $(K, \tau, \ell_{\min}) = (6, 0.4, 1)$. Difficulty is aggregated via event-conditioned attention, while execution uses global context. The total score $T = D + E$ is supervised with weights $w_{\text{Tsum}} = 1.0$, $w_D = 0.5$, $w_E = 0.1$ (ranking loss disabled). We train for 60 epochs (batch 32) using AdamW (decay $3 \times 10^{-4}$) and cosine annealing with learning rate $\in [1 \times 10^{-5}, 6 \times 10^{-5}]$.

**Rhythmic Gymnastics (RG).** Using VST features and event IDs (Ball, Clubs, Hoop, Ribbon), we use a transformer with $d_{\text{model}} = 320$, 8 attention heads, and 2 encoder layers. We adopt apparatus-specific pseudo-segment and decoding hyperparameters: `uniform_segments`$=(4, 12, 6, 12)$, and $(K, \tau, \ell_{\min}) = (8, 0.40, 3)$, $(12, 0.45, 1)$, $(8, 0.40, 2)$, and $(6, 0.55, 2)$ for Ball, Clubs, Hoop, and Ribbon, respectively. $w_{\text{Tsum}} = 0.6$, $w_D = 0.2$, $w_E = 0.9$, and a light ranking regularization $w_{\text{rankT}} = 0.05$. We apply per-event $z$-normalization and mixed-precision training. Optimization uses AdamW with learning rate $2 \times 10^{-5}$, no weight decay, cosine annealing to $5 \times 10^{-6}$, a 3-epoch warm-up, gradient clipping at 1.5, batch size 24, and 80 training epochs.

## 4. More Ablation Studies

**Rule-consistent totalization vs. direct-$T$ variants.** We study whether enforcing the judging rule helps total-score prediction. FineGrade uses rule-consistent totalization, $\hat{T} = \hat{D} + \hat{E}$. We compare it with two direct-$T$ baselines that share the same encoder, boundary head, segment decoding, and training schedule, and differ only in the final predictor. *Dagg-T* predicts $\hat{T}$ from the segment-aggregated representation used by the difficulty branch, while *Direct-T* predicts $\hat{T}$ directly from the global context $\mathbf{G}$ through an event-conditioned head $\phi_{\text{evt}}^{(T)}(\mathbf{G})$. We report per-apparatus SRCC and validation MSE.

Table 1. FineGym-AQA results for different totalization strategies. We report per-apparatus SRCC, Fisher-averaged SRCC, and MSE. Best in bold.

| | Per-apparatus SRCC↑ | | | | Average | |
|---|---|---|---|---|---|---|
| **Setting** | VT | FX | BB | UB | *SRCC*↑ | *MSE*↓ |
| $T$ **Contribution** | | | | | | |
| Dagg-T | 0.578 | 0.602 | 0.778 | 0.719 | 0.678 | 1.279 |
| Direct-T | 0.645 | 0.643 | 0.770 | 0.746 | 0.706 | 0.929 |
| Rule-consistent | **0.746** | **0.716** | **0.794** | **0.750** | **0.753** | **0.752** |

Table 1 shows that rule-consistent totalization performs best on all four apparatus, and also achieves the highest average SRCC and the lowest MSE. Among the two direct baselines, *Direct-T* is better than *Dagg-T*, suggesting that when predicting $T$ directly, global context is more effective than simply aggregating segment features. Still, both direct variants are clearly weaker than $\hat{T} = \hat{D} + \hat{E}$, showing that explicit score decomposition is important for accurate assessment.

## 5. More Visualizations

**Qualitative Examples of Temporal Alignment.** Figure 2 verifies the model's interpretability: predicted actionness aligns well with ground-truth intervals, and attention focuses on critical skills (*e.g.*, landings) rather than background.

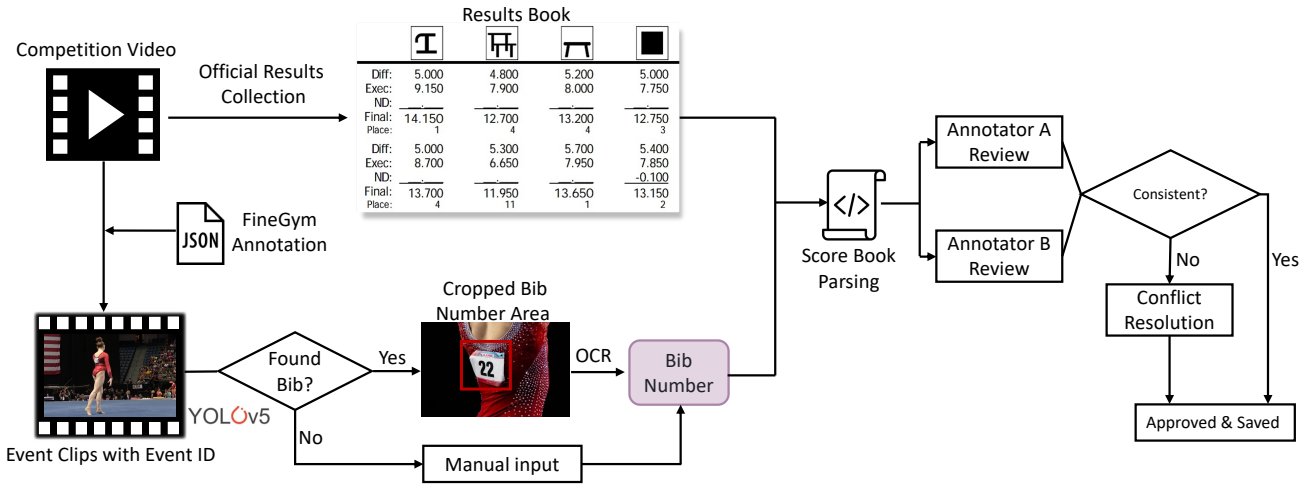



Figure 1. Event clips from competition videos are linked to parsed official scores using athlete bib numbers, which are extracted via YOLOv5 and OCR (or manual input). The matched video-score data then undergoes a dual-annotator review and conflict resolution process to ensure accuracy before being saved.

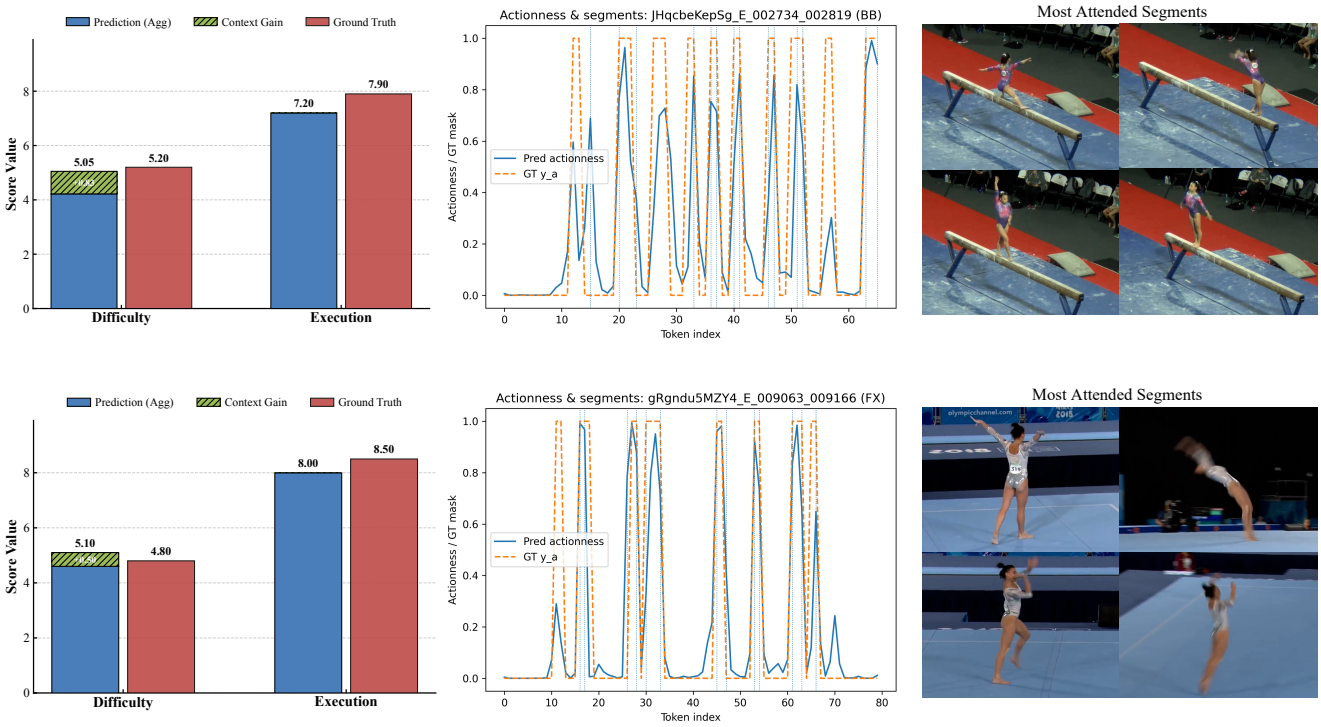


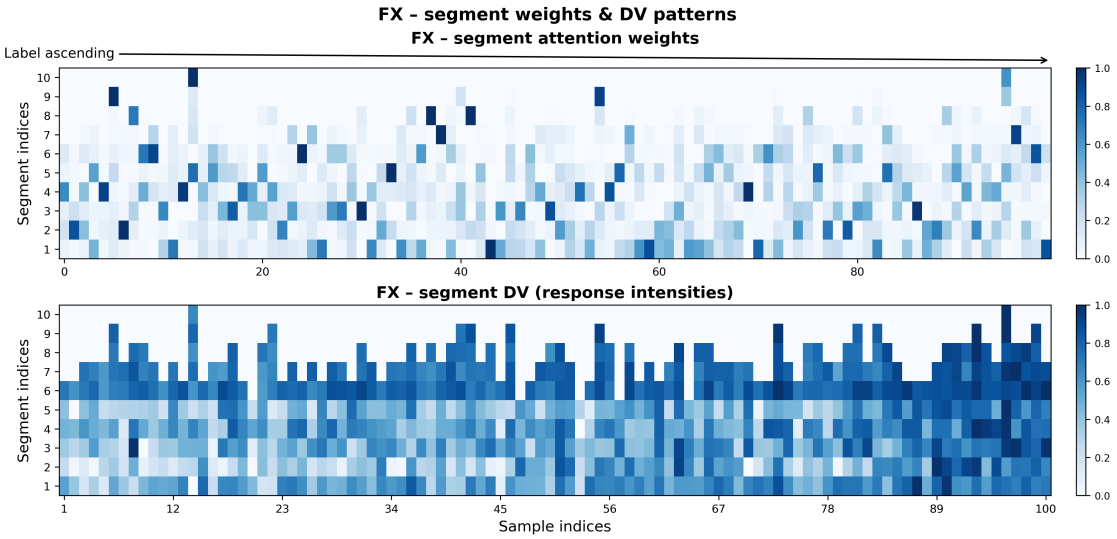

Figure 2. Qualitative visualization of score prediction and temporal attention on Balance Beam (top) and Floor Exercise (bottom). **Left:** Comparison between the predicted Difficulty/Execution scores and the Ground Truth. **Middle:** The predicted actionness curves (blue) align well with the ground-truth action intervals. **Right:** The most attended segments visualized as keyframes.

**Visualizations of Segment Attention and Response Intensities.** Further visualizations for FX, BB, and UB (Figures 3a–3c) show that segment difficulty value response intensities increase with higher scores (sorted x-axis), while attention weights remain sparse. This correlation confirms the model effectively learns grade-aware representations across different apparatuses.

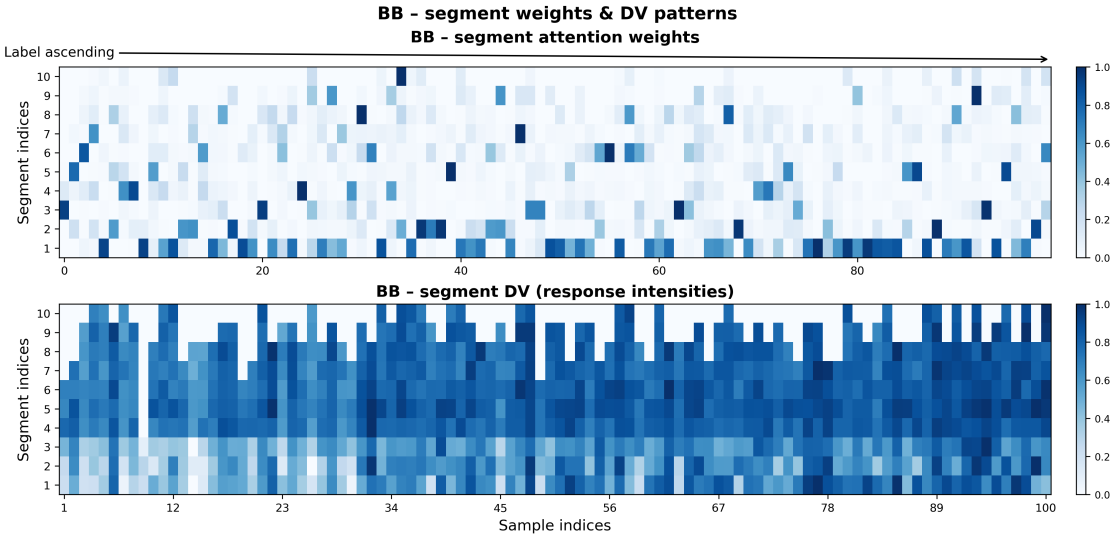
## References


[1] International Gymnastics Federation: 2022-2024 code of points (Women's artistic gymnastics) (2022).

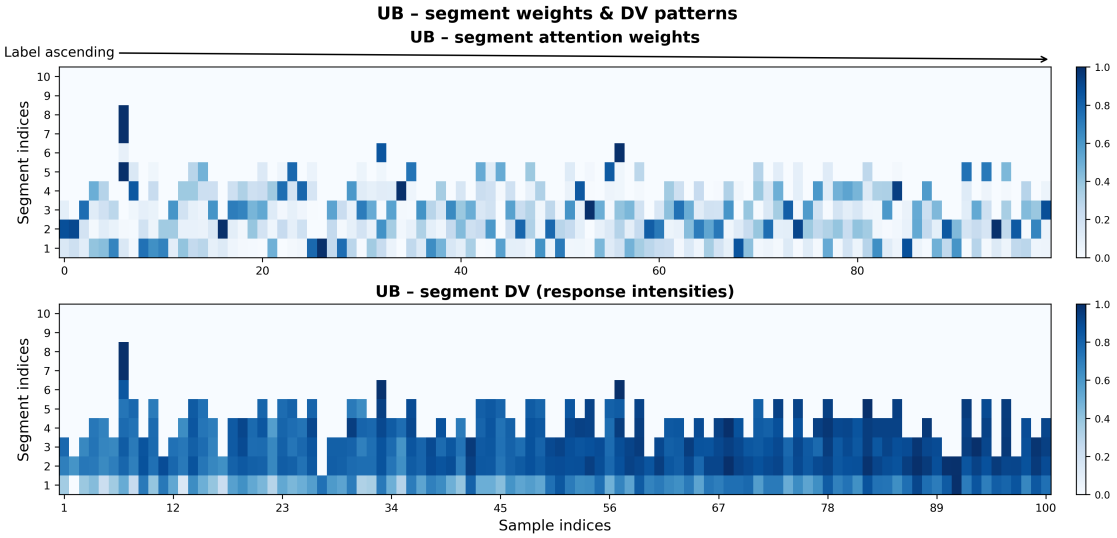
[2] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. *CVPR*, pages 3232–3241, 2022.




(a) Floor Exercise (FX)




(b) Balance Beam (BB)




(c) Uneven Bars (UB)

Figure 3. Qualitative visualization across different gymnastic apparatuses.