Abstract

Long-term action quality assessment is a task of evaluating how well an action is performed, namely, estimating a quality score from a long video. Intuitively, long-term actions generally involve parts exhibiting different levels of skill, and we call the levels of skill as performance grades. For example, technical highlights and faults may appear in the same long-term action. Hence, the final score should be determined by the comprehensive effect of different grades exhibited in the video. To explore this latent relationship, we design a novel Likert scoring paradigm inspired by the Likert scale in psychometrics, in which we quantify the grades explicitly and generate the final quality score by combining the quantitative values and the corresponding responses estimated from the video, instead of performing direct regression. Moreover, we extract grade-specific features, which will be used to estimate the responses of each grade, through a Transformer decoder architecture with diverse learnable queries. The whole model is named as Grade-decoupling Likert Transformer (GDLT), and we achieve state-of-the-art results on two long-term action assessment datasets.\(^1\)

1. Introduction

Action quality assessment (AQA) is a task to evaluate how well a specific action is performed and is usually modeled as a score regression task. Due to its rich application scenarios in the real world, such as sport events [16, 27, 30–32, 37, 43–45], surgical training [10–12, 21, 41] and daily skills [8, 9, 18], AQA has attracted growing attention from the computer vision community.

Compared with actions that only take a few seconds (e.g., diving), AQA of long-term actions (e.g., figure skating) is more challenging since they contain richer and more complex information. Intuitively, a long video is very likely to exhibit different levels of skill (e.g., excellent, good, fair or poor) at different parts [9], and we call the levels of skill as performance grades. For example, a perfect air twist, a substandard leg lifting, and a fall fault may occur in the same long-term action (figure skating). Therefore, we conceive that the quality score should be determined by the comprehensive effect of different grades exhibited in a video. In other words, we suppose that there exists an inherent mapping from grades to scores. This observation has hardly been discussed in pre-
Previous works [21, 27, 43, 45], and these existing works use MLP to directly regress the score from video representations, ignoring this inherent complexity.

In this work, we aim to explicitly model the influences of different grades on the score. To this end, we propose a novel scoring paradigm, named Likert scoring, which is inspired by the well-known Likert scale [19] in psychometrics and sociological investigation. A scale is a psychometric tool for quantitatively evaluating the psychological state of the respondent, which consists of several related questions or statements about different aspects. The respondent is asked to evaluate how well he/she agrees with each statement. Then the agreement degrees of each statement will be converted into quantified scores, and all scores are added to get a total score, which indicates the respondent’s mentality. In the context of this paper, we treat assessing a complex action as filling a “scale”, whose “statements (questions)” refer to the inherent performance grades. The input video is then required to “answer” the questions that how well it matches each grade, i.e., the response intensities are estimated for each grade from the video. These intensities will be combined with the pre-quantified scores to determine the final quality score. The underlying insight here is to evaluate a complex objective (i.e., action quality) by explicitly measuring several inherent components, which is consistent with the Likert scale. A brief illustration of this idea is shown in Figure 1.

Moreover, to fill the “scale”, we need “evidence” for each question (i.e., the information related to each grade from the video) to generate responses. For this purpose, we disentangle video features into different grade-aware features, which contain the grade-specific information. This procedure is called grade decoupling. Inspired by DETR [3], this step is implemented by a Transformer [39] decoder, which ingests a video feature sequence and a set of learnable vectors serving as the prototypes of various grades, and the grade-specific semantics are extracted from video features by these prototypes via the cross-attention mechanism.

Formally, we name our whole model as Grade-decoupling Likert Transformer (GDLT), which is composed of a standard Transformer [39] encoder-decoder architecture and a Likert Scoring Module (LSM). The former consists of a Temporal Context Encoder (TCE) and a Grade-aware Decoder (GAD). In the TCE, we leverage the self-attention mechanism to better explore the rich context information for each segment, which is critical for long video understanding. Then the GAD and LSM will perform grade decoupling and Likert scoring respectively. In summary, our main contributions are two-fold:

1. A novel assessment paradigm named Likert scoring inspired by psychological research is proposed to explore the comprehensive effect of different grades on the score.
2. A Transformer [39] encoder-decoder architecture is introduced to perform grade decoupling, which aims to extract grade-specific features used for Likert scoring from the input video. To the best of our knowledge, it is the first work to adopt the Transformer in AQA.

To evaluate our idea, we conduct experiments on two public long-term action assessment datasets: Rhythmic Gymnastics [45] and Fis-V [43]. Our model achieves state-of-the-art results on both datasets, demonstrating its effectiveness.

2. Related Work

Action Quality Assessment. AQA is generally regarded as a regression problem [11, 16, 21, 27, 28, 30–32, 40, 43–45], i.e., estimating a quality score for an action. Some early works [31, 32] directly adopt support vector regression to perform regression with the hand-crafted discrete cosine transform or deep C3D [38] features as input. To achieve a more accurate assessment, recent works [11, 16, 27, 28, 37, 40, 43–45] aim to solve some specific problems in AQA. For example, Tang et al. [37] utilize the label distribution learning to model score uncertainty. However, problems in long-term AQA are still relatively unexplored [43, 45]. Xu et al. [43] propose two LSTMs [14] to learn both local and global information. Zeng et al. [45] leverage static posture information to enhance video motion features and design a graph-based attention module for long-term temporal modeling. In this work, we explore various grades of performance implied in long videos and propose a novel scoring paradigm considering these grades instead of directly regressing the score.

However, some daily activities such as tying a tie have no professional criteria for accurate scoring. Doughty et al. [8, 9] address this issue by regarding AQA as a pairwise ranking problem, i.e., to determine which of a given pair of videos is better. Note that in [9], they propose separately modeling high-skill and low-skill parts in a video, and design loss functions to constrain the relationships between these two parts between a pair of videos, which is similar to our work. However, our proposed model is generalized to multiple grades instead of binary ones and can be used for direct score estimation.

Transformer. Transformer [39] was first introduced by Vaswani et al. for machine translation and sequence modeling. It proposes a self-attention mechanism allowing each element to see the whole sequence and to update itself by aggregating information from other elements. Due to its advanced ability to model global relationships, Transformer has dominated the natural language processing field [6, 33],
Figure 2. The overall framework of our proposed GDLT. The backbone extracts feature sequence from video segments, and the TCE enhances it by the context information. The GAD maintains a set of learnable vectors serving as prototypes of performance grades and exploits them to extract grade-aware features from context-enhanced video features. Finally, the grade-aware features are used to generate response intensities, which will be combined with quantitative values to calculate the final score.

3. Our Approach

In this section, we introduce our proposed Grade-decoupling Likert Transformer (GDLT) in detail. We first describe some preliminaries of our work in Section 3.1. Then we introduce three main components of GDLT, i.e., Temporal Context Encoder (TCE), Grade-aware Decoder (GAD), and Likert Scoring Module (LSM) in Section 3.2, Section 3.3 and Section 3.4, respectively. Figure 2 illustrates the overall framework of the GDLT.

3.1. Preliminaries

**Problem Formulation.** We first formulate the AQA problem. Following the practice in the real world (e.g., sports competitions), the action quality is measured by a score, which is a non-negative real number, and a higher score indicates better action quality. Naturally, the model is required to learn a mapping from videos to scores under the supervision of human expert annotations. Following [45], we normalize the labels to the interval [0,1] for more stable training.

**Feature Extraction.** Following the practice in long-term action understanding [15, 43, 45, 46, 48], we build GDLT upon the feature sequences extracted from non-overlapping video segments, each of which consists of several consecutive frames. The features are extracted via a well-designed video backbone (e.g., I3D [5], TSM [20], and VST [23]). Then, a 2-layer MLP is applied for reducing the dimension of backbone features. We denote the obtained feature sequence of a video with \( T \) segments as \( \{f_t\}_{t=1}^T \) where \( f_t \in \mathbb{R}^d \), serving as the input for GDLT.

**Grade Definition.** As described in Section 1, the grade is the level of quality. In this work, we define \( K \) grades, indexed from 1 to \( K \), to indicate action quality from bad to good with ascending index. Note that these grade indices are consistent with the subscripts of the grade prototypes (see Section 3.3), quantitative values (see Section 3.4) and other relevant symbols, namely, a relevant symbol with subscript \( k \) corresponds to the \( k \)-th grade.

3.2. Temporal Context Encoder

Since the features are independently extracted from the video segments, each \( f_t \) only contains information of a very small temporal region (i.e., current segment) and lacks oliganxthe context information. Therefore, a Transformer [39] encoder is adopted to first enrich segment-wise representations \( \{f_t\}_{t=1}^T \). The context information of each segment is obtained through weighted aggregation among all segment features, and the weights are determined by the semantic correlations between the current segment and others. This procedure is called the self-attention mechanism. Then the context information is added back to the original \( f_t \), and the summed vectors are passed into a small feed-forward network for further fusion. Multiple encoders can be stacked to gradually aggregate and refine context semantics. We denote the final context-enhanced features as \( \{f^{ctx}_t\}_{t=1}^T \), which will be used by the Grade-aware Decoder.
Figure 3. Illustration of the Grade-aware Decoder and grade decoupling mechanism. The shapes of important tensors are shown in gray. ⊗ denotes matrix multiplication. ⊕ denotes element-wise sum and is omitted in some residual connections for brevity. Query, Key and Value are three different linear projections.

3.3. Grade-aware Decoder

In the Grade-aware Decoder, we aim to extract the information related to different grades from context-enhanced video features \( \{f_t^{ctx}\}_{t=1}^T \). For this purpose, we introduce a set of \( K \) learnable vectors \( \{p_k\}_{k=1}^K \) as the prototypes of \( K \) performance grades to learn the distinct characteristics of them. Then inspired by DETR [3], the interaction between \( \{p_k\}_{k=1}^K \) and \( \{f_t^{ctx}\}_{t=1}^T \) is implemented by a parallel non-autoregressive and non-masked version of Transformer [39] decoder that consists of three parts: self-attention, cross-attention and a small feed-forward network (FFN). The self-attention mechanism is first applied for mining the relationships among \( K \) prototypes. We denote the updated prototypes after the self-attention as \( \{\hat{p}_k\}_{k=1}^K \). Then they will be used to extract the relevant information from video feature sequences through cross-attention, and this procedure is called grade decoupling. Figure 3 shows the details of GAD.

Grade Decoupling. The grade decoupling is implemented by the cross-attention mechanism. Specifically, the module takes \( \{f_t^{ctx}\}_{t=1}^T \) and \( \{\hat{p}_k\}_{k=1}^K \) as input, and first generates queries from \( \{\hat{p}_k\}_{k=1}^K \) while keys and values are transformed from \( \{f_t^{ctx}\}_{t=1}^T \) via three different linear projections:

\[
q_k = W_q \hat{p}_k, \quad k_t = W_k f_t^{ctx}, \quad v_t = W_v f_t^{ctx},
\]

where \( \{q_k\}_{k=1}^K \), \( \{k_t\}_{t=1}^T \) and \( \{v_t\}_{t=1}^T \) indicate queries, keys and values, respectively. After that, the semantic correlation between \( k \)-th grade and \( t \)-th video segment is measured by dot-product similarity between corresponding query-key pair:

\[
a_{k,t} = \frac{q_k^T k_t}{\sqrt{d}},
\]

where \( \sqrt{d} \) serves as a scaling factor. It shows how much the \( t \)-th segment is related to the \( k \)-th performance grade. The softmax function is then applied along temporal dimension \( t \) to produce normalized attention weights \( \hat{a}_{k,t} \) for information aggregation among values:

\[
\hat{p}_k^{agg} = \sum_{t=1}^T \hat{a}_{k,t} v_t.
\]

The above equation is applied for pooling video features via grade-dependent weights. Therefore, the results can be regarded as a kind of “pure substance” containing information only related to specific grades in the video, ideally.

We then leverage the obtained \( \{\hat{p}_k^{agg}\}_{k=1}^K \) to activate video-agnostic prototypes \( \{p_k\}_{k=1}^K \) by adding \( \{\hat{p}_k^{agg}\}_{k=1}^K \) back to \( \{p_k\}_{k=1}^K \), and the summed vectors are further refined by the FFN. Multiple decoders can also be stacked and the output of one layer serves as the input queries for the next one. The output of the last GAD layer is denoted as \( \{\hat{p}_k^{agg}\}_{k=1}^K \), and we call it grade-aware features.

Diversity of the Grade-aware Features. Intuitively, different grade prototypes should focus on different semantic patterns, so the grade-aware features should have low correlation. Therefore, we exploit a diversity loss to regularize them explicitly, inspired by [17, 36]. Specifically, we adopt a triplet loss [34] to ensure that the grade-aware features of different grades are far enough apart. Given a batch of \( B \) videos, we rewrite \( \{p_k^{att,(i)}\}_{k=1}^K \) as \( \{p_k^{att,(i)}\}_{k=1}^K \), where \( i = 1, 2, \ldots, B \). Each triplet consists of a grade-aware feature \( p_k^{att,(i)} \) of the \( k \)-th grade and \( i \)-th video as an anchor, a positive sample with the same grade and a negative one with a different grade. Hence, for each \( p_k^{att,(i)} \), we search for the hardest positive and negative pair distances \( D^+_{i,k} \) and \( D^-_{i,k} \) with:

\[
D^+_{i,k} = \max_j \text{dist}(p_k^{att,(i)}, p_k^{att,(j)}), \ j \neq i, \\
D^-_{i,k} = \min_m \text{dist}(p_k^{att,(i)}, p_m^{att,(n)}), \ m \neq k,
\]

where \( \text{dist}(\cdot, \cdot) \) is a pairwise distance metric. We use the cosine distance here:

\[
\text{dist}(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}.
\]

Then the diversity loss is defined as:

\[
\mathcal{L}_{\text{div}} = \frac{1}{B K} \sum_{i=1}^B \sum_{k=1}^K \max(0, D^+_{i,k} - D^-_{i,k} + \alpha),
\]

measured by dot-product similarity between corresponding query-key pair:

\[
a_{k,t} = \frac{q_k^T k_t}{\sqrt{d}},
\]

where \( \sqrt{d} \) serves as a scaling factor. It shows how much the \( t \)-th segment is related to the \( k \)-th performance grade. The softmax function is then applied along temporal dimension \( t \) to produce normalized attention weights \( \hat{a}_{k,t} \) for information aggregation among values:

\[
\hat{p}_k^{agg} = \sum_{t=1}^T \hat{a}_{k,t} v_t.
\]

The above equation is applied for pooling video features via grade-dependent weights. Therefore, the results can be regarded as a kind of “pure substance” containing information only related to specific grades in the video, ideally.

We then leverage the obtained \( \{\hat{p}_k^{agg}\}_{k=1}^K \) to activate video-agnostic prototypes \( \{p_k\}_{k=1}^K \) by adding \( \{\hat{p}_k^{agg}\}_{k=1}^K \) back to \( \{p_k\}_{k=1}^K \), and the summed vectors are further refined by the FFN. Multiple decoders can also be stacked and the output of one layer serves as the input queries for the next one. The output of the last GAD layer is denoted as \( \{\hat{p}_k^{agg}\}_{k=1}^K \), and we call it grade-aware features.

Diversity of the Grade-aware Features. Intuitively, different grade prototypes should focus on different semantic patterns, so the grade-aware features should have low correlation. Therefore, we exploit a diversity loss to regularize them explicitly, inspired by [17, 36]. Specifically, we adopt a triplet loss [34] to ensure that the grade-aware features of different grades are far enough apart. Given a batch of \( B \) videos, we rewrite \( \{p_k^{att,(i)}\}_{k=1}^K \) as \( \{p_k^{att,(i)}\}_{k=1}^K \), where \( i = 1, 2, \ldots, B \). Each triplet consists of a grade-aware feature \( p_k^{att,(i)} \) of the \( k \)-th grade and \( i \)-th video as an anchor, a positive sample with the same grade and a negative one with a different grade. Hence, for each \( p_k^{att,(i)} \), we search for the hardest positive and negative pair distances \( D^+_{i,k} \) and \( D^-_{i,k} \) with:

\[
D^+_{i,k} = \max_j \text{dist}(p_k^{att,(i)}, p_k^{att,(j)}), \ j \neq i, \\
D^-_{i,k} = \min_m \text{dist}(p_k^{att,(i)}, p_m^{att,(n)}), \ m \neq k,
\]

where \( \text{dist}(\cdot, \cdot) \) is a pairwise distance metric. We use the cosine distance here:

\[
\text{dist}(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}.
\]

Then the diversity loss is defined as:

\[
\mathcal{L}_{\text{div}} = \frac{1}{B K} \sum_{i=1}^B \sum_{k=1}^K \max(0, D^+_{i,k} - D^-_{i,k} + \alpha),
\]
 Score Generation. The final score is generated through a linear combination of quantitative values and response intensities among grades. Note that the score should be determined by the proportion of each grade to ensure that it falls within a valid interval (i.e., [0,1]). Hence we normalize \( \{ \hat{w}^g_k \}_{k=1}^K \) such that the sum is 1 to obtain new weights \( \{ w^g_k \}_{k=1}^K \):

\[
 w^g_k = \frac{\hat{w}^g_k}{\sum_{i=1}^K \hat{w}^g_i}. \tag{9}
\]

Finally, the quality score \( s \) is calculated as:

\[
 s = \sum_{k=1}^K w^g_k s^g_k. \tag{10}
\]

 Loss Function. To directly minimize the errors between estimated scores and labels, we adopt the mean-squared error (MSE) loss \( \mathcal{L}_{\text{MSE}} \) to train our model, together with the diversity loss term \( \mathcal{L}_{\text{div}} \) described in Section 3.3:

\[
 \mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{div}}, \tag{11}
\]

where \( \lambda \) is a trade-off hyper-parameter.

4. Experiments

We conduct experiments on two datasets: Rhythmic Gymnastics [45] and Fis-V [43] to evaluate our model. We first briefly introduce the datasets and common metrics. Then, we describe our implementation details and present the results. After that, we perform ablation studies to further analyze our model in depth and conduct some visualization for intuitive understanding.

4.1. Datasets and Metric

Rhythmic Gymnastics (RG). The RG dataset contains a total of 1000 videos of 4 rhythmic gymnastics actions with different apparatuses, i.e., ball, clubs, hoop, and ribbon. The length of each video is approximately 1.6 minutes with a frame rate of 25. There are 200 videos for training and 50 for evaluating in each kind of action. Following the practice of [45], we train individual models for each kind.

Figure Skating Video (Fis-V). The Fis-V dataset has 500 videos of ladies’ singles short program of figure skating. Each video is approximately 2.9 minutes long with a frame rate of 25. We follow the official split which has 400 videos for training and 100 for testing. All videos are annotated with two scores, namely, Total Element Score (TES) and Total Program Component Score (PCS), according to the competition rule. Following [43], we train two independent models for predicting these two scores.

Note that Fis-V is a substitute for MIT-Skating [32] and UNLV-Skating [31], which are also about figure skating but much smaller (150/171 videos, respectively). Thus, we no longer conduct experiments on them.
### Table 1. Comparisons of GDLT with other methods on RG and Fis-V datasets. Avg. is the average SRCC across all classes computed using Fisher’s z-value. † indicates the results of our reimplementation. ‡ indicates that the higher the metric, the better. Best results are in bold, second best are underlined.

| Methods                  | Features                          | Rhythmic Gymnastics |  |  | Fis-V |  |
|--------------------------|-----------------------------------|---------------------|---|---|      |---|
|                          |                                   | SRCC↑               |  |  |      |   |
|                          |                                   | Ball | Clubs | Hoop | Ribbon | Avg. | TES | PCS | Avg. |
| C3D+SVR [31]            | C3D [38]                          | 0.357 | 0.551 | 0.495 | 0.516 | 0.483 | 0.400 | 0.590 | 0.501 |
|                          |                                   | - - - | - - - | - -  | - -  | -    | 0.650 | 0.780 | 0.721 |
| MS-LSTM [43]            | C3D [38]                          | 0.515 | 0.621 | 0.540 | 0.522 | 0.551 | -    | -    | -    |
|                          |                                   | 0.621† | 0.661† | 0.670† | 0.695† | 0.663† | 0.660† | 0.809† | 0.744† |
| ACTION-NET [45]         | I3D [5] + ResNet [13]             | 0.684† | 0.732† | 0.733† | 0.754† | 0.728† | 0.694† | 0.809† | 0.757† |
|                          | VST [23] + ResNet [13]            | 0.746 | 0.802 | 0.765 | 0.741 | 0.765 | 0.685 | 0.820 | 0.761 |

**Metric.** Following previous works [31, 32, 43, 45], we adopt the Spearman’s rank correlation coefficient (SRCC) $\rho$ as the evaluation metric, which measures the monotonicity between the predicted series and the ground-truth series. It’s defined as follows:

$$\rho = \frac{\sum_i (x_i^r - \bar{x}^r)(y_i^r - \bar{y}^r)}{\sqrt{\sum_i (x_i^r - \bar{x}^r)^2 \sum_i (y_i^r - \bar{y}^r)^2}},$$

where $x^r$ and $y^r$ indicate the rankings of two series respectively. It ranges from -1 to 1 and the higher is the better. In addition, the average SRCC across classes (the word “class” refers to both action types in RG and score types in Fis-V) is calculated from individual per-class SRCCs using Fisher’s z-value as in [11, 27, 29, 37, 44].

**4.2. Implementation Details**

**Feature Extraction.** As described in Section 3.1, we first divide the video into non-overlapping segments, each of which is composed of 32 consecutive frames. Due to the rapid development of vision Transformer in recent years, we adopt a newly developed Video Swin Transformer (VST) [23] pretrained on Kinetics-600 [4], which is extended from Swin Transformer [22], as our backbone. Note that we don’t fine-tune it, following previous works on long-term action understanding [15, 43, 45, 48]. For mini-batch training, the number of segments is fixed to 68 for RG and 124 for Fis-V. If a video has more segments, we select continuous segments where the start position is randomly determined in each training iteration, as in [43, 45]. All segments are used when testing.

**Experimental Settings.** We use the 1-layer TCE and 2-layer GAD all with single-head attention to implement GDLT. The dimension of the latent space $d$ is 256, and the number of grades $K$ is set as 4 for all classes. We use the SGD with a momentum of 0.9 to optimize all models. The batch size is 32 and the learning rate is 0.01, and we then gradually decrease it to 0.0001 by a cosine annealing strategy [24]. For better convergence, we set different epochs for different models: 250/400/500/150/320/400 for RG(Ball) / RG(Clubs) / RG(Hoop) / RG(Ribbon) / Fis-V(TES) / Fis-V(PCS). The $\lambda$ in Equation (11) is 1.0 for RG and 0.5 for Fis-V. The $\alpha$ in Equation (6) is 1.0 for all models. To regularize the models, we use a dropout of 0.3/0.7 for RG/Fis-V and the weight decay is set as 1e-4. See more details in the supplementary material.

**4.3. Comparison with the State-of-the-art**

Table 1 shows the assessment results of our model and previous state-of-the-art methods on RG and Fis-V datasets. For fair comparison, we reimplement [43, 45] on the same VST features as ours. As shown in Table 1, our model outperforms the current state-of-the-art method ACTION-NET [45], especially on RG (by 0.037 on average), and achieves average improvements of 0.102 on RG and 0.017 on Fis-V compared with MS-LSTM [43]. Note that they both directly regress the score from the global feature of a video, so the results demonstrate the effectiveness of modeling the latent grades. Remarkably, ACTION-NET utilizes additional static image features to assist the dynamic video features. Instead, our GDLT uses video features only but still achieves superior results.

**4.4. Ablation Studies**

**Likert Scoring Paradigm.** To evaluate our proposed Likert scoring (LS) paradigm, we compare it with the common practice of AQA that directly regresses the score from the video-level global description via MLP. Hence, we adopt the common average pooling (AVG) and attention pooling (ATT) to generate this global description from the context-enhanced features $\{f_t^{ctx}\}_{t=1}^T$, as two baselines. The attention unit consists of two fully-connected layers with ReLU and softmax activation functions [9, 26, 35, 45]. At each time step $t$, it takes the feature $f_t^{ctx}$ as input and outputs a weight for aggregation.

Additionally, the output of Transformer [39] decoder (i.e., GAD) can be seen as a set of response features corresponding to specific semantic patterns [3, 17, 25, 36, 46]. Therefore, to further show that the superiority is achieved...
### Table 2. Ablation studies on the Likert scoring paradigm. AVG, ATT, TD-IS, TD-CS, TD-AS, and TD-LS indicate average pooling, attention pooling, Individual-Scoring, Concatenating-and-Scoring, Averaging-and-Scoring, and our Likert-Scoring, respectively. Best results are in bold, second best are underlined.

<table>
<thead>
<tr>
<th>Variants</th>
<th>Rhythmic Gymnastics</th>
<th>Fis-V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ball Clubs</td>
<td>Hoop</td>
</tr>
<tr>
<td>TCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG</td>
<td>0.773</td>
<td>0.754</td>
</tr>
<tr>
<td>ATT</td>
<td>0.711</td>
<td>0.685</td>
</tr>
<tr>
<td>TD-IS</td>
<td>0.715</td>
<td>0.701</td>
</tr>
<tr>
<td>TD-CS</td>
<td>0.697</td>
<td>0.719</td>
</tr>
<tr>
<td>TD-AS</td>
<td>0.705</td>
<td>0.787</td>
</tr>
<tr>
<td>TD-LS</td>
<td>0.746</td>
<td>0.802</td>
</tr>
</tbody>
</table>

Table 3. Ablation studies on the impact of TCE and $L_{div}$.

<table>
<thead>
<tr>
<th>Variants</th>
<th>Rhythmic Gymnastics</th>
<th>Fis-V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ball Clubs</td>
<td>Hoop</td>
</tr>
<tr>
<td>GDLT w/o TCE</td>
<td>0.725</td>
<td>0.693</td>
</tr>
<tr>
<td>GDLT w/o $L_{div}$</td>
<td>0.723</td>
<td>0.755</td>
</tr>
<tr>
<td>GDLT</td>
<td>0.746</td>
<td>0.802</td>
</tr>
</tbody>
</table>

Figure 5. Comparison with different numbers of grades. Best viewed in color.

exact by the scoring paradigm instead of the Transformer decoder, we construct three additional baselines (prefixed by “TD”) that generate scores from these response features by some common manners:

- **Individual-Scoring (TD-IS).** This baseline individually regresses scores from each response feature through different MLPs and then averages them.

- **Concatenating-and-Scoring (TD-CS).** This baseline concatenates all response features as the global representation and regresses the score from it directly.

- **Averaging-and-Scoring (TD-AS).** This baseline is similar to TD-CS but TD-AS generates the global representation by averaging all response features.

The results are shown in Table 2. Our proposed Likert scoring achieves best or second-best performance on all classes and outperforms others with a large margin on average, showing its robustness and effectiveness. Especially, compared with TD-IS, TD-CS, and TD-AS, the results show that we establish a more direct and meaningful connection between the response features and the final score than them, as explained in Section 1.

Moreover, similar to [9], we have an interesting finding that the inclusion of the attention unit decreases the performance from naive average pooling in some cases. We think it’s due to the huge gap between key segment selection and score regression. On the contrary, our model explicitly links the pooled features to specific grades. This operation can be regarded as an intermediate bridge, which alleviates the above gap and leads to superior results.

**Impact of TCE.** From Table 3, we can observe significant performance drops on both RG (-0.05) and Fis-V (-0.063) when removing the TCE from the full model. It demonstrates that the context information is important for long-term video understanding.

**Impact of $L_{div}$.** As shown in Table 3, performance declines when $L_{div}$ is not applied. $L_{div}$ provides additional regularization to assist the learning of GAD, which lacks of any direct supervision signal (only video-level score labels are provided).

**The Number of Grades $K$.** The number of grades $K$ is critical. As shown in Figure 5, $K = 4$ is suitable for all classes. We observe that the performance drops on most classes when increasing $K$, because too many grades may bring ambiguity to the model. Remarkably, the performance at $K = 2$ is relatively poor, which shows that the good/bad binary modeling [9] isn’t enough for complex action.

### Quantitative Strategies.

In Equation (7), we uniformly set the quantitative values $\{s^g_k\}_{k=1}^K$. We call this method as Uniform-Interval (UI), and examine two other possible methods here (note that for covering the entire score interval $[0,1]$, $s^g_0$ and $s^g_K$ must be 0 and 1):

- **Uniform-Sample (US).** We quantify the grades so that the ground-truth scores of samples in training set are uniformly distributed in $K - 1$ intervals.

- **Learnable-Width (LW).** We make the quantitative values learnable by taking the widths of $K - 1$ intervals as a part of trainable parameters. When scoring, we apply the softmax function for making them non-negative and sum up to 1, and generate quantitative values by normalized widths.

As shown in Table 4, the simplest method UI achieves the best average performance. Remarkably, making quantitative values learnable doesn’t improve the model since it may be difficult to optimize the model when both the values and combination weights are constantly changing.

### 4.5. Qualitative Analyses

**Visualization of Cross-attention Weights.** To figure out the patterns on which grade prototypes focus, we show
in Figure 6 the cross-attention weights computed by Equation (2) of each prototype on a video feature sequence in the last GAD layer. The different fluctuations of weight curves demonstrate different attention patterns. Specifically, the 1st grade prototype gives high weight to the moment when an athlete falls (marker a), which indicates poor performance. The 2nd-grade prototype detects more trivial parts that cannot be given high scores (marker b). The curve of the 3rd grade is relatively stable since its quantitative value $s_g^3 (0.667)$ is closest to the average label score of the dataset, so its grade pattern might be common. Finally, some technical movements related to high skill are attended by the prototype of the highest grade, such as air twist (marker c) and spinning (marker d).

**Visualizations of Response Intensities.** Figure 7 shows how the response intensities $\{\hat{w}_{kg}\}_{k=1}^{K}$ estimated by Equation (8) of a trained model change with the label scores. We find that the intensity of low grades decreases almost monotonically with the increment of sample scores, while the intensity of high grades increases, which is in line with human experience. See more visualizations in the supplementary material.

### 5. Conclusion

In this work, we propose a novel Grade-decoupling Likert Transformer (GDLT) to explore the comprehensive effect of different grades exhibited in the video on the score. For this purpose, a new scoring paradigm named Likert scoring is proposed, in which we regard the quality score as the combination between quantified grades and corresponding responses estimated from the video. Besides, a Transformer decoder is adopted to extract the grade-specific information, which will be used for response estimation, from the video via diverse learnable queries. The state-of-the-art results on two long-term AQA datasets demonstrate the effectiveness of our model.

### 6. Acknowledgment

This work was supported partially by the NSFC (U21A20471, U1911401, U1811461), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), and the Key-Area Research and Development Program of Guangzhou (202007030004).
References


