This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

From Beats to Scores: A Multi-Modal Framework for Comprehensive Figure Skating Assessment

Fengshun Wang Qiurui Wang* Dan Chen
Capital University of Physical Education and Sports
{wangfengshun2023,wangqiurui,chendan2018}@cupes.edu.cn

Abstract

Accurate quantitative evaluation of the Technical Elements Score (TES) and Program Components Score (PCS) in figure skating requires exceptional skill and professionalism, making it a highly challenging task. It requires not only meticulous observation of athletes' technical movements but also the ability to appreciate and assess artistic elements according to the current scoring criteria. The multi-modal large language model (MLLM) is rapidly advancing in its ability to understand various forms of information, such as images, videos, and audio. Previous research efforts have mainly utilized audio and video in isolation, but effective evaluation in figure skating requires a unified approach. In our work, we developed a multi-modal method that first identifies sub-movements through audio-guided videos by leveraging the phenomena of "hit the beat" and integrates audio and video features specifically tailored for figure skating. This integration enhances the coordination between these modalities. Building upon this foundation, we propose a comprehensive assessment model that utilizes multi-modal series representation learning to derive TES and PCS scores and generate text-based competition evaluations based on video, audio, and contextual prompts. Extensive experiments have proven that our proposed method has state-of-the-art scoring ability and generalization performance. Our code is available at https://github.com/ycwfs/Figure-Skating-Quality-Assessment.

1. Introduction

The assessment of figure skating performances is a complex task, generally requiring expert human judges to evaluate both technical proficiency and artistic expression. However, recent advancements in Deep Learning and Language Models have facilitated the development of systems capable of automatically assessing figure skating quality, promising to enhance objectivity and efficiency in scoring processes. Early work focused on capturing spatiotemporal features from video data. Britosh et al. [26] introduce a framework using 3D convolutional neural networks (C3D) for feature extraction, combined with support vector regression (SVR) and long short-term memory (LSTM) networks for score prediction. Building on this, Xu et al. [35] propose a deep architecture combining Self-Attentive LSTM and Multi-scale Convolutional Skip LSTM. This method efficiently learned both local and global sequential information in figure skating videos, demonstrating the importance of capturing multi-scale temporal dependencies. Beyond video analysis, recent efforts have been made to enhance semantic representation for scoring figure skating. Du et al. [10] propose a semantics-guided network that leverages knowledge transfer between semantic and visual domains, improving the interpretation of actions. Moreover, approaches incorporating pose estimation, such as the Skeleton-Based Action Quality Assessment by Li et al. [18], have been pivotal in capturing the nuanced motion details essential for accurate assessment. Recognizing the importance of both visual and audio cues in figure skating, Xia et al. [33] introduce a novel architecture that extends the MLP framework into a multi-modal fashion. Their approach uses a designed memory recurrent unit (MRU), effectively learning long-term representations. Ji et al. [15] employ an uncertainty score disentanglement solution and subaction localization to decouple PCS-oriented and TES-oriented representations from skating sequences, enabling differential learning for the two score prediction types. In addition, several papers [15, 21, 26, 33] have made significant contributions through dataset creation which leading the development of figure skating AQA.

The recent advancements in Multi-modal Large Language Models (MLLMs) have significantly enhanced the capabilities of artificial intelligence in understanding and synthesizing information across various modalities, including text, image, and video. These developments are particularly relevant for applications such as figure skating quality assessments, where complex visual and semantic data

*indicates the corresponding author

must be processed and interpreted. A notable advancement is the development of general MLLMs, As is seen in the InternVL model family, it enables vision-language task performance by incorporating a high-quality multi-modal dataset [7]. Such features are beneficial for multi-modal understanding tasks, often involving diverse linguistic and visual elements.

Previous research has explored the use of audio and video in analyzing performance data, yet it has largely overlooked the direct alignment and integration of these different modalities. Moreover, existing temporal action localization modules are typically applied directly for scoring without modification. Our approach introduces an innovative framework featuring an audio-guided action localization module. This module leverages the "hit the beat" phenomenon, where athletes' movements synchronize with musical beats during performances, allowing for more natural movement and transition segmentation. Furthermore, our study investigates the potential of integrating various modalities to enhance evaluation accuracy in figure skating. We propose a method that learns contextual representations critical for understanding and generating multi-modal sequences. This work addresses a gap in existing research, which often treats audio and visual data separately without fully utilizing their synchronized integration within a single evaluative framework.

The main contributions can be summarized as follows:

- We propose an audio-guide sub-actions segmentation module that can automatically separate movements and transitions to obtain more fine-grained motion information for scoring.
- We developed an efficient method to fuse audio and video features of figure skating, addressing the challenge of coordinating different modalities in multi-modal prediction.
- We devise a multi-modal series understanding and generation method that can learn multi-modal contextual representation besides generating language and numerical assessment to evaluate athletes' performance accurately.

2. Related Works

Action Quality Assessment (AQA) has emerged as a significant area of research in computer vision and sports analytics. This field focuses on developing automated systems to evaluate the quality of human actions, particularly in sports[6, 9, 15, 20, 22, 24, 34, 37, 38, 42, 43, 46, 47], physical activities, medical care[2, 19, 39, 41]. Early AQA research laid a crucial foundation using traditional machine learning and hand-crafted features[28], establishing the fundamental principles for computationally assessing action quality. Recognizing the importance of temporal dynamics in action quality, researchers have leveraged transformer architectures[1] to capture long-range dependencies, significantly advancing temporal modeling in AQA. The field has evolved towards fine-grained analysis and interpretability, with studies like those by Yu et al. [36] and Doughty et al. [23] providing detailed insights and aligning models with human-defined evaluation criteria. Multitask learning approaches have enhanced AQA by exploiting shared information across related tasks. Parmar and Morris [27] demonstrated the effectiveness of this method in improving overall performance. Innovative graph-based techniques, such as those developed by Pan et al. [25], have enabled more sophisticated modeling of the complex interactions underlying action quality. Addressing the inherent ambiguity in action assessment, researchers like Tang et al. [29] have introduced uncertainty modeling, shifting from point estimates to more nuanced score distributions. The integration of multiple data sources has led to more comprehensive assessments, as exemplified by Liu et al.'s [43] work on multi-modal approaches to AQA. Recent advancements in AQA tackle emerging challenges, with Chen et al. [6] addressing AI-generated content, Kumar et al. [24] exploring neuro-symbolic methods for more robust and interpretable systems and Gedamu et al. [11] utilizing visual-semantic to parse temporal dimension for assessing actions.

Multi-modal Large Language Models Recent advancements in Multi-modal Large Language Models (MLLMs) have significantly enhanced the field of artificial intelligence, particularly in tasks involving multi-modal information understanding and evaluation. For instance, Chen et al. [5] introduces a large-scale vision-language foundation model that demonstrates powerful visual capabilities. The paper [7] presents InternVL 1.5, an MLLM that bridges the capability gap between open-source and proprietary commercial models in multi-modal understanding. It introduces improvements such as a strong vision encoder, dynamic high-resolution processing, and high-quality bilingual datasets. In the video domain, Wang et al. [30] introduces a family of large-scale video-language models designed for high-quality video descriptions. Wang et al. [32] presents a new family of video foundation models that are talented in video recognition, video-text tasks, and video-centric dialogue. Addressing the challenge of assessing MLLMs, Chen et al. [4] introduces a novel benchmark for evaluating MLLMs' ability to assist judges across diverse modalities. The study reveals that while MLLMs show remarkable human-like discernment in pair comparisons, there are still challenges in scoring evaluation and batch ranking tasks.

These advancements in MLLMs' multi-modal information understanding and scoring capabilities have significant implications for tasks like figure skating assessment. The ability of these models to process and analyze complex visual and temporal information, combined with their capacity to generate detailed descriptions and evaluations, sug-

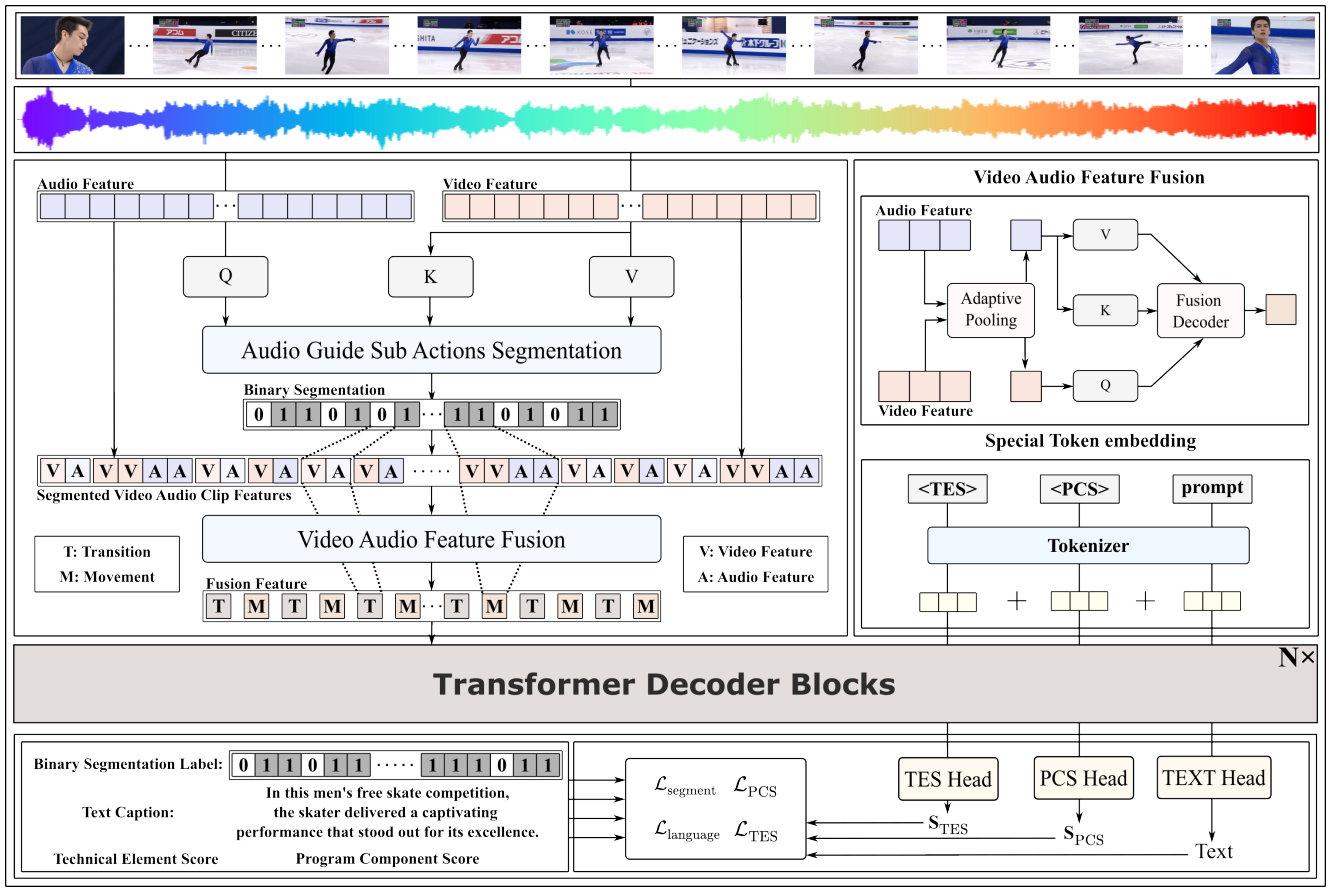


Figure 1. **Overview of our method.** We begin by identifying sub-actions using the audio-guide and constructing a series of multi-modal features. Next, we fuse the corresponding video and audio feature clips to generate fusion tokens. The input series is then concatenated with special tokens and a text prompt. After performing context representation learning, we are able to regress scores and generate evaluation texts.

gests potential applications in judging sports performances. However, there are still challenges in achieving consistent and unbiased scoring across different evaluation tasks, which would need to be addressed for reliable use in competitive sports judging.

3. Methodology

In this section, we introduce our multi-modal audio-guide multi-task figure skating quality assessments architecture which is shown in Fig. 1, we first segment sub-actions with the audio-guide and construct multi-modal feature series, then we fuse the corresponding video audio feature clips to obtain fusion tokens, the input series were concatenated with special tokens and text prompt, After the context representation learning, we can regress scores and generate evaluation texts.

3.1. Preliminaries

Problem Definition The objective of traditional Action Quality Assessment is to obtain numerical scores from given videos, but in the field of figure skating, the quality of an athlete's performance in competition depends on complexity criteria according to the International Skating Union[1]. For instance, the Technical Element Scores need to evaluate each movement in the show, and Program Component Scores should assess the overall program content which includes the interpretation of music and transitions especially. In our method, we extract the audio $A \in \mathbb{R}^T$ from the given video $V \in \mathbb{R}^{HWC \times T}$ and generate additional text descriptions $I \in \mathbb{R}^L$. Our objective is to predict the accurate TES and PCS scores using multi-modal input.

Featue Extractor We use pre-trained 3D Conv and 2D Conv backbones to extract feature vectors for video and

[1]https://www.isu.org/

its corresponding audio. We then apply two linear layers with Relu activation to project them to the same embedding space, resulting in video and audio features $F_v, F_a \in \mathbb{R}^{T\times D}$, where D denotes the embedding dimension.

3.2. Audio-Guide Sub Actions Segmentation

A complete figure skating program includes many movements, and each movement and its transitions are important for scoring. Based on the phenomena of "hit the beat", we use an audio feature to guide sub-actions segmentation.

The Audio-Guide attention mechanism performs cross-modality attention between the audio and video features, generating a unified hidden representation F_h. The attention mechanism learns to focus on specific audio signals that guide the segmentation of actions in the video:

$$F_h = \text{CrossAttention}(F_a, F_v, F_v), \tag{1}$$

where $F_h \in \mathbb{R}^{T\times D}$, with D being the shared embedding dimension.After the attention mechanism, the hidden states F_h are passed through the segmentation Head to predict the segment-wise action probabilities:

$$\mathcal{P} = \sigma(\text{Linear}(\text{ReLU}(\text{Linear}(F_h)))), \tag{2}$$

where $\mathcal{P} \in \mathbb{R}^L$ represents the action segmentation probabilities for each segment, and σ is the sigmoid activation function that outputs values between 0 and 1.

The segmentation decision is made by applying a learnable threshold τ to the predicted probabilities:

$$C_i = \mathbb{1}(\mathcal{P}_i > \tau) = \begin{cases} 1, & \text{if } \mathcal{P}_i > \tau, \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where C is a binary tensor of shape L, indicating whether each frame contains a sub-action (1) or transition (0).

To further process the localized sub-actions, consecutive frames with the same classification in C are grouped into segments. For the segment k, both video and audio features are pooled:

$$\tilde{V}_k = \text{AdaptivePool}(F_v^k), \tilde{A}_k = \text{AdaptivePool}(F_a^k), \tag{4}$$

where $\tilde{V}_k, \tilde{A}_k \in \mathbb{R}^D$ are the pooled segment features for the k-th segment.

3.3. Video Audio Feature Fusion

For the video-audio fusion, we apply a Fusion-Decoder to effectively integrate audio and video features, thereby enhancing the multi-modal learning process. The decoder treats the segment video features $\tilde{V}_k$ as key-value pairs, while the segment video features $\tilde{A}_k$ is used as the query. This setup allows the decoder to align and fuse relevant audio information into the video representations. The process is formally described as:

$$F_{va}^k = \text{LN}(\text{Fusion-Decoder}(\tilde{A}_k, \tilde{V}_k, \tilde{V}_k)), \tag{5}$$

Here, F_{va}^k represents the fused representations which are normalized through Layer Normalization (LN) after the Fusion-Decoder has processed them. This ensures that the learned multi-modal representations maintain robustness and are able to leverage the complementary strengths of both video and audio modalities to support downstream tasks. Through this fusion strategy, we achieve a coherent integration where video cues can enhance the perception of audio content and vice versa, leading to an improved understanding of multi-modal inputs.

3.4. multi-modal Embedding Preparation

Token Embedding For managing raw text prompts, we utilize a Byte-Pair Encoding (BPE) tokenizer to transform the text into input IDs, denoted as $I \in \mathbb{R}^{L'}$. Additionally, we define special tokens, such as <TES> and <PCS> denoted as E_{tes}, E_{pcs}, to mark contextual or structural elements within the data.

To construct comprehensive token sequences capturing multi-modal information, we systematically concatenate the fused video and audio features from N segments in one video, with the special tokens E_{tes}, E_{pcs}, followed by the encoded text tokens. This sequential combination results in a robust token sequence $F_e \in \mathbb{R}^{L_{max}}$ that effectively integrates diverse media inputs into a cohesive format.

Formally, the enriched embeddings are represented as:

$$F_e = \text{Concat}(F_{va}^N, E_{tes}, E_{pcs}, E_I, E_{pad}), \tag{6}$$

where F_e reflects the integration of multi-modal inputs and padding embeddings E_{pad} to ensure uniform sequence lengths. This concat strategy enhances the model's capability to process and understand complex, mixed-modality information, paving the way for better performance in downstream tasks.

3.5. Context Representation Learning

In our context representation learning framework, we employ a multi-layer transformer decoder architecture to effectively capture interdependencies and contextual information within sequential data, where we transform input features F_e through a series of operations that capture contextual dependencies. The critical components of this architecture are formalized as follows:

$$\begin{aligned} \hat{F} &= \text{DP}(\text{LN}(F_e + \text{MHA}(F_e, F_e, F_e; Mask))) + F_e, \\ F_r &= \text{DP}(\text{LN}(\hat{F} + \text{FFN}(\hat{F}))) + \hat{F}, \end{aligned} \tag{7}$$

where $\text{MHA}(\cdot)$ denotes the Multi-Head Attention mechanism applied to the input F_e, used for capturing interdependencies across the features. $\text{LN}(\cdot)$ is the Layer Normalization applied to stabilize the feature distribution both before and after MHA. $\text{FFN}(\cdot)$ denotes the Feedforward Network operation that processes each feature independently.

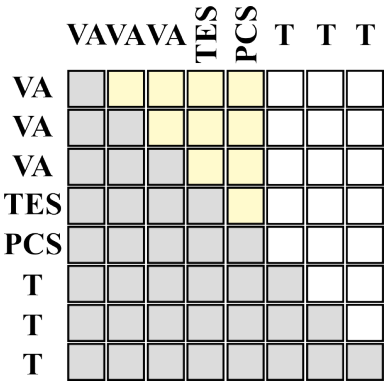


Figure 2. The attention mask we adopt

DP$(\cdot)$ is the Drop Path technique applied to randomly bypass certain layers for regularization. The residual connections are incorporated to form the output representations, which improves the flow of gradients during training. After that, we get the feature vectors with rich contextual representation $F_r \in \mathbb{R}^{L_{max}}$.

This comprehensive operation effectively integrates multiple transformation layers into a coherent representation, enhancing the model's capacity to understand and process complex sequential data.

In our architecture, the attention mask $Mask$ is crucial for controlling the visibility of tokens during the Multi-Head Attention (MHA) process. It prevents text tokens from being visible to the TES and PCS tokens while still allowing interdependencies within specific groups to be processed effectively.

The attention mechanism with the mask is defined as:

$$\text{MHA}(\text{Q}, \text{K}, \text{V}; \text{Mask}) = \text{softmax}(\frac{\text{QK}^T}{\sqrt{d_k}} + \text{Mask}) \cdot \text{V} \quad (8)$$

where: Q, K, V represent the Query, Key, and Value matrices derived from the input F_e. d_k is the dimensionality of the Key vectors, used to scale the dot products. $Mask$ is an additive mask where elements are set to $-\infty$ for the masked positions, which can effectively prevent attention to undesired tokens.

In the provided attention matrix Fig. 2, the attention mask is visually depicted. The mask is designed to ensure symmetric scoring through the exposed special yellow parts and to maintain causal text generation using the masked white parts. It achieves alignment by selectively occluding certain tokens during computation, which enforces task-specific constraints that enhance contextual understanding.

3.6. Scores Regression and Text Generation

Once the representation F_r is derived from the decoder layers, we perform the following operations:

Scores Regression The TES and PCS tokens are utilized for regression purposes. The regression process involves the extraction of specific states for TES and PCS from the representation F_r:

$$h_{\text{TES}} = F_r[:, :, p-2], \quad h_{\text{PCS}} = F_r[:, :, p-1], \quad (9)$$

where p refers to the index for the TES and PCS tokens, ensuring their separation from the text sequence. To perform regression for TES and PCS scores, the extracted states undergo a transformation through a series of linear operations:

$$\begin{aligned} S_{\text{tes}} &= W_2^{\text{TES}} \cdot \text{ReLU}(W_1^{\text{TES}} \cdot h_{\text{TES}} + b_1^{\text{TES}}) + b_2^{\text{TES}}, \\ S_{\text{pcs}} &= W_2^{\text{PCS}} \cdot \text{ReLU}(W_1^{\text{PCS}} \cdot h_{\text{PCS}} + b_1^{\text{PCS}}) + b_2^{\text{PCS}}, \end{aligned} \quad (10)$$

where W_1^{TES}, W_2^{TES}, b_1^{TES}, b_2^{TES} and similar parameters for PCS are learnable weights and biases in the regression network.

Text Generation For text generation, the language model head transforms the relevant portion of the features, excluding TES and PCS tokens. This is transformed via a linear layer for the vocab distribution:

$$S_{\text{t}} = W_{\text{lm}} \cdot F_r[:, :, p:] \quad (11)$$

where W_{lm} is the weight matrix mapping the hidden representations to the vocabulary size, enabling subsequent text generation tasks.

Overall, this dual-path architecture combining score regression and text generation facilitates robust learning and contextual understanding within linguistic and scoring tasks.

3.7. Optimization

In the optimization process, we focus on enhancing three key aspects: sub-actions segmentation, score regression for TES and PCS, and the generation of coherent textual outputs. This involves employing carefully defined loss functions to create an integrated optimization framework as follows:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{TES}}^{\text{MSE}} + \beta \mathcal{L}_{\text{PCS}}^{\text{MSE}} + \lambda \mathcal{L}_{\text{segment}}^{\text{BCE}} + \delta \mathcal{L}_{\text{language}}^{\text{MSE}} \quad (12)$$

Here, $\mathcal{L}_{\text{TES}}^{\text{MSE}}$ and $\mathcal{L}_{\text{PCS}}^{\text{MSE}}$ represent the Mean Squared Error losses associated with the regression of TES and PCS scores, respectively. And $\mathcal{L}_{\text{segment}}^{\text{BCE}}$ is the Binary Cross Entropy loss. Meanwhile, $\mathcal{L}_{\text{language}}^{\text{CE}}$ denotes the Cross-Entropy loss employed for accurate text generation. The α, β, γ and δ are parameters that are designed to balance the different parts of loss $\mathcal{L}$. Notice that the following conditions should be satisfied: $\alpha + \beta + \gamma + \delta = 1$.

Dataset	Method	Spearman Correlation (⇑)		Mean Square Error (⇓)	
		TES	PCS	TES	PCS
Fis-V	M-BERT [17]	0.68	0.82	27.73	12.38
	MS-LSTM [35]	0.65	0.78	19.49	8.41
	Skating-Mixer [33]	0.68	0.82	19.57	7.96
	Semantic-Guide [10]	0.70	0.83	19.05	7.96
	Action-Net [44]	0.81	0.70	-	-
	GDLT [34]	0.82	0.69	-	-
	LUSD-Net [15]	0.82	0.68	-	-
	Ours	**0.84**	**0.85**	**16.32**	**7.82**
FS1000	M-BERT [17]	0.79	0.75	131.28	15.28
	MS-LSTM [35]	0.86	0.80	94.55	11.03
	Skating-Mixer [33]	0.88	0.82	81.24	9.47
	Sementic-Guide [10]	0.89	0.85	79.08	8.40
	Ours	**0.90**	**0.87**	**78.42**	**7.85**
FineFS	TSA [36]	0.60	0.78	-	-
	MS-LSTM [35]	0.55	0.61	-	-
	GDLT [34]	0.64	0.79	-	-
	LUSD-Net [15]	0.73	0.84	-	-
	Ours	**0.76**	**0.88**	**97.95**	**26.21**

Table 1. Comparison with the state-of-the-art approaches on three datasets

4. Experiments and Results

4.1. Datasets

We use the benchmark Fis-V [35], FS1000[33], FineFS[15] to evaluate our figure skate quality assessment model.

The fis-V dataset contains 500 videos about ladies' singles short program with 149 players from more than 20 countries. The FS1000 dataset is larger, with 1604 videos from elite competitions covering all figure skating program types. Each video has about 5000 frames at a frame rate of 25 and is annotated with ground-truth scores. The FineFS dataset encompasses a total of 1,167 figure skating examples: 729 are from short programs, each lasting around 2 minutes and 40 seconds, and 438 are from free skating, each approximately 4 minutes in duration. Among these, there are 570 samples from male athletes and 597 from female athletes. All the videos are recorded at a frame rate of 25 frames per second. We divided these benchmarks into training and testing sets according to their original configuration.

Text Description Generate Due to the figure skating dataset currently only containing numerical scores, we need additional text-assisted description, therefore, we first summarize the total score of the competitions, then we define a few thresholds according to the quantile of the total score, for each threshold interval we prepare a rich corpus containing descriptions of athletes' competitive performance. Finally, we map the scores of each sample to the corresponding interval to obtain its textual description label.

4.2. Implementation Details

We input 24 continuous frames at a rate of 24 frames per second into an I3D model [3] that has been pre-trained on the Kinetics-400 dataset [16] to extract RGB features. Additionally, the audio features are extracted using VGGish [13], which has been pre-trained on AudioSet [12], for every 1 second to align with the RGB features. Furthermore, the fusion decoder, as indicated in Eq. (5), integrates both modalities and is configured with 32 transformer blocks to enable comprehensive cross-modal representation and interaction. The multi-modal decoder consists of 64 blocks and our model is trained for 100 epochs with a batch size of 1 and a learning rate of $2e-5$ on a single NVIDIA A800 GPU. The Byte-Pair Encoding (BPE) tokenizer has a vocabulary size of 49,413, with a token embedding dimension of 1024. The prompt is universally set to "How did the ath-

lete perform in the show?". The maximum token length is constrained to 100, providing an equilibrium between computational efficiency and the capacity to represent intricate interactions. The weights within the loss function $\alpha, \beta, \gamma, \delta$ is set at 0.3, 0.3, 0.2, and 0.2, respectively.

4.3. Comparison with the SOTA methods

Tab. 1 presents a comparison of various methods for different datasets: Fis-V, FS1000, and FineFS. Each method is evaluated based on two metrics: Spearman Correlation (higher values indicate better performance) and Mean Square Error (lower values are better) across two tasks, TES and PCS. It is important to note that the Fis-V and FS1000 datasets lack temporal annotations for each movement. Therefore, we continue to train them without the segment loss function using the model that pre-trains on the FineFS dataset.

Our method demonstrates superior performance across all the evaluated datasets. Specifically, in the Fis-V dataset, our approach outperforms existing methods with all evaluation metrics, especially at the higher TES Spearman Correlation. For the FS1000 dataset, our model again achieves notable performance improvements. We obtain Spearman Correlation values of 0.90 and 0.87 for TES and PCS, respectively, along with Mean Square Errors of 78.42 and 7.85.

Lastly, on the FineFS dataset, our method maintains its superior ranking, achieving a Spearman Correlation of 0.76 for TES and 0.88 for PCS, with Mean Square Errors of 97.95 and 26.21. The high correlation scores, particularly for PCS, underscore our model's capability to grasp the nuances of scoring criteria more effectively than previous approaches.

The consistent performance across all datasets validates the generalization and efficiency of our approach. Our method's ability to handle datasets without temporal annotations, by leveraging pre-trained models on FineFS and refining them without segment loss, underscores its adaptability and comprehensive design. Furthermore, the use of advanced techniques such as audio-guided segmentation contributes to robust performance, particularly in complex scoring assessments where fine details and temporal dynamics are crucial.

4.4. Ablation Study

In this section, we conduct ablation studies on the FineFS dataset, which provides detailed temporal annotations for each movement.

4.4.1 Ablation of Audio-guide segmentation strategy

In this part, we present the ablation study of the Audio-guide segmentation strategy. We compare different methods including Video-guide, Audio only, Video only, and Audio-guide. The performance is evaluated using two metrics: SP (Spearman Correlation) and MSE (Mean Squared Error). SP is further evaluated using TES (Temporal Event Segmentation) and PCS (Positional Context Segmentation). The results are summarized in Tab. 2

Method	SP		MSE	
	TES	PCS	TES	PCS
Audio only	0.65	0.69	101.32	29.82
Video only	0.71	0.74	99.85	28.98
Video-guide	0.72	0.84	98.52	27.89
Audio-guide	**0.76**	**0.88**	**97.95**	**26.21**

Table 2. Ablation of Audio-guide segmentation strategy

Method	SP		MSE	
	TES	PCS	TES	PCS
causal attention	0.72	0.86	101.25	29.32
full attention	**0.76**	**0.88**	**97.95**	**26.21**

Table 3. Ablation of attention mask

The results indicate that the Audio-guide strategy outperforms other methods across all metrics. Specifically, the Audio-guide method achieves a TES of 0.76 and PCS of 0.88 in SP, and an MSE of 97.95 and 26.21 in TES and PCS respectively. This demonstrates the effectiveness of the Audio-guide approach in improving both spatial precision and reducing mean squared error compared to the other strategies.

4.4.2 Ablation of attention mask in text generation

In this section, we present an ablation study focusing on the impact of different attention mechanisms before text tokens on the model's performance. We specifically compare causal attention and full attention. The results of this comparison are summarized in Table 3. The results show that the full attention mechanism consistently outperforms causal attention in both SP and MSE across the evaluated tokens. Overall, these findings suggest that the use of full attention significantly enhances the model's ability to scoring through special tokens more effectively than causal attention, emphasizing the importance of the chosen attention mechanism in model performance.

4.4.3 Ablation of the matching degree between score and textual description

This ablation study investigates how the matching quality of textual descriptions affects the performance evaluation

Method	SP		MSE	
	TES	PCS	TES	PCS
w/o text	0.74	0.87	101.72	28.12
Misaligned	0.69	0.85	105.42	31.98
Aligned	**0.76**	**0.88**	**97.95**	**26.21**

Table 4. Ablation of matching degree between score and textual description.

Loss		SP		MSE	
$\mathcal{L}_{seg}$	$\mathcal{L}_{lan}$	TES	PCS	TES	PCS
✓		0.74	0.86	98.12	26.84
	✓	0.72	0.85	98.56	28.24
✓	✓	**0.76**	**0.88**	**97.95**	**26.21**

Table 5. Ablation study: impact of loss

of models. We assess three configurations: (1) a model without textual descriptions output, (2) scenarios where textual descriptions are misaligned with the scores (e.g., poor-quality descriptions match high-scoring samples), and (3) scenarios where textual descriptions are aligned with the scores (e.g., good-quality descriptions match high-scoring samples). The results presented in Table 4 provide the following insights: Without textual descriptions, the model performs lower accuracy. When poor-quality descriptions match high-scoring samples, SP scores and MSE worsen, highlighting the risk of misalignment in descriptive guidance. With high-quality matching descriptions, SP and MSE scores improve significantly, showing that well-aligned descriptions enhance both accuracy and coherence. In summary, this study underscores the importance of aligned textual descriptions for optimal model performance.

4.4.4 Ablation of loss function

In this ablation study, we evaluate the impact of different components of the loss function on our model's performance. Specifically, we investigate the individual and combined effects of segmentation loss ($\mathcal{L}_{\text{seg}}$) and language loss ($\mathcal{L}_{\text{lan}}$). Tab. 5 summarizes our findings. It is important to note that this study is based on situations where both the $\mathcal{L}_{\text{TES}}$ and $\mathcal{L}_{\text{PCS}}$ were applied.

From the results, it is evident that including the segmentation loss ($\mathcal{L}_{\text{seg}}$) alone achieves a TES of 0.74 and a PCS of 0.86 under the SP criterion, and 98.12 and 26.84 under the MSE criterion. On the other hand, using only the language loss ($\mathcal{L}_{\text{lan}}$) results in a slight decrease in performance with a TES of 0.72 and a PCS of 0.85 under SP, and 98.56 and 28.24 under MSE. Interestingly, the combination of both loss functions ($\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{lan}}$) yields the best performance across all metrics, with TES and PCS scores of 0.76 and 0.88 respectively under SP, and 97.95 and 26.21 under MSE. This demonstrates the scoring benefits of integrating movement segmentation and text generation components in training the model.

Method	SP		MSE	
	TES	PCS	TES	PCS
Qwen2-VL [31]	0.70	0.75	239.5	230.14
MiniCPM-V2.6 [40]	0.75	0.80	190.77	191.53
InterVL2 [5]	0.80	0.86	103.77	99.12

Table 6. MLLM's overall performance in scoring

4.5. How Does the General MLLM Perform in Figure Skating Scoring

We can evaluate the scoring ability in the general Multi-modal large language model by generating additional text-type score reports. Specifically, we select some leading open-source MLLM as our evaluation object from the leader box on OpenCompass[8], which include Qwen2-VL[31], InternVL2[5], MiniCPM-V2.6[40]. We construct the fine-tuned dataset from the video and the text we generated, and then we use the swift toolkit [45] to supervise fine-tuning the Multi-modal Large Language Model above. Our fine-tuning method of choice is Lora [14] with the rank set to 8 and alpha to 16. We conduct training for 10 epochs with a batch size of 12. The results in Tab. 6 indicate that the general multi-modal large model struggles with scoring due to a lack of adequate datasets in its original pre-training. After fine-tuning, we observed that InterVL2 demonstrated the strongest scoring ability among the models tested. However, it still performed below our method. This suggests that it requires a comprehensive dataset and a more extensive training scale to improve its evaluation capabilities.

5. Conclusion

In conclusion, our study introduces a novel, multi-modal approach to the quantitative evaluation of figure skating performance. By aligning the "hit the beat" phenomenon with visual sub-movement analysis, our model provides a comprehensive framework for accurately assessing TES and PCS scores, while generating insightful text-based commentary that enriches traditional evaluation practices. Our results highlight the model's superior performance, demonstrating both state-of-the-art (SOTA) scoring accuracy and strong generalization capabilities across various figure skating scenarios. This advancement not only represents a step forward in multi-modal fusion for figure skating assessment but also sets a foundation for future work in enhancing the transparency and artistic appreciation of competitive evaluations.

References

[1] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action Quality Assessment with Temporal Parsing Transformer, 2022. 2

[2] XB Bruce, Yan Liu, Keith CC Chan, Qintai Yang, and Xiaoying Wang. Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer's progression. *Pattern Recognition*, 119:108095, 2021. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark, 2024. 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. 2, 8

[6] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. *arXiv preprint arXiv:2406.06087*, 2024. 2

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites, 2024. 2

[8] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 8

[9] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 42–52, 2024. 2

[10] Zexing Du, Di He, Xue Wang, and Qing Wang. Learning Semantics-Guided Representations for Scoring Figure Skating. *IEEE Transactions on Multimedia*, pages 1–11, 2023. 1, 6

[11] Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Visual-semantic alignment temporal parsing for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[13] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 8

[15] Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted Uncertainty Score Disentanglement Network for Action Quality Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8590–8597, 2023. 1, 2, 6

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[17] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *International Conference on Learning Representations*, 2021. 6

[18] Hui-Ying Li, Qing Lei, Hong-Bo Zhang, and Ji-Xiang Du. Skeleton Based Action Quality Assessment of Figure Skating Videos. In *2021 11th International Conference on Information Technology in Medicine and Education (ITME)*, pages 196–200, 2021. 1

[19] Jianwei Li, Jun Xue, Rui Cao, Xiaoxia Du, Siyu Mo, Kehao Ran, and Zeyan Zhang. FineRehab: A Multi-modality and Multi-task Dataset for Rehabilitation Analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3184–3193, 2024. 2

[20] Yuan-Ming Li, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Continual action assessment via task-consistent score-discriminative feature distribution modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[21] Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal Segmentation of Fine-gained Semantic Action: A Motion-Centered Figure Skating Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2163–2171, 2021. 1

[22] Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. RICA2: Rubric-Informed, Calibrated Assessment of Actions, 2024. 2

[23] Hitoshi Matsuyama, Nobuo Kawaguchi, and Brian Y Lim. Iris: Interpretable rubric-informed segmentation for action quality assessment. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 368–378, 2023. 2

[24] Lauren Okamoto and Paritosh Parmar. Hierarchical NeuroSymbolic Approach for Comprehensive and Explainable Action Quality Assessment, 2024. 2

[25] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action Assessment by Joint Relation Graphs. 2

[26] Paritosh Parmar and Brendan Tran Morris. Learning to Score Olympic Events. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 76–84, Honolulu, HI, USA, 2017. IEEE. 1

[27] Paritosh Parmar and Brendan Tran Morris. 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*, pages 304–313, 2019. 2

[28] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014. 2

[29] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-Aware Score Distribution Learning for Action Quality Assessment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9836–9845, 2020. 2

[30] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for Training and Evaluating Large Video Description Models. 2

[31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[32] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding, 2024. 2

[33] Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-Mixer: Long-Term Sport Audio-Visual Modeling with MLPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):2901–2909, 2023. 1, 6

[34] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert Scoring with Grade Decoupling for Long-term Action Assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3222–3231, 2022. 2, 6

[35] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to Score Figure Skating Sport Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590, 2020. 1, 6

[36] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2949–2958, 2022. 2, 6

[37] Jinglin Xu, Yongming Rao, Jie Zhou, and Jiwen Lu. Procedure-aware action quality assessment: Datasets and performance evaluation. *International Journal of Computer Vision*, pages 1–22, 2024. 2

[38] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14628–14637, 2024. 2

[39] Long Yao, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. A Contrastive Learning Network for Performance Metric and Assessment of Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3790–3802, 2023. 2

[40] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone, 2024. 8

[41] Bruce X.B. Yu, Yan Liu, Xiang Zhang, Gong Chen, and Keith C.C. Chan. EGCN: An Ensemble-based Learning Framework for Exploring Effective Skeleton-based Rehabilitation Exercise Assessment. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 3681–3687, 2022. 2

[42] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware Contrastive Regression for Action Quality Assessment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7899–7908, 2021. 2

[43] Ling-An Zeng and Wei-Shi Zheng. Multimodal Action Quality Assessment. *IEEE Transactions on Image Processing*, 33:1600–1613, 2024. 2

[44] 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. 6

[45] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. 8

[46] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. *arXiv preprint arXiv:2404.13999*, 2024. 2

[47] Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang. Magr: Manifold-aligned graph regularization for continual action quality assessment. *arXiv preprint arXiv:2403.04398*, 2024. 2