Supplementary Material for "Language-Guided Audio-Visual Learning for Long-Term Sports Assessment"

Huangbiao Xu[1,2], Xiao Ke[1,2]*, Huanqi Wu[1,2], Rui Xu[1,2], Yuezhou Li[1,2], Wenzhong Guo[1,2]
[1]Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350108, China
kex@fzu.edu.cn,{huangbiaoxu.chn, wuhuanqi135, xurui.ryan.chn, liyuezhou.cm}@gmail.com

1. Datasets

To fully validate the effectiveness of our method, we perform extensive experiments on four public long-term sports assessment benchmarks including FS1000 [17], Fis-V [19], Rhythmic Gymnastics (RG) [22] (three audio-visual datasets), and LOGO [23] (one visual-only dataset).

FS1000. The FS1000 dataset consists of 1,000 training videos and 247 validation videos, covering eight categories of figure skating competitions: men's/ladies'/pairs' short programs, men's/ladies'/pairs' free skating, and ice dance rhythm/free dances. It provides Total Element Score (TES) and Total Program Component Score (PCS), along with five additional scores: Skating Skills (SS), Transitions (TR), Performance (PE), Composition (CO), and Interpretation of Music (IN). Each video contains approximately 5,000 frames at 25 frames per second. Notably, FS1000 is the first figure skating sport assessment dataset to encourage audio-visual learning, enabling rule-consistent multimodal learning. Following [6, 17], we train separate models for each score type.

Fis-V. The Figure Skating Video (Fis-V) dataset consists of 500 videos of ladies' singles short program performances in figure skating. Each video is approximately 2.9 minutes long and recorded at 25 frames per second. Following the official split, 400 videos are allocated for training and 100 for testing. Each video is annotated with two scores: Total Element Score (TES) and Total Program Component Score (PCS), in accordance with competition regulations. Consistent with prior works [5, 6, 17–19, 21, 24], we train separate models to predict each score.

Rhythmic Gymnastics (RG). The RG dataset contains 1,000 videos of rhythmic gymnastics performances involving four apparatuses: ball, clubs, hoop, and ribbon. Each video is approximately 1.6 minutes long and recorded at 25 frames per second. The dataset is divided into 200 training videos and 50 evaluation videos for each action type. Consistent with prior studies [5, 18, 21, 22, 24], we train separate models for each action type.

LOGO. The LOGO dataset is a multi-person, long-term video dataset comprising 150 training samples and 50 testing samples. The videos are sourced from 26 artistic swimming events, each featuring 8 athletes and averaging 204.2 seconds in duration. LOGO provides formation labels to represent group dynamics among athletes and includes detailed annotations of action procedures. The LOGO provides only visual RGB maps, so we validate the effect of the domain-specific action knowledge introduced by our MAG² module on visual semantic learning on this dataset.

Methods	Spearman Correlation (↑)			Mean Square Error (↓)		
	TES	PCS	RG-Avg.	TES	PCS	RG-Avg.
MLP	0.892	0.874	0.824	70.86	7.44	4.80
Shared-Transformer	0.909	0.882	0.831	66.59	6.65	4.68
Dual-Transformer	**0.920**	0.888	0.840	**64.56**	6.79	4.53
Action Graph (Ours)	0.917	**0.892**	**0.849**	64.89	**6.39**	**4.47**

Table 1. Different ways of introducing action knowledge.

2. More Ablation Studies

In this section, we will further add some ablation studies to determine the experimental details. All ablation studies are performed on the FS1000 and RG if not specifically stated.

Different ways of introducing action knowledge. We construct text prompt sets based on action terms defined in the rules of sports events, leveraging language to introduce action knowledge and facilitate action understanding. To utilize this knowledge for accurate audio-visual learning, we design a multidimensional action graph guidance (MAG²) module. This module employs graph neural networks to propagate information across graph nodes and effectively integrates knowledge into visual and audio features through constructed edges. As shown in Tab. 1, our

*Corresponding author.

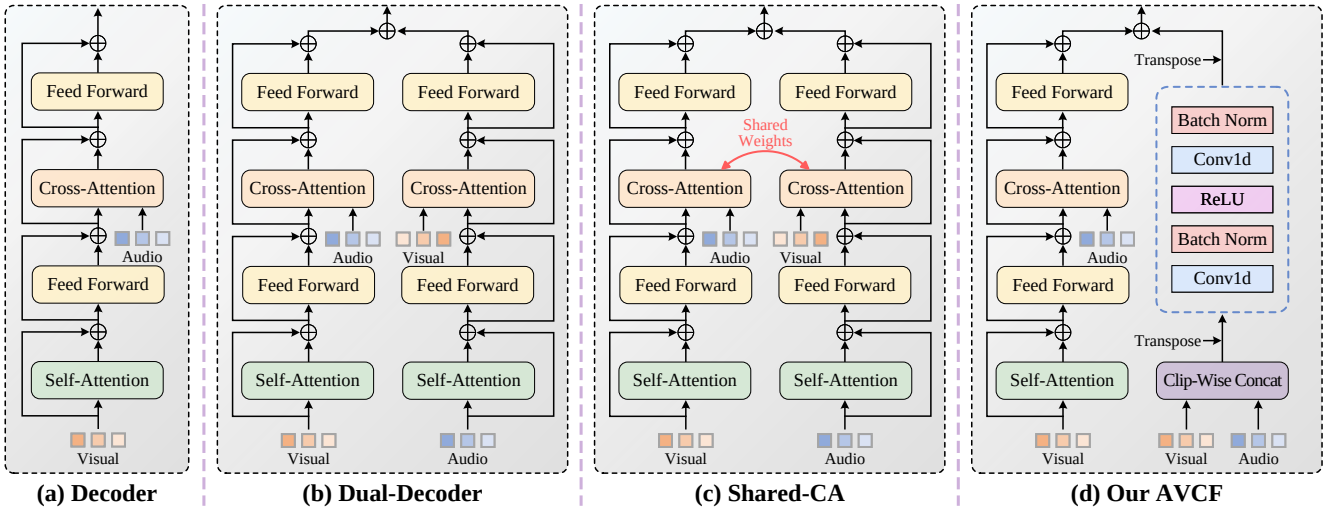


Figure 1. Framework comparison of cross-modal fusion modules: (a) Transformer decoder—visual features serve as "queries," while audio features act as "keys-values." (b) Dual Transformer decoders—one decoder uses visual features as "queries" and audio features as "keys-values," while the other uses audio features as "queries" and visual features as "keys-values." (c) Shared cross-attention—the cross-attention blocks in the two decoders from (b) share weights. (d) Our audio-visual cross-modal fusion (AVCF) module—a decoder captures global consistency between actions and music, while modeling the local matches through a convolutional block on a clip-by-clip basis.

#N	Spearman Correlation (↑)			Mean Square Error (↓)		
	TES	PCS	RG-Avg.	TES	PCS	RG-Avg.
3	0.878	0.875	0.819	69.86	6.93	4.85
4	0.892	0.871	0.828	68.04	6.70	4.80
5	0.903	0.867	0.821	66.87	6.63	4.71
6	**0.917**	0.892	**0.849**	**64.89**	6.39	**4.47**
7	0.898	0.888	0.836	65.47	**6.33**	4.69
8	0.887	**0.894**	0.830	66.21	6.41	4.64

Table 2. Different number of grade prompts N.

#	Methods	Spearman Correlation (↑)			Mean Square Error (↓)		
		TES	PCS	RG-Avg.	TES	PCS	RG-Avg.
(a)	Decoder	0.894	0.876	0.817	66.03	6.62	4.76
(b)	Dual-Decoder	0.882	0.866	0.813	70.20	7.11	5.04
(c)	Shared-CA	0.911	0.883	0.824	66.86	6.94	4.85
(d)	AVCF (Ours)	**0.917**	**0.892**	**0.849**	**64.89**	**6.39**	**4.47**

Table 3. Different audio-visual fusion modules.

action graph outperforms commonly used information aggregation methods, such as MLP and Shared-Transformer. Although the Dual-Transformer also demonstrates competitive performance, it incurs significantly higher computational costs (1.20M parameters and 0.59G FLOPs). Our MAG² efficiently processes complex relationships within graph structures to transfer action knowledge, offering significant advantages in computational efficiency and scalability. This also allows our MAG² to be seamlessly integrated as a plug-and-play module to enhance the performance of existing methods.

Different number of grade prompts N. Our dual-branch prompt-guided grading module (DPG) employs two sets of grade prompts to map visual and audio-visual features into N visual and $2N$ audio-visual grades. The effect of varying N is shown in Tab. 2. The results indicate that performance improves steadily as N increases from 3 to 6, particularly for the MSE metric, highlighting the importance of fine-grained grade modeling for accurate assessment. However, when N exceeds 6, performance begins to decline, likely due to the subtle distinctions among an excessive number of grade patterns, which may lead to confusion.

Different audio-visual fusion modules. To evaluate the consistency between actions and musical rhythm, we propose a novel audio-visual cross-modal fusion (AVCF) module for long-term sports assessment. This module emphasizes both global and clip-wise alignment between actions and music. As shown in Tab. 3, we compare AVCF with prevalent Transformer-based cross-modal fusion modules [3, 4, 9, 10, 13, 20]. To illustrate the differences between the modules, we provide a framework comparison in Fig. 1. The results demonstrate that our method outperforms existing approaches, which can be attributed to AVCF's ability to align each video clip within the global cross-modal fusion framework, adhering to the assessment criteria for long-term sports events. Additionally, for fine-grained clip-wise alignment, we leverage convolutional blocks to capture local details, achieving excellent performance while maintaining low model complexity.

#	Methods	Spearman Correlation (↑)								Mean Square Error (↓)							
		TES	PCS	SS	TR	PE	CO	IN	Avg.	TES	PCS	SS	TR	PE	CO	IN	Avg.
(1)	CoFInAl* [24]	0.835	0.830	0.838	0.836	0.814	0.829	0.819	0.829	81.65	16.05	0.56	0.63	0.71	0.41	0.54	14.36
(2)	CoFInAl* [24] + Texts	0.847	0.834	0.842	0.844	0.816	0.834	0.823	0.835$^{\uparrow 0.7\%}$	80.86	12.69	0.50	0.58	0.68	0.40	0.50	13.74$^{\downarrow 4.3\%}$
(3)	CoFInAl* [24] + MAG²	0.858	0.844	0.848	0.851	0.822	0.838	0.827	0.842$^{\uparrow 1.6\%}$	79.78	11.82	0.40	0.44	0.63	0.34	0.44	13.41$^{\downarrow 6.6\%}$
(4)	QTD* [5]	0.876	0.845	0.850	0.857	0.827	0.845	0.841	0.849	137.09	17.48	0.51	0.73	0.80	0.91	0.98	22.64
(5)	QTD* [5] + Texts	0.884	0.850	0.854	0.860	0.835	0.849	0.846	0.855$^{\uparrow 0.7\%}$	131.51	14.80	0.44	0.69	0.73	0.84	0.90	21.42$^{\downarrow 5.4\%}$
(6)	QTD* [5] + MAG²	0.889	0.858	0.861	0.864	0.840	0.857	0.853	0.861$^{\uparrow 1.4\%}$	119.74	13.89	0.41	0.60	0.66	0.83	0.88	19.57$^{\downarrow 14\%}$
(7)	PAMFN* [21]	0.897	0.885	0.856	0.866	0.855	0.867	0.845	0.868	104.89	10.05	0.39	0.52	0.78	0.40	0.56	16.80
(8)	PAMFN* [21] + Texts	0.899	0.887	0.859	0.867	0.859	0.866	0.850	0.871$^{\uparrow 0.3\%}$	104.04	9.69	0.37	0.50	0.69	0.38	0.51	16.60$^{\downarrow 1.2\%}$
(9)	PAMFN* [21] + MAG²	0.904	0.889	0.862	0.869	0.861	0.869	0.856	0.874$^{\uparrow 0.7\%}$	101.18	8.78	0.31	0.43	0.66	0.38	0.42	16.02$^{\downarrow 4.6\%}$
(10)	**MLAVL† (Ours)**	0.917	0.892	0.895	0.895	0.876	0.885	0.878	**0.892**$^{\uparrow 2.1\%}$	64.89	6.39	0.23	0.24	0.50	0.25	0.26	**10.39**$^{\downarrow 23\%}$
(11)	Texts→LV	0.908	0.887	0.895	0.889	0.873	0.881	0.873	0.887$^{\downarrow 0.7\%}$	68.10	6.55	0.23	0.25	0.50	0.25	0.27	10.88$^{\uparrow 4.7\%}$
(12)	Texts→LV+PE	0.909	0.876	0.892	0.888	0.870	0.883	0.867	0.884$^{\downarrow 0.9\%}$	69.78	7.03	0.23	0.25	0.51	0.26	0.29	11.19$^{\uparrow 7.7\%}$

Table 4. Effects of grade-related texts and applying our texts and MAG² to existing methods on the FS1000 dataset. LV: Learnable Vectors, PE: Positional Embeddings. The **bold** / underline indicate the best / second-best results. The red / blue is performance increase / decrease.

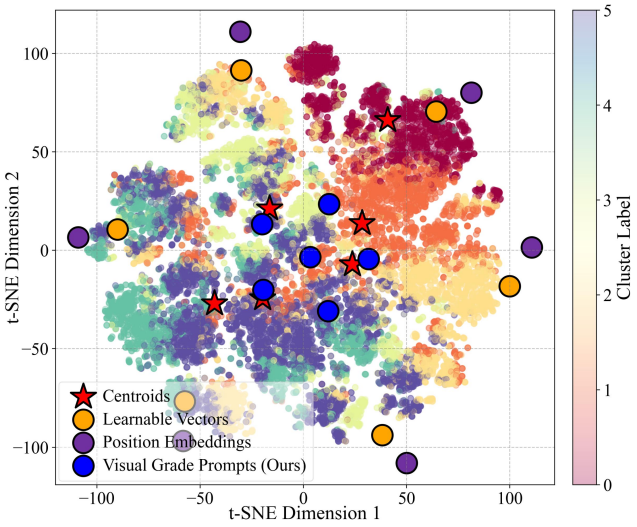


Figure 2. The t-SNE visualization of the initial stage of grade pattern learning on the FS1000.

Effects of applying our texts and MAG² to existing methods. To validate the effectiveness of our designed texts and MAG², we embed them in three 2024 SOTA works [5, 21, 24], with the results displayed in Tab. 4 (1)-(10). It can be seen that both our texts and MAG² significantly improve the performance. Even compared to the best performance of the enhanced existing methods (0.874 Avg. Sp. Corr. and 13.41 Avg. MSE), our methods significantly improve by 2.1% and 23%. This is a further fair comparison and demonstrates the strong effectiveness of our method.

Effects of grade-related texts. To validate the effect of introducing grade-related text prompts, we replace the text embeddings with common alternatives: learnable vectors [5, 18, 24] (LV in Tab. 4 (11)) and positional embeddings [5] (PE in (12)). The performance of all categories drops significantly, especially the MSE metric, demonstrating that our grade-related textual semantics effectively models the quality-aware space. To intuitively demonstrate the advantages of texts, we visualize the t-SNE for the initial stage of grade pattern learning in Fig. 2. Our text prompts provide precise class semantics, aligning closer to target class centers early in training.

Settings	Spearman Correlation (↑)			Mean Square Error (↓)		
	TES	PCS	RG-Avg.	TES	PCS	RG-Avg.
'a photo of' [14]	0.896	0.878	0.838	65.77	6.55	4.61
'human action of' [16]	0.912	**0.894**	0.845	65.01	**6.23**	**4.44**
'a video of' (Ours)	**0.917**	0.892	**0.849**	**64.89**	6.39	4.47
'a music of'	0.911	0.887	0.840	65.39	6.41	4.59
'a musical rhythm of'	0.913	0.886	0.836	65.48	6.53	4.54
'a music suitable for' (Ours)	**0.917**	**0.892**	**0.849**	**64.89**	**6.39**	**4.47**
Trainable Text Encoder	0.923	0.898	0.858	63.54	6.21	4.36
Trainable Prompt Learning	0.929	0.906	0.865	62.96	6.20	4.18

Table 5. Different text prompt settings.

Different text prompt settings. In this work, we aim to introduce action knowledge in a cost-effective manner. To achieve this, we manually design fixed text prompt templates and freeze the text encoder, thereby avoiding the high computational cost associated with training the backbone network and implementing prompt learning. Additionally, the choice of text template settings plays a crucial role in introducing textual semantics. To evaluate their impact, we conduct experiments on various text prompt configurations, with the results presented in Tab. 5. Notably, the prompt 'a video of' outperforms 'a photo of' in the visual template, likely because the term "video" emphasizes temporal information. Similarly, 'human action of' proves effective in introducing action-specific knowledge. Among the

Methods	Features	Spearman Correlation (↑)								Mean Square Error (↓)							
		TES	PCS	SS	TR	PE	CO	IN	Avg.	TES	PCS	SS	TR	PE	CO	IN	Avg.
M-BERT† (Late) [12]	TF [1]+AST [7]	0.79	0.75	0.80	0.81	0.80	0.80	0.76	0.79	131.28	15.28	0.44	0.43	0.67	0.47	0.55	21.30
MLP-Mixer† [17]	TF [1]+AST [7]	0.88	0.82	0.80	0.81	0.80	0.81	0.81	0.82	81.24	9.47	0.35	0.35	0.62	0.37	0.39	13.26
SGN† [6]	TF [1]+AST [7]	0.89	0.85	0.84	0.85	0.82	0.85	0.83	0.85	79.08	8.40	0.31	0.32	0.61	0.33	0.37	12.77
PAMFN†* [21]	C3D [15]+VGGish [8]+I3D [2]	0.88	0.85	0.85	0.84	0.85	0.85	0.83	0.85	108.43	11.27	0.42	0.59	0.73	0.51	0.49	17.49
	TF [1]+AST [7]+I3D [2]	0.90	**0.89**	0.86	0.87	0.86	0.87	0.85	0.87	104.89	10.05	0.39	0.52	0.78	0.40	0.56	16.80
MLAVL† (Ours)	C3D [15]+VGGish [8]+BERT [11]	0.90	0.85	0.84	0.85	0.85	0.85	0.84	0.86	71.18	10.29	0.37	0.28	0.68	0.43	0.48	11.96
	C3D [15]+VGGish [8]+CLIP [14]	**0.92**	0.87	0.85	0.86	0.87	0.85	0.85	0.87	67.51	7.66	0.30	0.29	**0.49**	0.32	0.34	10.99
	TF [1]+AST [7]+BERT [11]	0.91	**0.89**	0.89	0.89	**0.88**	0.88	0.87	0.89	67.73	6.52	0.24	0.27	0.53	0.27	0.34	10.84
	TF [1]+AST [7]+CLIP [14]	**0.92**	**0.89**	**0.90**	**0.90**	**0.88**	**0.89**	**0.88**	**0.90**	**64.89**	**6.39**	**0.23**	**0.24**	0.50	**0.25**	**0.26**	**10.39**

Table 6. Different modal backbones on the FS1000 dataset. The **bold** / underline / blue indicate the best / second-best / third-best results. * indicates our reimplementation based on the official code. † indicates using audio information. **TF stands for Timesformer.**

audio templates, 'a music suitable for' achieves the best performance in our design. While training the text encoder yields better results, it incurs a higher computational cost (25.20M parameters and 395.31G FLOPs). "Trainable Prompt Learning," which introduces learnable tokens such as "a [XXX] video of," offers additional performance improvements but at the expense of increased computational requirements. These results also demonstrate the potential of our approach. However, since this study prioritizes cost efficiency, we use fixed prompts and a frozen encoder.

Different modal backbones. Features extracted from pre-trained backbones are crucial for long-term sports assessment tasks with poor intra-class discrimination. Stronger features provided by more advanced backbones enhance assessment performance, as demonstrated in prior studies [5, 18, 21, 23, 24]. Multimodal learning methods, which involve multiple backbones, may be even more sensitive to backbone quality. To evaluate the impact of different backbones, we present results in Tab. 6. Compared to C3D [15]+VGGish [8], the more powerful Timesformer [1]+AST [7] combination extracts stronger audio-visual features, significantly improving the performance of PAMFN [21] and our MLAVL. Furthermore, our method outperforms state-of-the-art approaches even with the classical BERT [11] text encoder, highlighting its effectiveness in leveraging language to introduce action knowledge for guiding audio-visual learning. Performance is further enhanced when using the CLIP [14] text encoder, which incorporates richer cross-modal knowledge.

3. Additional Implementation Details

Compute resources. All experiments are conducted on an RTX 3090 GPU with PyTorch 2.4.1 and a 2.40GHz CPU. For instance, using a batch size of 64 and 500 epochs with visual, audio, and textual features extracted from pre-trained backbones, training on the RG dataset requires approximately three and a half hours.

Label normalization. Our DPG module is designed to evaluate visual and audio-visual performance by utilizing grade prompts to model distinct grade patterns. Each pattern corresponds to a specific quality level, with grade weights defined as $\mathbf{W}n^{\mathbf{v}} = \frac{n-1}{N-1}$ and $\mathbf{W}n^{\mathbf{v}\cdot\mathbf{a}} = \frac{n-1}{2N-1}$. Following prior works [5, 18, 21, 22, 24], we normalize the score label range to [0,1] using a constant ξ. Formally, for all real score labels $\{s_i\}_{i=1}^{P}$ in a dataset, the normalized labels s'_i are computed as s_i/ξ. The value of ξ is determined by the maximum score in the training set. In our experiments, ξ is set to 130/60/10/10/10/10/10 for FS1000's TES/PCS/SS/TR/PE/CO/IN, and to 45/40/25 for Fis-V(TES)/Fis-V(PCS)/RG, respectively.

To ensure fair comparisons with existing methods, predicted scores are multiplied by ξ to revert to the original score range when calculating the MSE metric. This approach is consistent with our reimplementation of prior works [5, 18, 21, 22, 24]. For the LOGO dataset, however, we adhere to the experimental setup of existing methods and evaluate the R-ℓ_2 metric, which is not constrained to a specific score range.

Epoch of training. We adopt a cosine annealing strategy to dynamically adjust the learning rate during training. Consistent with prior works [18, 21, 22, 24], we use dataset-specific epoch settings across various models to achieve better convergence. Specifically, for the FS1000 dataset, the epochs are set to 510/540/390/380/400/450/450 for TES/PCS/SS/TR/PE/CO/IN, respectively. For the Fis-V dataset, the epochs are set to 420/440 for TES/PCS, while for the RG dataset, the epochs are 620/340/450/710 for Ball/Clubs/Hoop/Ribbon, respectively. For the LOGO dataset, our method is integrated as a plug-in to existing approaches and follows the experimental setup reported in the original method.

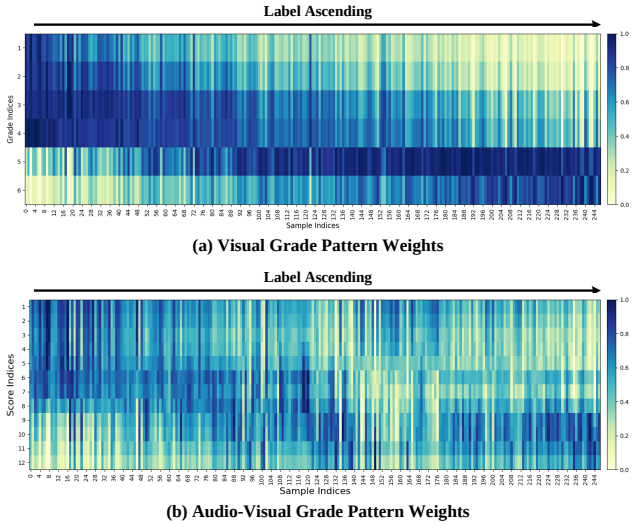
$\mathcal{L}_{TL}$ between projected text features. Pre-trained backbones for different modalities typically extract features with varying dimensions. To facilitate interaction between multimodal features, we project visual, audio, and textual features into a unified dimension d using token projection net-

Sports	Action Category					
Rhythmic Gymnastics	walking	running	leaping	jumping	skipping	ballet feet
	ballet hand	dance steps	bending	360° turn	split	one leg stand
	30 sec. without stopping	spirals	release	circles	mills	throw and catch
	bouncing 3 times	roll and recover 3 feet	roll and spin	pass through	swinging	figure 8
	jump through	hopping	sliding	swaying	turning	stretching
	twisting	balance				
Figure Skating	alternating backward crossovers	alternating forward crossovers	back spin	bunny hop	change of edge	slide chasse
	cross stroke	dance mode	drop Mohawk	drop three	Dutch Waltz	entry
	extended facing hold	dance steps	free sides	freestyle mode	gliding	half-flip
	half-Lutz	hockey stop	pumping	moving	one-foot snowplow	one-foot spin
	pivoting	power forward crossovers	power skating	rolling	scratch spin	slaloming
	snowplow stop	spiral	straight line holds	straight line spirals	stroking	swing
	swing roll	toe-loop jumping	two-foot spin	Waltz jump	Waltz hold	Waltz three
Artistic Swimming	ballet leg single	ballet leg alternate	ballet leg double	twist	spinning	twirling
	twist spin	spin up	combined spin	continuous spinning	boost action	cadence action
	crane action	ibis action	Eiffel Tower action	Catalina action	Catalac action	helicopter action
	flamingo action	flamingo bent knee	stingray action	Rio straight leg	manta ray action	knight action
	London action	swan action	swanita action	albatross action	goeland action	barracuda action
	blossom action	somersault back pike	barracuda bent knee	flying fish action	barracuda airborne split action	somersault back tuck
	kip action	seagull action	Kip bent knee	kip-swirl action	somersault front pike	elevator action
	somersub action	aurora action	subalina action	subilarc action	ballerina action	lagoon action
	Gaviata action	heron action	butterfly action	Neptunus action	Catalina reverse	side fishtail split
	Minerva action	porpoise action	front Ariana	walkover front	prawn action	water drop action
	cyclone action	Ipanema action	Saturn action			

Table 7. Domain-specific action texts employed in the three assessed sports scenes.

works. The four sets of textual features are represented as $\{f_m^{\text{t·v}}\}_{m=1}^{M}$, $\{f_m^{\text{t·a}}\}_{m=1}^{M}$, $\{f_n^{\text{t·v}}\}_{n=1}^{N}$, $\{f_n^{\text{t·v·a}}\}_{n=1}^{2N}$. To preserve the original textual semantics during the projection process, we apply a triplet loss $\mathcal{L}_{TL}$ to each set of textual features, ensuring that their discriminative properties are maintained. Specifically, $\mathcal{L}_{TL}^{\text{text}} = \mathcal{L}_{TL}(f_m^{\text{t·v}}) + \mathcal{L}_{TL}(f_m^{\text{t·a}}) + \mathcal{L}_{TL}(f_n^{\text{t·v}}) + \mathcal{L}_{TL}(f_n^{\text{t·v·a}})$. The balancing weights for these losses are equally used λ_1.

M action texts for different action scenes. We construct M multidimensional action texts using the rule files provided by the official websites of international sports associations. Specifically, M is set to 32, 42, and 63 for rhythmic gymnastics, figure skating, and artistic swimming, respectively. Tab. 7 provides a detailed breakdown of the '`[category]`' fields used in the text templates. In future work, the number of text prompt sets (M) can be expanded, and domain-specific action knowledge can be incorporated from additional perspectives.



(a) Visual Grade Pattern Weights

(b) Audio-Visual Grade Pattern Weights

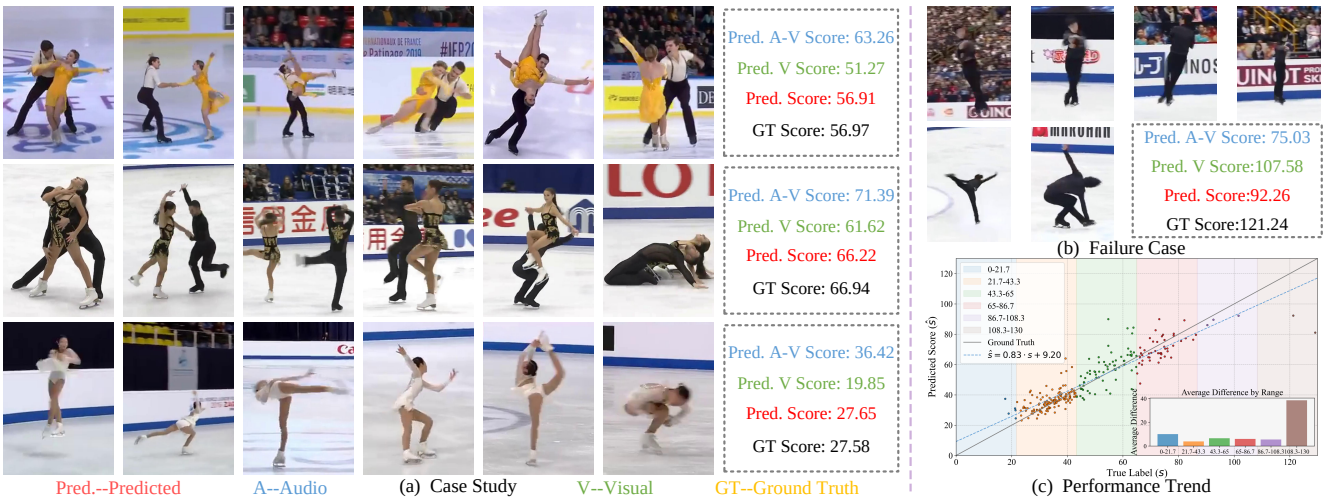
Figure 3. Visualization of grade weights on the FS1000 (TES).

4. More Visualizations

Here, we provide more visualizations to illustrate the contribution of our proposed designs to audio-visual learning, and sports assessment.

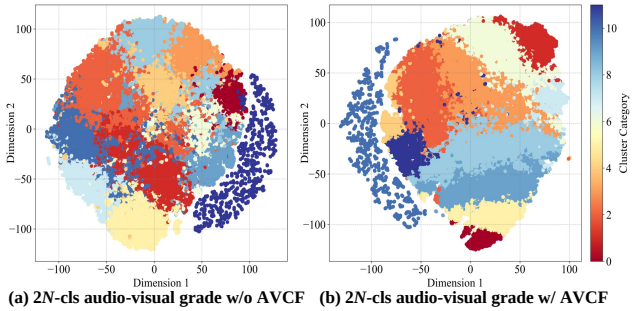
Visualization of grade pattern weights. The methods [5, 18, 24] of modeling grade patterns aim to understand action performance and aggregate video clips into patterns of corresponding quality. After normalizing the weights across grade patterns, videos with lower quality scores should receive higher weights for lower grades, whereas higher scores tend to have higher higher-grade weights. As illustrated in Fig. 3, the grade patterns developed by our method effectively capture the action performance associated with the respective quality levels, highlighting the effectiveness of our approach.

Case study. We show some samples containing inputs and assessed scores in Fig. 4 (a) to visualize the effectiveness of our approach for audio-visual learning and sports assessment. It can be seen that the single visual modality tends to ignore some bright performances and produces limited



(a) Case Study (b) Failure Case (c) Performance Trend

Figure 4. Some qualitative examples, including (a) case study, (b) failure case, and (c) performance trend.



(a) 2*N*-cls audio-visual grade w/o AVCF **(b) 2*N*-cls audio-visual grade w/ AVCF**

Figure 5. The t-SNE plots contrasting with and without our AVCF.

scores, while more accurate scores are achieved after integrating the coordination of the athlete's movements with the music. This demonstrates the importance of investigating assessment models that conform to the rules of real-life long-term sport.

Failure case and performance trend. To explore the limitations of our approach, we analyze the performance trends for different score ranges in FS1000 (TES), as shown in Fig. 4 (c). We observe that the score error is largest in the highest score range and show one of the failure samples in Fig. 4 (b). Our method underperforms in the highest score range, likely due to the limited training data in this range, failing to model accurate score mapping.

Visualization of fusion effects of our AVCF. Compared to existing Transformer-based cross-modal fusion modules, our approach integrates the global context-capturing capability of Transformers with the local detail-capturing ability of convolutional blocks. This design aligns with the requirements of long-term sports assessment, which emphasizes action-music consistency at both the overall and clip-specific levels. As illustrated in Fig. 5, our AVCF module enhances the fusion of visual and audio features more effectively than existing methods, resulting in more accurate and discriminative modeling of audio-visual grades.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824, 2021. 4

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 4

[3] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. 2

[4] Lu Chi, Guiyu Tian, Yadong Mu, and Qi Tian. Two-stream video classification with cross-modality attention. In *ICCVW*, pages 4511–4520, 2019. 2

[5] Xu Dong, Xinran Liu, Wanqing Li, Anthony Adeyemi-Ejeye, and Andrew Gilbert. Interpretable long-term action quality assessment. In *BMVC*, 2024. 1, 3, 4, 5

[6] Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE TMM*, 26:4987–4997, 2024. 1, 4

[7] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 4

[8] 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 *ICASSP*, pages 131–135, 2017. 4

[9] Jenhao Hsiao, Yikang Li, and Chiuman Ho. Language-guided multi-modal fusion for video action recognition. In *ICCVW*, pages 3158–3162, 2021. 2

[10] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022. 2

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, 2019. 4

[12] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *ICLR*, 2020. 4

[13] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *NeurIPS*, pages 11449–11461, 2021. 2

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3, 4

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 4

[16] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[17] 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. In *AAAI*, pages 2901–2909, 2023. 1, 4

[18] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *CVPR*, pages 3232–3241, 2022. 1, 3, 4, 5

[19] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE TCSVT*, 30(12):4578–4590, 2019. 1

[20] Yating Xu, Conghui Hu, and Gim Hee Lee. Rethink cross-modal fusion in weakly-supervised audio-visual video parsing. In *WACV*, pages 5603–5612, 2024. 2

[21] Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE TIP*, 33:1600–1613, 2024. 1, 3, 4

[22] 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 *ACM MM*, pages 2526–2534, 2020. 1, 4

[23] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, pages 2405–2414, 2023. 1, 4

[24] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *IJCAI*, pages 1771–1779, 2024. 1, 3, 4, 5