FSBench: A Figure Skating Benchmark for Advancing Artistic Sports Understanding

Supplementary Material

1. Dataset Contents

1.1. RGB Data

Considering that most video-based LLMs typically use 8 or 16 frames for frame extraction, this section separately presents the visualizations for extracting 8 frames and 16 frames, showcasing both the overall performance of the entire match and individual executions, as shown in Figure 1, 2, 3 and 4.

1.2. 3D Human Motion Data

FSAnno provides a structured format encompassing 3D human motion information, with a focus on Skinned Multi-Person Linear Model (SMPL)[4]. Each frame's data is organized as a dictionary, capturing a comprehensive range of motion and visual features, such as 2D and 3D joint positions, appearance features, 2D bounding boxes, and camera parameters. SMPL data includes parameters such as betas (10 shape coefficients representing individual body shape variations), body_pose (23 joints with 3×3 rotation matrices representing body posture), and global_orient (a 3×3 rotation matrix for global orientation). These parameters are stored in a high-dimensional vector format, with the pose field capturing them as a 229-dimensional vector and detailed further under the SMPL format to facilitate reconstructing the 3D human body shape and pose.

1.3. Skeleton Data

The skeleton data modality in FSAnno provides a structured representation of human motion using a tensor with the shape N×C×T×V×M, where N represents the number of samples, C denotes the three channels (x-coordinate, y-coordinate, and confidence score, with coordinates normalized between -1 and 1), T indicates the temporal dimension (up to 1500 frames, with zero-padding applied to shorter sequences for uniformity), V corresponds to 25 keypoints representing major joints in the human body, and M is set to 1, representing a single athlete per sample. The skeleton structure captures the spatiotemporal dynamics of motion through predefined keypoints, including landmarks such as the head, shoulders, elbows, wrists, hips, knees, and ankles, organized into a connected graph structure.

2. Dataset Quality Control

To ensure the accuracy and consistency of annotations in the figure skating dataset, we implemented a rigorous quality verification process. The dataset is sourced from publicly available figure skating competitions and includes three modalities: RGB video, 3D Human Motion, and skeleton data. The 3D Human Motion and skeleton data were extracted from the RGB video data using advanced motion capture and pose estimation algorithms. Annotation information for the dataset was entirely derived from official competition documents or visible cues in the videos and was completed and validated by a team of five experts with relevant domain expertise.

During the annotation phase, each expert meticulously annotated key information in the dataset based on official documents, ensuring high consistency between the annotations and the visible content in the videos. Upon completion, all annotations underwent a double-review process by the expert team to eliminate potential errors or ambiguities. For contentious or uncertain cases, the team prioritized in-depth discussions and consensus evaluations to ensure the quality and reliability of the annotations.

To validate the experts' mastery of the annotation process, a systematic training phase was conducted prior to the formal annotation phase. During training, each expert completed multiple rounds of practice tasks and received standardized guidance on annotation rules and review procedures. The large-scale annotation process officially began only after all experts met the expected standards.

Furthermore, ambiguous or uncertain samples within the data were carefully screened, and potentially misleading examples were excluded to minimize the risk of mislabeling. Through this rigorous process, we ensured the consistency and reliability of annotations across all three modalities—RGB video, 3D Human Motion, and skeleton data—providing high-quality data support for subsequent research and applications.

3. Evaluation Metrics

In addition to accuracy and AutoDQ mentioned in the paper, FSBench also considers the following two evaluation metrics [6] for evaluation tasks: **Mean Squared Error (MSE)** and **Spearman Correlation Coefficient** (ρ).

3.1. Mean Squared Error (MSE)

MSE is used to evaluate the absolute error between the predicted scores and the ground truth. It is defined as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i represents the ground truth score, $\hat{y}_i$ is the predicted score, and N is the total number of samples.

A lower MSE indicates that the predicted scores are closer to the ground truth, reflecting better performance of the model. However, MSE is sensitive to large deviations due to the squaring of errors.

3.2. Spearman Correlation Coefficient

Spearman correlation coefficient (ρ) evaluates the ranking consistency between the predicted scores and the ground truth scores. It is calculated as:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{2}$$

where d_i is the difference between the ranks of the i-th predicted score and the ground truth score, and N is the number of samples.

The value of ρ ranges from -1 to 1, where:

- $\rho = 1$: The predicted ranking is identical to the ground truth ranking.
- $\rho = -1$: The predicted ranking is completely opposite to the ground truth ranking.
- $\rho = 0$: No correlation between the predicted and ground truth rankings.

A higher ρ indicates better ranking consistency, which is particularly important for competitive scenarios where relative rankings matter more than absolute scores.

By combining MSE and ρ, we evaluate the assessment score's absolute accuracy and their consistency with the ground truth rankings. This dual evaluation ensures a robust assessment of the scoring model's performance.

4. More Experiment Details

We provide more detailed test results on FSBench. From the table 1, it is evident that apart from GPT-4 (75.8%) and GPT-3.5-turbo (61.2%), other mainstream open-source RGB-based video LLMs (such as onellm, Chatuniv, Video-chatgpt, and Video-LLAMA) demonstrate limited understanding of figure skating knowledge, as their accuracy scores on the FSBench prior knowledge test are all below 50%. Notably, Video-chatgpt (27.58%) and Video-LLAMA (26%) perform particularly poorly, highlighting their lack of expertise in this domain.

The task of segmenting an entire figure skating performance is extremely challenging for current large language models. Beyond the limited frame-level information that these models typically learn, there has been a longstanding lack of comprehensive, annotated datasets like FSAnno that specifically focus on figure skating. Such datasets are critical for enabling large models to develop a deeper understanding of figure skating knowledge. Without access to this level of detailed and domain-specific annotation, existing models struggle to effectively interpret and analyze the complex temporal and contextual dynamics of a full figure skating performance.

Table 1. Comparison of LLMs' accuracy (%) on the FSBench prior knowledge test.

Method	Acc
GPT4	75.8
onellm [1]	43.75
GPT3.5-turbo	61.2
Chatuniv [2]	28.42
Video-chatgpt [5]	27.58
Video-LLaMA [7]	26

The evaluation of an entire figure skating performance is also an extremely challenging task for current large language models. As shown in Figure 10, the models presented fail to demonstrate the ability to accurately assess figure skating competitions. For instance, the scores provided by Video-chatgpt often exceed reasonable ranges, indicating a lack of understanding of the scoring criteria. Similarly, LLaMA-vid tends to give nearly identical results for many performances, showing an inability to distinguish between varying levels of execution and artistic quality. These issues highlight the current limitations of LLMs in handling nuanced and domain-specific evaluation tasks like figure skating scoring.

Table 2. Comparison of LLMs' score on the FSBench entire performance comment test. We hire GPT-4 as the judge to score the generated comments on a scale of 1 to 10.

Method	Score
LLaMA-VID [3]	2.96
Chatuniv	1.36
Video-chatgpt	3.72
Video-LLaMA	2.3

For the task of generating commentary for an entire performance, we utilized GPT-4 as our benchmark judge and instructed it to provide scores within a range of 1 to 10. As shown in Table 2, the performance of the presented models is unsatisfactory. For example, LLaMA-VID and Video-LLAMA show limited variability in their scoring, while Chatuniv performs significantly worse with a low score of 1.36, indicating its inability to generate meaningful commentary or assessments for figure skating competitions.

As shown in Table 3, compared to the comment score of the entire performance, the scores of individual elements have decreased. The decline in performance for single-action commentary tasks can be attributed to the increased difficulty of these evaluations. Unlike full-program assessments, single-action analysis requires precise recognition of action categories and the ability to detect deduction points. However, the evaluated models lack these critical capabil-

Table 3. Comparison of LLMs' score on the FSBench individual element comment test. We hire GPT-4 as the judge to score the generated comments on a scale of 1 to 10.

Method	Score
LLaMA-VID	2.07
Chatuniv	0.95
Video-chatgpt	2.6
Video-LLaMA	1.61

ities, limiting their effectiveness in providing accurate and meaningful commentary for individual elements.

For the individual-element recognition task, considering that our categories are divided into two levels, we conducted tests for each level separately. Specifically, the two levels are: spin, jump, and sequence, as well as numerous subcategories under these three categories, with a total of 20 subcategories combined.

Table 4. Comparison of LLMs' accuracy (%) on the FSBench individual element recognition test (3 Classes).

Method	All	Spin	Jump	Sequence
LLaMA-VID	31.14	97.92	0.44	0
Chatuniv	39.04	38.9	52.19	3.57
Video-chatgpt	33.11	77.08	17.11	1.19
Video-LLaMA	31.58	1	0	0
Motion-GPT	31.36	33.33	40.35	3.57

Table 5. Comparison of LLMs' accuracy (%) on the FSBench individual element recognition test (20 Classes).

Method	Acc
LLaMA-VID	11.4
Chatuniv	12.7
Video-chatgpt	9.87
Video-LLaMA	10.09
Motion-GPT	8.48

As shown in Tables 4 and 5, the results highlight the challenges faced by current LLMs in individual element recognition tasks. In the 3-class recognition test (Table 4), the models demonstrate reasonable accuracy for the "Spin" category. However, performance on the "Jump" and "Sequence" categories is notably poor, with most models achieving near-zero accuracy for "Sequence." In the more fine-grained 20-class recognition test (Tables 5), the accuracy of all models declines significantly, with scores ranging from 9.87% (Video-chatgpt) to 12.7% (Chatuniv). These results suggest that while current LLMs may perform relatively well on broader categories, they struggle to accurately recognize and classify diverse subcategories, underscoring the limitations of their understanding in domain-specific tasks like figure skating.

Table 6. Comparison of LLMs' accuracy (%) on the FSBench individual element assessment test.

Method	Acc
LLaMA-VID	71.27
Chatuniv	77.85
Video-chatgpt	60.09
Video-LLaMA	61.87

The classification tasks in Table 6 evaluate the models' ability to predict the GOE (Grade of Execution) scores in figure skating as positive, negative, or zero. The results indicate that although these models show lower accuracy in identifying complex movements, they perform better in classifying the positive or negative value of GOE (e.g., Chatuniv achieves 77.85%). This may be attributed to the fact that GOE prediction relies more on assessing the overall execution quality and identifying major errors rather than precise movement recognition.

Different models perform significantly differently across various tasks. GPT4 demonstrates the best performance in the Prior Knowledge Test with an accuracy of 75.8%, while Chatuniv achieves the highest score in the Element Assessment Test (77.85%). Other models, such as Video-chatgpt and Video-LLAMA, also perform reasonably well in this test, with scores of 60.09% and 61.87%, respectively. However, it is important to note that the Element Assessment Test only presents results for simpler tasks. For more challenging tasks, such as predicting GOE scores, the models' performance aligns with their capability in predicting overall performance scores. Currently, these large models struggle to handle such complex tasks.

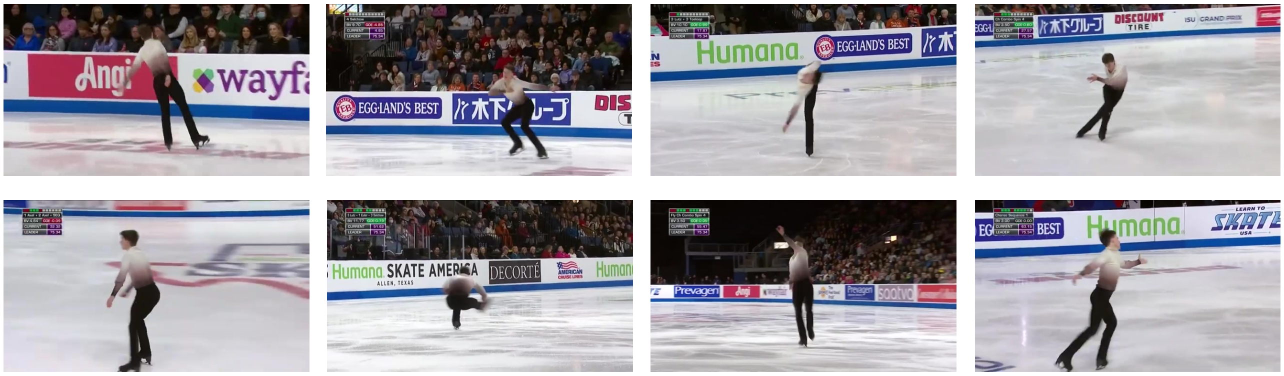

Figure 1. 8 frames extracted from the entire performance of an athlete.

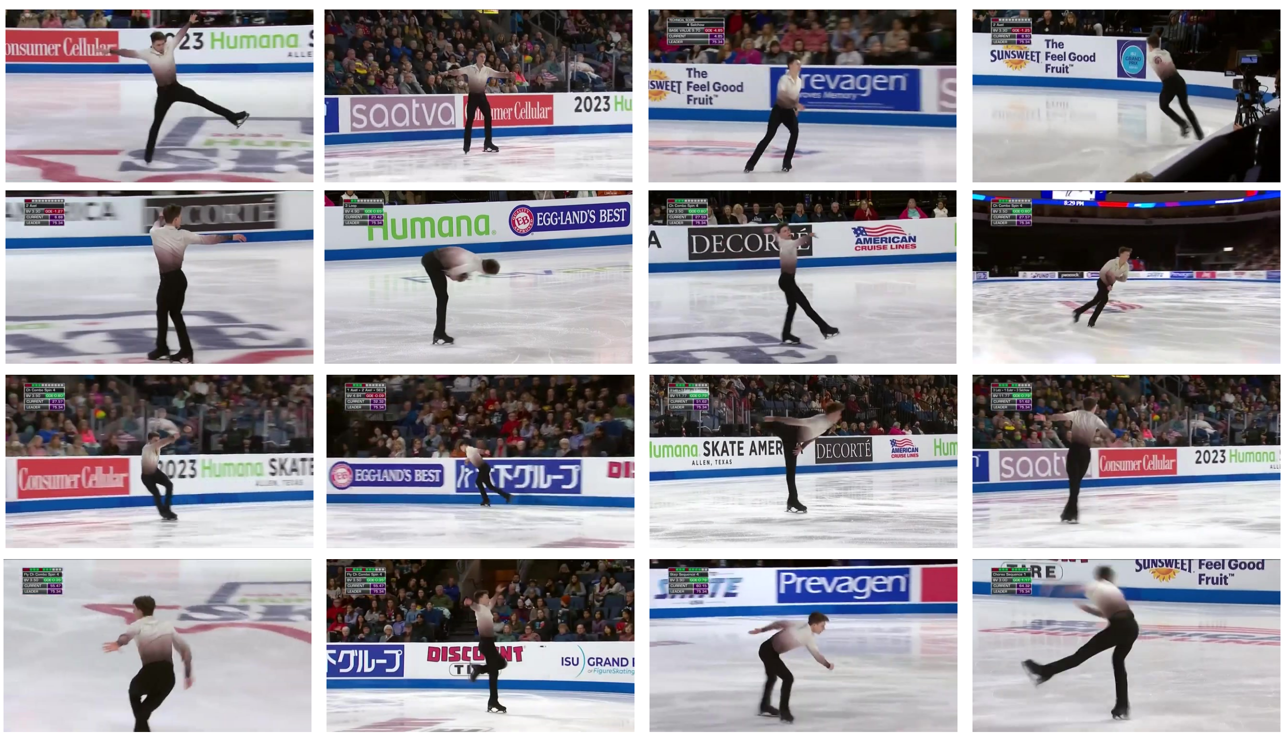

Figure 2. 16 frames extracted from the entire performance of an athlete.

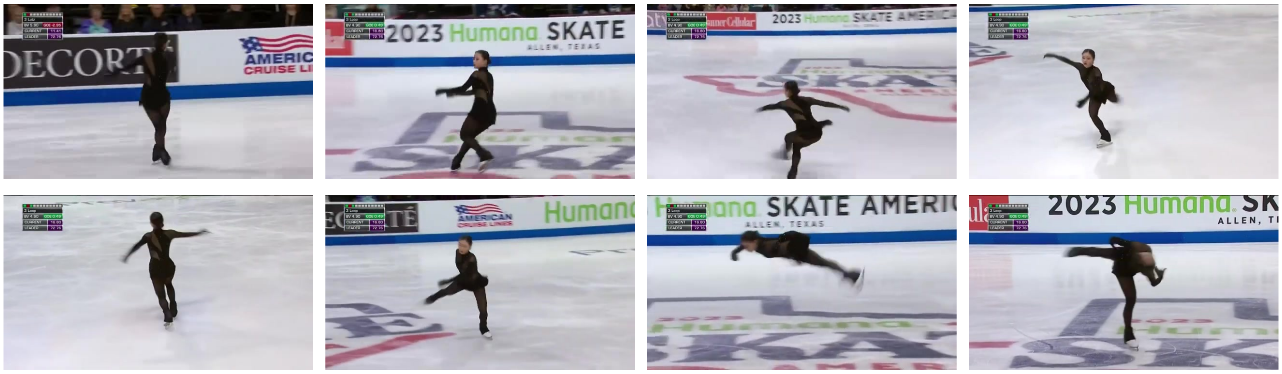

Figure 3. 8 frames extracted from a single execution of an athlete.

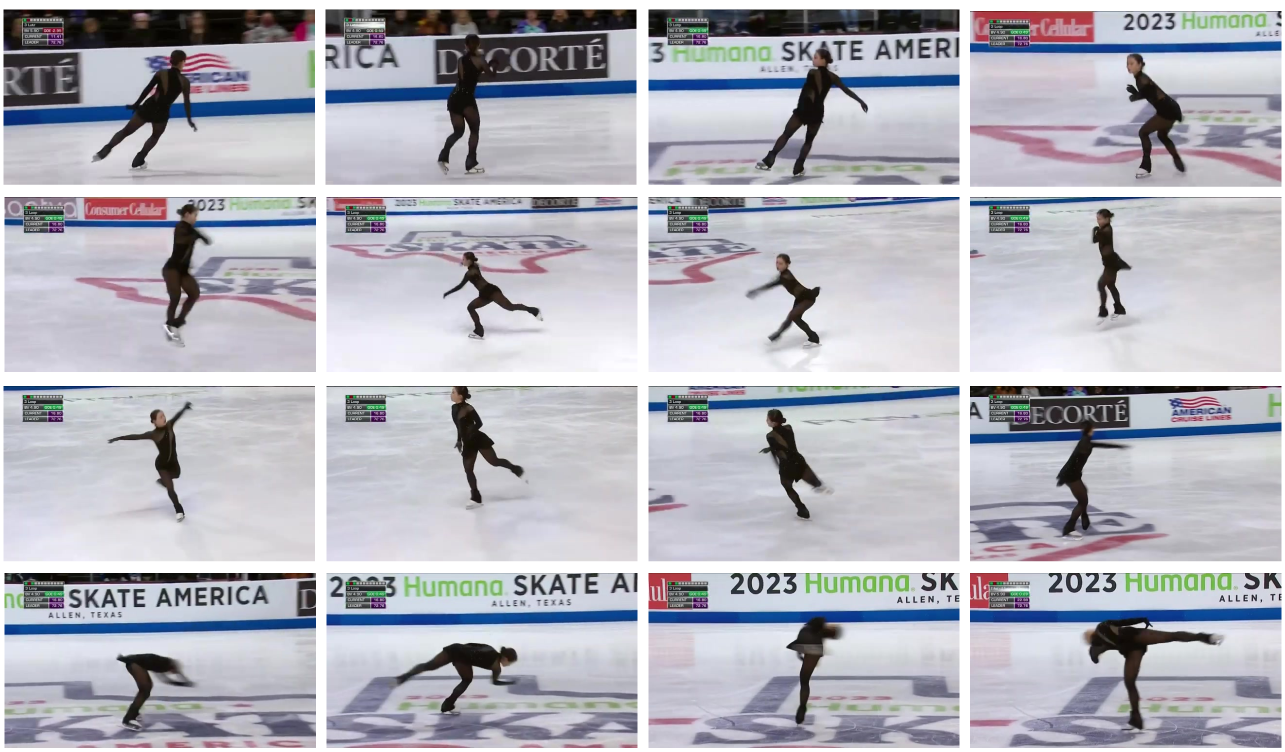

Figure 4. 16 frames extracted from a single execution of an athlete.

Event Information

1. *In which city was the World Figure Skating Championships 1931 held?*
A) Sofia **B) Berlin** C) New York City D) Jeonju

2. *Who won the bronze medal in the Women's singles category at the World Figure Skating Championships held in 2003?*
A) Hilde Holovsky B) Elene Gedevanishvili C) Oskar Uhlig **D) Fumie Suguri**

3. *Which year did the World Figure Skating Championships take place in Stockholm?*
A) 1964 B) 1897 **C) 1905** D) 2019

4. *What medal did Debi Thomas win in the World Figure Skating Championships 1986 competition?*
A) Gold B) Bronze C) Silver

5. *In which city was the European Figure Skating Championships 2014 held?*
A) Beijing B) Milan C) Edmonton **D) Budapest**

Figure 5. FSBench Prior Knowledge Testing Task Presentation: Event Information.

Rules

1. *Which element is performed on one foot with the free leg extended at hip level or higher?*
A) Spiral B) Spin C) Step sequence D) Jump

2. *How is the scoring range for a Grade of Execution (GOE) in figure skating?*
A) 0 to 10 B) 1 to 6 **C) −5 to +5** D) −3 to +3

3. *What determines a spin's base value when marked with a 'V'?*
A) Number of positions B) Speed of the spin **C) Fulfilling specific** requirements D) Number of revolutions

4. *Which is not a factor in the program components score (PCS)?*
A) Costume design B) Choreography C) Interpretation of the music D) Timing of the elements

5. *What is not allowed in men's figure skating costumes?*
A) Tights B) Trousers C) Shorts D) Decorative elements

Figure 6. FSBench Prior Knowledge Testing Task Presentation: Rules.

Individual Elements Comment

*You are a **professional figure skating commentator** tasked with providing concise and professional commentary based on the specified movements. Your commentary should clearly identify **the type of movement**, such as jumps, spins, or footwork, and **highlight its characteristics**, especially noting the challenge of high-difficulty elements like triple or quadruple jumps. **Analyze the quality of execution**, including technical aspects like takeoff, rotation, and landing, as well as details such as posture adjustments, edge control, and fluidity. Additionally, relate the movement to the overall performance, commenting on its alignment with the music and **how it enhances the program's flow or expressiveness**. The language should be vivid and concise, balancing professionalism with audience engagement to ensure clarity and appeal.*

Figure 7. The evaluation prompt case for the "Individual Elements Comment" task.

Entire Performance Assessment

knowledge = "The Technical Element Score (TES) evaluates the difficulty and execution of technical elements. Each element has a base value, modified by the Grade of Execution (GOE), scored from -5 to +5, which adjusts the base value by ±10% (Singles and Pairs) or ±16% (Ice Dance). The highest and lowest GOE scores are discarded, and the average of the remaining scores is added to the base value. Deductions are applied for excess elements, edge errors ("!" or "e"), under-rotations ("<" or "<<"), and downgraded combinations/sequences (+COMBO or +SEQ). Jumps in the second half of a program earn a 10% bonus to their base value. The Program Component Score (PCS) assesses the artistic and skating quality of the performance, scored on a 0.25 to 10.00 scale and averaged using the same trimmed mean method. PCS consists of three subcomponents: Composition (CO), which evaluates the program's design, transitions, spatial use, and alignment with music; Presentation (PR), which assesses expression, energy, synchronization, and musical timing; and Skating Skills (SK), which measures blade control, flow, balance, and power. PCS scores are multiplied by event-specific factors, and deductions are made for violations such as music, costume, or time issues. These combined scores provide a comprehensive evaluation of technical precision and artistic excellence in figure skating."

question = "Score the performance by evaluating TES based on the Base Value (BV) and Grade of Execution (GOE) of each element without a fixed range, and assign scores between 0.25 and 10 for CO, PR, and SK."

format = "Provide the scores in the following FORMAT, with each item's name followed by a colon and the corresponding score: TES: [score], CO: [score], PR: [score], SK: [score]"

Figure 8. The evaluation prompt case for the "Entire Performance Assessment" task.

Entire Performance Comment

You are a **professional figure skating commentator** tasked with providing comprehensive and engaging commentary on a figure skating competition. Your analysis should **balance technical precision and artistic insight** to help the audience appreciate the performance and understand the scoring criteria. Evaluate **technical elements** such as jumps, identifying their types (e.g., Axel, Lutz) and difficulty (e.g., triple, quad), while assessing the cleanliness of takeoff, stability during rotation, and the elegance of the landing. Comment on spins by noting their speed, stability, transitions between positions, and any innovative combinations. Analyze step sequences for their complexity, smoothness, and alignment with the music, highlighting difficult edge transitions and multidirectional steps. Beyond technical analysis, provide insights into the **program's artistry** by evaluating skating skills, such as speed, edge control, and fluidity; transitions between elements, noting their creativity and natural flow; and the **overall performance**, including **emotional expression**, **use of gestures, and connection with the music**. Highlight how the skater interprets the music's rhythm and mood through their movements. Use vivid and concise language to illustrate both the technical execution and artistic impact, ensuring clarity for viewers while enhancing their enjoyment of the performance.

Figure 9. The evaluation prompt case for the "Entire Performance Comment" task.

Entire Performance Assessment

knowledge = "The Technical Element Score (TES) evaluates the difficulty and execution of technical elements. Each element has a base value, modified by the Grade of Execution (GOE), scored from -5 to +5, which adjusts the base value by ±10% (Singles and Pairs) or ±16% (Ice Dance). The highest and lowest GOE scores are discarded, and the average of the remaining scores is added to the base value. Deductions are applied for excess elements, edge errors ("!" or "e"), under-rotations ("<" or "<<"), and downgraded combinations/sequences (+COMBO or +SEQ). Jumps in the second half of a program earn a 10% bonus to their base value. The Program Component Score (PCS) assesses the artistic and skating quality of the performance, scored on a 0.25 to 10.00 scale and averaged using the same trimmed mean method. PCS consists of three subcomponents: Composition (CO), which evaluates the program's design, transitions, spatial use, and alignment with music; Presentation (PR), which assesses expression, energy, synchronization, and musical timing; and Skating Skills (SK), which measures blade control, flow, balance, and power. PCS scores are multiplied by event-specific factors, and deductions are made for violations such as music, costume, or time issues. These combined scores provide a comprehensive evaluation of technical precision and artistic excellence in figure skating."

question = "Score the performance by evaluating TES based on the Base Value (BV) and Grade of Execution (GOE) of each element without a fixed range, and assign scores between 0.25 and 10 for CO, PR, and SK."

format = "Provide the scores in the following FORMAT, with each item's name followed by a colon and the corresponding score: TES: [score], CO: [score], PR: [score], SK: [score]"

LLaMA-VID

"pred": "TES: 90, CO: 9, PR: 9, SK: 9"

Chatuniv

"pred": "The given information does not provide any specific details about the skater's performance, so it is not possible to evaluate their performance based on the given information."

Video-ChatGPT

"pred": "TES: 90, CO: 90, PR: 90, SK: 90."

Video-LLaMA

"pred": "First, we see a young man skating on an ice rink. He is wearing a black and white shirt and is performing a figure skating routine. Next, we see a man skating on the ice rink, wearing a black and white shirt. He is performing a figure skating routine and is wearing a black and white shirt. Then, we see a man skating on the ice rink, wearing a black and white shirt. He is performing a figure skating routine and is wearing a black and white shirt. Finally, we see a man skating on the ice rink, wearing a black and white shirt. He is performing a figure skating routine and is wearing a black and white shirt. Throughout the video, we see the man performing various figure skating routines, including jumps and spins."

Figure 10. The "Entire Performance Assessment" task evaluation case.

Entire Performance Comment

You are a professional figure skating commentator tasked with providing comprehensive and engaging commentary on a figure skating competition. Your analysis should balance technical precision and artistic insight to help the audience appreciate the performance and understand the scoring criteria. Evaluate technical elements such as jumps, identifying their types (e.g., Axel, Lutz) and difficulty (e.g., triple, quad), while assessing the cleanliness of takeoff, stability during rotation, and the elegance of the landing. Comment on spins by noting their speed, stability, transitions between positions, and any innovative combinations. Analyze step sequences for their complexity, smoothness, and alignment with the music, highlighting difficult edge transitions and multidirectional steps. Beyond technical analysis, provide insights into the program's artistry by evaluating skating skills, such as speed, edge control, and fluidity; transitions between elements, noting their creativity and natural flow; and the overall performance, including emotional expression, use of gestures, and connection with the music. Highlight how the skater interprets the music's rhythm and mood through their movements. Use vivid and concise language to illustrate both the technical execution and artistic impact, ensuring clarity for viewers while enhancing their enjoyment of the performance.

LLaMA-VID

"The skater in the video is a professional figure skater, and he is performing a routine on the ice. He starts by executing a series of jumps, showcasing his agility and precision. He then moves on to perform a spin, demonstrating his control and balance. The skater then executes a step sequence, displaying his technical skills and artistic expression. Throughout the routine, the skater maintains a high level of control and fluidity, seamlessly transitioning between elements. His movements are in perfect sync with the music, and he demonstrates a deep connection with the audience. Overall, the skater's performance is a testament to his skill and dedication to the sport, and he is a true master of figure skating."

Video-ChatGPT

As a professional figure skating commentator, I would analyze the performance of the skater in terms of technical execution, artistic interpretation, and overall performance. I would start by noting the skater's jumps, identifying their types and difficulty, and assessing the cleanliness of their takeoff, stability during rotation, and the elegance of their landing. I would also evaluate the skater's spins, noting their speed, stability, transitions between positions, and any innovative combinations. Additionally, I would analyze the skater's step sequences for their complexity, smoothness, and alignment with the music, highlighting difficult edge transitions and multidirectional steps.\n\nBeyond technical analysis, I would provide insights into the skater's artistry by evaluating their speed, edge control, and fluidity. I would also observe the skater's transitions between elements, noting their creativity and natural flow. Furthermore, I would highlight the skater's overall performance, including emotional expression, use of gestures, and connection with the music. By doing so, I would help the audience appreciate the skater's performance and understand the scoring criteria, ultimately enhancing their enjoyment of the competition."}

Video-LLaMA

"In this video, we see a young man skating on an ice rink. He is wearing a black and white outfit and is performing various jumps and spins. The video is captured in slow motion, allowing us to see the skater's movements in great detail.\n\nThe skater starts by performing a series of jumps, including a triple jump and a quadruple jump. He lands smoothly and gracefully, showing off his technical skills. Next, he performs a series of spins, including a backspin and a frontspin. He rotates quickly and smoothly, demonstrating his control and precision.\n\nThroughout the video, we see the skater performing various combinations of jumps and spins, showcasing his artistic skills. He moves with fluidity and grace, and his movements are in perfect harmony with the music. The video also captures the skater's facial expressions and body language, adding to the overall artistic impact of the performance.\n\nOverall, this video provides a detailed and engaging look at the technical and artistic aspects of figure skating. The skater's movements are precise and graceful, and his ability to interpret the music through his movements is evident throughout the video."

Figure 11. The "Entire Performance Comment" task evaluation case.

Individual Element Comment

You are a professional figure skating commentator tasked with providing concise and professional commentary based on the specified movements. Your commentary should clearly identify the type of movement, such as jumps, spins, or footwork, and highlight its characteristics, especially noting the challenge of high-difficulty elements like triple or quadruple jumps. Analyze the quality of execution, including technical aspects like takeoff, rotation, and landing, as well as details such as posture adjustments, edge control, and fluidity. Additionally, relate the movement to the overall performance, commenting on its alignment with the music and how it enhances the program's flow or expressiveness. The language should be vivid and concise, balancing professionalism with audience engagement to ensure clarity and appeal.

Video-LLaMA

"First, we see a young woman figure skater performing a routine on an ice rink. She is wearing a black and white dress and has a white skate on her foot. The camera captures her movements in slow motion, allowing us to see the intricate details of her performance. Next, we see a young man figure skater performing a routine on the same ice rink. He is wearing a black and white suit and has a white skate on his foot. The camera captures his movements in slow motion as well, giving us a clear view of his technique. Then, we see a young woman figure skater performing a routine on the ice rink. She is wearing a black and white dress and has a white skate on her foot. The camera captures her movements in slow motion, allowing us to see the intricate details of her performance. Finally, we see a young man figure skater performing a routine on the ice rink. He is wearing a black and white suit and has a white skate on his foot. The camera captures his movements in slow motion, giving us a clear view of his technique. Throughout the video, we see various movements, including jumps, spins, and footwork. The skaters' posture, edge control, and fluidity are all highlighted, making their movements look effortless and precise."

Chatuniv

"In the first image, the female skater is performing a jump, showcasing her athleticism and control as she soars through the air. Her landing is crucial, and she executes it with precision, demonstrating her mastery of the basic skating techniques. In the second image, the skater is performing a spin, showcasing her grace and control as she twirls on the ice. Her movements are fluid and well-coordinated, highlighting her skill in executing complex skating maneuvers. Overall, both skaters display impressive athleticism and artistry, captivating the audience with their performances."

Video-ChatGPT

"In the video, a woman is seen performing a series of jumps and spins on the ice. She starts by jumping and then rotating her body while spinning on the ice. She then lands back on the ice and continues to perform more jumps and spins. The quality of her execution is excellent, with precise takeoffs, rotations, and landings. She also demonstrates good posture adjustments and edge control, which are essential for maintaining balance and fluidity in her performance. Overall, her performance is impressive and demonstrates her skill and dedication to the sport."

LLaMA-VID

"The skater in the video is performing a beautiful and graceful routine on the ice. She starts by executing a series of graceful footwork moves, showcasing her precision and control. As the routine progresses, she executes a series of spins, demonstrating her ability to maintain balance and control while moving at high speeds. The skater then performs a series of jumps, displaying her strength and agility as she soars through the air. Throughout the routine, the skater maintains a fluid and seamless movement, demonstrating her mastery of the sport. Overall, the skater's performance is a testament to her skill and dedication to the sport of figure skating."

Figure 12. The "Individual Element Comment" task evaluation case.

References

[1] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 2

[2] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 2

[3] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 2

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1

[5] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[6] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2909, 2023. 1

[7] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2