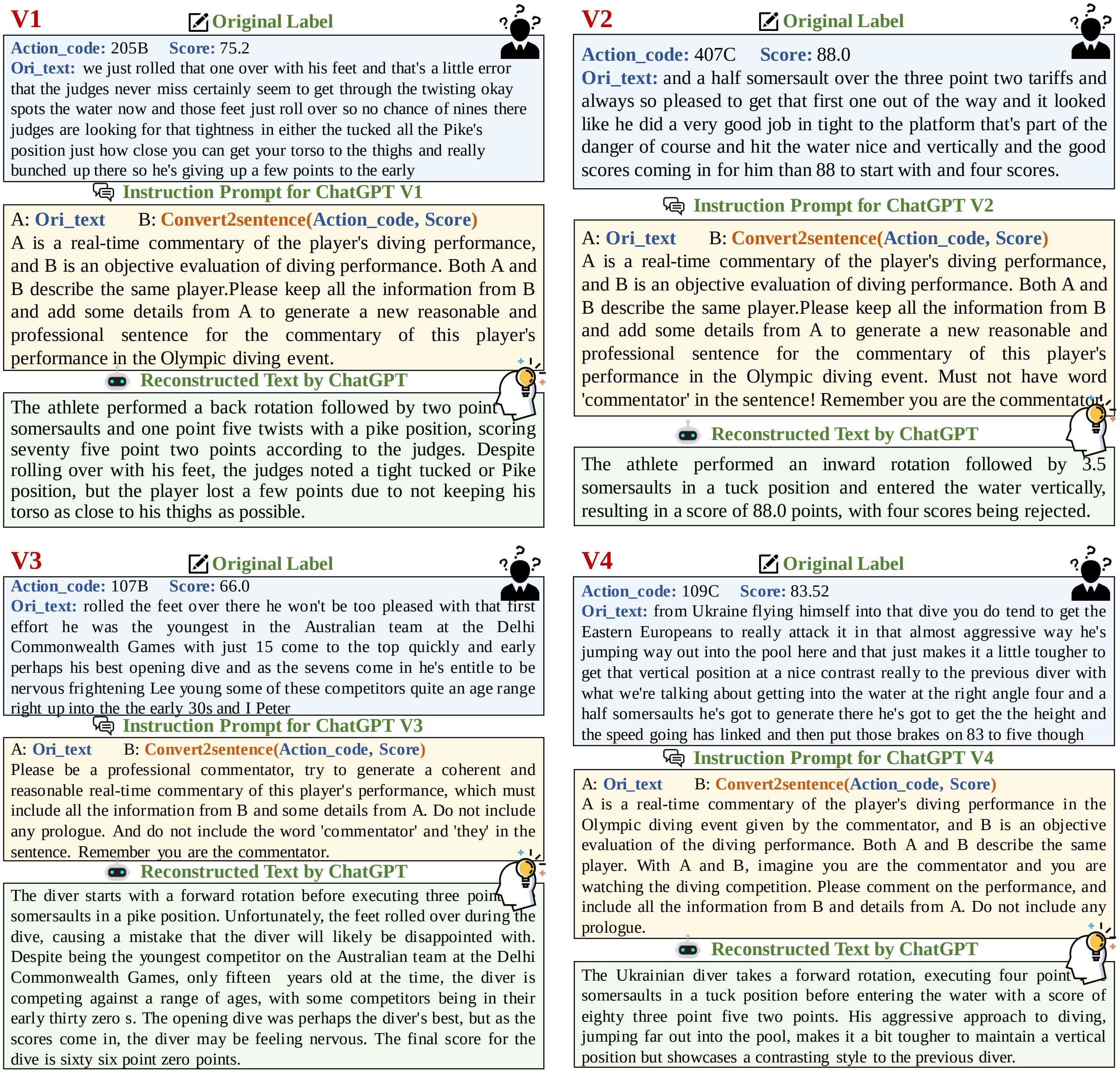
Narrative Action Evaluation with Prompt-Guided Multimodal Interaction

Supplementary Material



V1

Original Label

Action_code: 205B **Score:** 75.2

Ori_text: we just rolled that one over with his feet and that's a little error that the judges never miss certainly seem to get through the twisting okay spots the water now and those feet just roll over so no chance of nines there judges are looking for that tightness in either the tucked all the Pike's position just how close you can get your torso to the thighs and really bunched up there so he's giving up a few points to the early

Instruction Prompt for ChatGPT V1

A: **Ori_text** B: **Convert2sentence(Action_code, Score)**

A is a real-time commentary of the player's diving performance, and B is an objective evaluation of diving performance. Both A and B describe the same player.Please keep all the information from B and add some details from A to generate a new reasonable and professional sentence for the commentary of this player's performance in the Olympic diving event.

Reconstructed Text by ChatGPT

The athlete performed a back rotation followed by two point somersaults and one point five twists with a pike position, scoring seventy five point two points according to the judges. Despite rolling over with his feet, the judges noted a tight tucked or Pike position, but the player lost a few points due to not keeping his torso as close to his thighs as possible.

V2

Original Label

Action_code: 407C **Score:** 88.0

Ori_text: and a half somersault over the three point two tariffs and always so pleased to get that first one out of the way and it looked like he did a very good job in tight to the platform that's part of the danger of course and hit the water nice and vertically and the good scores coming in for him than 88 to start with and four scores.

Instruction Prompt for ChatGPT V2

A: **Ori_text** B: **Convert2sentence(Action_code, Score)**

A is a real-time commentary of the player's diving performance, and B is an objective evaluation of diving performance. Both A and B describe the same player.Please keep all the information from B and add some details from A to generate a new reasonable and professional sentence for the commentary of this player's performance in the Olympic diving event. Must not have word 'commentator' in the sentence! Remember you are the commentator!

Reconstructed Text by ChatGPT

The athlete performed an inward rotation followed by 3.5 somersaults in a tuck position and entered the water vertically, resulting in a score of 88.0 points, with four scores being rejected.

V3

Original Label

Action_code: 107B **Score:** 66.0

Ori_text: rolled the feet over there he won't be too pleased with that first effort he was the youngest in the Australian team at the Delhi Commonwealth Games with just 15 come to the top quickly and early perhaps his best opening dive and as the sevens come in he's entitle to be nervous frightening Lee young some of these competitors quite an age range right up into the the early 30s and I Peter

Instruction Prompt for ChatGPT V3

A: **Ori_text** B: **Convert2sentence(Action_code, Score)**

Please be a professional commentator, try to generate a coherent and reasonable real-time commentary of this player's performance, which must include all the information from B and some details from A. Do not include any prologue. And do not include the word 'commentator' and 'they' in the sentence. Remember you are the commentator.

Reconstructed Text by ChatGPT

The diver starts with a forward rotation before executing three poin somersaults in a pike position. Unfortunately, the feet rolled over during the dive, causing a mistake that the diver will likely be disappointed with. Despite being the youngest competitor on the Australian team at the Delhi Commonwealth Games, only fifteen years old at the time, the diver is competing against a range of ages, with some competitors being in their early thirty zero s. The opening dive was perhaps the diver's best, but as the scores come in, the diver may be feeling nervous. The final score for the dive is sixty six point zero points.

V4

Original Label

Action_code: 109C **Score:** 83.52

Ori_text: from Ukraine flying himself into that dive you do tend to get the Eastern Europeans to really attack it in that almost aggressive way he's jumping way out into the pool here and that just makes it a little tougher to get that vertical position at a nice contrast really to the previous diver with what we're talking about getting into the water at the right angle four and a half somersaults he's got to generate there he's got to get the the height and the speed going has linked and then put those brakes on 83 to five though

Instruction Prompt for ChatGPT V4

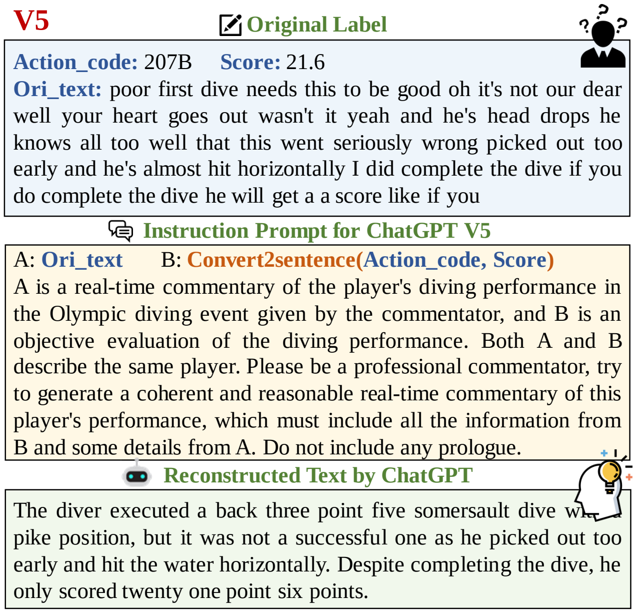
A: **Ori_text** B: **Convert2sentence(Action_code, Score)**

A is a real-time commentary of the player's diving performance in the Olympic diving event given by the commentator, and B is an objective evaluation of the diving performance. Both A and B describe the same player. With A and B, imagine you are the commentator and you are watching the diving competition. Please comment on the performance, and include all the information from B and details from A. Do not include any prologue.

Reconstructed Text by ChatGPT

The Ukrainian diver takes a forward rotation, executing four point somersaults in a tuck position before entering the water with a score of eighty three point five two points. His aggressive approach to diving, jumping far out into the pool, makes it a bit tougher to maintain a vertical position but showcases a contrasting style to the previous diver.

Figure 1. **Prompt (a).** The five versions of prompts we use to guide ChatGPT in generating different forms of text. And comparisons of the text before and after re-annotation.


V5
Original Label
Instruction Prompt for ChatGPT V5
Reconstructed Text by ChatGPT

Action_code: 207B **Score:** 21.6

Ori_text: poor first dive needs this to be good oh it's not our dear well your heart goes out wasn't it yeah and he's head drops he knows all too well that this went seriously wrong picked out too early and he's almost hit horizontally I did complete the dive if you do complete the dive he will get a a score like if you

A: **Ori_text** B: **Convert2sentence(Action_code, Score)**

A is a real-time commentary of the player's diving performance in the Olympic diving event given by the commentator, and B is an objective evaluation of the diving performance. Both A and B describe the same player. Please be a professional commentator, try to generate a coherent and reasonable real-time commentary of this player's performance, which must include all the information from B and some details from A. Do not include any prologue.

The diver executed a back three point five somersault dive with a pike position, but it was not a successful one as he picked out too early and hit the water horizontally. Despite completing the dive, he only scored twenty one point six points.

Figure 2. **Prompt (b).** The five versions of prompts we use to guide ChatGPT in generating different forms of text. And comparisons of the text before and after re-annotation.

1. Dataset

1.1. Five Versions of Prompt

In our main paper, we mention that we design five versions of prompts for both datasets so that ChatGPT can generate five different evaluative texts for each video. In Figure 1 and Figure 2, we show in detail the five prompt forms we use and more comparisons of captions before and after our re-annotation.

1.2. Action Code

In diving competition, the dives are referred to by a schematic system of three- or four-digit numbers. The letter to indicate the position is appended to the end of the number. The first digit of the number indicates the dive group. For groups 1 to 4, the number consists of three digits and a letter of the alphabet. The third digit represents the number of half-somersaults. The second digit is either 0 or 1, with 0 representing a normal somersault, and 1 signifying a ”flying” variation of the basic movement (i.e. the first half somersault is performed in the straight position, and then the pike or tuck shape is assumed).

2. Qualitative Results

In this section, we provide more qualitative results of our methods. Figure 3, Figure 4, and Figure 5 demonstrates that our method can generate sentences which maintain both narrative flexibility and evaluation rigor.

3. Implementation Details

3.1. Context-Aware Prompt Learning.

We set the number of classes, n_emb in K-class to 100. And we use CLIP[2] text encoder to get the prompt embeddings. The Context-Aware Transformer has 8 cross-attention layers, each containing 8 heads. And we sum the prompt embeddings before and after refining through a residual structure, where the hyper-parameter γ_1 is initialized with 10^{-3}.

3.2. Score-Guided Tokens Learning.

We use video swin-transformer-base as our video encoder in all of the experiments. The video swin-transformer takes 32 frames sampled from the video as the inputs. The Score-Aware Transformer has the same structure as the Context-Aware Transformer. And there is also a residual structure, where the hyper-parameter γ_2 is initialized with 10^{-3}.

3.3. Multimodal-Aware Text Generator.

A transformer encoder is used as the text decoder. Moreover, we use a three-layer multi-layer perceptron (MLP) with a ReLU activation function for the regressor.

3.4. Experimental Setting.

Following [1, 3–5], in the training stage of both the narrative action assessment and action quality assessment, we split videos into 75 percent for training and 25 percent for evaluation. We implemented and trained our method with the Pytorch toolbox and run on a Linux machine with Nvidia GeForce RTX 3090.

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. In *ECCV*, pages 422–438, 2022. 2

[2] 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, 2021. 2

[3] 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 *CVPR*, pages 9839–9848, 2020. 2

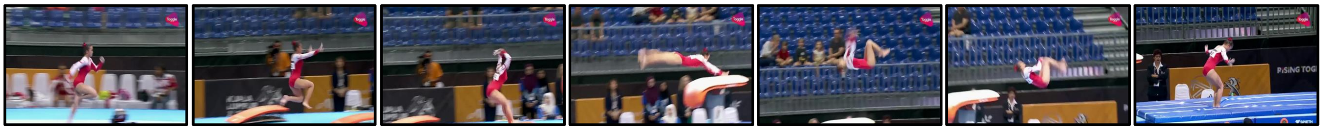
[4] Dezhong Xu, Heng Fu, Lifang Wu, Meng Jian, Dong Wang, and Xu Liu. Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention. *Access*, 8:65689–65698, 2020.

[5] 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 *CVPR*, pages 2949–2958, 2022. 2

Prediction	Ground-Truth
the athlete executes a flawless **back rotation** before launching into a series of **three point five somersaults** with a **pike position**. the dive is truly remarkable, and the entry into the water is **flawless**, earning a score of **ninety one point eight points**. a truly **remarkable performance** by the athlete.	The athlete executed a flawless **back rotation** before transitioning into a stunning **three point five somersaults** with a **pike position**. Their well-timed entry into the water was a testament to their **impeccable technique and athleticism**. The judges took notice and awarded a tremendous score of **ninety one point eight points**, putting the athlete in contention for a top spot on the leaderboard.

Prediction	Ground-Truth
the athlete executed a flawless **vault**, showcasing his exceptional **skills and precision**. the judges were impressed, awarding a score of **fifteen point one** for his outstanding performance.	With a difficult **vault** to execute, this athlete soared through the air with **great skill and precision**. Their flawless execution was well-deserving of the impressive score of **fifteen point zero**, which reflects their impressive diving performance.

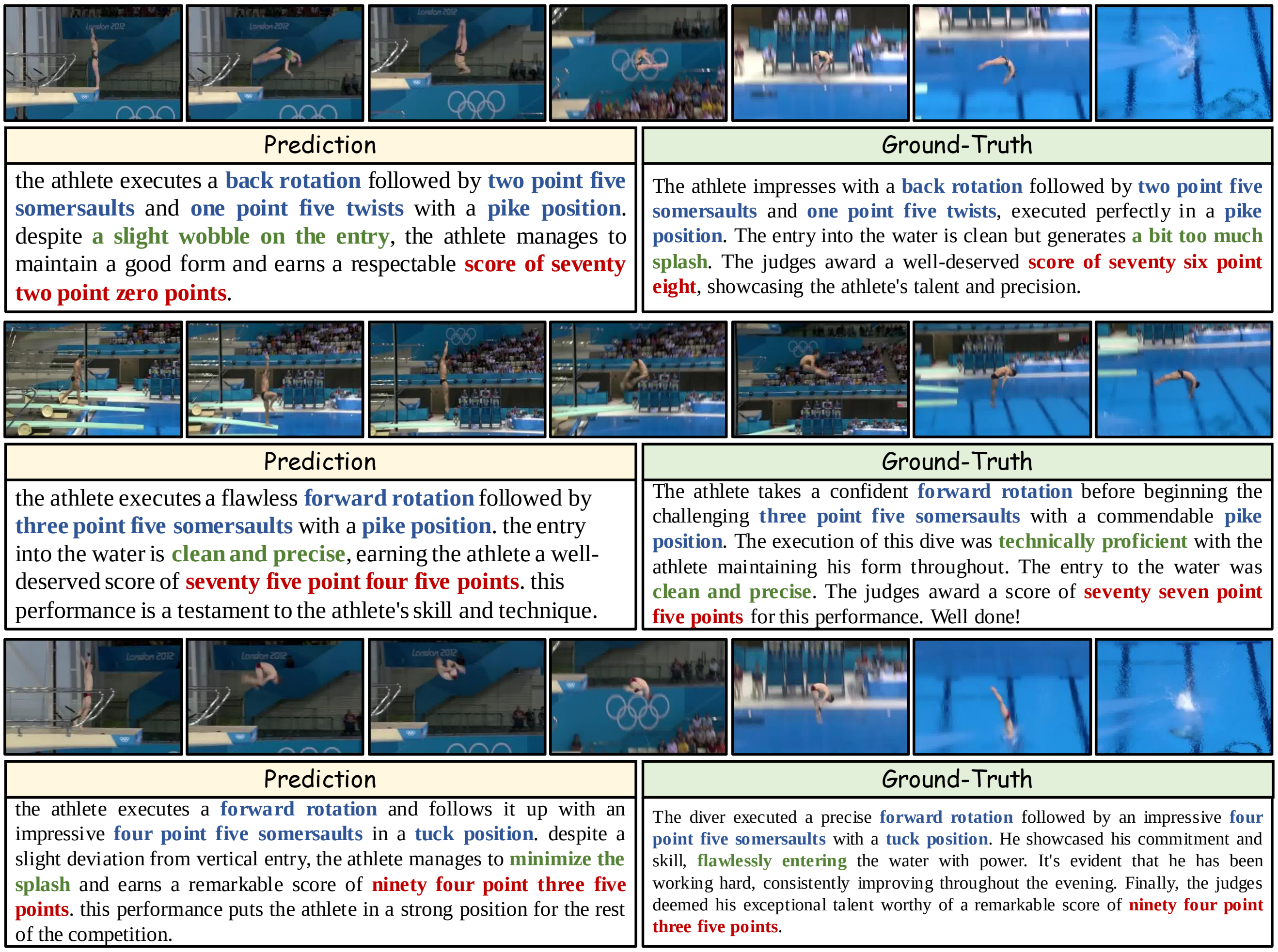
Figure 3. **Qualitative Results (a).** Our model can generate detailed narrations to describe and evaluate the actions comprehensively. Notably, the model can analyze the quality of actions by pointing out the details of the execution.



Prediction	Ground-Truth
the athlete executed a **handspring forward on**, followed by a **tucked salto forward off**, displaying **impressive technique and control**. the judges were impressed with the athlete's technique and awarded them a score of **twelve point three three** for her outstanding performance.	With an impressive **handspring forward on**, the athlete demonstrated **great precision and control** in her **tucked salto forward off**. The judges were clearly impressed, awarding the dive with a score of **twelve point five**. It's safe to say that this was an exceptional performance from the athlete.

Prediction	Ground-Truth
the athlete executes a **back rotation** followed by **three point five somersaults** in a **tuck position**. despite a slight splash on the entry, the athlete manages to maintain a **good form and control** throughout the dive. the judges award a score of **sixty four point three five points** for this dive.	The diver executes a phenomenal **back rotation** before taking off for **three point five somersaults** with a **tuck position**. The complexity of this dive requires tremendous skill and control, and the athlete showcases that perfectly. As she enters the water, the judges are sure to take note of the **impressive form and technique** on display. She will be thrilled with her score of **sixty four point three five points**.

Figure 4. **Qualitative Results (b).** Our model can generate detailed narrations to describe and evaluate the actions comprehensively. Notably, the model can analyze the quality of actions by pointing out the details of the execution.



Prediction	Ground-Truth
the athlete executes a **back rotation** followed by **two point five somersaults** and **one point five twists** with a **pike position**. despite **a slight wobble on the entry**, the athlete manages to maintain a good form and earns a respectable **score of seventy two point zero points**.	The athlete impresses with a **back rotation** followed by **two point five somersaults** and **one point five twists**, executed perfectly in a **pike position**. The entry into the water is clean but generates **a bit too much splash**. The judges award a well-deserved **score of seventy six point eight**, showcasing the athlete's talent and precision.
the athlete executes a flawless **forward rotation** followed by **three point five somersaults** with a **pike position**. the entry into the water is **clean and precise**, earning the athlete a well-deserved score of **seventy five point four five points**. this performance is a testament to the athlete's skill and technique.	The athlete takes a confident **forward rotation** before beginning the challenging **three point five somersaults** with a commendable **pike position**. The execution of this dive was **technically proficient** with the athlete maintaining his form throughout. The entry to the water was **clean and precise**. The judges award a score of **seventy seven point five points** for this performance. Well done!
the athlete executes a **forward rotation** and follows it up with an impressive **four point five somersaults** in a **tuck position**. despite a slight deviation from vertical entry, the athlete manages to **minimize the splash** and earns a remarkable score of **ninety four point three five points**. this performance puts the athlete in a strong position for the rest of the competition.	The diver executed a precise **forward rotation** followed by an impressive **four point five somersaults** with a **tuck position**. He showcased his commitment and skill, **flawlessly entering** the water with power. It's evident that he has been working hard, consistently improving throughout the evening. Finally, the judges deemed his exceptional talent worthy of a remarkable score of **ninety four point three five points**.

Figure 5. **Qualitative Results (c).** Our model can generate detailed narrations to describe and evaluate the actions comprehensively. Notably, the model can analyze the quality of actions by pointing out the details of the execution.