Narrative Action Evaluation with Prompt-Guided Multimodal Interaction

Shiyi Zhang[1,*], Sule Bai[1,*], Guangyi Chen[2], Lei Chen[3], Jiwen Lu[3], Junle Wang[4], Yansong Tang[2,†]
[1] Shenzhen International Graduate School, Tsinghua University
[2] Carnegie Mellon University, Pittsburgh PA, USA
[3] Department of Automation, Tsinghua University [4] Tencent
{sy-zhang23@mails.,bsl23@mails.,tang.yansong@sz.}tsinghua.edu.cn

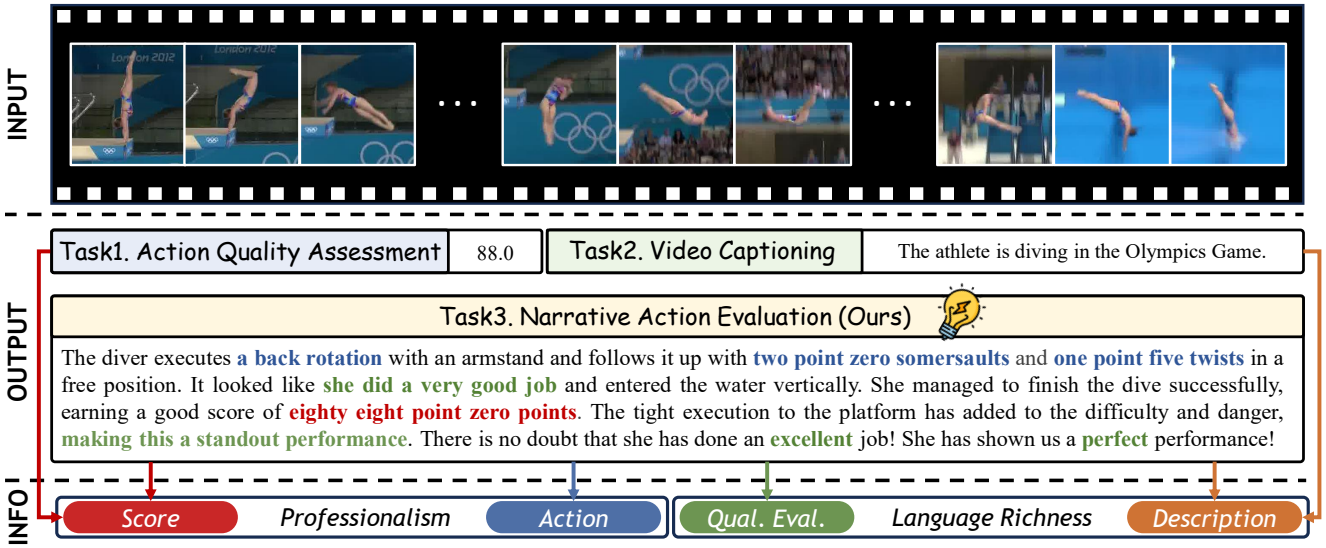


Figure 1. A comparison of our proposed narrative action evaluation (NAE) task with action quality assessment (AQA) and video captioning. The three lines in the figure represent the input video, the outputs of the three tasks, and the information contained in each task. In comparison to AQA, NAE provides rich language *descriptions*. When compared to Video Captioning, NAE includes much more evaluation information such as ***scores***, ***actions***, and ***qualitative evaluations***, which is often rigorous and granular. In general, NAE aims to strike a balance between the professionalism of assessment information and the richness of language. This duality is both the characteristic and challenge of NAE.

Abstract

In this paper, we investigate a new problem called narrative action evaluation (NAE). NAE aims to generate professional commentary that evaluates the execution of an action. Unlike traditional tasks such as score-based action quality assessment and video captioning involving superficial sentences, NAE focuses on creating detailed narratives in natural language. These narratives provide intricate descriptions of actions along with objective evaluations. NAE is a more challenging task because it requires both narrative flexibility and evaluation rigor. One existing possible solution is to use multi-task learning, where narrative language and evaluative information are predicted separately. However, this approach results in reduced performance for individual tasks because of variations between tasks and differences in modality between language information and evaluation information. To address this, we propose a prompt-guided multimodal interaction framework. This framework utilizes a pair of transformers to facilitate the interaction between different modalities of information. It also uses prompts to transform the score regression task into a video-text matching task, thus enabling task interactivity. To support further research in this field, we re-annotate the MTL-AQA and FineGym datasets with high-quality and comprehensive action narration. Additionally, we establish benchmarks for NAE. Extensive experiment results prove that our method outperforms separate learning methods and naive multi-task learning methods. Data and code are released at here.

1. Introduction

Recent years have witnessed substantial advancements in action description methods, including video captioning[13–15, 17, 38, 39, 44] that describes content (Task2 in Figure

*Equal contribution
†Corresponding author

1) and quality assessment[4, 23, 23, 24, 33, 35, 37, 40, 43] that evaluates quality (Task1 in Figure 1). However, most of the works only provide a depiction of actions from a singular perspective. How to balance information from multiple dimensions and provide a more comprehensive, rich, and professional evaluation remains blank. As shown in Figure 1, we introduce the Narrative Action Evaluation (NAE) task which aims to utilize narrative language to holistically evaluate actions from multiple perspectives, such as action depictions, professional objective evaluations, and qualitative evaluations. This approach merges the language richness of narrative description with the professionalism of expert analysis, thereby providing a multifaceted and thorough evaluation. The NAE presents numerous practical applications in the real world. For example, these narrative assessments could serve as an AI commentator providing professional descriptions of sports events, or as a fitness coach offering real-time feedback in natural language for physical actions.

As shown in Figure 1, the task of Narrative Action Evaluation is notably complex as it necessitates a delicate equilibrium between narrative flexibility and evaluation precision. Striking this balance is a substantial challenge as these two objectives sometimes conflict with each other. To be concrete, Figure 1 shows that some evaluation information (such as scores and actions) that require very high accuracy often only occupies a small proportion of the output sentence. However, language models often pursue diversity in the generation, which contradicts the rigor of evaluation information. To resolve this contradiction, models should be guided to focus on the evaluation information during text generation. An intuitive solution is using multi-task learning, which generates the text and predicts evaluation information at the same time. Similarly, [23] utilized a multi-task learning paradigm in which tasks are parallel to each other. Through the relatively independent training process of three tasks, the backbone shared by the tasks was refined. However, experiments have shown that such a paradigm improves every single task very little and even weakens the performance in some cases. The reason for this phenomenon is that such a multi-task learning paradigm, which only trains multiple tasks in parallel, may ultimately lead to a lack of interaction between features that focus on different tasks or even different modalities. Finally, it may confuse the model when balancing multiple tasks or even lose sight of one another.

In order to solve this problem, we propose a new framework, Prompt-Guided Multimodal Interaction, to encourage interactions between different modalities and tasks to aid the joint learning of description and evaluation. Specifically, we perform the first multimodal interaction by augmenting the learnable prompts which contain score information with video features through Context-Aware Prompt Learning. Then, in Score-Guided Tokens Learning, we formulate the score prediction as a video-text matching task, letting the video features perceive the score information of the language modality, and perform the second multimodal interaction under the guidance of the prompt. This process establishes the interaction between the score prediction task and the text generation task. Afterward, we combine the obtained multimodal embeddings with a learnable template, which contains the predicted score, action information, and the learnable prompt, as the input to the text decoder. Finally, we use the Tri-Token Attention Mask to guide the decoder to focus on professional evaluation information and video information in the input, ultimately generating the narrative evaluations.

Additionally, we find that the existing datasets are insufficient for supporting research in NAE. For instance, text labels in MTL-AQA[23], transcribed directly from video speeches, are too colloquial, incoherent, and noisy to be effectively used as texts for NAE. To propel further research in NAE, we re-annotate the MTL-AQA and FineGym[29] datasets as shown in Figure 2, ensuring high-quality and comprehensive action narration. We will make our code and data publicly available to support further progress on the NAE task. Moreover, we establish benchmarks using these datasets, demonstrating that our framework significantly outperforms the approaches based on previous state of the arts.

The contributions of this paper can be summarized as: (1) We propose a new task, Narrative Action Evaluation, which aims to generate professional commentary to evaluate action execution. And we re-annotate MTL-AQA and FineGym datasets for this task. (2) To tackle NAE, we propose a new framework, Prompt-Guided Multimodal Interaction, which uses prompts to integrate information from different modalities, realizing better mutual promotion between multiple tasks. (3) Experiments show that our framework surpasses the baselines based on previous state-of-the-art approaches.

2. Related Work

Action Quality Assessment. AQA is a task of evaluating the quality of actions performed in videos across various domains such as sports events[9, 12, 22–24, 26, 27, 35, 43], healthcare[18, 30, 41, 42, 45, 46], and others. Most existing approaches[7, 9, 18, 24, 27, 30, 33, 35, 40] treat AQA as a regression task, employing diverse video representations as input and training the model with scores in a supervised manner. For example, Xu *et.al.* [37] utilizes fine-grained action information to assist in score prediction. Bai *et.al.* [2] constructs learnable action queries to encode action information in videos. While the regression paradigm has shown impressive performance in predicting precise scores, it falls short in providing comprehensive evaluations due to the single score output. Parmar *et.al.* [23] tackles the issue by proposing a new dataset with caption labels and reformulating AQA as a multi-task parallel learning paradigm. Nonetheless, the captions within the dataset are considerably noisy and informal, thus limiting the model's capacity to generate professional commentary. To solve this, we re-

annotate the captions and design a new framework to enable the multimodal information interaction guided by prompts.

Video Captioning. Video Captioning[5, 13, 25] is a task of generating language descriptions of the video content. The most common method for video captioning is based on the encoder-decoder architecture. Several works[1, 17, 20, 31, 38, 39, 44] propose to use different vision encoders[8, 10, 11, 32, 36] to extract features from video frames, and a language decoder to generate captions using the visual features. Recently, SwinBERT[14] utilizes Video Swin Transformer[16] as the vision encoder and a Transformer-based module[6] as the text decoder, resulting in an end-to-end model which directly takes video frames as input. While video captioning models can effectively capture and describe visual content, they struggle to provide detailed and reliable assessments of action quality. In this paper, we propose a prompt-guided multimodal interaction multi-task learning approach, which can precisely depict the video content and narratively evaluate the quality of actions.

3. Dataset

To facilitate research on the NAE task, we construct new video-text pair datasets. Specifically, we re-annotate MTL-AQA [23] and FineGym [29] datasets with rich narrative texts for videos, including multidimensional evaluation information such as scores, actions, and qualitative evaluations.

As shown in Figure 2, we take MTL-AQA as an example to explain our re-annotation process. The MTL-AQA dataset is a multi-task AQA dataset that provides three types of annotations for each video: action codes, scores, and text transcribed from video audio (an example is noted as **Original Label** in Figure 2). Among them, the action codes reflect all fine-grained professional action information performed by athletes. Although MTL-AQA already includes video-text pairs, the quality of its text labels is often poor because they are directly transcribed from the audio of the videos and contain a lot of interference information (an example is noted as **Ori_text** in Figure 2). In addition, the commentary information in the videos often only contains subjective evaluations from commentators and tone words; there is generally a lack of professional rigorous evaluation information such as action types and scores in the text. To solve this problem, we integrate and reconstruct the textual information in MTL-AQA with ChatGPT [19] using prompts. The prompts as shown in Figure 2 instruct ChatGPT to generate texts including professional evaluation information while preserving details from the original transcription captions. Additionally, to make generated texts more diverse, we design five versions of prompts for both datasets so that ChatGPT can generate five different evaluative texts for each video (an example of the reconstructed text is shown in Figure 2). Table 1 shows experiments using video captioning methods on the datasets before and after our re-annotation, the significantly better results also validate the effectiveness of our re-annotation.

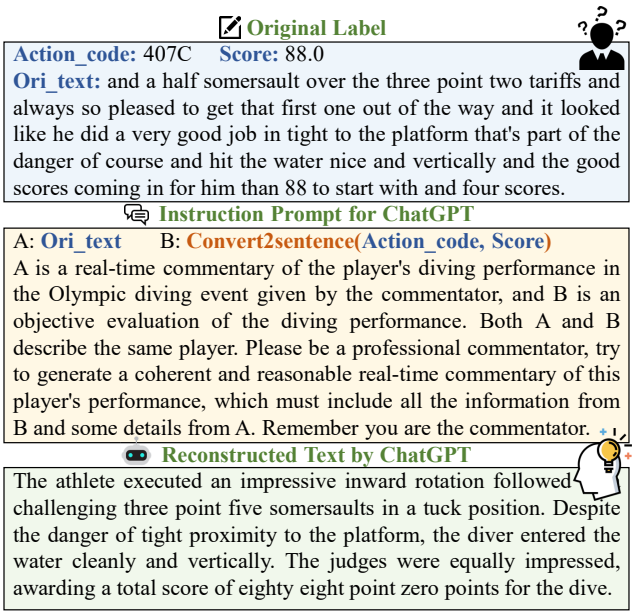


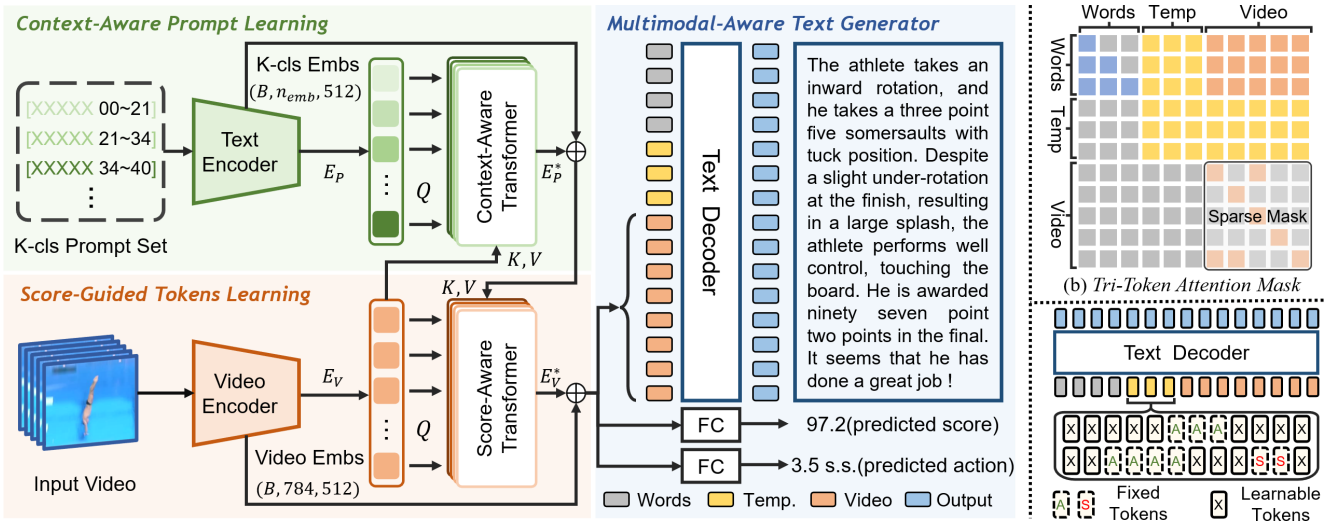
Figure 2. The process of re-annotating a sample using ChatGPT, based on existing action and score labels. *Convert2sentence(act, score)* constructs a pre-fixed template to insert the action and score information into the template to generate a complete sentence.

Table 1. Comparison of video captioning results using video captioning methods on the datasets before and after our re-annotation.

Method	MTL-AQA			FineGym		
	B4	M	C	B4	M	C
Trained with the original narration						
VLTinT[39]	2.5	12.4	6.1	2.1	10.5	5.7
SwinBert[14]	3.3	12.3	7.3	3.0	11.9	6.6
Trained with our re-annotated narration (Ours)						
VLTinT[39]	22.5	19.9	14.4	15.4	17.6	14.2
SwinBert[14]	40.2	26.4	16.2	27.4	21.0	17.8

FineGym offers fine-grained action information for each video. During the re-annotation, we first annotate scores and transcription captions for each sample (taken from the original competition videos). Then following a similar approach mentioned above with MTL-AQA, we insert the actions, scores, and transcription captions into prompts that are inputted into ChatGPT to generate narrative evaluative texts.

Finally, to ensure the accuracy of the evaluation information in the generated texts, we hire 8 professional divers and gymnasts to check the generated video-text pairs. Each pair is checked by two athletes to ensure that there are no changes made to the action and score information in the generated text and that the correspondence between scores and quantitative evaluations in the text is correct. For example, when a player scores 98 points, the quantitative evaluation in the text should be positive rather than negative. If there are any errors in a sample's text, we ask ChatGPT to iterate again until it passes the inspection. The entire annotation and checking process above takes about 150 hours to complete. For more details, please refer to supplementary materials.



(a) *Overview of the Prompt-Guided Multimodal Interaction Paradigm* (b) *Tri-Token Attention Mask* (c) *Learnable Template*

Figure 3. The left part shows an overview of our Prompt-Guided Multimodal Interaction paradigm. First, we send the K-class Prompts into the text encoder to get K-class Prompt Embeddings. After that, we perform Context-Aware Prompt Learning using the video features based on Context-Aware Transformer. Second, in Score-Guided Tokens Learning, we interact the video embeddings from the video encoder with the K-class Prompts mentioned above through Score-Aware Transformer. Thirdly, we utilize Multimodal-Aware Text Generator with the Tri-Token Attention Mask to integrate the multimodal tokens from Score-Guided Tokens Learning and generate the text. The upper right part shows the Tri-Token Attention Mask and the bottom right part shows the learnable template in Multimodal-Aware Text Generator.

4. Approach

4.1. Overview

An overview of our Prompt-Guided Multimodal Interaction is shown in the left part of Figure 3. Our framework consists of three parts. Firstly, in Context-Aware Prompt Learning, we employ a cross-attention module to refine the learnable prompt embeddings E_P, which contain score information, with contextual features from video embeddings E_V. During this stage, E_V acts as the prompt to guide E_P to perceive contextual information, thus fusing the language modality and video modality. This process can be represented as:

$$E_P^* = \textbf{MHCA}(E_P, E_V) + \gamma_1 E_P, \quad (1)$$

where $\textbf{MHCA}(,)$ indicates multi-head cross-attention whose first parameter denotes "*query*", and the second parameter means "*key*" and "*value*"; E_P^* means refined context-aware prompt embeddings; and γ_1 indicates learnable coefficient.

Secondly, in Score-Guided Tokens Learning, we adopt a similar approach to Context-Aware Prompt Learning. Specifically, we use the refined score-aware embeddings E_P^* mentioned above as the prompt to guide video embeddings E_V to perceive the score information, thus achieving the second multimodal interaction. This process can be represented as:

$$E_V^* = \textbf{MHCA}(E_V, E_P^*) + \gamma_2 E_V, \quad (2)$$

where E_V^* denotes refined video embeddings integrated with score-aware embeddings; γ_2 denotes learnable coefficient.

Finally, we use Multimodal-Aware Text Generator to generate the narrative evaluation in an autoregressive paradigm, whose input consists of word tokens already produced, learnable template tokens (bottom right of Figure 3), and refined video embeddings from Score-Guided Tokens Learning. The learnable template tokens contain score information and action information which is predicted by passing the refined video embeddings E_P^* through an MLP respectively. During the generation of narrative texts, we use the Tri-Token Mask (upper right of Figure 3) to guide the text generation process to focus on professional evaluation information from the learnable template and video information from the refined video embeddings. This process can be represented as:

$$Input = \textbf{Concat}(Word_{1:i-1}; Template; E_V^*), \quad (3)$$

$$Word_i = \textbf{Decoder}[\textbf{Mask}_{\textbf{TT}}(Input)], \quad (4)$$

where $Word_i$ indicates the i-th word during generation, $\textbf{Concat}(;;)$ denotes concatenation operation, $\textbf{Mask}_{\textbf{TT}}$ represents Tri-Token Mask. We will now introduce the three parts in our framework separately in the following texts.

4.2. Context-Aware Prompt Learning

Textual Prompt Construction. To guide the model in generating texts with accurate evaluation information, we incorporate the score into the generation process. One simple approach is to utilize the video features acquired from a pre-trained video backbone and simultaneously conduct score prediction and text generation using separate heads. However, this approach overlooks the distinctions between the score regression task and the text generation task. Furthermore, scores, actions, videos, and texts belong to distinct modalities; nevertheless, this approach fails to account for these modality differences. To tackle this, we reformulate

the score prediction problem as a video-text matching problem. Specifically, we employ the learnable prompt to convert numerical scores into textual form. To represent all possible scoring situations with a limited number of categories in the video-text matching problem, we cannot directly use text to represent specific scores because scores are continuous and there are infinitely many of them. Therefore, we use text to represent finite intervals of scores. When predicting scores, different score intervals represent different categories. These intervals do not overlap, and the beginning and end of two adjacent score intervals are the same. To this end, we obtain multiple score intervals with similar distribution probabilities in the sample space. Concretely, we first collect a list of all scores $\mathbf{S} = [S_1, ..., S_N]$ from training samples. Then, we sort the list in ascending order to obtain $\mathbf{S}^* = [S_1^*, ..., S_N^*]$. Given the interval numbers R, the partitioning algorithm gives the bounds of each interval $\mathcal{I}^r = (\zeta_{left}^r, \zeta_{right}^r)$ as:

$$\zeta_{left}^r = \mathbf{S}^*(\lfloor (N-1) \times \frac{(r-1)}{R} \rfloor), \tag{5}$$

$$\zeta_{right}^r = \mathbf{S}^*(\lfloor (N-1) \times \frac{r}{R} \rfloor), \tag{6}$$

where $\mathbf{S}^*(i)$ represents the i-th element of $\mathbf{S}^*$. Then, we express the score intervals of different categories in textual form, and connect a learnable prompt of a certain length in front, so as to obtain the K-class score-aware textual prompt. For example, the score range from 25.6 to 36.3 can be expressed as "[XXXXXX] twenty-five point six to thirty-six point three", where "[X]" is a learnable token. Until then, we obtain the K-class prompt set with score information.

Context-Aware Prompting. After constructing prompts with score information, our model augments the K-class prompt embeddings from the text encoder with contextual information from video embeddings. Specifically, we use a transformer decoder to utilize the video features from the video encoder as "*key*" and "*value*", and the K-class prompt embeddings from the text encoder as "*query*", to refine the prompt embeddings with context information from the video, so as to integrate prompt embeddings with context information. Then these prompt embeddings will be used in the video-text matching during Score-Guided Tokens Learning.

4.3. Score-Guided Tokens Learning

In 4.2, we have obtained score-aware prompt embeddings that have perceived video context information. In Score-Guided Tokens Learning, we use a cross-attention module called Score-Aware Transformer, which is symmetrical to Context-Aware Transformer in 4.2, to integrate video embeddings from the video encoder with prompt embeddings.

In the Score-Aware Transformer, we aim to enhance the attention of input video information toward score-aware prompt embeddings that correspond to the video. This ensures the accurate integration of score information into the video features. As mentioned above, we use score intervals to convert the score prediction task into a classification task. So we supervise the input video to focus more on its corresponding score interval using cross-entropy loss, which is represented as, $\mathcal{L}_{CE_S} = -\sum_i p_i log\hat{p}_i$, where p_i indicates the possibility that the predicted score belongs to the i-th interval defined in 4.2. This process completes the filter of text-based score information and refines video information.

Then, we obtain the video tokens that incorporate score information. We merge these video tokens with word tokens and template tokens, then input them into the text decoder for text generation. Details of the template tokens will be explained in 4.4. Additionally, we pass the video tokens through two heads that predict the score and action respectively. The score head is an MLP. The action head consists of multiple MLPs, and each MLP corresponds to different parts' actions in videos. The MSE loss $\mathcal{L}_{MSE}$ and CE loss $\mathcal{L}_{CE_A}$ supervises the score and action prediction respectively.

4.4. Multimodal-Aware Text Generator

Following [14], we use a Transformer-based generator as the text decoder to generate the natural language description. As shown in the bottom right of Figure 3, the text decoder has input from multiple modalities, which include word tokens already generated, score-aware video tokens from Score-Guided Tokens Learning, and learnable template tokens. This template includes three parts: learnable tokens, unlearnable action, and score tokens. Specifically, in 4.3, we use two heads to predict the scores and the fine-grained action categories for each part in videos respectively. Therefore, we can obtain a score, and action types correspond to different parts of videos. We convert them into textual form and insert them as fixed tokens into the template. Learnable tokens are inserted between these fixed tokens, concatenating action, and score information into a complete sentence.

Moreover, we use the Tri-Token Attention Mask (as shown in the upper right of Figure3) in the text decoder. Specifically, during the generation process, the decoder attends to tokens already generated and all the template tokens and video tokens. The template tokens attend to themselves as well as all the video tokens. Meanwhile, following [14], a Sparse Mask is used when refining the video tokens to save the computing cost. To be concrete, the Sparse Mask is learnable. Assume that the number of video tokens is M, and V is the learnable attention mask of size $M \times M$ governing the attention among the video tokens. Then we use the sparse loss to address the redundancy among video tokens, which can be represented as:

$$\mathcal{L}_{SPARSE} = \lambda \times \sum_{i}^{M} \sum_{j}^{M} |V_{i,j}|, \tag{7}$$

where λ represents the regularization hyperparameter and $V_{i,j}$ represents the activation values of the attention mask.

Table 2. **Comparison with previous video captioning methods on two benchmarks for NAE task.** Except for the NAE metric (mAP), we also specifically compare the accuracy of professional information in the generated text using AQA and Action Classification metrics. The scores and actions are extracted from the generated text. We also utilize Video Captioning metrics to assess the quality of the generated text.

Method	MTL-NAE							FineGym-NAE						
	NAE	AQA		Captioning			Action	NAE	AQA		Captioning			Action
	mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc	mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc
C3D-AVG[23]	0.157	0.843	1.032	16.4	18.6	13.6	0.89	0.051	0.606	3.76	10.0	9.8	11.6	0.80
MSCADC[23]	0.074	0.797	1.601	16.7	18.4	13.3	0.84	0.025	0.583	4.42	10.2	10.3	12.6	0.76
UniVL[17]	0.166	0.836	1.086	16.4	18.3	13.6	0.87	0.057	0.604	3.81	11.0	10.7	13.3	0.79
VLCap[38]	0.197	0.851	0.867	19.8	18.7	13.9	0.90	0.086	0.627	3.07	13.6	12.1	13.5	0.81
VLTinT[39]	0.214	0.868	0.820	22.5	19.9	14.4	0.90	0.094	0.640	2.33	15.4	17.6	14.2	0.84
SwinBert[14]	0.261	0.881	0.706	40.2	26.4	16.2	0.92	0.118	0.656	2.13	27.4	21.0	17.8	0.85
Ours	**0.383**	**0.943**	**0.340**	**42.2**	**28.2**	**20.5**	**0.97**	**0.162**	**0.749**	**1.55**	**28.9**	**23.7**	**20.7**	**0.93**

5. Experiment

5.1. Experimental Setup

Evaluation Metrics. We want our model to accurately provide evaluation information (such as score and action) and rich descriptions (as in Video Captioning), intuitively.

To simultaneously consider the performance of score prediction, action prediction, and text generation, we propose to measure the mean Average Precision (AP) across a range of thresholds for evaluation metrics in AQA [40], Action Classification, and Video Captioning [34]. For AQA we use intersection over relative ℓ_2-distance (R-ℓ_2) thresholds 0.003, 0.005, 0.010, 0.015, 0.020. For captioning we use CIDEr score thresholds .05, .10, .15, .20, .25. For Action Classification we use the average Acc of the classification results of all actions in a sentence with thresholds 0.25, 0.50, 0.75, 1. We adopt CIDEr since the idea that good captions should not only be similar to the reference captions in terms of word choice and grammar but also in terms of meaning and content [34]. We calculate the average precision for all the possible combinations of thresholds and report the average of APs. Thus, the mAP values range from 0 to 1.

In our experiments, we also use mainstream metrics for traditional tasks to compare with our task and method. For AQA, we use Spearman's rank correlation (ρ) [33, 37, 40, 43] and R-ℓ_2 to measure the rank correlation and relative distance (following previous works, we multiply R-ℓ_2 by 100). And for Captioning, we also use BLEU[21], METEOR[3] to assess the language generation capability of the models. Among these metrics, the lower the R-ℓ_2, the better the performance. While for all other metrics, the higher, the better.

Implementation Details. We conduct experiments on our re-annotated datasets, referred to as **MTL-NAE** and **FineGym-NAE** respectively. Following the original split for MTL-AQA in [23], we divide both datasets into a training set and a test set at a ratio of 3:1 (5295 for training and 1765 for testing in MTL-NAE, 4665 for training and 1560 for testing in FineGym-NAE). We use the Video Swin Transformer [16] pre-trained on the Kinetics600 dataset and CLIP [28] model with pre-trained weights provided by Huggingface. Other components are randomly initialized. We employ the AdamW optimizer and include a learning rate warm-up during the first 10% of training steps followed by linear decay. More details can be found in the supplementary materials.

5.2. Results on Narrative Action Evaluation

As shown in Table 2, we conduct NAE experiments on two re-annotated datasets. The results show that the mAP for the NAE task of our method is significantly better than previous methods (achieving improvements of 46.7% and 37.3% on MTL-NAE and FineGym-NAE respectively). To demonstrate more precisely that our method can generate sentences that balance evaluation accuracy and linguistic richness, we extract score and action information from the generated sentences to calculate evaluation metrics for the AQA and Action Classification tasks. We also calculate metrics for the Captioning task. It can be seen that our method outperforms previous methods in these three subtasks, especially in predicting professional evaluation information (scores and action types). The main reason is that previous methods ignore the accuracy of professional information when generating text. Due to the flexibility of text generation, professional evaluation information often only accounts for a small proportion of sentences. Therefore, traditional methods cannot accurately provide professional evaluations during text generation. Our method, however, can guide multi-task interactions and introduce professional information into the text generation to generate accurate evaluation information.

5.3. Ablation Study

Effectiveness of Our Multi-Task Learning Paradigm. To verify the superiority of our method compared to the traditional multi-task learning paradigm, we conduct comparisons with the multi-task learning method, C3D-AVG, and MSCADC, from [23]. These methods jointly train a backbone with independent branches. Our approach differs from them by coupling the tasks closer with prompt-guided multimodal interaction. We compare the strategies

Table 3. **Comparison with existing multi-task learning methods on MTL-NAE.** We change the backbone in [23] to Video Swin Transformer[16] for fair comparisons. For AQA and Action, methods with $*$ use scores and actions from their regression and classification heads. Ours uses scores and actions from the generated sentences except for the AQA-only case in the first line.

Method	Tasks	NAE	AQA		Captioning			Action
		mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc
AVG*	AQA	-	0.903	0.512	-	-	-	-
	Cap	0.149	-	-	16.5	18.6	13.7	-
	Cap+AQA	0.151	0.910	0.472	16.2	18.4	13.6	-
	Cap+AQA+Cls	0.157	0.912	0.542	16.4	18.6	13.7	0.97
MSC*	AQA	-	0.863	0.840	-	-	-	-
	Cap	0.071	-	-	17.0	18.6	13.5	-
	Cap+AQA	0.072	0.857	0.847	16.5	18.3	13.4	-
	Cap+AQA+Cls	0.074	0.860	0.842	16.7	1804	13.3	0.84
Ours	AQA	-	0.909	0.633	-	-	-	-
	Cap	0.341	0.897	0.569	40.9	27.7	17.8	0.92
	Cap+AQA	0.379	0.940	0.346	41.7	27.6	19.5	0.93
	Cap+AQA+Cls	**0.383**	**0.943**	**0.340**	**42.2**	**28.2**	**20.5**	**0.97**

Table 4. **Ablation study of different components.** *CAT.* means Context-Aware Transformer, and *SAT.* is Score-Aware Transformer.

w/o Module	NAE	AQA		Captioning			Action
	mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc
CAT.	0.352	0.919	0.428	41.46	27.37	19.08	0.95
SAT.	0.347	0.886	0.589	41.73	27.69	19.68	0.95
Ours	**0.383**	**0.943**	**0.340**	**42.23**	**28.22**	**20.54**	**0.97**

on the MTL-NAE dataset with the visual backbone in [23] changed to Video Swin Transformer [16] for a fair comparison. As shown in Table 3, our multi-task learning method brings significant performance gain to single-task learning. Specifically, compared to the single AQA training mode, our method shows a significant performance improvement, with R-l2 decreasing by 46.3%. Compared to the single Captioning training mode, all metrics have improved, with CIDEr increasing by 15.2%. However, the method in [23] improves little and even brings degradation to their single-task method. These results demonstrate the effectiveness of our proposed prompt-guided multimodal interaction multi-task learning.

Effect of Different Framework Components. We then investigate the effectiveness of various components in our framework. As shown in Table 4, we observe the following two facts. Firstly, without the Context-Aware Transformer, the model fails to integrate textual score information with contextual information in videos. Consequently, removing the Context-Aware Transformer results in a 25.9% increase in R-ℓ_2 and a drop from 20.54 to 19.08 in CIDEr, ultimately leading to an 8.8% decrease in mAP. Secondly, if video information is not guided to perceive language modality scores by the Score-Aware Transformer, the integration of video information and score information becomes impossible, leading to a significantly worse performance in AQA metrics

Table 5. **Ablation study of loss functions on MTL-NAE dataset.**

w/o Loss	NAE	AQA		Captioning			Action
	mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc
CE	0.308	0.881	0.652	41.62	27.32	18.96	0.97
MSE	0.316	0.876	0.651	41.56	27.35	19.22	0.97
CE+MSE	0.293	0.851	0.783	41.38	27.44	18.74	0.97
Ours	**0.383**	**0.943**	**0.340**	**42.23**	**28.22**	**20.54**	**0.97**

Table 6. **Ablation study of different kinds of templates on MTL-NAE dataset.** *w/o Lr.* means the template without learnable tokens. *All Lr.* denotes the template composed entirely of learnable tokens.

Template	NAE	AQA		Captioning			Action
	mAP	$\rho\uparrow$	R-$\ell_2\downarrow$	B4	M	C	Acc
w/o Lr.	0.364	0.935	0.366	41.02	27.77	19.54	0.95
All Lr.	0.258	0.898	0.605	36.90	25.99	14.21	0.90
Ours	**0.383**	**0.943**	**0.340**	**42.23**	**28.22**	**20.54**	**0.97**

where ρ drops from 0.943 to 0.886 and the R-ℓ_2 increases by 73.2%, ultimately resulting in a 9.4% decrease in the mAP.

Effect of Different Loss Functions. We adopt two loss functions to incorporate the score information, namely mean squared error loss (MSE), which directly regulates predicting scores, and cross-entropy loss (CE), which plays a vital role in regulating the Score-Aware Transformer. We conduct experiments by removing one or both of them, and the results are shown in Table 5. We observe that no matter which loss function is removed, it will lead to a significant decrease in the final AQA metrics, and then result in a significant decrease in the NAE metric mAP. This result proves the effectiveness of both loss functions. The NAE task has a high sensitivity to the accuracy of professional information. Even though these two loss functions related to score prediction have little impact on text generation performance, they can supervise the model to generate more accurate evaluation information, thus improving the performance of NAE task.

Effect of Learnable Template. To demonstrate the effectiveness of our template design, we compared the performance of models under three different templates in Table 6. The three templates are *w/o Lr.* (only contains fixed evaluation information tokens and without learnable tokens), *All Lr.* (only contains learnable tokens), and *ours* (contains both evaluation tokens and learnable tokens). It can be seen that *w/o Lr.* introduces evaluation information into the text generation process, thus significantly outperforms *All Lr.*, which includes no evaluation information. While its performance is worse than the template inserted with the learnable tokens.

5.4. Analysis on Action Quality Assessment

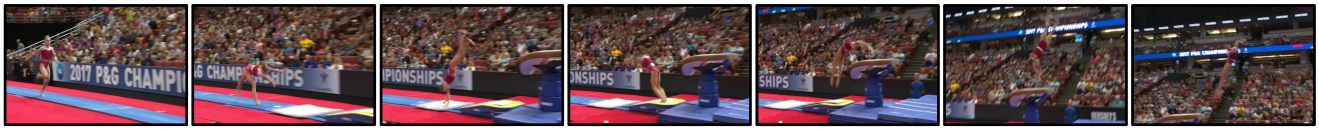
While the primary objective of the NAE task is to generate detailed narrations in language form, we also compare the

Prediction

The athlete executes a **back rotation** followed by **three point five somersaults** with a **pike position**, displaying **impressive control and precision**. the entry into the water is **flawless**, earning the athlete a well-deserved score of **ninety one point eight points**.

Ground-Truth

The athlete executes a **back rotation** followed by a flawless **three point five somersaults** with precise **pike positioning**. The entry into the water was **perfect**, leaving no splash and proving **impressive control**. This performance deems him a strong contender with a score of **ninety one point eight points**.



Prediction

With a powerful run-up and a powerful approach, the athlete executed a **flawless round-off, flic-flac on, stretched salto backward with two turns off**. the judges were impressed with the athlete's technique and awarded them a score of **fourteen point two** for her **outstanding** performance.

Ground-Truth

This athlete's performance was **commendable**, displaying **remarkable** skill in executing a **round-off**, **flic-flac on**, **stretched salto backward with two turn off**. The final result was an impressive score of **fourteen point four**, reflecting the athlete's talent and dedication to the craft.

Figure 4. **Qualitative results.** Our model can generate detailed narrations including ***scores***, ***actions***, and ***qualitative evaluations*** to describe and evaluate the actions comprehensively. Notably, the model can analyze the quality of actions by pointing out the details of the execution.

Table 7. **Comparison with state-of-the-art AQA methods on two benchmarks.** Although the NAE task aims to generate comprehensive natural language assessments, our framework outperforms all of the methods that use a single video as input on both datasets.

Method	MTL-NAE		FineGym-NAE	
	$\rho\uparrow$	R-$\ell_2\downarrow$	$\rho\uparrow$	R-$\ell_2\downarrow$
Methods with single input video				
C3D-LSTM[24]	0.849	-	0.641	-
C3D-AVG-MTL[23]	0.904	-	0.701	-
USDL[33]	0.923	0.468	0.726	1.82
MUSDL[33]	0.927	0.541	0.729	1.78
Ours	**0.943**	**0.340**	**0.749**	**1.55**
Methods with several input videos				
CoRe[40]	0.951	0.260	0.754	1.34
TPT[2]	0.961	0.238	0.764	1.27

performance of the score prediction with existing state-of-the-art methods in traditional score-based AQA [2, 23, 33, 40]. The results are shown in Table 7. Notably, our model attains the result of 0.943 on Spearman's rank correlation and 0.340 on relative ℓ_2-distance on MTL-NAE, surpassing all the methods that use a single input video. Besides, our model achieves comparable performance with recently proposed methods that require additional exemplar videos [2, 40], which predict score differences by comparing multiple input videos. Such a paradigm is suitable for comparative tasks like predicting score difference, but not for our NAE task that needs to focus on the information of a single video since the input of multiple videos may introduce noises. These experimental results prove that our approach can predict the score accurately, even only using a single video as the input.

5.5. Qualitative Results

In Figure 4, we display qualitative examples of our model. We observe that our model is capable of generating detailed narrations that describe the corresponding action categories and scores. Notably, our model can assess and analyze the quality of actions, as indicated by phrases such as "entry...flawless" and "impressive control", highlighting commendable execution and areas where improvements can be made. Besides, our predicted action categories and scores are accurate and basically the same as the ground truth. More qualitative results can be found in supplementary materials.

6. Conclusion

In this paper, we have introduced the Narrative Action Evaluation (NAE) task, which aims to generate professional commentary to assess action executions maintaining both narrative flexibility and evaluation rigor. To address this task, we have proposed a Prompt-Guided Multimodal Interaction multi-task learning framework, which interacts and integrates different tasks and information from different modalities, thereby achieving performance improvement in the NAE task and multiple subtasks. To facilitate further research in this field, we have re-annotated the MTL-AQA and FineGym datasets and established benchmarks for the NAE problem. Extensive experimental results have demonstrated the power and efficiency of our model with respect to baselines based on the previous state-of-the-art methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62125603, Grant 62321005, and Grant 62336004, and in part by CCF-Tencent Rhino-Bird Open Research Fund.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, pages 12487–12496, 2019. 3

[2] 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, 8

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[4] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *ICCV*, pages 2177–2185, 2017. 2

[5] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In *IJCAI*, page 2, 2019. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 3

[7] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR*, pages 7862–7871, 2019. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3

[9] Andrew S Gordon. Automated video assessment of human performance. In *AI-ED*, 1995. 2

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[12] Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič. Trajectory based assessment of coordinated human activity. In *ICVS*, pages 534–543, 2003. 2

[13] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *NeurIPS*, 2021. 1, 3

[14] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, pages 17949–17958, 2022. 3, 5, 6

[15] Sheng Liu, Zhou Ren, and Junsong Yuan. Sibnet: Sibling convolutional encoder for video captioning. *TPAMI*, 43(9): 3259–3272, 2020. 1

[16] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 3, 6, 7

[17] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1, 3, 6

[18] Anand Malpani, S Swaroop Vedula, Chi Chiung Grace Chen, and Gregory D Hager. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In *IPCAI*, pages 138–147, 2014. 2

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 3

[20] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, pages 10870–10879, 2020. 3

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 6

[22] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 2

[23] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *CVPR*, pages 304–313, 2019. 2, 3, 6, 7, 8

[24] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPR*, pages 20–28, 2017. 2, 8

[25] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *CVPR*, pages 8347–8356, 2019. 3

[26] Matej Perše, Matej Kristan, Janez Perš, and Stanislav Kovačič. *Automatic evaluation of organized basketball activity using bayesian networks*. 2007. 2

[27] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014. 2

[28] 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. 6

[29] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 2, 3

[30] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of osats using sequential motion textures. 2014. 2

[31] Botian Shi, Lei Ji, Zhendong Niu, Nan Duan, Ming Zhou, and Xilin Chen. Learning semantic concepts and temporal alignment for narrated video procedural captioning. In *ACM MM*, pages 4355–4363, 2020. 3

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017. 3

[33] 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, 6, 8

[34] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[35] Vinay Venkataraman, Ioannis Vlachos, and Pavan K Turaga. Dynamical regularity for action analysis. In *BMVC*, pages 67–1, 2015. 2

[36] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 3

[37] 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, 6

[38] Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In *ICIP*, pages 3656–3661, 2022. 1, 3, 6

[39] Kashu Yamazaki, Khoa Vo, Sang Truong, Bhiksha Raj, and Ngan Le. Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2211.15103*, 2022. 1, 3, 6

[40] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, pages 7919–7928, 2021. 2, 6, 8

[41] Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *MMAR*, pages 19–24, 2011. 2

[42] Qiang Zhang and Baoxin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *TPAMI*, 37(6):1206–1218, 2014. 2

[43] 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*, 2023. 2, 6

[44] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *CVPR*, pages 9837–9846, 2021. 1, 3

[45] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan Essa. Automated assessment of surgical skills using frequency analysis. In *MICCAI*, pages 430–438, 2015. 2

[46] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, and Irfan Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *IJCARS*, 13(3): 443–455, 2018. 2