Assessing LVLM alignment for the Evaluation of Automatic Video Commentary Generation

Erica K. Shimomoto[1] Edison Marrese-Taylor[1,2] Ichiro Kobayashi[1,3]
Hiroya Takamura[1], Yusuke Miyao[1,3]
[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]The University of Tokyo [3]Ochanomizu University
{kidoshimomoto.e, edison.marrese}@aist.go.jp, koba@is.ocha.ac.jp
takamura.hiroya@aist.go.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

Video commentaries are a set of timed subtitles for videos that describe their contents and add information useful to the viewer. Automatically generating commentaries from videos is a challenging task, and evaluation has primarily relied on automatic n-gram overlap metrics. However, such metrics are not ideal due to the open-ended nature of the task. This inadequacy has also been observed in other tasks, leading to the LLM-as-a-Judge approach. However, the use of vision LLMs (LVLMs) for video-based tasks evaluation remains mostly unexplored. Thus, we explore LVLMs to evaluate video commentaries and perform a human evaluation. Our results show that the LVLMs can stably evaluate our task while agreeing with humans.

1 Introduction

Video Commentary Generation aims to generate a set of timed subtitles commenting on the contents of a given video, mimicking the live commentaries we often see in sports matches and gaming live streamings. Such commentaries can describe the actions and objects in the video, as well as include additional information regarding the contents, making spectators more excited, more immersed, and better informed about what they are viewing [1].

We find previous works generating commentary on specific domains, such as sports [2, 3] or video games [4, 5], with models often relying on field-specific information to aid the generation. We also find works tackling the open-domain setting of this task, where the goal is to enable models to generate commentaries for videos containing actions in many situations [6, 7].

One open problem within this task is the evaluation. Evaluation for video commentary generation has so far relied on n-gram overlap metrics, such as BLEU. However, commentators may choose to talk about a given subject at different times and attend to different points [6], and, therefore, the usefulness of evaluation schemes based on text similarity is limited.

This inadequacy of standard evaluation metrics has also been observed in other tasks, leading to the LLM-as-a-Judge approach [8]. LLMs have been shown to align with human preference in text-to-text applications, but we find their use in multimodal applications is still limited [9]. While attempts to use LVLMs to evaluate video commentaries exist [7], they solely rely on image-based LVLMs and are based on a QA setting.

Therefore, this paper explores LVLMs to evaluate video commentaries, more specifically, Gemini 2.5-Flash and GPT 4-o. We ask LVLM judges to evaluate commentaries based on a constructed rubric that accounts for the aspects that previous works suggest make a good quality commentary [2, 3, 6, 7].

Furthermore, to understand how well these models can evaluate video commentaries, we perform an extensive human evaluation. Utilizing the same key aspects, we ask human annotators to evaluate the quality of human-generated and automatically-generated video commentary.

Our results show that both LVLM judges can stably evaluate the quality of generated video commentary, reliably agreeing with each other (Kendall's $\tau \approx 0.60$). Furthermore, we observe that LVLM judges are in agreement with human evaluation ($\tau \approx 0.37$), similar to the agreement between humans ($\tau \approx 0.36$). Our results suggest that LVLMs may be a feasible alternative to standard metrics and human evaluation for video-based tasks.

2 Related Work

Video Commentary Generation To the best of our knowledge, the task of automatically generating video commentaries was first proposed in the context of racing car video game streams [4], releasing the first dataset annotated for this task, which consisted of gameplay videos aligned with transcribed spoken commentaries. Other works have also worked on automatically generating commentary for sports matches [10, 3]. Further work focused on a similar task in an open-domain setting [6], detailing the construction of a dataset of transcribed commentary aligned with videos containing human actions in a variety of domains based on videos from ActivityNet [11]. The lack of domain-specific information led to low performance, an issue that was later solved by incorporating spatial features obtained by object detectors [7].

Video Commentary Evaluation Early works on the task have mainly relied on n-gram overlap metrics [4, 6], such as BLEU score. However, these metrics are not ideal for evaluating video commentaries, as commentators are free to talk about different aspects of the video at different timings, leading to low inter-annotator agreement [6]. To tackle this problem, image LVLMs have been employed to obtain reference-free commentary evaluation, such as the answerability score [7]. This method relies solely on the visual modality to generate a set of questions for each input video. The answerability score varies according to how many questions can be answered based on the target commentary, revealing commentaries that may not describe what is in the video or are unrelated to the video. However, this metric still relies on ground-truth timestamps to select segments from which questions will be derived, potentially not covering the whole video duration, and it overlooks other aspects that characterize a good quality commentary, such as helpfulness and engagement.

3 Methodology

3.1 Commentary Data

Our main target of evaluation is the Open-Domain Live Commentary Dataset [6]. This dataset that was constructed by asking crowdsource workers to comment a subset of the videos in ActivityNet [12], and later improved [7]. In total, this dataset contains approximately 25K human commentary utterances, covering 6,506 videos (3,631 for training and 2,875 for validation) with a total of 220 hours of annotated visual content. In this paper, we only evaluate data from the validation set.

Furthermore, we also evaluate automatic generated commentaries from the videos in the validation set, using the following methods:

1. **LVLM**: We use Gemini 2.5-Flash to generate video commentaries in a zero-shot fashion, by asking the model to comment the video. Prompt details can be found in §A. We chose Gemini 2.5-Flash to generate these commentaries given its native capabilities of directly processing videos.
2. **Encoder-decoder** [6]: A BERT-size encoder-decoder architecture that relies on an offline I3D video encoder.
3. **Spatial-graph** [7]: A model that augments the above encoder-decoder architecture with spatial information in the form of objects detected in key frames through a spatial graph parameterized by a neural network.

Encoder-decoder and spatial-graph commentaries were generated from videos in the validation set of the dataset, using checkpoints provided by the respective authors.

3.2 Qualities of a good video commentary

Visual content is often accompanied by objective statements or subjective remarks about the events in the video given by a commentator, aiming to help the spectators understand events in the videos. They also excite spectators and make them more immersed [1].

Based on these observations, we establish three key aspects to evaluate video commentaries:

- **Content:** Video commentaries should describe the contents of the video and add complementary related information.
- **Helpfulness:** Information brought by the video commentary should help viewers better understand the contents of the video.
- **Engagement:** Video commentaries should incite curiosity and grab viewers attention.

3.3 LVLM-as-a-judge

We adopt the LLM-as-a-Judge (LJ) approach [9] and utilize Gemini 2.5-Flash, an LVLM that can natively support

Prompt for LVLM Evaluation

You are given the following input video and output commentary. The output is the transcription of a live commentary made about the input video where each line contains the [start, end] timestamps in seconds of the commentary and the commentary utterance. The commentary was made as the video was playing and, therefore, there is some delay between the commentary and the happenings in the video. Rate the quality of the output commentary with respect to three aspects:
Content: Does the commentary talk about things that are happening in the video or about things related to it?
Helpfulness: Does the information brought by the commentary help viewers better understand what is happening in the video?
Engagement: Does the commentary grab the viewer attention, increasing chances the viewer will watch the video until the end?
Output a score of 1 - 5 for each aspect.
{input-video}
{commentary}

Figure 1: Prompt for our Gemini-based LLM-as-a-Judge evaluation, where {input-video} is a placeholder for the raw video, in the case of GEMINI 2.5-FLASH, or a sequence of video frames, in the case of GPT 4-o.

video for evaluation. Since, to the best of our knowledge, we are the first ones to utilize such an approach to evaluate a video-to-text generation task, we assess the capabilities of this model by considering an alternative, image-based judge like GPT-4o. For both models, we enable reasoning tokens and obtain three different answers for each evaluated commentary. For GPT-4o, we use low image resolution. We assess the consistency by computing standard deviation and inter-annotation agreement (IAA).

The LJ evaluation is based on a constructed rubric that accounts for the key aspects established in Section 3.2. We present the video commentary, including the corresponding timestamps of each utterance, alongside the video, and ask the LVLM to rate the quality of the commentary using a 1-to-5 score for the established key aspects. Fig. 1 shows the evaluation prompt we used.

As GEMINI 2.5-FLASH can directly accept video as inputs, we include the whole video along the prompt. However, GPT 4-o only accepts images and, therefore, we input a sequence of video frames. To match the amount of information processed by GEMINI 2.5-FLASH, frames for GPT 4-o are sampled at 1fps.

3.4 Human Evaluation

To verify the validity of the evaluations given by the LVLM-judges, we conduct an extensive human evaluation. Utilizing the same key aspects introduced in Section 3.2, we ask human annotators to evaluate the quality of the video commentaries. We developed a visual interface where the annotators can watch the video with the embedded video commentary as subtitles and rate each aspect in a 1-to-5 Likert scale. We carry out the study in two steps.

First, we run a small scale experiment using 1% of the videos sampled from the evaluation set, to assess the feasibility of utilizing crowdsourcing for such an evaluation. These videos are separately evaluated by two volunteer expert human annotators. They are co-authors and hence well-informed about the details of the project.

Then, we run a large-scale study by crowdsourcing via the Prolific platform, with 5% of the videos. We provide our workers with general guidelines about the purpose and scope of the study, as well as precise instructions on how to evaluate each aspect, detailed in §B. Only native English speakers participated in our study. Each worker was informed about the details of the project, and each agreed to participate in our study on their own volition.

To make the duration of each annotation task manageable, we only allow each worker to evaluate a total of 25 video-commentary pairs, roughly equivalent to one hour of work per annotator (approx. 54 minutes). Compensation is an hourly rate of 8 USD (or 6 GPB), which is approximately equivalent to the minimum salary where this study is hosted. A total of 116 workers participated in our study. Of them, 59 were female, and 55 were male.

4 Results

Table 1 show our results. Along with LVLM-judges and human scores, we also include BLEU scores against the ground-truth annotations in the Live Commentary Dataset, as well as the answerability score [7].

LVLM-as-a-Judge. For our LJ evaluation, we denote $\mathbf{LJ}^i_{\text{Gemini}}$, $\mathbf{LJ}^i_{\text{GPT}}$ $i \in [1, 2, 3]$ as single predictions for the GEMINI 2.5-FLASH and GPT-4o judges. Table 2 reports Kendall's Tau (τ) and Spearman correlation coefficient (r_s). We also report the percentage of examples where there is an exact match (**EM**).

Data	BL	ρ	$\mathbf{LJ}_{\text{Gemini}}$				$\mathbf{LJ}_{\text{GPT}}$				h_{cs}			
			C	H	E	All	C	H	E	All	C	H	E	All
Human Commentary	-	0.446	2.48	1.91	1.56	1.98	3.36	2.85	2.48	2.90	3.88	3.38	3.19	3.40
LVLM	2.501	0.479	3.48	3.29	3.26	3.35	4.12	4.09	3.78	3.99	4.10	3.83	3.63	3.85
ENCODER-DECODER [6]	2.779	0.432	1.22	1.11	1.00	1.11	-	-	-	-	-	-	-	-
SPATIAL-GRAPH [7]	**30.035**	**0.561**	1.25	1.00	1.00	1.08	-	-	-	-	-	-	-	-

Table 1: Evaluation results of human commentaries against generated commentaries. BL is short for BLEU; ρ is the answerability score[7]; LJ is short for LLM-as-a-Judge, with C, H and E short for Content, Helpfulness and Engagement, respectively; h_{cs} is the human evaluation through crowdsource.

We see both models are remarkably stable in evaluating the quality of generated video commentary, reliably agreeing with each other ($\tau \approx 0.6$.) We find this is comparable to similar studies [13, 14]. Furthermore, we observed Gemini responses had a standard deviation of 0.17 for all aspects. For GPT, the value was 0.01 for Content and Engagement, but 1.36 Helpfulness. This shows there is variance in how each aspect is evaluated. However, these numbers also show how stable the LVLM evaluations are for our task.

Human Evaluation. Based on the performance in terms of LJ, for the human study, we focused on evaluating human and LVLM commentaries. As Table 2 shows, our expert human annotators (denoted as $h_e^i,\ i \in [1,2]$) exhibit high IAA with each other in our pilot study, with τ and r_s values of 0.405 and 0.468, respectively, comparable to previous work [15].

Furthermore, we find that when aggregated via average, results of our pilot crowdsourcing study (denoted as h_{sc}) are also in high agreement with expert annotators, with a Kendall's τ of 0.315. These results indicate that crowdsourcing is a valid mechanism to evaluate the quality of video commentary, validating our planned large-scale study. Final scores are reported in Table 1.

Finally, we observe that LVLMs show high agreement with human evaluators, where the agreement between crowdsource workers and LVLMs is comparable to their agreement with the human experts. The agreement is even higher with expert annotators, indicating that LVLMs may be an alternative to human evaluation.

Overall Discussions. Supervised models achieved higher BLEU scores, but received very low scores from LVLM-judges, indicating that supervised models can mimic the training data, but their generalization capacities are still limited. The higher answerability score shows that SPATIAL-GRAPH can better cover the contents of the videos, also reflected in a higher Content score by $\mathbf{LJ}_{\text{Gemini}}$, but lacks in the other two aspects. This is likely due to the human commentary on which it was trained also does not perform well in terms of Helpfulness and Engagement.

Videos	Annotator(s)	EM	τ	r_s
100%	$\{\mathbf{LJ}^i_{\text{Gemini}}\}_{i=1,2,3}$	0.579	0.687	0.768
	$\{\mathbf{LJ}^i_{\text{GPT}}\}_{i=1,2,3}$	0.579	0.651	0.719
	$\{\mathbf{LJ}_{\text{Gemini}}, \mathbf{LJ}_{\text{GPT}}\}$	0.095	0.493	0.636
5%	$\{h_{cs}, \mathbf{LJ}_{\text{Gemini}}\}$	0.071	0.321	0.428
	$\{h_{cs}, \mathbf{LJ}_{\text{GPT}}\}$	0.318	0.370	0.486
1%	$\{h^i_e\}_{i=1,2}$	0.350	0.405	0.468
	$\{h_e, h_{cs}\}$	0.092	0.315	0.414
	$\{h_e, \mathbf{LJ}_{\text{Gemini}}\}$	0.088	0.427	0.548
	$\{h_e, \mathbf{LJ}_{\text{GPT}}\}$	0.318	0.393	0.505

Table 2: Inter-annotation agreement and aggregated Likert scores for our LJ and human evaluation experiments, where h_e denotes the expert annotators and h_{cs} denotes aggregated crowdsourced human labels.

Interestingly, we observe that LVLM generated commentaries offer significantly higher *Engagement* scores compared to HUMAN COMMENTARY. LLMs are trained to generate responses that users prefer, and this can benefit video commentary generation.

5 Conclusions

This paper focused on the evaluation of video commentaries. This evaluation is not trivial, as commentators may decide to comment on a given subject from the video at different times and attend to different points. Thus, we explored the use of LVLMs to evaluate video commentaries based on three aspects: content, helpfulness, and engagement. We also performed an extensive human evaluation.

Our results show that both LVLM judges are remarkably stable in evaluating the quality of generated video commentary, reliably agreeing with each other (Kendall's $\tau \approx 0.60$). Furthermore, we observe that LVLM judges are in agreement with human evaluation with $\tau \approx 0.37$, similar to the agreement between humans ($\tau \approx 0.36$). Our results suggest that LVLMs may be a feasible alternative to standard metrics and human evaluation for video-based tasks.

Acknowledgement

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain".

References

[1] Michael Schaffrath. Mehr als 1:0! Bedeutung des Live-Kommentars bei Fußballübertragungen– eine explorative Fallstudie [more than 1:0! the importance of live commentary on football matches – an exploratory case study]. **Medien und Kommunikationswissenschaft**, Vol. 51, No. 1, pp. 82–104, 2003.

[2] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating Live Soccer-Match Commentary from Play Data. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, No. 01, pp. 7096–7103, July 2019.

[3] Byeong Jo Kim and Yong Suk Choi. Automatic baseball commentary generation using deep learning. In **Proceedings of the 35th Annual ACM Symposium on Applied Computing**, pp. 1056–1065. Association for Computing Machinery, New York, NY, USA, March 2020.

[4] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Generating Racing Game Commentary from Vision, Language, and Structured Data. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 103–113, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[5] Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Audio commentary system for real-time racing game play. In **Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations**, pp. 9–10, Prague, Czechia, September 2023. Association for Computational Linguistics.

[6] Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. Open-domain Video Commentary Generation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 7326–7339, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Erica K. Shimomoto, Edison Marrese-Taylor, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. Introducing spatial information and a novel evaluation scheme for open-domain live commentary generation. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 10352–10370, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[9] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In **Forty-first International Conference on Machine Learning**, 2024.

[10] Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In **Proceedings of the 32nd ACM International Conference on Information and Knowledge Management**, pp. 5391–5395, 2023.

[11] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 961–970, 2015.

[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 961–970, 2015.

[13] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Pairwise Neural Machine Translation Evaluation. In Chengqing Zong and Michael Strube, editors, **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 805–814, Beijing, China, July 2015. Association for Computational Linguistics.

[14] Noy Sternlicht, Ariel Gera, Roy Bar-Haim, Tom Hope, and Noam Slonim. Debatable Intelligence: Benchmarking LLM Judges via Debate Speech Evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 18861–18880, Suzhou, China, November 2025. Association for Computational Linguistics.

[15] Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. Finding Replicable Human Evaluations via Stable Ranking Probability. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 4908–4919, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

A Video Commentary Generation using Gemini

Prompt for Gemini video commentary generation

System: You are an excited live commentator. Your mission is to comment on videos, describing the main events as they happen in the video while also including interesting related information. Provide the commentary sentences and their respective start and end timestamps in seconds. Commentary sentences should be uttered within the total video length.

User: Provide one commentary sentence for the following segments of the video (start, end) in seconds.

Figure 2: Prompts for our Gemini commnetary generation.

B Crowdsource Instructions

In the following, we show the detailed instructions given to the crowdsource workers:

- Content
  - Score 1: Non-factual commentary: Commentary talks about things completely unrelated to the contents of the video / Very long sentences: When the amount of information given cannot be digested within the duration of the commentary.
  - Score 2: Key activity missing: Commentaries fail to identify the key activity, despite getting some details right. For example, it mentions a bicycle in a city when the video shows a motorcycle in a city.
  - Score 3: Wrong details: Key activity is correctly identified, but commentaries mention non-factual events, e.g., talks about a non-existing cheering crowd.
  - Score 4: Temporal misalignment: Commentaries talk about factual things in the video but in the wrong time, e.g., utterance lasts longer or
  - Score 5: None of the above issues found.
- Helpfulness
  - Score 1: Non-helpful information: Commentary talks about things that confuses the viewer. You could watch the video without the commentary. / Too much information: Commentary includes too many details that do not help better understanding what happens in the video.
  - Score 2: Plain description: The commentary consists of just a description of what happens in the video.
  - Score 3: Efficient summary: Commentaries efficiently summarizes the information shown in the video.
  - Score 4: Commentaries efficiently summarize the information in the video while and also mention things that might not be obvious to all viewers, e.g., celebrity or character names, sport-related information, it names an unfamiliar sport.
  - Score 5: It explains how something works, which helps the viewers understand what is going on in the video
- Engagement
  - Score 1: *'I can hear what's happening, but I feel nothing.'*: Passive Commentary: The commentary consists of just a description of what happens.
  - Score 2: *'This is useful, but I'm not really drawn in.'*: Informative but distant: The commentary adds factual context or explanations but has few cues about what matters most and little to no reactions.
  - Score 3: *'I'm interested, but not consistently hooked.'*: Engaging at Key Moments: Emphasis, insight, or reaction at key moments, with some guidance of attention.
  - Score 4: *'I'm pulled in and want to keep watching.'*: Actively Engaging: Consistently adds meaning beyond the visuals, with frequent attention-directing cues. Feels conversational and alive.
  - Score 5: *'I'm absorbed: this commentary elevates the video.'* Compelling, Memorable: Strong narrative flow across the video with insightful reactions. Commentary and visuals feel inseparable.