Data Augmentation for Open-Domain Live Commentary Generation

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

Abstract

This paper proposes automatic data augmentation for Live Commentary Generation using Large Vision-Language Models (LVLMs). This task aims to generate a set of timed subtitles commenting on the contents of a given video, describing the actions and objects in the video, and including additional information. Collecting data for live commentary generation can be an expensive and time-consuming task. Therefore, we propose a less labor-intensive alternative by utilizing LVLMs to generate artificial live commentary data based on frames extracted from videos. Our results using a simple live commentary generation model reveal that training on a combination of original and augmented data has the potential for performance improvement in this task in terms of BLEU score.

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]. This task is similar to video captioning, as comments may include descriptions of what is happening in videos. However, it differs significantly because the timing and contents of utterances may vary significantly, as a commentator can choose what to say, when to say, and how to say things.

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. Furthermore, 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 a variety of situations [6, 7]. The latter poses a more challenging setting, as it cannot use domain-specific features, which have proven essential for attaining good performance.

Collecting data for live commentary generation can be expensive and time-consuming, as shown by [6]. Moreover, their work suggests annotators have a lower degree of agreement than other video-to-text tasks, such as dense video captioning. Therefore, we propose to utilize vision-and-language models, specifically those from the LLaVA family, to generate artificial commentary data based on frames extracted from videos.

We further assess the efficacy of this augmentation by training a simple model for the open-domain version of the task. Our results show that while augmented commentaries differ from human-annotated ones, training on a combination of original and augmented data has the potential to improve performance in this task in terms of BLEU score.

2 Related Work

Live Commentary Generation To the best of our knowledge, the task of automatically generating live commentaries was first proposed in the context of racing car videogame streams [4], releasing the first dataset annotated for this task, which consisted of gameplay videos aligned with transcribed spoken commentaries. Other works, including [8] and [3], have also worked on automatically generating commentary for sports matches. Soon after, [6] tackled a similar task in an open-domain setting, detailing the construction of a dataset of transcribed commentary aligned with videos containing human actions in a variety of domains constructed using videos from ActivityNet [9]. As no domain-specific information was used, they achieved considerably poorer performance. Aiming to compensate for the lack of domain-specific information, [7] proposed to incorporate spatial features obtained by object detectors, bringing the open-domain performance comparable to the in-domain works.

Data augmentation via LLMs Data augmentation using LLMs is a less labor-intensive strategy to overcome issues related to data collection, such as the high costs and inaccuracies in human annotation. LLMs have been used to augment data by automatically annotating existing data [10] with performance on pair to human-annotated labels. On the other hand, they can also increase data variety by modifying data of a specific label [11]. Such capabilities can be helpful in tasks such as question answering [12] and counterfactual generation [13].

LLMs can also be useful to generate new synthetic datasets, especially in settings where it is difficult to collect data, such as in dialogue tasks [14]. Furthermore, LLM-based augmentation can also be combined with human supervision [15].

Large Vision and Language Models (LVLMs) can also be used for augmentation in multimodal settings. For example, [16] uses LVLMs to automatically generate natural language descriptions of a dataset's domains and augment the training data via language-guided image editing.

3 Proposed approach

Figure 1 illustrates our proposed augmentation approach. It generates artificial live commentaries utilizing Large Vision-Language Models (LVLM). Concretely, we propose to rely solely on the visual input. Given the aforementioned annotator disagreement issue, the intuition behind our proposal is that while the nature of the live commentary task means each annotator independently decides when and what to speak, leading to each commentator producing significantly different output utterances, it is reasonable to assume there will be a set of common-ground facts shared across annotators, which depend only on the video's visual contents.

To provide LVLMs with relevant visual information from a video, in this paper, we follow [7] and construct "visual summaries", which consist of frames sampled from the video, sorted temporally and arranged in a grid-like fashion. Each frame is labeled with an index from 1 to n, where n is the total number of frames sampled, which is added in the top left corner of each frame, as shown in Figure 2.

Although LVLMs have proven effective in a wide variety of downstream tasks, including image captioning and visual question answering [17], early experiments where we directly prompted these models to generate live commentary from visual summaries suggested that this cannot yet be achieved directly. Therefore, we propose to decompose the generation of live commentary into a set of simpler steps so that each task more closely resembles the training scheme of the LVLM. Concretely, we propose the three steps below. Actual prompts given to the models are shown in the Appendix.

Fetching Global Context We obtain a draft of the live commentary by prompting an LVLM with a visual summary of the entire video and requesting it to generate a description of the video contents in the style of live commentary. To obtain this global visual summary, we propose a *sparse frame sampling* approach, where we only use one frame from each one of the video segments defined by the timestamps of the annotations in the data.

Fetching Local Contexts We prompt an LVLM to obtain a *detailed description* of what happens on a short segment of a given video, constructing a visual summary with a set of *densely-sampled frames* from this segment. We define segments by following the timestamps of the annotations in the data, but we stress that alternative approaches are also possible. In contrast to the previous step, here we request the model to focus only on the main action that can be identified and to be succinct in the reply.

Composing live commentary We prompt an LLM to compose live commentaries based on the global and local contexts obtained during the previous steps. We request the model to generate one utterance for each one of the provided video segments, feeding the local/global contest in JSON format.

4 Experiments

Original Data We work with the dataset created by [6]. This dataset was built on top of the videos in the ActivityNet Dataset [18], where human annotators were asked to record commentary narrations of the videos in

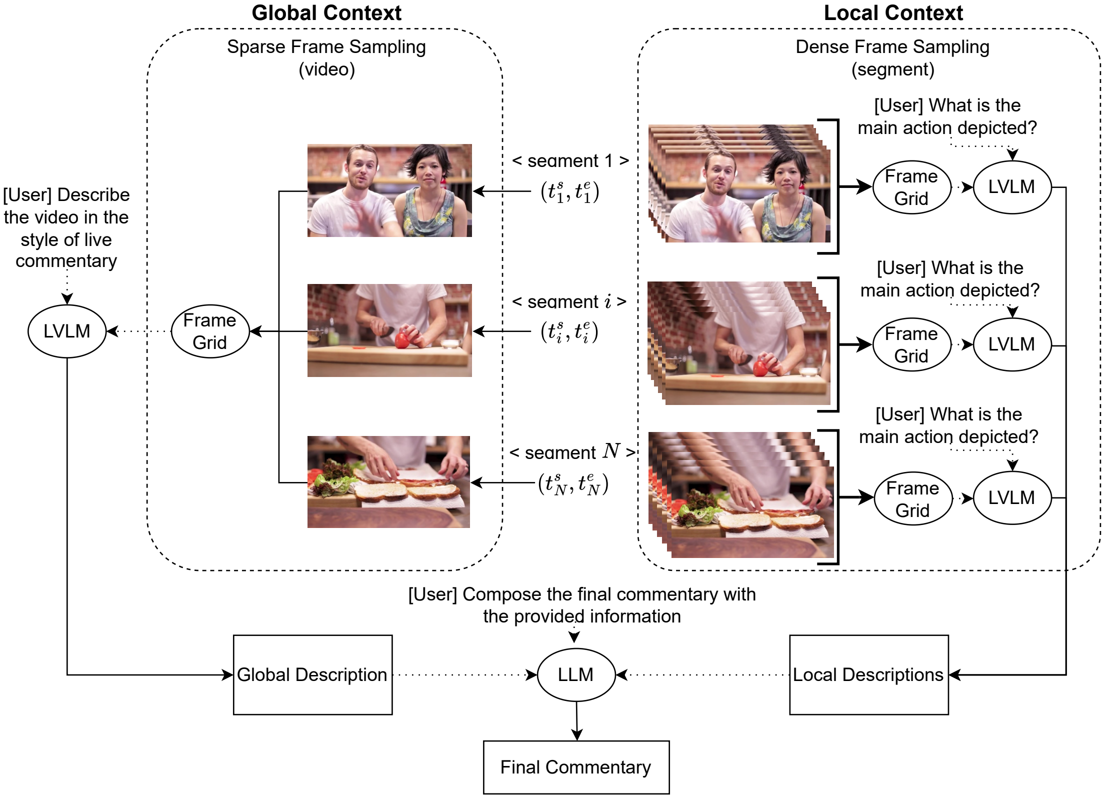
Global Context
Sparse Frame Sampling (video)
Local Context
Dense Frame Sampling (segment)
[User] Describe the video in the style of live commentary
LVLM
Frame Grid
< segment 1 >
(t_1^s, t_1^e)
< segment i >
(t_i^s, t_i^e)
< segment N >
(t_N^s, t_N^e)
[User] What is the main action depicted?
Frame Grid
LVLM
[User] What is the main action depicted?
Frame Grid
LVLM
[User] What is the main action depicted?
Frame Grid
LVLM
[User] Compose the final commentary with the provided information
Global Description
LLM
Local Descriptions
Final Commentary

Figure 1 Summary of our approach for artificial live commentary data generation. It consists of three main steps: (1) Generating a global commentary via an LVLM based on a visual summary of the entire video; (2) Generation of short, precise descriptions of key scenes via an LVLM by dense sampling frames, and (3) Composition the final commentary via an LLM, which refactors the outputs of (2) based on the global perspective offered by the (1).

English under two settings: (1) without prior knowledge and (2) after watching them once. These audios were then transcribed into text. It consists of 25k commentaries, covering a total of 6,771 videos. In our experiments, we considered only the annotations in setting (1), resulting in 6,854 commentaries for training and 6142 commentaries for validation.

Data Augmentation We use LLaVA v1.6 (34 B) [17] as our LVLM to generate the commentary blueprint and to obtain fine-grained scene descriptions. This choice is based on [7], who empirically found that this model can handle context from up to 8 frames. We thus use this number of frames to construct our visual summaries. We also use their technique to identify sharp frames, selecting the sharpest frame closest to the middle of a given segment. As a commentary composer LLM, we used Llama3 (8 B) [19] (*Meta-Llama-3-8B-Instruct*). Furthermore, we quantize the above models to 4 bits via QLoRA [20] in order to fit our GPU memory.

Finally, given our available computational resources, we augmented commentaries for only 2205 training videos and 1785 from validation videos.

Video Commentary Generation with Augmented Data To understand how the augmented data affects performance, we trained the model proposed by [6]. We followed the authors and used an offline video encoding function, using the features released by [21]. The model was trained with a maximum learning rate of 10^{-4} with Adam and a linear annealing for 5% of the epochs, with a batch size of 8 using 4 NVIDIA V100 GPUs. During inference, we utilize beam search with a beam size of 5. For evaluation, we resort to BLEU-4 separately for segments for different annotators.

In this experiment, we used three variations of the data:

- Original: Original dataset proposed by [6].
- Augmented: Data augmented using our proposed method.
- Original + Augmented: The combination of the original data and augmented data.

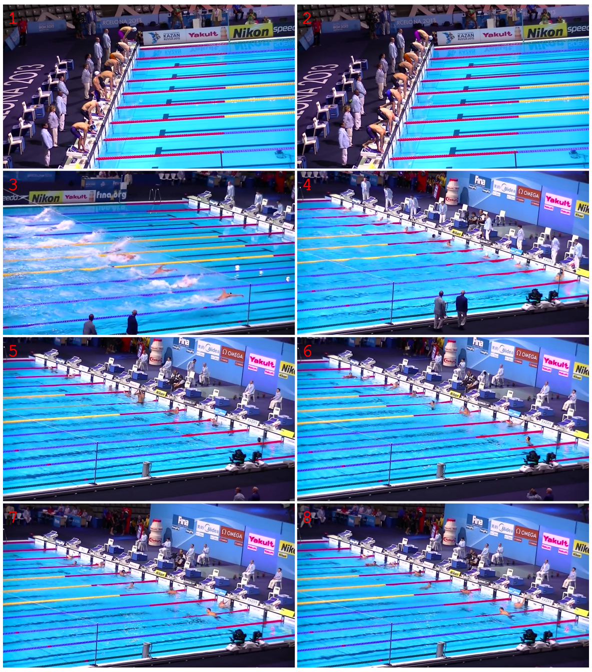

Figure 2 Example of a “visual summary” fed to the LVLM for video *v_r3dM-5cZ7e8*‘ from the dataset released by [6].

Table 1 Performance on Commentary generation in terms of BLEU score when training and testing with data augmented following our proposed method

Train Data	**Test Data**		
	Original	Augmented	Original + Augmented
Original	2.28	0.23	1.55
Augmented	0.52	6.57	3.58
Original + Augmented	**2.36**	**6.75**	**4.78**

Table 1 shows the results. When training only with the original or augmented data, the model achieves the best performance when evaluating the corresponding test data. A possible explanation might be that despite using the same videos and timestamps, human-annotated data (original) and LLMs’ augmented data are quite different.

Furthermore, we can see that results when evaluating only on the original or only on the augmented data improve when training on the combined data. This result indicates that there may be some synergy between original and augmented data. However, additional research is needed to understand better how they complement each other.

5 Conclusions

This paper proposes an automatic data augmentation method for Live Commentary Generation using Large Vision-Language Models (LVLM). Our augmentation strategy consists of three steps: (1) it generates a global commentary via an LVLM based on a visual summary of the entire video; (2) it generates short, precise descriptions of key scenes via an LVLM by dense sampling frames; (3) it composes the final commentary via an LLM, based on the global and local perspectives.

Our results from training a simple model on both original and augmented data showed that while our strategy generated commentary-like samples, they differed from human-annotated commentaries. Still, by training the model with a combination of original and augmented data, results improve when evaluating only the original data or only the augmented data. This result suggests that augmented commentaries may complement information provided by the original commentaries.

It is important to note that our study only augmented commentary for a third of the original dataset. There is no guarantee that training on augmentation on all videos will lead to performance improvements. Furthermore, while BLEU is a standard metric used in this task, it may not be able to account for commentaries focusing on different parts of the video, e.g., the description of action vs. the description of the background/environment depicted in the video, as pointed out by [6]

A natural progression of this work is to utilize other metrics to evaluate live commentary generation, e.g., as proposed by [7], to better understand the impact of the augmented data.

Acknowledgements

This paper is based on results obtained from the project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) was used for experiments. Finally, we are also grateful to the NVIDIA Corporation, which donated one of the GPUs used in this research.

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 Kido 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] 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.

[9] 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.

[10] Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In **International Conference of the Italian Association for Artificial Intelligence**, pp. 151–171. Springer, 2023.

[11] Damir Korenčić, Ivan Grubišić, Gretel Liz De La Peña Sarracén, Alejandro Hector Toselli, Berta Chulvi, and Paolo Rosso. Tackling covid-19 conspiracies on twitter using bert ensembles, gpt-3 augmentation, and graph nns. In **MediaEval 2022: Multimedia Evaluation Workshop**, pp. 243–247, 2023.

[12] Arijit Chowdhury and Aman Chadha. Generative data augmentation using LLMs improves distributional robustness in question answering. In Neele Falk, Sara Papi, and Mike Zhang, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 258–265, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[13] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. DISCO: Distilling counterfactuals with large language models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5514–5528, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. A unified dialogue user simulator for few-shot data augmentation. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 3788–3799, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 6826–6847, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[16] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 79024–79034. Curran Associates, Inc., 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, April 2023.

[18] 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.

[19] AI@Meta. Llama 3 model card. 2024.

[20] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[21] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. Dori: Discovering object relationships for moment localization of a natural language query in a video. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 1079–1088, 2021.

Global Context Prompt

The image presented contains a set of frames sampled from a video, with the frames arranged in a grid-like fashion, and each one labeled with a number in red. Please describe what is happening in the video as if it was the transcript of someone providing live commentary. As such, describe aspects of the actions to listeners who cannot see it for themselves, count the number of participants, and make educated guesses about the overall context of the video, such as the location where actions are taking place. If you believe the same person or object appears multiple times, make sure to account for it. Use the text displayed on the frames to help yourself. DO NOT mention specific frames.

Figure 3 Prompt fed to the LVLM to obtain a blueprint of the live commentary based on a sparse visual summary of the entire video.

Local Context Prompt

The image presented contains a set of frames sampled from a video. The frames are sorted temporally in a grid and each one is labeled with an index from 1 up to 8 in red. What is happening in this portion of the video? Focus on the main action you can identify. You are allowed to make educated guesses about the overall context of the video. Use text displayed on the images to help yourself. Reply in one sentence.

Figure 4 Prompt fed to the LVLM to obtain succinct, fine-grained details of the main action occurring on a given section of a video, based on a dense visual summary of the section.

Commentary Composition Prompt

The data below, provided in JSON format, contains automatically-generated descriptions of a video. The description denoted as 'global_context' present an overview or summary of the contents of the video to help you understand its overall context. The descriptions denoted as 'local_context' focus on the details of what is happening at specific segments of the video, each denoted by a specific index, and sorted by time.
{data}
Based on this information, generate a transcript in the style of live commentary for the video, making sure to tell a cohesive story and to provide an utterance for each input index of 'local_ context'. As such, describe aspects of the actions to listeners who cannot see it for themselves, and make educated guesses about the overall content of the video, such as the location where actions are taking place. Format your output in JSON format, with one entry for each input index. Reply with the JSON data output ONLY.

Figure 5 Prompt fed to the LLM to compose the final commentary based on the local and global contexts.