Commentary Generation from Multimodal Game Data for Esports Moments in Multiplayer Strategy Games

Zihan Wang
The University of Tokyo
zwang@tkl.iis.u-tokyo.ac.jp

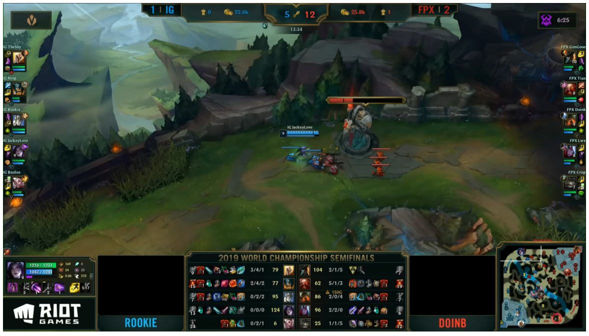
Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

Esports is a competitive sport in which highly skilled players face off in fast-paced video games. Matches consist of intense, moment-by-moment plays that require exceptional technique and strategy. These moments often involve complex interactions, including team fights, positioning, or strategic decisions, which are difficult to interpret without expert explanation. In this study, we set up the task of generating commentary for a specific game moment from multimodal game data consisting of a gameplay screenshot and structured JSON data. Specifically, we construct the first large-scale tri-modal dataset for *League of Legends*, one of the most popular multiplayer strategy esports titles, and then design evaluation criteria for the task. Using this dataset, we evaluate various large vision language models in generating commentary for a specific moment. We will release the scripts to reconstruct our dataset.

https://github.com/ArnoZWang/esports-trimodal

Screenshot of "BUILDING_KILL" event, presenting a player is taking an enemy building down:



Structured data record (in JSON format):

```
{
 "type": "BUILDING_KILL",
 "timestamp": 839426,
 "position": "x4318y13875",
 "killerId": 4,
 "assistingParticipantIds": "",
 "teamId": 200,
 "buildingType": "TOWERBUILDING",
 "laneType": "TOPLANE",
 "towerType": "OUTERTURRET"
}
```

Textual game commentary:
Fighting everywhere, Jackeylove is just farming for himself.

Figure 1: Commentary generation for esports moments from a gameplay screenshot and a structured JSON data record. This is from the contests of League of Legends.

1 Introduction

The esports industry has witnessed rapid global growth (Jenny et al., 2017; Hallmann and Giel, 2018), highlighted by its inclusion in the Asian Games since 2018 and the Olympic Esports Games in 2027 (IOC, 2025). Esports matches in popular titles like *League of Legends* (LoL), a leading multiplayer online battle arena (MOBA) game, feature a continuous sequence of highly skilled, moment-by-moment plays requiring exceptional technique and strategic decision-making. For audiences to fully appreciate these intense moments, expert commentary is essential, as it helps interpret complex player interactions, strategies, and game states.

Several studies have thus explored commentary generation from esports game data (§ 2) (Tanaka and Simo-Serra, 2021; Zhang et al., 2022; Wang and Yoshinaga, 2024). These studies typically rely on either structured data (*e.g.*, game logs, metadata) or visual information (*e.g.*, video frames) as input, and focus on generating commentary for matches. While such approaches have shown promise, they do not fully capture the granular nature of esports, where understanding often hinges on brief but critical moments. Although Zhang et al. (2022) studied moment-level commentary generation, they rely solely on structured data, failing to exploit the dual-modality nature of esports game records, which should be used together to interpret game moments.

To bridge this gap, this study sets up a task of generating commentary for a specific esports moment (Figure 1) by leveraging multimodal game

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1795–1807
December 20-24, 2025 ©2025 Association for Computational Linguistics

data, combining both static visual inputs (gameplay screenshots) and structured JSON data records. Motivated by recent advances in vision-language models (VLMs), our task explores how to effectively integrate these complementary modalities to produce concise, informative commentary and challenges VLMs on their ability to perform this integration. We construct the first large-scale trimodal dataset dedicated to League of Legends that includes aligned visual, structured, and textual data (§ 3). Additionally, we design evaluation metrics that account for the unique challenges of esports commentary, focusing on the accuracy, relevance, and interestingness of the generated text (§ 4.2).

We evaluate both open- and closed-source large vision-language models (VLMs) on our dataset to analyze their capabilities in cross-modal esports commentary generation (§ 4). Our experiments confirm the importance of combining visual and structured data to improve commentary quality and reveal the challenges still faced by current models, such as reducing inaccurate statements and capturing the strategic content of gameplay moments.

The contributions of this paper are as follows:

- We set up the task of commentary generation for League of Legends, one of the most popular esports titles, from multimodal game data (§ 3); we built a dataset to facilitate research.
- We designed several human evaluation criteria for this task (§ 4.2), and examined whether these criteria can be captured by existing automatic metrics and LLM-as-a-judge results.
- We evaluated strong VLMs on our task (§ 4.3), and confirmed the role of information provided by each data modality in esports commentary generation (§ 5).

2 Related Work

In this section, we review existing tasks of data-to-text generation for (e)sports games from game data. We first discuss text generation from structured data, followed by multimodal data.

Text generation for structured game data Prior studies have focused on generating textual summaries or commentaries for various games including physical sports (*e.g.*, basketball (Wiseman et al., 2017) and football (van der Lee et al., 2017; Taniguchi et al., 2019)), board games (*e.g.*, shogi (Kameko et al., 2015) and chess (Jhamtani et al., 2018)), and esports (*e.g.*, Dota 2 (Zhang et al., 2022) and League of Legends (Wang and Yoshinaga, 2024)). These data-to-text studies typically transcribe structured game data, such as match statistics, move expressions, and real-time action records, into textual descriptions. A notable difference among these data-to-text tasks is the information density in the input data; while physical sports provide mainly match statistics, board games and esports provide more detailed, play-by-play records, enabling commentary for each moment. However, compared to turn-based board games, esports games are fast-paced video games, making generation solely from structured data more challenging (Zhang et al., 2022).

Game video captioning Another rich source for game commentary generation is multimodal data, in particular, gameplay videos. Early studies primarily focused on match footage (video recordings) of physical sports such as tennis (Yan et al., 2016), basketball (Yu et al., 2018), baseball (Kim and Choi, 2020), and football (Mkhallati et al., 2023; Qi et al., 2023), where only videos provide detailed information for individual game moments. In these studies, automatic video captioning is used to generate descriptions of actions within the video. Meanwhile, some studies have leveraged esports match videos to perform play-by-play commentary generation and build visual data-to-text datasets for League of Legends (LoL-V2T (Tanaka and Simo-Serra, 2021)) and for Asseto Corsa, a racing game (Ishigaki et al., 2021). However, the LoL-V2T dataset does not include structured game data, despite the fact that such data is a valuable source for commentary generation. Although the Assetto Corsa dataset contains structured data, it focuses on a racing game, and the structured data contains relatively simple numerical metadata.

In this study, we focus on generating commentary for a specific esports moment in the multiplayer strategy game League of Legends. Unlike previous work on Dota 2 (Zhang et al., 2022) and LoL (Tanaka and Simo-Serra, 2021; Wang and Yoshinaga, 2024), we leverage both a structured data record and a screenshot (a single video frame) to combine the strengths of these modalities. Our proposed task and dataset offer a valuable benchmark for evaluating the capabilities of recent large vision-language models (VLMs), particularly in generating on-demand commentary for a specific gameplay moment.

Event type	Number	(Proportion)	# keys	Explanation
ITEM_PURCHASED	828	(24.0%)	3	The player purchases an item
ITEM_SOLD	60	(1.7%)	3	The player sells an item
ITEM_UNDO	20	(0.6%)	4	The player cancels the purchase of an item
ITEM_DESTROYED	704	(20.4%)	3	The player destroys an item
BUILDING_KILL	52	(1.5%)	8	The team destroys an enemy building
CHAMPION_KILL	100	(2.9%)	5	The player defeats an enemy champion
SKILL_LEVEL_UP	511	(14.8%)	4	The player upgrades a skill
WARD_PLACED	822	(23.8%)	4	The player places a ward
WARD_KILL	328	(9.5%)	3	The player destroys an enemy ward
ELITE_MONSTER_KILL	25	(0.7%)	4 or 5	The team defeats an elite monster
Total	3,450	(100.0%)		

Table 1: Statistics of all ten types of game events in our LoL19-trimodal dataset.

3 Tri-modal Commentary Generation Dataset for Esports Moments

We have built a multimodal commentary generation dataset for esports moments in League of Legends (LoL), one of the most popular multiplayer strategy games. LoL is featured as medal events in 2022 and subsequent Asian Games (Lam and Kuen Wong, 2024; Hong and Yi, 2024), highlighting its popularity and the practical value of developing a dataset.

We begin by introducing LoL basics, and next describe our methods for extracting multimodal game data with accompanying commentaries. Then, we explain how we align these data across modalities to obtain moment-commentary triples. Finally, we compare our dataset to existing benchmarks.

3.1 Basics of LoL Data

In LoL games,[1] two five-player teams compete on a game map. Each player controls a unique character called a "champion," with distinct abilities that grow stronger over time. The primary objective is to destroy the opposing team's base, called as "Nexus," while defending one's own. Throughout the game, players earn resources by defeating enemy champions, neutral monsters, and destroying enemy structures. These resources can be used to purchase items and upgrade champions' abilities, enhancing their impact on team strategy. Matches typically last from fifteen minutes to over an hour, depending on game pace, patch updates, and strategies (Claypool et al., 2017).

Compared to physical sports and board games, LoL presents greater complexity due to real-time actions, rules specific to digital games, and multimodal challenges. These factors pose key obstacles to game commentary generation.

3.2 Extracting Game Data and Commentaries

Following the existing studies on LoL commentary generation (Tanaka and Simo-Serra, 2021; Wang and Yoshinaga, 2024), we construct our dataset using matches from the highest-level tournament in the 2019 League of Legends World Championship. We adopt the same approach as the LoL19 dataset (Wang and Yoshinaga, 2024) to obtain the structured data and commentaries, extending their script[2] to additionally extract screenshots.

Specifically, we collect game logs via the Riot Games API of LoL[3] as structured data and screenshots (video frames) from YouTube game videos as additional input, while using subtitles as commentaries. These multimodal data are paired with the match IDs provided by the contest's Match History site.[4] Below, we detail these procedures.

Retrieving game data records via API The LoL games from the 2019 World Championship[5] record every player action, accessible via the official Riot Games API.[3] Using this data, we can reconstruct complete games. However, the full logs contain redundant information, resulting in large data volumes strain storage and processing resources.

To mitigate this issue, following LoL19 (Wang and Yoshinaga, 2024), we use the "event-based data frame" provided by the official API. In this data frame, key actions are recorded based on discrete events; Figure 1 shows an example of a game event, and Table 1 presents an overview of various

[1] All references to LoL in this paper refer to its main game mode, Summoner's Rift; other game modes are not used.

[2] https://github.com/ArnoZWang/esports-data-to-text

[3] https://developer.riotgames.com/

[4] https://lol.fandom.com/wiki/2019_Season_World_Championship/Match_History

[5] The match history site changed its method of counting games, which reduced the total from 220 to 120, even though the set of matches remained the same. One match was excluded due to data corruption, resulting in 119 valid matches.

Name	**Domain**	**Input Modal**	**# Games**	**# Examples**
RotoWire (Wiseman et al., 2017)	basketball	structured data	4,853	4,853
GameKnot (Jhamtani et al., 2018)	chess	structured data (game logs)	11,578	298,008
Dota2-Commentary (Zhang et al., 2022)	esports	structured data (game meta-data)	70	7,473
LoL19 (Wang and Yoshinaga, 2024)	esports	structured data (game logs)	220	3,490
LoL-V2T (Tanaka and Simo-Serra, 2021)	esports	video	157	9,723
LoL19-trimodal (ours)	esports	image, structured data (game logs)	120	3,450
Assetto Corsa (Ishigaki et al., 2021)	racing game	structured data, text, video	1,389	129,226
ScienceQA (Lu et al., 2022)	science	image, text	N/A	6,532

Table 2: Statistics of our LoL19-trimodal dataset and existing datasets for related generation tasks.

types of events in the collected data. Each event represents an update to the game state and includes a key named "type" indicating the event type. Different event types contain different sets of keys. The event-based data frames are stored in JSON format, as shown in Figure 1.

As in the LoL19 dataset, we linearize the structured input to improve consistency. For each key-value pair in the top-level list of each event, we recursively concatenate the value and the key using delimiter "|" and insert spaces between these key-value pairs. For example, the "BUILDING_KILL" event in Figure 1 is linearized as follows:

```
BUILDINGKILL|type 839426|timestamp
x4318y13875|position 4|killerId
|assistingParticipantIds 200|teamId
TOWERBUILDING|buildingType
TOPLANE|laneType OUTERTURRET|towerType
```

Collecting screenshots from game videos We collect screenshots from YouTube videos for the same matches as those used to obtain the structured data records. Specifically, we capture all screenshots (video frames) in PNG format at the same timestamps as the events in the structured data, resulting in paired data across two modalities.

Gathering commentaries from game videos Following LoL19 and LoL-V2T, we gather subtitles from YouTube contest videos. The subtitles are then split into utterances using line breaks given by YouTube's automatic speech recognition (ASR) results as cues, each of which utterances is regarded as one commentary segment. To verify data quality, we follow the approach from LoL19 (Wang and Yoshinaga, 2024) by randomly sampling 200 examples. The resulting word error rate (WER) is 7.2, which is comparable to the WERs[6] observed in human transcriptions from standard ASR datasets, confirming that the data is sufficiently clean for use. To further enhance quality, we applied simple LLM-based post-processing to recover missing punctuations and correct grammatical errors, as detailed in Appendix A.

3.3 Aligning Game Data and Commentaries

After obtaining paired game data (a structured data record and a screenshot for each event) and commentary segments, we align the inputs and outputs based on timestamps, resulting in a total of 171,460 examples from the 2019 League of Legends World Championship. Since each game contains hundreds to thousands of events, we apply random sampling during alignment to keep the dataset compact and suitable for efficient benchmarking:

Step 1: Divide the paired game data by timestamps into one-minute intervals, following LoL-V2T (Tanaka and Simo-Serra, 2021) and LoL19 (Wang and Yoshinaga, 2024), and randomly select one data point per interval.

Step 2: For each selected game data from Step 1, extract the commentary segment that begins immediately after its timestamp as the corresponding commentary, forming a single example (a multimodal triple) for our task.

This procedure yields approximately 30 examples per game, resulting in 3450 examples. To ensure the quality of timestamp-based alignment, we randomly selected 100 examples and manually verified whether the commentaries corresponded to the selected game data. Among them, 97% (97 out of 100) exhibited satisfactory alignment, with the commentaries accurately describing the corresponding game data; the remaining three showed minor misalignment but were still suitable for use.

Statistics Table 2 lists the statistics of our dataset and several existing datasets for related text generation tasks. The average number of words in

[6] https://github.com/syhw/wer_are_we

commentaries is 15.5. Our dataset is the first trimodal commentary generation dataset for multiplayer strategy esports, consisting of three modalities including images, structured data, and text. It contains a comparable number of examples to existing esports data-to-text datasets. Note that although our dataset is based on similar sources as LoL19 and LoL-V2T, it is neither a subset nor a superset of them. To support the research community while respecting copyright constraints, we take the common approach of releasing scripts for collecting and processing the data to reproduce our dataset.

4 Experiments

In this section, we conduct experiments on commentary generation for esports moments from multimodal game data using our LoL19-trimodal dataset. We evaluate the performance of open- and closed-source large vision-language models using automatic metrics and human judgments.

4.1 Settings

Dataset We split our LoL19-trimodal dataset into training, validation, and test sets with an approximate ratio of 8:1:1, resulting in 2750:350:350 examples.[7] The split follows the natural chronological order of the games.

Models We evaluate five vision-language models (**VLMs**) on our task. Three open-source models are Llama2-7B and Llama2-13B combined with BLIP-2 vision encoder (**Llama2-7B + BLIP-2** and **Llama2-13B + BLIP-2**) and **Qwen2.5-VL-7B**. Two closed-source models are **GPT-4o** and **Gemini-2.0-Flash**. For the Llama2 models, following the implementation used in MiniGPT-4 (Zhu et al., 2024), we use the pre-trained vision components of BLIP-2 (Li et al., 2023) as the vision encoder and Llama2 (Touvron et al., 2023) as the text decoder, which are fine-tuned on the training split using QLoRA (Dettmers et al., 2023), while Qwen2.5-VL-7B[8] and the closed-source models are evaluated under zero-shot settings. Refer to Appendix B for links to the five VLMs, training details, including hyperparameters, optimizer, and QLoRA settings for the open-source VLMs, as well as the task instructions for the closed-source models.

We also conducted an ablation study using several selected models by providing single modal inputs (§ 5.1). Additionally, **T5**[9] was used to evaluate generation solely from structured data inputs.

4.2 Evaluation Procedure

We use both automatic metrics and human judgments to evaluate the VLMs on our task. For automatic metrics, we adopt existing reference-based metrics commonly used in game data-to-text generation tasks. Given the unique characteristics of multiplayer strategy esports, we design three evaluation criteria for human judgments to comprehensively assess system outputs against the references.

Automatic metrics We adopt the automatic metrics used in recent data-to-text generation studies (Liu et al., 2025; Long et al., 2025; Zhang et al., 2024; Alonso and Agirre, 2024; Saha et al., 2023), especially in commentary generation for LoL19 (Wang and Yoshinaga, 2024), which built upon common metrics from related studies (§ 2). The metrics are **BERTScore** (Zhang et al., 2020),[10] **BARTScore** (Yuan et al., 2021),[11] **sacreBLEU** (Papineni et al., 2002; Post, 2018),[12] and **text distance** (normalized Damerau-Levenshtein (Brill and Moore, 2000)).[13] Although these metrics correlated with human judgments on strategic depth in the experiments on the LoL19 dataset (Wang and Yoshinaga, 2024), the latter two metrics have been reported to correlate poorly with human relevance scores for text generation tasks (Wiseman et al., 2017; Novikova et al., 2017). Therefore, we exclude evaluation results with these metrics from the main results; refer to Appendix D for the results.

Human judgments against reference The automatic metrics capture only a limited aspect of the quality of the system outputs. Given the unique characteristics of multiplayer strategy esports, we need diverse evaluation criteria that highlight different aspects of commentary quality. For instance, LoL19 (Wang and Yoshinaga, 2024) introduces strategic depth to assess the strategically relevant content in the system outputs. GameKnot (Jhamtani et al., 2018) conducts a human evaluation to measure correctness, relevance, and fluency.

[7] We manually check data alignment for all 350 examples in the test set, confirming there are no notable misalignments.

[8] We were unable to successfully fine-tune Qwen2.5-VL-7B, likely due to limitations in the current training setup.

[9] https://huggingface.co/t5-base

[10] https://pypi.org/project/bert-score

[11] https://github.com/neulab/BARTScore

[12] https://github.com/mjpost/sacrebleu

[13] https://github.com/life4/textdistance

Models	Human judgments against reference			Automatic metrics	
	Relevance	**Inspiration**	**Interestingness**	**BERTScore**	**BARTScore**
T5 (no visual input)	36.50	13.00	11.25	78.60	-5.60
Llama2-7B + BLIP-2	36.50	29.75	15.00	79.62	-5.01
Llama2-13B + BLIP-2	**47.75**	35.00	**20.00**	**81.19**	**-4.80**
w/o visual input	43.50	31.50	19.50	80.07	-4.92
w/o data record	40.00	32.50	19.00	80.10	-4.86
zero-shot settings					
Qwen2.5-VL-7B	39.50	30.00	17.50	79.87	-4.95
Gemini-2.0-Flash	39.50	30.00	20.50	80.01	-4.82
GPT-4o	41.00	**35.00**	20.75	80.10	-4.97
w/o visual input	40.25	29.75	18.50	80.05	-5.01
w/o data record	39.00	34.50	20.25	80.01	-5.00

Table 3: Results of commentary generation for esports moments with the LoL19-trimodal dataset: the best results are in **bold** and the second best results (except ablated models) are underlined.

In this study, we deconstruct the concept of strategic depth in LoL19 and elaborate on it through the following three criteria, each highlighting concrete aspects of moment-level commentary:

Relevance assesses how well the system output aligns with the input game data in terms of factuality. It verifies whether the content faithfully reflects the information in the input. High relevance ensures that the output is factual and consistent with the source data.

Inspiration evaluates the ability of the system output to provide creative insight. Commentary that analyzes of the game situation or a summary of the players' actions and intentions is considered more inspiring. High inspiration can enhance the content's usefulness in understanding the strategies employed by professional players.

Interestingness focuses on how captivating and engaging the system output is to the audience. It considers elements like clarity, narrative flow, and rhetorical devices. Interestingness often reflects the balance between creativity and readability, making the commentary appealing and enjoyable to watch.

We recruit three graduate students as human judges. They have substantial experience with our target game, League of Legends, and can readily comprehend its game commentaries in English. Each judge has watched at least three full seasons of the World Championship (especially, including the 2019 event), and have played more than 100 matches in the season preceding the annotation. All judges have previously reached at least *Platinum* rank, corresponding to approximately the top 30% of players. Additionally, an LLM-based agent (GPT-4o) acts as the fourth annotator to provide a consistent and scalable baseline for comparison across examples, and to explore the feasibility of integrating automated judgment into a more efficient evaluation pipeline.

The evaluation follows a comparison-based protocol. For 200 randomly selected examples, four annotators compared each system output with its corresponding reference and chose the better one for each aspect (refer to Appendix C for the complete instructions used for human evaluation). We report the percentage of times the system outputs were preferred, averaged across annotators. Scores range from 0 to 100, where 50 indicates parity with the reference.

4.3 Results

Table 3 lists the evaluation results of the five VLMs on our LoL19-trimodal dataset (the evaluation results of sacreBLEU and text distance are presented in Appendix D for reference only). The fine-tuned Llama2-13B + BLIP-2 outperformed other models including strong zero-shot VLMs, partially validating the usefulness of our dataset for training purposes. The VLMs achieve comparable performance to professional commentary in terms of relevance and struggle to compete in terms of inspiration, while they fall short when it comes to interestingness. Closed-source VLMs outperform fine-tuned open-source models in inspiration. In addition, the comparison between Llama2-7B + BLIP-2 and -13B revealed that larger VLMs are relatively effective.

Notably, the results from automatic metrics mostly align with human evaluations in terms of relevance. However, for the other two aspects, inspiration and interestingness, the automatic metrics show less consistency, indicating the importance of human judgments.

5 Analysis

In this section, we provide a more in-depth analysis of the results and discuss the remaining challenges associated with this task. In what follows, we focus on analyzing the results of two competing VLMs, including Llama2-7B + BLIP-2 and GPT-4o.

5.1 Ablation Study

To assess the impact of each input modality on commentary generation, we perform ablation studies which removes one modality from the input: visual input is replaced with a blank image, and structured data with a placeholder token. All model variants are trained and evaluated under the same LoL19-trimodal setup to ensure comparability.

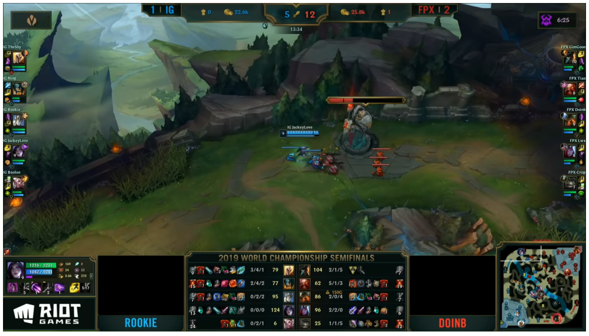
Table 3 presents the ablation results. Incorporating bi-modal inputs consistently improves performance across all evaluation metrics, demonstrating the benefits of multimodal integration for esports commentary generation. A closer examination of the human evaluation results reveals the distinct contributions of each modality: structured data plays a more crucial role in generating relevant commentary, whereas visual input contributes more to producing inspiring commentary.

Furthermore, we conduct a McNemar's test (for human evaluations) and a Wilcoxon signed-rank test (for automatic metrics) to assess the statistical significance of the difference between the Llama2-13B + BLIP-2 model with and without visual input. The resulting p-values for the three human evaluation metrics, including **Relevance**, **Inspiration**, and **Interestingness**, are 0.008, 0.008, and 0.052, respectively; the results for the automatic metrics, including **BERTScore** and **BARTScore**, are 0.016 and 0.040, respectively. Most results indicate statistically significant differences ($p < 0.05$), with several demonstrating even stronger significance ($p < 0.01$). These findings underscore the value of our tri-modal dataset and evaluation criteria.

5.2 Examples of System Outputs

Figure 2 shows an example of system outputs for the same gameplay moment as shown in Figure 1, regarding a BUILDING_KILL event. In this exapmle, the system outputs correctly describe the core event where a player is taking an enemy building down on the top lane of the game map. However, several issues are remaining, mainly including unclear expressions (**Llama2-13B + BLIP-2**: "*pushing too far forward*"), unusual terminology in natural language (**Llama2-13B + BLIP-2**: "*Building Kill*", **GPT-4o**: "*BUILDINGKILL*"), and low-information statements (**GPT-4o**: "*At 13:34 in Game 3 of the 2019 World Championship Semifinals ...*"). Additionally, it is worth noting that the systems tend to produce overly long yet insufficiently informative commentary. To generate more professional and concise outputs, we may need to increase the amount of training data or provide more detailed instructions that clearly describe the characteristics of human commentary.

Screenshot of "BUILDING_KILL" event, presenting a player is taking an enemy building down:



Structured data records:

```
{
 "type": "BUILDING_KILL",
 "timestamp": 839426,
 "position": "x4318y13875",
 "killerId": 4,
 "assistingParticipantIds": "",
 "teamId": 200,
 "buildingType": "TOWERBUILDING",
 "laneType": "TOPLANE",
 "towerType": "OUTERTURRET"
}
```

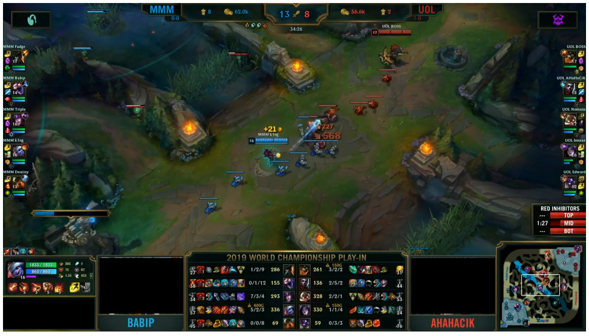
Reference: Fighting everywhere, Jackeylove is just farming for himself.

Llama2-13B + BLIP-2: The IG top laner has taken a risky move by pushing too far forward and getting caught out by the cunning FPX jungler, leading to a costly Building Kill for IG.

GPT-4o: At 13:34 in Game 3 of the 2019 World Championship Semifinals between Invictus Gaming (IG) and FunPlus Phoenix (FPX), IG's JackeyLove pressures the top lane alone and takes down FPX's outer turret, marking a BUILDINGKILL event in the TOPLANE–a small but crucial step toward regaining map control despite IG trailing 5-12 in kills and over 3k gold.

Figure 2: System outputs for BUILDING_KILL event.

Screenshot of “ITEM_PURCHASED” event, presenting a player is purchasing an item:



Structured data records:

```
{
 "type": "ITEM_PURCHASED",
 "timestamp": 2067324,
 "participantId": 7,
 "itemId": 2420
}
```

Reference: I think fudges gangplank has looked generally strong King is hitting close to full item.

Llama2-13B + BLIP-2: MMM player in the middle lane has just purchased an item, at timestamp 7, with participant ID 2420.

GPT-4o: At 34:26 in this 2019 World Championship Play-In match between MAMMOTH and Unicorns of Love, MMM holds a solid lead with an 8k gold advantage and map pressure. A recent ITEMPURCHASED event by participant 2420 (likely MMM King on Kai’Sa) further boosts their late-game power, strengthening their chances of closing out the game as they push mid against a struggling UOL defense.

Figure 3: System outputs for `ITEM_PURCHASED` event.

Figure 3 shows another example regarding an `ITEM_PURCHASED` event. Related types of events typically have only a limited impact on the overall progression of the game and are not usually the main focus of audience attention. As a result, we observe that human commentary tends to offer relatively general descriptions, rather than offering details for such events. This insight leads us to a broader discussion on the performance of generation across event types in the following section.

5.3 Quality Variations across Event Types

Based on our LoL19-trimodal and several prior studies (Zhang et al., 2022; Wang and Yoshinaga, 2024), we observe a dominance of certain event types. For example, “`ITEM_PURCHASED`” accounts for 24.0% of the dataset (Table 1), despite having limited impact on the game progression. Generating commentaries for such frequent yet less important events is challenging for both humans and models. In contrast, rarer but more informative events like “`BUILDING_KILL`” (1.5%) often correspond to pivotal moments that can affect the entire game. To address this imbalance, we further evaluate the model performance across each event type in LoL19-trimodal. Specifically, we compute the percentage of times the system outputs were preferred for examples of each event type, using the method explained in § 4.1.

Event type	Relevance	Inspiration	Interestingness
`ITEM_PURCHASED`	50.00	35.00	21.67
`ITEM_DESTROYED`	48.91	34.78	20.65
`CHAMPION_KILL`	70.83	41.67	29.17
`SKILL_LEVEL_UP`	40.63	35.42	19.80
`WARD_PLACED`	36.36	34.09	25.00
`WARD_KILL`	46.15	34.62	21.15

Table 4: Human judgements on results of Llama2-13B + BLIP-2 on LoL19-trimodal for common event types.

Table 4 lists the evaluation results for examples in common event types. Among these event types, “`CHAMPION_KILL`” stands out as a relatively information-rich and important event, and the Llama2-13B + BLIP-2 successfully generated better commentary than the reference. In the current LoL19-trimodal dataset, the original distribution of event types was preserved through the random sampling process (§ 3.3). However, since the importance of different event types varies, future versions of the dataset can be optimized by first manually specifying key event types (*e.g.*, `BUILDING_KILL`) to retain such events in full, while reducing overwhelming and less important ones.

5.4 Applicability of LLM-as-a-judge

Given the rapid advancement of LLM-as-a-judge evaluation, recent studies have shown that LLMs, such as GPT-4, can serve as reliable proxies for human evaluators across various natural language generation (NLG) tasks (Zheng et al., 2023; Gu et al., 2024). Accordingly, we incorporated an LLM-based agent into the procedure of human judgements to act as an additional annotator (§ 4.2).

To validate the reliability of this approach in our task, we computed both inter-human agreement and overall agreement including both humans and the LLM judge for the outputs of Llama2-13B + BLIP-2, using Fleiss’ κ (Fleiss, 1971). For the three evaluation aspects of **Relevance**, **Inspiration**, and **Interestingness**, the inter-human agreements are 0.492, 0.482, and 0.371, while the overall agreements, including the LLM annotator (GPT-4o), are

0.511, 0.440, and 0.357.[14] These similar scores demonstrate a reasonable alignment between humans and the LLM, supporting the feasibility of using LLM-based assessments as a supplementary evaluation method. In future work, the role of LLMs as evaluators can be further explored and extended to a larger test set.

6 Conclusions

This study set up a new task of commentary generation for esports moments from multimodal game data: structured game data and gameplay screenshots. We collected game data and commentaries from online sources for one of the most popular esports, League of Legends, to construct the first tri-modal commentary generation dataset aligned at the moment level. We also designed evaluation metrics to assess generation performance, reflecting the unique characteristics of esports and the role of game commentary. Using our dataset, we evaluated several strong vision-language models and showed that combining structured and visual data improves generation quality.

For future work, we plan to extend our dataset to other languages such as Chinese, Korean, and European languages, which are widely spoken in regions with large esports audiences. Another important direction is to build a tri-modal commentary dataset for live commentary generation, where the system detects events that require commentary and generates responses in real time.

Limitations

This work addresses commentary generation for esports moments in multiplayer strategy games. While the study introduces a novel dataset using the multimodal data collected from a representative esports game, League of Legends, it does not yet consider other esports genres at this stage.

Moreover, although our task is beneficial for the audience and players to obtain commentary for esports moments on demand, live commentary generation for esports game videos can also go a long way towards improving audience's viewing experience. Although our research does not focus on video-to-text generation at the current stage, our data extraction methods and scripts are capable of collecting dynamic video clips if needed. Alternatively, another potential approach is to use a sequence of game screenshots, where multiple images correspond to a single structured data input, combining the benefits of static images and dynamic videos.

In the end, recent advancements in esports commentary generation have mostly focused on English, despite the fact that esports enjoys a vast and growing global audience, particularly in regions such as China, South Korea, and various European countries (Mangeloja, 2019; Lam and Kuen Wong, 2024). This highlights the lack of multilingual support in esports-related NLG, which hinders model generalizability across languages and cultures. Although our dataset is English-only, commentary in other languages (*e.g.*, Chinese livestreams[15]) is readily available. With the rise of LLM-based machine translation (Xu et al., 2024), integrating multilingual data has become increasingly feasible, paving the way for future cross-lingual research in esports commentary generation.

Ethics Statement

In obtaining data for our LoL datasets, we strictly followed the policies of RiotGames API and YouTube. The former is the publisher of LoL game data records. The latter supplied the screenshots and subtitles of LoL contest videos. To respect copyright constraints, we take the common approach of releasing scripts for collecting and processing the data to reproduce our dataset (Wang and Yoshinaga, 2024; Tanaka and Simo-Serra, 2021).

Further ethical concerns related to the game content (*e.g.*, video game content rating) refer to ESRB Rating.[16] The game is rated as "Teen," indicating it is suitable for players aged 13 and older.

Acknowledgements

This work was partially supported by Google Research Grant with support from `Google.org`.

References

Inigo Alonso and Eneko Agirre. 2024. Automatic logical forms improve fidelity in table-to-text generation. *Expert Systems with Applications*, 238:121869.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 286–293,

[14] The low κ for interestingness may be the subjective nature of this aspect.

[15] `https://lpl.qq.com/`

[16] `https://www.esrb.org/ratings/32211/league-of-legends/`

Hong Kong. Association for Computational Linguistics.

Mark Claypool, Artian Kica, Andrew La Manna, Lindsay O'Donnell, and Tom Paolillo. 2017. On the impact of software patching on gameplay for the league of legends computer game. *The Computer Games Journal*, 6:33–61.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Kirstin Hallmann and Thomas Giel. 2018. eSports – Competitive sports or recreational activity? *Sport management review*, 21(1):14–20.

Jianping Hong and Jiandong Yi. 2024. Esportisation: The inclusion of esports in the hangzhou asian games. In *The Mediating Power of Sport: Global Challenges and Sport Culture in China*, pages 143–162. Emerald Publishing Limited.

International Olympic Committee IOC. 2025. Inaugural olympic esports games to be held in riyadh in 2027 – road to the games to start this year. Accessed: 2025-02-01.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Seth E Jenny, R Douglas Manning, Margaret C Keiper, and Tracy W Olrich. 2017. Virtual(ly) athletes: where esports fit within the definition of "sport". *Quest*, 69(1):1–18.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1671, Melbourne, Australia. Association for Computational Linguistics.

Hirotaka Kameko, Shinsuke Mori, and Yoshimasa Tsuruoka. 2015. Learning a game commentary generator with grounded move expressions. In *2015 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 177–184. IEEE.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 1056–1065, New York, NY, USA. Association for Computing Machinery.

Gigi Lam and Oscar Wai Kuen Wong. 2024. Cross-country comparison of the esports industry in china, south korea and japan. *Sport in Society*, pages 1–26.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. What is that talk about? a video-to-text summarization dataset for scientific presentations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6187–6210, Vienna, Austria. Association for Computational Linguistics.

Zefei Long, Zhenbiao Cao, Wei Chen, and Zhongyu Wei. 2025. EMGLLM: Data-to-text alignment for electromyogram diagnosis generation with medical numerical data encoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20470–20480, Vienna, Austria. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Esa Mangeloja. 2019. Economics of esports. *Electronic Journal of Business Ethics and Organization Studies*, 24(2).

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. SoccerNet-Caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5073–5084.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and Jie Tang. 2023. 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*, CIKM '23, page 5391–5395, New York, NY, USA. Association for Computing Machinery.

Swarnadeep Saha, Xinyan Yu, Mohit Bansal, Ramakanth Pasunuru, and Asli Celikyilmaz. 2023. MURMUR: Modular multi-step reasoning for semi-structured data-to-text generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11069–11090, Toronto, Canada. Association for Computational Linguistics.

Tsunehiko Tanaka and Edgar Simo-Serra. 2021. LoL-V2T: Large-scale esports video description dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4557–4566.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence*, pages 7096–7103.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Zihan Wang and Naoki Yoshinaga. 2024. Commentary generation from data records of multiplayer strategy esports game. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 263–271, Mexico City, Mexico. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *ICML*.

Fei Yan, Krystian Mikolajczyk, and Josef Kittler. 2016. Generating commentaries for tennis videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2658–2663. IEEE.

Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. 2018. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6006–6015.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Dawei Zhang, Sixing Wu, Yao Guo, and Xiangqun Chen. 2022. MOBA-E2C: Generating MOBA game commentaries via capturing highlight events from the meta-data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4545–4556, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haowei Zhang, Shengyun Si, Yilun Zhao, Lujing Xie, Zhijian Xu, Lyuhao Chen, Linyong Nan, Pengcheng Wang, Xiangru Tang, and Arman Cohan. 2024. OpenT2T: An open-source toolkit for table-to-text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–269, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *The eighth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The twelfth International Conference on Learning Representations*.

A Methods for Improving the Quality of Collected Commentaries

Although we have demonstrated that the collected commentaries are of usable quality via the word error rate (WER) metric (§ 3), we applied additional post-processing techniques to further improve their quality. Specifically, we employ GPT-based methods with minor manual corrections to enhance the commentary data, addressing inserting appropriate punctuations and correcting grammatical errors. Below is an example prompt used for this process:

Example prompt for improving the quality of collected commentaries

```
You are a knowledgeable expert in the field
of esports game League of Legends (LoL) and
its contests.

We would like you to read a commentary on
a piece of an LoL game.  Please review
the following commentary and improve its
quality by adding appropriate punctuations
and correct grammatical errors. Only output
the improved result in one single paragraph,
without explanatory sentences and quotation
marks.
```

For example, in Figure 1, the original commentary sentence collected from subtitles of game videos is "*fighting everywhere jackeylove is just farming for himself*". After post-processing, it becomes "*Fighting everywhere, Jackeylove is just farming for himself.*" as shown in Figure 1. We will release scripts for obtaining both styles of the game commentaries.

B Training Details

For the VLMs used in our experiments, we adopt the same settings as in LoL19 (Wang and Yoshinaga, 2024) and MiniGPT-4 (Zhu et al., 2024). Specifically, for T5,[17] we set the decoder dropout to 0.5, the number of training steps to 10,000, and the learning rate to 0.001; it applies Adam optimizer ($\beta 1 = 0.9$, $\beta 2 = 0.999$); all other hyperparameters follow their default settings. For Llama2-7B + BLIP-2[18] and -13B,[19] the training is based on the first-stage outcomes of MiniGPT-4; the pretrained vision encoder and LLM remain frozen, with only the linear projection layer undergoing training; Llama2 applies QLoRA (Dettmers et al., 2023) fine-tuning with 4-bit precision and a LoRA dropout of 0.1.

For Qwen2.5-VL-7B[20] and close-sourced models (GPT-4o[21] and Gemini-2.0-Flash[22]), we employ zero-shot prompting using the prompt as follows:

Prompt for generating commentaries

```
You are a knowledgeable expert in the field
of esports game League of Legends (LoL) and
its contests.

This  image  is  a  screenshot  showing  a
moment of a game in the 2019 Season World
Championship of League of Legends (LoL).

The following is the structured data record
that corresponds to the game moment shown in
this screenshot: [insert input structured
data here].

Based on the screenshot and the data record,
write a short commentary to describe them
within one or two sentences.
```

C Human Scoring Instructions

The instructions used for human scoring (§ 4.2) is as below:

Human scoring instructions

```
The    definition    of    [Relevance]    /
[Inspiration] / [Interestingness] is:

[It assesses how well the generated result
aligns with the input data.  It measures
whether  the  content  faithfully  reflects
the  information  provided  by  the  input.
High  relevance  ensures  that  the  output
is factual and consistent with the source
data.] /

[It evaluates the ability of the generated
result  to  provide  creative  insights.
Specifically,  commentary  that  includes
an  analysis  of  the  game  situation  or
a  summary  of  the  players'  actions  and
intentions  is  considered  more  inspiring.
High inspiration can enhance the content's
usefulness in understanding the strategies
employed by professional players.] /
```

[17] https://huggingface.co/google-t5/t5-base

[18] https://huggingface.co/meta-llama/Llama-2-7b

[19] https://huggingface.co/meta-llama/Llama-2-13b

[20] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

[21] https://chatgpt.com

[22] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

```
[It focuses on how captivating and engaging
the generated result is to the audience.
It considers elements like clarity,
narrative flow, and rhetorical devices.
Interestingness often reflects the balance
between creativity and readability, making
the commentary appealing and enjoyable to
watch.]

This image is a screenshot showing a
moment of a game in the 2019 Season World
Championship of League of Legends (LoL).

The following is the structured data record
that corresponds to the game moment shown
in this screenshot: [structured data].

Based on the screenshot and the data record,
there are two commentaries describing them:
[commentary A], [commentary B].

According to the definition of [Relevance]
/ [Inspiration] / [Interestingness], choose
which commentary has a higher [Relevance]
/ [Inspiration] / [Interestingness].
```

D Automatic Metrics

Models	sacreBLEU	Text distance ↓
T5 (no visual input)	3.2	74.60
Llama2-7B + BLIP-2	5.1	70.01
Llama2-13B + BLIP-2	**10.2**	**68.85**
w/o visual input	9.1	69.97
w/o data record	9.0	69.10
zero-shot settings		
Qwen2.5-VL-7B	5.9	72.00
Gemini-2.0-Flash	8.2	70.50
GPT-4o	6.2	69.37
w/o visual input	5.8	69.43
w/o data record	6.1	69.40

Table 5: Results of commentary generation for esports moments with the LoL19-trimodal dataset over sacreBLEU and text distance: the best results are in **bold** and the second best results (except ablated models) are underlined.

Metrics	Relevance	Inspiration	Interestingness
BERTScore	0.431	0.418	0.301
BARTScore	0.408	0.355	0.294
sacreBLEU	0.306	0.274	0.217
text distance	0.289	0.288	0.176
(inter-human)	0.492	0.482	0.371

Table 6: Results of agreement between human ratings and automatic metrics. The results are gathered by comparing the pairwise ranking of the same outputs, as determined by each metric, using Fleiss' κ.

Table 5 lists the evaluation results over sacreBLEU and text distance. As discussed in § 4.2, they are provided for reference only and are not primary metrics.

Furthermore, following § 5.4, we evaluate the agreement between human ratings and automatic metrics by comparing the pairwise ranking of the same outputs, as determined by each metric, using the outputs of Llama2-13B + BLIP-2. Table 6 lists the results. As can be observed from the results, the values of sacreBLEU and text distance are relatively low. This finding is consistent with the conclusion drawn in § 4.2, namely that automatic metrics such as sacreBLEU and text distance have been found correlate poorly with human ratings for language generation tasks (Wiseman et al., 2017; Novikova et al., 2017).