AI Powered Commentary and Camera Direction in E-Sports

Swathi Jamjala Narayanan*, Kevin Winston Joseph, Devansh Sirohi, Harsh Chaudhary, Hitesh Shivkumar

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore (India)

* Corresponding author: jnswathi@vit.ac.in

Received 6 May 2024 | Accepted 15 July 2025 | Published 19 February 2026

Abstract

Real-time, AI-driven commentary and camera direction provide revolutionary possibilities to improve spectator engagement and comprehension of live events in the rapidly advancing world of e-sports. This paper proposes an autonomous system designed to both generate dynamic commentary as well as control the spectator camera for live-streamed e-sports matches, specifically focusing on League of Legends (LoL), a popular Multiplayer Online Battle Arena (MOBA) game. It incorporates the use of GPT-4o with Vision and OpenAI's TTS API. Synchronization of commentary with real-time camera movements is one of the major challenges tackled. This is done using a camera tracking and scene change detection algorithm that effectively adjusts the commentary to changing scenes in real-time by utilizing computer vision techniques. Further, two neural architectures for AI-driven camera control: a 2D Convolutional-LSTM (Conv-LSTM) model that concentrates on independent spatial and temporal analysis, and a 3D CNN model that combines these features to forecast camera movements in a more comprehensive way are presented. Evaluations on fluency, relevance, and strategic depth metrics, show that our integrated system improves viewer experience by providing deep and coherent narratives that are contextually aligned with the game dynamics. The proposed models are evaluated quantitatively in capturing spectator camera movement patterns.

Keywords

AI-Driven Commentary, Camera Control, Computer Vision, E--sports Analytics, Neural Architectures.

DOI: 10.9781/ijimai.2026.6566

I. Introduction

As elucidated in [1], digital prowess and virtual battles have ushered in a new era of competitive entertainment, which has undergone a seismic shift in recent years. The traditional competitive gaming landscape has been upended by the rise of e-sports, or electronic sports, which have captured the attention of millions of people worldwide. In contrast to traditional sports, which are played on fields or courts, e-sports utilize the power of digital platforms. Participants compete fiercely in a wide range of video and computer games across different genres, including League of Legends and Counter-Strike: Global Offensive. From being thought of as a niche activity, it has developed into a multi-billion dollar industry that challenges traditional sports' dominance and captivates audiences with its unique combination of skill, strategy, and spectacle.

E-Sports is gaining popularity at a rate that surpasses generational and cultural divides, attracting a wide range of fans, players, and spectators. Millions of people tune in to watch the fierce bouts take place on virtual arenas, despite the fact that older generations may find it difficult to understand why virtual battles are so appealing. Furthermore, it is impossible to overstate the economic impact of e-sports, as their earnings surpass not just the combined earnings of traditional entertainment industries like music and film, but also rival them. The dynamics of this emerging industry become more evident upon digging deeper, demonstrating that e-sports is a cultural phenomenon that is here to stay and is drastically changing the competitive entertainment landscape.

The real-time generation of engaging commentary and intelligent camera control present significant technical challenges in e-sports broadcasting. Commentary requires understanding complex game states, strategic implications, and generating natural language in real-time. Meanwhile, camera control demands rapid identification of important game events across multiple locations. These traditionally human-operated tasks are increasingly difficult to scale with the growing e-sports industry. Recent advances in computer vision and large language models offer promising solutions - vision models can track game state and identify key moments, while generative AI can produce contextual commentary. When combined, these technologies could potentially match or exceed human capabilities in capturing the dynamic nature of e-sports competitions.

Please cite this article as:
S. J. Narayanan, K. W. Joseph, D. Sirohi, H. Chaudhary, H. Shivkumar. AI Powered Commentary and Camera Direction in E-Sports, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 6, pp. 116-125, 2026, http://doi.org/10.9781/ijimai.2026.6566

League of Legends (LoL) is a Multiplayer Online Battle Arena (MOBA) video game developed and published by Riot Games. Teams of players compete against one another in LoL's fast-paced, strategic contests, which take place in the mythical realm of Runeterra. Every player takes control of a different champion, a character with special skills and roles. Depending on the game mode, the goal is different, but usually it entails taking out the Nexus, the base structure of the other team. Players maneuver their champions, engage in combat with enemy minions and structures, and work together with their allies to secure objectives and outmaneuver opponents on a lanes-divided map throughout matches. Attracting millions of players and viewers worldwide, League of Legends has grown to become one of the most well-known and significant games in the e-sports sector because to its rich competitive scene, vast roster of champions, and complex gameplay mechanics. Our focus on LoL is for these reasons.

This paper proposes an intelligent system that leverages deep learning and computer vision to automatically interpret game events, control camera positioning to capture key moments, and generate engaging commentary. By combining convolutional neural networks for spatial understanding, LSTM networks for temporal dynamics, and large language models for natural commentary generation, we create an end-to-end solution for automated e-sports broadcasting.

Our method and its implications in detail are delineated in the parts that follow. Our approach is formulated after a thorough review of the work that has already been done on AI in e-sports commentary and camera control, as shown in Section II. The experimental setup, proposed system designs, neural networks trained, and the algorithm developed for scene change detection and camera tracking are all detailed in Section III. The performance of our suggested models and the effectiveness of our real-time system are highlighted in Section IV, including quantitative and qualitative findings of our experiments. In V, results are interpreted and further contextualized, limitations discussed, possible future directions suggested. Lastly, Section VI provides a summary of our research and highlights the ways in which AI-driven technologies can improve the viewer experience in live e-sports events.

II. Review of Literature

A. E-Sports AI Commentary

As the e-sports industry sees rapid growth, there is an increase in research efforts that focus on enhancing the spectator and commentator experience with the help of AI. [2] introduces an automated system for generating video game commentary, focusing on League of Legends. Utilizing generative AI, their system recognizes in-game events and produces vocal commentary, potentially easing the cognitive load for streamers. The system employs models for event detection and ChatGPT for commentary generation, with voices rendered via FakeYou. Preliminary evaluations show promise in accuracy and qualitative reception of generated commentary, highlighting the system's potential in enhancing streaming and spectator experiences in e-sports. [3] focuses on generating commentary for e-sports events by leveraging a large-scale data-to-text dataset. Their approach, aimed at enhancing the viewing experience for League of Legends spectators, involves sophisticated models that interpret structured game data to produce insightful and contextually relevant commentary. They highlight the potential of their system to provide a deeper, more engaging spectator experience, though recognizing the complexities and challenges in perfecting this AI-driven commentary generation.

[4] present "MCS," an in-battle commentary system for Multiplayer Online Battle Arena (MOBA) games, leveraging match statistics for real-time, coherent commentary generation. The system utilizes a GPT-2-based neural pipeline, integrating hero selection loss and restricted token decoding for accurate, dynamic commentary, showcasing promising results in enhancing the gaming experience and audience comprehension. Further, [5] examines the use of e-sports caster commentaries for generating text-based match recaps, emphasizing the use of Natural Language Generation (NLG) systems. The research explores user perceptions of NLG-generated text as accurate recaps of game highlights and assesses its support for viewer understanding of matches. The research offers insights into the potential of text-based recaps in enhancing spectator experiences, particularly in Dota 2, and discusses the implications for future NLG applications in e-sports.

[6] addresses the challenge of sentence punctuation in collaborative commentary generation for e-sports live-streaming. They introduce two strategies to improve the coherence of AI-generated commentary by optimizing sentence punctuation. Utilizing the Text-to-Text Transfer Transformer (T5) model, they demonstrate that their approach enhances the alignment of AI-generated commentary with the flow of the game, surpassing baseline methods in both objective metrics and subjective assessments. [7] Introduces a pioneering approach for generating racing game commentary, integrating vision, language, and structured data. Their method addresses the intricate task of aligning commentary with live game-play, posing a significant challenge due to the dynamic nature of racing games. The research contributes a novel dataset and baseline models, underscoring the complexity and potential of multi-modal data integration for real-time e-sports commentary generation.

[8] Delves into the potential of LLMs, specifically ChatGPT, in generating insightful commentary for fighting games, focusing on the DareFightingICE platform. They highlight the influence of various prompt components on the quality of generated commentary, revealing that certain exclusions enhance readability and vocabulary diversity, marking a significant stride in AI-driven game commentary generation.

B. Traditional Sports AI Commentary

In sports, particularly football, efforts have been made to create a robust, live AI commentary system. Particularly, [9] explores AI-generated live football game commentary using GPT-3.5. The system, trained on pre-written commentary templates and game state data, aims to offer dynamic and accurate narration, enhancing the gaming experience. Despite some inaccuracies and repetition, player feedback indicates a generally positive reception, highlighting the potential of AI in live sports commentary. Further improvements and the integration of advanced models like GPT-4 are suggested to enhance accuracy and variety.

[10] delves into generating live soccer-match commentary from play data, emphasizing the need for alignment between commentary and events, categorization of data attributes, and named entity handling. The research employs an encoder-decoder model with a gate mechanism and placeholder reconstruction for accurate name generation. Despite achieving promising results, challenges in event-commentary alignment and named entity identification persist, highlighting areas for future improvement. [11] introduce a sophisticated soccer game commentary system using Mixed Spatial and Temporal Attention. This approach integrates local and global attention mechanisms with temporal grouping to provide precise commentary in real-time. The system's effectiveness is demonstrated through rigorous testing, indicating its potential to revolutionize automated sports commentary by offering accurate, real-time insights during soccer matches.

C. General Purpose Video Comment Generation

Besides field-specific applications, efforts have been made to generate commentary in a general-purpose, domain-agnostic and

multi-modal fashion. [12] propose an open-domain approach to generating live video commentaries, extending beyond specific fields. They crafted a large dataset, analyzed current methodologies, and identified the essential trade-offs between textual and visual inputs. Their findings highlight the significance of external knowledge in this context, providing a foundation for future research in multimodal commentary generation. Another research work [13] presents "LiveBot," an innovative system for generating live video comments based on both visual and textual contexts. The system utilizes two neural network models to process video content and associated text, aiming to produce contextually relevant live comments. The research demonstrates the model's effectiveness through various performance metrics, highlighting its potential in enriching viewer interaction and engagement in live video streaming platforms.

[14] Introduces a unified Multi-Task approach for generating and positioning live video comments (danmu). Their novel framework combines a multi-modal fusion Transformer architecture, harnessing video, subtitles, audio, and existing comments to simultaneously train two subsystems: comment generation and density prediction. This end-to-end approach outperforms single-task baselines, indicating the mutual benefit of modeling danmu distribution and comment generation together. [15] focuses on enriching live video comments by integrating external knowledge. The model uses pre-trained encoders and decoders to assimilate diverse data sources, aiming to create more informative and contextually relevant comments. Their approach demonstrates notable improvements over traditional models, emphasizing the importance of external knowledge in generating meaningful live video commentary.

[16] Presents a novel model for generating live video comments, considering both the surrounding frames and existing live comments. This approach aims to produce human-like comments by capturing the essence of the video and the audience's input, achieving notable improvements over baseline models. The paper emphasizes the significance of integrating visual and textual contexts for a richer, more interactive live-streaming experience. [17] explores the generation of highlight-based bullet comments in live-streaming games to enhance audience mental well-being. The system utilizes game highlights and audience interactions, aiming to foster a positive viewing experience. Preliminary tests on a fighting game live-streamed on Twitch show the system's potential in reducing negative viewer emotions, promising a new avenue for improving mental health through interactive gaming.

D. Advancements in Video Processing Models

Recent advancements in computer vision have been largely driven by the successful adaptation of Transformer architectures to visual tasks. [18] demonstrated this potential with Vision Transformers (ViT), showing that the architecture could match or exceed state-of-the-art CNNs on image recognition tasks when pre-trained on sufficient data, despite lacking traditional CNN inductive biases. This breakthrough laid the groundwork for applying Transformer architectures to more complex visual tasks, particularly video understanding. Building on this foundation, [19] introduced TimeSformer, which effectively adapted Transformers to video tasks by separating spatial and temporal attention operations. Their "divided attention" approach achieved state-of-the-art results on major benchmarks while significantly reducing computational requirements compared to traditional 3D CNNs.

The field has continued to evolve with several innovative approaches to handling longer temporal sequences and improving semantic understanding. [20] developed MeMViT, addressing the limitation of short temporal windows in video understanding through a memory-augmented architecture. By processing videos sequentially and caching memory representations, MeMViT achieved 30x longer temporal support with minimal computational overhead, marking a significant advance in long-form video understanding. Similarly, [21] introduced MVP, enhancing visual pre-training by incorporating multimodal information through CLIP's vision branch, demonstrating substantial improvements in both classification and semantic segmentation tasks. This multimodal approach was further complemented by [22] VideoMAE, which adapted masked autoencoders for video pre-training. Their innovative tube masking strategy and high masking ratio (90-95%) leveraged temporal redundancy in video data, achieving state-of-the-art results even with limited training data.

These developments collectively represent a significant shift in video processing approaches, moving from traditional convolutional architectures to more flexible and efficient Transformer-based models that can better handle temporal relationships and semantic understanding. The success of these models across various benchmarks suggests that Transformer-based architectures, combined with innovative pre-training strategies and memory mechanisms, are becoming the dominant paradigm in video understanding tasks.

III. Methods

A. Experimental Setup

This paper proposes a real-time commentary and camera direction system for e-sports using a variety of software tools and frameworks. OpenCV was used to capture the live game feed. FFmpeg was then used to process generated audio and overlay the commentary onto the feed. To achieve realistic content and voice synthesis, the commentary was produced using OpenAI's GPT-4o and then transformed into spoken audio using the Elevenlabs TTS API.

Custom algorithms for spectator camera tracking and scene change detection are proposed. The dataset used to train our neural network models was gathered thanks in large part to these algorithms. Plots that demonstrate the effectiveness and analysis of our models were produced using Matplotlib, and diagrams were made using draw.io to help visualize the system architecture and data flow inside the setup.

Model training was conducted on a local machine equipped with a CUDA (v11.8) enabled GPU (NVIDIA GeForce RTX 2060 Max-Q), using PyTorch (v2.2.2). Batch size of 16 was used during training due to computational constraints, and the Adam optimizer with initial learning rate set to 1×10^{-3}.

B. Baseline Architecture

A baseline real-time commentary system is shown in Fig. 1 and built upon it in the following subsections. The process involves extracting frames from the live feed of a game at regular intervals of 15 seconds and passing it into GPT-4o with Vision capabilities. We implement a two-commentator system, where commentators alternate with each scene change or time interval expiry. This mimics traditional e-sports broadcasting formats where play-by-play and color commentators complement each other's analysis. The system prompt is as follows:

"You are a passionate, dynamic and professional League of Legends Esports commentator. Here is some data about the game that is currently happening - [RAG Contextual Data]. Use this data only for gaining insight. Don't repeatedly mention it during the commentary. It is simply for you reference. Given the previous commentary and the current game image, continue with crisp commentary, highlighting impactful plays, critical moments, or strategic moves in 50 words or less. Make the commentary exciting while staying relevant to the action in the image. Make sure your responses don't become repetitive or awkwardly dramatic. Be natural and fluent like a sports caster. [Memory Context Prompt]". The memory prompt looks like "You have a special structure called MEMORY which is a sentence of 150 words. This is used for you to maintain context for the next prompt completion.

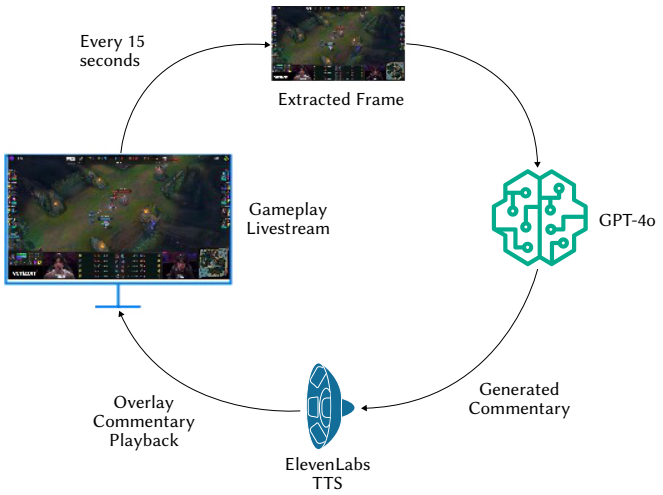

Fig. 1. Baseline Architecture of a Real-Time E-sports Commentary System.

Be careful and fill this structure with relevant information. You can replace whatever you want from this whenever you want. Output this structure after you are done modifying its contents based on the game state and the relevant information in the format MEMORY: memory of your choice. Remember that the quality of this memory is crucial for maintaining context and strategic commentary over long range".

This is followed with the last three pieces of commentary it generated in order to avoid repetition. Following the generation of textual commentary, our system utilizes the Elevenlabs (Text-to-Speech) API to convert the text-based commentary into spoken words. Leveraging advanced speech synthesis techniques, the TTS API ensures the production of natural-sounding speech. The resulting spoken commentary is seamlessly synchronized with the live-streamed gameplay, providing viewers with an immersive and engaging experience enriched with real-time insights generated by AI.

While the system works well as a baseline, one of the challenges faced is synchronizing the commentary with the movements of the in-game camera. Within the sampling interval of consecutive frames, the camera can shift to any part of the map (top, mid, bottom, river, or jungle), which makes previously generated commentary irrelevant and outdated. To address this, a scene change detection algorithm using classical computer vision operations as outlined below.

C. Scene Change Detection

The in-game minimap of League of Legends has a white rectangle demarcating the position of the spectator camera. Given a particular frame from a live game, the proposed scene change detection algorithm works in the following way:

1. Crop the image to keep only the minimap and convert to gray-scale.
2. Apply binarization using simple thresholding (190 was found to be effective).
3. Draw contours on the image, and filter for the contour with the largest area.
4. Check if the area of contour is large enough to be a valid camera rectangle (1000 square pixels was found to be effective).
5. Draw a rectangle around the contour(to convert it to a perfect rectangle and extract its top left corner co-ordinates, width, and height.
6. Calculate the Euclidean distance (in pixels) of the camera co-ordinates detected in the current frame and compare with that of the previous frame extracted.
7. If the camera moves more than a certain threshold, detect a scene change (distance threshold of 80 pixels was found to be effective).

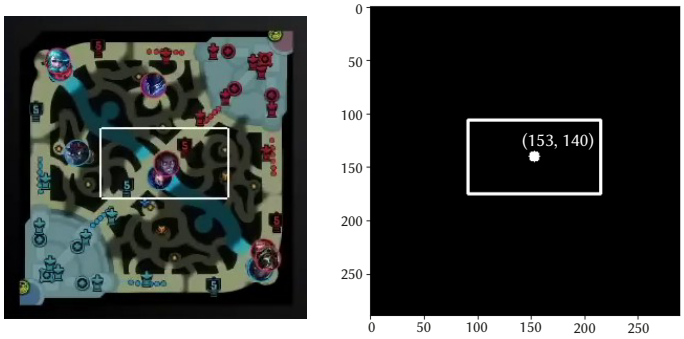

Fig. 2. Detecting Camera Position using Computer Vision Techniques.

Fig. 2 shows the minimap and the corresponding camera rectangle extracted using the proposed algorithm. A distance threshold (experimentally set to be 80 pixels) sensitive to minor errors in extracting camera co-ordinates (see Fig. 3). Rather than relying on fixed time intervals, this method intelligently extracts frames based on these scene changes, ensuring a dynamic and contextually relevant stream of commentary. Additionally, to improve real-time performance, a minimum interval is chosen (5 seconds) before which commentary cannot change, as well as a maximum interval (25 seconds) after which commentary must be refreshed, approximating a more natural flow and human-like coherence. This approach not only enhances the responsiveness of our system but also ensures a seamless viewing experience for audiences immersed in live-streamed gaming content.

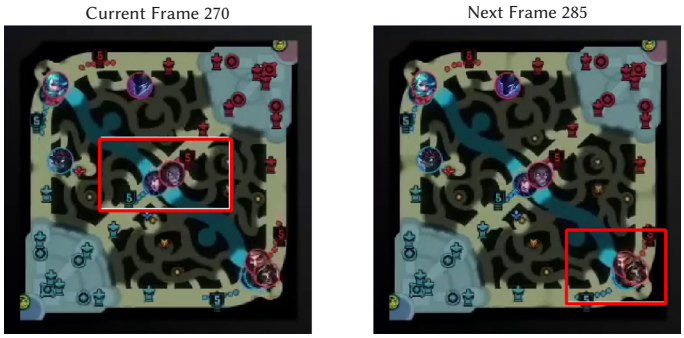

Fig. 3. Scene Change Detection Based on Euclidean Distance Thresholding (126.62 x 80 pixels).

D. AI Camera Dataset

Our dataset comprises gameplay footage captured from game two of the 2023 World Championship between SKT T1 and Weibo Gaming. The capture process was structured across three distinct game phases:

- Early Game (0-10 minutes): Captured footage focusing on lane matchups, initial jungle pathing, and early skirmishes. Camera positions during this phase predominantly centered around individual lanes and river objectives.
- Mid Game (10-20 minutes): Team rotations, objective control, and group engagements. Camera work in this phase balanced between strategic point coverage and teamfight positioning.
- Late Game (20+ minutes): Large-scale teamfights and base sieges, with camera positions clustering around major objectives and base defense/offense scenarios.

For each phase, we implemented a systematic capture process:

- Frame extraction at 1 Hz (one frame per second) from the live game feed.

- Precise camera centroid position annotation using our tracking algorithm.
- Minimap state preservation for spatial context.
- Scene transition documentation between major game events.

This segmented capture approach resulted in 1,717 annotated in-game minimaps. The distribution of camera positions throughout the match reveals dense clustering in activity hotspots - primarily the three lanes, with sparser distributions across jungle and river areas (see Fig. 4).

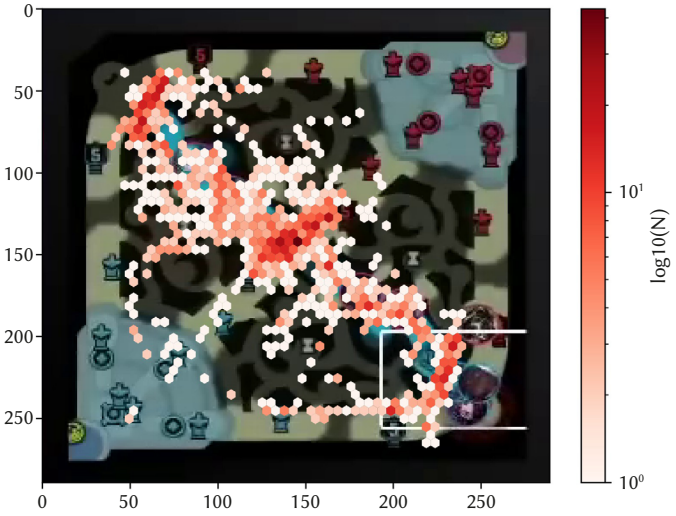

Fig. 4. Hexbin Plot of Distribution of Camera Positions

To prepare this data for our AI models, we used a collate function that transforms sample batches into time series with window size n = 15 (corresponding to 15-second gameplay segments). Each subsequent element is obtained by sliding this window one time step forward. The target variable for each time series is the camera centroid position at the next immediate time step. This structure yields 1,701 annotated image sequences (1,717 − n − 1).

Through analysis of the captured footage, we determined that scene changes occurred on average every 16.17 seconds. This insight informed our choice of a one-second sampling rate and 15-frame window size, ensuring our capture methodology effectively documented complete camera movement sequences.

E. Camera AI Architecture

Based on the collected dataset, the task of AI powered camera direction is proposed, with the goal of controlling the camera based on the activity in the minimap. The model must be capable of understanding and analyzing both spatial and temporal features of the minimap and camera movements, in order to effectively predict next positions of the camera. Two neural network architectures are proposed. The first, detailed in [23], employs a Conv2D-LSTM model. Unlike traditional LSTM models that struggle with spatial relationships in image sequences, this hybrid architecture segregates spatial and temporal modeling, leveraging Convolutional Neural Networks (Conv2D) for extracting hierarchical spatial features and Long Short-Term Memory (LSTM) networks for temporal dynamics. The Conv2D-LSTM framework enables comprehensive analysis by separately capturing spatial features and temporal dependencies within the gameplay footage.

Fig. 5 illustrates an overview of the 2D Convolutional-LSTM architecture. Both the minimaps and the camera centroids are normalized to [0, 1] by division with maximum pixel value (255) and image height/width (289) respectively. If B is the batch size, N the number of time steps, C the number of colour channels, H and W the height and width of the minimap image respectively, then the model receives sequences of images as an input tensor of the shape (B, N, C, H, W). In order to run 2D convolutions, this is reshaped to $(B \times N, C, H, W)$. Each Conv2d block consists of the 2d convolutions (3 ×3 filter) , followed by a ReLU activation to introduce non-linearity, and finally pooling. After convolutions, the tensor is reshaped to the original B and N dimensions, and each feature map is pooled, resulting in multidimensional time series representation of (B, N, D), where D is the number of features per minimap. Centroids of the minimaps are concatenated to their corresponding feature vector, resulting in a tensor of shape $(B, N, D + 2)$. This additional explicit spatial information of the camera centroid helps the model better understand spatio-temporal context of the game. This tensor is passed through the LSTM layer, after which a pooled output of each hidden state is used as final state representation. Finally, fully connected layers are used to predict the next centroid. The Mean Squared Error (MSE) loss with the Adam optimizer is used to train.

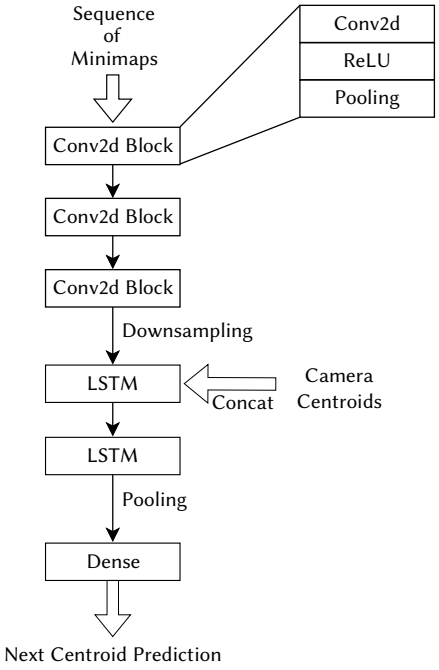

Fig. 5. Overview of Conv2d-LSTM Architecture.

In contrast, the second architecture, outlined in [24], adopts a 3D Convolutional approach. This architecture integrates spatial and temporal modeling into a unified framework, utilizing 3D convolutional layers (Conv3D) to simultaneously capture both spatial and temporal information. By employing Conv3D networks, this architecture achieves a holistic understanding of the gameplay dynamics, effectively modeling spatial-temporal interactions within the live-streamed footage. By exploiting the structure of the data, this model is more efficient in terms of learned parameter size as well as inference/training speed.

Fig. 6 shows an overview of the 3D Convolutional Neural Network Architecture, with each Conv3d block consisting of a 3d convolution operation (3 ×3 filter), followed by pooling, batch normalization, and ReLU activation for non-linearity. The normalization of minimaps and camera centroids remains the same as that in the conv2d-LSTM network, except there is no need to reshape the tensor before 3D Convolutions. This is due to the fact that 3D convolutions already take sequences of images as input to interpret the spatio-temporal context. This also eliminates the need for separate recurrent layers

to handle temporal dependencies. Thus, it is possible to directly downsample the features to pass into the fully connected layers to make predictions. Note that the camera centroids are appended to the flattened representation of features instead of each minimap's feature vector (like in the Conv2d-LSTM model), since 3D convolutions do not preserve the temporal dimension. Similar to training the conv2d-lstm network, the Mean Squared Error loss function with the Adam optimizer is used to train.

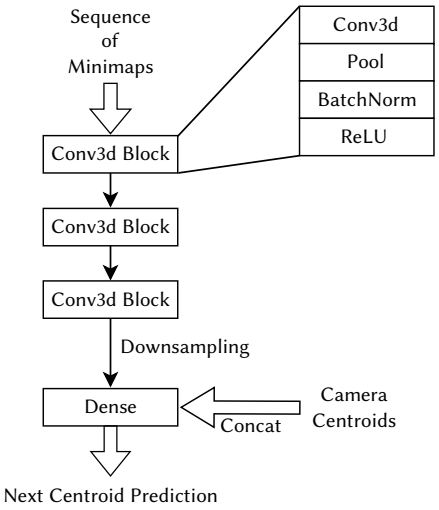

Fig. 6. Overview of Conv3d Architecture.

F. Enhanced Context Awareness

To improve the system's ability to maintain consistent and contextually relevant commentary throughout a match, we enhance our baseline architecture with two key components: a dynamic token memory system and pre-game knowledge augmentation.

Our implementation utilizes a 200-token cache that serves as the LLM's working memory throughout the match. This memory system, inspired by recent work in context management [25], allows the model to maintain and reference key information about ongoing game developments, strategic shifts, and narrative threads. The LLM dynamically manages this memory space, choosing which information to retain or discard based on its relevance to the current game state and commentary flow.

The pre-game knowledge integration is implemented through a focused Retrieval-Augmented Generation (RAG) approach [26]. The system augments the base prompt with crucial domain-specific information including champion-specific details (abilities, roles, power spikes, playstyle), team composition dynamics, etc.

This targeted augmentation ensures that the generated commentary is grounded in accurate game knowledge while maintaining computational efficiency. Unlike traditional RAG implementations that require real-time retrieval from large knowledge bases [27], our approach front-loads relevant information into the prompt, allowing for faster generation while still maintaining domain expertise in the commentary.

The combination of dynamic memory management and focused knowledge augmentation enables our system to produce commentary that is both technically accurate and narratively coherent, capable of referencing both immediate game developments and broader strategic contexts.

G. Proposed Architecture

Coalescing the camera tracking, scene change detection and AI camera to the baseline system, our proposed architecture is illustrated in Fig. 7. The spectator camera is controlled by the camera neural network, and on scene change detection, the current frame from the live feed is sent into the GPT-4o model to generate newer, more relevant commentary. This is overlaid onto the live feed in real-time using FFmpeg. Hence, an autonomous, real-time AI powered spectating and commentating system for the game League of Legends is achieved.

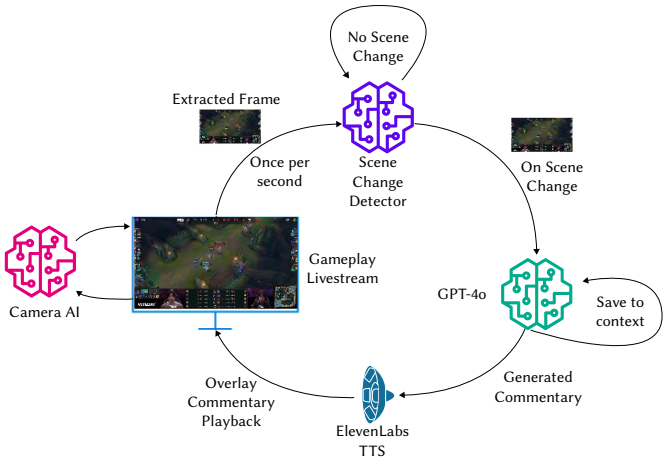

Fig. 7. Proposed Architecture with Scene Change Detection and AI Camera Direction.

IV. Results

A. Quantitative Evaluation of Scene Change Detection

The approximately 30 minute game of League of Legends was divided into 3 segments, the early, mid and late game respectively. For each of these approximately 10 minute long segments, the first 7.5 minutes were used for training and the final 2.5 minutes were used for evaluation. Training and evaluation was done using single time-step predictions. This was achieved using custom PyTorch Datasets and Dataloaders. The sequences of minimaps were created dynamically in memory (in batches) during training using the dataset in local storage. Fig. 8 & Fig. 9 show the training and evaluation loss vs epoch curves for the Conv2d-LSTM and the Conv3d models respectively. Losses are reported as average of each batch used in the epoch.

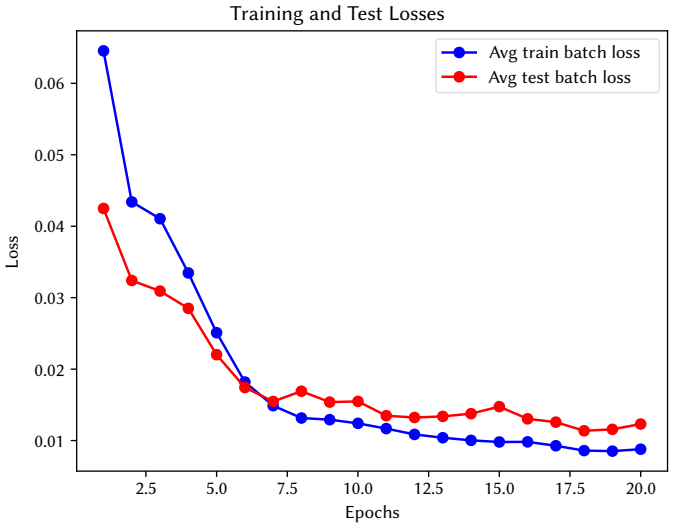

Fig. 8. Average batch loss versus epoch of Conv2d-LSTM.

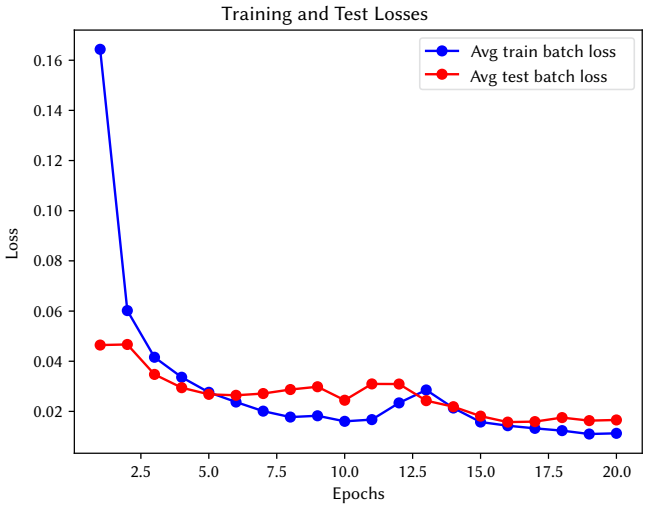

Fig. 9. Average batch loss versus epoch of Conv3d.

It is observed that both the Conv2d-LSTM model and the Conv3d model are able to learn the spatio-temporal dynamics of the game from the minimap and result in similar evaluation loss. However, it should be noted that the latter had 140,514 trainable parameters, whereas the former had 700,194 trainable parameters, approximately 5 times more complex. To interpret model loss in terms of pixel distances, it is de-normalized by multiplying with the width/height of the minimap image. This gives us the average error in pixel distance for a batch of 16, thus dividing, the average error for each individual prediction is calculated. These results are shown in Table I.

TABLE I. Interpreting Eval Loss of Models in Pixels

Metric	Conv2d-LSTM	Conv3d
Average Eval Batch Loss	0.0114	0.0157
Average Eval Batch Loss (denorm)	3.2856	4.5454
Average Loss per Forecast	0.2054	0.2841

B. Generated Commentary

In order to evaluate AI-generated commentary, it is contrasted with transcripts of the commentary from the official LoL World Championship broadcast. 3 particular frames from the beginning of game 2 of the 2023 Finals are randomly selected, and commentary is compared for the duration until a scene change occurs. Fig. 10 shows the 3 frames selected with specific timestamps and durations, followed by Table II comparing AI and human commentary for each frame segment. While both models showed similar capabilities in learning from game dynamics, the lesser complexity of the Conv3d model suggests a more efficient approach in handling spatio-temporal data, which is crucial in real-time applications like e-sports commentary. This finding aligns with recent studies [24] emphasizing the efficiency of 3D convolutions in capturing temporal patterns more succinctly than LSTM-based architectures.

V. Discussion

A. Qualitative Assessment of Baseline and Proposed Systems

The baseline and proposed systems are qualitatively evaluated on the 3 evaluation sets used for the camera AI architectures corresponding to the early, mid and late parts of the game. For both systems live videos are compiled for the duration of the set, overlaying commentary generated from each system. The evaluation is done using the following 3 metrics: fluency, relevance to context, and strategic depth.

(a) Frame @ 1:46, Duration: 10 seconds

(b) Frame @ 1:56, Duration: 8 seconds

(c) Frame @ 2:04, Duration: 9 seconds

Fig. 10. Frames at Different Scene Segments of the Game.

TABLE II. Comparison of AI (LLM With Memory+RAG) and Human Commentary at Different Scene Segments of the Game

Frame	AI Commentary	Human Commentary
1:46	Alice: Xiaohu, on Ahri, is skillfully managing the wave against Faker's Sylas in the mid lane. Both players are focusing on last-hitting minions ef-.	And obviously WeiWei gonna be way stronger early on, Oner mainly looking to cover in case they do uhh seeWeiWei show up in the bot side. Do note that they do have award on his blue-
1:56	Light and Crisp are aggressively pressuring Gumayusi and Keria in the bot lane! Light's Kalista is utilizing her mobility to.	So the T1 botlane is very aware as oh, the bot lane! Flash already out from Gumayusi as Light's relatively low as well but that's definitely a trade-
2:04	TheShy launches an aggressive trade against Zeus in the top lane! Aatrox's Deathbringer Stance is making these early skirmishes deadly. Both players are fiercely contesting the wa-.	That goes incredibly well here for Weibo. Zeus also taking a fair bit of damage towards this topside. Weibo Gaming, they are not backing down yet. Yeah, it's getting spicy here early on-

Results indicate that while both systems are comparable in fluency, the proposed system, leveraging the scene change detection and camera tracking algorithm, demonstrated stark superiority in relevance to context and by virtue, better strategic depth. The baseline system with periodic commentary generation could not keep up with dynamic camera shifts and movements, resulting in outdated or irrelevant commentary post a scene change. The proposed system, utilizing the scene change detection and camera tracking algorithm, ensured contextually relevant commentary, and in effect improving the strategic depth of the commentary. Furthermore, the API calls to GPT-4o and the TTS API remained roughly similar, reaching better performance at similar running costs.

B. Qualitative Evaluation Methodology

This study employed a qualitative assessment framework to systematically evaluate the system's performance across three principal dimensions:

1. Core Evaluation Metrics

- **Fluency**: The evaluation of linguistic coherence and the smoothness of commentary delivery, reflecting the natural flow of language.
- **Contextual Relevance**: An analysis of the commentaryâ€™s alignment with ongoing game events, assessing its appropriateness and contextual accuracy.
- **Strategic Depth**: The examination of the systemâ€™s capacity to provide meaningful tactical insights, enhancing the interpretability of game dynamics.

2. Comparative Analysis

A direct comparative analysis was undertaken between AI-generated commentary and human-generated commentary, employing timestamped game segments for a rigorous and systematic evaluation. Three distinct game frames were selected, each with predefined durations:

- **Frame 1**: Timestamp 1:46 (10 seconds duration)
- **Frame 2**: Timestamp 1:56 (8 seconds duration)
- **Frame 3**: Timestamp 2:04 (9 seconds duration)

The evaluation within these frames focused on two critical aspects:

- **Context Awareness**: The assessment of the system's comprehension of game dynamics and its ability to demonstrate strategic knowledge.
- **Transition Fluidity**: The analysis of the commentary's coherence and fluidity during scene transitions.

3. System Comparison

A comparative evaluation of the baseline system and the proposed system was conducted using the aforementioned core evaluation metrics. This analysis was performed across distinct stages of gameplay:

- Early game
- Mid game
- Late game

Performance was systematically assessed through the examination of compiled live video footage featuring overlaid commentary. This approach ensured a consistent and reproducible basis for comparative analysis.

C. Qualitative Comparison of Human Vs AI Generated Commentary

While there are similarities in the two commentaries generated, two key observations:

- **Context Awareness**: While GPT-4o demonstrates a general understanding of the game and players, it cannot commentate with the depth of knowledge possessed by human commentators, particularly when it comes to player-specific strategies, historical performance data, game meta or even broader tournament narratives.
- **Transition Fluidity**: During scene changes when the spectator camera swithches to a different part of the map, AI transitions are abrupt, often interrupting itself to talk about the next scene, whereas human commentary maintains its natural flow from the previous scene into the next.

D. Limitations

It is important to acknowledge two limitations of our research. First off, only single time-step forecasts were used to evaluate the camera control models. To potentially improve the robustness and practicality of our models, multi-step forecasting benchmarks are necessary. This in turn necessitates generating new minimaps with distinct camera demarcations for each forecast. This kind of generation would require simulating live League of Legends games using the AI-controlled spectator camera, which is currently not allowed as per Riot Games' Terms of Service (scripting violations); explicit consent for these kinds of activities has not yet been given.

Furthermore, the training of our camera models was limited to a smaller dataset batch size, due to computational constraints, something to address in the future given adequate resources. These aspects should be taken into account when analyzing the findings and possible uses of our research.

E. Future Work

Future directions for expanding on the work described in this research work seem promising. Scaling up of computational resources can extend our model to handle larger, more comprehensive datasets, which could improve the performance and adaptability of our system to various gaming environments and platforms. To achieve a reliable, full-scale application that can function well in real-world environments, this scale-up is necessary.

Furthermore, incorporating in-depth game insights into the live commentary could be feasible with a model converts natural language into targeted queries for the Riot Games API. Such a model might be able to provide automated game commentary with previously unexplored levels of interaction and information depth.

VI. Conclusion

This research introduced a comprehensive AI-driven system with automated camera direction and dynamic, contextually-aligned commentary to improve the live-streamed e-sports event viewing experience. Through the integration of cutting-edge AI tools like OpenAI's GPT-4o and Elevenlabs' TTS API with sophisticated neural network architectures, such as 2D Conv-LSTM and 3D CNN, a system that not only enhances the relevance and fluency of commentary but also synchronizes it with real-time camera movements is achieved. Our analyses show that these integrated systems have the potential to deliver captivating and immersive viewing experiences, pushing the boundaries of live broadcast automation.

Additionally, a new camera tracking and scene change detection algorithm was used to overcome the difficulties related to real-time commentary syncing and AI-driven content delivery, demonstrating the viability of employing AI to comprehend and react to dynamic visual inputs. While the work is constrained by available computational resources and legal requirements, it establishes a strong basis for future investigations into more intricate AI applications in the gaming and

entertainment sectors. Looking ahead, attaining higher practicality and robustness in AI-driven live streaming solutions would require scaling computer resources and increasing dataset capabilities in order to get past current constraints.

Finally, this paper contributes to the advancement of artificial intelligence in digital entertainment by showing how state-of-the-art technology can be used to improve user engagement and offer more in-depth narrative experiences in e-sports, an industry where the merging of technology and human creativity is still a rapidly growing field.

CRediT Authorship Contribution Statement

Swathi Jamjala Narayanan: Supervision, Conceptualization, Methodology, Writing – Review & Editing, Project administration.

Kevin Winston Joseph: Conceptualization, Methodology, Software, Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.

Devansh Sirohi: Software, Investigation, Data Curation, Validation.

Harsh Chaudhary: Software, Investigation, Writing – Review & Editing.

Hitesh Shivkumar: Software, Investigation, Visualization.

Data Statement

The data used in this study was derived from publicly available League of Legends World Championship 2023 gameplay footage. As this was a proof-of-concept study, the processed dataset is not publicly archived. The methodology for data collection and processing is fully described in the manuscript to enable reproduction.

Declaration Of Conflicts Of Interest

The authors declare no conflicts of interest.

References

[1] Block, F. Haack, "esports: a new industry," in *Proceedings of 20th International Scientific Conference Globalization and its Socio-Economic Consequences, SHS Web of Conferences,* Zilina, Slovak Republic, vol. 92, 2021, p. 04002, EDP Sciences, doi: https://doi.org/10.1051/shsconf/20219204002.

[2] N. Renella, M. Eger, "Towards automated video game commentary using generative ai," in *Proceedings of the AIIDE Workshop on Experimental Artificial Intelligence in Games, CEUR Workshop,* Salt Lake City, Utah, USA, vol. 3626, 2023, pp. 341–350. url: https://ceur-ws.org/Vol-3626/paper7.pdf.

[3] Z. Wang, N. Yoshinaga, "Esports data-to-commentary generation on large-scale data-to-text dataset," *arXiv preprint arXiv:2212.10935*, 2022.

[4] X. Qi, C. Li, Z. Liang, J. Liu, C. Zhang, Y. Wei, L. Yuan, G. Yang, L. Huang, M. Li, "Mcs: an in-battle commentary system for moba games," in *Proceedings of the 29th International Conference on Computational Linguistics,* Gyeongju, Republic of Korea, 2022, pp. 2962–2967. url: https://aclanthology.org/2022.coling-1.262/.

[5] O. Olarewaju, A. V. Kokkinakis, S. Demediuk, J. Robertson, I. Nölle, S. Patra, D. Slawson, A. Chitayat, A. Coates, B. Kirman, *et al.*, "Automatic generation of text for match recaps using esport caster commentaries," in *Proceedings of International Conference of Natural Language Computing, CS IT – Computer Science Conference Proceedings,* Sydney, Australia, 2020, pp. 117–131. Virtual conference, doi: https://doi.org/10.5121/CSIT.2020.101811.

[6] H. Huang, J. H. Xu, X. Ling, P. Paliyawan, "Sentence punctuation for collaborative commentary generation in esports live-streaming," in *Proceedings of IEEE International Conference on Consumer Electronics (ICCE),* Las Vegas, NV, USA, 2022, pp. 1–2. url: https://arxiv.org/abs/2110.12416.

[7] T. Ishigaki, G. Topić, Y. Hamazono, H. Noji, I. Kobayashi, Y. Miyao, H. Takamura, "Generating racing game commentary from vision, language, and structured data," in *Proceedings of the 14th International Conference on Natural Language Generation,* Aberdeen, Scotland, UK, 2021, pp. 103–113, doi: https://doi.org/10.18653/v1/2021.inlg-1.11.

[8] C. Nimpattanavong, P. Taveekitworachai, I. Khan, T. V. Nguyen, R. Thawonmas, W. Choensawat, K. Sookhanaphibarn, "Am i fighting well? fighting game commentary generation with chatgpt," in *Proceedings of the 13th International Conference on Advances in Information Technology,* ACM, Bangkok, Thailand, 2023, pp. 1–7, doi: https://doi.org 10.1145/3628454.3629551.

[9] M. Czaplicki, "Live commentary in a football video game generated by an ai," Master's thesis, University of Twente, Business IT BSc programme, Enschede, The Netherlands, July 2023. url: https://purl.utwente.nl/essays/96001.

[10] Y. Taniguchi, Y. Feng, H. Takamura, M. Okumura, "Generating live soccer-match commentary from play data," in *Proceedings of the AAAI Conference on Artificial Intelligence,* Honolulu, Hawaii, USA, vol. 33, 2019, pp. 7096–7103, doi: https://doi.org/10.1609/aaai.v33i01.33017096.

[11] C. Chan, C. Hui, W. Siu, S. Chan, H. A. Chan, "To start automatic commentary of soccer game with mixed spatial and temporal attention," in *TENCON 2022-2022 IEEE Region 10 Conference (TENCON),* Hong Kong SAR, China, 2022, pp. 1–6, IEEE, doi: https://doi.org/10.1109/TENCON55691.2022.9978078.

[12] E. Marrese-Taylor, Y. Hamazono, T. Ishigaki, G. Topić, Y. Miyao, I. Kobayashi, H. Takamura, "Open-domain video commentary generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* Abu Dhabi, United Arab Emirates, 2022, pp. 7326–7339, doi: https://doi.org/10.18653/v1/2022.emnlp-main.495.

[13] S. Ma, L. Cui, D. Dai, F. Wei, X. Sun, "Livebot: Generating live video comments based on visual and textual contexts," in *Proceedings of the AAAI Conference on Artificial Intelligence,* Honolulu, Hawaii, USA, vol. 33, 2019, pp. 6810–6817. url: https://arxiv.org/abs/1809.04938, doi: https://doi.org/10.1609/aaai.v33i01.33016810.

[14] H. Wu, G. J. F. Jones, F. Pitié, "Knowing where and what to write in automated live video comments: A unified multi-task approach," in *Proceedings of the 2021 International Conference on Multimodal Interaction,* Montréal QC, Canada, 2021, pp. 619–627, doi: https://doi.org/10.1145/3462244.3479942.

[15] J. Chen, J. Ding, W. Chen, Q. Jin, "Knowledge enhanced model for live video comment generation," in *2023 IEEE International Conference on Multimedia and Expo (ICME),* Brisbane, Australia, 2023, pp. 2267–2272, IEEE, doi: https://doi.org 10.1109/ICME55011.2023.00387.

[16] D. Dai, "Live video comment generation based on surrounding frames and live comments," *arXiv preprint arXiv:1808.04091*, 2018.

[17] J. H. Xu, Y. Cai, Z. Fang, P. Paliyawan, "Promoting mental well-being for audiences in a live-streaming game by highlight-based bullet comments," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE),* Kyoto, Japan, 2021, pp. 383–385, IEEE, doi: https://doi.org/10.1109/GCCE53005.2021.9621853.

[18] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] G. Bertasius, H. Wang, L. Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of 38th International Conference on Machine Learning,* Virtual (online), vol. 2, 2021, p. 4. url: https://arxiv.org/abs/2102.05095.

[20] C. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, C. Feichtenhofer, "Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* Santiago, Chile, 2022, pp. 13587–13597. url: https://arxiv.org/abs/2201.08383. doi: https://doi.org/10.1109/CVPR52688.2022.01322.

[21] L. Wei, L. Xie, W. Zhou, H. Li, Q. Tian, "Mvp: Multimodality-guided visual pre-training," 2022. [Online]. Available: https://arxiv.org/abs/2203.05175.

[22] Z. Tong, Y. Song, J. Wang, L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10078–10093, 2022. url: https://arxiv.org/abs/2203.12602.

[23] M. H. Mohd Noor, S. Y. Tan, M. N. Ab Wahab, "Deep temporal conv-lstm for activity recognition," *Neural Processing Letters*, vol. 54, no. 5, pp. 4027–4049, 2022. url: 10.1007/s11063-022-10799-5, doi: https://doi.org/10.1007/

s11063-022-10799-5.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision,* Santiago, Chile, 2015, pp. 4489–4497, doi: https://doi.org/10.1109/ICCV.2015.510.

[25] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, M. Yang, "Longrope: Extending llm context window beyond 2 million tokens," 2024. [Online]. Available: https://arxiv.org/abs/2402.13753.

[26] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2312.10997.

[27] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, "Dense passage retrieval for open-domain question answering," 2020. [Online]. Available: https://arxiv.org/abs/2004.04906.

Swathi Jamjala Narayanan

She received her Ph.D. from Vellore Institute of Technology in 2015. She is currently designated as Professor G in School of Computer Science and Engineering, VIT Vellore, India. She has 17 years of teaching experience in computer science. Her research interest includes Soft Computing, Pattern Recognition, Machine Learning and Data Mining. She has been awarded with the Best Ph.D Thesis by Computer Society of India. She is a member of the International Association of Engineers and also a lifetime member of the Computer society of India and the Soft computing research society.

Kevin Winston Joseph

He is currently in his final year of pursuing his B. Tech in Computer Science Engineering with a specialization in Data Science at Vellore Institute of Technology, Vellore, India. His research interests include foundational and applied deep learning.

Devansh Sirohi

He is currently in his final year of pursuing his B.Tech in Computer Science Engineering with a specialization in Data Science at Vellore Institute of Technology, Vellore, India. His current research interests are data mining, machine learning and cloud computing.

Harsh Chaudhary

He is currently pursuing his B.Tech in Computer Science with a specialization in Information Security at Vellore Institute of Technology, India, expected to graduate in 2024. With a keen interest in generative AI, NLP, and cybersecurity, he has actively engaged in research projects and academic pursuits.

Hitesh Shivkumar

He is currently pursuing his B.Tech in Computer Science at Vellore Institute of Technology, Vellore, India. His research areas include Generative AI, LLMs, Cloud Development, and Stochastic Modeling. He is actively engaged in multiple research projects along with academic pursuits.