A Small Video-Language Model for Sparse Action Spotting

Evan Zimmerman
(evanzimm@umich.edu)

Naveen Unnikrishnan
(naveenu@umich.edu)

Nicholas Mellon
(nbmellon@umich.edu)

Pranavkumar Mallela
(pmallela@umich.edu)

Vishal Chandra
(chandrav@umich.edu)

Abstract— **Sparse action spotting describes the challenge of retrieving moments with a specific semantic meaning in long untrimmed videos. Previous works addressing this task have typically relied on Convolutional Neural Networks (CNN) or feature-extractor-based methods. Our work tackles the sparse action spotting problem in the context of highlight detection for broadcast soccer video. In light of recent work, we explore a purely token-based, prompt-driven highlight generation framework using a fine-tuned Vision Language Model (VLM), Large Language Model (LLM), and large reasoning model, and investigate if the enhanced multimodal understanding of VLMs helps in better identifying important events in broadcast soccer matches. Our results show that a small fine-tuned VLM achieves better overall performance in identifying intervals for key actions in a soccer video over a larger LLM and a full-size state-of-the-art reasoning model with zero-shot prompting.**

I. Introduction

The problem our work addresses is the automatic generation of custom highlights (a.k.a. "exciting moment") reels in broadcast sports video, with a specific focus on soccer. Traditionally, the generation of these reels has been done manually—a labor-intensive and time-consuming process. We propose a system that, when given a video of a soccer match and associated commentary, can extract highlights corresponding to key semantic events (such as goals, corners, fouls, etc.). For example, a user should be able to prompt the system with a natural language query like "show me the goals scored in the second half" or "show me all the yellow cards," and receive accurate highlight segments in response.

This task falls under the broader problem of sparse action spotting—identifying temporally sparse, semantically meaningful events within long, untrimmed video sequences [4]. Prior work on this problem has largely focused on frame-level classification using Convolutional Neural Networks (CNNs) or feature-extractor methods [7]. However, the rise of the transformer architecture and self-attention mechanism begs the investigation of a prompt-driven, token-based approach to this task that leverages vision-language models (VLMs) or large language models (LLMs) for a richer, more flexible semantic understanding. These models are particularly compelling for a task like highlight detection, which requires reasoning over both spatial and temporal dimensions, and natural language as well (when commentary is included).

To investigate this, we construct a purely prompt-driven pipeline and compare the performance of a fine-tuned VLM, a large LLM, and a zero-shot state-of-the-art reasoning model. Our results demonstrate that even the smallest version of a moderately effective VLM outperforms larger and more powerful models in identifying highlight intervals. While sparse action spotting remains a highly challenging task for VLMs, the early, relative success of our approach is encouraging. With additional scale and further finetuning, more powerful VLMs may offer even greater potential for solving this problem.

II. Significance

In our work, we introduce a completely novel token-based method to query and return time intervals of significant events. This is in contrast to traditional classification methods that return whether or not a frame or clip contains an event (and if so, what exact time it happens; for example, the exact second when a ball crosses the goal line). Rather, we focus on identifying the interval during which the event unfolds. We argue that this better reflects how events like goals are understood semantically; they may not be best understood in a singular second, but instead contain moments of build-up or celebration, supported by video evidence or intonation changes in commentary.

This reframing also aligns well with the strengths of VLMs and LLMs, which are better equipped to reason over longer contexts. Concurrent work [18] by LG AI validates the usage of VLMs for sparse action spotting in soccer games.

Additionally, a benefit of this novel approach is the flexibility it allows for users. Querying a system with text for a specific type of highlight is challenging, if not impossible under traditional approaches.

Lastly, we note that the fact that the desired actions are very sparse (1-2 goals per game on average) emphasizes why this is a non-trivial problem.

III. Related Work

A. Highlight Generation

Automated sport highlight detection models have traditionally relied on sport-specific heuristics and domain knowledge to identify segments of importance. Xie et al [25] use common trends in football videos (namely camera angles, dominant colors, and motion) with hidden Markov models (HMMs) to identify segments of play. Rui et al [17] use SVMs to extract highlights for baseball matches using audio data to classify regions of *excitement*. However, such approaches do not necessarily translate well to other sports. For instance, while audio-based classification works reasonably well for baseball, the same approach might not work for sports like soccer where audiences tend to chant throughout the game to show support or rile the opposition.

More recent work in this area has focused on creating generalized (activity-agnostic) highlight generation solutions. Lin et al [14] use professional photographs as a prior to detect highlights of a video. While this works well for activities where common photographs closely align with relevant video segments (like weddings or skateboarding), it might not scale to team sports like soccer where there is often a disparity in published photographs (which are usually close-ups of individual players) versus highlight reels (which tend to be zoomed-out segments of gameplay). Santa et al [6] use deep learning to detect clips of interest from a combination of audio and video. While their approach works well for extracting a specific type of highlight for a specific sport (a goal in a soccer match), it is limited by the highlight event and sport it is trained on.

B. VLMs

Recent advances have prompted similar explorations using large language models. Maaz et al. [15] introduce Video-ChatGPT, demonstrating strong capabilities in generating human-like conversations about video content. Sun et al [20] introduce GPTSee for moment retrieval using similarity scores of generated descriptions and the natural language query. More recently, the Alibaba group released Qwen-2-VL and Qwen-2.5-VL, which are the current open source state-of-the-art models for video understanding [24], [21].

C. Sparse Action Spotting

SoccerNet is a large-scale dataset for soccer video understanding [5]. In conjunction with the SoccerNet dataset, there is a yearly challenge for sparse action spotting. For many years, this competition has been dominated by CNN-based approaches, with most research innovations centering on pooling methods and complex losses. However, more recently, a transformer-based method [19] surpassed the performance of all CNN approaches, with the caveat that it still operated on ResNet input features. That is, each frame of the video was pre-processed by a ResNet, its features aggregated per frame, and fed to the transformer network. Table I, largely borrowed from that work, compares many approaches and their input feature representations.

Method	Features	Avg-mAP
NetVLAD++ [12]	ResNet	53.4
AImageLab RMSNet [22]	ResNet-tuned	63.5*
Vis. Analysis of Humans	CSN-tuned	64.7†
Faster-TAD [2]	Swin-tuned	N/A
CALF [4], [3]	ResNet+PCA	40.7
NetVLAD++ [12]	ResNet+PCA	50.7
DU [19]	ResNet+PCA	63.8
DU+SAM+mixup w/o $\hat{\mathbf{D}}$ [19]	ResNet+PCA	72.1
DU+SAM+mixup [19]	ResNet+PCA	**72.2**
Zhou et al. [26]	Combination	73.8
DU [19]	Combination	75.7
DU+SAM+mixup w/o $\hat{\mathbf{D}}$ [19]	Combination	**77.3**
DU+SAM+mixup [19]	Combination	**77.3**

TABLE I

A comparison of prior approaches and their input features for SoccerNet action spotting [19]

Most recently, two works stand out to us as a possible paradigm-shift for this task. COMEDIAN [7] and SoccerCLIP [18] both deviate from CNNs using transformers and CLIP respectively as their backbones. While both methods consider vision and language, they do not investigate our goal of a "pure" token-based framework for the task. More specifically, COMEDIAN uses a pair of transformers—one spatial and one temporal—in its feature extraction and inference. SoccerCLIP in contrast co-optimizes language-image embeddings using a joint latent space as the basis of its inference. Both methods seem to be hinting at a shift to a unified token paradigm for this sparse task, which we begin to investigate in this work.

IV. Methodology

A. Implementation

The problem of sports highlight reel generation is closely related to detecting important events in natural videos and instruction videos which is much more well studied. Our proposed method borrows from successful approaches in these other areas, fusing audio and visual cues from gameplay footage. One major difference in this application, however, is that narrated or commentated sports footage contains many more natural language cues than something like a cooking video. This means in place of the more complex audio encoders employed in methods like [1], we can instead simply transcribe natural language to text and use this as an additional input in finetuning a pretrained VLM or video conversation model. We discuss this commentary input more in the dataset section.

With a video-language model that can better understand the complexities of sports matches, we can select clips of sports videos according to user queries. For example, if a user requests all moments a soccer player is offside, this can be discussed and tied back to specific video moments by the finetuned video conversation model.

We propose the usage of the model Qwen-2-VL [24] as the backbone of our pipeline for highlight detection. There are often moments in sports where complex spatial and temporal dependencies come into play, and understanding these has been lacking in typical VLMs and their extensions to video. However, based on Qwen-2-VL's performance on the OpenVLM Leaderboard [10], we have reason to believe that this model will be one of the most performant open-source models for our highlight detection purposes.

As demonstrated in Figure 2, we define a pipeline for generating timestamps based on user queries, which will then be used to trim a video to output the highlights the user wants. The pipeline takes the video and associated commentary transcript along with a user prompt which are fed into the VLM. The VLM generates timestamp pairs (depicting intervals of gameplay) which we parse and use to trim the video to generate the final highlight.

To achieve robust performance on this pipeline, we first fine-tune the model to our specifications as demonstrated in Figure 1. The finetuning pipeline takes the video and associated commentary transcript along with a user prompt which are fed into the VLM. The output of the VLM is then compared with the ground truth from our finetuning dataset to perform supervised finetuning. This is further discussed in the finetuning section below.

B. Dataset

For our experiments, we use the SoccerNet Action Spotting Dataset [4]. Action spotting involves finding all the actions from a predefined set of classes occurring in the videos. This task requires the model to gain a semantic understanding of the different action classes to effectively localize when and which soccer action occurs. This aligns well with our task of highlight detection. The dataset consists of 500 videos from past soccer games along with annotations for common soccer actions such as Penalty, Kick-off, Goal, Substitution, and Offside among 17 total classes. Each action is annotated with game timestamps allowing us to use these as the ground truth to finetune our VLM. The SoccerNet repository also has a dataset that includes commentary for each timestamp in a game [11]. This is an important piece of our pipeline, as when the model is prompted, the relevant commentary will be fed in as a natural language input, to provide extra description and context.

To execute finetuning, we build a finetuning dataset - a collection of individual finetuning samples. Each finetuning sample pertains to a 1-minute video chunk and a specific action (e.g., goal) and asks the model for the timestamps in the video where said action occurs. Finetuning samples are manufactured from SoccerNet Action Spotting labels [4], their corresponding video (also in the SoccerNet dataset), and the commentary transcript for that chunk of gameplay (from [11]). We use the SoccerNet Action Spotting Dataset [4] because it directly provides labels for when highlight actions occurred in match time. An example label from the dataset is shown in Listing 1.

Our finetuning dataset contains 28939 data points, where each data point corresponds to a 1-minute video chunk and an action. We split this dataset into 23151 training and 5788 testing samples. The structure of a sample in our dataset is given in Listing 2 and a complete example is given in Listing 3 in the appendix.

C. Finetuning

Finetuning is a technique that involves adjusting a pretrained model using domain-specific data to enhance

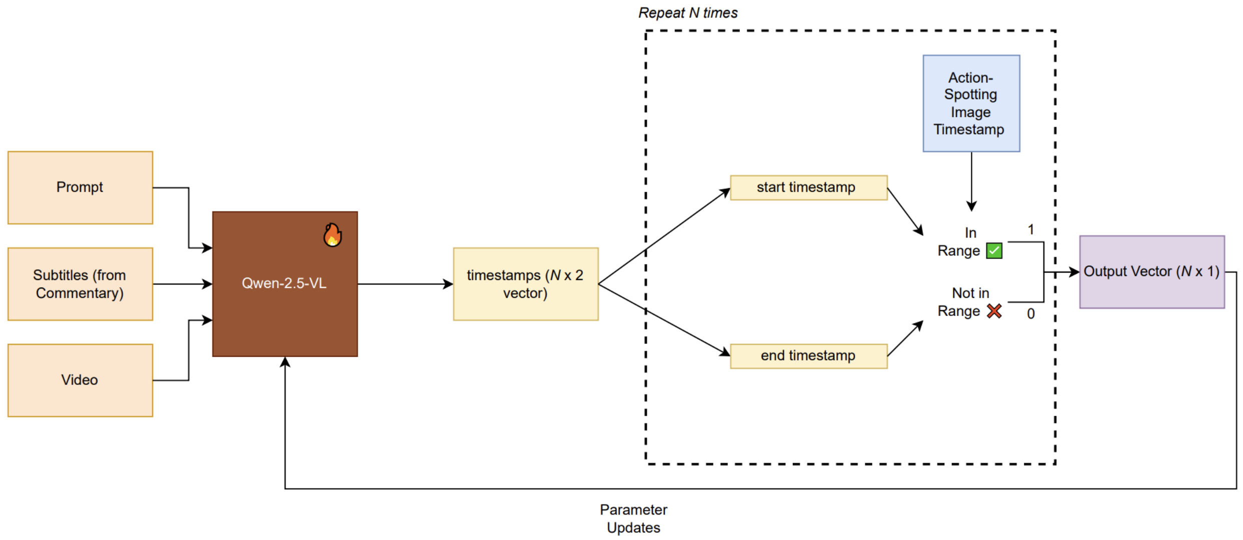

Fig. 1. The finetuning pipeline. As part of the finetuning dataset, a one minute video chunk, its associated commentary, and the prompt are given as input. The VLM learns to output a list of timestamps which is validated by the ground truth timestamps from the Action Spotting dataset. This is used to compute the token-level cross entropy loss, which is then used to update the VLM's unfrozen layer parameters.

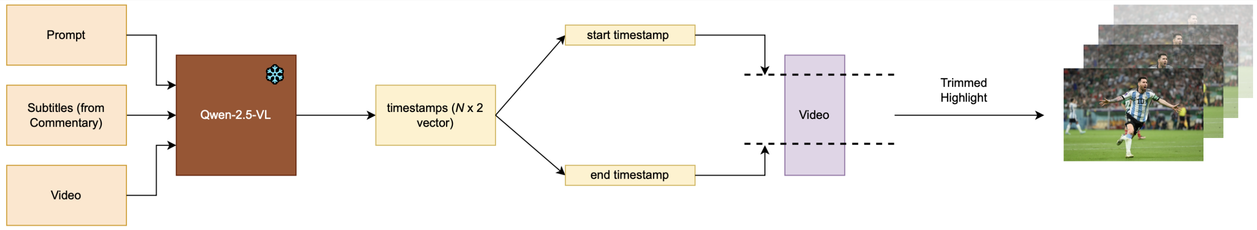

Fig. 2. The inference pipeline. A one minute video chunk, its associated commentary, and a prompt requesting for a specific highlight are given to the VLM at test time. As it learned the action spotting task during finetuning, the VLM produces a list of timestamps as output. These timestamps can be used to trim the video to obtain the desired highlight.

```
{
    "gameTime": "1 - 02:44",
    "label": "Goal",
    "position": "164960",
    "team": "home",
    "visibility": "visible"
}
```

Listing 1. Sample from the SoccerNet Action Spotting dataset

its performance on a targeted task. In our framework, finetuning Qwen-2-VL with soccer match videos accompanied with associated commentary transcripts and match-time labels of key events will help the VLM to better identify and localize highlights, such as goals or fouls. Our final objective is to leverage this improved understanding to enable prompt-driven retrieval of precise highlights from soccer matches. The finetuned VLM should demonstrate superior accuracy compared to an off-the-shelf version of the pretrained VLM for our highlight generation tasks.

The input to the model is the video chunk itself and the associated commentary transcript. A highlight-related query is given as prompt, and a ground truth answer of when the highlight occurred is used for supervised finetuning. An example of a finetuning sample is given in Listing 3 in the Appendix. We also take advantage of prompt engineering techniques and in-context learning examples, by providing a detailed system prompt to the model (passed in as part of the query). The system prompt is given in Listing 4 in the Appendix.

We ran our finetuning on a single NVIDIA A40 GPU on the Great Lakes computing cluster. Unfortunately, there were many limitations with this, turning the finetuning process into more of an engineering problem than research problem. Although we were able to run

```
"id": "2016-09-13 - 21-45 Barcelona 7 - 0
    Celtic/chunk_0.mp4/2",
"video": "chunk_0.mp4",
"conversations": [
  {
    "from": "human",
    "value": "<video> Here is match
        commentary for a 1 min segment of a
        match <commentary> What are all the
        match times that goals occur?"
  },
  {
    "from": "gpt",
    "value": "A goal occur at (02:39, 02:49)"
  }
]
```

Listing 2. A sample from our finetuning dataset

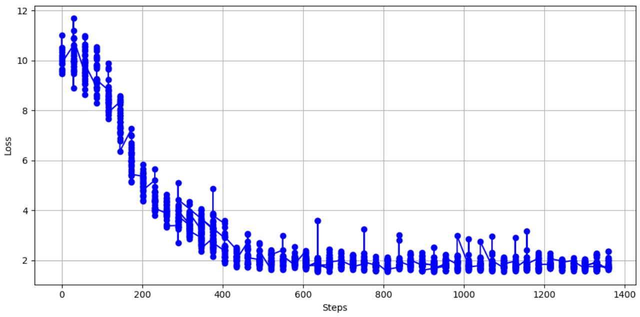

Fig. 3. Finetuning loss across training steps

inference on `Qwen2.5-VL-7B-Instruct` (the 7 billion parameter model) with 5-minute video chunks using this hardware, we had to severely quantize and downsize our model and video length to be able to finetune without CUDA OOM issues on the single available GPU. As we were unable to run finetuning on `Qwen2.5-VL-2B-Instruct` with 1-minute chunks and since the `autoawq` module lacks support for the `transformers` version containing the newer `Qwen2.5-VL-2B-Instruct-AWQ` model, we instead settled on finetuning the slightly older, quantized `Qwen2-VL-2B-Instruct-AWQ` model for 1-minute video chunks at 1 FPS. Moreover, to further allow our system to work under the Great Lakes cluster constraints, we used QLoRA [8]. QLoRA is a memory efficient fine-tuning technique that allows the tuning of a subset of a model's full parameters, by making use of low-rank adapters and quantization. It has been shown to approximate the results one may achieve with a full parameter fine-tune, except at a much reduced memory and computational cost.

We utilized Hugging Face's Transformer Reinforcement Learning framework [23]—specifically its Supervised Finetuning Trainer—as our finetuning implementation using standard token-level cross-entropy loss. In this setup, given an input sequence of tokens $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ from a vocabulary $\mathcal{V}$, the language model is trained to maximize the likelihood of each next token conditioned on the preceding ones. The model defines a probability distribution $P_\theta(x_t \mid x_{<t})$ over the vocabulary at each time step t, where θ are the model parameters and $x_{<t} = (x_1, \ldots, x_{t-1})$.

The objective is to minimize the negative log-likelihood over the training dataset, which is equivalent to minimizing the cross-entropy loss:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})$$

This loss quantifies the discrepancy between the true token distribution (which is a one-hot vector for the ground truth token) and the model's predicted distribution. More formally, for a batch of N sequences $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$, the loss becomes:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T^{(i)}} \log P_\theta \left(x_t^{(i)} \mid x_{<t}^{(i)} \right)$$

This formulation ensures that the model is optimized to assign high probability mass to the ground-truth next tokens across all sequences in the batch, thereby improving next-token prediction accuracy.

Lastly, as job wall-times were capped at 8 hours on Great Lakes, we split up finetuning over many repeated jobs, checkpointing after every 100 steps so that the next job can resume training from the checkpoint. We ran finetuning for a single epoch with 1360 steps with 8 samples per step. The loss curve during finetuning is given in Figure 3.

D. Evaluation Criteria

Assuming our model outputs a list of discrete start and end timestamps $[(s_1, e_1), \cdots, (s_m, e_m)]$ for a given query on some SoccerNet [4] video, we use SoccerNet's *Action Spotting Dataset* to construct the ground truth: a list of timestamps $[h_1, \cdots, h_n]$ for the action corresponding to the query (e.g., "goals by the red team"). At a high-level, our evaluation method must:

- **Reward successfully captured highlights**: For each i, if h_i falls within (s_j, e_j) for some j, reward with a positive score.

- **Penalize missed highlights (false-negatives)**: For each i, if h_i does not fall within any (s_j, e_j), apply a penalty.
- **Penalize false-positives**: For each j, if there is no h_i that falls between (s_j, e_j), apply a penalty.
- **Penalize non-sensible durations**: Highlight clips must neither be too long nor too short.

To satisfy the above criteria, we use a many-to-many IoU evaluation function. Let the model output a set of predicted intervals:

$$\mathcal{P} = \{(s_1, e_1), (s_2, e_2), \ldots, (s_m, e_m)\}$$

and let the ground-truth action timestamps be:

$$\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$$

To compute temporal Intersection-over-Union (IoU), we convert each ground-truth timestamp h_j into a ground-truth interval of fixed length l, centered at h_j:

$$\mathcal{G} = \left\{ \left(h_j - \frac{l}{2}, h_j + \frac{l}{2} \right) \right\}_{j=1}^{n}$$

For a predicted interval (s_i, e_i) and a ground-truth interval $(g_j^{(s)}, g_j^{(e)})$, the temporal IoU is defined as:

$$\text{IoU}(i, j) = \frac{\max(0, \min(e_i, g_j^{(e)}) - \max(s_i, g_j^{(s)}))}{\max(e_i, g_j^{(e)}) - \min(s_i, g_j^{(s)})}$$

Given an IoU threshold θ, we define the set of matching pairs:

$$M = \{(i, j) \mid \text{IoU}(i, j) \geq \theta\}$$

This many-to-many matching scheme allows multiple predictions to match multiple ground-truth intervals. We define:

$$\begin{aligned} \text{TP} &= |M| \\ \text{FP} &= |\{i \in \{1, \ldots, m\} \mid \forall j,\ \text{IoU}(i, j) < \theta\}| \\ \text{FN} &= |\{j \in \{1, \ldots, n\} \mid \forall i,\ \text{IoU}(i, j) < \theta\}| \end{aligned}$$

From these, we compute standard detection metrics: Precision, Recall, and F1 score. In our implementation, we choose the threshold $\theta = 0.3$.

V. Experimental Setup

We use a traditional LLM and an LLM reasoning model to compare with the VLM. The LLM we use, `Mistral-7B`, is an open-weight language model designed for high efficiency and strong reasoning [13]. It is built with Grouped Query Attention and Sliding Window Attention which enables faster inference and longer context handling. This makes it well-suited to sparse action spotting in memory constrained environments.

We use `o4-mini` as the reasoning model [16]. This is OpenAI's latest lightweight reasoning model (released April 16, 2025). Though its exact parameter count is undisclosed, it's designed as a compact yet powerful successor to the `o3` series.

Each model answered the prompts from the test subset of the finetuning dataset, and the outputs of each model were compared to their corresponding ground truths using the evaluation criteria described in Section IV-D. While the VLM is given both the video and commentary as input, both `Mistral-7B` and `o4-mini` rely exclusively on the commentary transcript to isolate highlights.

VI. Results and Discussion

Table II shows that across all models, precision, recall, and F1 scores are low, suggesting that a purely token-based approach to the task of extracting intervals for key actions from a long video is a challenging task for VLMs and LLMs. However, despite having the least parameters of all tested models, the finetuned Qwen2-VL-2B scored the highest precision, recall and F1 score on our experiment. This result suggests that the finetuned Qwen2-VL-2B is able to extract relevant information from video input, and that finetuning may have helped Qwen2-VL-2B in better understanding how actions appear in broadcast soccer video. Given this outcome, we hypothesize that a larger and more recent version of the Qwen-VL model with access to more hardware resources for finetuning would produce better results.

Also, we note that for this task, false negatives are worse than false positives – flagging extraneous intervals for key actions is less of an issue than entirely missing key actions, so we do not find recall being significantly higher than precision to be a major issue. However, given that precision is quite low, we can conclude that the models are labeling nearly all intervals in a video as the key action being queried from the prompt. To address this issue we believe that using shorter chunks and more frames per second could be beneficial for a VLM to understand what chunk of video actually constitutes the action in question.

Table III highlights which actions each model was the most effective at identifying. Interestingly, o4-mini was relatively successful in identifying substitutions and goals, purely from commentary, compared to the other two models (F1 scores of 0.2667 and 0.2222 respectively). We reason that o4-mini's success with goals and substitutions (and lack of success with other actions) is due to the nature of commentary and the model's superior language understanding combined with reasoning ability. Events like goals and substitutions are often clearly discussed by commentators when they occur, so it is understandable that the model with the best language understanding would be better at picking up actions that are often accompanied by descriptive commentary. The fact that Qwen2-VL-2B did not excel on these actions that are clearly described in commentary, despite having commentary input as well, emphasizes that the finetuned VLM is relying on contents in the video input to decide what constitutes a key action. We take this as evidence that the finetuned VLM's enhanced video understanding is the difference-maker in identifying actions with minimal commentary.

Model	Precision	Recall	F1 Score
`Qwen2-VL-2B`	**0.0324**	**0.1795**	**0.0549**
`o4-mini`	0.0207	0.1132	0.0350
`Mistral-7B`	0.0186	0.1136	0.0320

TABLE II

OVERALL PERFORMANCE METRICS OF EACH MODEL.

Model	Action 1	Action 2	Action 3
`Qwen2-VL-2B`	shots on target (0.1052)	goal (0.0889)	substitution (0.0559)
`Mistral-7B`	goal (0.0889)	corner (0.0407)	shots on target (0.0387)
`o4-mini`	substitution (0.2667)	goal (0.2222)	shots on target (0.0541)

TABLE III

TOP 3 PREDICTED ACTIONS FOR EACH MODEL WITH F1 SCORES DISPLAYED BENEATH ACTION NAMES.

VII. LIMITATIONS AND FUTURE WORK

There are several limitations of our study that should be discussed. First, the main limitation is the resource constraints we were under in the Great Lakes environment. As previously discussed, this meant we had to use a quantized version of Qwen-2-VL-2B. However, we firmly believe that our results would significantly improve if we had the GPU and memory resources to be able to fine-tune the SOTA Qwen-2.5-VL-72B model. The 36× increase in parameter count is a clear path for improvement.

A separate limitation of our study is the assumptions we make in how we want outputs to be returned. We implicitly assume that the VLM uses the "score bug" in the top left corner of the video screen along with the timestamps in the commentary transcript to understand the internal game time. But, there is debate on how well language models reason with temporal dependencies and concepts in video, and if it is even possible for them to understand time in certain contexts [9]. We suggest that our results may improve by passing in a separate video clip of the cropped score bug to the VLM, to hopefully force the model to attend to and reason about internal game time.

VIII. CONCLUSION

We proposed and developed a novel, completely token-based VLM pipeline for sparse action spotting. Specifically, we used Qwen-2-VL-2B as a baseline model to conduct highlight interval identification experiments on soccer game footage from the SoccerNet dataset. Our methodology and results demonstrate that while there is a way to go before fully token-based highlight generation pipelines are viable, there seems to be promise in the pathway, as small VLMs surpass much larger LLMs on returning the time frames of key actions.

AUTHOR CONTRIBUTIONS

The entire team ideated and finalized the high-level approach. Evan and Pranav developed the finetuning dataset creation script and Naveen developed the script to generate chunked videos. Evan, Pranav, Nicholas, and Vishal worked on the finetuning strategy. Naveen and Nicholas ran finetuning and evaluation for Qwen-2. Pranav and Evan setup and ran evaluation for Mistral. Nicholas set up and ran evaluation for o4-mini. Naveen and Vishal focused on and wrote scripts for result analysis and evaluation. All co-authors contributed to the progress report write-up. All co-authors equally contributed to this project.

REFERENCES

[1] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. Tim: A time interval machine for audio-visual action recognition, 2024.

[2] Shimin Chen, Chen Chen, Wei Li, Xunqiang Tao, and Yandong Guo. Faster-TAD: Towards temporal action detection with proposal generation and classification in a unified network. *preprint arXiv:2204.02674*, 2022.

Contribution	Evan	Naveen	Nicholas	Pranav	Vishal
Ideation	✓	✓	✓	✓	✓
Finetuning Strategy	✓		✓	✓	✓
Finetuning Dataset Creation	✓	✓		✓	
Qwen-2 Finetuning and Evaluation		✓	✓		
Mistral Evaluation	✓			✓	
o4-mini Evaluation			✓		
Result Analysis / Evaluation	✓	✓			✓
Report Write-Up	✓	✓	✓	✓	✓

TABLE IV
PROJECT CONTRIBUTIONS BY TEAM MEMBERS

[3] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas Moeslund. A context-aware loss function for action spotting in soccer videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519, 2021.

[5] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2 : A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.

[6] Francesco Della Santa and Morgana Lalli. Automated detection of sport highlights from audio and video sources. *arXiv preprint arXiv:2501.16100*, 2025.

[7] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers, 2023.

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

[9] Xi Ding and Lei Wang. Do language models understand time? *arXiv preprint arXiv:2412.13845*, 2024.

[10] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

[11] Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A. Riegler, Pål Halvorsen, and Mubarak Shah. Soccernet-echoes: A soccer game audio commentary dataset, 2024.

[12] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[14] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. Videogenic: Identifying highlight moments in videos with professional photographs as a prior. In *Proceedings of the 16th Conference on Creativity & Cognition*, CC '24, page 328–346, New York, NY, USA, 2024. Association for Computing Machinery.

[15] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[16] OpenAI. Openai o4-mini system card, 2024.

[17] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, page 105–115, New York, NY, USA, 2000. Association for Computing Machinery.

[18] Yoonho Shin, Sanghoon Park, Youngsub Han, Byoung-Ki Jeon, Soonyoung Lee, and Byung Jun Kang. Soccer-clip: Vision language model for soccer action spotting. *IEEE Access*, 13:44354–44365, 2025.

[19] João V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors, 2022.

[20] Yunzhuo Sun, Yifang Xu, Zien Xie, Yukun Shu, and Sidan Du. Gptsee: Enhancing moment retrieval and highlight detection via description-based similarity features. *IEEE Signal Processing Letters*, 31:521–525, 2024.

[21] Qwen Team. Qwen2.5-vl, January 2025.

[22] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. RMS-Net: Regression and masking for soccer event spotting. In *International Conference on Pattern Recognition (ICPR)*, 2021.

[23] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi

Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[25] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recognition Letters*, 25(7):767–775, 2004.

[26] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and Transformer based temporal detection. *arXiv preprint arXiv:2106.14447*, 2021.

Appendix

```
{
  "id": "germany_bundesliga/2015-2016/2015-09-26 - 16-30 1. FSV Mainz 05 0 - 3 Bayern
      Munich/elephant/2_037.mkv/84",
  "video": "germany_bundesliga/2015-2016/2015-09-26 - 16-30 1. FSV Mainz 05 0 - 3 Bayern
      Munich/elephant/2_037.mkv",
  "conversations": [
    {
      "from": "human",
      "value": "<video>\nHere is match commentary for a 1 min segment of a match [[’36:58’,
          ’37:01’, ’also with a goal difference of plus 20.’], [’37:01’, ’37:09’, ’That
          would be the best record in Bundesliga history after seven games by any team.’],
          [’37:09’, ’37:11’, ’has reached.’], [’37:11’, ’37:19’, ’Bayern have already
          managed seven wins twice, in 1995, 1996 and 2012, 2013.’], [’37:19’, ’37:25’, ’But
          the goal difference would be even better this year than in the first two.’],
          [’37:25’, ’37:35’, ’Borussia Dortmund would of course have no chance of taking the
          lead for Bayern’], [’37:35’, ’37:39’, \"We’ll be back on matchday seven against
          Darmstadt tomorrow.\"], [’37:39’, ’37:41’, ’Eckball Brosinski.’], [’37:41’,
          ’37:53’, ’Brosinski wanted to tunnel there for the lamb, but of course he brings a
          lot’], [’37:53’, ’37:55’, ’Experience and routine with.’], [’37:55’, ’37:59’,
          \"It’s not that easy for the Bayern captain to just put the ball between the two
          of them.\"]]\nUtilize the video clip and commentary of this segment to accurately
          find all the match times that Free-kick occurs in this segment."
    },
    {
      "from": "gpt",
      "value": "A Free-kick occurs at (37:05, 37:17)."
    }
  ]
}
```

Listing 3. A complete finetuning sample

```
You are a Vision Language Model specialized in identifying an action in a soccer
   match from a fixed set of action classes as it occurs in a given soccer video.
   The action classes are corner, shots on target, goal, clearance, foul,
   free-kick, and substitution. Your task is to observe the input video and
   commentary carefully and respond to the prompt. Prompts will ask for the match
   times of an action. You are to respond with a series of sentences that describe
   the time (start, end), in real match time (minutes:seconds), when the action
   occurred. If the action does not occur in the input video, simply state that in
   your response. The video contains broadcast footage from soccer games between
   players of two teams distinguished by their team uniform. The commentary will
   come as a list of comments in the format [start_time, end_time, ’comment’], and
   time will be in the format minutes:seconds. Focus on delivering accurate
   timestamps based on the live game time displayed in the video. Absolutely avoid
   additional explanation. Here are some example prompts and answers in the format
   you are expect to follow:

Prompt: <video>Here is match commentary for a 1 min segment of a match
   <commentary>. Utilize the commentary and video clip of this segment to
   accurately find all the match times that Goals occur in this segment. Answer: A
   goal occurs at (2:15, 2:27).
Prompt: <video>Here is match commentary for a 1 min segment of a match
   <commentary>. Utilize the commentary and video clip of this segment to
   accurately find all the match times that Shots on target occur in this segment.
   Answer: A shot on target occurs at (2:15, 2:27).
Prompt: <video>Here is match commentary for a 1 min segment of a match
   <commentary>. Utilize the commentary and video clip of this segment to
   accurately find all the match times that Corners occur in this segment. Answer:
   A corner occurs at (0:10, 0:22).
Prompt: <video>Here is match commentary for a 1 min segment of a match
   <commentary>. Utilize the commentary and video clip of this segment to
   accurately find all the match times that Goals occur in this segment. Answer: A
   goal occurs at (1:05, 1:18). A goal occurs at (1:45, 1:55).
Prompt: <video>Here is match commentary for a 1 min segment of a match
   <commentary>. Utilize the commentary and video clip of this segment to
   accurately find all the match times that Fouls occur in this segment. Answer: A
   foul occurs at (2:16, 2:23). A foul occurs at (2:44, 2:52).
```

Listing 4. VLM System Prompt Template