MLLM-Driven Highlight Reel Generation for Ultimate Frisbee Games

Farah Shahbaz Heather Szczesniek Megan Ja
fshahbaz@stanford.edu, hsz@stanford.edu, meganja@stanford.edu

Abstract

Manually editing highlight moments such as layouts in ultimate frisbee games is a time-consuming and labor-intensive process, especially for long videos exceeding 1 hr like frisbee matches. Layouts, where players dive to catch for the disc, are among the most exciting and sought-after actions to capture in a frisbee game. This project presents a proof of concept for an automated approach that uses multimodal large langauge model (MLLM) with adaptive keyframe sampling (AKS) to detect layout highlights in frisbee games based on carefully designed prompts. The multimodal model performance is compared to a traditional object detection YOLOv8 model. Our mutlimodal model outperformed YOLOv8, achieving a perfect recall (1.000) and the highest F1-score (0.583) for the prompt recognizing diving actions toward a frisbee as the key layout indicator. The multimodal model performance significantly improved with video grid splitting, but demonstrated sensitivity to linguistic variations of prompts. In contrast, YOLOv8 model struggled with detecting small, fasting-moving objects such as frisbee discs, particularly in low-resolution or occluded objects. These findings highlight the potential of multimodal models to enhance the reliability of highlight detection in sports videos, while significantly reducing editing time.

1. Introduction

In today's fast-paced digital world, highlight reels for sports leagues have become an essential marketing tool to maintain fan engagement on social media. However, creating these highlight reels is a time-consuming and labor-intensive process that requires hours of manual editing of footage. In addition, expert knowledge of sports teams and rules is required to identify key moments in a game. Our project aims to automate the generation of highlight reels for ultimate frisbee games by detecting layout highlights using a Multimodal Large Language Model (MLLM) integrated with a novel adaptive keyframe sampling (AKS) method. A layout is when a player dives for the disc and can be one of the coolest and most exciting moments in a frisbee game. Automating highlight selection for frisbee games has not been attempted before, which offers a promising solution to save hours of manual editing while improving the quality and reliability of layout detection.

Automating layout detection in ultimate frisbee games faces two core challenges. First, the dynamic nature of layouts involving rapid body movements like jumping, diving, and catching disc requires precise disc detection and motion recognition, which existing object detection like YOLO models and pose estimation models often struggle with. Second, the length of frisbee games demands long-video analysis, which exceeds the token limits of standard MLLMs. This is usually tackled by oversimplifying sampling of clips that could significantly overlook key moments. To address these challenges, we proposed a multimodal framework combining video-based MLLMs with AKS to automatically detect layout highlights in frisbee games. The multimodal model detected layouts based on carefully crafted prompts fed into the MLLM without requiring object labels. This mutlimodal approach is also compared to the performance of layout detections using YOLOv8 model to demonstrate the limitations of traditional object detection models in detecting frisbee discs.

2. Related Works

Previous CS231n course projects have provided relevant references for our topic of highlight generation. The automatic game highlight detection project employed 2D Convolutional Neural Network (CNN), 3D Residual Network CNN (ResNet3D), and Vision Transformer (ViT) models to detect soccer highlights using manually labeled clips from full-length games [1]. The results show that all models achieved moderate recall, but low precision due to the complexity of event detection in dynamic environments like soccer, which could inform similar performance in frisbee layout detection [1]. The rock-climbing pose estimation project used a vision transformer for pose estimation (ViTPose) and YOLOv8 for human detection to analyze rock climbing techniques from videos [2]. The project results show high precision values, but low recall values due to failures in pose estimation [2].

Building upon insights from previous course projects, it is evident that computer vision-based methods using deep

learning for object detection and pose estimation face several limitations when applied to dynamic sports like frisbee layouts. Various factors are reported in the literature that make the task of detecting and tracking both players and balls very difficult. These include similar appearances of objects, complex occlusions, varying background, lower pixel resolution of distant and smaller objects in the frame, unpredictable movements, unstable camera motion, and motion blur [3]. These challenges are directly applicable to frisbee where players often move rapidly, layouts can be obscured, and frisbee discs can be blurred or too small to detect. The model combining object detection with pose estimation to identify successful layouts might fail due to the mentioned reasons.

Given the limitations of object detection and pose estimation models in sports, exploring alternative approaches to detecting highlights in videos becomes imperative. One promising direction is the adoption of multimodal models that integrate different data types such as visual, audio, and text. Such multimodal models have been investigated in highlight detection in sport, including a multimodal system using 2D CNNs on Mel-spectrogram for audio and grayscale video frames to detect highlights in soccer games [4]. This approach achieved 89% accuracy for audio-based detection and 83% for video-based detection [4]. To enhance performance, the paper developed an ensemble model that averages the audio and video scores to effectively reduce both false positive and false negative detections compared to single-modality models [4]. Other efforts include introducing MLLM models such as the Highlight-CLIP (HL-CLIP) that leverages the pretrained CLIP model to improve highlight detection in videos [5]. HL-CLIP fine-tunes CLIP's vision and text encoders to detect video highlights by predicting saliency scores between video games and text queries [5]. However, processing long videos such as frisbee games remains challenging due to computational constraints with CLIP's ViT encoder or even 2D CNN audio and text ensemble models that only work effectively on short clips.

Video-based MLLMs mainly sample a small number of tokens from input data to not exceed the maximal token limit of MLLMs. To mitigate this length constraint, the Adaptive Keyframe Sampling (AKS) method has been proposed to maximize the useful information during keyframe selection in analyzing videos with MLLMs [6]. This is done by maximizing the relevance between the keyframes and the prompt as well as the temporal coverage of keyframes throughout the long video [6]. The AKS method has significantly reduced the computational cost, which allows multimodal models to process long videos more efficiently, while preserving key useful information [6]. In our project, we use the AKS method with video-based MLLMs to automate highlight detection and reel generation in frisbee games focusing on identifying key layout moments.

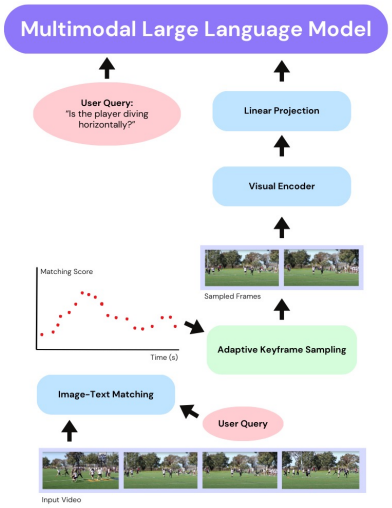

Figure 1: Schematic diagram for the multimodal framework combining video-based MLLMs with AKS to detect layout highlights in frisbee games

3. Methods

Our method was based on a multi-modal framework, as shown in Figure 1, combining video-based MLLMs with AKS to first detect highlight layouts in frisbee games. The training starts by feeding the raw videos of the frisbee games into the image-text encoders of BLIP. The raw video was trained at 1 fps as recommended by the adaptive keyframe sampling paper to reduce computational costs. The role of the image-text encoder is to compute a relevance score for each candidate frame against the query prompt. The adaptive keyframe selection (AKS) algorithm then acts as a smart filter before the MLLM step by selecting the most useful keyframes. This is done by maximizing the relevance and coverage scores. Based on recursive temporal binning, the algorithm splits the video into segments and allocates keyframe slots proportionally for each segment. The algorithm then performs hierarchical selection by selecting frames with highest relevance scores in each segment. If the segment has no high-scoring frames, then the algorithm redistributes the slots to action-rich segments. To match the token limit of the MLLM LLaVA-Video model, the AKS algorithm is run to select 64 frames from each video.

The most critical part of our multimodal framework was designing the prompts to guide the MLLM LLaVA model. The keyframes selected by the AKS algorithm were fed into the LLaVA-Video model with task-specific prompts. This step requires careful prompt engineering to let the MLLM focus on layout-specific features. The prompts were itera-

Table 1: Description of video prompts and experimental setup

#	Prompt	Video Length (min)	Video Grid
P1	Is someone diving to try to catch a frisbee?	5	None
P2	Is any player on the field horizontal?	5	None
P3	Is an ultimate frisbee player doing a layout?	5	None
P4	Is there someone diving for the frisbee and, if so, are they on offense or defense?	5	None
P5	Is any player's body parallel to the ground?	5	None
P6	Is someone diving towards a frisbee?	5	None
P7	Do two people jump high to compete to catch the frisbee?	5	None
P8	Is someone diving towards a frisbee?	5	split into 4 quadrants
P9	Do two people jump high to compete to catch the frisbee?	5	split into 4 quadrants
P10	Is someone diving towards a frisbee?	30	split into 4 quadrants

tively refined by gradually introducing specificity and context as shown in Table 1. The model was trained first on broad prompts aimed at general layout detection including: "P1: Is someone diving to try to catch a frisbee?" and "P2: Is any player on the field horizontal?". The refined prompts then added more specific context including: "P4: Is there someone diving for the frisbee and, if so, are they on offense or defense?". In addition, "P7: Do two people jump high to compete to catch the frisbee?" was tested to capture a different type of highlight known as skies where two players jump in competition to catch the disc. The responses to these prompts were then used to identify and extract timestamps from key layout moments, which could be subsequently aggregated for highlight video compilation.

In total, 10 prompts (Yes or No questions) were tested in this project, 7 of which were unique (P1 -P7). The remaining 3 prompts (P8 - P10) were repeated across different experimental setups to evaluate the effect of video length and grid resolution on our multimodal model (explained in section 4). The prompts were intentionally designed to capture a range of visual tasks needed to detect layouts, including action recognition, assessing spatial orientation, and object interaction. For example, P1 targets action recognition of someone diving to a catch a frisbee, P2 attempts to assess the spatial orientation of a player's body in relation to the field, and P7 focuses on object interaction between two players competing to catch the frisbee. These prompts were also designed to test the prompt sensitivity to linguistic variations. For example, "P2: Is someone diving to catch a frisbee?" versus "P6: Is someone diving towards a frisbee?" and "P2: Is any player on the field horizontal?" versus "P5: Is any player's body parallel to the ground?".

The layout detection pipeline was implemented by adapting the Adaptive Keyframe Sampling (AKS) for Long Video Understanding algorithm from its official GitHub repository to fit the specific demands of our project [7]. AKS was particularly well-suited for our application in detecting layout highlights in frisbee games because it strategically balances relevance and coverage. This approach ensures that the final keyframe set is both semantically rich and temporally diverse, which are critical qualities for generating reliable responses from the MLLM. Therefore, AKS was selected over uniform or random sampling because it is proven to prioritize rare, high-value events, while minimizing redundancy in the visual input [6].

To evaluate the effectiveness of our multimodal model with AKS, we conducted a comparative experiment using YOLOv8 (You Only Look Once), a real-time object detection model.YOLO detects objects in a single forward pass by dividing each frame into a grid and predicting the bounding boxes and class probabilities for each cell simultaneously [8]. We trained YOLOv8 on layout videos to assess the performance of the model in detecting players and frisbee discs during layout moments. This comparative experiment allowed us to assess the limitations of traditional object detection models in capturing dynamic actions like layouts against our proposed multimodal approach with AKS sampling.

The performance of our multimodal model to detect key layout highlights was evaluated using quantitative metrics computed based on the manually annotated dataset. A confusion matrix was constructed to assess layout detection (successful vs. unsuccessful). The accuracy, precision, recall, and F1-score were computed as our main quantitative metrics to evaluate our model as shown in Eq.1-4. Although these metrics provide qualitative performance measures, it is critical to acknowledge that their reliability inherently depends on the qualitative judgment of the human annotators who labeled the ground truth. Our evaluation also includes qualitative analysis of baseline YOLOv8 experiments to emphasize on the limitations of traditional detection models for sports highlight generation, specifically for frisbee games.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

4. Dataset

The main training data of our multimodal layout detection model was collected from the Mixed Final of the 2024 USA Ultimate National Championships featuring the match between Hybrid and Sprocket [9]. Due to the time constraint of the project, we only extracted two test videos of different durations including 5 min video and 30 min video to analyze the performance of our model for different video lengths. Each video was segmented into 5-second clips to provide enough time to capture the layout moments, but also keep it short enough to avoid excessive content. For evaluation purposes, a dataset was manually prepared by annotating the 5-second clips in both the 5 min video and 30 min video to establish the ground truth. This is done by manually labeling successful layouts and no layouts to each 5-second clip. Prompts P1 - P9 were tested using the 5 min video that contained 6 successful layouts, while prompt P10 was tested using the 30 min video that contained 24 successful layouts.

For prompts P8 - P10, we introduced an additional processing step to investigate the effect of video grid resolution on the performance of the model. In this experiment setup, the full video frame was divided into 4 quadrants and each quadrant was split into 5-second clips. This grid-based approach aimed to increase the grid resolution, especially for detecting smaller, fast-moving objects like a frisbee disc. The underlying mechanism was that in most successful layouts, the action of a layout and the disc should appear in at least one quadrant of a given keyframe.

To validate the need for an alternative multimodal approach, we conducted two trials with YOLOv8 to detect players and discs. The first trial was trained on vertical YouTube shorts of layout shots in frisbee games [10]. Different data augmentation was experimented on the first trial including zooming, cropping, and contrast adjustment. The second trial was trained on higher resolution horizontal videos (1080p) [11]. The successful detection of players and frisbee discs was qualitatively evaluated for both trials at different confidence levels.

5. Experiments

5.1. Multimodal Model with AKS

Table 2 shows the performance metrics including accuracy, precision, recall, and F1-score for all prompts tested to detect layout highlights in frisbee games using the multimodal model with AKS. Accuracy serves as the overall general indicator of the model's effectiveness by measuring the ratio of correctly predicted clips to the total number of clips. The accuracy across prompts range from a low of 0.138 (P4) to a high 0.957 (P8), which suggests that some prompts were significantly more effective in guiding the MLLM to detect specific layout features.

Table 2: Performance metrics for all prompts

Prompt #	Accuracy	Precision	Recall	F1-score
P1	0.879	0.400	0.333	0.364
P2	0.897	-	-	-
P3	0.862	0.333	0.333	0.333
P4	0.138	0.077	0.667	0.138
P5	0.897	-	-	-
P6	0.879	0.429	0.500	0.462
P7	0.828	0.000	0.000	0.000
P8	**0.957**	**0.412**	**1.000**	**0.583**
P9	0.914	0.136	0.750	0.231
P10	0.924	0.040	0.167	0.065

Nearly all prompts achieved accuracy scores above 0.800, except for P4 "Is there someone diving for the frisbee and, if so, are they on offense or defense?". This prompt also had a significantly high number of false positives (48) compared to the rest of the prompts, as shown in the confusion matrix in Figure 2. This could be due to the fact that Prompt P4 is compound and semantically more complex than the rest, as it not only requires the model to detect the diving action, but also to infer the player's role (offense or defense). Therefore, this suggests the model's limitation in handling compound prompts and understanding role recognition within dynamic events like frisbee.

The results overall show the critical role of prompt engineering in achieving reliable model performance not only in terms of accuracy, but also in precision, recall, and F1-score. Precision measures the ratio of true positive predictions compared to all positive predictions, while recall measures the ratio of true positive predictions compared to all actual positive labels. Prompts P2 and P5, which aimed to assess the spatial orientation of a player's body in relation to the field, failed to detect any layouts. As a result, precision, recall, and F1-scores were not computed for these prompts because they produced neither true positives nor false positives in their respective confusion matrices in Figure 2. This outcome suggests that spatial body orientation might not be a significant factor to be considered for detecting layouts using MLLMs.

Prompts focused on diving actions, such as P1, P6, and P8 achieved higher precision and recall compared to P3, which specifically asked about layouts. This suggests that

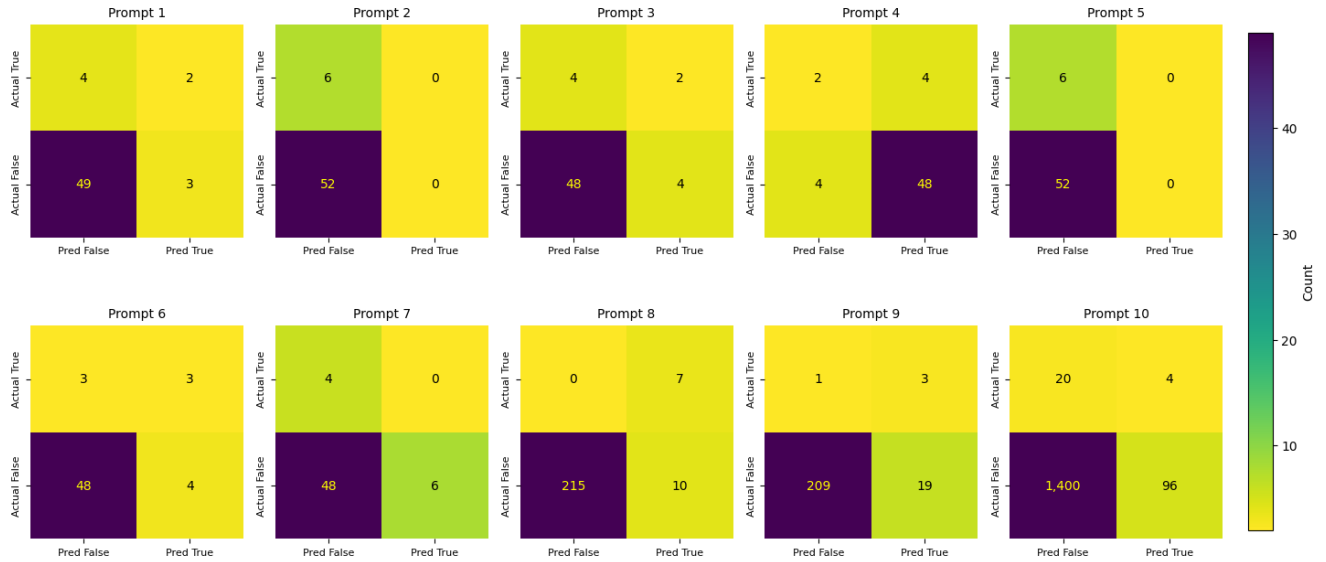

Figure 2: Confusion matrices for all prompts

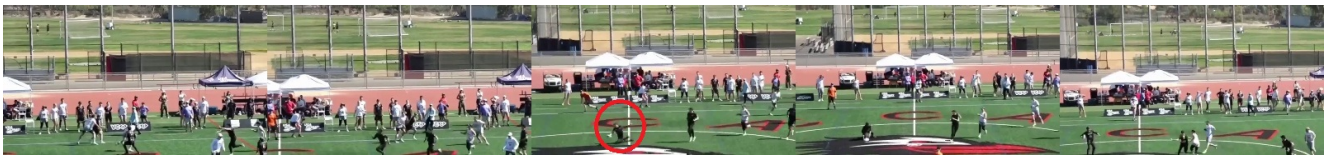

Figure 3: Successfully detected layout using the multimodal model with AKS for prompt P8

the model is more effective at identifying general movements like diving rather than more context-dependent actions such as layouts in frisbee. In addition, the linguistic variation between P1 ”Is someone diving to try to catch a frisbee?” and P6 ”Is someone diving towards a frisbee?” highlights the model’s sensitivity to prompt phrasing. While both prompts target diving actions, P6’s more general phrasing led to improved precision and recall compared to P1. This indicates that even subtle differences in wording could influence how effectively the model responds to prompts.

Recall, which is a measure of the true positive rate, is a critical metric for the task of highlight detection. The best-performing model on the 5 min video with no grid splitting was prompt P6 and yet it only achieved a recall of 0.500. Upon closer inspection to the layout frames, we observed that prompt P6 mainly detected layouts that are zoomed in closest to the camera. This finding motivated us to experiment with splitting the video grid into 4 quadrants, which significantly improved both the accuracy and recall of prompts P8 and P9 compared to P6 and P7, respectively.

Prompt P8 significantly outperformed all other prompts including P9 in the grid-split videos with highest recall of 1.000 and highest F1-score of 0.583. While P9 targeted a different type of highlight, skies, it did not perfrom as well as P8, which focused on diving actions. This performance gap may indicate that layouts are more visually distinctive and prominent within the gameplay, making them easier for the model to detect. The results suggest that diving related actions are more reliably captured by the model than other highlight types like skies, which might be more context-dependent.

Prompt P8 also interestingly detected 7 layouts out of the actual 6 present in the 5 min video as shown in the confusion matrix in Figure 2. This happened because one layout spanned across two grid quadrants in the frame that led to this duplicate detection. However, this minor duplication could be resolved by implementing a simple post-processing rule that considers only one layout detection per set of 4 grid quadrants in a frame. This will ensure that overlapping detections are merged into a single event. The model for prompt P8 only mislabeled 10 out 225 clips as shown in the confusion matrix in Figure 2. Upon inspecting the false positives, we found that 3 out of 10 were instances considered as ”skies” moments when two players jump very high in competition for the disc. This indicates that the model sometimes mistook another common highlight for a layout, which is still worthy and valuable for editors creating highlight reels.

The model for prompt P8 detected 16 clips (equivalent to 1 min and 20 sec) as potential layouts in the 5 min video, 6 of which were true layouts. This means a video editor would only need to review 26.6% of the total footage to find all layout moments, which significantly reduces the time spent reviewing film to capture key highlights. This level of performance could be highly useful for content creators making highlight reels. In addition, Figure 3 shows an example of a successfully detected layout within a 5 sec clip in the red circle for prompt P8. While the underlying clip was divided in grids during processing to test P8, the clip shown in Figure 3 is an aggregated segment to provide a clear view of the detected highlight.

The video length had a significant impact on multimodal model performance. For example, prompt P8 achieved a strong precision and recall on a 5 min video, while the same prompt in P10 with identical grid-splitting strategy performed poorly with very low precision and recall on a 30 min video. One key reason for this drop in performance is difference in the relative density of layouts in the videos between P8 and P10. For P8, the 5 min video contained 6 layouts, which could give 1 layout every 50 sec. In contrast, for P10, the 30-min video had 24 layouts, which could give 1 layout every 75 sec. This lower frequency of relevant events in the longer video makes it difficult for the model to correctly identify true positives among a much larger pool of mostly irrelevant content even with AKS sampling method.

Overall, detecting layouts in longer videos with multimodal model proposed in this paper remains feasible by dividing them into shorter segments. For example, processing a 5 min video with grid splitting took approximately 10 mins, and the 30 min video took roughly 1 hr. By splitting the longer videos into smaller chunks, the model's performance can be preserved, while remaining within the model's input capacity limits. However, further testing on a wider range of video lengths, diverse datasets, and additional prompts is necessary to fully assess the mutlimodal model's capacity. Due to time constraints, our current evaluation focused on just two video durations, while we prioritized prompt experimentation.

5.2. YOLO Experiments

The results for the first trial of frisbee YouTube shorts trained on YOLOv8 overall showed poor performance in detecting players and frisbee discs using the standard parameters of a confidence level of 0.3. The detection of players was improved when the confidence level was reduced to 0.1, but the model struggled detecting the disc. In the efforts of improving the detection of discs, we experimented with zoomed-in shots where the player was successfully detected as a "person" and partially identified the disc as a "sports ball" as shown in Figure 4a. In the following frame of the same clip, the model failed to detect the disc as it was

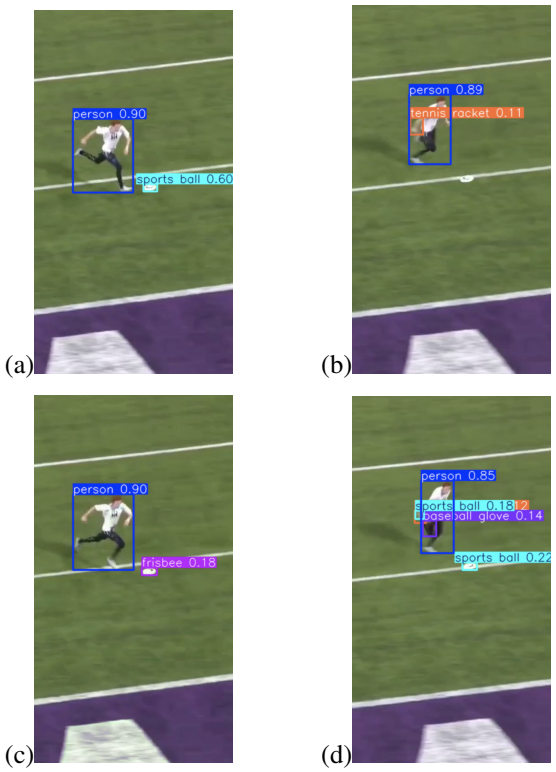

Figure 4: YOLO detection results on vertical videos for **(a)** zoomed-in shot successfully detecting players and ball, **(b)** following frame in zoomed-in shot failing to detect frisbee disc, **(c)** high-contrast shot successfully detecting players and frisbee disc, and **(d)** following frame in the high-contrast shot failing to detect frisbee disc

blended in with the white line as shown in Figure 4b.

We also experimented with increasing the contrast in the zoomed-in shots where the model successfully identified the disc as a "frisbee" but at a low confidence level in Figure 4c. In the following frame of the same clip, the model again failed to detect the disc even with increased contrast as it blended with the white line as shown in Figure 4d. The results also showed several misclassification errors such as detecting the player's hand as a "tennis racket" (Figure 4b) and the player's hand as a "baseball glove" (Figure 4d).

The results for the second trial with higher resolution, horizontal videos of frisbee games showed improved performance compared to the first trial with lower quality, vertical videos. The model successfully detected frisbee discs with high confidence for close-up shots in horizontal videos as shown in Figure 5a and b. Increasing the confidence level from 0.36 to 0.77, the model was able to reduce the misclassification errors as shown in Figure 5. However, the model still failed to detect frisbee discs in the second trial for zoomed-out frames. Our experiments revealed core challenges in using YOLO for reliable highlight detection, es-

(a)

(b)

Figure 5: YOLO results on higher resolution, horizontal videos at confidence level **(a)** 0.36 and **(b)** 0.77.

pecially in identifying frisbee discs during layout moments. While the model could identify frisbees in some zoomed-in frames, the detection confidence remained inconsistently low in low quality, vertical videos due to the small object size, motion blur during fast throws, occlusions, and poor contrast against background. Even though horizontal videos improved the performance for zoomed-in shots, the model failed to detect the frisbee disc on wider zoomed-out shots.

We could potentially further improve the performance by raising the confidence levels to filter noise and false positives and process only high confidence detections of frisbee discs. However, this approach could severely compromise the recall value, which is an important qualitative metric for identifying all key moments for layout detection. Due to the trade-off between confidence level and recall as well as YOLO's inability to handle small objects and occlusions, we decided that YOLO suits real-time applications rather than precision tasks like frisbee highlight detection. This limitation is also data-dependent, as higher resolution inputs and zoomed-in spatial features might mitigate the effects, but not eliminate the core issue in detecting small objects in dynamic environments.

Our experiments reveal clear differences between the YOLOv8 object detection model and the multimodal model with AKS in layout highlight detection for frisbee games. The multimodal model significantly outperformed the YOLOv8 model, especially for prompt P8, which achieved a perfect recall (1.000). The YOLO model proved to be limited for layout detection in frisbee games because YOLO inherently fails to understand temporal sequences or object interactions in dynamic actions like layouts. Therefore, YOLO model only focuses on object detection to capture what is present in a frame and not what happened. In addition, YOLO's reliance on object visibility means it could often miss highlights when objects are occluded, blurred, or momentarily hidden.

On the other hand, the multimodal model with AKS is designed to handle the challenges of traditional object detection models. This is done by considering both human pose estimation and temporal attention to detect motion patterns over time. Rather than depending on labels, the multimodal approach focuses on coordinated human motion patterns such as diving or catching. This means that it analyzes the keyframes centered more around action recognition than object detection, which could detect the essence of an event across time even if the object is not always visible. Overall, the results show that the multimodal model with AKS is a more reliable and scalable solution for identifying key highlight moments like layouts in frisbee games.

6. Conclusions

This project presents a proof of concept demonstrating the effectiveness of multimodal MLLMs with AKS in detecting highlight-worthy actions, such as layouts in frisbee games. The multimodal model with AKS significantly outperformed traditional detection methods such as YOLOv8 models in detecting layout actions in ultimate frisbee, which struggled with detecting small, fast moving objects such as frisbee dics. The multimodal model excelled in idenitfying diving actions to detect layouts, as shown by prompt "Is someone diving towards a frisbee?", which achieved a perfect recall of 1.000 and the highest F1-score of 0.583 across all prompts. Grid splitting played a crucial role in improving the detection performance of the model by helping the model better localize and focus on action-specific regions within each grid in a frame. Our results show that the multimodal model not only improves the reliability of detection precision, but also narrows down footage editors need to review to just 26.6% of the full video, which significantly reduces editing time.

However, the study has several limitations. It was conducted on a small dataset for a single game due to time constraints, which restricts the generalizability of the results. The model also showed sensitivity to linguistic variation in prompts. In addition, the multimodal model's performance may have been limited by the narrow range of clip lengths and action types included in the dataset. Future work could focus on expanding the dataset to include more games, teams, filming styles, and camera angles to better test the model's generalizability. Varying clip lengths could be tested including 3s, 5s, and 10s to investigate the effect of temporal resolution needed for different types of highlights. In addition, increasing prompt diversity integrated with a feedback loop could improve the prompt's clarity for a more accurate and reliable highlight detection.

7. Contributions and Acknowledgments

Farah - conceptualization, research, data analysis, writing, and reviewing. **Heather** - conceptualization, data collection, preprocessing, data analysis, and reviewing. **Megan** - conceptualization, research, code implementation, and reviewing.

We would like to thank Yunfan Jiang, our project mentor, for his constructive feedback and assistance. We would also like to thank Chaitanya Patel, Head CA, for guiding us in formulating the project idea.

References

[1] Fang Shu and Haoxing Yang. *Automatic Soccer Game Highlight Detection*. en. CS231n Course Project. Stanford University, 2024.

[2] Jerry Qu. *Using Pose Estimation to Analyze Rock Climbing Technique*. en. CS231n Course Project. Stanford University, 2024.

[3] Banoth Thulasya Naik, Mohammad Farukh Hashmi, and Neeraj Dhanraj Bokde. "A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions". en. In: *Applied Sciences* 12.9 (Jan. 2022). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 4429. ISSN: 2076-3417. DOI: 10.3390/app12094429. URL: https://www.mdpi.com/2076-3417/12/9/4429 (visited on 05/14/2025).

[4] Francesco Della Santa and Morgana Lalli. *Automated Detection of Sport Highlights from Audio and Video Sources*. arXiv:2501.16100 [cs] version: 2. Jan. 2025. DOI: 10.48550/arXiv.2501.16100. URL: http://arxiv.org/abs/2501.16100 (visited on 05/14/2025).

[5] Donghoon Han et al. *Unleash the Potential of CLIP for Video Highlight Detection*. arXiv:2404.01745 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2404.01745. URL: http://arxiv.org/abs/2404.01745 (visited on 05/14/2025).

[6] Xi Tang et al. *Adaptive Keyframe Sampling for Long Video Understanding*. arXiv:2502.21271 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2502.21271. URL: http://arxiv.org/abs/2502.21271 (visited on 05/14/2025).

[7] Tim Tang. *AKS code repository*. https://github.com/ncTimTang/AKS. Accessed: 2025-05-10. 2025.

[8] Juan Terven and Diana Cordova-Esparza. *A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS*. Feb. 2024. URL: https://arxiv.org/abs/2304.00501 (visited on 05/20/2025).

[9] Ultiworld - Ultimate Frisbee. *Hybrid vs Sprocket — Mixed Final — 2024 USA Ultimate National Championships*. https://www.youtube.com/watch?v=VssSNlS37LU. YouTube video; Accessed: 2025-05-20. 2024.

[10] espn. *Some of the greatest ultimate frisbee plays*. https://www.youtube.com/shorts/umdGjsdsQQ8. YouTube video; Accessed: 2025-05-10. 2023.

[11] UFA Ultimate Frisbee Association. *Ultimate Frisbee Top 10 Plays — 2023 UFA season*. https://www.youtube.com/watch?v=cso9580Oo1k&t=11s. YouTube video; Accessed: 2025-05-10. 2023.