Video Summarization for Soccer Match

Review Report - 6

March 1 st 2025 - April 21st 2025

Uchit N M - PES1UG22CS661

Arya Gupta - PES1UG22CS111

Faculty Mentors:

Prof. Prasad H B prasadhb@pes.edu

Dr. Gowri Srinivasa gsrinivasa@pes.edu

Prof. Preet Kanwal preetkanwal@pes.edu

Background :

This report presents the progress made from March 1st 2025 to April 21st 2025, building upon the developments documented in the *PESU_Ittiam_Report_5_Feb_2025.*

As per the discussion in the previous meet held on 3rd March 2025, the PESU team focused on leveraging audio to build a system that can summarize sports match commentary and identify key events along with their timestamps. While traditional keyword-based methods provided a foundational approach, they were limited in their ability to fully capture the contextual complexity of the commentary. So we shifted focus to transformer based architectures as they are suitable for these tasks, exploring the advanced capabilities of Large Language Models (LLMs), which offered enhanced context understanding and facilitated more accurate structured output generation.

Our evaluation included both open-source and API-based LLMs—such as Deepseek, Claude2, LLama, Gemma, Gemini, and ChatGPT—where we assessed their ability to handle context, extract insights, and integrate seamlessly into a scalable pipeline. We developed a Proof of Concept (PoC) using Gemini-2.0-flash, the Gemini model is selected for its large context window (100k tokens), free api availability, and ease of integration. Through careful prompt engineering and iterative testing, we refined our approach, successfully delivering structured summaries and accurately tagging events with timestamps. This effort laid the foundation for an effective summarization system, achieving an impressive 92.85% score using TF-IDF Cosine Similarity, reflecting excellent event coverage and semantic alignment for the final video highlight created using this method.

Additionally looking at the video counterpart, efforts were put into evaluating the performance of object detection models trained to identify key elements like players, goalkeepers, referees, and the ball in football match footage. To incorporate the suggestions given by the ITTIAM team where we were asked to explore the limitations of Yolov8 (3rd March 2025), a new dataset was curated with clips from various leagues and academic matches, including the Indian Super

League, Saudi League, and Major League Soccer, covering both professional and academic-level matches.

To ensure a reliable evaluation, we implemented a Temporal Majority Voting approach, which assesses object detections across sequences of frames rather than individually. This method effectively addresses challenges such as motion blur and occlusion, offering a more accurate evaluation of the model's real-world performance.

Recap: A list of reports submitted by Team PESU to Team Ittiam

1. PESU_Ittiam_Report_01 (Initial Work)
2. PESU_Ittiam_Report_02 (Architecture based work)
3. PESU_Ittiam_Report_03 (Proposed Framework)
4. PESU_Ittiam_Report_04 (Analyzing soccer match data and developing summarization methods.)
5. PESU_Ittiam_Report_05 (Video and Commentary analysis)

Table of Contents

Table of Figures

Section - 1: Leveraging Audio for Summary Generation

Post the previous meet, we have focused on developing an efficient system for summarizing sports match commentary and extracting key events with their corresponding timestamps.

After we had recognized the limitations of traditional keyword-based retrieval approaches, we had a comprehensive exploration of various Large Language Models (LLMs). This section details our findings about LLM —from open-source models with varying parameter sizes to API-based solutions—analyzing their performance in understanding commentary context, identifying key events, and providing structured outputs.

Through systematic prompt engineering and comparative analysis, we sought to identify the most suitable LLM approach that has balanced accuracy, context handling, and integration feasibility for our specific use case, ultimately leading to a viable proof-of-concept implementation.

The Explored LLMs:

- **Deepseek R1 (Distilled Models) - 8B parameters**
  Due to limited context size the entire transcript had to be chunked into parts with a overlap of 500 tokens, but the output was not similar to as that was achieved on the website with the entire model
- **Claude2 Alpaca 13B parameters**
- **LLama**
  The output was generalized to the match even when the queries were specific to events
- **Gemma 2B**

This activity was aimed at understanding what could be our go-to LLM once we have met the compute and storage requirements.

Keeping in mind the end goal of the project, a proof of concept would help us visualize what we are aiming at. To do this we explored a few API based LLMs to parse the transcript and help us

pinpoint the timestamp of events.
The LLMs explored were

- Claude
- Gemini (gemini 2.0-flash)
- Deepseek
- ChatGPT

The Proof of Concept (PoC)

A thorough experimentation with prompt design and engineering has been done in order to achieve an optimal structured output.

Among the mentioned, Gemini-2.0-flash is used due the benefits of getting a **context window of 1,00,000 tokens** and **free source** and its **ease of integration** in the pipeline.

Approach 1 : Goal Event Extraction

This experimental approach focuses solely on **goal events**. The timestamps for each goal were extracted from the **match transcript** and later verified by matching them with respect to the **original video**.

Prompt & Results

Q: Provide the timestamps for the goal events in the match.

A: Here are the timestamps for the goal events in the **Belgium vs. Japan** match, according to the provided transcript:

- **[0:52:48]** – Haraguchi scores for Japan (Japan 1 - 0 Belgium)
- **[0:57:05]** – Inui scores for Japan (Japan 2 - 0 Belgium)
- **[1:14:27]** – Vertonghen scores for Belgium (Japan 2 - 1 Belgium)
- **[1:19:00]** – Fellaini scores for Belgium (Japan 2 - 2 Belgium)
- **[1:39:16]** – Chadli scores for Belgium (Belgium 3 - 2 Japan)

Approach 2 : Duration-Based Goal Events

In this approach, the focus shifted from **specific goal timestamps** to **time durations**. Instead of just marking the goal moment, the extracted timestamps covered the **build-up play before the goal** and the **celebration phase after the goal**.

Prompt & Results

Q: Retrieve the start and end timestamps for all goal events in the match, ensuring that the extracted segments include context both before and after the goal. Specifically, include the build-up play leading to the goal and the post-goal celebration phase. The output should provide structured start and end timestamps for each goal event.

A: Here is the structured output of goal events, including their start and end timestamps:

- **Goal 1 (Japan - Haraguchi)**
  - **Start:** 0:52:36
  - **End:** 0:54:00
- **Goal 2 (Japan - Inui)**
  - **Start:** 0:56:58
  - **End:** 0:58:23
- **Goal 3 (Belgium - Vertonghen)**
  - **Start:** 1:13:14
  - **End:** 1:15:10
- **Goal 4 (Belgium - Fellaini)**
  - **Start:** 1:18:19
  - **End:** 1:19:55

- **Goal 5 (Belgium - Chadli)**
    - **Start:** 1:38:41
    - **End:** 1:40:26

Approach 3: Context-Rich Goal Events

In this approach, the focus was on capturing the **full context** around each goal event. The extracted timestamps included:

- **The build-up play before the goal**
- **The goal itself**
- **The celebration and the reaction of the opposing team**

Timestamps were recorded in **duration format (start - end)** to provide a structured and comprehensive view of the event.

Prompt & Results

Prompt 1

Q: Retrieve the start and end timestamps for all goal events in the match. Ensure that the extracted segments include context both before and after the goal. Specifically, include the build-up play leading to the goal and the post-goal celebration phase. The output should provide structured start and end timestamps for each goal event. Please give only timestamps, no context or descriptions. Give in the format: **[start timestamp] - [end timestamp]**.

A:

- **[0:52:46] - [0:53:04]**
- **[0:57:02] - [0:57:25]**
- **[1:14:24] - [1:14:31]**
- **[1:19:00] - [1:19:12]**

- **[1:39:08] - [1:39:27]**

Prompt 2

Q: Can you get me a start and end timestamp as I am trying to summarize the entire video based on goal events? I want to include all the before and after context of the goal event so that the **inclination of the goal to celebration is considered**.

A: Okay, here are the timestamps for the goal events, including some context before and after each goal, to capture the inclination and celebration:

- **Japan's First Goal (Haraguchi)**
  - **Start:** [00:52:30]
  - **End:** [00:53:00]
- **Japan's Second Goal (Inui)**
  - **Start:** [00:57:02]
  - **End:** [00:57:25]
- **Belgium's First Goal (Vertonghen)**
  - **Start:** [01:14:13]
  - **End:** [01:15:01]
- **Belgium's Second Goal (Fellaini)**
  - **Start:** [01:19:00]
  - **End:** [01:19:55]
- **Belgium's Third Goal (Chadli)**
  - **Start:** [01:39:02]
  - **End:** [01:40:37]

Prompt 3

Q: Can you get me the start and end timestamp as I am trying to summarize the entire video based on goal events? I want to include all the before and after context of the goal event so that the **initiation of the goal to celebration is considered**.

A: Okay, here are the start and end timestamps for each goal event, including some context before and after, based on the provided transcript:

- **Japan Goal 1 (Haraguchi)**
  - **Start:** 0:52:42
  - **End:** 0:54:00
- **Japan Goal 2 (Inui)**
  - **Start:** 0:56:47
  - **End:** 0:58:20
- **Belgium Goal 1 (Vertonghen)**
  - **Start:** 1:13:14
  - **End:** 1:15:10
- **Belgium Goal 2 (Fellaini)**
  - **Start:** 1:18:25
  - **End:** 1:19:43
- **Belgium Goal 3 (Chadli)**
  - **Start:** 1:38:38
  - **End:** 1:40:41

Approach 4 : Context-Optimized Goal Events

In the fourth approach, after reviewing and analyzing the outputs from the third approach, the duration of the background context was examined and marked. It was observed that event clips typically began airing **5 to 10 seconds** before the actual event occurred. Based on this pattern, a **constant of 8 seconds** was established to optimize contextual coverage, and prompts were fine-tuned accordingly.

In this approach, the model first analyzed the commentary to determine the **relevant contextual starting point**.

$$max(base\_constant(8\ sec), contextual\_start\ (extracted\ by\ the\ model))$$

This ensured that each clip captured the **key moments leading up to the event** while maintaining the same endpoint as before, covering the **goal celebration and the opposing team's reaction**.

Prompt & Results

Prompt 1

Q:
Retrieve the start and end timestamps for all goal events in the match. For each goal:
Start timestamp should be either 8 seconds before the actual goal or from where meaningful build-up play begins, whichever provides more context
End timestamp should extend until all goal-related context concludes, including celebrations, arguments, VAR reviews, or any other relevant aftermath
Format each entry as: [start timestamp] - [end timestamp] - [scoring team name]
Please provide only the formatted timestamps without additional descriptions or commentary.

A:

- [0:04:39] - [0:06:13] - Portugal
- [0:25:17] - [0:27:04] - Spain
- [0:46:04] - [0:47:23] - Portugal
- [0:57:51] - [0:59:30] - Spain
- [1:01:34] - [1:02:44] - Spain
- [1:31:18] - [1:32:45] - Portugal

The extracted timestamps were then processed using **regex** and structured into a **hash table**, which was later exported as a JSON file (**timestamps.json**).

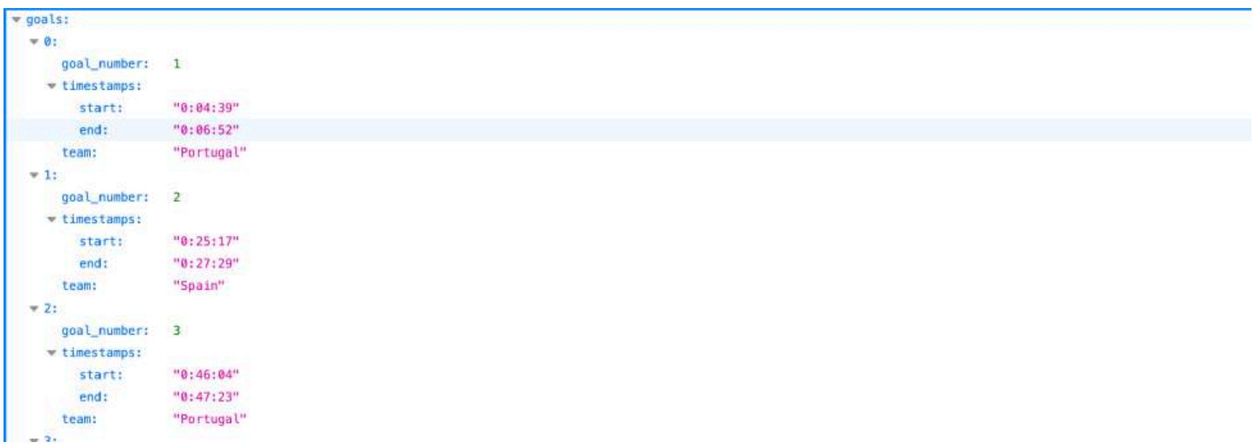


Fig 1.1 Prompt results (timestamps.json) approach 4

Prompt 2

Q: Retrieve the start and end timestamps for all goal events in the match. Start timestamp should be either 8 seconds before the actual event or from where meaningful build-up play begins, whichever provides more context End timestamp should extend until all goal-related context concludes, including celebrations, arguments, VAR reviews, or any other relevant aftermath. Make sure proper context is maintained and the entire goal event is covered. The output should provide structured start and end timestamps for each goal event. Please give only timestamps only, no context or descriptions. Give in the format: [start timestamp] - [end timestamp] - [team name] - [type] - [small descriptions]

A: *(Exported as **timestamps.json**)*

This structured output allowed for **efficient processing** in later stages of the analysis pipeline.

Fig 1.2 Prompt results approach 4 (2)

Approach 5 : Enhanced Event Extraction & Highlights Video

Building on **approach 4**, this method **refined the extraction structure** to improve the **final video output**. Additionally, the event list was **expanded** to include:

- ➔ **Goals**
- ➔ **Fouls**
- ➔ **Replacements**
- ➔ **Missed Goals**

A **transition frame** was introduced in the final video, placed between event clips to:

- Inform viewers about the **upcoming event**
- Highlight **key players** or **major actions**

This approach resulted in a **proof of concept (POC)**, culminating in a **highlight video lasting 20.44 minutes**.

Prompt & Results

Prompt

Q:

Retrieve the start and end timestamps for all goal, foul, replacements and missed goal events in the match.Start timestamp should be either 8 seconds before the actual event or from where meaningful build-up play begins, whichever provides more context End timestamp should extend until all goal-related context concludes, including celebrations, arguments, VAR reviews, or any other relevant aftermath. Make sure proper context is maintained and the entire goal event is covered. The output should provide structured start and end timestamps for each goal event. Please give only timestamps only, no context or descriptions. Give in the format: [start timestamp] - [end timestamp] - [team name] - [type] - [small descriptions]

A)

Extracted Timestamps

Okay, I understand. Here's the output with just the timestamps, team name, event type, and a brief description, formatted as requested:

[0:04:39] - [0:06:13] - Portugal - goal - Penalty Goal by Ronaldo
[0:25:17] - [0:27:04] - Spain - goal - Costa scores after foul claim
[0:37:51] - [0:38:00] - Spain - missed goal - Iniesta hits the bar
[0:46:04] - [0:47:23] - Portugal - goal - Error leads to goal by Ronaldo
[0:57:51] - [0:59:30] - Spain - goal - Costa scores from free kick
[1:00:01] - [1:00:08] - Portugal - foul - Guesh goes down
[1:00:54] - [1:01:04] - Spain - foul - Iniesta's touch cut out
[1:01:34] - [1:02:44] - Spain - goal - Nacho scores
[1:02:58] - [1:03:11] - Spain - foul - challenge on Ronaldo
[1:06:51] - [1:06:53] - Spain - foul - perhaps a little easily
[1:11:11] - [1:11:25] - Portugal - replacements - Bruno Fernandez off, xia Mario on
[1:12:50] - [1:13:46] - Portugal - replacements - Bernardo Silva off, Ricardo kesma on
[1:13:16] - [1:13:37] - Spain - replacements - Iniesta off, Thiago comes on
[1:14:18] - [1:14:34] - Spain - missed goal - Glorious Chance for goal
[1:19:00] - [1:19:03] - Portugal - foul - in the end on kesma
[1:20:07] - [1:20:31] - Spain - replacements - Diego Costa off, Lucas Vasquez on
[1:22:06] - [1:22:14] - Portugal - replacements - Andre Silva on
[1:22:14] - [1:22:30] - Portugal - replacements - W Das off, Andre Silva on
[1:22:30] - [1:22:39] - Portugal - foul - Hand ball from Ricardo kesma
[1:23:56] - [1:23:59] - Spain - missed goal - time Andre Silva
[1:29:32] - [1:30:10] - Spain - replacements - HIAGO aspes on, vasquez comes on
[1:30:07] - [1:30:10] - Portugal - foul - Ronaldo goes down
[1:31:18] - [1:32:45] - Portugal - goal - Ronaldo hattrick
[1:33:14] - [1:33:17] - Spain - missed goal - a simply extraordinary night

JSON Raw Data Headers

Save Copy Collapse All Expand All Filter JSON


```
0:
    start_timestamp:   "0:04:39"
    end_timestamp:     "0:06:30"
    team_name:         "Portugal"
    type:              "Goal"
    description:       "Ronaldo scores penalty after Nacho foul. Includes penalty build-up, goal, and initial celebrations."
1:
    start_timestamp:   "0:19:00"
    end_timestamp:     "0:19:51"
    team_name:         "Spain"
    type:              "Foul"
    description:       "Sergio Busquets receives a yellow card for a foul on Guedes. Includes foul and reaction."
2:
    start_timestamp:   "0:25:10"
    end_timestamp:     "0:27:00"
    team_name:         "Spain"
    type:              "Goal"
    description:       "Diego Costa scores after a challenge on Pepe, VAR check. Covers the challenge, goal, protest and VAR confirmation."
3:
    start_timestamp:   "0:29:14"
    end_timestamp:     "0:29:54"
    team_name:         "Portugal"
    type:              "Foul"
    description:       "Bruno Fernandes gets a yellow card. Includes foul, assistant raise flag and showing yellow card."
4:
    start_timestamp:   "0:45:50"
    end_timestamp:     "0:47:23"
    team_name:         "Portugal"
    type:              "Goal"
    description:       "Ronaldo scores his second goal due to a De Gea error. Includes Guedes shot, De Gea fumble, goal, and initial reactions."
5:
    start_timestamp:   "0:58:01"
    end_timestamp:     "0:59:53"
    team_name:         "Spain"
    type:              "Goal"
```


Fig 1.3 Prompt results approach 5

The above represent the outcomes and results when worked with Gemini 2.0 Flash.

To compare if the other LLMs which we had limited access, we worked with approach 5:

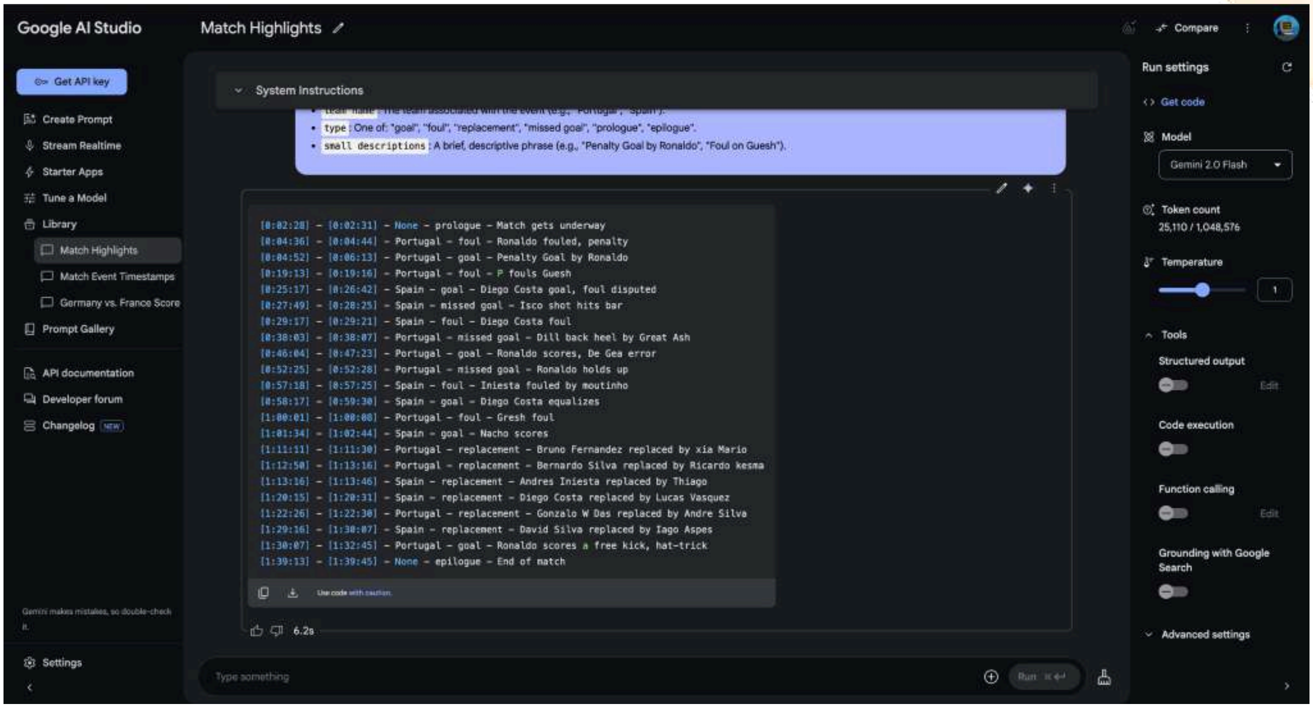


Fig 1.4 . Gemini 2.0 flash

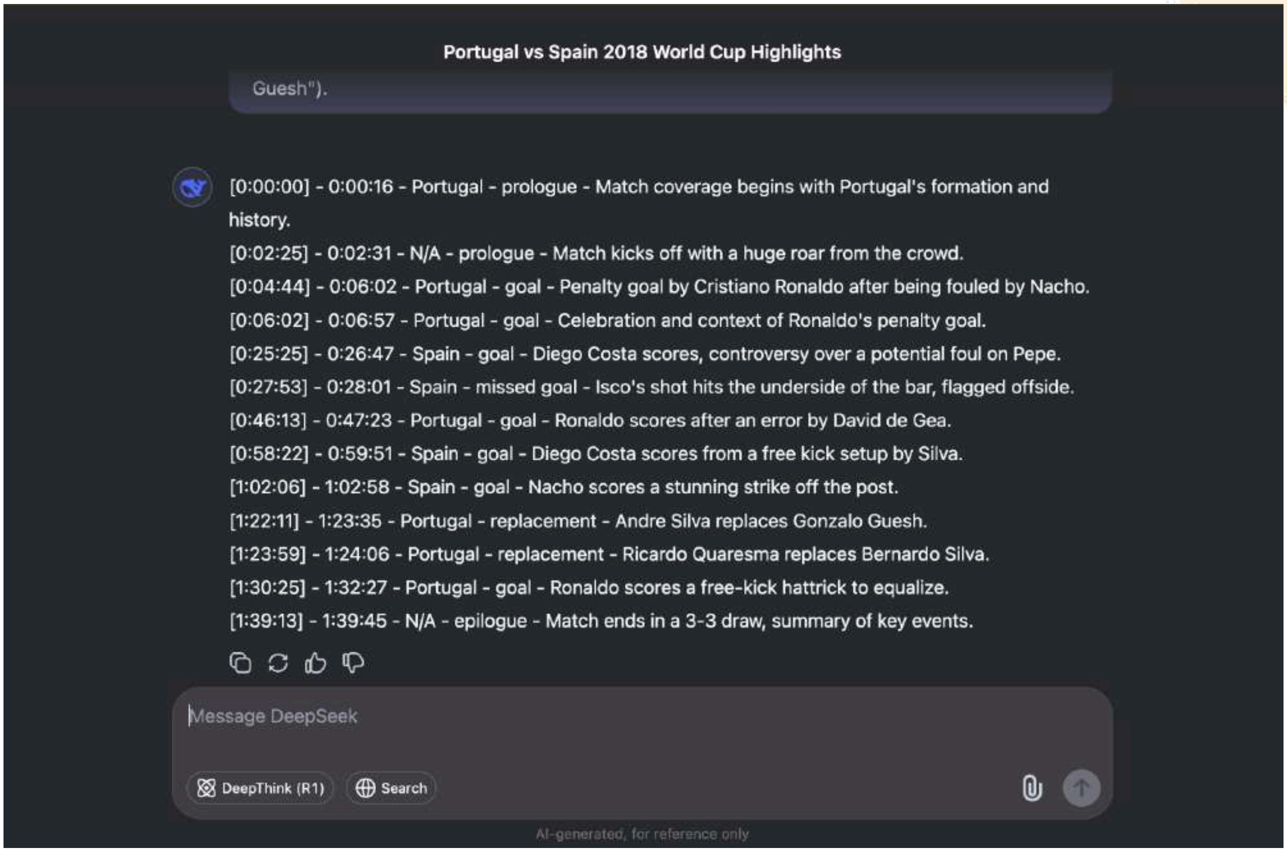


Fig 1.5. Deepseek

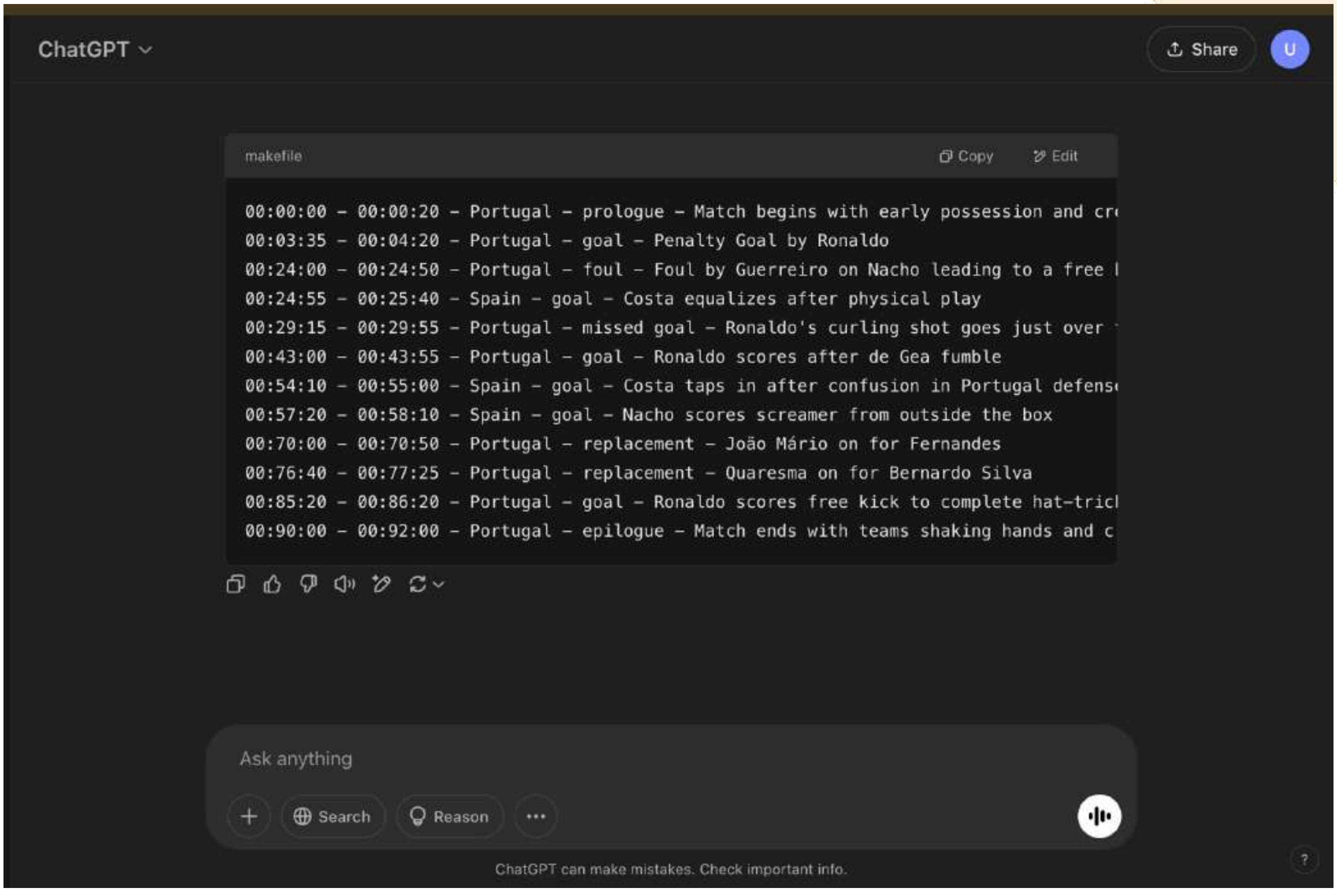


Fig 1.6. ChatGPT

The figures above display the marked timestamps of various events across different platforms. After thorough verification and comparison with the main video, these timestamps have been confirmed to be contextually and semantically accurate, corresponding precisely to the respective events. This demonstrates the effectiveness of the prompt (approach 5) and ultimately leads to a positive outcome and the PoC.

Final Processing

All trimmed clips were **merged using FFmpeg**.

- Since the **trims** were sourced **directly from the original video**, they **remained perfectly synchronized** with the **audio**.
- The **transition frames** provided **smooth storytelling** throughout the highlights video.

This structured approach **enhanced the final output**, ensuring a **seamless viewing experience**.

Verification of the created Highlights:

Our video summarization pipeline was evaluated using multiple similarity metrics to ensure a comprehensive assessment of its performance against ground truth summaries. The validation results demonstrate promising alignment between our generated summaries and the reference standards. Ref. Table 1.1

Metric	Score	Inference
TF-IDF Cosine Similarity	0.9285	Indicates excellent semantic alignment; the model captures key concepts effectively.
SBERT Similarity Score	0.7494	Shows strong semantic similarity using Sentence-BERT for contextual meaning.
ROUGE-1	0.6221	Measures unigram (single word) lexical overlap.
ROUGE-2	0.4738	Measures bigram (two-word sequence) lexical overlap.
ROUGE-L	0.4908	Measures longest common subsequence for overall structural similarity.
BLEU Score	0.1979	Stricter metric focusing on n-gram precision; lower scores are typical in summarization tasks.

Table 1.1. Evaluating Semantic and Lexical Alignment of Generated Summaries Using NLP Metrics

Here,
The **Sentence-BERT** framework, which captures contextual meaning beyond simple word overlap. **ROUGE** scores validate the lexical overlap between generated and reference summaries. **BLEU** is known to be stricter for summarization tasks as it focuses on precision of n-gram matches.

The collective results indicate that our model produces summaries that effectively capture both the semantic meaning and key content elements of the ground truth summaries. The particularly strong TF-IDF Cosine similarity demonstrates that our approach successfully identifies and includes the most informative terms from the source videos.

These validation metrics confirm that our video summarization solution is performing well and producing reliable, content-relevant summaries that closely match human-created reference standards.

Visual Validation:

Fig 1.7. Event Coverage Mapping

For visual validation, both the broadcasted and generated highlights were thoroughly inspected. Events were listed, and their contextual coverage was observed. Subsequently, a mapping of these events was performed.

Fig 1.7 illustrates this mapping between the broadcasted highlights and the generated highlights. Under each column—left for the broadcast highlights and right for the generated highlights—the covered events are listed.

The common events are matched across both columns to assess how effectively the models captured them, along with a complete listing of all identified events.

Output Drive link -

Highlight using gemini ai This Drive folder contains the constructed highlight clips, the proof-of-concept (POC).

Broadcast Extended Highlight link -

Portugal 3-3 Spain | Extended Highlights | 2018 FIFA World Cup

Section - 2: Testing out the model over the newly curated dataset

The following section describes the work done to understand the limitations of models trained for player and ball detection as showcased in the previous meet held on 3rd March 2025.

This work focused on detecting and tracking essential elements, namely the **ball, players, goalkeeper, and referee**, using advanced object detection techniques. The models were implemented using **YOLO-based object detection frameworks** which worked with team classification models.

To ensure a comprehensive assessment, a newly curated dataset of match clips was compiled, encompassing a diverse range of gameplay scenarios. The evaluation specifically included lesser-known matches, along with academic-level games and professional tournaments, to validate the model's adaptability across different match conditions. This approach ensured that the model was not solely trained on standard international tournaments or teams with widely recognized jersey colors but was adaptable to a broader spectrum of football games.

Composition of new dataset:

Clips	**Source**
Indian Super League	(India)
Clip 1	https://www.youtube.com/watch?v=yCLE0MN9dLQ
Clip 2	https://www.youtube.com/watch?v=yCLE0MN9dLQ
Clip 3	https://www.youtube.com/watch?v=3Da5FzjJ3_Q
Clip 4	https://www.youtube.com/watch?v=3Da5FzjJ3_Q
Clip 5	https://www.youtube.com/watch?v=GRepKwfM730
Clip 6	https://www.youtube.com/watch?v=GRepKwfM730
Saudi League	(UAE)
Clip 7	https://www.youtube.com/watch?v=zcZELssSxk0
Clip 8	https://www.youtube.com/watch?v=zcZELssSxk0
Clip 9	https://www.youtube.com/watch?v=5oWrRCrhqV0
Clip 10	https://www.youtube.com/watch?v=5oWrRCrhqV0
Clip 11	https://www.youtube.com/watch?v=0E8tHlTw-64
Clip 12	https://www.youtube.com/watch?v=0E8tHlTw-64
Major League Soccer	(USA)
Clip 13	https://www.youtube.com/watch?v=VwKahcr97RA
Clip 14	https://www.youtube.com/watch?v=VwKahcr97RA
Clip 15	youtube.com/watch?v=ja3sFThVDTQ
Clip 16	youtube.com/watch?v=ja3sFThVDTQ

Table 2.1 Dataset

Indian Super League

Clip 1

Clip 2

Clip 3

Clip 4

Clip 5

Clip 6

Saudi Pro League

Clip 7

Clip 8

Clip 9

Clip 10

Clip 11

Clip 12

Major League Soccer

Clip 13

Clip 14

Clip 15

Clip 16

Fig 2.1 Sample Clips

Temporal Majority Voting Approach for Object Detection Evaluation

To effectively evaluate our trained YOLOv8 models, we developed a systematic evaluation process combining model outputs, manual annotation, and statistical analysis. Our methodology focuses on determining detection accuracy across the video rather than isolated frames, providing a more robust assessment of real-world performance.

Evaluation Process

Our evaluation consisted of three key phases:

1. **Model Output Generation**: We ran video clips through our trained YOLOv8 models to generate object detection results.
2. **Ground Truth Creation**: We manually annotated frames in Roboflow to establish accurate ground truth data, identifying the correct number of objects and their proper classifications.
3. **Performance Analysis**: A mathematical pipeline was developed for comparison and analyzing 12 objects across 720 frames to evaluate model performance.

Temporal Majority Voting Methodology:

Rather than evaluating each frame in isolation, we implemented a temporal majority voting system that considers object detections across a sequence of frames. This approach helps overcome temporary occlusions, motion blur, and other frame-specific challenges that might affect detection quality.

Let:

- F_r: Reference frame index (e.g., frame 150)
- T: Number of frames to look before and after F_r (hyperparameter)
  → In our case, $T = 30$
- $W = 2T + 1$: Total number of frames considered (window size)
  → In our case, $W = 61$
- O_r: Ground truth label of the object under consideration in frame F_r
  → e.g., "player"
- P_i: Model's predicted label for the object in frame i, where $i \in [F_r - T, F_r + T]$
- V: Set of all predicted labels across W frames
  → $V = \{P_i \mid i = F_r - T, ..., F_r + T\}$
- $\text{mode}(V)$: The majority label predicted across the frames (i.e., majority vote)

Let L be the final predicted label from the majority vote:

$$L = \text{mode}(\{P_i \mid i \in [F_r - T, F_r + T]\})$$

In case the set is not unimodal, ties are resolved arbitrarily. (Empirically, such cases have been rare.)

Then your final evaluation is:

$$\mathbf{y} = \begin{cases} \text{Correct}, & \text{if } L = O_r \\ \text{Incorrect}, & \text{if } L \neq O_r \end{cases}$$

a) Models Output

b) Ground Truth

Fig 2.2. Validation of Model Output Against Ground Truth

Analysis :

The following figures provide a frame-wise validation of object detection and classification performance for the newly curated clips. Each consists of a bar chart (left) showing the number of frames where a specific object was either correctly classified or misclassified, helping quantify the model's accuracy. Alongside, a reference frame from the video (right) visually highlights the object under evaluation using a cyan-colored bounding box along with an arrow pointing to the target object. These visualizations enable both quantitative (frame count) and qualitative (visual) assessment of the model's ability to correctly identify key entities, such as players and referees, in dynamic and crowded scenes.

Object 1 - Clip 5 (Player):

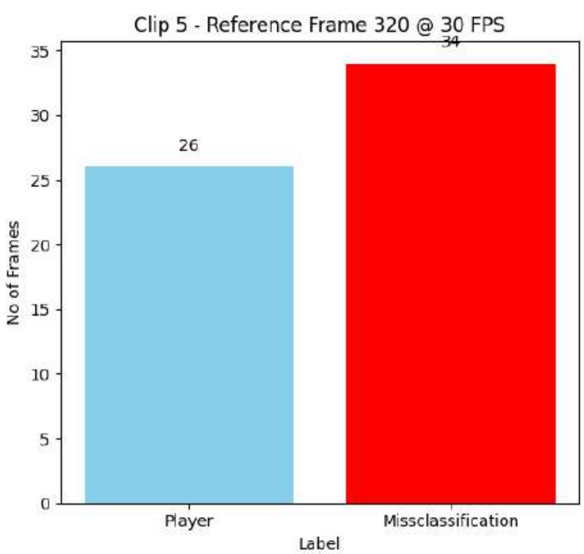


Fig 2.3. Validation Of frame 320 - clip 5

Object 2 - Clip 5 (Referee):

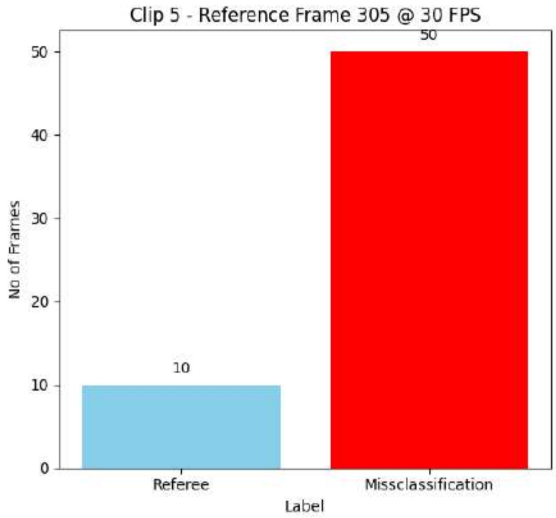


Fig 2.4. Validation Of frame 305 - clip 5

Object 3 - Clip 2 (Referee):

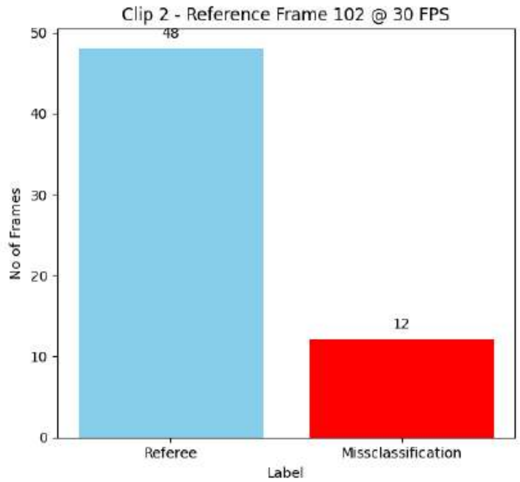


Fig 2.5. Validation Of frame 102 - clip 2

Object 4 - Clip 2 (Player):

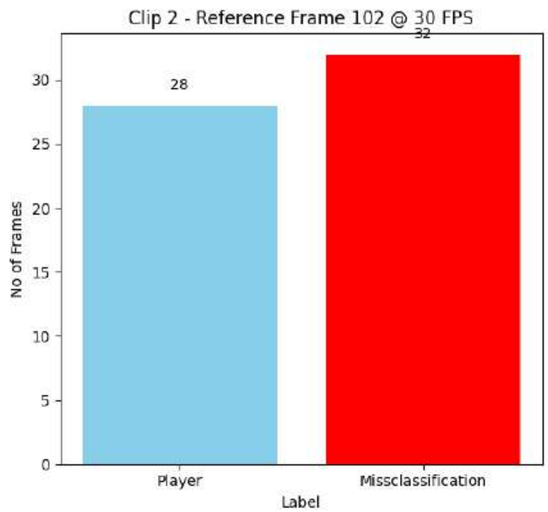


Fig 2.6. Validation Of frame 102 - clip 2

Object 5 - Clip 12 (Player):

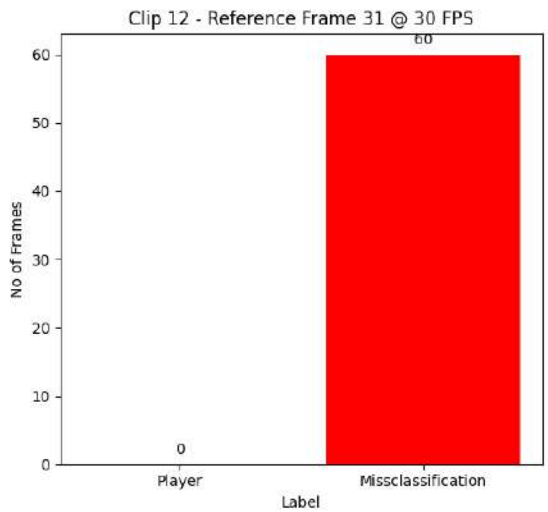
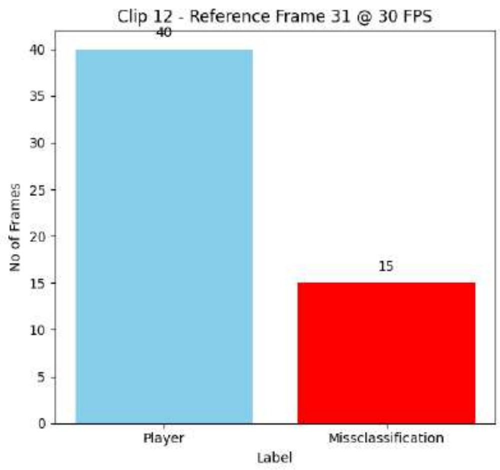


Fig 2.7. Validation Of frame 31 - clip 12

Object 6 - Clip 12 (Player):



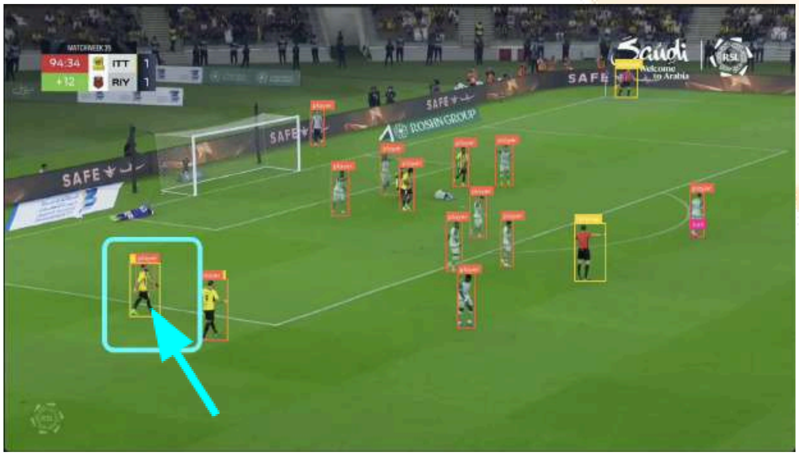
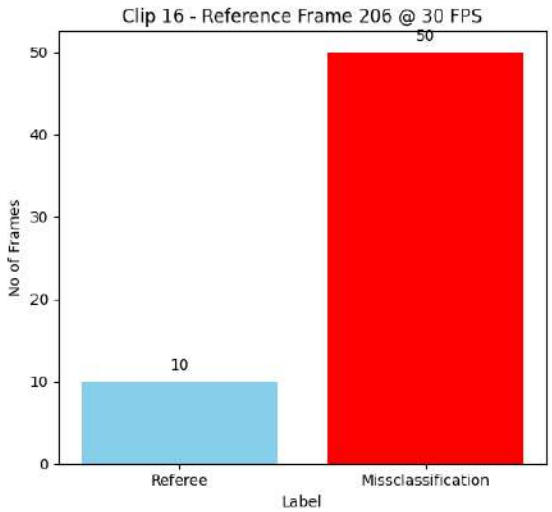


Fig 2.8. Validation Of frame 31 - clip 12

Object 7 - Clip 16 (Referee):



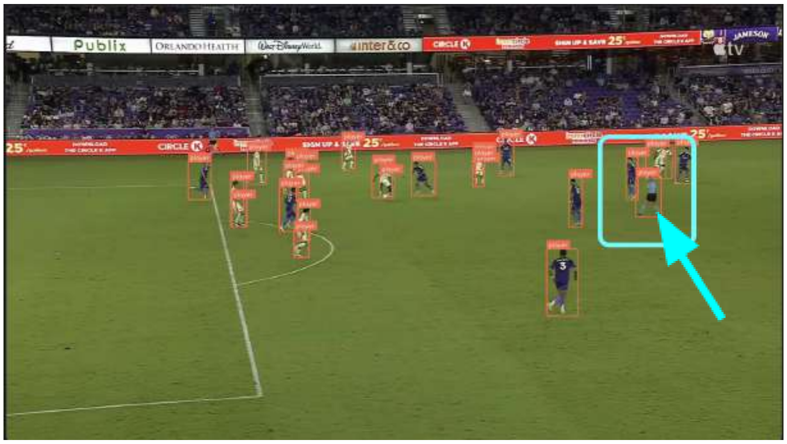
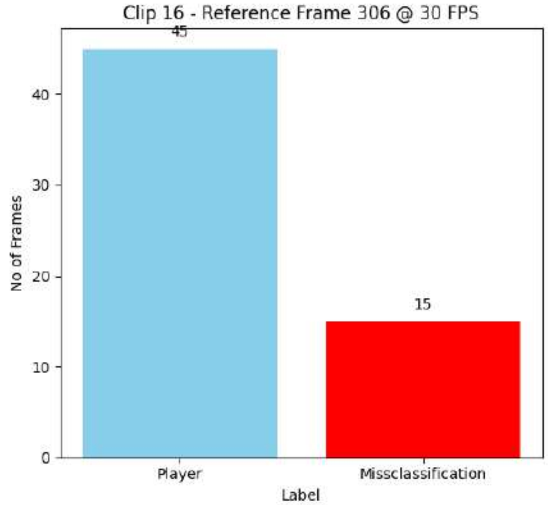


Fig 2.9. Validation Of frame 206 - clip 16

Object 8 - Clip 16 (Player):



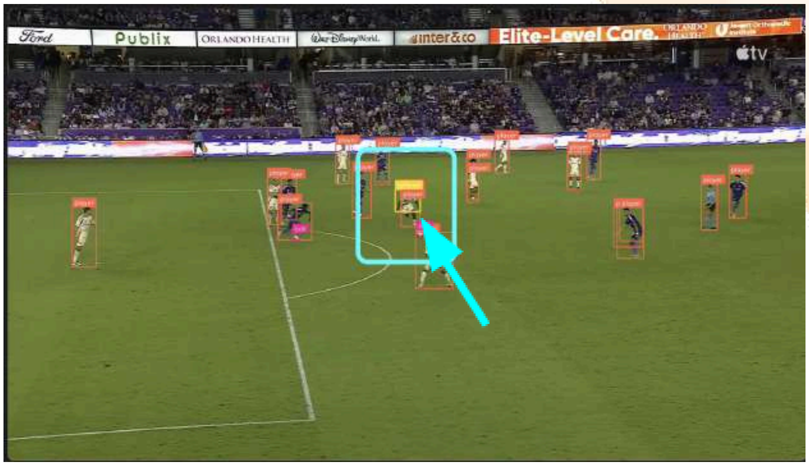


Fig 2.10. Validation Of frame 306 - clip 16

Object 9 - Clip 7 (Referee):

Fig 2.11. Validation Of frame 41 - clip 7

Object 10 - Clip 10 (Player):

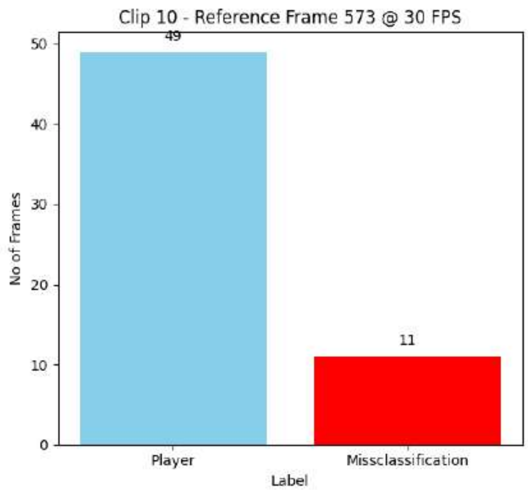


Fig 2.12. Validation Of frame 573 - clip 10

Object 11 - Clip 10 (Referee):

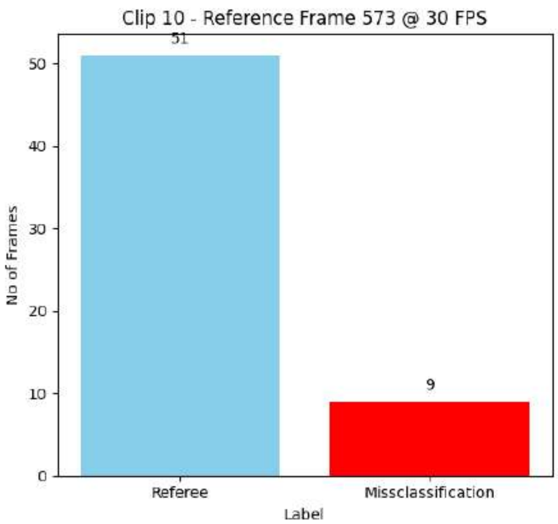


Fig 2.13. Validation Of frame 573 - clip 10

Object 12 - Clip 10 (Player):

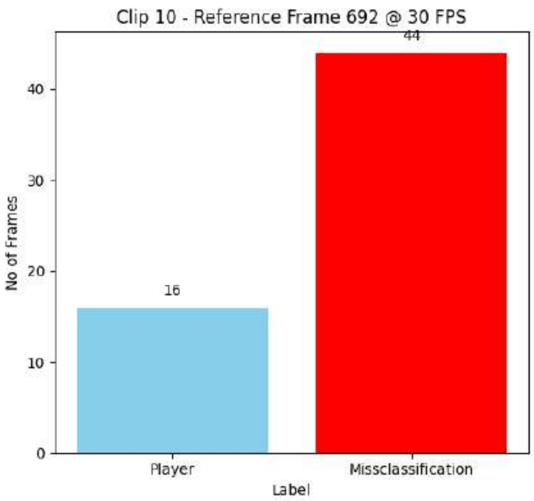


Fig 2.14. Validation Of frame 692 - clip 10

Inference from the Analysis:

The model exhibits reduced accuracy and performance under the following conditions:

- Dense clustering: When multiple players or objects are concentrated in a small area, the model struggles to differentiate between them.
- Object overlap: Overlapping elements lead to misclassification or missed detections.
- Non-bird's-eye view angles: Frames captured from perspectives other than a top-down (bird's-eye) view impair the model's spatial interpretation.
- Color similarity: The model often confuses players' shoes with the ball when they share similar colors.
- Airborne ball scenarios: In some frames, the model fails to accurately detect the ball when it is in mid-air.
- Player orientation: The model performs poorly when players appear horizontally rather than vertically in the frame.

The complete analysis for 12 objects across 720 frames : GTV.pdf

The heading indicates the **Clip Number** from the curated dataset. The line below specifies the **Ground Truth Class** for the clip. This is followed by a particular **Frame Number**, along with a label indicating whether the predicted class for that frame matches the ground truth.

- A **green label** indicates a correct match (prediction = ground truth).
- A **red label** indicates a mismatch (prediction ≠ ground truth).

Conclusion and Future Work:

As a continuation of the current work, we plan to further enhance the system by leveraging locally hosted Large Language Models (LLMs), using potential model architectures such as DeepSeek, LLaMA, or Gemma. By fine-tuning and configuring these models to extract event timelines effectively, we aim to eliminate reliance on API-based solutions such as GeminiAI, which is used for the initial proof of concept. This transition will provide greater control and enable the development of a domain-specific solution using open-source LLM frameworks, paving the way for an in-house model tailored to our needs while minimizing external dependencies.