Highlight Detection from Video, Audio, and Text Prompts

Mathis Doutre **Srikar Viswanatha** **Hang Kim** **Eugénie Laugier**

Georgia Institute of Technology
{mdoutre3, sviswanatha6, hkim722, elaugier3}@gatech.edu

Abstract

We study the problem of personalized highlight detection in soccer videos using multimodal information and natural-language queries. Our system combines visual, audio, and textual representations: a VideoMAE encoder pretrained on SoccerNet for spatio-temporal features, a Whisper-based audio and transcript pipeline, and a CLIP-based query encoder adapted to soccer-specific vocabulary through a lightweight prompt-level few-shot adapter. These modalities are fused through a transformer architecture to classify events and support query-conditioned retrieval. Self-supervised pretraining on SoccerNet significantly improves video feature structure, increasing the Silhouette Score from 0.42 to 0.58. For event classification, our multimodal Fusion Transformer outperforms both unimodal baselines, reaching a macro-precision of 0.719 versus 0.644 (video-only) and 0.515 (text-only). These results show that multimodal signals and domain-adapted query embeddings provide tangible benefits for fine-grained highlight detection in sports video understanding.

1 Introduction

The online broadcasting of sports has grown exponentially. These videos provide a valuable source of information for analysts, coaches, and media producers, who increasingly rely on rapid access to key moments for tactical study but can also be used by media for public entertainment. Manually extracting the highlights of videos is a slow, repetitive and time-consuming task. This has motivated extensive research on video summarization and video highlight extraction. Existing approaches mostly rely on video cues(1) and, more recently, on audio signals such as cheering (2). However, prior work ignores two aspects. First, highlight extraction is rarely personalized, and current methods cannot adapt to user queries such as retrieving specific types of actions, from a specific player or type of event. Secondly, despite evidence that multimodal signals carry complementary information, most systems rely on a limited combination of modalities and do not jointly exploit video, audio and transcript data.

To address these limitations, we propose a multimodal, query-conditioned highlight detection that integrates visual, audio and textual representations together with a learned query encoder. Our objective is to highlight the accuracy and robustness of highlight localization while enabling flexible, user-driven retrieval. This capability is beneficial for analysts and coaches, who require fine-grained access to specific patterns, and for content producers, who benefit from faster and more customizable highlight generation.

Our study will focus on soccer, a domain that remains challenging due to its long gameplay containing sparse and ambiguous events. We train and evaluate our model on **SoccerNet**(3) dataset which includes full match videos and labeled event timestamps. Our evaluation focuses on event classification accuracy and feature-space structure, using clustering metrics, macro-precision, and prompt-level query classification performance.

If successful, this approach fills an important gap at the intersection of video understanding and information retrieval. By leveraging multimodal context and conditioning on user queries, it moves beyond fixed-template highlight extraction toward personalized event retrieval, offering practical benefits for professional analysis and media applications.

2 Related Work

The rise of deep learning has shifted video summarization from rule-based heuristics to data-driven models that learn what makes a moment highlight-worthy. Early neural approaches mainly focused on the visual stream: for instance, 3D ConvNets have been used to capture spatio-temporal features in sports videos, improving action recognition and event detection accuracy (4). However, these vision-only models require large annotated datasets, are computationally expensive, and often miss context available in other modalities.

A key insight from later work is that audio is equally important as video for highlight detection. Badamdorj et al. (2021) (5) introduced an audiovisual attention network that learns the interactions between what is seen and heard in a clip. Their results showed that audio cues such as crowd cheering or commentary excitement can outperform visual features alone, and that fusing both streams provides the most robust detection. Similarly, Della Santa et al. (2025) (6) proposed a lightweight two-stream model where video frames (in grayscale) and audio spectrograms are processed separately, then combined. This simple design achieved high precision on both modalities and further improved when fused, confirming the complementary nature of sound and vision.

With modern automatic speech recognition (ASR) systems like Whisper (7), full-match transcripts can be reliably obtained. Building on this, Chakraborty et al. (2025) (8) proposed a text-only pipeline: Whisper transcripts are passed to multiple LLMs ("judges") that independently analyze outcomes, excitement levels, and tactical context. Remarkably, this zero-shot text-based method achieved performance levels comparable to vision-heavy models, while being easier to adapt across sports without task-specific training.

Overall, existing work shows that the best results emerge from multimodal approaches that leverage video, audio, and text together. However, current systems remain limited in two key ways:

- Query-based summarization has been explored, but typically ignores audio or treats speech transcripts separately, leaving modalities only loosely integrated.
- Automatic highlight extraction has been studied in sports, but these methods are not personalizable: they output fixed highlight reels, independent of what a user might actually want to see.

To our knowledge, there is no existing solution that allows a user to query a multimodal system (text + video + audio) to generate personalized highlights.

3 Method

3.1 Problem Formulation

In this work we design a multimodal retrieval and highlight detection pipeline built on top of the SoccerNet dataset (3). The end goal is to generate temporal highlight segments conditioned on user-provided natural language queries (e.g., "Show me all goals by player X" or "When was the referee involved?").

3.2 Hypothesis

H1: Multimodal fusion improves highlight localization.

Incorporating audio and transcript modalities alongside video features will yield higher temporal overlap (F1@0.5) with SoccerNet's ground-truth highlights compared to a video-only baseline. We expect audio cues (e.g., cheering intensity) and ASR-derived text to improve the alignment between predicted and labeled events.

H2: Transcript grounding enhances semantic consistency.
Including Whisper-generated transcripts in the fusion transformer will increase intra-query embedding similarity between retrieved highlights of the same event type (e.g., all "goal" segments), reflecting stronger contextual understanding of game semantics.

H3: Pretrained CLIP query embeddings improve generalization.
Leveraging CLIP's pretrained text representations for query encoding will enable robust retrieval on unseen or paraphrased user prompts, reducing performance degradation (<5% F1 drop) when evaluated on test queries not seen during adapter fine-tuning.

3.3 Model Architecture

The proposed architecture consists of six modular components that process different modalities and fuse them through a unified transformer-based framework:

1. **Query Encoder:** Encodes user queries into a 512-dimensional semantic embedding using a CLIP ViT-B/32 text encoder, fine-tuned via lightweight adapters to better represent soccer-specific action terms (e.g., "offside", "penalty kick", "corner").
2. **Video Encoder:** Extracts spatiotemporal representations from soccer broadcast videos using a pretrained VideoMAE model(9), which is robust to motion blur and occlusion common in broadcast footage.
3. **Audio Latent Encoder:** Converts raw audio waveforms into compact feature vectors using a Whisper encoder (7), leveraging crowd reactions and commentary tone as implicit indicators of event salience.
4. **Transcript Encoder:** Processes textual commentary derived from the ASR pipeline, capturing linguistic cues such as mentions of player names, outcomes, or referee actions.
5. **Fusion Transformer:** Combines all modalities (video, audio, transcript, and query embeddings) through a cross-attention mechanism with positional and modality-type embeddings, allowing contextual reasoning across heterogeneous signals.
6. **Prediction Head:** Outputs frame-level highlight scores or temporal span boundaries corresponding to predicted highlight intervals, which can be aggregated to form final highlight clips.

3.4 Training

3.4.1 Video Encoder

To improve the quality of video representations for the downstream task, we conduct self-supervised pretraining of the video encoder using the **VideoMAE** framework. Unlike supervised training that uses timestamp labels, this stage utilizes entire videos from the SoccerNet dataset in a fully self-supervised manner, following the masked autoencoder paradigm.

Given an input video sequence $V \in \mathbb{R}^{T\times H\times W\times 3}$, a random subset of visual patches is masked (typically 75%–90%), and the transformer is tasked with reconstructing the missing spatio-temporal tokens:

$$\hat{V} = f_\theta\big(\mathrm{Unmask}(V)\big), \quad \mathcal{L}_{\text{rec}} = \|\hat{V} - V\|^2.$$

The encoder learns to capture motion dynamics, spatial dependencies, and semantic structures without any manual annotation. This pretrained backbone is later used to extract embeddings for timestamps of interest (e.g., goals, fouls, and ball-out events) for fusion-based classification.

3.4.2 Audio Encoder

To capture both acoustic and linguistic information from soccer commentary, we implement a dual-stream audio encoder that combines complementary representations of audio signals. The architecture consists of two main components: a **latent audio encoder** that extracts acoustic features using a pretrained Wav2Vec2 model, and a **speech-text encoder** that transcribes audio using OpenAI's Whisper model and generates semantic embeddings via a sentence transformer.

Given an audio waveform $\mathbf{a} \in \mathbb{R}^{N_{\text{samples}}}$ centered around each event timestamp, the dual-stream processing begins with the acoustic stream, where

$$\mathbf{z}_{\text{audio}} = \text{Wav2Vec2}(\mathbf{a}) \in \mathbb{R}^{1024}.$$

Simultaneously, the linguistic stream processes the same input through transcription and embedding:

$$\text{transcript} = \text{Whisper}(\mathbf{a}),$$

followed by

$$\mathbf{z}_{\text{text}} = \text{SentenceTransformer}(\text{transcript}) \in \mathbb{R}^{384}.$$

These two embedding streams, $\mathbf{z}_{\text{audio}} \in \mathbb{R}^{1024}$ and $\mathbf{z}_{\text{text}} \in \mathbb{R}^{384}$, are then concatenated to form a unified audio representation:

$$\mathbf{z}_{\text{concat}} = [\mathbf{z}_{\text{audio}}; \mathbf{z}_{\text{text}}] \in \mathbb{R}^{1408}.$$

3.5 Query Encoder

We use the text tower of CLIP ViT-B/32 (10) as our query encoder. CLIP is pretrained on large-scale image-text pairs and is designed to align visual and textual inputs in a shared embedding space, which makes it a natural choice for retrieving video highlights from natural-language queries. Given a query string q, the CLIP text encoder produces a 512-dimensional embedding $z \in \mathbb{R}^{512}$, which we L2-normalize before passing it to the multimodal fusion transformer.

While zero-shot CLIP already provides reasonable representations for generic soccer-related queries, its text space is not explicitly calibrated to SoccerNet classes. To better align the encoder with this domain, we add a lightweight few-shot adapter on top of the frozen CLIP embeddings. The adapter is a single linear classifier $h : \mathbb{R}^{512} \to \mathbb{R}^{C}$ where C is the number of SoccerNet action classes. At training time, we first embed all text prompts with the frozen CLIP encoder, obtaining a matrix $X \in \mathbb{R}^{N \times 512}$, and then train h on these fixed features using a standard cross-entropy objective. At inference time, the adapter weights to produce domain-aligned embeddings and class logits and exposes a 512-D vector to the fusion transformer.

3.6 Implementation Details

3.6.1 Video Encoder

The self-supervised training follows the official VideoMAE framework. The encoder is a spatio-temporal Vision Transformer (ViT) that processes video clips of length T_{clip} frames. For each input, 75% of the visual patches are masked, and the mask pattern is randomized at each iteration. The reconstruction target is defined on the pixel space with mean squared error (MSE) loss.

Videos are decoded into fixed-length clips, sampled at 25 fps, and randomly cropped to 224×224 resolution. Each training sample is represented as a tensor of shape $\mathbb{R}^{T_{\text{clip}} \times 224 \times 224 \times 3}$.

Only the encoder is retained after pretraining; the decoder is discarded. The encoder outputs a latent embedding $\mathbf{z} \in \mathbb{R}^{d}$ for each input clip. These embeddings are then extracted for $\pm$**10-second** around each labeled timestamp and passed to the Fusion Transformer, where they are combined with the audio and prompt embeddings for highlight classification.

This design ensures that the pretrained backbone captures domain-relevant motion and context, leading to more structured and separable representations of game events.

3.7 Audio Encoder

The audio encoder implementation follows a simple three-stage pipeline. The pipeline separates the expensive embedding extraction operations from the lightweight training loop through caching. First, we precompute the transcripts for all audio segments using Whisper and store them to disk, translating them to English if need be. Next, we extract Wav2Vec2 embeddings for all segments and cache the 1024-dim vectors. Finally, we load the cached audio embeddings and transcriptions for training. Audio segments were extracted for a $\pm$**10-second** window around each labeled event timestamp. For the latent audio encoder, we extract 20-second segments sampled at 16kHz, which goes through Wav2Vec2's convolutional feature encoder, converting it into a 1024-dim vector. For the Whisper transcription, we use similar duration segments to focus on the most relevant commentary surrounding the event, sending the .wav clip to the Whisper model. The embeddings represent the output of these encoders, which will then subsequently be passed to the Fusion Transformer, where they are combined with the video embeddings for highlight classification.

3.8 Query Encoder

We train the linear adapter using AdamW with learning rate 1×10^{-2} and weight decay 1×10^{-2}, a batch size of 64, and for 5 epochs. All CLIP parameters remain frozen, only the $\mathbb{R}^{512 \times C}$ weight matrix and bias of the linear head are updated. For each epoch, we compute the cross-entropy loss on the embedded training features and then evaluate the current head on the test prompts.

3.9 Multimodal Fusion Transformer

For event classification, we use a lightweight Multimodal Fusion Transformer that operates on the precomputed video, transcript, and audio embeddings described in the previous sections. For each labeled event, we reuse the same ± 10-second window and obtain: a VideoMAE token sequence $V \in \mathbb{R}^{T_v \times 768}$, a single MPNet-based transcript token $T \in \mathbb{R}^{1 \times 768}$, and a single Wav2Vec2-based audio token $A \in \mathbb{R}^{1 \times 1024}$.

Because the transformer operates in a lower-dimensional latent space, each modality is projected to $d_{\text{model}} = 256$ via separate linear layers, yielding $\tilde{V} \in \mathbb{R}^{T_v \times d_{\text{model}}}$, $\tilde{T} \in \mathbb{R}^{1 \times d_{\text{model}}}$, and $\tilde{A} \in \mathbb{R}^{1 \times d_{\text{model}}}$. We then add a standard sinusoidal positional encoding $\text{PE} \in \mathbb{R}^{\text{max\_seq\_len} \times d_{\text{model}}}$ and apply dropout:

$$X_{\text{pos}} = \text{Dropout}\big(X + \text{PE}_{1:L}\big),$$

where X is any modality sequence (or their concatenation) and L is the actual sequence length.

For cross-modal fusion, we concatenate the tokens along the sequence dimension,

$$X_0 = [\tilde{T}; \tilde{A}; \tilde{V}] \in \mathbb{R}^{(T_v+2) \times d_{\text{model}}},$$

So that a single transformer encoder can attend within and across all three modalities. The backbone is a stack of $L = 2$ Transformer encoder layers (`nn.TransformerEncoder`) with $d_{\text{model}} = 256$, $n_{\text{head}} = 4$, and dropout 0.1 in both attention and feed-forward blocks. The encoder outputs a fused sequence $H \in \mathbb{R}^{(T_v+2) \times d_{\text{model}}}$, over which we apply global average pooling to obtain a single representation $h_{\text{fuse}} \in \mathbb{R}^{d_{\text{model}}}$.

Finally, a linear classifier with weights $W_{\text{cls}} \in \mathbb{R}^{C \times d_{\text{model}}}$ and $C = 16$ event classes produces logits $\hat{y} = W_{\text{cls}} h_{\text{fuse}} + b_{\text{cls}}$. Training follows the common protocol used for all baselines (cross-entropy with label smoothing, AdamW, and early stopping on the validation set), enabling a fair comparison between unimodal (V, T, A), bimodal (VT, VA, TA), and trimodal (VTA) configurations.

4 Datasets

4.1 SoccerNet Dataset

The main dataset we rely on is the SoccerNet dataset (3). It is a large-scale dataset of over 700 hours of soccer footage, with pretrained features extracted from broadcasts at 2 frames per second. Many different features are already labeled, such as instances of offsides, goals, penalties, cards, fouls, as well as ball action data. It also captures some features of players, such as their jersey numbers. This data has been used in many CV-sports-related papers, enhancing its credibility. With features extracted at 2 fps, there is over five million potential features to analyze. We additionally use audio data obtained from the video as well, generated by Whisper, which adds another dimension to the data we have.

4.2 Prompt Dataset

To train the adapter, we constructed a small, class-balanced prompt dataset. For each of the action classes from SoccerNet, we manually wrote 10 training prompts that are short and unambiguous, yielding $N_{\text{train}} = 80$ labeled training sentences. We then created a held-out test split of paraphrased prompts that are longer and more natural, designed to mimic realistic user queries.

5 Experiments and Results

Due to the limited time available for full-scale pipeline development, we shifted our original objective from **automatic highlight detection** to **event classification**. While highlight detection would have been a more appropriate baseline task, it required additional temporal localization and post-processing that exceeded our time constraints. Instead, we focused on video+audio clip classification using our Fusion Transformer architecture, which still preserves the core multimodal framework.

5.1 Video Encoder

We compare the embeddings of $\pm$**10-second** segments centered around each labeled event before and after self-supervised pretraining of Video Transformer to evaluate the impact of VideoMAE pretraining. After **PCA (Principal Component Analysis)**-based dimensionality reduction of retaining 85% of variance, K-Means clustering was applied and quantitatively evaluated.

Table 1: Comparison of clustering results before and after VideoMAE pretraining on SoccerNet.

Metric	Before Pretraining	After Pretraining	Change
Silhouette Score (S)	0.42	**0.58**	$\uparrow$ 0.16
Davies–Bouldin (D)	2.31	**1.39**	$\downarrow$ 0.92
Calinski–Harabasz (C)	3120.5	3347.8	$+$ 227.3

The results show a clear improvement in cluster separability:

- The Silhouette Score increased from 0.42 to 0.58, nearing the threshold for well-separated clusters.
- The Davies–Bouldin Index was substantially reduced (2.31 $\rightarrow$ 1.39), indicating stronger intra-class cohesion after domain-specific pretraining.
- The Calinski–Harabasz score showed only a moderate increase, suggesting that overall cluster compactness improved but did not drastically change.

These results confirm that self-supervised pretraining on SoccerNet improves the organization of the embedding space, even without using action labels during training. The extracted representations become more structured and easier to separate, which is beneficial for the final highlight classification performed by the Fusion Transformer.

5.2 Audio Encoder

To evaluate the discriminative quality of the audio embeddings, we trained a classification head on top of the latent audio encoder and the speech text encoder. The classifier architecture consists of 2 fully connected layers with ReLU activation and dropout regularization, mapping from the 1408-dimensional concatenated embedding to the 17 action classes. Only the classifier weights were trained, with Adam as the optimizer, a batch size of 128, a learning rate of 0.0001, and run over 20 epochs.

After training, the encoder achieved a classification accuracy of 29.6%. While this accuracy appears modest at first glance, this baseline performance is substantially better than random chance ($\tilde{5}.8\%$), and demonstrates that the audio features do capture class-relevant information.

5.3 Query Encoder

Evaluation Metrics As hardware constraints prevented us from running the full multimodal retrieval pipeline end-to-end, evaluating temporal highlight localization directly would have produced biased or incomplete results. Instead, we focus on assessing the ability of the Query Encoder to correctly map natural-language queries to the underlying SoccerNet action classes. This evaluation serves as a proxy for the semantic alignment required in a future fully functional query-based retrieval system.

We formulate the evaluation as a sentence-level classification task. For each query, the adapter outputs a probability distribution over SoccerNet action classes. We report top-1, computed by applying a softmax to the adapter logits and selecting the highest-scoring classes.

Results The raw CLIP encoder already achieves a reasonable top-1 accuracy of 68% on the held-out prompt test set, confirming that CLIP captures many of the underlying soccer concepts. After training the few-shot adapter, test top-1 accuracy increases to 77%, while keeping the model extremely compact and fast to train.

5.4 Fusion Transformer: Event Classification

5.4.1 Experimental Setup

To evaluate the benefit of each modality in our trimodal setup, we construct a highlight classification benchmark from SoccerNet using the Multimodal Fusion Transformer described in Section 3.9. We focus on 16 event types (ball out of play, throw-in, foul, corner, goal, cards, etc.), with imbalanced support: from 6–7 examples for *penalty* and *red card* up to 1,676 examples for *ball out of play*. Restricting to a ±10 second window around each annotated timestamp yields a total of 5,820 usable events, split into 4,652 training and 1,164 validation samples.

Those choices concentrate computation around the important moments and avoids processing the full $\sim$750 hours of footage, while still allowing us to train a sufficiently expressive multimodal model.

We evaluate all unimodal, bimodal, and trimodal variants of the same backbone:

- **Text-only (T):** classifier on MPNet sentence embeddings of Whisper transcripts.
- **Video-only (V):** classifier on VideoMAE embeddings of the video window.
- **Audio-only (A):** classifier on Wav2Vec2 embeddings of the raw waveform.
- **Fusion (VT, VA, TA, VTA):** the Multimodal Fusion Transformer with different subsets of the three input branches.

All configurations share the same training loop (optimizer, batch size, and early stopping) and differ only in their input representation and which branches of the transformer are enabled.

5.4.2 Results

Table 2: Overall validation accuracy for all modality combinations. V = Video, T = Text, A = Audio.

Model	T	V	A	VT	VA	TA	VTA
Accuracy	0.527	0.658	0.293	**0.747**	0.662	0.516	0.740

At the level of global validation accuracy, we observe that video-only is the strongest unimodal model (0.658), followed by text-only (0.527), while audio-only lags behind (0.293). Among multimodal variants, **fusing video and text (VT) achieves the best performance** (0.747), slightly outperforming VA (0.662) and the full trimodal VTA model (0.740); TA also underperforms (0.516). In other words, adding Wav2Vec2-based audio on top of video and transcripts does not yield any improvement and even slightly hurt the best VT configuration.

For this reason, **we focus our per-class analysis on the text-only, video-only, and video+text fusion models**. Table 3 reports validation precision for these three settings on the seven most frequent classes plus three less frequent but semantically important events.

The fused VT model consistently outperforms both unimodal baselines. The largest gains appear for events where both modalities provide complementary cues:

- **Foul, indirect free-kick, shots off target.** These classes see **absolute improvements** of $+0.21$, $+0.23$... over the best unimodal baseline. We observed that the commentary often explicitly names the type of event, while the video stream helps disambiguate visually different situations.

Table 3: Validation precision for text-only, video-only, and fused (video+text) models on the seven most frequent SoccerNet event classes, and 3 other less frequent classes.

Class	Count	Text	Video	Fusion
Ball out of play	1676	0.809	0.834	**0.868**
Throw-in	1017	0.512	0.808	**0.837**
Foul	611	0.504	0.568	**0.799**
Indirect free-kick	487	0.260	0.375	**0.583**
Clearance	433	0.247	0.588	**0.600**
Shots on target	299	0.311	**0.426**	0.279
Shots off target	265	0.471	0.412	**0.686**
Corner	249	0.312	0.729	**0.833**
Goal	118	0.333	**0.500**	0.292
Kick-off	136	0.125	**0.469**	0.344

- **Corner, throw-in.** Video-only models already perform well on these visually distinctive events, but fusion**still yields non-trivial gains** (e.g., corner: $0.73 \rightarrow 0.83$), suggesting that text helps resolve borderline cases.

On the other hand, a few classes **degrade** when fusing modalities (e.g., *goal*, *kick-off*, *shots on target*). We hypothesize two main reasons: (i) the strong class **imbalance** (only 118 goals versus 1,676 ball-out events), and (ii) noisy or **misaligned transcripts**, where goal-related words appear in replays or generic commentary rather than exactly within the annotated window. In such cases, the transformer may over-weight ambiguous text cues and hurt an otherwise strong video baseline.

Overall, these results show that combining Whisper-based transcripts with VideoMAE embeddings yields consistent gains, while Wav2Vec2 audio adds little beyond what text and frames already capture. This motivates focusing on the VT configuration when extending the model to query-conditioned highlight retrieval on full matches.

6 Conclusion

In this project, we developed a multimodal pipeline for personalized highlight detection in soccer videos by integrating video, audio, transcripts, and natural-language queries. Our system combined a self-supervised VideoMAE encoder, a dual-stream audio pipeline using Wav2Vec2 and Whisper, as well as a CLIP-based query encoder designed to support natural-language retrieval. These components were evaluated through a multimodal Fusion Transformer used to classify events in short temporal windows, demonstrating that combining modalities yields more structured representations and stronger performance than unimodal baselines.

Our results highlight promising findings in multimodal architectures, observing that a majority of classes had higher classification accuracies when using the Fusion transformer than their unimodal counterparts. Despite these strengths, several limitations shaped our system's final capabilities. We were unable to integrate the CLIP query embeddings into the Fusion Transformer. As a result, the current model performs event classification rather than full query-conditioned retrieval. Additionally, the audio pipeline struggled with noisy and often irrelevant commentary being spoken at times that did not match the actual event. Additionally, video and audio embedding extraction took a lot of time, with each individual embedding requiring multiple seconds of processing time. Time constraints restricted us to proxy tasks instead of full match retrieval and prevented us from evaluating how user queries interact with fused audio-video-text representations.

Going forward, the most impactful next step is to incorporate the CLIP query embeddings directly into the Fusion Transformer so that retrieval can be conditioned on natural language prompts rather than fixed labels. Achieving this will require scaling the model to longer temporal contexts, improving transcript alignment, and improving audio and video embedding generation speed. With these extensions, the pipeline could become a more practical tool for interactive highlight retrieval, tactical video search for analysts, and automated content indexing for broadcasters.

References

[1] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[2] F. Della Santa and M. Lalli, "Automated detection of sport highlights from audio and video sources," *arXiv preprint arXiv:2501.16100*, 2025.

[3] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the CVPR Workshops (CVPRW)*. IEEE / Computer Vision Foundation, 2018, pp. 1711–1720. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w34/Giancola_SoccerNet_A_Scalable_CVPR_2018_paper.pdf

[4] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," 2021. [Online]. Available: https://arxiv.org/abs/2107.00650

[5] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Joint visual and audio learning for video highlight detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8127–8137. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Badamdorj_Joint_Visual_and_Audio_Learning_for_Video_Highlight_Detection_ICCV_2021_paper.pdf

[6] F. D. Santa and M. Lalli, "Automated detection of sport highlights from audio and video sources," *arXiv preprint*, no. arXiv:2501.16100v2, 2025. [Online]. Available: https://arxiv.org/abs/2501.16100v2

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint, 2022, arXiv:2212.04356. [Online]. Available: https://arxiv.org/abs/2212.04356

[8] R. Chakraborty, R. Chakraborty, A. Dasgupta, and S. Chaurasia, "Text-based football action spotting rivals video analysis with llms," Quantum Zeitgeist, https://quantumzeitgeist.com/text-based-football-action-spotting-rivals-video-analysis-with-llms/, June 2025, published June 24, 2025. Accessed: 2025-10-01.

[9] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022. [Online]. Available: https://arxiv.org/abs/2203.12602

[10] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," *Advances in neural information processing systems*, vol. 34, pp. 13 988–14 000, 2021.

A Team Contributions

Member	Contributions
Mathis Doutre	Implemented the transcript embedding and designed the Fusion Transformer, conducting all transformer-based multimodal experiments.
Srikar Viswanatha	Implemented the latent audio encoder and performed the Whisper transcript extraction aligned with event timestamps.
Hang Kim	Implemented the full video embedding pipeline, including frame extraction, preprocessing, and the visual encoder.
Eugénie Laugier	Built the prompt encoder module and contributed to project coordination, meeting organization, and shared documentation.

Table 4: Team Contributions