Multi-task Contrastive Learning for Multimodal Soccer Foul Retrieval

Ahmad Sait Mohammed Bremoo Iyas Baroud Rayan Ghabashi Mazen Albalawi
Silvio Giancola Bernard Ghanem
King Abdullah University of Science and Technology (KAUST)
{ahmed.sait, mohammed.bremoo, iyas.baroud, rayan.ghabashi, mazen.balawi}@kaust.edu.sa
{silvio.giancola, bernard.ghanem}@kaust.edu.sa

Abstract

We address the task of multimodal retrieval for referee decision support in soccer, focusing on the detection and analysis of foul incidents. Leveraging X-CLIP, a video–language model, we construct a system that embeds both video clips and textual descriptions into a shared representation space and performs efficient similarity search with FAISS. To adapt pretrained representations to the soccer domain, we propose a multi-task contrastive learning framework that incorporates video–text alignment, foul-type supervision, and severity-aware objectives. Our retrieval system supports both text-to-video and video-to-text queries, providing interpretable outputs via a Streamlit-based interface. Experiments on the SoccerNet-XFouls benchmark highlight the potential of contrastive fine-tuning for domain-specific discriminability, while t-SNE visualizations reveal that overly complex objectives such as triple-loss training can hinder stable convergence. These findings point to the need for simplified or targeted supervision strategies. Overall, our work contributes a practical retrieval framework and offers insights into adapting vision–language models for fine-grained sports analytics.

1. Introduction

Sports video understanding is a central challenge in computer vision, with applications ranging from highlight generation and tactical analysis to referee decision support. A key task in this domain is video retrieval, where the goal is to efficiently retrieve relevant events from large collections of match footage based on natural language or video queries. In soccer, foul-related events are particularly important: identifying their type, severity, and the corresponding referee decision (no card, yellow card, or red card) is valuable for both automated analytics and human-in-the-loop support.

Recent advances in vision–language pretraining have enabled joint multimodal representations of images, video, and text. In particular, Microsoft's X-CLIP [8], an extension of CLIP [1] to the video domain, uses a dual-encoder architecture to embed both videos and their corresponding textual descriptions into a shared multimodal embedding space. Each indexed instance in this space fuses visual and textual information into a single vector, enabling retrieval between multimodal embeddings rather than separately between video and text. To enable scalable search, these embeddings are stored in a FAISS [9] index with cosine similarity as the distance metric.

In this work, we investigate the use of X-CLIP for soccer foul retrieval. We leverage the SoccerNet-XFouls [3] dataset, which provides video clips of match events along with corresponding video-question-answer triplets. For our retrieval system, we condense the answers into a single textual description for each clip, forming the text modality that corresponds to the video. We first establish a zero-shot baseline, applying pretrained X-CLIP without domain adaptation. While retrieval is feasible in this setting, similarity scores between relevant and non-relevant candidates are often too close, limiting discriminability. To address this, we explore contrastive fine-tuning on this domain-specific dataset to better align soccer video–text pairs in the shared embedding space.

Our contributions are threefold: (i) we present a video retrieval system for soccer fouls using X-CLIP embeddings, (ii) we evaluate its zero-shot performance on the SoccerNet-XFouls dataset, and (iii) we improve retrieval discriminability through contrastive fine-tuning.

The following sections describe our system architecture, dataset processing, and experimental evaluation.

2. Related Work

From VARS with MVFouls to XVARS with XFouls. Early work on soccer referee assistance grounded in video understanding focused on building structured benchmarks for foul incident analysis. Held et al. introduced VARS on top of the MVFouls dataset to study automated support for Video Assistant Referees, formulating future retrieval tasks

over multimodal evidence (video and text) [4]. Building on the limitations observed in SoccerNet-MVFouls (e.g., narrower coverage and weaker textual supervision), the same line of work advanced to XVARS with the SoccerNet-XFouls dataset [3], expanding the number and diversity of incidents and strengthening the cross-modal alignment between vision and language. This progression mirrors the trend in video-language research toward richer annotations and evaluation protocols specifically tailored to decision-support scenarios in sports officiating.

Question–Answer augmentation. A key contribution in the transition from VARS/SoccerNet-MVFouls to XVARS/SoccerNet-XFouls is the explicit addition of question–answer (QA) supervision. Held et al. augment foul incidents with standardized questions (e.g., who committed the offence, what was the action class, was there contact, what was the severity/offence category) and canonical answers aligned with event metadata [3]. This enables both retrieval-as-answering and answer grounding, facilitating evaluation not only of nearest-neighbour retrieval quality but also of decision-relevant semantic fidelity across modalities. In our system, these QA-style labels provide structured targets that complement similarity search, allowing downstream analyses of action class, severity, offence type, and contact.

Retriever systems. Retriever systems map queries and candidates into a shared representation space and rank candidates by similarity. Classical retrievers rely on sparse lexical matching (e.g., BM25), whereas modern neural retrievers learn dense or late-interaction representations. Dense Passage Retrieval (DPR) learns dual encoders optimized for in-batch negatives [6], while ColBERT introduces late interaction to preserve token-level matching signals at scale [7]. In multimodal settings, contrastive pre-training (e.g., CLIP and its video extensions such as X-CLIP) aligns images/videos and text for cross-modal retrieval [2, 8]. Practical large-scale retrieval is commonly accelerated with FAISS for approximate nearest neighbour search [5]. Our project follows this paradigm by using an X-CLIP-based encoder and FAISS-backed indexing for efficient multimodal similarity search over foul incidents.

3. Methodology

System Overview. We propose a multimodal retrieval system for soccer foul analysis that leverages contrastive learning to enable both video-based and text-based search capabilities. The system consists of three main components: (1) a fine-tuned X-CLIP model for multimodal understanding, (2) a FAISS-based similarity search engine, and (3) a Streamlit web interface for user interaction. The architecture enables users to search through a database of soccer incidents using either uploaded video clips or natural language descriptions to retrieve similar foul incidents.

3.1. Data Preparation and Annotation

Dataset Structure. Our dataset comprises soccer match footage where each action is stored as a sequence of 8 extracted frames for each clip in a hierarchical directory structure:

```
data/
 {split}/
    action_{id}/
       clip_0/
          frame_000001.jpg
          frame_000002.jpg
          ...
       clip_1/
    annotations.json
```

Each soccer incident is annotated with four key dimensions. The **Action Class** categorizes the foul type (e.g., “Standing tackling”, “Sliding tackle”), while **Severity** provides a numerical scale (1.0-5.0) indicating the seriousness of the incident. The **Offence** field classifies the violation (e.g., “Offence”, ‘No Offence”), and **Contact** indicates whether physical contact occurred (‘With contact”, “Without contact”).

Text Summarization Pipeline. To enhance the multimodal understanding, we employ a text summarization pipeline using LLaMA-3 to process [3]’s question-answer pairs for each foul instance. The system processes four question-answer pairs for each incident and generates concise, single-sentence summaries that capture the essential details of the foul (e.g. The defender’s tackle on the leg prevented a team from regaining possession or scoring, justifying a yellow card punishment). This summarization process creates rich textual descriptions that complement the visual content.

3.2. Model Architecture

Base Model: X-CLIP. We utilize Microsoft’s X-CLIP (Cross-CLIP) model as our foundation, specifically the `microsoft/xclip-base-patch32` variant. X-CLIP is a vision-language model that extends CLIP’s capabilities to video-text understanding through cross-modal contrastive learning. The model processes video frames through a Vision Transformer (ViT) backbone and text through a text encoder, learning aligned representations in a shared embedding space.

For video input, we implement a robust frame extraction and sampling strategy. Videos are processed using OpenCV

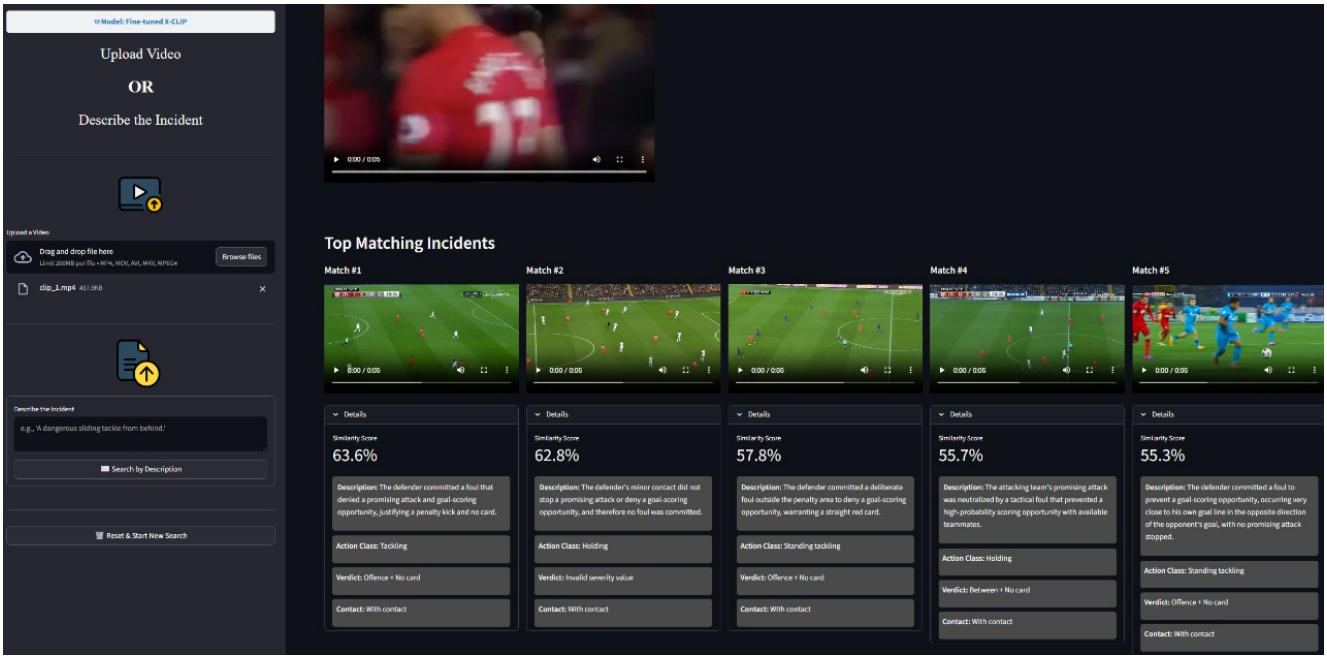


Figure 1. The Multimodal Foul Retriever User Interface.

to extract individual frames, followed by temporal sampling where we select 8 frames evenly distributed across the video duration using linear interpolation. Each frame is converted to RGB format and resized to 224×224 pixels. Multiple clips within an action are processed separately and their embeddings are averaged to create a unified representation.

Text inputs undergo preprocessing that includes concatenation of description and annotation labels, truncation to 300 characters to prevent sequence length issues, and tokenization through X-CLIP's tokenizer with appropriate padding and truncation.

3.3. Multi-Task Contrastive Learning

Triple Contrastive Loss. We introduce a triple contrastive learning objective that extends traditional video-text contrastive learning with domain-specific supervision:

$$\mathcal{L}total = \alpha \cdot \mathcal{L}VT + \beta \cdot \mathcal{L}Foul + \gamma \cdot \mathcal{L}Severity \quad (1)$$

where $\mathcal{L}VT$ represents the video-text contrastive loss, $\mathcal{L}Foul$ denotes foul-type-based video-video contrastive loss, $\mathcal{L}_{Severity}$ indicates severity-based video-video contrastive loss, and α, β, γ are weighting parameters with default values of 0.5, 0.25, and 0.25 respectively.

Video-Text Contrastive Loss The standard video-text contrastive loss follows the InfoNCE formulation:

$$\mathcal{L}VT = \frac{1}{2}(\mathcal{L}v2t + \mathcal{L}_{t2v}) \quad (2)$$

where:

$$\mathcal{L}v2t = -\log \frac{\exp(sim(v_i, t_i)/\tau)}{\sum j = 1^N \exp(sim(v_i, t_j)/\tau)} \quad (3)$$

Here, $sim(\cdot, \cdot)$ represents cosine similarity, τ is the temperature parameter (0.07), and N is the batch size.

Domain-Specific Contrastive Losses. We introduce two additional contrastive objectives that leverage the structured annotations. The Foul-Type Contrastive Loss encourages videos with similar foul types to have similar embeddings while pushing apart videos with different foul types. The Severity Contrastive Loss promotes clustering of videos based on severity levels, enabling the model to learn fine-grained distinctions in foul seriousness.

Loss Mode Configurations The system supports seven different loss configurations ranging from standard video-text contrastive learning (`vt_only`) to full triple contrastive learning (`triple`) and various combinations including foul-type only (`foul_only`), severity only (`severity_only`), and mixed objectives (`vt_foul`, `vt_severity`, `foul_severity`).

3.4. Training Methodology

Dataset Preparation We create video-text pairs by associating each video clip with its corresponding textual description. The dataset includes both the original annotation

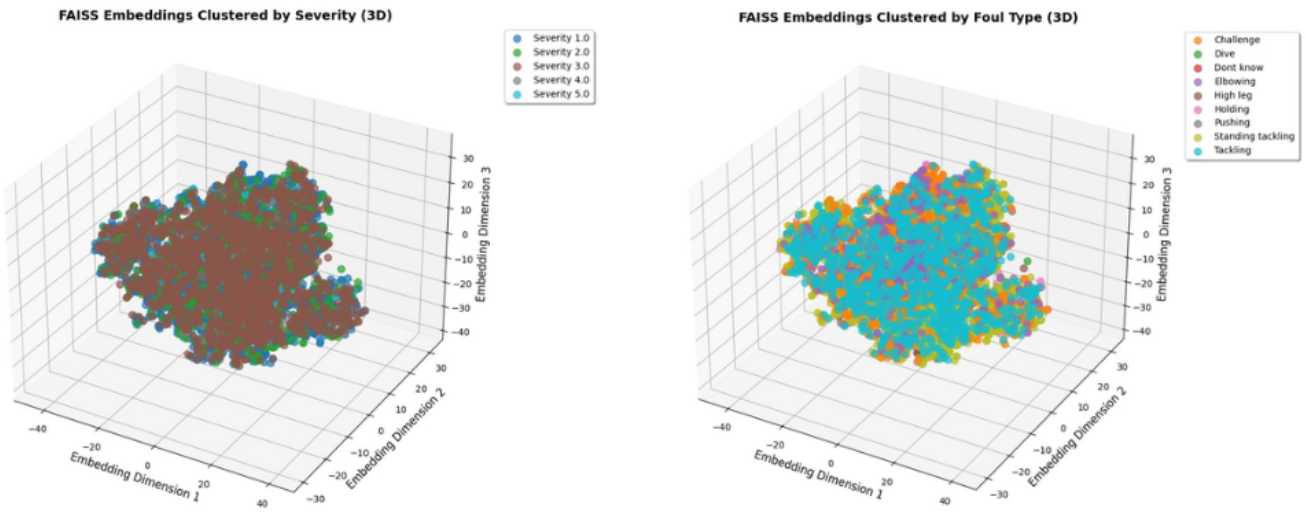


Figure 2. FAISS embeddings clustered by Foul Type and Severity

labels and the generated summaries, providing rich multimodal supervision for the training process.

Training Configuration Our training setup employs AdamW optimizer with a learning rate of 1e-5 and weight decay of 0.01. We use a batch size of 8, limited by GPU memory constraints, and implement a linear warmup for 100 steps followed by linear decay. Early stopping is configured with a patience of 3 epochs based on validation loss. We employ a 90/10 split of the combined dataset for training and monitoring, respectively. The training process begins with a warmup phase for gradual learning rate increase, followed by main training with full triple contrastive learning objectives. The best model is saved based on monitoring loss.

3.5. Retrieval System Architecture

The system generates embeddings for both videos and text queries through the fine-tuned X-CLIP encoders. Video embeddings are processed through the video encoder, while text embeddings utilize the text encoder. All embeddings undergo L2-normalization for consistent cosine similarity computation.

We employ FAISS (Facebook AI Similarity Search) for efficient similarity search using IndexFlatIP (Inner Product) for cosine similarity. The system automatically offloads to GPU when available and stores pre-computed indices on disk for fast loading and retrieval.

The retrieval process begins with query processing where video frames are extracted or text is tokenized. The query is then converted to an embedding using the fine-tuned model, followed by FAISS performing k-nearest neighbor search with k=5. Results are ranked by cosine similarity scores, and associated annotations and descriptions are retrieved for display.

3.6. User Interface Design

Streamlit Web Application The system provides an intuitive web interface built with Streamlit that supports video upload through a drag-and-drop interface for various formats (MP4, MOV, AVI, MKV), natural language description input with example placeholders, and results display in a grid layout showing top-5 similar incidents with similarity scores and detailed information in expandable sections.

Model Selection and Result Visualization Users can choose between the base model (zero-shot X-CLIP) for baseline comparison and the fine-tuned model for enhanced performance. As shown in Figure 1, each retrieved result displays a percentage-based similarity score, generated video preview from extracted frames, comprehensive foul classification including action class, severity, offence, and contact information, along with the generated text description of the incident.

3.7. Results

Embedding Space Visualization. We further evaluate the learned representations by projecting the embedding space into two dimensions using t-SNE. Ideally, distinctive clusters should emerge, reflecting semantic separability of foul incidents by action class and severity. However, as shown in Figure 2, the model did not converge to well-formed clusters even after multiple training attempts. This suggests that the triple-loss objective may introduce excessive complexity, preventing stable optimization. Simpler loss configurations (e.g., video–text alignment with

foul-type only, or video–text alignment alone) may provide stronger supervision signals and yield more coherent embeddings. Future work will explore these combinations to balance alignment across modalities with domain-specific discriminability.

4. Conclusion

In this work, we presented a multimodal retrieval system for soccer foul analysis based on X-CLIP embeddings and FAISS-backed similarity search. Starting from a zero-shot baseline, we demonstrated that while pretrained video–language representations provide a feasible starting point, similarity scores were often insufficiently discriminative for fine-grained foul retrieval. To address this, we introduced a multi-task contrastive learning framework that incorporates video–text alignment, foul-type supervision, and severity-based objectives.

Our experiments highlighted both the promise and the challenges of this approach. Although fine-tuning improved retrieval performance qualitatively, visualization of the learned embedding space with t-SNE revealed that the triple-loss configuration struggled to produce well-separated clusters. This suggests that overly complex objectives may hinder convergence, and that simpler or more targeted configurations could yield more coherent embeddings.

Future work will focus on exploring alternative loss formulations, integrating larger and more diverse soccer datasets, and extending the evaluation with quantitative retrieval metrics such as Recall@K and mean reciprocal rank. Beyond retrieval, we also envision the integration of our system into decision-support pipelines for referees and analysts, where transparent, interpretable retrieval can complement human judgment in high-stakes scenarios.

References

[1] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision. volume 139, pages 8748–8763, 2021. 1

[2] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[3] Anthony Cioppa Silvio Giancola Bernard Ghanem Marc Van Droogenbroeck Jan Held, Hani Itani. X-vars: Introducing explainability in football refereeing with multi-modal large language model. In *CVsports*, 2024. 1, 2

[4] Silvio Giancola Abdullah Hamdi Bernard Ghanem Marc Van Droogenbroeck Jan Held, Anthony Cioppa. Vars: Video assistant referee system for automated soccer foul analysis on mvfouls. *arXiv preprint arXiv:2304.04617*, 2023. 2

[5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. In *IEEE Transactions on Big Data*, 2019. FAISS. 2

[6] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020. 2

[7] Omar Khattab and Ion Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, 2020. 2

[8] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1, 2

[9] Chengqi Deng Jeff Johnson Gergely Szilvasy Pierre-Emmanuel Mazaré Maria Lomeli Lucas Hosseini Hervé Jégou Matthijs Douze, Alexandr Guzhva. The faiss library. 2024. 1