Introduction This report presents the methodology, design rationale, and implementation details developed for Round 1 of the EdgeFleet.AI recruitment assessment. The objective of the assignment is to construct a functional cricket ball detection and tracking system using a single fixed-camera video input, closely resembling real-world sports analytics pipelines. The proposed solution integrates video preprocessing, multimodal grounding-based localization, and video-level segmentation and tracking. By leveraging recent advances in vision–language models and foundation segmentation models, the pipeline is designed to be modular, reproducible, and robust to common broadcast-specific artifacts present in cricket videos.

Problem Understanding Broadcast cricket videos typically contain several visual elements such as scorecards, channel logos, branding overlays, and user interface components positioned near the top and bottom of the frame. These elements are irrelevant to ball motion analysis and can introduce noise, false positives, or distract downstream models. Additionally, the cricket ball itself poses a challenging detection problem due to its small size, high speed, frequent motion blur, and partial occlusions. As a result, a straightforward single-model approach is insufficient, motivating the need for a carefully designed multi-stage pipeline that combines complementary model capabilities for reliable detection and tracking.

Video Preprocessing An initial inspection of the provided videos revealed prominent borders and overlays that do not contribute meaningful information for ball tracking. Including these regions increases computational overhead and degrades model performance. To address this, a center-cropping strategy was applied uniformly across all videos, retaining 0.8 of the original width and 0.7 of the original height while preserving the spatial center of the frame. This preprocessing step effectively removes distracting UI elements, condenses the spatial focus to the pitch and ball trajectory region, and improves the signal-to-noise ratio for subsequent stages. All cropped videos are saved and consistently reused throughout the pipeline to ensure reproducibility.

Pipeline Design The overall system follows a multi-model boosting strategy in which the outputs of one model are used to guide and refine another. The pipeline begins with video preprocessing, followed by initial ball localization using a multimodal grounding-based large language model. This coarse localization is then passed to a video-level segmentation and tracking model, and the resulting track is further refined using point-based supervision derived from the grounding model. The final stage generates trajectory visualizations and structured annotations. **Multimodal Ball Localization using MOLMO** Traditional object detectors often struggle to localize the cricket ball without extensive task-specific training. To overcome this limitation, the pipeline

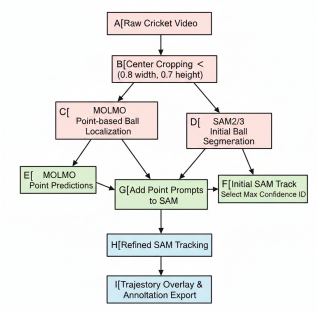

Figure 1. Inference pipeline Description

employs MOLMO, a multimodal grounding-based model capable of localizing objects using natural language prompts. Video frames are sampled at a fixed interval and provided to MOLMO with the prompt "ball." The model produces point-based predictions indicating the ball's location, each associated with an object identifier and confidence score. Although sparse, these predictions are semantically grounded and provide reliable cues that serve as weak supervision for downstream tracking.

Video Segmentation and Tracking with SAM3 For video-level segmentation and propagation, the pipeline utilizes SAM3 (Segment Anything Model v3) from Meta AI Research. SAM3 supports prompt-based segmentation, temporal propagation across frames, and object ID consistency, making it well suited for refining object tracks once an approximate initialization is available. The model is initially prompted with the text "ball" on sampled frames, producing multiple candidate tracks. The track exhibiting the highest confidence and temporal consistency is selected as the initial ball track.

Cross-Model Refinement Strategy While SAM3 provides strong segmentation capabilities, it may drift in challenging scenarios such as rapid motion, blur, or partial occlusion. In contrast, MOLMO offers semantically precise but sparse point predictions. To exploit the complementary strengths of both models, MOLMO's ball point predictions are injected as additional prompts into the current SAM3 track. SAM3 then re-segments and propagates the refined object mask across frames. This iterative refinement process improves spatial accuracy, enhances temporal consistency, and significantly increases robustness against tracking failures.

Conclusion This project demonstrates a hybrid vision–language and segmentation-based approach to cricket ball tracking that effectively addresses the challenges posed by broadcast video artifacts and fast-moving, small objects. The modular design further enables future extensions such as physics-aware trajectory modeling, multi-camera integration, and real-time deployment.