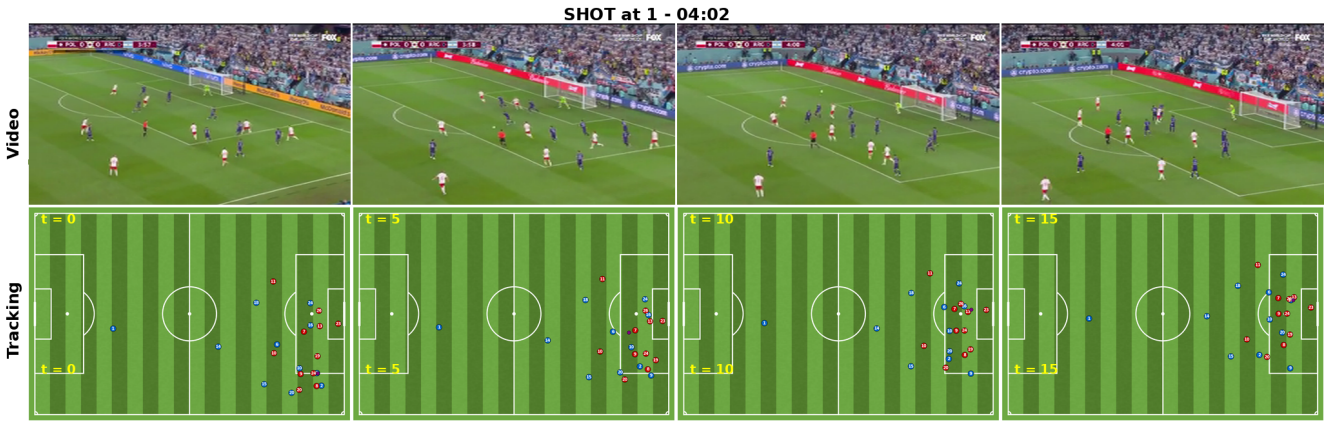

Figure 6. **Video vs Tracking Input for a SHOT Event.** Top: broadcast frames at four timesteps showing camera motion and partial field coverage. Bottom: synchronized tracking data on a standardized pitch, providing complete spatial context for all players regardless of camera position.

8. Visualization of data synchronization

Figure 6 illustrates the two input modalities for a *SHOT* event at match time 04:02. The top row shows four broadcast frames sampled at timesteps $t \in \{0, 5, 10, 15\}$ within the observation window. The camera pans and zooms to follow the play, introducing viewpoint changes, partial occlusions, and background clutter that the video encoder must contend with. The bottom row shows the corresponding tracking data projected onto a standardized pitch template, where red and blue markers denote players from opposing teams and the purple marker indicates the ball position. The tracking representation abstracts away visual noise and captures the spatial configuration of all 22 players in a canonical coordinate frame, making tactical patterns such as defensive compactness and attacking movement directly readable. This example highlights a key advantage of tracking modality: while the broadcast view loses several players outside the frame at each timestep, the bird's-eye tracking view preserves the full team-level spatial context throughout the observation window.

9. Data scaling

Figure 7 shows performance versus training set size for both modalities, using GIN + MaxPool + Positional edges for tracking and VideoMAEv2-B finetuned for video (mean over 5 seeds for tracking). Tracking achieves 67.0% with only 5 matches and plateaus around 35 matches (78.5%), while video achieves 41.6% with 5 matches and also plateaus around 35 matches (62.4%), after which both modalities fluctuate slightly. The gap narrows from 25.4% at 5 matches to 16.1% at 35 matches, indicating that video benefits more from additional data but remains substantially behind. Tracking maintains superior performance across all data regimes, suggesting that structured positional representations require less training data to reach strong performance.

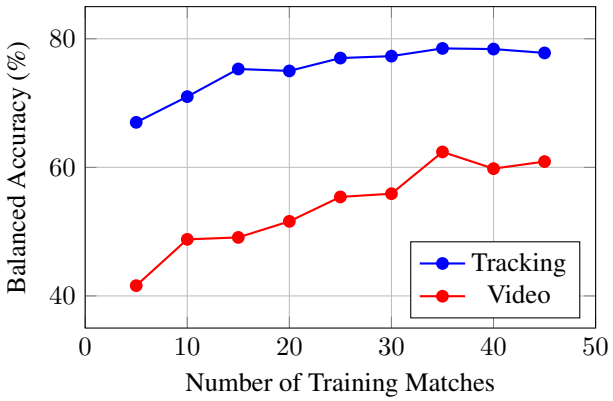

Figure 7. **Performance *vs*. Number of Training Matches for Video and Tracking Modalities.** Both modalities plateau around $\approx$ 35 matches, but tracking reaches strong performance (67.0%) with as few as 5 matches, whereas video starts at 41.6%. The gap narrows from 25.4% (5 matches) to 16.1% (35 matches).