Streaming Video Crime Anticipation with Spatio-Temporal Causal Reasoning

Supplementary Material

1. Online Human Verification Platform

To ensure annotation quality described in main paper, we developed a web-based annotation platform with three modules, to validate distinct annotation types: temporal events, spatial relationships and causal dependencies.

Temporal Event Verification. (Figure 1) The first module validates the binary classification of events into criminal acts versus criminal precursors. Assessors review the video alongside LLM-generated preliminary labels and event timelines. By cross-referencing the visual content with the crime type and event descriptions, they verify whether each event constitutes a crime act or serves as a precursor. This process corrects model misclassifications and ensures precise demarcation of the causal event chain.

Spatial Relationship Verification. (Figure 2) The second module refines entity-level spatial annotations through a multi-stage workflow. The interface displays detection results with bounding boxes, depth maps for geometric validation, and frame-by-frame tracking. Assessors first validate and correct object detection outputs, adjusting mislabeled entities and imprecise boxes. Next, they proceed to entity tracking and verification, where they rectify identity switches, disappears and ensure trajectory consistency across frames. Finally, the module provides a depth quality check, allowing assessors to discard frames with corrupted or invalid depth estimates, thereby maintaining spatial annotation integrity.

Causal Relation Verification. (Figure 3) The third module distinguishes causal dependencies from coincidental temporal sequences. Assessors analyze candidate events presented with their temporal windows and textual descriptions. For each pair, they determine whether the preceding event logically triggers or influences the subsequent event, rather than merely co-occurring. This verification ensures that only samples exhibiting predictive causality, where earlier events serve as meaningful precursors are retained in the dataset, while temporally adjacent but causally unrelated sequences are discarded.

2. Task Example Case

To illustrate the dataset structure, we provide representative samples for both the Spatio-Temporal Causal Reasoning (STCR) tasks and the downstream crime anticipation tasks. Figure 4 presents the five STCR tasks, showing the input query format and ground truth annotations for each hierarchical reasoning level from local causal prediction (Task 1-2) to global structural reasoning (Task 3-4) and entity-level attribution (Task 5). Figure 5 demonstrates the three downstream evaluation tasks: future crime detection, classification, and temporal prediction, displaying their query structures and target label formats.

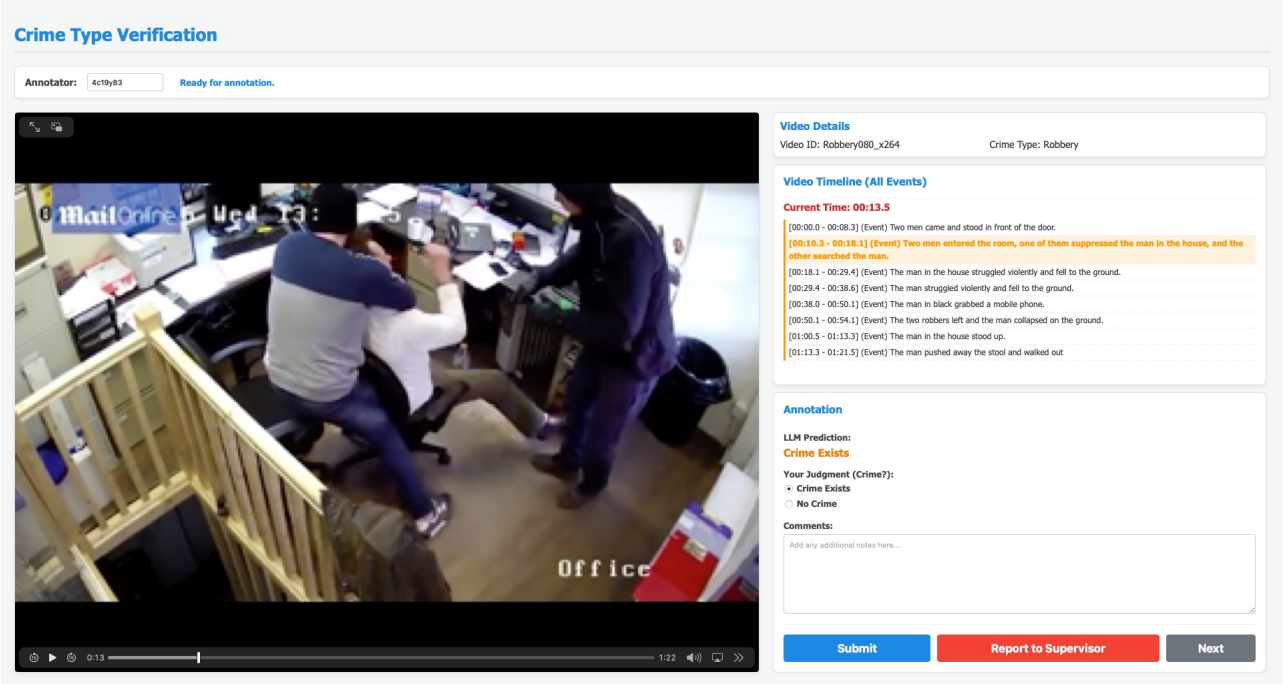

Figure 1. Visualization of the Crime Type Verification Interface. The platform employs a model-assisted workflow where assessors review video segments to validate binary causal labels, explicitly distinguishing between precursor events and actual crimes to rectify misclassifications generated by the LLM.

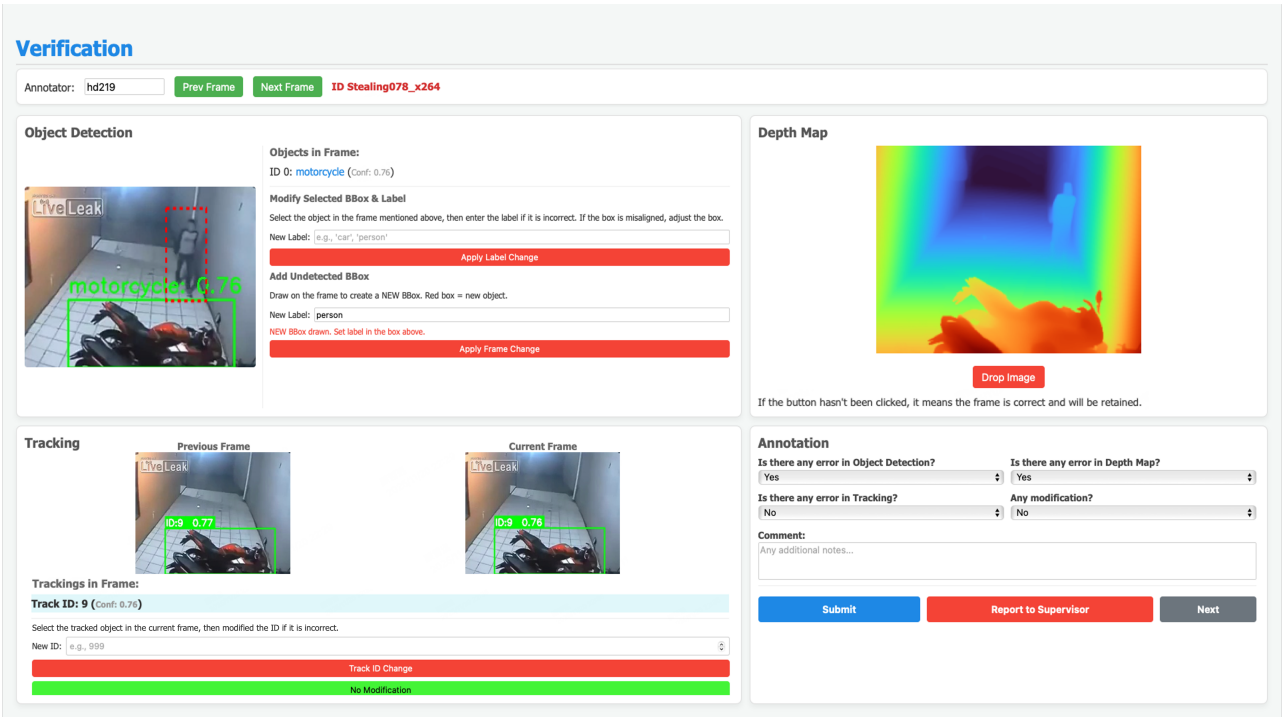

Figure 2. Visualization of the Spatial Relationship Verification Interface. The platform enables assessors to refine spatial annotations by manually correcting entity detection and tracking trajectories while filtering out frames with invalid depth estimations to ensure precise spatial modeling.

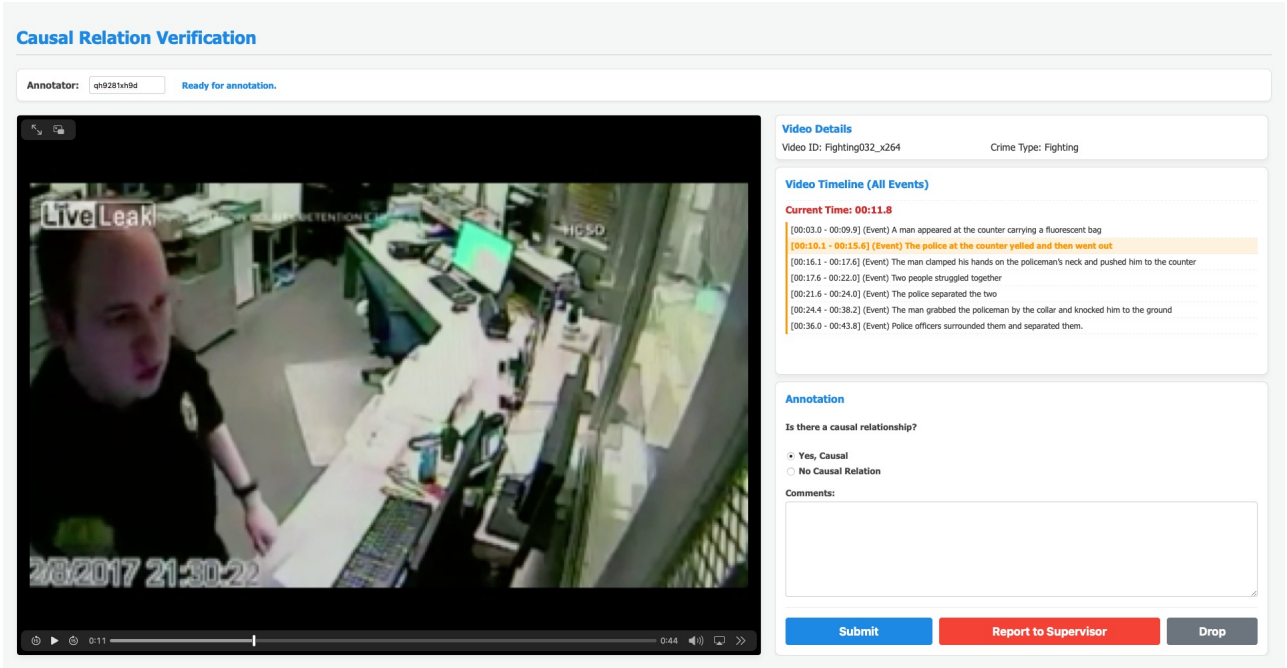

Figure 3. Visualization of the Causal Relation Verification Interface. The platform requires assessors to explicitly validate the logical causal linkage between sequential events to ensure the dataset captures causal precursors.

STCR Task1:

Query Time: 17.8
Question: Anticipate the next event based on the current event scene.
Ground Truth: Man poured gasoline on the car, mainly on the rear and top of the car.

STCR Task2:

Query Time: 02:04.0
Question: Predict the spatial layout for the upcoming 4.0-second interval.
Ground Truth: person (TrackID 2) is at the front-top-right of gun (TrackID 1) (Distance: 42.8 units).

STCR Task3:

Query Time: 52.3
Question: Based on the information available up to this point, please reorganize the following events to reflect the correct chronological sequence. Events: 1: A woman got off the car from the back door and walked to the door of the house next to her. 2: The door of a white car parked in front of the house opened. 3: The man on fire kept running, running from one side to the other and back again, but the fire never went out. 4: The man came out to the door again and lit a fire and burned himself. 5: A man in black got out of the car at the front door and poured gasoline on the door of the house.
Ground Truth: $2 \rightarrow 1 \rightarrow 5 \rightarrow 4 \rightarrow 3$.

STCR Task4:

Query Time: 05:30.0
Question: What causal events preceded the event happening right now in the video? 1. A police car came and parked next to the police car at the beginning. 2. There are two silver cars driving opposite each other on the road and passing each other. 3. Many police cars and police officers arrived one after another. 4. The police moved around the man in white with a gun and spoke on a walkie-talkie. 5. The man in white was subdued by two people and dragged out of the left side of the screen.
Ground Truth: 1. A police car came and parked next to the police car at the beginning. 3. Many police cars and police officers arrived one after another. 4. The police moved around the man in white with a gun and spoke on a walkie-talkie.

STCR Task5:

Query Time: 07:24.0
Question: What group of entities most likely caused the event happening right now in the video? Entities: gun (TrackID 1), person (TrackID 1), person (TrackID 2), person (TrackID 3), car (TrackID 1).
Ground Truth: gun (TrackID 1), person (TrackID 1), person (TrackID 2).

Figure 4. Visualization of the five hierarchical Spatio-Temporal Causal Reasoning tasks. The figure presents a sample input query alongside its corresponding ground-truth answer for each reasoning level ranging from local causal prediction to entity attribution.

Downstream Detection (AUC):

Query Time: 16.0
Question: Looking ahead 2.0 seconds, will any criminal activity be detected?
Ground Truth: Yes, a crime is positive for the upcoming interval.

Downstream Classification (WF1):

Query Time: 12.0
Question: Classification of the anticipated crime event.
Ground Truth: The anticipated crime event is Shooting.

Downstream Prediction (TimeDiff):

Query Time: 36.0
Question: Please estimate the time remaining until the incident.
Ground Truth: Forecast: 13.8 seconds until the event.

Figure 5. Visualization of the downstream crime anticipation tasks. The figure displays queries and ground-truth targets for future crime detection, crime type classification, and time-to-crime prediction.