Designing Multi-Robot Ground Video Sensemaking with Public Safety Professionals

Puqi Zhou
Computer Science
George Mason University
Fairfax, VA, USA
pzhou@gmu.edu

Ali Asgarov
Computer Science
Virginia Tech
Blacksburg, VA, USA
aliasgarov@vt.edu

Aafiya Hussain
Computer Science
Virginia Tech
Blacksburg, VA, USA
aafiyahussain@vt.edu

Wonjoon Park
Computer Science
University of Maryland
College Park, MD, USA
wpark814@umd.edu

Amit Paudyal
Computer Science
George Mason University
Fairfax, VA, USA
apaudya@gmu.edu

Sameep Shrestha
Computer Science
George Mason University
Fairfax, VA, USA
sshres32@gmu.edu

Chia-Wei Tang
Computer Science
Virginia Tech
Blacksburg, VA, USA
cwtang@vt.edu

Michael Lighthiser
George Mason University
Fairfax, VA, USA
mlighthi@gmu.edu

Michael Hieb
C5I Center
George Mason University
Fairfax, VA, USA
mhieb@gmu.edu

Xuesu Xiao
Computer Science
George Mason University
Fairfax, VA, USA
xiao@gmu.edu

Chris Thomas
Computer Science
Virginia Tech
Blacksburg, VA, USA
chris@cs.vt.edu

Sungsoo Ray Hong
Information Sciences and Technology
George Mason University
Fairfax, Virginia, USA
shong31@gmu.edu

Abstract

Videos from fleets of ground robots can advance public safety by providing scalable situational awareness and reducing professionals' burden. Yet little is known about how to design and integrate multi-robot videos into public safety workflows. Collaborating with six police agencies, we examined how such videos could be made practical. In Study 1, we present the first testbed for multi-robot ground video sensemaking. The testbed includes 38 events of interest relevant to public safety, a dataset of 20 robot patrol videos (10 day/night pairs) covering EoI types, and 6 design requirements aimed at improving current video sensemaking practices. In Study 2, we built MRVS, a tool that augments multi-robot patrol video streams with a prompt-engineered video understanding model. Participants reported reduced manual workload and greater confidence with LLM-based explanations, while noting concerns about false alarms and privacy. We conclude with implications for designing future multi-robot video sensemaking tools.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → *Computer vision*; • **Applied computing** → *Evidence collection, storage and analysis.*

Keywords

Human-Centered Design, Video Sensemaking, Ground Robotics Fleet, Public Safety

ACM Reference Format:
Puqi Zhou, Ali Asgarov, Aafiya Hussain, Wonjoon Park, Amit Paudyal, Sameep Shrestha, Chia-Wei Tang, Michael Lighthiser, Michael Hieb, Xuesu Xiao, Chris Thomas, and Sungsoo Ray Hong. 2026. Designing Multi-Robot Ground Video Sensemaking with Public Safety Professionals. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790679

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790679

1 Introduction

Public safety professionals remain understaffed and disproportionately exposed to injury through in-person operations. US national data show that public safety professionals experience injury rates more than four times higher than the average occupation [127]. A Department of Justice study of 18 local agencies reported nearly 1,300 injuries in a single year, causing 6,000 missed workdays and about $2 million in overtime costs [50]. Advances in robotics and

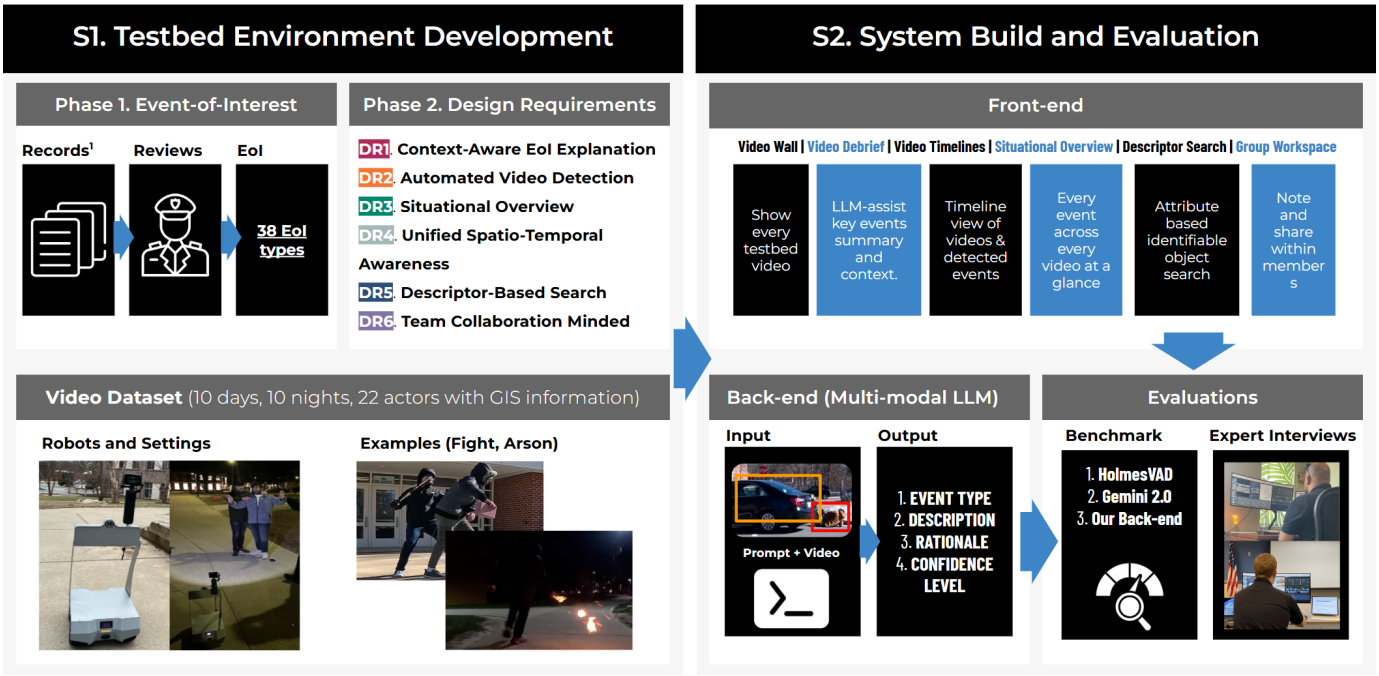

Figure 1: Research flow: (1) Identify 38 EoIs and 6 DRs from records[1] (13,234 crime records from three US campuses and 10 research anomaly video datasets [1, 74, 80, 96, 97, 99, 103, 134, 144, 158]) with five public safety professionals survey reviews and interviews; (2) Create a 20-video multi-robot testbed simulating these EoIs; (3) Build the MRVS system with front-end interface and multimodal LLM back-end; (4) Evaluate via benchmarking and expert interviews with nine professionals.

semi-autonomous control now make fleets of ground robots capable of providing scalable situational awareness for police operations [73]. The use of video technologies for situational awareness is not new: agencies have long expanded their reliance on digital imagery, from fixed surveillance cameras [53] to in-car systems [100], body-worn cameras [140], and aerial drones [44]. Most recently, ground robots are being tested in public sector operations [6, 27]. However, early deployments highlight misalignment risks. The New York Police Department's K5 patrol robot, for example, was deployed for six months but retired after two months of use due to limited effectiveness and imposed operational burden [25].

While designs that connect public safety professionals with ground-robot video footage promise both societal and technological benefits, there has not been much research in how to make such connections practical and aligned with operational contexts. Within HCI, public safety professionals have long been studied, but research has focused mainly on fixed surveillance systems [53, 90], crime prediction [42], community engagement [41, 93], and decision making [39, 129]. A recurring insight across these studies is the need to involve professionals in the design process [130], since interventions developed without their input risk adding burdens rather than support [126]. Beyond the public safety domain, HCI research on video interaction has explored video learning [55, 157], video editing [49, 63], and video assessment [52, 128], showing how users can be supported in extracting insights from large volumes of video data. Yet despite these advances, video analysis in public safety remains manual and labor-intensive [121, 126]. In short, enabling agencies to make practical use of multi-robot video streams requires substantial adaptation grounded in user-centered design to bridge emerging robotic and AI capabilities with operational needs.

While ground robotics could play a transformative role in high-stakes public-safety operations, we still lack a clear understanding of how these systems should be designed, built, and evaluated in ways that reflect public-safety professionals' real practices, constraints, and decision-making needs. To mitigate this gap, this work first presents a **testbed environment**[1] for multi-robot video sensemaking , co-designed with experienced public-safety professionals, that future researchers can build upon (Study 1, see Fig. 1, S1). This testbed addresses the current gap by providing a first-of-its-kind design and evaluation environment capable of supporting multi-robot, multi-video workflows in realistic public-safety operational settings. Next, this work assesses the testbed's practicality with public safety professionals through the development of an interactive **Multi-Robot Video Sensemaking System**, (**MRVS**; Study 2, see Fig. 1, S2), conducted in collaboration with six local and state public safety agencies. MRVS is the first interactive system that operationalizes multi-robot video sensemaking for public safety by tightly aligning a multimodal AI back-end and a human-centered front-end with public safety professionals' procedures, concerns, and adoption constraints.

In S1, we (1) identify the types of events captured by ground robots that are relevant to public safety operations (see Fig. 1, Event-of-Interest), (2) construct a multi-robot video dataset with ground-truth labels and context (see Fig. 1, Video Dataset), and (3) derive

[1]The testbed is available at: https://github.com/Puqi7/MRVS_VideoSensemaking

design requirements for MRVS-like tools (see Fig. 1, Design Requirements). To define **Events of Interest (EoIs)**, we analyze three years of crime records from three US campuses and 10 public anomaly video datasets in collaboration with five public safety professionals, producing a taxonomy of 38 visually observable events aligned with their operational context. Building on the EoIs taxonomy, we create a **Video Dataset** of 10 daytime and 10 nighttime patrol videos, recorded with a ground robot and 22 actors, each with a duration of 22–30 minutes. Finally, we collaborate with the same five professionals to derive six **Design Requirements (DRs)** that specify how future systems and interfaces can better support situational awareness with robot-captured video.

In S2, we assess the testbed's practicality by designing, implementing, and evaluating MRVS, a novel system integrating a multimodal LLM back-end and an interactive front-end. (see Fig.1, Back-end), while the interface provides six core features driven from DRs: Video Wall, Video Debrief, Video Timeline, Situational Overview, Descriptive Search, and Group Workspace (see Fig.1, Front-end). Benchmark evaluation shows that the back-end achieved F1 score gains of 6% (day), 23% (night), and 15% overall compared to Gemini 2.0 [119], and substantially outperformed HolmesVAD [153]. To evaluate the front-end, we conduct expert interviews with nine public safety professionals. They report that MRVS improved situational awareness, investigation speed, and reduced manual video analysis effort, while also raising concerns about false alarms and privacy risks in MRVS deployment. Synthesizing findings from both studies, we outline implications for designing future robot-enabled video sensemaking tools for public safety operations.

This work presents the testbed environment for studying video-sensemaking interfaces generated by multiple ground robots in public safety contexts, enabling future researchers to build and evaluate systems using the expert-validated, labeled video dataset with reproducible design requirements. MRVS is the first interactive system grounded in public safety professionals' practices, providing a reference point for future MRVS-like systems that leverage multi-robot video workflows for public safety. MRVS demonstrates that AI-augmented multi-video sensemaking can enable one operator to effectively supervise multiple robots—a critical capability for scaling safety operations under chronic staffing constraints. This work offers the following contributions:

- **Testbed environment:** We introduce the first multi-robot video sensemaking testbed with (i) a taxonomy of 38 events of interest, (ii) a public ground-robot videos dataset, and (iii) six design requirements derived from public safety professionals.
- **System artifact:** We develop MRVS, a multi-video sensemaking system with back-end model and front-end interface innovations.
- **Back-end and Front-end Evaluation:** We evaluate back-end through benchmarking against baselines and front-end via expert interviews, providing technical and human-centered insights.
- **Implications for design:** We identify practical, ethical, and privacy considerations in aligning video sensemaking systems with public safety priorities, offering directions for future research.

2 Related Work

Gaining situational awareness through ground robots is an emerging problem space. While few studies have addressed it directly, related work in HCI and Computer Vision (CV) offers useful insights. In HCI, collaborations with public safety professionals have highlighted the value of user-centered design in safety-critical settings [13, 41, 45]. Other HCI studies developed video interaction tools to support sensemaking [29], though rarely in public safety contexts. HCI research has also emphasized ethical and privacy concerns in video surveillance [19]. Meanwhile, computer vision has shown longstanding interest in abnormal event detection [153], mostly from fixed surveillance perspectives. Yet the event types examined were not formally elicited from public safety professionals. In this review, we examine each domain in turn, highlighting both their progress and the gaps that remain.

HCI: Working with Public Safety Professionals. When introducing technological aids, including AI-driven tools, in safety-critical contexts, grounding design in the situated practices of public safety professionals is critical. Prior work highlights a persistent misalignment between sophisticated video technologies and interpretation expertise, underscoring the need for systems that augment rather than replace professional judgment [121, 126]. To be adopted, such tools must integrate seamlessly into existing workflows rather than impose additional burdens [40]. Studies show that officers value explainability framed in policing terms [24, 45] and moderate forms of AI assistance that reduce cognitive load while maintaining professional control [5, 141]. Adoption is shaped by organizational dynamics and stakeholder engagement [57], reinforcing the importance of trust [108], accessibility [4], and institutional fit [129].

Although ground robots are emerging in public safety, no HCI study has examined this space; related work focuses on other domains. Agencies have expanded sociotechnical infrastructures with diverse video sources—from fixed CCTV [155] to body cameras [140], drones [44], and automated analytics [6, 100], yet empirical understanding of how these tools reshape frontline policing remains limited [10]. Research shows surveillance technologies often increase, rather than alleviate, officers' cognitive burden when monitoring massive, fragmented streams, while data-driven approaches overlook experiential knowledge essential in high-stakes contexts [90, 130]. While AI systems have been piloted in areas such as patrol planning [129], crime forecasting [111], and crime mapping [42], questions remain about aligning emerging multi-robot videos with frontline professionals' sensemaking practices.

HCI: Human-Video Interaction. Extensive research has advanced interactive designs for video streams and content, yet public safety investigations still rely on manual, labor-intensive practices. Many systems support video sensemaking, including tools for navigation [150], retrieval [124], and collaborative analysis across multiple streams [110, 145]. Recent work has explored chunk-based editing of interview footage [63], instruction-following video interactions [29], and crowd-assisted annotation [58]. Systems such as ChronoViz [30] and CrossA11y [70] indicate that aligning notes, metadata, and multimodal cues deepens interpretation. Other studies highlight the heavy cognitive demands of reviewing complex video [67, 122], challenges of assessing video credibility [48], and interface designs for multi-video layouts [3, 90]. Despite this progress, design assumptions behind these systems are often rooted in creative or interaction settings rather than policing, leaving challenges of fragmented, distributed multi-robot video largely unexplored.

HCI: Ethical and Privacy Considerations. HCI research has examined the ethical and privacy implications of surveillance, including community perceptions of being monitored, when adopting intelligent, AI-driven video analysis for public safety. Studies show that while robotic patrols and AI detection tools promise safer engagements [60, 84], they risk harm when introduced without transparency or oversight [94, 109]. Public backlash to San Diego's covert streetlight program illustrates how hidden monitoring undermines trust and fuels community demands for accountability [139]. Research notes that video capture in outdoor public environments is often considered acceptable [149], but only when safeguards are in place to balance safety benefits with privacy protection [7].

CV: Abnormality Detection. The computer vision community has extensively developed video understanding models, classifiers, and datasets for public safety, often framed as "anomaly detection". However, most efforts focus on fixed-perspective surveillance footage rather than ground-robot video, and event types have not been defined in consultation with public safety professionals, limiting their value as practical benchmarks. Advanced vision models [135, 156] trained on fixed video perform poorly on noisy, mobile streams. Core challenges include context-dependent anomaly definitions, data imbalance, and real-time constraints [36, 47]. Methods range from unsupervised normality learning [137], to weakly-supervised video-level MIL [116, 117], and *supervised* frame-level labeling [79], with advanced such as autoencoder reconstruction [43], frame prediction [69], adversarial training [104], memory models [2], and transformers for temporal reasoning [148]. Recent work explores recurrent autoencoders [132] and vision-language models such as Video-LLaMA2 [18] and Gemini [119], enhancing scene reasoning while raising efficiency and deployment challenges [89, 146]. Existing abnormal resources include campus footage [134], online clips [97], or long-form surveillance video [96].

As we reviewed the four domains, we found that each offers useful directions for designing practical tools in multi-robot video sensemaking, yet key gaps remain. First, no design or artifact research has addressed how public safety professionals might interact with multiple video streams from the moving perspectives of ground robots. Second, event types relevant to real-world public safety operations have not been systematically defined in collaboration with professionals. Third, without such event definitions, no datasets exist to simulate multi-robot video sensemaking scenarios for research and design. Addressing these gaps requires a testbed environment to support the development and evaluation of new tools in the multi-robot video sensemaking problem space.

3 S1. Formative Study

We aim to build a testbed environment grounded in the operational realities of public safety work and identify professionals' design requirements for future MRVS-like systems development. We conducted a two-phase formative study to ground our design in the realities of public safety practice with five public safety professionals combining a survey (Phase 1) and expert interviews (Phase 2) to anchor our design in real-world practice. All study procedures were reviewed and approved by our university's IRB, and informed consent was obtained from all participants. Specifically, this study addressed the following two research questions:

- **RQ1.** Which events captured by multi-robot video are perceived as relevant and urgent for public safety?
- **RQ2.** How do public safety professionals conduct video investigations with technological aids, what challenges remain, and how do they collaborate during the process?

Beyond the two RQs, we explored professionals' perceptions of multi-robot video sensemaking as an emerging technology, what societal and ethical considerations arise in its deployment, and what further issues future researchers should keep in mind when engaging with this domain.

3.1 Participants and Recruitment

Police work is high-stakes, time-sensitive, and bound by strict confidentiality and protocols, making direct recruitment for our studies challenging. To apply a human-centered design approach to building MRVS into public safety workflow, we first established trust and long-term engagement with police agencies. Over six months, three authors presented the MRVS vision seven times across multiple police departments in one US state, gradually building relationships with three representatives from different agencies. These representatives agreed to support recruiting other police officers for future interviews and system evaluations. To facilitate recruitment, we created an outreach package including a short explanatory video, a slide deck outlining the project scope, and a one-page summary. Agency representatives circulated these materials internally to ensure confidentiality and organizational compliance. Participants completed a screening survey capturing gender, department, role, video-investigation experience, and years of service. This trust-based recruitment approach yielded five active-duty participants spanning patrol, dispatch, investigation, and supervisory roles, with 7–22 years of experience (see Table 1). All five participants completed both Phase 1 and Phase 2.

PID	Rank	Agency	Years Exp.	Primary Focus Areas
P1	Lieutenant	Arlington County Police Department	22	Real-time crime, canine, SWAT, drones, civil disturbance
P2	Captain	City of Fairfax Police Department	18	Patrol, criminal investigation, community service
P3	Detective	Manassas City Police Department	8	Post-incident investigation, drones
P4	Detective	George Mason University Police Department	7	Post-incident investigation
P5	Sergeant	Virginia State Police Department	15	Real-time crime, post-incident investigation, drones

Table 1: Study 1 Participant Profiles.

3.2 Phase 1. Events of Interest (EoIs)

Phase 1 addressed RQ1 through an online survey designed to identify EoIs relevant to MRVS multi-robot video public safety scenarios and the collection of contextual resources.

3.2.1 Methodology. We constructed an online EoI survey instrument (see supplement) and distributed it to five public-safety professionals to obtain a practitioner-grounded set of EoIs (Table 2). Our EoI derivation process consisted of three steps: (1) preparing the set of potential EoIs, (2) gathering public-safety professionals' feedback on their validity and priority, and (3) analyzing the results.

Step 1. EoIs Preparation: We first constructed a candidate set of EoIs balancing operational realism, practical relevance, and reduced participant burden through a three-step process: (1) *Data collection*: We assembled an initial pool of candidate records from two sources: (a) 13,234 daily crime logs collected over three years across three universities in one U.S. states and (b) ten public research video datasets on anomaly detection and crime video analysis [1, 74, 80, 96, 97, 99, 103, 134, 144, 158]. For the daily crime logs, we extracted the existing fields *Nature of Case*, *Brief Description*, and *Offenses*. For the video datasets, we collected available *event types* and *descriptions*; when unavailable, two authors independently reviewed videos, drafted descriptions, and resolved differences through consensus coding [113] to complete our initial record archive. (2) *Filtering records*: Using descriptions from both sources, we removed records that (a) could not be visually detected from robot-captured footage or (b) would be impractical to simulate (e.g., sexual offenses). (3) *Classification*: With input from our public-safety co-author, we reviewed all remaining records and defined event types aligned with U.S. public-safety standards. Category naming drew from the Clery Act [14], a legal standard for campus-crime classification, and event types used in existing abnormal-event video datasets. To categorize the 40 candidate events of interest, five authors conducted an in-person collaborative affinity diagramming session [77]. Authors iteratively clustered EoIs and developed high-level themes through synchronous discussion, informed by a coauthor's public safety domain expertise. Each EoI was assigned only after unanimous consensus, followed by a secondary review ensuring themes were mutually exclusive and differentiated. This produced six categories: Property and Environmental Incidents, Public Order Disturbances, Vehicle and Mobility Incidents, Suspicious or Unusual Behavior, High-Risk Threats, and Miscellaneous Activities. Finally, taxonomy was refined with concise definitions and representative visual cues derived during data collection, with example images sourced from public materials to ground the survey.

Step 2. EoIs Survey: With this survey instrument prepared, we invited public safety professionals to complete the online EoI survey via email. Participants first viewed the six category definitions with example events and representative visual cues, then identified any missing event types within each category in open text to refine the taxonomy. For each of the 40 candidate EoIs, participants answered three questions: (1) "Is this event relevant to public safety?" (binary), (2) "How important do you consider this event for public safety?" (7-point Likert), and (3) "How urgent do you consider this event for public safety?" (7-point Likert). These questions were designed to assess each EoI's relevance to our problem space and to capture perceived urgency and importance to establish EoI priority.

Step 3. EoIs Analysis: For each candidate EoI, we calculated the proportion of officers who judged it as relevant, the mean (M) and standard deviation (Std) of both urgency and importance ratings. Analysis proceeded in three steps: (1) excluding events that were either unanimously deemed irrelevant or judged relevant by only one officer with low urgency and importance ratings ($M < 2$, $Std < 0.5$); (2) applying k-means clustering to group the remaining events by their mean urgency and importance ratings with k=4 selected via an elbow analysis [123]; and (3) adding events from open-ended responses, defining priorities through Phase 2 expert interviews.

3.2.2 Results. The Phase 1 survey produced a final taxonomy of 38 EoI types (see Table 2), grouped into four operational levels: *emergency, urgent, moderate, advisory*. Eight candidate events were excluded: six unanimously judged irrelevant and two judged relevant by only one officer with consistently low urgency and importance ratings (e.g., *Climbing fence*). Six additional events emerged from open-ended responses, including *Incident exposure* and *Propping open doors*, capturing context-specific concerns beyond predefined survey EoIs. K-means clustering revealed a clear boundary between response-critical events and routine observations: *Emergency* events showed the strongest agreement, with the highest mean ratings with relatively low dispersion ($M_{Imp} \approx 5.84$; $Std_{Imp} \approx \mathbf{1.64}$; $M_{Urg} \approx 6.27$; $Std_{Urg} \approx \mathbf{1.79}$), indicating consensus on immediate threats, whereas disagreement concentrated in the *Moderate* presenting the largest variance across both dimensions ($Std_{Imp} = \mathbf{2.39}$; $Std_{Urg} = \mathbf{2.88}$), suggesting triage sensitivity to situational context. Based on these, police co-authors proposed the final four-level taxonomy to reflect operational prioritization. This four-level structure served as the foundation for Phase 2, where expert interviews successfully validated these clustering levels and enriched them with additional metadata (e.g., entity categories).

3.3 Phase 2. Design Requirements (DRs)

Phase 2 addressed RQ2 through in-person semi-structured interviews [62] with the same five participants. The interviews produced six DRs for designing tools that enhance public safety professionals' current video sensemaking practices by enabling the use of multi-robot video sensemaking capabilities.

3.3.1 Methodology. Each interview followed a semi-structured format designed to ensure consistency and elicit participants' perspectives on current practices and potential system implications. A prepared slide deck guided the discussions across three themes. The first revisited Phase 1 results, prompting reflections on the EoIs survey, event types deemed irrelevant, and types of professionals believed ground-robot footage should capture more extensively. The second focused on video investigation practices, addressing current technology use, challenges faced, desired functionalities, and visual cues professionals rely on to identify people or objects. The third examined system implications, exploring concerns about adopting MRVS-like capabilities, anticipated benefits, and suggestions for future development. Each sessions lasted 38–65 minutes, were audio-recorded with consent, automatically transcribed, manually reviewed, and corrected by the lead author for accuracy.

We analyzed transcripts using Braun and Clarke's thematic analysis framework [9], informed by Saldaña's coding methodology [106].

No.	Event	Priority	Occur.	Label	No.	Event	Priority	Occur.	Label
1	Arson	Emergency	3	Crime	20	Trespassing	Urgent	0	Crime
2	Burglary	Emergency	0	Crime	21	Drunkenness	Urgent	10	Civil
3	Robbery	Emergency	17	Crime	22	Snatching Bag	Urgent	4	Civil
4	Assault	Emergency	13	Crime	23	Bag Left Behind	Urgent	1	Civil
5	Shooting	Emergency	1	Crime	24	Indecent Exposure	Urgent	0	Civil
6	Explosion	Emergency	0	Crime	25	Unattended Domestic Animals	Urgent	3	Civil
7	Kidnapping	Emergency	4	Crime	26	Medical Emergencies	Urgent	2	Civil
8	Weapon Holding	Emergency	4	Crime	27	Illegal Parking	Moderate	3	Civil
9	Destruction/Damage/ Vandalism	Urgent	9	Crime	28	People Falling	Moderate	11	Civil
10	Theft from Vehicle	Urgent	6	Crime	29	Person Smoking	Moderate	3	Civil
11	Theft from Building	Urgent	2	Crime	30	Prohibited U-turns	Moderate	2	Civil
12	Motor Vehicle Theft	Urgent	2	Crime	31	Jaywalking	Moderate	8	Civil
13	Abuse	Urgent	2	Crime	32	Cars Stopping on Road	Moderate	3	Civil
14	Brawling	Urgent	13	Crime	33	Harassment/Stalking	Moderate	3	Crime
15	Crowds Escaping	Urgent	0	Civil	34	Loitering	Advisory	2	Civil
16	Obstructing Justice	Urgent	0	Crime	35	Crowd Gathering	Advisory	1	Civil
17	Carrying Suspicious Object	Urgent	5	Civil	36	Wrong-way Driving	Advisory	2	Civil
18	Hit and Run	Urgent	7	Crime	37	Wearing Face Mask	Advisory	3	Civil
19	Road Accidents	Urgent	6	Civil	38	Propping Doors Open	Advisory	1	Civil

Table 2: Events Classification with Occurrence Count in the testbed environment

The lead and fourth authors independently conducted open coding across all transcripts. To ensure analytical rigor, we assessed inter-rater reliability on the pre-reconciliation codes, achieving substantial agreement (Cohen's κ = 0.72). The authors then met to discuss discrepancies and iteratively refined and consolidated codes into higher-level themes, producing a finalized codebook. These themes were synthesized into six DRs specifying how MRVS should support public safety video work.

3.3.2 Results. We developed seven interrelated themes illustrating how public safety professionals engage with multiple videos in practice (T1-T7). Across themes (T1-T6) surface tensions between frontline needs and existing technological support. We present each theme alongside its corresponding design requirement (DR1–DR6) and one additional insightful design consideration.

T1. ***Context-Sensitive Recipes for Event Detection.*** While detecting an event type such as a parking violation may appear straightforward, the "recipe" for defining and justifying an event is highly context-dependent, hinging not only on the immediate situation but also on agency policies, spatial constraints, and operational norms. For example, a parked vehicle might be a delivery or a crime signal in preparation (P2). Likewise, gatherings acceptable in public may be treated as suspicious loitering on private property (P4). Professionals emphasized that EoIs shift with organizational capacity and priorities; the same event may draw attention in one agency but not another, depending on staffing and coverage (P1-P3). EoIs are also distinguished between real-time alerts and post-incident review (N=3/5) : for instance, *"Skateboarding isn't alert-worthy, but it matters later"* (P3), and *"Parking needs no response, but we want the data"* (P5). These reflections highlight that event significance is inseparable from institutional and temporal context. Accordingly, professionals expect video sensemaking systems to move beyond simple flagging by explaining why behavior is abnormal (P2, P3) and simplifying cognitively heavy taxonomies into clearer groupings such as *"vehicle, people, others"* (P4), for clearer criminal-civil distinctions, adopting to finalize our EoIs results (details see Table 2).
DR1. Context-Aware EoIs Explanation. MRVS should support flexible, contextual EoIs tailored to agency needs, with transparent AI reasoning and explanations.

T2. ***The Grind of Single-Video Analysis.*** Investigating a single video remains one of the most cognitively exhausting and time-consuming tasks in public safety work, revealing the need for more effective ways to navigate and interpret footage. Public safety professionals described spending hours manually scanning through footage to locate mere seconds of relevant content, typically fast-forwarding at 8x or 16x speed (N=4/5). This requires sustained focus and risks missing subtle cues, especially in low-light or low-resolution footage. As P3 put it, *"I could be looking for 15 seconds of video in 3 or 4 hours of footage… I can't blink, or I'll miss it"*. Agencies sometimes delegate this labor to junior staff to preserve the senior staff's time (P1), yet chronic understaffing still leads to overload and leaves lower-urgency cases backlogged (P3, P5). Technical flaws further undermine trust: misaligned timestamps and interfaces lacking frame-accurate scrubbing make review difficult

and not useful in their workflow (N=3/5). As P4 remarked, *"I've seen videos say 2 AM when it's clearly daylight—how can we use that in court?"* Participants expressed cautious interest in AI tools, stressing that decision-making must remain human-controlled. They sought support that surfaces key moments with a transparent rationale. As P2 emphasized, *"Quick detection is not enough. I need to know why it thinks this is evidence."* Concerns include unreliable outputs creating additional verification work and over-reliance on unexplainable results that could undermine legal admissibility (P1, P4). Professionals valued the small and accurate assistance over automation, which may simply shift verification burdens onto them (N=5/5).
DR2. Automated Video Detection. MRVS should enable efficient review through automated abnormal event summaries, AI reasoning, and timeline navigation to reduce manual effort.

T3. ***The Juggle of Multi-Video Sensemaking.*** Making sense of multiple video streams compounds the difficulty of situational awareness, as professionals must monitor several feeds simultaneously while extracting and connecting dispersed events into a coherent picture. While additional cameras provided wider coverage and multiple angles, participants described the experience of juggling them as cognitively overwhelming (N=4/5). They struggled to decide which screen deserved attention, often missing important activity elsewhere, and found it difficult to piece together a continuous storyline across feeds in real time (P2, P5). As P3 explained, *"I can pull four videos to watch at once, but I can't fully make sense of each one"*. The fragmented video sources came from different platforms, forcing professionals to switch between systems and disrupting workflow (N=2/5). As P1 noted, *"We have someone monitoring calls, pulling up whatever camera they can, but it's all disconnected. You have to jump system to system, with no single place showing everything that matters"*. Participants envisioned tools that could relieve this "juggling" by moving beyond raw video display toward an integrated, event-centric overview. Similar to a police report, which has a high-level overview and recorded anomalies with rank urgency. As P3 summarized, *"Help me see everything at a glance. Flag what's abnormal, show me what's important first"*.
DR3. Situational Overview. MRVS should offer a unified overview across video streams for event comparison, urgency prioritization, and anomaly detection to support rapid decisions.

T4. ***When Space and Time Collide.*** Video investigation is hindered by the collision of space and time, as professionals must determine which sources captured a given location and whether the footage is still accessible (P3, P4). This tension between spatial coverage and temporal accessibility creates a substantial burden, forcing them to manually reconstruct whether useful video exists. Current infrastructures are fragmented: traffic cameras offer real-time monitoring but often *"don't record"* (P4); private business footage, though critical, can take days or weeks if owners are unavailable (P2). Even when accessible, fixed cameras leave blind spots in mid-blocks, trails, and rural roads, making it unclear whether an incident was ever captured (N=3/5). Professionals emphasized that the first step in post-incident work is simply confirming whether useful video exists, a process that requires manually scanning across six to seven disconnected systems—body-worn, in-car, drone, traffic, and private feeds—all with separate logins and inconsistent retention. This verification process is time-consuming and cognitively demanding, often complicated by vague reporting, such as *"a business was broken into overnight (P2)."* Participants envisioned that instead of piecing together coverage maps, an integrated platform could synchronize all video sources onto a unified map and timeline views. Such a system would help professionals to quickly identify where and when coverage is available, close geographic gaps, and streamline the reconstruction of movement timelines.
DR4. Unified Spatio-Temporal Awareness. MRVS should synchronize fragmented feeds in an interface combining map and timeline views for seamless verification and reconstruction.

T5. ***Attribute Without a Place to Search.*** Even when professionals know which attributes of a person or vehicle they want to locate, current systems can not operationalize those descriptors, forcing long manual video scanning. Rather than depending on faces or license plates, they look for relatively stable attributes such as hair, clothing, pants, shoes, or vehicle features like model, make, color, and visible damage (N=5/5). As P1 emphasized, *"People change jackets easily even after minor crime, but they don't usually change shoes."* Similarly, P3 noted, *"Nine times out of ten, I find someone by their pants or shoes, not their face."* Yet these clues are often vague or incomplete. A witness may only recall a red hoodie, leading to hours of scanning across feeds. Vehicle tracking presents comparable challenges: plates are easily obscured or swapped, making color, model, and make more reliable identifiers (P2). Despite their importance, current video systems cannot link descriptors across sources, leaving professionals with labor-intensive searches (N=3/5). Participants emphasized the need for descriptor-driven search that connects attributes across footage and provides transparent explanations of which features were matched and why. Such functionality would improve efficiency and align with policy constraints and public concerns, limiting facial recognition (N=3/5).
DR5. Descriptor-Based Search. MRVS should support search for people and vehicles using appearance attributes to locate entities across video data quickly.

T6. ***Gaps in Collaborative Sensemaking.*** Teams struggle to coordinate their findings and share interpretations across video analysis tasks, underscoring the need for better collaborative support. Currently, though video investigations are highly collaborative, often spanning multiple teams, shifts, and extended periods, there is little support for persisting progress, sharing annotations, or enabling seamless handovers between public safety professionals. Critical updates were typically shared informally through texts, emails, or verbal briefings, which were prone to being lost, misunderstood, or overlooked (N=2/5). P3 noted *"We just email clips around and hope the next person watches the right part."* This lack of continuity resulted in duplicate work, missed leads, and inconsistent conclusions, particularly during long investigations or shift changes. The professionals expressed frustration at the need to manually track timestamps, notes, and clip links, which then had to be transferred into emails or printable formats to share with the team (P4). Such fragmented workflows created unnecessary barriers in situations where speed, accuracy, and team coordination were critical. These findings highlight the need for systems facilitating team-based workflows by supporting persistent annotations, shared workspaces, and cross-shift progress tracking, ensuring the

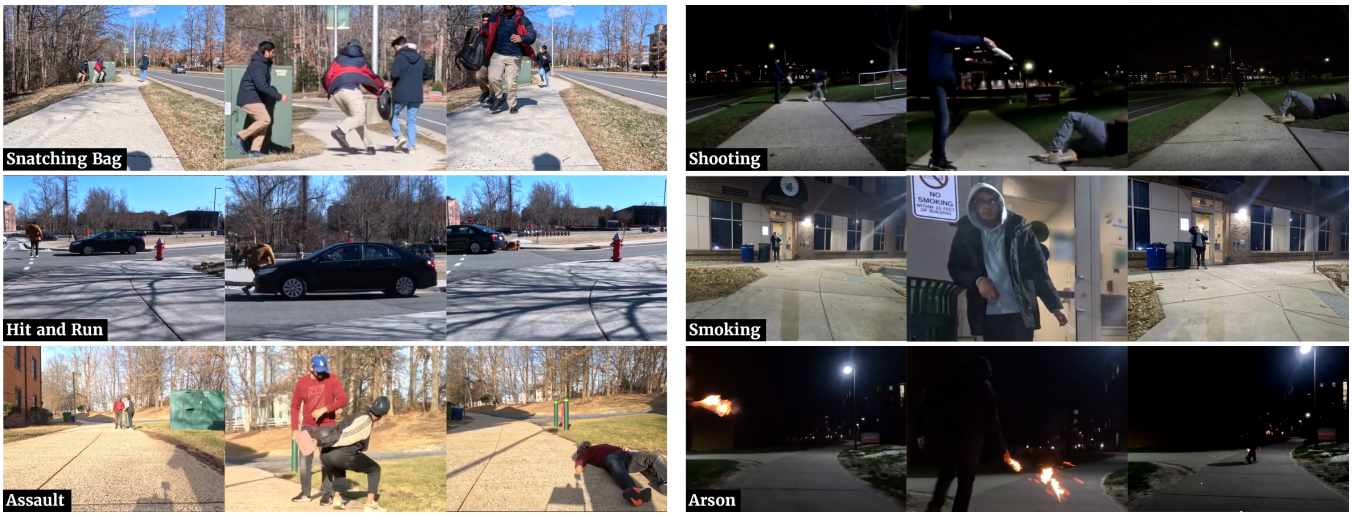

Figure 2: Examples of different anomalies in our testbed shown in sequences. Each second column is manually zoomed in.

continuity and collective sensemaking throughout investigations.
DR6. Team Collaboration Minded. Building on T6, MRVS should facilitate collaborative annotation, validation, and event management to support workflows and continuity across shifts.

T7. ***Socio-Technical Preconditions for Patrol Robot Scenario.*** Professionals recognized patrol robots' potential as force multipliers for police work, noting their deterrent effect similar to parking marked cars in high-crime areas. However, they emphasized that successful deployment requires more than just technical readiness. Key concerns included public acceptance and privacy (N=4/5). Participants stressed the need for transparency, clear operational boundaries, and context-appropriate privacy protections with robots minimizing identifiable capture during routine patrol while allowing richer footage during authorized investigations (P2). Several noted that while police-styled robots could enhance deterrence *"Knowing the robot is patrolling here will disperse some people. You could claim that as a win (P4)"* , preventing them from becoming attack targets in reality needs to be carefully considered. Practical deployment challenges centered on cost sustainability, hardware robustness, and reliable performance in rough terrain. P1 noted: *"Budget's a huge thing and changes every year—you have to make cuts."*
Design Consideration: Ensure patrol robot deployment respects community norms and institutional capacity through transparency logs, operational boundaries, and durable infrastructure that balances effectiveness with public trust.

4 Video Dataset

We present a first-of-its-kind video dataset[2] captured by a patrolling ground robot to establish a testbed environment for multi-robot video sensemaking. Unlike existing anomaly video datasets, ours is grounded in co-defined EoIs with public safety professionals, enacted in real-world environments, and recorded under police-guided patrol scripts specifying locations and event types. This dataset offers a realistic benchmark for evaluating future systems leveraging robot-captured video in public safety. Representative scripted EoIs are shown in Fig. 2.

Grounded in the refined EoI types in Table 3.3.2, we captured 156 scripted events across 10 campus zones and 10 day/night sessions with 22 student actors, resulting in 20 non-stop patrol videos around 30 minutes each, containing 833 people and 1,537 vehicles. For video capture, we utilized a Frodo Zero ground robot [32] equipped with a GoPro Hero 11 [38], recording 4K-resolution videos without audio during video taking. Each video contains GPS coordinates and timestamps. Our video production process is as follows.

Step 1. Preparation Before the main video production, we prepared four elements to prioritize ecological validity, ensuring footage appeared natural, realistic, and reusable for future studies. Specifically, we focused on defining plausible patrol routes, creating naturalistic acting guidelines, conducting dry-run pilots, and securing public safety professional oversight. First, robot patrol routes were designed to cover campus "hot spots" identified by the eighth author with public-safety domain expertise, remain fully wheel-accessible, and avoid spatial overlap. Second, the acting scripts specified naturalistic behavior, such as maintaining a normal walking pace, avoiding exaggerated glances toward the robot, and performing each EoI at varying distances, locations, and occlusion levels to reflect real-world visual complexity. We balanced the distribution of EoI types and ensured that each EoI was represented at least once in the dataset. Third, we conducted dry-run recordings using two scripted routes and seven actors, reviewing two 30-minute pilot videos to tune robot speed, route length, and ensure actor behaviors appeared natural. Finally, the eighth author confirmed these behaviors aligned with actual campus incidents, leading us to refine both patrol routes and acting scripts before final production. We note that some abnormal events cannot be naturally captured, so we followed prior work [23, 76] using actor-performed scenarios to ensure systematic coverage of rare and critical events.

Step 2. Video Capture Before each capture, actors rehearsed assigned locations and were briefed on where, when, and how to perform each EoI naturally within everyday campus settings. To

[2] Dataset is available at: https://huggingface.co/datasets/Puqi7/MRVS_anomaly_long_video_dataset

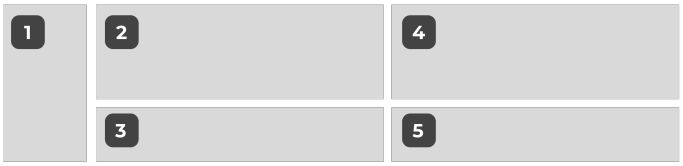

Figure 3: MRVS interface layout and corresponding design requirements.

synchronize events with the robot's continuously moving field of view, a coordinator followed the robot (out of frame) and provided silent temporal cues, ensuring that events unfolded as the robot arrived without actors breaking character or visibly "waiting". All recordings took place in an active, unannounced campus environment: while the *foreground* events were scripted, the *background* included uncontrolled pedestrian traffic, vehicles, lighting variation, and other ambient activities to preserve ecological validity. After collecting the videos, we annotated each recording with ground-truth labels, including EoI types, timestamps, and event durations.

To simulate a scenario in which ten videos are captured simultaneously, we designed spatially non-overlapping patrol routes, each assigned a spatiotemporal index (route ID and timestamp ID) corresponding to an individual video. We collected ten patrol sessions using the same robot across these routes over ten days under comparable weather conditions and at the same time of day, thereby emulating simultaneous multi-robot coverage of the environment; all recordings were completed within one month to maintain consistent seasonal lighting and foliage. For each route, we recorded a daytime and nighttime pair of videos. To avoid unrealistic identity collisions, actors changed outfits across sessions, and scripts were designed to prevent impossible location jumps or the same individual appearing simultaneously in different videos. Finally, three authors and four volunteers reviewed all videos and scripts to confirm the absence of cross-video identity conflicts and to ensure that appearances across days remained visually distinguishable.

Step 3. Post Production: Anonymization and Distribution All identifiable faces and license plates were blurred using ORB-HD Deface [91] for facial anonymization and EgoBlur [98], which is based on Faster R-CNN [102], for license-plate anonymization. This produces a research-safe foundation for evaluating timeline-based filtering, real-time alerting, descriptor-based search, and event detection. The dataset contains video only, with no audio, to avoid potential First Amendment concerns [31]. This research-safe multi-robot video dataset includes (a) anonymized video files (with all faces and license plates blurred) and (b) metadata describing EoI types, timestamps, and corresponding GPS information.

5 MRVS

We built MRVS to faithfully embody the DRs identified in S1. Developing such a tool posed the two key challenges raised consistently across S1 themes and echoed in prior work. First, providing scalable real-time insights from video investigations requires a strong backend capable of balancing speed and reliability [51, 152]. In S1, professionals described the "grind" of manually reviewing massive, dispersed footage under time pressure (T2, T3). Unreliable automated results force additional verification work, compounding operator burnout [41, 92]. Second, delivering these insights to public safety professionals demands a simple and trustworthy flow of information; the system must minimize false alarms yet surface all noteworthy cases so that no critical issue is overlooked [61, 118]. In high-stakes scenarios, professionals stressed the need to understand why context-aware events are flagged (T1, T4), as false alarms, and opaque reasoning erode trust in safety-critical alerting systems [26, 65, 85]. Addressing these challenges required balancing two tensions: building an advanced backend generating all possible cases, while presenting them to avoid overwhelming users. Our novelty lies in balancing the two demands: advancing video understanding models to be fast, accurate, and trustworthy, and structuring their output into clear, actionable insights through the front-end. To this end, we introduce the backend pipeline leveraging *multimodal LLM* prompt-engineered with the EoI taxonomy and professionals' domain-specific analysis protocol, and a front-end that presents noteworthy events selectively in a *structured format* of (1) event type, (2) explanation, (3) rationale, and (4) confidence level. Together, these components enable public safety professionals to build situational awareness across both area and time. Embedding the DRs into our system, we organized them into four recurring user workflows: **F1.** Browsing and investigating detected EoIs for situational awareness (DR1–DR3), **F2.** Reasoning across time and space (DR4), **F3.** Locating objects of interest through targeted search (DR5), and **F4.** Collaborating with team members (DR6).

5.1 Frontend: Interactive Sensemaking and Collaborative Investigation

In this section, we explain how features in MRVS are designed, built, and embedded to support the four flows and design requirements. Our interface dashboard use five divisions (D1~D5) as shown in Fig. 3. In describing the four flows, we will explain how we utilized each division using the features.

F1. Browsing and investigating detected EoIs (DR1–DR3). MRVS supports browsing through *Video Debrief* (D2) and *Situational Overview* (D4). **Video Debrief** (Fig. 4, left) structures detected EoIs within a single video as chapter-like chunks, each with title, description, rationale, confidence score, and a representative frame. A prioritized summary and timeline navigation bar at the top allows quick access, while clicking the representative frame synchronizes all components to the event's start time. **Situational Overview** (Fig. 4, right) aggregates EoIs across multiple videos, showing events as cards with robot ID, time, and priority. Cards can be filtered, played, marked, or synchronized with the global timeline. These components reflect DR1-DR3, reducing manual video review and enhancing situational awareness by surfacing structured, contextual information from the back-end.

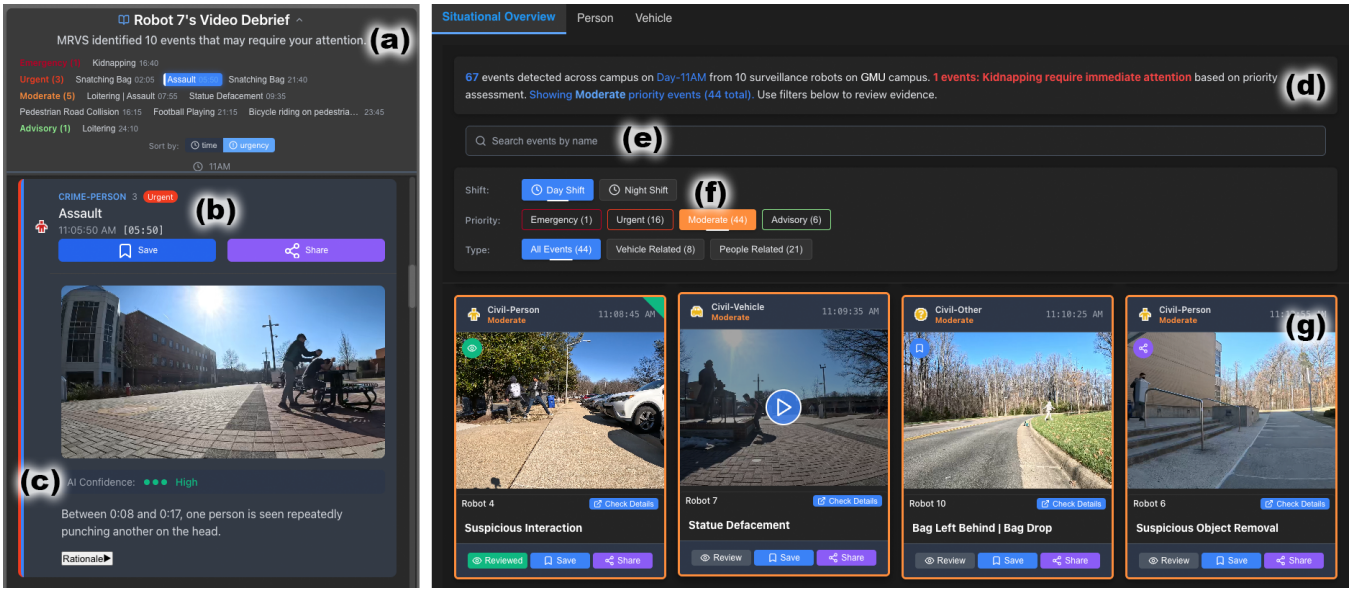

Figure 4: F1.Browsing and investigating detected EoIs for situational awareness. Left: Professionals begin with a robot-level video debrief *(a)*, where detected events are grouped by priority and sorted by time/urgency. Selecting an event opens an inspectable card *(b)* with triage actions (save/share), and a representative keyframe; with model confidence and rationale *(c)*. Right: The situational overview summarized events across robots *(d)*, supports keyword search *(e)*, and filtering by shift, priority, and event type *(f)*. Filtered events are presented as cards for rapid scanning *(g)* containing mark reviewed, save, or share items, quickly viewed event video segment, and clicking "Check Details" links a card to deeper inspection on the card.

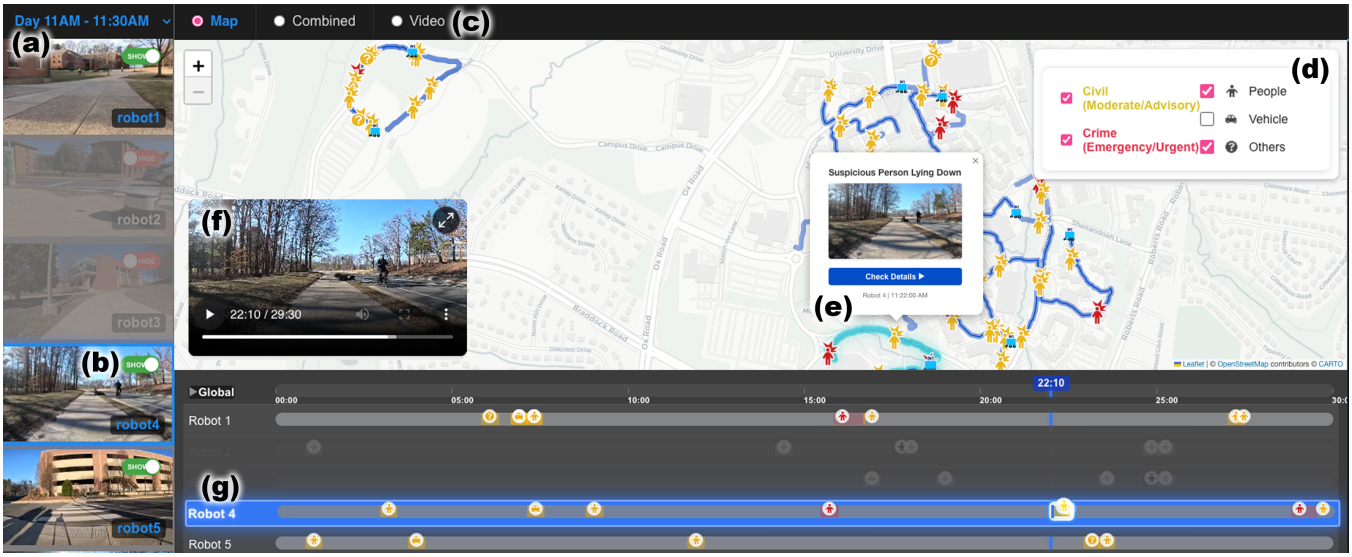

Figure 5: F2. Reasoning across time and space. Professionals adjust the day/night window time for videos *(a)* and browse multiple robots via the video list, with toggles to show/hide videos linked to timeline and trajectory *(b)*. Three layout options display map, video debrief, and video *(c)*. The legend supports event type and entity filtering *(d)*. Robot trajectories can be selected, with icons linking to keyframe popups and a "Check Details" for deeper inspection *(e)*. Users can quickly preview corresponding video segments *(f)*. A global timeline serves as a shared temporal reference *(g)*, while per-robot timelines with synchronized playheads enable aligned cross-robot comparison and time-jump navigation for spatiotemporal reasoning.

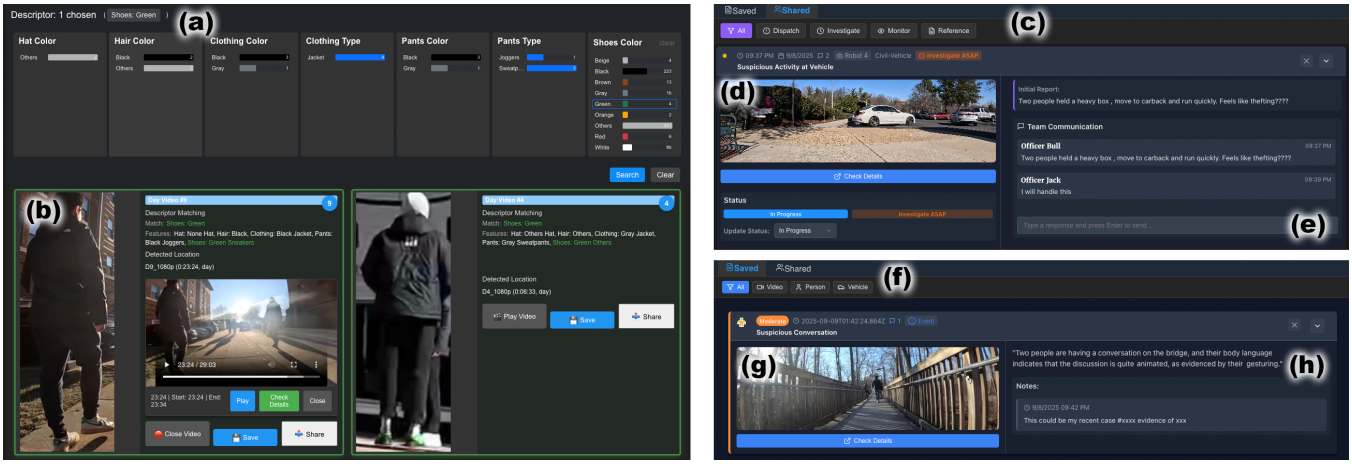

Figure 6: Left F3. Locating objects of interest through targeted search. Professionals build queries using appearance descriptors *(a)*; results appear as ranked cards with keyframes, matched attributes, and clip-level evidence for verification *(b)*. Right F4. Collaborating with team members (DR6). Events can be shared to a team workspace *(c)*, where items become trackable work cards with status/priority and evidence entry points *(d)*. Team members coordinate via a shared communication thread *(e)*. Saved events in personal workspace are organized by entity-level filter *(f)*, displaying as detailed note cards with save time and evidence keyframes with entry points *(g)* and model event descriptions with personal notes *(h)*.

F2. Reasoning across time and space (DR4). To support spatio-temporal reasoning, MRVS links all feeds through a synchronized *Video Timeline* (D3), *Map Navigation* (D2), and the *Video Wall*, together in Fig. 5. The **Video timeline** provides a global, color-coded view of EoI types and priorities, allowing professionals to scan activity, filter by category, and jump into synchronized playback. The map visualizes robot real-time updating trajectories with pinned EoIs locations; selecting a marker retrieves the corresponding video timestamp. Each EoI is represented by a simplified icon (person/vehicle/other) determined by its properties, positioned on the map using the robot's GPS at the time of the event. All three components are interconnected. D2 further supports three in-division layouts that differ in the location and size of the map navigator, video navigator, and video debrief for different analytical focuses. This reflects DR4 coupling and enables professionals to track how events emerge, propagate, and connect across both space and time.

F3. Locating objects of interest through targeted search (DR5). MRVS enables targeted search through a **Descriptor Searcher** (D4) see Fig. 6. Person and Vehicle appearance features, such as clothing, vehicle type, or color, are derived from back-end descriptors with visual shown. Search results are shown as cards with images, metadata, and match scores, each linked to the corresponding video clip. This reflects DR5 in filtering and comparison functions that help professionals narrow results and trace objects across multiple videos, reducing time spent on manual scanning.

F4. Collaborating with team members (DR6). Collaboration is supported through a **Group Workspace** (D5) see Fig. 6, where professionals can save, annotate, and share EoI cards/chapters or search results. Each entry keeps its contextual details: event type, time, and robot ID, so that teams can review evidence consistently. Professionals can leave notes, validate detections, and organize cases for follow-up, enabling shared situational awareness and coordinated decision-making for DR6.

5.2 Back-end: MLLM-based Video Sensemaking and Descriptor Searching

The back-end of MRVS converts testbed multi-robot footage into structured, reviewable information about EoIs and object visual attributes, and outperformed the state-of-the-art traditional computer vision anomaly detection model, HolmesVAD [153] and LLM Gemini [119]. By combining multi-object detection and tracking with multimodal LLM reasoning, our back-end model could detect and classify EoIs, describe objects, generate confidence-scored event types with descriptions and rationales, extract key preview thumbnails, and generate stable appearance-based descriptors such as clothing or vehicle type, providing the foundation for front-end video debrief, situational awareness, and descriptor-based search.

5.2.1 MLLM-based Video Sensemaking (DR1-DR3). To help professionals quickly grasp long videos, the MRVS back-end employs a *"surveillance persona"* LLM incorporating the EoIs taxonomy from S1 and follows domain-specific analysis protocols as an *"expert analyst"*. This persona reviews videos in short clips, reasons about key objects and activities, and generates structured event cards with event type, description, rationale, confidence level, and a representative frame. These outputs reduce professionals' cognitive load and remove manual browsing. With the confidence level and AI-generated rationale, professionals can also assess how the AI reached its conclusions and evaluate the reliability of its evidence.

Specifically, our approach builds upon the capabilities of Gemini 2.0 Flash [119] as a baseline based on its leading performance on established video understanding benchmarks [107, 136, 142] and enhances it with multi-modal reasoning and spatial-temporal localization. The model was configured with a temperature of 1.0 to balance response creativity with output consistency and reliability. We guide attention to objects of interest with point prompting [20, 21]. Specifically, we preprocess the video streams using the BoxMOT [11] framework, which integrates real-time multi-object

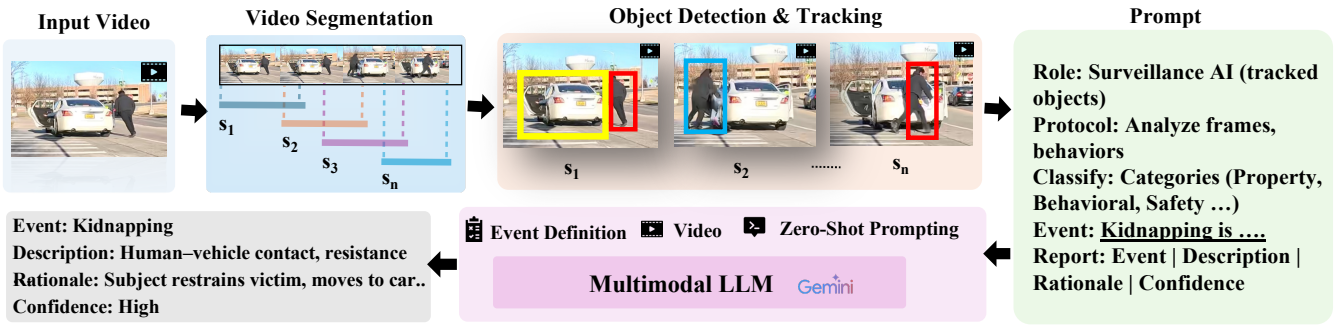

Figure 7: MRVS backend EoIs detection process, support front-end Video Debrief & Situational Overview.

tracking with YOLO [131] based detectors. Each frame is enriched with bounding boxes $\mathcal{B} = \{b_i = (x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i})\}_{i=1}^{N}$ and class labels $C = \{c_i\}_{i=1}^{N}$ for detected entities. These annotated frames form an object-aware video stream $\mathcal{V}^*$, which is then processed by a multimodal large language model (MLLM) using both **role prompting** and **temporal segmentation**. Inspired by [59], we introduce a system-level persona prompt, defining the model as a specialized surveillance agent (VSAI-9000) with domain-specific analysis protocols. This role-based formulation enhances zero-shot reasoning and structured event classification. The input video is temporally segmented into overlapping intervals $\mathcal{S} = \{s_j\}_{j=1}^{n}$ (e.g., 10 seconds each), where consecutive segments share partial temporal overlap to ensure continuous coverage of key activities and smooth transition detection. The 10 second interval duration was determined through pilot testing with 10 videos, balancing sufficient temporal context for meaningful event identification against computational efficiency and the model's context window limitations. For each interval, the pipeline (Fig. 7) returns the following: the event type $e \in \mathcal{E}$, a natural language description $\mathcal{D}(e)$, an explanatory rationale $\mathcal{R}(e)$, a confidence level $l(e) \in \{high, medium, low\}$, and representative frame(s) $\mathcal{K}(e)$ selected via keyframe extraction. A keyframe is selected as the frame with the highest count of distinct objects. This frame serves as the clearest snapshot of the interval. The model is instructed to state its confidence as one of three discrete levels, *high*, *medium*, or *low*, based on how certain it is about the inferred event type and description. This confidence level is passed unchanged to the front-end (e.g. shown as "High" in Fig. 3) and used as an ordinal indicator of reliability. This design follows recent work showing that large language models can express reasonably calibrated uncertainty in natural language without exposing logits or explicit probability scores [56, 66].

5.2.2 Descriptor-based Search (DR5). To support targeted investigation, the system back-end provides a *"descriptor-based search"* function allowing professionals to locate specific people or vehicles from partial descriptions. It extracts stable visual attributes, such as clothing color or vehicle type and color, paired with cropped object images. This enables professionals to quickly filter and verify candidates without relying on memory or manual review.

Objects are extracted from videos using BoxMOT [11], combining YOLO-based object detectors [101] with DeepOCSort [81] for multi-object tracking. LLAVA 1.6 [68] verifies detection semantic correctness through binary questions, discarding negatives. For cars and persons, descriptor-based attributes are extracted through structured queries, with cropped first-appearance images stored for validation. Predicted attributes are re-evaluated on saved crops, and rejected outputs are placed in the *"other"* category.

Descriptor-based search matches the selected car or person against all objects sharing similar descriptor values. We extract descriptors with LLAVA 1.6 [68], which chooses from predefined options or answers yes/no. For non-binary descriptors, we include *"other"* and *"unclear"* so low-confidence cases avoid overly specific labels. We then run a second yes/no round (converting option questions to binary) to filter out low-confidence answers. This removes low-confidence answers and improves robustness by reducing the impact of weak predictions on search accuracy.

To address missed tracking, image similarity checks with a contrastive model [125] compare same-class objects sharing identical descriptors (shirt and pants color for people; body color for vehicles). Trajectories are merged when similarity surpasses a high threshold (0.95), with final descriptor assignments resolved through majority voting. The system preserves the detection crop exhibiting the maximum bounding box area per trajectory.

6 Study 2: Summative Study

In this summative study (S2), we conducted two evaluations assessing MRVS's practical utility for back-end technical performance and public safety professionals' video sensemaking tasks. In algorithmic performance evaluation, we compared two state-of-the-art anomaly detection models as baselines against our improved LLM-enhanced module. We conducted an expert review with 9 public safety professionals at their respective offices to evaluate how the system's frontend supports real-world investigative workflows.

6.1 Algorithm Evaluation

6.1.1 Method. To emulate the sliding-window logic of an online detector, we divide every video into 30 s segments with a 5 s overlap; therefore, a new segment starts every 25 s. We denote each segment by s_i where $i = 1, \ldots, N$ and N is the total number of segments across all videos. For each segment s_i the proposed model generates a text prompt that is matched against a predefined *incident taxonomy* comprising eight high-level categories (*Vehicle & Mobility*, *Public-Order Disturbance*, *High-Risk Threat*, *Suspicious Behaviour*, etc.). Each category contains several fine-grained events, e.g. *Hit-and-Run*, *Loitering*, or *Shooting*. The pipeline outputs at most one

Period	# Videos	Duration (min)	Normal	Abnormal	Method	Precision	Recall	F1
Day	10	250	400	218	HolmesVAD [153]	0.223	0.008	0.016
					Gemini 2.0 [119]	**0.769**	0.351	0.469
					Ours	0.505	**0.534**	**0.497**
Night	10	220	404	161	HolmesVAD [153]	0.074	0.012	0.021
					Gemini 2.0 [119]	**0.623**	0.368	0.438
					Ours	0.440	**0.792**	**0.540**
Overall	20	470	804	379	HolmesVAD [153]	0.015	0.011	0.018
					Gemini 2.0 [119]	**0.696**	0.359	0.453
					Ours	0.473	**0.663**	**0.519**

Table 3: Detection performance on 20 videos (470 min). Best scores per metric (per period) are in bold.

incident label per segment. Three trained annotators independently labeled every segment. A segment was marked *abnormal* only if at least two annotators agreed that some event from the taxonomy was visible; otherwise it was marked *normal*. Comparing the model prediction with the ground truth yields four mutually exclusive outcomes per segment: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

The four counts above form the confusion matrix. Because abnormal events are relatively rare compared to normal segments (Table 3), overall accuracy is misleading. For example, a system that always predicts "normal" would achieve high accuracy but zero utility for anomaly detection. We therefore report three standard evaluation metrics used in information retrieval, classification, and video anomaly detection under class imbalance [22, 83, 116]:

- Precision (P), defined as $TP/(TP + FP)$, measures the proportion of segments predicted as abnormal that are truly abnormal. High precision reduces false alarms and unnecessary verification effort.
- Recall (R), defined as $TP/(TP + FN)$, measures the proportion of truly abnormal segments that are successfully detected. High recall is critical in safety-critical surveillance, where missed incidents can have serious consequences.
- F1 Score, defined as the harmonic mean of P and R, $2PR/(P + R)$, summarizes the trade-off between precision and recall into a single number and penalizes systems that perform well on only one of the two [22, 83]. In video anomaly detection, where positives are rare and both false alarms and missed detections matter, F1 is widely used as a primary comparison metric [116].

To benchmark our method, we compare against two baselines: (1) HolmesVAD [153], a recent state-of-the-art model specifically designed for explainable video anomaly detection, and (2) Gemini 2.0 [119], a powerful multimodal large language model, which we adapt as a strong baseline for anomaly detection by leveraging its multimodal reasoning capabilities. This setup enables a meaningful comparison between specialized video models and the emerging class of general-purpose LLMs.

6.1.2 Results. Table 3 summarizes the detection performance across 20 videos (470 minutes), broken down by time of day (Day and Night). For each method, we report precision, recall, and F1 score, with the best value per metric highlighted in bold. Our method's ability to detect nearly twice as many anomalies results in substantially better overall detection performance and practical utility. It consistently outperforms both HolmesVAD, a state-of-the-art video anomaly detection model, and Gemini 2.0, a strong multimodal LLM baseline we adapted for this task. During the day, our method achieves the highest recall (0.534) and F1 score (0.497), successfully balancing the precision-recall tradeoff—unlike Gemini 2.0, which favors precision (0.769), suffers from much lower recall (0.351), ultimately missing a significant number of anomalies. At night, the superiority of our approach becomes even more evident, achieving the best precision (0.440), recall (0.792), and F1 score (0.540). In contrast, both baselines deteriorate sharply, especially HolmesVAD, which drops to a recall of just 0.012. Overall, our method achieves the highest F1 score (0.519) and recall (0.663), confirming its robustness across diverse lighting conditions. These results validate our design choices and demonstrate that a task-specialized approach significantly outperforms both specialized and general-purpose multimodal baselines in real-world anomaly detection scenarios.

The lower precision of our method compared to Gemini 2.0 (e.g. 0.505 vs. 0.769 during the day and 0.440 vs. 0.623 at night) reflects an intentional design trade-off. We configured MRVS to more liberal in proposing candidate anomalies so as to minimize false negatives in our safety-critical setting. Essentially, extra false positives increase screening workload for public safety professionals where missed incidents correspond to unobserved threats. In practice, MRVS is intended to be used as a decision-support tool in which flagged segments are reviewed and validated by officers. In this setting, the substantial gains in recall and F1 are more important than maximizing precision alone.

6.2 Expert Review

The goal of expert review was to address our research questions regarding professionals' perceptions, use experiences, and reflections on the adoption of MRVS in their investigative workflows. Specifically, this study aimed to explore:

- **RQ1**: How might the adoption of the MRVS system transform current police agency practices, including expected benefits, risks, and its influence on team collaboration?
- **RQ2**: How do professionals perceive the functionality of each system component, including AI-generated debriefs and descriptor-based object tracking?
- **RQ3**: What are the considerations, challenges, and design opportunities for integrating MRVS-like systems into future public safety workflows?

6.2.1 Participants and Recruitment. We recruited nine experienced public safety professionals (P1–P9) from five local police departments, including patrol officers, detectives, captains, and real-time crime center officers. All participants had at least 5 years of field experience (see Table 4) and were familiar with video investigation in their daily work, but had limited exposure to AI-augmented video analysis systems. Participants were recruited through institutional

partnerships and volunteered to participate without compensation.

6.2.2 Procedure. We designed a structured evaluation procedure combining task-based expert use of MRVS with post-task interviews to assess its utility in realistic policing scenarios. Sessions were conducted in person with nine public safety professionals at their offices, using our monitor for consistency. After a brief introduction and demonstration of MRVS 's core components: multi-robot video playback with timelines and maps, LLM-powered debrief and situational overview, clickable color-coded icons, descriptor-based search, and a collaborative workspace—participants were invited to freely explore the system to gain familiarity. They were then asked to complete three investigative tasks simulating realistic police scenarios, followed by **semi-structured interviews** examining how MRVS compared with current workflows (RQ1), perceptions of specific system components (RQ2), and broader views on MRVS's role in future policing (RQ3). Sessions lasted between **75 and 120 minutes**. Given the open-ended, scenario-based task design, participants used MRVS's features in varying combinations and sequences. Consequently, we report findings as a holistic system evaluation rather than a comparative analysis of features across tasks.

- **Task 1**: Urgent Incident Response and Verification. Locate and verify video evidence of an ongoing fight reported on a campus bridge, saving key findings with notes for team coordination.
- **Task 2**: Routine Patrol Review and Event Validation. Review AI-detected events from a single time shift, validate significant events versus false alarms, and share verified cases for further action (15-minute session).
- **Task 3**: Descriptor-based Search and Suspect/Vehicle Identification. Use witness descriptions to search, locate, and verify suspects or vehicles, sharing matches to support investigations.

6.2.3 Expert Interview Analysis. We analyzed the expert interviews using a data-driven thematic analysis approach [87, 106]. All sessions were audio recorded, transcribed independently open-coded by two trained authors using the Glaser method, with short memos capturing contextual details [37]. To validate coding quality, we assessed inter-rater reliability on the pre-reconciliation codes, achieving substantial agreement (Cohen's $\kappa = 0.78$). They then compared codes and iteratively clustered them into higher-level categories via constant comparison [106], resolving ambiguities by revisiting transcripts and, when needed, consulting the corresponding author. Together, the two coders and the corresponding author diagrammed relationships among categories and refined them into themes organized around our three research questions (RQ1-RQ3).

6.2.4 Front-end evaluation results.

RQ1: MRVS Overview *Empowers Police Workflows and Strengthens Team Collaboration.* Participants overwhelmingly described MRVS as essential support for the cognitively demanding work of video sensemaking, both in real-time dispatch and post-incident investigations. Professionals emphasized that analyzing fragmented, multi-source footage, whether real-time or post-incident, was among the most exhausting and resource-intensive aspects of their work, often requiring hours of scanning and attention to subtle cues (N=7/9). MRVS was seen as a *force multiplier* that eased these burdens, helping them prioritize attention, surface relevant segments, and focus on verification and decision-making amid staff shortages (N=7/9). Professionals noted that by minimizing time spent on exhaustive searches, the system freed them to focus on higher-value interpretive tasks critical to effective investigations. Beyond supporting individual sensemaking, MRVS also addressed longstanding challenges in coordinating video investigations across shifts, teams, and timeframes. Professionals explained that cases often unfolded over extended periods, requiring asynchronous handovers and collaborative sensemaking (N=3/9). Existing tools offered little support for this, leading to duplicated efforts, missed leads, or inconsistent interpretations (N=5/9). The system's collaborative workspace, including tagging, annotations, action items, and shared status updates, enabled teams to continue work seamlessly across shifts and off-days, supporting situational continuity and maintaining accountability without relying solely on informal or verbal handoffs.

However, these perceived efficiencies came with clear expectations for procedural integrity and accountability (N=3/9). Participants emphasized that AI-driven features, such as suspect identification and abnormal events detection, were valuable aids for narrowing focus during investigations but could not displace human judgment (N=5/9). Professionals further advocated for safeguards to ensure legitimate use (N=4/9), such as requiring case numbers before searches (P1), enforcing strict access controls (P2), and conducting regular audits (P4, P6), emphasizing aligning the system with broader legal and institutional protocols. These accounts highlight how professionals negotiated the system's role as a support for existing video sensemaking practices, embracing its ability to streamline workflows and strengthen team coordination, while insisting such systems operate within the procedural, legal, and professional boundaries central to public safety work. Participants described MRVS as *"easy to learn (P1, P3-P8)"* and highly intuitive, all professionals completed the three evaluation tasks without assistance. They highlighted the natural interactive flow *"I don't need to go back to the video, just click what I'm investigating on the timeline, map, or search, and the video jumps there (P4)"* and fit to their workflow *"Everything I need is right there, easy to find the alert, check it, and share through right near clicks (P7)."* Meanwhile, professionals (N=5/9) recommend splitting the current interface into two coordinated views to better support smaller monitors.

RQ2: LLM-based Video Debriefs *Support Rapid Evidence Review while Reinforcing Public Safety Professionals Control and Trust.* Participants widely viewed AI-generated video debriefs as valuable accelerators for routine evidence review, particularly under high-pressure, resource-constrained conditions. Professionals appreciated features such as direct video jumps (N=5/9) and AI-curated summaries (N=7/9) for reducing cognitive load and enabling faster triage, while emphasizing these tools must operate within user-controlled workflows to maintain trust and accountability (N=5/9). As P6 noted, *"Even when it's wrong, it saves time because I can jump right to the clip and watch the footage."* AI-generated confidence scores were especially valued for guiding attention and helping professionals calibrate their review effort: *"If it's low confidence, I know I need to double-check everything carefully; if it's high, I can rely more on the detection and just validate (P1)"*. Participants also welcomed transparent AI explanations embedded within the debriefs to help them assess system reasoning, even if imperfect (N=6/9).

PID	Rank	Agency	Years	Primary Focus Areas
P1	Detective	George Mason University Police Department	7	Post-incident investigation
P2	Detective	Manassas City Police Department	12	Post-incident investigation
P3	Detective	Manassas City Police Department	8	Post-incident investigation, drones
P4	Sergeant	Virginia State Police Department	16	Real-time crime, post-incident investigation, drones
P5	Detective	City of Fairfax Police Department	5	Post-incident investigation, dispatcher, patrol
P6	Detective	City of Fairfax Police Department	25	Post-incident investigation
P7	Captain	City of Fairfax Police Department	18	Patrol, criminal investigation, community service
P8	Captain	Fairfax County Police Department	16	Real-time crime, community service
P9	Sergeant	George Mason University Police Department	13	Patrol, real-time crime

Table 4: Study 2 Participant Profiles.

However, professionals cautioned that AI should remain a support tool, never replacing human judgment: *"AI can't testify in court; we still have to make the final call (P3)"*. Several participants raised legal concerns (N=4/9) about over-reliance on AI-curated evidence potentially introducing risks in court proceedings (*P3*).

Participants framed MRVS as a user-controlled system where AI accelerates routine tasks and enhances situational awareness, while ensuring critical judgments, prioritization, and accountability firmly remain with the user. Some professionals further recommended the system allow setting custom alerts for high-risk incidents like assaults, with human oversight prioritized even in low-confidence detections (*P4, P7*). Although MRVS sometimes surfaced incorrect or missed events, all participants felt it substantially reduced workload versus exhaustive manual review. The metadata-rich cards (keyframes, icons, event types, confidence rationales; Fig. 4, left) let them dismiss many false alarms "at a glance" (N=6/9): *"I realized this was wrong just look the image (P3)"*. For missed events, officers valued that the detected events, together with the synchronized map–timeline view, provided a spatial–temporal scaffold narrowed manual checks to specific time windows or locations *"even when it misses one, I know exactly where to start(P6)."* Beyond reducing workload, participants noted recurring failure modes in the underlying detectors: many missed events were *small, brief, poorly lit, or near frame edges*, especially from ground-robot viewpoints where important behaviors occurred in the periphery rather than center (*P3, P6, P9*). These patterns led officers to treat MRVS alerts as useful starting points rather than exhaustive coverage, reinforcing the need for continued human review of subtle or peripheral incidents.

RQ2: Descriptor-Based Search *is Valuable for Targeted Investigation.* Public safety professionals described the descriptor-based search feature as a powerful capability for narrowing suspects and focusing attention on relevant entities. Participants valued AI's support in filtering vehicles and persons, recognizing that even imperfect narrowing saved significant effort (N=5/9): *"Even if it just helps me rule out a few wrong ones, that's a huge help (P1)"*. Professionals accepted that AI is better at broad categories (e.g., car type, color) than fine-grained details (e.g., brand) (N=4/9), and recommended providing flexible filtering mechanisms, such as the ability to exclude certain features rather than only include them (P3, P5). Beyond people and vehicles, participants emphasized the need to search for and track specific objects such as backpacks, bags, or items left behind across hours of footage (N=2/9).

Participants also stressed the need for progressive filtering workflows, starting from broad descriptors—like light vs. dark vehicles—before narrowing to specific details, reflecting the often vague information provided in real-world situations (*P4*). However, they consistently emphasized that human interpretation remains essential for complex identifications and ambiguous cases (N=7/9): *"You still have to eyeball it yourself when things look alike. The system can't make those calls for you (P3)"*. To support situational awareness, professionals recommended integrating search results directly into the map, allowing them to visualize suspects or vehicles over space and time: *"If every time I do a search, the matches show up on the map, I can start building the suspect's trail right away (P7)"*. While professionals acknowledged trade-offs between AI accuracy and workload reduction, they stressed the system must avoid excessive wrong detection (N=3/9), with low-confidence results always requiring manual verification to preserve procedural rigor (*P4*).

RQ2: Situational Overview *Enhances Situational Awareness and Professionals Safety.* Participants consistently described the Situational Overview as a valuable feature for enhancing situational awareness and officer safety by enabling rapid comprehension of complex scenes. Professionals appreciated the card-style key images and synchronized timeline and map functions, which allowed them to decide when to skim, when to validate, and when to investigate in depth (N=5/9). This flexibility supported their preference for workflows where they *"quickly scan and choose what to focus on"* rather than following linear review processes.

Participants valued the system's filtering and marking capabilities (N=4/9), empowering them to adjust the interface to investigative priorities and felt MRVS provided *"a lot of options to sort and manage"* while keeping them in control (*P9*). Professionals emphasized that AI-assisted situational awareness should remain transparent and user-led, ensuring they retained discretion over how to explore, interpret, and act on information. Professionals suggested future enhancements, such as question-based video querying, to streamline information retrieval without sacrificing context or control (N=2/9). Overall, professionals framed the Situational Overview as a user-controlled, AI-assisted tool reducing cognitive load while supporting proactive and reactive policing tasks.

RQ2: Collaboration Workspace *Supporting Flexible Police Practices.* Participants described current collaboration practices as fragmented and inefficient, relying on calls, emails, and paper trails that often led to information loss and uncertainty. They saw MRVS as a valuable opportunity to centralize and streamline coordination through a shared workspace, particularly praising the envisioned status update function as critical for cross-role and cross-agency collaboration (*P1–P8*). Professionals highlighted the benefits of packaging case materials for easy sharing (*P3*), but stressed the need for police-centered collaboration workflows (N=5/9), including tiered permissions(*P2, P5*), secure sharing with external partners(*P1, P4*), and safeguards to ensure accountability(*P3*).

They recommended features such as video sharing links (*P1, P5*), downloadable reports (*P2*), and case folders integrating all related events and clips (*P4, P5*), ensuring supervisors could monitor progress without disrupting investigators. Importantly, professionals cautioned that AI-generated events should first undergo human review to avoid overwhelming frontline professionals (*P9*). They emphasized balancing awareness with workload by tailoring notification settings by role (N=4/9). Across these suggestions, participants underscored MRVS must support controlled, case-centered collaboration while minimizing friction and protecting rigor.

RQ2: Seamless Spatial and Temporal Information. Participants emphasized the importance of tightly integrated spatial and temporal information to support rapid situational awareness and investigation workflows (P1-P4, P8). While they appreciated MRVS's current map and timeline features, they identified key areas for improvement to better align with policing practices and decision-making under time pressure. They emphasized that visual icons and color-coding provide a clear overview of location and time without even seeing any text (N=6/9).

Participants highlighted the need for clearer geographical context, recommending that robot identifiers use location-based names rather than numerical codes to help locate video sources (N=4/9). They also proposed enhancing map interactivity by enabling direct selection of areas of interest, clickable patrol routes for instant timeline synchronization (P1-P3, P6-P8), and automatic highlighting of related events and movements when selecting a suspect or vehicle in descriptor-based searches (N=3/9). Overall, participants urged that future designs prioritize intuitive, police-centered interactions that seamlessly connect spatial and temporal data for efficient incident response.

RQ3: Envisioned Personalized MRVS *Supporting Flexible Police Practices.* Professionals consistently envisioned MRVS not as a one-size-fits-all tool, but as a modular, adaptable workspace that aligns with their situational needs and investigative styles. They emphasized the system should mirror their workflows, enabling adjustments to robot labels, event priorities, alert thresholds, and interface layouts (N=6/9). For example, patrol robots should be named by location (e.g., "North Plaza") to support situational awareness, and EoIs should be user-adjustable plug and play module each agency could tailor to its policies, staffing levels, and local context (e.g., time of day, crowd density) (N=2/9). Participants also highlighted that collaboration features should reduce reporting effort between colleagues and supervisors by making it easy to see who is working on which case and how far each has progressed (N=3/9).

Professionals emphasized retaining agency and adaptive control over AI-generated alerts, preferring user-defined filters or role-based templates to avoid overload, especially in frontline tasks. Investigators, supervisors, and real-time crime center officers were seen as needing different default views, filters, and notification templates for which alerts appear, how they are grouped, and how much detail is shown (N=2/9). They envisioned a minimalist, customizable interface where components like timeline or descriptor search could be rearranged, only expanded as needed. Many preferred distributing the interface across multiple screens to separate functions (e.g., real-time alerts, communications, post-incident review), aligning with existing dual-monitor setups (N=6/9). This vision reflects the expectation MRVS should adapt to professionals' practices, supporting flexibility, transparency, and discretion, while preserving autonomy and accountability in public safety.

RQ3: Envisioned Deployment and Human Robot Interaction (HRI) *Preventive, Hot-Spot-Oriented, and Specialized Use Cases.* Professionals described MRVS-like systems as primarily preventive resources, highlighting value in detecting emerging risks and discouraging undesirable behavior rather than only reconstructing incidents (N=3/9). They pointed to use cases such as flagging deteriorating infrastructure and repeatedly frequented locations where people *"hang around for no good reason"*, noting that visible patrol robots could help disperse groups while complementing traditional patrol work. Participants emphasized that deployment should follow empirical risk patterns rather than uniform coverage (N=4/9). They cited nightlife districts, dimly lit alleys, and known hot spots during peak hours as priority zones, *"I could see a bunch of robots in Old Town on a Friday night (P2)"*, and identified missing-person searches as a high-stakes scenario where MRVS could reduce time-to-locate and officer exposure. Across these contexts, they articulated a spectrum of desired HRI modes from low-level steering (precise locations or camera angles) to high-level directives such as *"search this area for any anomalies (P7)"*, including multi-robot coordination and integration with fixed cameras and drones (N=6/9).

7 Discussion

We reflect on our core contributions in relation to prior work, present implications for designing future MRVS-like systems, and discuss limitations.

7.1 Reflection in Relation to Prior Work

In this section, we situate our contributions within existing research, highlighting how our testbed and system advance prior work across HCI, public-safety, and computer vision. We reflect on how both the testbed and the MRVS system were informed by existing research and how they can serve as a foundation for future MRVS studies.

7.1.1 Testbed environment. Core components in our testbed environment are (1) professional-validated EoI types, Design Requirements of MRVS-like systems, and (3) a ground-robot video dataset.

- **The EoIs Types** (see Table 2) were informed by prior anomaly detection video datasets that define "anomaly" from a computer vision perspective as statistical rarity [82], researcher-defined categories [99], or labeled from online surveillance footage [144], valuable for benchmarking but not directly actionable in public safety practice. We drew EoIs from real campus crime logs (following Clery Act [14]) on the foundation of existing anomaly datasets [1, 75, 80, 96, 97, 99, 103, 134, 144, 158], then refined with frontline professionals to yield operationally meaningful, context-rich categories (e.g., "Incident Exposure"). This approach grounds "what to detect" in professionals' operational realities, translating their priorities into actionable EoIs for future system and model design. However, our EoIs remain campus-centric, reflecting a limited jurisdictional perspective and excluding non-visual or hard-to-simulate incidents. We therefore position it not as a universal ontology, but as a reusable approach: MRVS-like systems should expose EoIs as configurable, locally governed "recipes"

rather than fixed, one-size-fits-all classes. The EoI types serve as a reusable template that other jurisdictions can adapt, extend to additional modalities (e.g., audio, text), and use to drive MRVS-like systems and upstream workflows (e.g., dispatch [71, 88, 120]).

- **Design Requirements** derived from S1 were informed by HCI's formative–summative methodological approach, which emphasizes human-centered design for specialized professionals [33, 63, 147]. Our methodological choice shows that domain stakeholders must remain central in defining what systems attend to, why, and under what constraints [34, 35, 46]. The derived design requirements offer guidance for creating MRVS-like systems that incorporate one or more components of multi-robot operation, video analysis, and decision-making support under attention and time constraints. While we instantiated these requirements in particular ways within our implementation, final designs may differ depending on designers' insights and local operational constraints. These requirements also generalize to applications that share characteristics with our user groups, including other high-stakes experts and broader multi-video sensemaking scenarios.
- **The Video Dataset** presents ground robot-captured, EoIs-driven, actor-performed videos in campus environments. It complements existing anomaly video datasets which predominantly use short, event-oriented clips recorded by fixed surveillance cameras [159] or generated in virtual scenes [1, 116]. Following established practice in computer vision and video anomaly detection, our actor-performed simulation approximates rare, safety-critical events that are difficult or unethical to capture opportunistically, similar to canonical datasets that deliberately record scripted or simulated activities in real environments [8, 75, 86]. For example, BEHAVE [8] provides multi-person interaction videos with actors performing chasing and fighting scenarios, the UMN crowd-panic dataset [86] contains staged escape events, and the CUHK Avenue benchmark [75] records abnormal behaviors on a university campus. These actor-driven benchmarks have been used extensively to develop methods that generalize to more realistic "in-the-wild" surveillance data, including large-scale real-world anomaly datasets such as UCF-Crime [116]. More broadly, work in human action recognition and synthetic video generation (e.g., UCF101, PHAV) shows that models trained on actor-performed or procedurally generated videos improve performance on realistic benchmarks [23, 112], indicating that controlled yet diverse scenarios can yield transferable representations. Our dataset follows the same pattern: actor-scripted EoIs, co-designed with a police co-author, are embedded in an operational campus environment with uncontrolled bystanders, vehicles, lighting, and occlusions. This approach balances ecological validity with experimental control while ensuring coverage of the defined EoIs.

7.1.2 MRVS System Architecture. As a manifestation of the testbed environment, MRVS integrates reusable modules be utilized in future systems for multi-video investigation, ground-robot operations, EoIs detection and decision support, and workflows tailored to public safety professionals. Many anomaly detection systems treat model-centric outputs (e.g., anomaly scores [69, 116], clip labels [16, 78], or captions [17, 54]) as the primary product, leaving triage, evidence gathering, and justification to manual operator work. By contrast, MRVS treats these as intermediate signals and centers meta-rich artifacts that make reasoning visible: temporal localization, event descriptions, keyframes, and textual rationales that can be checked against the source footage. Even with imperfect back-end precision, these artifacts are designed to help operators move from "a possible event" to "a defensible judgment" with lower cognitive and coordination cost [28], aligning with high-stakes HCI arguments that accountability depends on legible evidence trails rather than end-to-end automation [12, 105]. MRVS situates the back-end outputs within an interactive front-end that supports the practical mechanics of professional sensemaking reasoning visible to the human decision-maker. While our current instantiation centers on ground-robot videos, the architecture is source-agnostic: it operates over time-stamped video with optional geospatial metadata, making the same interaction patterns immediately transferable to existing single- or multi-camera infrastructures CCTV, body-worn, in-car, or drone footage, which is a plausible near-term deployment pathway beyond robot-equipped agencies.

7.2 Implications for Designing Future MRVS-like Systems

Drawing on public safety professionals' reflections on MRVS in S1 and S2, we outline five design directions for future MRVS-like systems: adaptive workflow enhancement, specialized application deployment, stronger robotics and vision capabilities, and attention to societal considerations.

7.2.1 Enhancing MRVS Capabilities Within Public-Safety Scenarios. To advance MRVS within existing public-safety workflows, we suggest future systems should incorporate adaptive features that tailor prioritization, scheduling, and operational support to each agency's unique context and EoI needs. EoIs can vary substantially by jurisdiction agencies: an activity that is concerning in one city may be routine in another, and situational factors (e.g., crowd density, time of day, weather, and season) shape both incident likelihood and what officers consider noteworthy. EoIs also vary by the location as not distributed uniformly, as public-safety concerns cluster unevenly across space, with certain neighborhoods or recurring incidents concentrating risk. A practical direction is therefore to make "*plug-and-play*" EoIs configurable, enabling agencies to fine-tune priorities without requiring technical expertise. More broadly, an adaptive MRVS agent capable of contextualizing data in real time could yield better decision-making outcomes and more relevant alerts for officers in the field.

7.2.2 Beyond Patrol: Deploying MRVS Across Specialized Context. We envision that MRVS can be extended to a wide range of real-world scenarios beyond event detection, supporting specialized, high-stakes operations such as missing-person search, night patrol, and disaster recovery. These scenarios shift design goals from continuous monitoring with specific EoIs toward rapid coverage, time-to-locate, and resilience under degraded conditions (e.g., low light, debris, unstable connectivity). Designing for such deployments implies MRVS-like systems should be rugged, quickly deployable, and configurable to mission-specific EoIs, while offering tools for flexible path planning, on-the-fly reprioritization, and collaboration with human teams operating under stress and uncertainty.

7.2.3 Advancing Core Robotic Foundations for MRVS. We identify key robotics side implications for advancing MRVS-like systems, including providing flexible control, multi-robot coordination, and robust field-ready platforms to ensure operability and effectiveness. Multiple abstraction control levels from precise waypoints or camera angles to high-level directives [95] such as "search this area for anomalies" and supports "condition-triggered" handoffs between autonomous and manual modes for both individual robots and the fleet. Participants anticipated scaling to multi-robot collaboration, including integration with existing surveillance infrastructure and coordination with other robotic platforms such as drones. MRVS-like systems should dynamically allocate robotic coverage based on real-time and historical data about hot spot activity. MRVS-like systems should deliver high-quality video streams in different formats [72, 143, 151] to match task demands. MRVS-like deployment holds dual preventive value: persistent patrol enables earlier intervention through anticipatory evidence collection and deters criminal activity through visible presence. Such moving robotics systems could also help surface deteriorating infrastructure, recurrent hot spots, and patterns of suspicious loitering before they escalate. Yet sustained preventive deployments raise operational requirements: participants worried about vandalism in areas with heightened anti-police sentiment and about software-level tampering that could compromise system integrity. They therefore underscored tamper-resistant design, protective mechanisms for critical components, robust physical infrastructure, and minimizing ongoing human intervention during routine operations.

7.2.4 Advancing Computer Vision Foundations for MRVS. To better support public-safety workflows, we highlight computer-vision directions that foreground investigator semantics, quantify uncertainty, and address mobile ground-robot video constraints. *Descriptor-based search* should reason about attributes and object relations when doing retrieval, not just rank by pure visual similarity. Pure embedding retrieval (e.g. SigLIP-style [125]) is efficient, but brittle for our application, where dominant colors overwhelmed semantics in these settings, but MLLMs produced attribute descriptions that better matched officer intent. Similarly, existing methods do not quantify the confidence of matches well. New embedding techniques that encode richer attribute–object relations, capture uncertainty, and remain robust under illumination and low-light variation (as surfaced in S2) are needed. *Re-identification research* likewise needs to move beyond its current emphasis on people and vehicles [76, 133]. In practice, investigators often need to find specific objects across hours of heterogeneous footage, where identity cues can be mutable or context-dependent (e.g., a convertible with its top open versus closed) and sensing conditions vary widely. New re-ID capabilities must therefore prioritize investigator-relevant semantics, explicitly handle mutability versus invariance, and remain robust across viewpoint, illumination, and camera-quality shifts. One possible direction is to leverage MLLM's zero/few-shot capabilities to make embeddings robust with respect to these semantics.

Finally, *robust anomaly detection* for ground-robot video must address peripheral and brief events. In S2, many missed events were small, short, poorly lit, or near frame edges, suggesting that models trained on centered, sustained actions allocate insufficient attention to the periphery where ground-robot anomalies frequently occur. Beyond data augmentation, model intervention strategies such as attention intervention [15, 64, 154] and steering [114, 138] offer practical routes to increase sensitivity to subtle events. Additional extensions include active perception and tool usage (e.g., models choosing to zoom). To reduce false positives without overwhelming operators, agentic loops triaging low-confidence anomaly proposals can identify candidates for secondary verification and refinement.

7.2.5 Building Responsible and Trustworthy MRVS for Society. Drawing from our study, we suggest that long-term adoption of MRVS requires embedding ethical and accountable design into core functionality, particularly transparency, reliability, and operability in high-stakes contexts. Participants stressed that the operational benefits of MRVS-like systems are inseparable from governance. They valued preventive capabilities, yet warned that many EoIs (e.g., "hanging around" or "loitering") are socially and historically charged, and uncritical use risks reinforcing over-policing of marginalized groups [10, 129]. Scholarship on predictive policing and big-data surveillance likewise warns that concentrating sensing resources in "hot spots" can create feedback loops in which historically over-policed communities are subject to even more intensive monitoring, regardless of actual harm levels [10, 94]. These concerns suggest that MRVS-like systems must make what is monitored, where, and under what escalation criteria explicit and contestable, rather than treating deployment as a purely technical optimization problem.

Privacy and public acceptance emerged as context-dependent constraints on sustained deployment. Participants cautioned that robots may feel intrusive without a clear protective purpose, but are welcomed for targeted tasks like inspections or missing-person searches, and argued that privacy protections should vary by task. Beyond individual expectations, responsible deployment must address group-level harms, including who is most likely to be continuously observed, how long footage is retained, and whether deployment patterns reinforce existing inequalities [10, 139].

Our results highlighted a practical accountability tension: AI assistance is valuable under time pressure, yet officers remain responsible for decisions and must be able to justify actions taken (or not) in response to system alerts. To support accountable policing, MRVS-like tools should embody explainable and transparent AI by making recommendations easy to inspect, contest, and document. Interfaces should provide brief, human-readable rationales for why clips are flagged, plus fast controls to validate, dismiss, or override suggestions without re-watching long segments. At the deployment level, MRVS should make its monitoring logic, data retention policies, and limitations legible to agencies and communities, with safeguards against biased over-policing across groups.

7.3 Limitations and Future Work

We note several limitations that should be considered when interpreting this work. First, our testbed relies on a simulated multi-robot patrol setting and actor-performed EoIs rather than organically occurring events captured by deployed patrol robots. Although this actor-performed approach follows established practice in anomaly-detection datasets, it may under-represent edge cases, adversarial behavior, or longer-term behavioral adaptation to robot presence. Second, the robots in our study were teleoperated, and video was

pre-recorded, so we did not evaluate end-to-end autonomy, navigation failures, or how officers would share attention between live robot control and MRVS interfaces. Third, our participants were drawn from a limited set of agencies in one U.S. state and evaluated a single campus environment during one season, limiting the generalizability of our EoI taxonomy and design requirements. Fourth, our evaluation examined the integrated MRVS experience rather than isolating the causal impact of interface components; future work could employ controlled, feature-by-feature comparisons to understand specific feature effectiveness under different tasks.

Despite these limitations, the testbed environment and prototype offer a reusable scaffold for future MRVS research. Near-term deployments may layer MRVS-like interfaces on existing CCTV, in-car, body-worn, or drone footage in jurisdictions where patrol robots are not yet available. Future work should also expand EoIs and data collection to additional jurisdictions, seasons, and modalities (e.g., audio, text), and examine long-term adaptation as MRVS systems transition from lab to operational deployment.

8 Conclusion

In this work, we present MRVS, a human-AI collaborative system for multi-robot video sensemaking that addresses both practical and technical challenges of integrating ground robot footage into public safety workflows. By combining insights from public safety professionals with advances in computer vision and robotics, we contribute a real-world testbed environment and a novel system that demonstrates how AI-driven video analysis enables scalable situational awareness with ground robot footage. Our goal is to advance intellectual contributions for researchers in HCI, computer vision, and HRI who seek to develop impactful, socially grounded AI systems for public safety professionals. With continued research, this work also aims to provide practical and meaningful capabilities for public safety professionals pursuing more efficient, scalable, data-driven operations—and, ultimately, for citizens who benefit from safer, more resilient communities.

Acknowledgments

We are grateful to the Northern Virginia police departments for their support and domain guidance, especially Detective Cooper Knight, Major Bull Hudson, Lieutenant Mohammed Tabibi, Captain John O'hare, and Captain Andrew E. Hawkins, as well as other colleagues who provided feedback and guidance. We thank the George Mason University undergraduate students for their help with video dataset collection. This work was supported by the National Science Foundation (NSF2128867, NSF2350352), the Army Research Office (W911NF2320004, W911NF2520011), Google DeepMind (GDM), Clearpath Robotics, FrodoBots Lab, Raytheon Technologies (RTX), Tangenta, the Mason Innovation Exchange (MIX), and Walmart. We used generative AI tools for proofreading and editing under full human supervision, in line with the GAIDeT taxonomy (2025) [115]. The authors are solely responsible for the final manuscript and outcomes.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20143–20153.

[2] Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. 2025. AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM. *arXiv.org* (2025).

[3] Khulood Alkhudaidi, Tish Burke, Rachel Boll, Shruti Mahajan, Erin T Solovey, and Jeanne Reis. 2025. Perceptions and Preferences: Deaf ASL-Signing Users' Insights on Video Elements, Styles and Layouts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[4] Gowri Saini Balasubramaniam, Clara Belitz, and Anita Say Chan. 2024. Bridging Informational Divides: A Community-Centered Analysis of "Public Safety" Surveillance Technology. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[5] David H Bayley and Egon Bittner. 1984. Learning the skills of policing. *Law & Contemp. Probs.* 47 (1984), 35.

[6] Oliver Bendel. 2023. Robots in Policing. In *Social Robots in Social Institutions*. IOS Press, 135–144.

[7] Jack S Benton, James Evans, Miranda Mourby, Mark J Elliot, Jamie Anderson, J Aaron Hipp, and David P French. 2023. Using video cameras as a research tool in public spaces: addressing ethical and information governance challenges under data protection legislation. *Journal for the Measurement of Physical Behaviour* 6, 2 (2023), 145–155.

[8] Scott Blunsden and Robert Fisher. 2010. The BEHAVE Video Dataset: Ground Truthed Video for Multi-Person Behavior Classification. *Annals of the BMVA* 4 (2010), 1–12.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Sarah Brayne. 2017. Big data surveillance: The case of policing. *American sociological review* 82, 5 (2017), 977–1008.

[11] Mikel Broström. 2023. *BoxMOT: pluggable SOTA tracking modules for object detection, segmentation and pose estimation models*. doi:record/7629840

[12] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *In Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[13] Dan Calacci, Jeffrey J Shen, and Alex Pentland. 2022. The cop in your neighbor's doorbell: Amazon ring and the spread of participatory mass surveillance. *In Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–47.

[14] Clery Center. 2026. The Clery Act. https://www.clerycenter.org/the-clery-act Accessed: 2026-02-09.

[15] Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Hengtao Shen. 2025. Attention Hijackers: Detect and Disentangle Attention Hijacking in LVLMs for Hallucination Mitigation. *ArXiv* abs/2503.08216 (2025).

[16] Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024. Prompt-Enhanced Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18319–18329. doi:10.1109/CVPR52733.2024.01734

[17] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. 2023. TEVAD: Improved Video Anomaly Detection With Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 5549–5559.

[18] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. arXiv:2406.07476 [cs.CV] https://arxiv.org/abs/2406.07476

[19] Eric Corbett and Graham Dove. 2024. Signs of the Smart City: Exploring the Limits and Opportunities of Transparency. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.

[20] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, Fang Zeng, Dajiang Zhu, Wei Liu, Quanzheng Li, Lichao Sun, Shu Zhang Tianming Liu, and Xiang Li. 2024. SAMAug: Point Prompt Augmentation for Segment Anything Model. arXiv:2307.01187 [cs.CV] https://arxiv.org/abs/2307.01187

[21] Qiyuan Dai and Sibei Yang. 2024. Curriculum Point Prompting for Weakly-Supervised Referring Image Segmentation. arXiv:2404.11998 [cs.CV] https://arxiv.org/abs/2404.11998

[22] Jesse Davis and Mark Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. ACM, 233–240. doi:10.1145/1143844.1143874

[23] César Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López Peña. 2017. Procedural Generation of Videos to Train Deep Action Recognition Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2594–2604.

[24] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2019. Understanding law enforcement strategies and needs for combating human trafficking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[25] Mack DeGeurin. 2024. NYPD retires big, egg-shaped subway surveillance robot—for now. *Popular Science* (February 2024). After a nearly six-month long trial, the New York Police Department is ending its use of an eye-catching "K5" mobile surveillance robot..

[26] Stephen R Dixon and Christopher D Wickens. 2006. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors* 48, 3 (2006), 474–486.

[27] Daniel S Drew. 2021. Multi-agent systems for search and rescue applications. *Current Robotics Reports* 2 (2021), 189–200.

[28] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[29] Sidong Feng, Chunyang Chen, and Zhenchang Xing. 2023. Video2Action: Reducing human interactions in action annotation of app tutorial videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[30] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D Hollan. 2011. ChronoViz: a system for supporting navigation of time-coded data. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 299–304.

[31] Freedom Forum. 2026. Recording in Public. https://www.freedomforum.org/recording-in-public/. Accessed: 2025-09-07.

[32] Frodobots. 2026. Frodobots Online Store. https://shop.frodobots.com/. Accessed: 2026-02-09.

[33] Amrita Ganguly, Chuan Yan, John Joon Young Chung, Tong Steven Sun, Yoon Kiheon, Yotam Gingold, and Sungsoo Ray Hong. 2024. ShadowMagic: Designing Human-AI Collaborative Support for Comic Professionals' Shadowing. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[34] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2024. Going beyond xai: A systematic survey for explanation-guided learning. *Comput. Surveys* 56, 7 (2024), 1–39.

[35] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning eyes between humans and deep neural network through interactive attention alignment. *In Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[36] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, F. Khan, M. Popescu, and M. Shah. 2020. A Background-Agnostic Framework With Adversarial Training for Abnormal Event Detection in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[37] Barney Glasser. 1992. Basics of grounded theory analysis: Emergence vs. forcing. *Mill Valley, CA* (1992).

[38] GoPro. 2026. GoPro HERO11 Black. https://gopro.com/en/us/shop/cameras/hero11-black/CHDHX-111-master.html Accessed: 2026-02-09.

[39] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *In Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–25.

[40] Stuart W Hall, Amin Sakzad, and Kim-Kwang Raymond Choo. 2022. Explainable artificial intelligence for digital forensics. *Wiley Interdisciplinary Reviews: Forensic Science* 4, 2 (2022), e1434.

[41] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[42] MD Romael Haque, Katherine Weathington, and Shion Guha. 2019. Exploring the impact of (not) changing default settings in algorithmic crime mapping-a case study of milwaukee, wisconsin. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*. 206–210.

[43] Mahmudul Hasan, Jonghyun Choi, J. Neumann, A. Roy-Chowdhury, and L. Davis. 2016. Learning Temporal Regularity in Video Sequences. *Computer Vision and Pattern Recognition* (2016).

[44] Daojing He, Sammy Chan, and Mohsen Guizani. 2017. Drone-assisted public safety networks: The security aspect. *IEEE Communications Magazine* 55, 8 (2017), 218–223.

[45] Elize Herrewijnen, Meagan B Loerakker, Marloes Vredenborg, and Paweł W Woźniak. 2024. Requirements and attitudes towards explainable ai in law enforcement. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 995–1009.

[46] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *In Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[47] Jyothi Honnegowda, Komala Mallikarjunaiah, and Mallikarjunaswamy Srikantaswamy. 2024. An Efficient Abnormal Event Detection System in Video Surveillance Using Deep Learning-Based Reconfigurable Autoencoder. *Ingénierie des Systèmes d'Information* (2024).

[48] Emelia May Hughes, Renee Wang, Prerna Juneja, Tony W Li, Tanushree Mitra, and Amy X Zhang. 2024. Viblio: Introducing credibility signals and citations to video-sharing platforms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[49] Mina Huh, Ding Li, Kim Pimmel, Hijung Valentina Shin, Amy Pavel, and Mira Dontcheva. 2025. VideoDiff: Human-AI Video Co-Creation with Alternatives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[50] International Association of Chiefs of Police. 2013. *Reducing Officer Injuries Study Fact Sheet.* Fact Sheet NCJ 245196. Bureau of Justice Assistance (BJA), U.S. Department of Justice. https://www.ojp.gov/ncjrs/virtual-library/abstracts/reducing-officer-injuries-study-fact-sheet

[51] M Erdem Isenkul. 2025. Energy-aware deep learning for real-time video analysis through pruning, quantization, and hardware optimization. *Journal of Real-Time Image Processing* 22, 3 (2025), 125.

[52] Shomik Jain, D Calacci, and Ashia Wilson. 2024. As an AI Language Model," Yes I Would Recommend Calling the Police": Norm Inconsistency in LLM Decision-Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 624–633.

[53] Niall Jenkins. 2015. 245 million video surveillance cameras installed globally in 2014. *IHS Technology* (2015).

[54] Yalong Jiang and Liquan Mao. 2024. Vision-Language Models Assisted Unsupervised Video Anomaly Detection. arXiv:2409.14109 [cs.CV] https://arxiv.org/abs/2409.14109

[55] Qiao Jin, Yu Liu, Ruixuan Sun, Chen Chen, Puqi Zhou, Bo Han, Feng Qian, and Svetlana Yarosh. 2023. Collaborative online learning with vr video: Roles of collaborative tools and shared video control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[56] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel R. Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Joshua Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221* (2022).

[57] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. 2024. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. *In Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–24.

[58] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 4017–4026.

[59] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. In *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 4099–4113. doi:10.18653/v1/2024.naacl-long.228

[60] KPEL News. 2024. One Louisiana Community Is Now Deploying Robot Police Dogs. https://kpel965.com/lake-charles-robot-police-dog-louisiana/ Accessed: 2025-05-07.

[61] Sumeet Kumar, Hakan Erdogmus, Bob Iannucci, Martin Griss, and João Diogo Falcão. 2018. Rethinking the future of wireless emergency alerts: A comprehensive study of technical and conceptual improvements. *In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–33.

[62] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[63] Mackenzie Leake and Wilmot Li. 2024. ChunkyEdit: Text-first video interview editing via chunking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[64] Kenneth Li, Oam Patel, Fernanda Vi'egas, Hans-Rüdiger Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *ArXiv* abs/2306.03341 (2023). https://api.semanticscholar.org/CorpusID:259088877

[65] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[66] Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. *arXiv preprint arXiv:2205.14334* (2022).

[67] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. ConceptScape: Collaborative Concept Mapping for Video Learning. In *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173961

[68] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[69] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2017. Future Frame Prediction for Anomaly Detection - A New Baseline. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017).

[70] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. In *In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 14 pages. doi:10.1145/3526113.3545703

[71] Yiren Liu, Ryan Mayfield, and Yun Huang. 2023. Discovering the hidden facts of user-dispatcher interactions via text-based reporting systems for community safety. *In Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–31.

[72] Yu Liu, Puqi Zhou, Zejun Zhang, Anlan Zhang, Bo Han, Zhenhua Li, and Feng Qian. 2024. Muv2: scaling up multi-user mobile volumetric video streaming via content hybridization and sharing. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 327–341.

[73] Zirui Liu, Haichun Sun, and Deyu Yuan. 2025. Automatic analysis of alarm embedded with large language model in police robot. *Biomimetic Intelligence and Robotics* (2025), 100220.

[74] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal Event Detection at 150 FPS in Matlab.

[75] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2720–2727.

[76] Liping Lu, Zihao Fu, Duanfeng Chu, Wei Wang, and Bingrong Xu. 2025. CLIP-SENet: CLIP-based Semantic Enhancement Network for Vehicle Re-identification. *ArXiv* abs/2502.16815 (2025). https://api.semanticscholar.org/CorpusID:276575653

[77] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *IFIP conference on human-computer interaction*. Springer, 231–248.

[78] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. 2023. Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 8022–8031. doi:10.1109/CVPR52729.2023.00775

[79] Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. 2025. Sherlock: Towards Multi-scene Video Abnormal Event Extraction and Localization via a Global-local Spatial-sensitive LLM. *The Web Conference* (2025).

[80] Ke Ma, Michael Doescher, and Christopher Bodden. 2015. Anomaly detection in crowded scenes using dense trajectories. *University of Wisconsin-Madison* 2 (2015).

[81] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International conference on image processing (ICIP)*. IEEE, 3025–3029.

[82] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1975–1981. doi:10.1109/CVPR.2010.5539872

[83] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[84] Gabriela Marcu, Iris Lin, Brandon Williams, Lionel P Robert Jr, and Florian Schaub. 2023. " Would I Feel More Secure With a Robot?": Understanding Perceptions of Security Robots in Public Spaces. *In Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–34.

[85] Angela Mastrianni, Hua Cui, and Aleksandra Sarcevic. 2022. " Pop-Up Alerts are the Bane of My Existence": Designing Alerts for Cognitive Aids Used in Time-Critical Medical Settings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[86] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal Crowd Behavior Detection Using Social Force Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 935–942.

[87] Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*. Springer, 25–48.

[88] Scott Munro, Lucie Ollis, Carin Magnusson, Jill Maben, and Cath Taylor. 2025. Video livestreaming in emergency trauma dispatch: an observational study of technological integration with clinical decision-making in prehospital enhanced care services. *Scandinavian journal of trauma, resuscitation and emergency medicine* 33, 1 (2025), 108.

[89] C Muralidharan, V Arulalan, K Kishore Anthuvan Sahayaraj, et al. 2024. Enhanced Real-Time Abnormal Event Detection in Video Surveillance for Safety and Security. In *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*. IEEE, 1–5.

[90] Cuong Nguyen, Wu-chi Feng, and Feng Liu. 2016. Hotspot: Making computer vision more effective for human video surveillance. *Information visualization* 15, 4 (2016), 273–285.

[91] ORB-HD. 2025. deface. https://github.com/ORB-HD/deface

[92] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[93] Sangkeun Park, Emilia-Stefania Ilincai, Jeungmin Oh, Sujin Kwon, Rabeb Mizouni, and Uichin Lee. 2017. Facilitating pervasive community policing on the road with mobile roadwatch. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3538–3550.

[94] Tawana Petty, Mariella Saba, Tamika Lewis, Seeta Peña Gangadharan, and Virginia Eubanks. 2018. Reclaiming our data: interim report, Detroit. (2018).

[95] David Porfirio, Mark Roberts, and Laura M. Hiatt. 2025. Uncertainty Expression for Human-Robot Task Communication. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1698–1707.

[96] Mantini Pranav, Li Zhenggang, et al. 2020. A day on campus-an anomaly detection dataset for events in a single camera. In *Proceedings of the Asian Conference on Computer Vision*.

[97] Yuanbin Qian, Shuhan Ye, Chong Wang, Xiaojie Cai, Jiangbo Qian, and Jiafei Wu. 2025. UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6577–6585.

[98] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. 2023. EgoBlur: Responsible Innovation in Aria. arXiv:2308.13093 [cs.CV]

[99] Bharathkumar Ramachandra and Michael Jones. 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2569–2578.

[100] Brian A Reaves. 2015. Local police departments, 2013: Equipment and technology. *Washington, DC: Bureau of Justice Statistics* (2015).

[101] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[102] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV] https://arxiv.org/abs/1506.01497

[103] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. 2020. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2626–2634.

[104] Pankaj Raj Roy, Guillaume-Alexandre Bilodeau, and Lama Seoud. 2021. Local anomaly detection in videos using object-centric adversarial learning. In *International Conference on Pattern Recognition*. Springer, 219–234.

[105] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[106] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.

[107] Mohammadreza Salehi, Jae Sung Park, Tanush Yadav, Aditya Kusupati, Ranjay Krishna, Yejin Choi, Hannaneh Hajishirzi, and Ali Farhadi. 2024. ActionAtlas: A VideoQA Benchmark for Domain-specialized Action Recognition. arXiv:2410.05774 [cs.CV] https://arxiv.org/abs/2410.05774

[108] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[109] Hyorim Shin, Junho Choi, and Changhoon Oh. 2024. Delivering the Future: Understanding User Perceptions of Delivery Robots. *In Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–24.

[110] Frank Shipman, Andreas Girgensohn, and Lynn Wilcox. 2003. Generation of interactive multi-level video summaries. In *Proceedings of the eleventh ACM international conference on Multimedia*. 392–401.

[111] Ric Simmons. 2016. Quantifying criminal procedure: how to unlock the potential of big data in our criminal justice system. *Mich. St. L. Rev.* (2016), 947.

[112] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv preprint arXiv:1212.0402* (2012).

[113] Anselm L. Strauss. 1987. *Qualitative Analysis for Social Scientists*. Cambridge University Press, Cambridge, UK.

[114] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. *ArXiv* abs/2205.05124 (2022). https://api.semanticscholar.org/CorpusID:248693452

[115] Yana Suchikova, Natalia Tsybuliak, Jaime A Teixeira da Silva, and Serhii Nazarovets. 2025. GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Accountability in Research* (2025), 1–27.

[116] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6479–6488.

[117] Shengyang Sun and Xiaojin Gong. 2023. Long-Short Temporal Co-Teaching for Weakly Supervised Video Anomaly Detection. *IEEE International Conference on Multimedia and Expo* (2023).

[118] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. Alert fatigue in security operations centres: Research challenges and opportunities. *Comput. Surveys* 57, 9 (2025), 1–38.

[119] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[120] Ivanna S Terrell, Michael D McNeese, and Tyrone Jefferson Jr. 2004. Exploring cognitive work within a 911 dispatch center: Using complementary knowledge elicitation techniques. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 605–609.

[121] David Thacher. 2008. Research for the front lines. *Policing & society* 18, 1 (2008), 46–59.

[122] Mirko Thalmann, Alessandra S Souza, and Klaus Oberauer. 2019. How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45, 1 (2019), 37.

[123] Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (1953), 267–276. doi:10.1007/BF02289263

[124] Bao Tran Gia, Tuong Bui Cong Khanh, Khoa Tran Nhat, Kien Luu Trung, Thuyen Tran Doan, Khiem Le Tran Trong, Tien Do, and Thanh Duc Ngo. 2023. Integrating multiple models for effective video retrieval and multi-stage search. In *Proceedings of the 12th International Symposium on Information and Communication Technology*. 1003–1010.

[125] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).

[126] Joe Tullio, Elaine Huang, David Wheatley, Harry Zhang, Claudia Guerrero, and Amruta Tamdoo. 2010. Experience, adjustment, and engagement: the role of video in law enforcement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1505–1514.

[127] U.S. Bureau of Labor Statistics. 2014. *Police Officers: Injury, Illness, and Fatality Statistics, 2014.* Technical Report. U.S. Department of Labor. https://www.bls.gov/iif/factsheets/archive/fatal-occupational-injuries-police-officers-2014.htm Accessed: 2026-02-08.

[128] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[129] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[130] Nitya Verma and Lynn Dombrowski. 2018. Confronting social criticisms: Challenges when adopting data-driven policing strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[131] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. YOLOv10: Real-Time End-to-End Object Detection. arXiv:2405.14458 [cs.CV] https://arxiv.org/abs/2405.14458

[132] Bokun Wang and C. Yang. 2022. Video Anomaly Detection Based on Convolutional Recurrent AutoEncoder. *Italian National Conference on Sensors* (2022).

[133] Dong Wang, Qi Wang, Weidong Min, Di Gai, Qing Han, Longfei Li, and Yuhan Geng. 2024. SAM-driven MAE pre-training and background-aware meta-learning for unsupervised vehicle re-identification. *Comput. Vis. Media* 10 (2024), 771–789. https://api.semanticscholar.org/CorpusID:271979951

[134] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. 2020. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence* 43, 6 (2020), 2141–2149.

[135] Tian Wang, Meina Qiao, Jie Chen, Chuanyun Wang, Wenjia Zhang, and Hichem Snoussi. 2018. Abnormal global and local event detection in compressive sensing domain. *AIP Advances* 8, 5 (2018).

[136] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2025. LVBench: An Extreme Long Video Understanding Benchmark. arXiv:2406.08035 [cs.CV] https://arxiv.org/abs/2406.08035

[137] X. Wang, Zhengping Che, Ke Yang, Bo Jiang, Jian-Bo Tang, Jieping Ye, Jingyu Wang, and Q. Qi. 2020. Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[138] Xintong Wang, Jingheng Pan, Liang Ding, Longyue Wang, Longqin Jiang, Xingshan Li, and Christian Biemann. 2024. CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:276421262

[139] Cedric Deslandes Whitney, Teresa Naval, Elizabeth Quepons, Simrandeep Singh, Steven R Rick, and Lilly Irani. 2021. HCI tactics for politics from below: Meeting the challenges of smart cities. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[140] Morgan C Williams Jr, Nathan Weil, Elizabeth A Rasich, Jens Ludwig, Hye Chang, and Sophia Egrari. 2021. Body-worn cameras in policing: Benefits and costs. (2021).

[141] James J Willis and Stephen D Mastrofski. 2018. Improving policing by integrating craft and science: what can patrol officers teach us about good police work? *Policing and society* 28, 1 (2018), 27–44.

[142] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. arXiv:2407.15754 [cs.CV] https://arxiv.org/abs/2407.15754

[143] Nan Wu, Kaiyan Liu, Ruizhi Cheng, Bo Han, and Puqi Zhou. 2024. Theia: Gaze-driven and perception-aware volumetric content delivery for mixed reality headsets. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 70–84.

[144] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference*. Springer, 322–339.

[145] Jinyue Xia, Vikash Singh, David Wilson, and Celine Latulipe. 2014. Exploring the design space of multiple video interaction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. 276–285.

[146] Zhen Xu, Xiaoqian Zeng, Genlin Ji, and Bo Sheng. 2022. Improved Anomaly Detection in Surveillance Videos with Multiple Probabilistic Models Inference. *Intelligent Automation and Soft Computing* (2022).

[147] Chuan Yan, John Joon Young Chung, Yoon Kiheon, Yotam Gingold, Eytan Adar, and Sungsoo Ray Hong. 2022. FlatMagic: Improving flat colorization through AI-driven design for digital comic professionals. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–17.

[148] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. 2023. Video Event Restoration Based on Keyframes for Video Anomaly Detection. *Computer Vision and Pattern Recognition* (2023).

[149] Yaxing Yao, Huichuan Xia, Yun Huang, and Yang Wang. 2017. Free to fly in public spaces: Drone controllers' privacy perceptions and practices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 6789–6793.

[150] Zihao Yu, Nicholas Diakopoulos, and Mor Naaman. 2010. The multiplayer: multi-perspective social video navigation. In *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*. 413–414.

[151] Ding Zhang, Puqi Zhou, Bo Han, and Parth Pathak. 2022. M5: Facilitating multi-user volumetric content delivery with multi-lobe multicast over mmWave. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 31–46.

[152] Haoyu Zhang, Ganesh Ananthanarayanan, Peter Bodik, Matthai Philipose, Paramvir Bahl, and Michael J Freedman. 2017. Live video analytics at scale with approximation and {Delay-Tolerance}. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. 377–392.

[153] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024. Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. arXiv:2406.12235 [cs.CV] https://arxiv.org/abs/2406.12235

[154] Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2023. Tell Your Model Where to Attend: Post-hoc Attention Steering for LLMs. *ArXiv* abs/2311.02262 (2023). https://api.semanticscholar.org/CorpusID:265033525

[155] Qingyang Zhang, Hui Sun, Xiaopei Wu, and Hong Zhong. 2019. Edge video analytics for public safety: A review. *In Proceedings of the IEEE* 107, 8 (2019), 1675–1696.

[156] Yuxing Zhang, Jinchen Song, Yuehan Jiang, and Hongjun Li. 2023. Online Video Anomaly Detection. *Italian National Conference on Sensors* (2023).

[157] Yuanhao Zhang, Yumeng Wang, Xiyuan Wang, Changyang He, Chenliang Huang, and Xiaojuan Ma. 2025. CoKnowledge: Supporting Assimilation of Time-synced Collective Knowledge in Online Science Videos. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[158] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 589–597.

[159] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. 2024. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems* 37 (2024), 89943–89977.