

Agentic Multimodal Violence Detection and Human-Gated Emergency Dispatch

ASMAE BAALA¹, Mostafa Hanoune², Mohssine Bentaib³

Laboratory of Artificial Intelligence and Systems (LIAS), Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco^{1,2}

Laboratory of Information Technologies and Modeling (LTIM), Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco³

Abstract—Smart cities deploy dense camera networks, yet automated video analysis usually stops at a detection label, leaving the choice of responder to an overloaded human operator. This study presents an end-to-end multi-agent architecture that carries a violence label from ingestion to an audited, operator-confirmed dispatch decision. Two analyst agents (built with the Agno framework and Google Gemini 3.7 Flash) examine the video and the audio of a surveillance window, each returning a typed threat assessment. A deterministic, rule-based function, with no language model involved, fuses the two verdicts through an entertainment gate and a cross-modal conflict check. Flagged verdicts wait for operator confirmation, routing to police, ambulance, or fire follows fixed rules, and every decision is logged. On a balanced 200-clip XD-Violence subset, fused recall is 0.95 (95% confidence interval 0.89-0.98) at specificity 0.85, with no significant gain over either modality alone or over a single multimodal call, and the first routed responder is a required agency for 70 of the 95 detected violent clips. These figures use a disclosed prompt that takes each clip at face value. Under the production prompt, which tells the analysts to ignore staged violence, both analysts judged every clip to be staged entertainment, such as a film scene, so not one violent clip was flagged (fused recall 0.00). A silent failure in the model's client library, which turned rate-limit errors into "no threat" verdicts (19 of 30 clips in an earlier run), is characterized and fixed. The integration is feasible and auditable; real surveillance footage remains untested.

Keywords—Multi-agent systems; vision-language models; violence detection; video surveillance; smart cities; emergency dispatch; human-in-the-loop; explainable artificial intelligence

I. INTRODUCTION

Urban camera networks now generate far more footage than any control room can watch, so the limiting resource in video surveillance is operator attention rather than sensor coverage. Community-in-the-loop deployments of alerting "smart" surveillance are already being trialed in urban and campus settings [1], and the appeal is obvious: if a system can recognize a violent act and summon the right responder, response latency falls from minutes of human scanning to seconds of machine triage.

That appeal carries a physical cost. A system empowered to call police, an ambulance, or the fire service pays for every mistake in real-world consequences, and the empirical record is discouraging: an audit by the Office of the New York City Comptroller of the NYPD's ShotSpotter gunshot-detection

deployment confirmed only 13% of alerts as genuine gunfire [2]. An automated dispatcher inheriting a false-positive rate of that magnitude would flood a Public Safety Answering Point (PSAP) and could place people in danger when responders arrive armed. Any credible detection-to-dispatch system must therefore be designed around the asymmetry of its two error modes rather than around a single aggregate accuracy figure.

The perception task addressed here is binary violence detection: labeling a short surveillance window as violent (a threat) or non-violent (benign). On that task in isolation, training-free prompting of vision-language models (VLMs) already achieves state-of-the-art area-under-curve results [3], and this work makes no attempt to beat that bar. What the literature does not do, as Section II establishes, is carry the label forward: detection work stops at a score, while the small body of work on language-model agents for emergency response acts on textual or telephonic input rather than on video. The step between them (deciding, auditably and under human supervision, which responder to summon) has received little attention.

This study addresses that gap with a multi-agent system that carries a clip from ingestion through multimodal analysis, deterministic fusion, a human-in-the-loop (HITL) gate, and audited dispatch to the correct responder or responders. The contributions are:

- an end-to-end detection-to-dispatch loop implemented on the Agno multi-agent framework [4] with Google Gemini 3.7 Flash [5] (the original run used Gemini 2.5 Flash) as the multimodal reasoning model;
- structured, explainable verdicts, in which a video analyst and an audio analyst each emit a typed ThreatAssessment record (validated with the Pydantic library) rather than free text, so that all downstream logic is deterministic and every decision is traceable to named evidence;
- a deterministic, rule-based fusion and safety layer (an entertainment gate, cross-modal conflict detection, and severity levels) behind a HITL gate, following established guidance for consequential automated decisions [6];
- an explicit mapping from operational requirements to architectural mechanisms (Section III), which makes the design rationale auditable rather than implicit;

- a loop-centered empirical evaluation on a real 200-clip XD-Violence subset, with interval estimates and paired exact tests, an ablation study that tests each modality and each gate in isolation against a single-call audio-visual baseline, responder-routing accuracy against the corpus category codes, bounds on the human-gated operating point, and a characterization of a silent client-side failure mode that threatens any VLM-based safety pipeline;
- a released deterministic pipeline, an agentic team variant in which a coordinator model decides which agent to call, and an offline test suite that verifies the decision logic without model access, so the safety-critical component is reproducible independently of a vendor model whose answers vary from run to run.

The claims are deliberately narrow. This work does not advance detection accuracy; it does not constitute a certified PSAP, Enhanced 911 (E-911), or Computer-Aided Dispatch (CAD) integration; and, as Section IX sets out, its detection figures are measured on film-derived clips under an evaluation prompt and do not establish fitness for deployment. The contribution is the integration, its auditability, and an honest measurement of where the integrated loop currently breaks.

The remainder of the study is organized as follows. Section II reviews related work and positions the contribution. Section III derives operational requirements and a threat model. Section IV details the architecture. Sections V and VI present the evaluation methodology and results. Section VII discusses what the evidence supports, Section VIII the ethical and regulatory posture, and Section IX the limitations, before Section X concludes.

II. RELATED WORK

The system sits at the intersection of five literatures: VLM-based video anomaly detection, real-time and edge violence recognition, language-model agents for emergency response, evaluation datasets and reliability studies, and the governance of public-space video analytics. Each is reviewed in turn before the present work is positioned against them.

A. Vision-Language Models for Video Anomaly Detection

Vision-language and large language models (LLMs) increasingly target explainable video anomaly detection. EventVAD [9] achieves state-of-the-art training-free results through event-aware decomposition, and ASK-Hint [3] reaches state-of-the-art area-under-curve performance using fine-grained structured prompting, establishing that competitive detection requires no task-specific training. A parallel strand emphasizes interpretability: VERA [10] verbalizes learned explanations through a VLM, AssistPDA [11] frames an online surveillance assistant, a knowledge-graph and graph-attention system unifies detection, retrieval, and explanation while explicitly disclaiming dispatch [12], and short-window sliding learning enables real-time LLM-based auto-labeling [13]. This literature is methodologically mature and, without exception, terminates at a detection or explanation output. No member of it issues, routes, or audits a response.

B. Real-Time and Edge Violence Recognition

A second line, run on edge devices (small computers on or near the camera), prioritizes latency and energy over semantic richness. An embedded framework combining a convolutional neural network (CNN) with a gated recurrent unit (GRU) reaches 26 frames per second at 96.9% accuracy on a Raspberry Pi 5 with alert generation [14], and compact 3D CNNs have been shown to match VLMs fine-tuned with a lightweight method (LoRA) on accuracy while consuming markedly less energy in federated settings [15], [16]. On throughput, energy, and benchmarked accuracy, these systems are well ahead of the present one, but they emit a class label without a justification, and they neither distinguish sanctioned from malicious physical contact nor select a responder. The present design treats this tier as complementary rather than competing: a cloud VLM is invoked only on motion-positive windows (Section IV-B), and Section X identifies replacing the pixel-difference gate with such an edge classifier as the natural next step.

C. Language-Model Agents for Emergency Response

Agentic decision support for public safety is emerging but is not video-driven. LLM agents for smart city management enhance decision support over structured municipal data [7], and DispatchMAS fuses a clinical taxonomy with agents for emergency medical services dispatch [8], operating on call transcripts rather than surveillance imagery. Comparative benchmarking of agent frameworks for crisis solutions across SmolAgents, LangGraph, AutoGen, Agno, and CrewAI [23] informed the choice of runtime. The gap is thus symmetric to the detection literature: these systems act, but not from pixels.

D. Datasets and Reliability

Evaluation in this field rests on three corpora: UCF-Crime [17], a large weakly supervised anomaly benchmark; XD-Violence [18], the only large-scale violence corpus distributed with genuine audio, introduced together with a weakly supervised audio-visual detection network (HL-Net); and RWF-2000 [19], a real-world fight dataset. Reliability studies temper the optimism of the detection results above. A reality check of multimodal LLMs used without task-specific training documents a collapse in recall on real-world footage, with peak F1 ranging from 0.09 to 0.64 depending on prompting [20], and a transferable spatial attack achieves over 93% success against violence classifiers [21]. Both findings directly motivate design choices in Section IV: the first argues against trusting a single verdict, hence the use of two analysts, either of which can raise a threat; the second argues that no perception layer should be the sole arbiter of a dispatch.

E. Governance of Public-Space Video Analytics

Legal analyses of AI-based crowd analysis find substantive gaps in the regime formed by the General Data Protection Regulation (GDPR) and the EU AI Act [22], and Regulation (EU) 2024/1689 [25] heavily restricts real-time remote biometric identification in publicly accessible spaces. The design deliberately stays outside that category (it performs act detection, not person identification), and its audit log, evidence recorder, and HITL gate are intended as a defensible compliance posture rather than a claim of compliance. Documented robustness disparities in face detection across skin

tone, age, gender, and lighting [26] remain an uncharacterized risk for the underlying model, as Section IX concedes.

F. Positioning

To the best of the authors' knowledge, no prior work combines VLM-based violence detection, joint audio-visual reasoning, explainable structured output, a conflict and human-in-the-loop gate, and responder routing in a single operating loop. Detection work stops before response [3], [9]-[13], [18]; dispatch work acts but not from video [7], [8]; edge methods are fast but neither explain nor route [14]-[16]. Table I summarizes this coverage, with one caveat stated plainly: its first six rows describe the integration this study targets and were chosen for that reason, so the fact that only the present system covers all of them follows from how those rows were defined and is not evidence of superiority. The last three rows record dimensions on which the cited systems are stronger than the present one: detection benchmarked on full standard test sets, on-device inference, and a domain-specific dispatch taxonomy.

TABLE I. CAPABILITY COVERAGE OF REPRESENTATIVE PRIOR WORK AND OF THIS WORK

Capability	Even tVA D [9]	ASK - Hint [3]	HL- Net [18]	CNN - GR U [14]	Sm ar tCit y [7]	Disp atch- MA S [8]	This wor k
1. Video detection	Yes	Yes	Yes	Yes	No	No	Yes
2. Joint A/V	No	No	Yes	No	No	No	Yes
3. Structured output	Parti al	Parti al	No	No	Yes	Yes	Yes
4. HITL / conflict gate	No	No	No	No	No	Parti al	Yes
5. Dispatch	No	No	No	No	Plan ned	Yes (med ical)	Yes (P/A/ F)
6. Live operation	No	No	Parti al	Yes	No	Yes	Parti al
7. Full-benchmark detection	Yes	Yes	Yes	Parti al	-	-	No
8. On-device inference	No	No	No	Yes	-	-	No
9. Dispatch taxonomy	-	-	-	-	-	Yes	Parti al

^a Yes = present, Partial = limited, No = absent, and - = not applicable or not reported. Rows 1-6 describe the integration this study targets and were chosen for that purpose; rows 7-9 are dimensions on which prior work is stronger. Joint A/V denotes joint audio and video reasoning; P/A/F denotes police, ambulance, and fire. The present system's live operation is windowed, with second-scale latency (Section IV-B).

III. REQUIREMENTS AND THREAT MODEL

Because the system's outputs are physical, the operational requirements are stated before the architecture, so that each mechanism in Section IV can be traced to the requirement that motivates it. Table II records that mapping.

A. The Asymmetry of Errors

A missed violent act and a false alarm are not symmetric costs, but neither is free, and their relative weight depends on what happens downstream. A false negative means an assault proceeds unwitnessed: the system has added nothing beyond a conventional camera. A false positive, if it reaches a responder

unfiltered, consumes a finite emergency resource and may bring an armed response to a non-event, which the ShotSpotter audit [2] shows is the empirically dominant failure mode of deployed alerting systems. The resolution adopted here is to separate the two decisions: perception is tuned for recall, and the cost of the resulting false positives is absorbed by a human gate rather than by a responder. This is why the recall-focused either-analyst rule of Section IV-D and the HITL gate of Section IV-E are architecturally distinct components, and why their effects are reported separately in Section VI.

B. Operational Requirements

Six requirements follow. R1, recall priority: the perception layer should prefer flagging an ambiguous window to dismissing it, since a human can veto but cannot recover a discarded window. R2, determinism where auditable: the decision to dispatch must be a reproducible function of the analysts' structured verdicts, even though perception itself is delegated to a model whose answers can vary from run to run. R3, explainability: every verdict must cite named evidence, because a control-room operator confirming a dispatch needs grounds, not a bare score. R4, human authority over consequential action: no dispatch may fire from perception alone, per established human-in-the-loop guidance [6]. R5, auditability: every decision, confirmation, and dispatch must be recoverable after the fact for accountability and for post-incident review. R6, cost and latency discipline: cloud VLM calls are paid per use, so the system must not forward inactive windows.

C. Threat Model and Non-Goals

The threat model assumes a cooperative operator and a non-adversarial environment. It explicitly excludes the transferable spatial perturbations of [21], coordinated attempts to exhaust operator attention through induced false positives, and compromise of the notification channels. The vendor model is assumed to be available but neither deterministic nor reliably well-behaved under load, an assumption Section VI-A shows to be necessary. Person re-identification, face recognition, and tracking are non-goals, excluded by design rather than merely unimplemented.

TABLE II. OPERATIONAL REQUIREMENTS AND THE ARCHITECTURAL MECHANISMS THAT SATISFY THEM

Req.	Requirement	Mechanism	Section
R1	Recall priority	Threat raised if either analyst flags it	IV-D
R2	Determinism where auditable	Rule-based fusion with no language model; fixed routing table	IV-D, IV-E
R3	Explainability	Typed ThreatAssessment with cited evidence	IV-C
R4	Human authority over action	Confirm-or-reject HITL gate in every front end; conflict withhold	IV-E
R5	Auditability	IncidentDecision record; append-only log; evidence bundle	IV-D, IV-E
R6	Cost and latency discipline	Motion gate; per-threat cooldown	IV-B

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The system is a single end-to-end loop carrying a surveillance stream from raw pixels to an audited, human-gated dispatch. It is built on the Agno multi-agent framework [4] with Google Gemini 3.7 Flash [5] (the original run used Gemini 2.5 Flash) as the multimodal reasoning model: two specialist analyst agents feed a deterministic decision core. Fig. 1 traces the loop; the subsections below follow it stage by stage before contrasting the two orchestration variants.

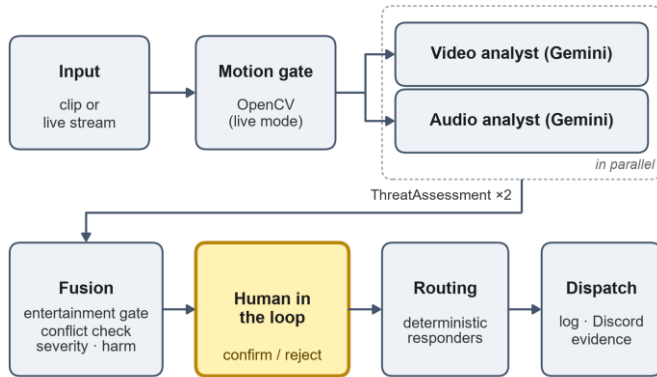


Fig. 1. End-to-end architecture: motion gate, parallel video and audio analysts, deterministic fusion and safety layer, human-in-the-loop gate, and routed dispatch, with every step written to an audit trail and an evidence bundle.

A. Design Principles

Three principles, traceable to Section III, constrain every stage. First, recall over precision (R1), since a missed act of violence costs more than a false alarm a human vetoes. Second, determinism where auditable (R2): perception is delegated to a VLM whose answers can vary from run to run, but the decision to dispatch is a plain, reproducible function of the analysts' structured verdicts. Third, separation of perception, decision, and action, so that each is independently testable, which is what allows the decision logic to be covered by an offline test suite that requires no model access.

B. Ingestion, Windowing, and the Motion Gate

The system ingests a live source (HLS, RTSP, MJPEG, or a resolved YouTube-live URL) and grabs a short window, five seconds by default, re-encoded to a Gemini-compatible MP4 container for windowed, near-real-time analysis. This is window-level rather than 30-frame-per-second per-frame processing, by design: the target is incident-level latency in seconds, not frame-synchronous tracking.

Each window is first screened by a deliberately cheap local computation satisfying R6. The motion score is the mean absolute inter-frame difference over grayscale, blurred frames, on a 0-255 scale, and a window is forwarded to the VLM only when that score exceeds a threshold τ , set to 3.0 by default. A per-threat cooldown suppresses repeated alerts for a single ongoing incident. This gate is a coarse activity filter and not a violence detector: it carries no semantic judgment, which remains entirely downstream. Its purpose is to avoid paying for model calls on empty scenes.

C. Dual-Modality Analysts and Structured Output

A forwarded window is passed to two specialist analyst agents. Because the clip, audio track included, is sent to the model as a single file, both agents attend to the same media artifact, each restricted to its own modality. Each returns a ThreatAssessment carrying a threat flag, a threat type, a confidence, an entertainment flag, a graded harm level, context notes, and cited evidence. Validating the output against this typed record, rather than parsing free text, is what makes the downstream logic deterministic and satisfies R3.

The video analyst verifies the presence of human subjects or weapons, analyzes body movement for violence, and reads scene context (scoreboards, crowd arrangement, uniforms) to distinguish malicious intent from sanctioned entertainment. The audio analyst flags screaming, shouting, and breaking glass, and notes cheering or laughter as entertainment cues. A fallback step turns any response that cannot be read as a valid assessment into a benign, low-confidence assessment rather than crashing the loop. Section VI-A reports, at length, why that apparently prudent default turned out to be the most dangerous line of code in the system.

The analysts are steered by prompt engineering alone, with no fine-tuning, and their instructions combine six techniques. 1) Role and modality framing: each agent is told which module it is and which track it covers, so that the two verdicts remain independent evidence. 2) A staged procedure: the video analyst first checks for the presence of people or weapons, then analyzes body movement, then reads context, so that scene context is weighed only once a subject is found. 3) Schema-constrained output: every ThreatAssessment field carries a natural-language description that the model receives as part of the requested output format (a JSON schema), so the schema itself instructs, and the response is validated on return. 4) Operational definitions: entertainment is tied to observable cues (scoreboards, structured crowds, sports uniforms, stadium lighting; cheering or laughter in the audio) rather than left to the model's discretion, and the recall-oriented variant defines confidence as the probability that a genuine threat is present rather than as self-certainty. 5) A graded scale for observed harm: four ordered levels (none, minor, serious, life-threatening), each tied to observable examples, such as a slap or a shove after which everyone stays on their feet for minor and a person knocked down who does not get up for serious, with a tie-break toward the more severe grade when the victim's condition cannot be observed. 6) Separation of judgment from action: the model grades what it observes but never names a responder; the deterministic router converts the grade into an ambulance or not (Section IV-E), which keeps that consequential choice out of the model's discretion (R2) and auditable (R5). The rubric was introduced after a live test in which two people slapping each other, recognized with high confidence, were routed to police and ambulance because the earlier rule added an ambulance to every high- or critical-severity alert. It was fixed before the evaluation run and was not tuned on its clips. For the evaluation of Section VI, the analysts and the single call were run under an evaluation-only variant of these instructions, described in Section V-C.

D. Deterministic Fusion: Entertainment Gate, Conflict Detection, Severity

The two assessments are combined by a rule-based function that calls no language model, satisfying R2. A threat is judged genuine when at least one modality flags it and not all flagging modalities judge the scene to be sanctioned entertainment; the second clause is the entertainment gate, which prevents a boxing match or a film from triggering a dispatch. This either-analyst rule is the direct expression of R1.

A cross-modal conflict is raised when fusion confirms a threat but at least one modality reads the scene as sanctioned entertainment, that is, when the analysts disagree about whether the observed violence is real. This mechanism was not planned but was forced by observation: repeated runs of the same clip through the same model returned differing verdicts, and the reliability literature reports the same fragility [20]. A conflicted verdict is therefore withheld from automatic action and routed to a human. The conflict flag changes what happens to a verdict, never the verdict itself, so the conflict gate leaves the detection view unchanged by construction and acts only on the auto-dispatch view (Section VI-C). Fusion confidence is the maximum confidence over flagging modalities, mapped to four severity bands (critical at 0.85 and above, high at 0.70, medium at 0.50, low below 0.50); the bands label the urgency of an alert but do not select responders. The fused harm level is likewise the highest harm grade among the flagging modalities (Section IV-C). The fused output is an IncidentDecision record carrying verdict, severity, harm level, confidence, and rationale, which is what makes the loop reproducible and reviewable (R5).

E. Routing, Human-in-the-Loop, Dispatch, and Evidence

Routing is deterministic and maps a confirmed threat to a subset of {police, ambulance, fire} in two steps. First, a keyword lookup on the fused threat type selects the lead agency: assault, fight, weapon, shooting, violence, scream, robbery, aggression, or breaking-glass cues select police; fire or smoke selects fire; medical or collapse cues select ambulance; and police is the safe default for an unmapped confirmed threat. The first matching keyword wins, which makes the lookup reproducible but literal: a label such as gunfire matches the fire key, and an unmapped label such as explosion or vehicle collision falls back to police. Section VI-D measures the consequences of that literalness against the corpus category codes. Second, an ambulance is added only on medical grounds: when the fused harm level is serious or life-threatening, or when the label names an event that presumably involves casualties (a shooting or gunfire, a stabbing, an explosion, a fire, a collision or crash, an accident, or an injury or bleeding). A mutual slap with no visible injury therefore routes to police alone, whereas the same altercation ending with a person on the ground who does not get up adds an ambulance. Severity, being a confidence band rather than a measure of injury, does not enter routing.

No dispatch fires from perception alone (R4). In the single-clip front end, a HITL gate presents the fused decision for confirm-or-reject [Fig. 2(a)]. The near-real-time Live Monitor applies the same rule to a stream: each motion-positive window that fuses to a genuine threat is placed in a review queue

together with both verdicts, an evidence frame, and the proposed responders, and it is dispatched only when an operator confirms it; a rejected alert is logged and never sent. Conflicted verdicts are always queued [6]. The per-threat cooldown applies to queued and dispatched alerts alike, so one ongoing incident produces one review item rather than a flood. For offline experiments only, the Live Monitor exposes an explicitly labeled research mode, off by default, that dispatches non-conflicted verdicts without confirmation; this was the Live Monitor's behavior in the original submission, and Section VIII explains why it is not a deployment configuration. The queueing rule is a plain function, checked by the offline test suite.

On confirmation, the dispatch stage packages a validated SecurityAlert, appends it to an append-only log, and delivers a severity-colored message per responder channel, skipping unconfigured channels and catching delivery errors so that a notification failure never breaks the pipeline. An evidence recorder persists a per-incident bundle. This is a proof-of-concept notification layer and not a certified PSAP, E-911, or CAD integration, although end-to-end delivery has been verified live (Fig. 3).

F. Implementation and Orchestration Variants

The multi-agent layer is built on Agno [4] for its multimodal and structured-output support, which avoids hand-written integration code; the choice is consistent with comparative benchmarking of agent frameworks for emergency-response tasks [23]. Every agent uses Agno's OpenAI-compatible chat model class against an AI gateway, so the model is a single configuration constant rather than a code dependency. Each analyst is bound to the model with structured output. The dispatcher is a separate agent exposing a single callable tool [24], a decomposition forced by a platform constraint: Gemini cannot combine structured output and function calling (tool use) within one agent.

Two orchestration variants expose the same components. A deterministic pipeline, used throughout the evaluation, invokes the stages in a fixed order. In the agentic team variant, a leader model decides which agent to call, so the order of the stages is best-effort rather than guaranteed. The deterministic pipeline is preferred for the safety-critical path: run-to-run variation in VLM answers is unacceptable in a dispatch decision, whereas a plain function remains reproducible, auditable, and testable offline. The agentic variant is retained because it is the more natural expression of the architecture and because it will benefit from future improvements in orchestration reliability.

V. EVALUATION METHODOLOGY

A. What is Measured, and Why

The evaluation targets the loop (detection, responder routing, and the action taken on each verdict) rather than detection area-under-curve in isolation, because the contribution is the integration and because training-free VLM prompting already leads on detection alone [3], [9]. Following the error asymmetry of Section III-A, the quantities that matter are recall, since missed threats are safety-critical, and the false-alarm rate, since each false alarm risks an unnecessary armed or medical response [2]. The ground-truth label is binary, so

detection is scored as a binary task; routing is scored separately against the corpus's category codes (Section V-C).

Three views of every experiment are reported. The detection view asks whether the fused IncidentDecision flags a genuine threat. The auto-dispatch view describes unsupervised operation: a flagged threat counts as dispatched only if the conflict gate does not withhold it, so a withheld true threat counts as a miss. That view is therefore a lower bound on what the supervised system would do, not an estimate of it. The human-gated view, which is how every front end now operates, sends every flagged verdict to an operator. No human operator was measured in this study, so that view is reported as bounds under stated operator policies (Section VI-E) rather than as a point estimate. Routing accuracy is reported separately (Section VI-D).

B. Metrics and Interval Estimation

Detection quality uses standard binary-classification metrics against the ground-truth label. With TP, FP, TN, and FN denoting the numbers of true positives, false positives, true negatives, and false negatives (integer clip counts summing to N), the reported quantities are precision, recall, specificity, the false-alarm rate $FPR = 1 - \text{specificity}$, F1, and accuracy, all unitless on [0, 1]:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

Recall and specificity are given priority over F1 and accuracy, because the latter two average across an asymmetry that matters operationally. Every proportion is accompanied by a 95% Wilson score interval, which remains appropriate for small counts and for proportions near 0 or 1 where the normal approximation fails. Because every condition is evaluated on the same clips, comparisons between conditions are paired and are tested with the exact two-sided McNemar test on per-clip correctness, which uses only the clips on which the two conditions disagree. Conclusions are drawn only where such a test, or non-overlapping intervals, support them.

C. Dataset, Sampling, and Configuration

The evaluation uses a balanced 200-clip subset, 100 violent and 100 normal, deterministically sampled with a fixed seed from the public XD-Violence mirror (jhermg/xd-violence on Hugging Face): real H.264 clips with native audio. The subset contains the 30 clips of the original free-tier run, so that the two runs can be compared clip for clip (Section VI-A), although the model also differs between them. XD-Violence [18] is chosen because it is the only large-scale corpus distributed with genuine audio, without which the audio analyst, and therefore the fusion mechanism, cannot be evaluated at all. Each clip carries up to three category codes, from which the binary label is derived; the violent clips

comprise fighting (B1, 49), shooting (B2, 42), riot (B4, 6), abuse (B5, 3), car accident (B6, 12), and explosion (G, 20); 28 clips carry more than one code and are counted under each. For routing, each code is mapped to the agency that must be notified: police for fighting, shooting, riot, and abuse; ambulance for a car accident; and fire for an explosion. A detected clip is routed correctly when its first routed responder is one of its required agencies, and is covered when the routed set contains all of them. The single-call baseline uses the same model and the same output schema, receives the combined instructions of both analysts, and sees the audio and video tracks in one call; it is scored with and without the same entertainment gate. Evaluation on RWF-2000 [19] and UCF-Crime [17] remains future work, and no generalization beyond the present sample is claimed.

The rerun used Gemini 3.7 Flash through an OpenAI-compatible AI gateway, with up to 4 clips processed at a time and temporary errors retried after increasing waits, on Python 3.12 or later with agno 2.5 or later and the openai Python library 3.31 or later. The gateway carries each clip inside the request itself, encoded as text (base64), so the 42 of 200 scored clips larger than 13.5 MiB were re-encoded once, whole and with the soundtrack, to H.264 at 10 frames per second with 48 kbps mono AAC audio; every other clip was sent unchanged. The original 30-clip run used Gemini 2.5 Flash on the free tier of the Gemini Developer API, whose rate limits produced the failure of Section VI-A. The gateway does not serve that model, so every comparison with the original run spans both the model change and the corrected retry logic. The detection results of Section VI use an evaluation prompt that differs from the production prompt of Section IV-C in one respect. Each analyst, and the single call, is told to treat every clip as real footage from a public-safety camera and to judge the events at face value: cinematic framing, editing, recognizable actors, a film soundtrack, or a studio mix are not evidence that the events are harmless, and only a sanctioned activity seen or heard in the scene itself, such as a refereed bout, a match, or a class, counts as entertainment. The staged procedure, the harm rubric, and the output schema are unchanged, and the production system keeps the production prompt. The change was made after the production prompt cleared the film-derived clips as staged (Section VI-B); the evaluation prompt was fixed before its run and was not tuned on its results. Table III summarizes published results from the literature for context; these numbers are not directly comparable to the present ones, since they are obtained on different corpora, at far larger n, and for the detection task alone.

TABLE III. PUBLISHED BASELINES FROM THE LITERATURE, FOR CONTEXT ONLY

Method	Corpus	Reported result	Task scope
Flow-Gated [19]	RWF-2000	87.25% accuracy	Detection only
[3], [9]	UCF-Crime / XD-Violence	State of the art (training-free)	Detection only
[14] (edge)	Not reported	96.9% accuracy, 26 frames/s on a Raspberry Pi 5	Detection only
[20]	ShanghaiTech	Peak F1 0.09 to 0.64	Reliability study

^b These results are obtained on different corpora, at substantially larger n, and for the detection task in isolation; they are not directly comparable to Tables IV and V.

VI. RESULTS

A. A Silent Failure Mode in the Vendor API Client

The first result is a measurement hazard that generalizes to any safety pipeline built on a pay-per-use VLM service, and it is reported before the metrics because it conditions them.

The model's client library does not always report an error when the service answers with a rate-limit (HTTP 429) or service-unavailable (HTTP 503) response. In the original free-tier run, on 19 of 30 clips (63%, 95% CI 46-78%), it returned the raw error message from the service as plain text in the position where a structured assessment was expected. The fallback step of Section IV-C caught these and marked them [non-response], which is what made the problem detectable at all, but that same fallback turns an unreadable response into a benign, low-confidence assessment. Had the marker not been logged, every rate-limited window would have been silently recorded as "no threat detected," deflating recall and producing a system that fails silently and unsafely under load. Because the original detection metrics therefore rested on the 11 clips that escaped the rate limits, a selection made by the rate limits rather than by design, every result below comes from a complete rerun with corrected retry logic.

The general lesson is that in a recall-critical pipeline, a safe-looking default must never be the same as the "no threat" answer. Turning an unreadable response into a benign verdict makes it look, downstream, exactly like a confident all-clear. The correct behavior is to pass on an explicit "unknown" state that the fusion layer treats as neither threat nor clearance, and that surfaces to an operator as a coverage gap. The evaluation code has since been changed to retry this response pattern as an exception, and also a second pattern observed under provider overload, in which a call ends with no content at all; any deployment of a pay-per-use VLM in a safety path should likewise check the type of every response rather than trust the client library to report errors.

In the rerun, all 200 clips were scored, and none carried a non-response from any analyst; on the original 30 clips, 0 remained affected, against 19 before. Because a contaminated clip carries no genuine verdict, detection metrics are reported on the 200 clips (100 violent, 100 normal) for which the video analyst, the audio analyst, and the single call all returned a genuine assessment, so that every condition is compared on identical clips. Operational measurements are reported over all 200 scored clips.

B. Detection and Auto-Dispatch Performance

Table IV reports the evaluation-prompt detection and auto-dispatch views on the clean $n = 200$. Fused recall is 0.95 (95 of 100 violent clips, 95% CI 0.89-0.98), with specificity 0.85 (85 of 100 normal clips cleared, 95% CI 0.77-0.91). In unsupervised operation, auto-dispatch recall falls to 0.94 because the conflict gate withholds 1 of the 95 true positives for human review. Of the 5 violent clips that fusion misses, 4 were labeled entertainment by at least one analyst, which Section VI-C examines. These figures describe face-value judgment of film-derived clips under the evaluation prompt; they do not establish deployability (Section VII-C) and are not directly comparable with the training-free detectors of [3] and

[9], whose results are obtained on other corpora and with other metrics (Table III).

Before Table IV was produced, the same 200 clips were scored under the production prompt of Section IV-C, the baseline prompt of the original run, which instructs both analysts to set genuine threats apart from sanctioned entertainment; that run is reported here as a finding. On its 200 clean clips, every condition (each analyst alone, fusion, and the single call with or without the gate) has recall 0.00 (0 of 100 violent clips) at specificity 1.00. The video analyst labels all 200 clips as entertainment, and so does the audio analyst; the video analyst's notes or cited evidence name film, movie, or cinematic cues on all 200, in some cases naming the film itself. The violence itself was seen: in a later reading of the outputs, on 56 violent clips that fusion cleared, the video analyst nonetheless graded the observed harm serious or life-threatening. The failure lies in what the corpus label means, not in what the model sees: XD-Violence is drawn largely from films, and an analyst that correctly recognizes a film scene as staged is, by the corpus label, missing a threat. Table IV therefore uses the evaluation prompt of Section V-C, which removes that judgment and leaves the procedure, the harm rubric, and the output schema unchanged; the production system keeps the production prompt, and how it performs on real surveillance footage, where the judgment matters, is untested.

The two runs together locate where detection quality is decided. The loop, the fusion rule, and the harm rubric are identical in both; only the analysts' instruction on staged footage differs, and fused recall moves from 0.00 to 0.95. Detection quality is therefore a property of the perception configuration and of its fit to the corpus, not of the loop architecture, whose decision, action, and audit layers do not change when the perception prompt does. Prompt validation at scale [3], [20], on real surveillance footage, and a lightweight on-camera trigger [14] remain the direct remedies (Section X).

TABLE IV. DETECTION AND AUTO-DISPATCH METRICS ON THE CLEAN $N = 200$ XD-VIOLENCE SUBSET (EVALUATION PROMPT), WITH 95% WILSON SCORE INTERVALS

View	Accuracy	Precision	Recall	Specificity	F1
Detection	0.900 [0.85, 0.93]	0.864 [0.79, 0.92]	0.950 [0.89, 0.98]	0.850 [0.77, 0.91]	0.905
Auto-dispatch	0.895 [0.84, 0.93]	0.862 [0.79, 0.91]	0.940 [0.88, 0.97]	0.850 [0.77, 0.91]	0.900

^c Counts: detection TP = 95, FP = 15, TN = 85, FN = 5; auto-dispatch TP = 94, FP = 15, TN = 85, FN = 6, over 100 violent and 100 normal clips. Each proportion is shown above its 95% Wilson interval; F1 is a derived quantity and is given without an interval.

C. Ablation: Modalities, Gates, and a Single-Call Baseline

Table V recomputes the decision under seven conditions on the same clean $n = 200$: each modality in isolation, fusion without gates, fusion with both gates (the detection view), the auto-dispatch view, and the single-call baseline with and without the entertainment gate. The conflict gate never changes the detection verdict, only whether it may be dispatched automatically, so fusion with the entertainment gate alone is identical to fusion with both gates by construction and is

shown once. Table VI gives paired exact McNemar tests, computed separately on violent clips, where they compare recall, and on normal clips, where they compare specificity, because a recall gain and a specificity loss can cancel in overall accuracy.

Complementarity is tested here on paired clips rather than read off a point difference. Of the 100 violent clips, 93 are flagged by both analysts, 0 by the video analyst alone, 2 by the audio analyst alone, and 5 by neither. Video alone reaches recall 0.93 (95% CI 0.86-0.97) and audio alone 0.95 (0.89-0.98); fusion reaches 0.95 (0.89-0.98). Neither gain is significant on paired violent clips ($p = 0.500$ against video, $p = 1.000$ against audio), so this sample does not establish that the modalities are complementary. Restricted to the original 30 clips, where the earlier claim rested on a difference of one clip, the rerun (with Gemini 3.7 Flash in place of Gemini 2.5 Flash) has 30 clean clips (15 violent), with fused recall 1.00 against 1.00 for video and 1.00 for audio. The cost falls on the normal clips: the video analyst flags 7 of the 100 and the audio analyst 13, so fusion's specificity is 0.85, against 0.93 for video and 0.87 for audio (paired $p = 0.008$ and $p = 0.500$).

The conflict gate determines what the system would do unsupervised. It withholds 1 of the 95 true positives that fusion detects and 0 of its 15 false positives, lowering auto-dispatch recall from 0.95 to 0.94, with specificity unchanged. This is the central tension of the design, and it is not presented as a positive result. The gate is correct, since it prevents an unpredictable model from commanding an armed response on a verdict its own analysts dispute, but in unsupervised operation it discards 1% of what fusion catches, which is why every front end now sends flagged verdicts to an operator (Section VI-E).

The entertainment gate is judged by the verdicts it changes. On the 200 clean clips, at least one analyst labels the scene as entertainment on 19 (5 violent, 14 normal). The original manuscript reported this kind of labeling rate as the rate at which the gate fired, which overstated its effect: the gate changes a verdict only when every analyst that flags a threat also calls the scene entertainment. On this sample it changes no verdict, so fusion with and without it is identical (Table V). Far more consequential is what the analysts do before the gate. Of the 5 violent clips that fusion misses, 4 were labeled entertainment by at least one analyst; the gate suppressed none of them, since no analyst flagged a threat at all. Under the production prompt the gate was never reached, since no analyst flagged a threat on any clean clip (Section VI-B).

The single-call baseline tests whether splitting the analysis between two specialists is worth its cost. One call that sees both tracks under the combined instructions of the two specialists reaches recall 0.93 (95% CI 0.86-0.97) and specificity 0.89 (0.81-0.94); with the same entertainment gate, recall 0.93 and specificity 0.89, against 0.95 and 0.85 for two-analyst fusion. Against the gated single call, the paired difference is not significant ($p = 0.500$) on violent clips and not significant ($p = 0.219$) on normal clips (Table VI). The single call costs about 10,517 tokens and a median of 44.3 s per clip, against 20,235 tokens and 48.8 s for the two analysts running in parallel. On this sample, therefore, the two-analyst design is

not shown to detect better than a single multimodal call; its justification rests on the per-modality evidence and the cross-modal conflict signal it exposes to the decision layer, which a single call cannot provide, rather than on a measured detection gain.

TABLE V. ABLATION ON THE SAME CLEAN N = 200: MODALITY ISOLATION, GATE CONTRIBUTION, AND A SINGLE-CALL BASELINE

Condition	n	Accuracy	Precision	Recall	Specificity	F1
Video only	200	0.930	0.930 [0.86, 0.97]	0.930 [0.86, 0.97]	0.930 [0.86, 0.97]	0.930
Audio only	200	0.910	0.880 [0.80, 0.93]	0.950 [0.89, 0.98]	0.870 [0.79, 0.92]	0.913
Fusion, no gates	200	0.900	0.864 [0.79, 0.92]	0.950 [0.89, 0.98]	0.850 [0.77, 0.91]	0.905
Fusion, both gates (detection)	200	0.900	0.864 [0.79, 0.92]	0.950 [0.89, 0.98]	0.850 [0.77, 0.91]	0.905
Auto-dispatch (conflicts withheld)	200	0.895	0.862 [0.79, 0.91]	0.940 [0.88, 0.97]	0.850 [0.77, 0.91]	0.900
Single call (audio + video)	200	0.910	0.894 [0.82, 0.94]	0.930 [0.86, 0.97]	0.890 [0.81, 0.94]	0.912
Single call + entertainment gate	200	0.910	0.894 [0.82, 0.94]	0.930 [0.86, 0.97]	0.890 [0.81, 0.94]	0.912

^d Each proportion is shown above its 95% Wilson score interval. Fusion with the entertainment gate alone equals fusion with both gates by construction, since the conflict gate never changes the detection verdict. All conditions are evaluated on identical clips; Table VI gives the paired tests.

TABLE VI. PAIRED EXACT MCNEMAR TESTS ON THE CLEAN N = 200

Comparison	Violent clips (recall)	Normal clips (specificity)	All clips (accuracy)
Fusion vs video	2 / 0 $p = 0.500$	0 / 8 $p = 0.008$	2 / 8 $p = 0.109$
Fusion vs audio	0 / 0 $p = 1.000$	0 / 2 $p = 0.500$	0 / 2 $p = 0.500$
Video vs audio	0 / 2 $p = 0.500$	8 / 2 $p = 0.109$	8 / 4 $p = 0.388$
Fusion vs single call	2 / 0 $p = 0.500$	1 / 5 $p = 0.219$	3 / 5 $p = 0.727$
Fusion vs single call + gate	2 / 0 $p = 0.500$	1 / 5 $p = 0.219$	3 / 5 $p = 0.727$

^c Each cell gives the clips that only the first condition classifies correctly / only the second, and the two-sided exact p-value. Violent clips compare recall, normal clips specificity, and all clips accuracy. Fusion is fusion with both gates; no correction for multiple comparisons is applied.

D. Routing Accuracy Against Corpus Category Codes

Routing is scored on the violent clips that the fused decision detects, against the reference of Section V-C. Of the 100 clean violent clips, fusion detects 95; for 70 of these the first routed responder is a required agency (primary accuracy 0.74, 95% CI 0.64-0.81), and for 64 the routed set contains every required agency (coverage 0.67). End to end, 70 of the 100 violent clips are both detected and routed to a correct first responder (0.70, 95% CI 0.60-0.78). The routing errors are 11 clips requiring police routed first to fire, 11 clips requiring fire routed first to police, 2 clips requiring ambulance routed first to police, and 1 clip requiring ambulance and fire routed first to police. Of these, 9 carry a threat label that matches no routing

keyword and so falls back to police (for example, “explosion” and “explosive_attack”), 2 carry a label in which a routing keyword occurs only inside another word, which the plain text matching of Section IV-E misroutes (“gunfire”), and 14 carry a threat label that itself names an incident handled by another agency (for example, “armed_assault” and “explosion and armed assault”); 11 errors therefore stem from the lookup of Section IV-E and 14 from perception, through the fused label. Table VII breaks detection and routing down by category.

TABLE VII. DETECTION AND ROUTING BY XD-VIOLENCE CATEGORY ON THE CLEAN N = 200

Category (n)	Video	Audio	Fusion	Single call	Routed correctly
Normal, A (100)	93/100	87/100	85/100	89/100	-
Fighting, B1 (49)	44/49	45/49	45/49	44/49	43/45
Shooting, B2 (42)	42/42	42/42	42/42	42/42	30/42
Riot, B4 (6)	5/6	5/6	5/6	5/6	5/5
Abuse, B5 (3)	3/3	3/3	3/3	3/3	3/3
Car accident, B6 (12)	12/12	12/12	12/12	12/12	0/12
Explosion, G (20)	19/20	20/20	20/20	19/20	2/20

^f. Detection columns give the clips each condition classifies correctly out of the category’s clean clips: violent clips detected or, for normal clips, clips cleared. A clip with several codes is counted under each. Routed correctly counts detected clips whose first routed responder is that category’s reference agency (Section V-C); for a multi-code clip this is stricter than the primary accuracy in the text, which accepts any of its required agencies. Fusion is fusion with both gates; the single call is gated.

E. Human-Gated View: Review Queue and Operator Bounds

Every front end now sends each flagged verdict to an operator, so the operating point of the deployed loop depends on a person who was not part of this study. Table VIII therefore brackets it rather than estimating it. Over the 200 clean clips, fusion places 110 alerts in the review queue (0.55 per clip), of which 95 are true and 15 false (queue precision 0.86, 95% CI 0.79-0.92); 1 alert is conflicted, and it is a true alert. An operator who confirms every alert reproduces the detection view (recall 0.95, specificity 0.85); an operator who confirms exactly the true alerts keeps that recall and removes every false alarm; an operator who rejects everything dispatches nothing. Recall under human gating is therefore capped by perception at 0.95, whatever the operator does, and the operator’s contribution is confined to suppressing false alarms. The unsupervised auto-dispatch view, at recall 0.94, is a lower bound that counts every withheld true threat as a miss; reviewing only the 1 conflicted alert, perfectly, would recover recall 0.95. How quickly and how accurately real operators clear such a queue is the open question on which the design depends, and it is left to a user study.

TABLE VIII. OPERATING POINTS OF THE HUMAN GATE ON THE CLEAN N = 200: UNSUPERVISED VIEWS AND OPERATOR BOUNDS

Operating view	Recall	Precision	Specificity
Auto-dispatch, conflicts never reviewed	0.940 [0.88, 0.97]	0.862 [0.79, 0.91]	0.850 [0.77, 0.91]
Auto-dispatch, conflicts reviewed perfectly	0.950 [0.89, 0.98]	0.864 [0.79, 0.92]	0.850 [0.77, 0.91]
Human-gated, confirm-all operator	0.950 [0.89, 0.98]	0.864 [0.79, 0.92]	0.850 [0.77, 0.91]

Human-gated, perfect operator	0.950 [0.89, 0.98]	1.000 [0.96, 1.00]	1.000 [0.96, 1.00]
Human-gated, reject-all operator	0.000 [0.00, 0.04]	-	1.000 [0.96, 1.00]

^g No operator was measured. The three human-gated rows are bounds under the stated operator policies, computed from the same per-clip fused decisions; the auto-dispatch rows describe unsupervised operation, in which conflict-free alerts are dispatched without review. Proportions carry 95% Wilson score intervals; precision is undefined when nothing is dispatched.

F. Operational Characteristics

Table IX reports operational measurements. Latency and token figures are taken over all 200 scored clips; label, gate, and conflict figures depend on verdict content and are taken over the 200 clean clips. Through the gateway, median end-to-end latency for the two-analyst pipeline is 48.8 s per clip (range 11.4-630 s); the two analysts run in parallel, so this is the time to prepare the clip (re-encoding it when it exceeds the request size limit) plus the slower of the two calls, including any retry waits. The single call takes a median of 44.3 s. The pipeline consumes about 20,235 tokens per clip and the single call about 10,517, a ratio of 1.9; all calls together consumed 6,150,360 tokens. Cross-modal conflicts arise on 1 of the 200 clean clips, and the entertainment gate changes no verdict.

A recall-oriented prompt variant was begun in the original free-tier run as a side-by-side comparison on identical clips, but rate limits capped it at n = 5, of which two were contaminated. It was not repeated in the rerun, and no conclusion is drawn from it; the prompt comparison the rerun does report is the one between the production prompt and the evaluation prompt of Section V-C (Section VI-B). Completing the recall-oriented comparison at full scale remains future work.

TABLE IX. OPERATIONAL CHARACTERISTICS OF THE RERUN

Quantity	Value
Clips sampled/scored/failed	200 / 200 / 0
Non-response, any analyst (Section VI-A)	0 (0.00 [0.00, 0.02])
Median latency, two-analyst pipeline	48.8 s (range 11.4-630 s)
Median latency, single call	44.3 s (range 10.9-623 s)
Mean tokens per clip, pipeline / single call	20,235 / 10,517
Total tokens, all calls	6,150,360
Entertainment label, any analyst (clean)	0.10 [0.06, 0.14]
Verdicts changed by the entertainment gate (clean)	0 of 200
Cross-modal conflicts (clean)	0.01 [0.00, 0.03]

^h Proportions carry 95% Wilson score intervals. Latency and tokens are over all scored clips; label, gate, and conflict figures are over the clean clips, because they depend on verdict content. The entertainment label counts clips on which at least one analyst calls the scene entertainment, which is not the same as the gate changing a verdict. Output tokens are as reported by the gateway and include the model’s reasoning (thinking) tokens.

G. Observed Live Behavior

Two pieces of observed live evidence complement the offline metrics. End-to-end per-responder dispatch was verified live (Fig. 3): a physical_assault and_gunfire alert at 95% confidence from the audio modality routed to police and ambulance, and a fire_incident alert at 98% from the video modality routed to fire and ambulance. Both alerts were captured under the earlier routing rule, which added an ambulance to every high- or critical-severity alert; under the

current harm-graded rule (Section IV-E) both would still include an ambulance, because gunfire and fire are presumed to involve casualties. Both front ends were then recaptured on the current build (Fig. 2). In single-clip mode [Fig. 2(a)], a clip flagged by the video analyst as a physical assault fuses to a critical threat with police proposed, and the decision waits at the confirm-or-reject step. The Live Monitor, in its default mode [Fig. 2(b)], exercised the near-real-time path: a window passing the motion gate with a score of 12.78 was flagged by both analysts at 95% confidence with serious harm, and was placed in the review queue with police and ambulance proposed and its evidence frame attached; nothing was sent, since dispatch requires the operator's confirmation. These are demonstrations of mechanism, not measurements of accuracy, and they are presented only as evidence that the action and audit layers function end to end.

(a) **Video Analyst** **Audio Analyst**

THREAT **clear**

Fused decision

Genuine threat: **YES** Severity: **CRITICAL** Confidence: 85%

Rationale: Threat detected by video; not gated as entertainment. Observed harm: minor.

Action

Proposed dispatch → **police**

Human-in-the-loop: confirm before responders are notified.

Confirm dispatch **Reject**

(b) ☐ Research mode: auto-dispatch without confirmation

Proposed dispatch → police, ambulance queued for human review (awaiting operator confirmation).

Review queue

INC-BB517BC9 · 20:16:42 · physical_assault · severity critical · confidence 95% · via video, audio

Proposed responders: **police** **ambulance**

Confirm & dispatch **Reject (no dispatch)**

time	motion	verdict	action
20:16:42	12.78	THREAT (physical_assault)	queued INC-BB517BC9 for review → police, ambulance

Fig. 2. The application (current build): (a) single-clip mode showing analyst verdicts, the fused decision, and the human-in-the-loop confirm-or-reject step; (b) the Live Monitor in its default mode (research mode off), where a motion-gated window classified as a critical physical assault is held in the review queue, with its evidence frame and proposed responders, until an operator confirms or rejects it.

(a) **#police-dispatch**

CRITICAL ALERT — physical_assault_and_gunfire
physical_assault_and_gunfire detected via audio.
Confidence: 95% Sources: audio Responder: police
Recommended action: Dispatch police, ambulance to the monitored location.
Today at 5:44 PM

(b) **#ambulance-dispatch**

CRITICAL ALERT — physical_assault_and_gunfire
physical_assault_and_gunfire detected via audio.
Confidence: 95% Sources: audio Responder: ambulance
Recommended action: Dispatch police, ambulance to the monitored location.
Today at 5:44 PM

(c) **#fire-dispatch**

CRITICAL ALERT — fire_incident
fire_incident detected via video.
Confidence: 98% Sources: video Responder: fire
Recommended action: Dispatch fire, ambulance to the monitored location.
Today at 5:58 PM

(d) **#fire-dispatch**

CRITICAL ALERT — fire_incident
fire_incident detected via video.
Confidence: 98% Sources: video Responder: ambulance
Recommended action: Dispatch fire, ambulance to the monitored location.
Today at 5:58 PM

Fig. 3. Per-responder notifications. Police and ambulance both receive the physical assault and gunfire alert (95%, audio); fire and ambulance both receive the fire alert (98%, video). Each message reports confidence, source modality, responder, and recommended action.

VII. DISCUSSION

A. What the Evidence Supports

Two claims are supported. The integration is feasible: a clip traverses ingestion, dual-modality analysis, deterministic fusion, a human gate, routing, and audited dispatch, and this was observed end to end. The decision layer is reproducible and testable independently of the model, which is the design's most transferable property, and its routing errors can be enumerated against a reference (Section VI-D). The rerun does not, however, support the original manuscript's claim that fusing the two analysts outperforms either alone: the recall gain is significant against neither modality (Section VI-C). The two-analyst structure is justified here by the per-modality evidence and the conflict signal it exposes to the decision layer, not by a measured detection gain.

One claim is explicitly not supported: that the system is ready for use in a control room. Fused recall of 0.95 at specificity 0.85 is measured on film-derived clips that the analysts were told to judge at face value; under the production prompt, which leaves that judgment to the model, fused recall on the same clips is 0.00. Neither result comes from real surveillance footage; the sample is a single corpus judged by a single model, and no operator was measured. The safety gate that makes the system responsible also makes it, without an attentive operator, marginally less effective.

B. The Recall-Trust Tension

The gap between detection recall (0.95) and auto-dispatch recall (0.94) is not a defect to be engineered away but a structural property of any design that refuses to let an unpredictable model act alone. Raising auto-dispatch recall means either improving perception until the modalities agree more often, or loosening the conflict rule. Only the first is acceptable here. The conflict rule exists because run-to-run disagreement was observed on identical input; loosening it would convert observed model instability directly into dispatched responses. The corollary is that the human review queue is not an optional convenience in this architecture but its essential component, which is why every front end now routes flagged verdicts through it, and that its design [queue latency, operator workload, and the interface of Fig. 2(a)] deserves the same empirical scrutiny given here to the detector; Section VI-

E bounds what an operator can contribute but does not measure it.

C. Threats to Validity

Internal validity is limited by the contamination described in Section VI-A, now 0% of scored clips, and by the size of the clean sample ($n = 200$), which still leaves the fused-recall interval 0.09 wide; this is mitigated with interval estimates and paired exact tests. Construct validity is limited in three ways: the binary violence label collapses a heterogeneous set of acts into one class, which the per-category breakdown of Table VII only partly undoes; XD-Violence is drawn largely from films, which under the production prompt led the analysts to clear every violent clip as staged (Section VI-B), while the evaluation prompt that removes that judgment measures face-value perception of film footage rather than operation on real surveillance video; and the routing reference of Section V-C is a mapping from category codes chosen for this study rather than an operational ground truth. The assumption that XD-Violence clip boundaries approximate operational windows is also untested. External validity is limited to a single corpus and a single model, and the reported detection figures use an evaluation prompt rather than the production prompt; no claim is made about other models, other cities, or other camera geometries. Finally, the ground truth is the corpus label, which is not the same as an operational definition of an incident warranting dispatch, a gap that any real deployment study would have to close with the responder organization itself.

VIII. ETHICS, PRIVACY, AND REGULATORY POSTURE

This work's central commitment is that a consequential dispatch decision should rest with a person rather than with a model alone. The ShotSpotter audit [2] found only 13% of gunshot alerts confirmed as actual gunfire, and an inherited false-positive rate of that magnitude would flood a PSAP and endanger people when responders arrive armed. The fusion is therefore recall-focused but withholds trust: in every front end a flagged verdict reaches a responder only after an operator confirms it [Fig. 2(a)], and a conflicted verdict can never be dispatched automatically. In the version described in the original submission, the Live Monitor was an exception that auto-dispatched non-conflicted verdicts; it now queues them for confirmation by default as well (Section IV-E). Unconfirmed dispatch survives only as an explicitly labeled research mode, off by default, which still queues every conflicted verdict; it exists for offline experiments and is not a deployment configuration.

Continuous analysis of public video is a mass-surveillance capability regardless of the intent behind it, and the evidence recorder's persisted clips are a privacy liability in the absence of strict retention governance. The system implements no face recognition, re-identification, or tracking: the task is act detection, not person identification. The regulatory environment remains unsettled [22]. Real-time remote biometric identification in publicly accessible spaces is heavily restricted under Regulation (EU) 2024/1689 [25], a category the design avoids by construction, but automated processing of public video still engages GDPR obligations, so the human in the loop forms part of a defensible compliance posture rather than a claim of compliance. No Data Protection Impact

Assessment, which a real deployment would require, has been conducted; the audit log and evidence bundle are necessary inputs to such an assessment rather than substitutes for it.

There is also an asymmetry of benefit and risk. The population subject to this analysis does not opt in, and the harms of a false positive fall on individuals while the benefits of a true positive accrue diffusely. That asymmetry is an argument for deployment only under institutional oversight with published error rates, and it is why this study reports the production-prompt failure of Section VI-B alongside the evaluation-prompt results rather than the better of the two alone.

IX. LIMITATIONS

The reasoning core is a model whose run-to-run variability motivates the conflict-detection and HITL mechanisms rather than being resolved by them; multimodal LLMs used without task-specific training also lose much of their recall on real-world anomaly detection [20]. No fairness property has been measured: Gemini 3.7 Flash is uncharacterized across demographic groups for this task, and vision systems show documented robustness disparities across skin tone, age, gender, and lighting [26]. Violence classifiers can be deliberately fooled, with a transferable spatial attack achieving over 93% success [21], and neither robustness nor any defense has been evaluated. The per-responder chat notifications (webhooks) are not a certified PSAP, E-911, or CAD integration. Detection performance is measured on 200 clean clips of a single, film-derived corpus under an evaluation prompt; on the same clips the production prompt reaches fused recall 0.00 (Section VI-B), so neither figure estimates performance on real surveillance footage. No human operator was measured, so the human-gated operating point is bounded rather than known (Section VI-E). The motion gate is a pixel-difference heuristic that will degrade under camera shake, lighting change, and weather. Latency is dominated by a pay-per-use cloud service, so the measured figures describe the experimental conditions rather than a provisioned deployment. This is a research prototype whose mitigations are necessary but demonstrably not sufficient.

X. CONCLUSION AND FUTURE WORK

This study has presented an end-to-end system closing the loop from surveillance-video violence detection to human-confirmed emergency dispatch. Rather than stopping at detection [3], [9]-[13] or dispatching from non-video signals [7], [8], the contribution is the integration itself: multimodal Gemini analyst agents emitting typed assessments, a deterministic recall-focused fusion function with an entertainment gate and cross-modal conflict detection, a human-in-the-loop gate in every front end, and logged per-responder dispatch, fronted by a near-real-time Live Monitor that queues alerts for confirmation. Alongside the architecture, the study contributes an explicit requirements-to-mechanism mapping, an ablation against a single-call audio-visual baseline with paired exact tests, interval-estimated detection and routing metrics on a 200-clip audio-inclusive corpus, bounds on the human-gated operating point, and the characterization of a silent client-side failure mode in which rate-limit errors are

mistaken for benign verdicts, a hazard expected to recur in any safety pipeline built on a pay-per-use VLM.

The rerun does not confirm the original finding that fusing the modalities outperforms either alone; the two-analyst design is justified by the evidence and the conflict signal it exposes rather than by a detection gain. The conflict gate, though necessary, withholds 1% of what fusion catches in unsupervised operation, which makes the operator's review queue the essential component. Detection quality is decided by the perception configuration and its fit to the corpus rather than by the loop: on the same film-derived clips, fused recall is 0.00 under the production prompt and 0.95 under an evaluation prompt that judges footage at face value. Future priorities follow directly: a production prompt that separates sanctioned entertainment from genuine incidents without relying on film cues, validated on real surveillance footage; completing the recall-oriented prompt comparison at full scale; extending evaluation to RWF-2000 [19], UCF-Crime [17], and real surveillance footage, beginning with a smart-city case study in Casablanca, Morocco, conducted under Moroccan personal-data law (Law 09-08, under the supervision of the CNDP, the national data-protection authority); prompt engineering validated at scale [3], [20]; widening the routing vocabulary to the corpus categories; replacing the pixel-difference gate with a hybrid trigger that runs a lightweight CNN on or near the camera, such as [14], to cut both cost and latency; passing on an explicit "unknown" state in place of the benign fallback; real PSAP and CAD integration addressing the GDPR and EU AI Act gaps identified in [22]; protection against deliberate attacks such as [21]; a user study of the human review queue itself; and scaling out to many cameras.

REFERENCES

- [1] S. Yao, B. Rahimi Ardabili, A. Alinezhad Noghre, C. Neff, and H. Tabkhi, "Real-world community-in-the-loop smart video surveillance: A case study at a community college," 2023, arXiv:2303.12934.
- [2] Office of the New York City Comptroller, "Audit report on the New York City Police Department's ShotSpotter gunshot detection system," New York, NY, USA, 2024.
- [3] S. Zou et al., "Unlocking vision-language models for video anomaly detection via fine-grained prompting (ASK-Hint)," 2025, arXiv:2510.02155.
- [4] Agno (agno-agi), "Agno: An open-source Python framework for building multi-agent systems," v2.5, 2025. [Online]. Available: <https://github.com/agno-agi/agno>
- [5] Google DeepMind, "Gemini 3.7 Flash model card," Aug. 2026. [Online]. Available: <https://deepmind.google/models/model-cards/gemini-3-7-flash/>
- [6] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Gener. Comput. Syst.*, vol. 135, pp. 364-381, 2022.
- [7] A. Kalyuzhnaya et al., "LLM agents for smart city management: Enhancing decision support through multi-agent AI systems," *Smart Cities*, vol. 8, no. 1, art. 19, 2025.
- [8] X. Li et al., "DispatchMAS: Fusing taxonomy and artificial intelligence agents for emergency medical services," *BMC Emerg. Med.*, 2026.
- [9] Y. Shao et al., "EventVAD: Training-free event-aware video anomaly detection," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2025, arXiv:2504.13092.
- [10] M. Ye, W. Liu, and P. He, "VERA: Explainable video anomaly detection via verbalized learning of vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, arXiv:2412.01095.
- [11] Z. Yang et al., "AssistPDA: An online video surveillance assistant for pedestrian detection and anomaly analysis," 2025, arXiv:2503.21904.
- [12] W.-D. Jiang, C.-Y. Chang, and D. S. Roy, "Detection, retrieval, and explanation unified: A violence detection system based on knowledge graphs and graph attention networks," 2025, arXiv:2501.06224.
- [13] S. Jung et al., "Short-window sliding learning for real-time violence detection via LLM-based auto-labeling," in *Proc. IEIE*, 2025, arXiv:2511.10866.
- [14] M. Salman et al., "An embedded deep learning framework for real-time violence detection and alert generation," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-44939-x.
- [15] S. Thuau, S. Haidar, and R. Chelouah, "Frugal federated learning for violence detection: A comparison of LoRA-tuned VLMs and personalized CNNs," 2025, arXiv:2510.17651.
- [16] S. Thuau, S. Haidar, and R. Chelouah, "Federated learning for video violence detection: Complementary roles of lightweight CNNs and vision-language models for energy-efficient use," in *Proc. IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, 2025, arXiv:2511.07171.
- [17] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6479-6488.
- [18] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 322-339.
- [19] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large-scale video database for violence detection," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2020, pp. 4183-4190.
- [20] S. Yao et al., "Are multimodal LLMs ready for surveillance? A reality check on zero-shot anomaly detection in the wild," 2026, arXiv:2603.04727.
- [21] M. Rahman, S. Ahmed, and M. Hasan, "Towards evaluating robustness of violence detection in videos using cross-domain transferability," *J. Inf. Secur. Appl.*, vol. 77, art. 103583, 2023, doi: 10.1016/j.jisa.2023.103583.
- [22] E. Veltmeijer and C. Gerritsen, "Legal and ethical implications of AI-based crowd analysis: The AI Act and beyond," *AI Ethics*, 2025, doi: 10.1007/s43681-024-00644-x.
- [23] P. Choure and S. Prajapat, "Agentic AI for emergency response and comparative analysis of SmolAgents, LangGraph, AutoGen, Agno AGI and CrewAI for crisis solution," in *Lect. Notes Netw. Syst.*, 2026, doi: 10.1007/978-3-032-16791-0_10.
- [24] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, arXiv:2302.04761.
- [25] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, 2024, CELEX:32024R1689.
- [26] S. Dooley et al., "Robustness disparities in face detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, 2022, arXiv:2211.15937.