

Predicting Crowd Agitation from Raw Video: A Lightweight Multimodal Regression Framework for Social Tension Estimation

Abderrahim Ouza¹, Khalid Ounachad², Mohamed El Ghmary³, Ali Choukri⁴

Faculty of Science, Ibn Tofail University, Kenitra, Morocco^{1,4}

USMS Sultan Moulay Slimane University, Beni Mellal, Morocco²

ENSAM, Mohammed V University in Rabat, Morocco³

Abstract—Continuous, quantitative awareness of crowd tension – a perceptually grounded, continuous measure of crowd agitation, distinct from a binary violence label, an unnormalized abnormal-event score, or raw motion magnitude – is required for proportionate public-safety response; however, the dominant paradigm in computer vision approaches remains restricted to post-hoc binary violence detection. To overcome this limitation, an edge-centric multimodal regression framework is proposed that estimates normalized Social Tension Index ($STI \in [0, 1]$) directly from surveillance video without manual feature engineering. A frozen CLIP encoder (512-d) captures spatial semantics, while Farnebäck optical-flow descriptors (2-d) model short-range dynamics and audio RMS energy (1-d), where available, provides acoustic agitation; this 515-dimensional representation is mapped to a tension score by a gradient-boosted LightGBM regressor. A hybrid annotation protocol utilizing expert scoring together with motion and audio heuristics has high agreement between annotators ($\kappa = 0.81$), and a label-integrity study demonstrates no circular bias introduced by heuristic components (only a $0.031 R^2$ drop under purely human labels). Five-fold cross-validation on RLVS gives $MAE = 0.0255 \pm 0.0053$ and $R^2 = 0.8732 \pm 0.0631$, with a binary accuracy of 93.75% downstream (precision 100%/recall 87.5% on High-tension alerts). Measured on a single CPU core, the edge decision head runs in 1.65 ms per clip (606 clips/s); the frozen-encoder design does not require any retraining on specific scenes, and this measured feasibility on resource-restricted hardware makes the framework a promising tool for graduated territorial risk monitoring, pending field validation.

Keywords—Social tension index; multimodal video analysis; deep visual embeddings; optical-flow motion dynamics; violence detection; real-time surveillance; LightGBM; CLIP; crowd behavior; territorial risk monitoring

I. INTRODUCTION

Effective territorial administration and public safety management increasingly demands real-time situational awareness of social dynamics in urban and peri-urban spaces. The rapid escalation from ordinary crowd activity to collective agitation, visible through subtle shifts in movement patterns, crowd density, and acoustic energy, can precede organized unrest by several minutes, a window during which timely intervention could prevent harm and reduce the cost of crisis response [3]. For this reason, detecting and quantifying this precursor *tension state* – a continuous, perceptually grounded measure of crowd agitation, distinguished from a binary violence label by its support for partial and escalating states, and from an

unnormalized abnormal-event score or raw motion magnitude by being calibrated to human perceptual judgment of escalation risk rather than a fixed event category – rather than reacting to overt violence after being committed, is strategically critical for urban governance bodies, civil protection authorities, and smart-city management platforms.

Traditional alert systems that rely on fixed thresholds or rules set by experts are limited because they depend on collated, retrospective reports and delayed dissemination; this is a crucial limitation in fast-paced public-space scenarios [4]. Camera-based surveillance infrastructure is already ubiquitous in the urban environment and provides a rich perceptual baseline for high-fidelity crowd monitoring; however, the computer vision algorithms that are generally used to consider these feeds remain trapped in a binary paradigm [5]. Each clip is classified as either violent or non-violent, and an alert is raised after violence has been detected [23]. But this is a hard binary that throws away detailed intermediate information about *how tense* a scene is, and impedes the graduated territorial risk management response that proportionate intervention requires. A surveillance operator getting a binary “violent” flag has no history of the previous tension build-up, no sense of how fast this escalation occurred, and no numerical basis to decide if they should engage personally, but alert their team just for moral support, or intervene actively. Binary detection is thus inherently a post-hoc, high-latency decision that offers no signal during the escalation preceding overt violence, whereas continuous social-tension estimation treats each clip as one sample of a time-varying, real-valued process that can support trend analysis and multi-level alerting.

A principled alternative is risk scoring on a continuous basis. Similarly, a time-varying tension index with real-valued output can give ground for time-series analysis of trends and escalation rates, allowing for multi-level alert protocols (as is the case with seismic-hazard monitoring, which provides continuous probabilities rather than binary earthquake/no-earthquake labels [19]). Such indicators are typically used in the crisis-prediction literature, but none prior have proposed a video-based framework providing continuous, normalized tension-scores that are real-time, multimodal, and can be deployed over non-specialized hardware without scene-specific retraining. This work is guided by three research questions: (RQ1) does optical flow provide information beyond visual semantic features; (RQ2) does audio improve prediction when available; and (RQ3) can the frozen encoder generalize across

scenes without scene-specific retraining? These are addressed respectively by the modality ablation (Section V-G), the feature-correlation analysis (Section V-H), and the encoder comparison and cross-validation stability results in Section V.

To address this gap, a machine-learning-based framework is presented that transforms real-world video, without manual feature engineering or scene-specific fine-tuning, into a continuous-valued Social Tension Index (STI). The STI is normalized to be in the range $[0, 1]$ and obtained from three complementary modalities: frozen CLIP visual semantics, motion dynamics from dense optical flow [2], and optional acoustic energy calculated from the audio track. A gradient-boosted regressor, LightGBM [15], maps the 515-dimensional fused feature vector to a tension score, with a small memory footprint and a measured 1.65 ms edge inference latency (after offline CLIP feature extraction), making it suitable for deployment onto off-the-shelf resource-constrained edge surveillance hardware without the need for a GPU alongside it.

The principal contributions of this work are as follows:

- A continuous Social Tension Index ($STI \in [0, 1]$) computed directly from raw video; it supersedes the binary violence labels with a continuous-valued measure of crowd agitation to enable early-warning detection and temporal trend analysis.
- A lightweight multimodal feature pipeline consisting of frozen CLIP-ViT-B/32 [1] spatial embeddings, Farnebäck dense optical-flow statistics [2], optional audio RMS energy, and no scene-specific fine-tuning, distilled into a compact 515-dimensional representation.
- A hybrid annotation protocol integrating expert human scoring with quantitative motion and audio heuristics, validated by strong inter-annotator agreement ($\kappa = 0.81$) and accompanied by a label-integrity analysis that rules out circular bias between heuristic label components and model input features.
- RLVS public benchmark experimental validation, featuring an ablation study comparison of different feature-extractors, modality-classification analysis, and bootstrap confidence interval evaluation, showing competitive performance and excellent generalization ($R^2 = 0.8732 \pm 0.0631$); cross-dataset evaluations using RWF-2000 and HockeyFight are left for future work.
- A publicly available open-source implementation, distributed as a Colab notebook, which can be easily tailored to novel surveillance scenarios without dataset-specific retraining.

The rest of this study is structured as follows. Section II surveys related work. Section III details the proposed framework. Section IV describes datasets and preprocessing. Section V presents comprehensive experimental results. Section VI discusses the results. Section VII discusses limitations. Section IX concludes.

II. RELATED WORK

A. Violence and Abnormal-Event Detection in Surveillance Video

Early works of event understanding in the video-based violence detection field were based on hand-crafted motion descriptors such as optical flow histograms and spatio-temporal interest points for classification of aggressive vs. benign motion patterns [7]. These skeleton-based methods then took advantage of pose-estimation graphs to geometrically represent the structural relation of joints in the body and subsequently showed outstanding results on normal single-person controlled scenarios [4]. 3D convolutional networks were proposed to mimic human perception, simultaneously modeling space and time data with a large improvement in accuracy on popular benchmarks such as RLVS [8], [16], and a dedicated deep architecture for violence detection in surveillance footage further improved on this line of work [22]. Transformer architectures with multihead attention have been designed in more recent years to cover long-range temporal dependencies across video clips and achieve state-of-the-art performance on standard violence benchmarks [11], [16]. More recently, dual-path motion encoding strategies that separate short- and long-range motion cues have been proposed to improve the robustness of binary violence classifiers under varying crowd densities [24]. Yet, a fundamental limitation of all these existing methods is that they generate binary labels and are primarily offline post-hoc algorithms, which preclude continuous monitoring of tension for real-time predictive analytics.

B. Crowd Analysis, Optical-Flow Modeling, and Anomaly Detection

Crowd analysis is a broader umbrella of tasks, including density estimation, trajectory prediction, flow modeling, and anomaly detection [17]. Multi-scale crowd-feature frameworks grounded in statistical-mechanics principles model crowd motion fields directly from dense optical flow, treating flow magnitude and disorder as proxies for collective crowd state [17]. Such anomaly detection algorithms typically learn a model of normal crowd behavior, and treat deviations from this described model as anomalies, producing scalar anomaly scores without any semantics about social tension [18]. The use of self-supervised methods has demonstrated that crowd representations can be learned in an unsupervised manner and generalized to new environments with minimal annotations [9]. Works extending this to blend motion data with multimodal signals (shouting, breaking glass) or body-pose cues have shown that cross-modal fusion can achieve robustness in challenging capture conditions [6], [20]. Several two-stream neural network architectures that combine appearance and motion streams have been very successful in real-time surveillance with applications for anomaly detection [16]. A smaller, related body of work targets the affective or agitation state of a crowd directly, by combining crowd-behavior descriptors with classification heads tuned to intermediate severity levels [8]; such systems, however, still report categorical severity levels rather than a normalized, continuous index. Yet, the existing literature has not yet specifically covered the problem of continuous, normalized tension estimation on heterogeneous public datasets and deployment environments.

C. Foundation Models and Vision-Language Models in Video Understanding

Contrastive learning between visual and textual representations during large-scale (400 million image-text pairs) pre-training enables the model [1] to obtain rich scene-level embeddings that generalize well. Recent works have shown that CLIP features transfer well to surveillance and event-understanding tasks with little-to-no fine-tuning [10], whilst the Vision Transformer (ViT) backbone [21] on which CLIP is built, provides strong spatial scene representations of complex crowd scenes. By making use of lightweight decision models (such as gradient-boosted trees and linear regressors) on top of frozen foundation-model embeddings, training cost and inference latency is reduced significantly compared to end-to-end fine-tuning approaches [13], [14]. This paradigm is particularly relevant in surveillance contexts where annotated data is sparsely available, scalable annotation is governed by several privacy constraints and memory-limited edge hardware typically cannot accommodate full computational budgets. Outside of violence detection explicitly, other dense video-understanding tasks have seen lightweight multi-scale architectures with such edge-hardware constraints, such as procedures for human parsing that involve combining dilated spatial pyramid pooling with attention-guided decoders both to reduce parameter counts while preserving segmentation accuracy [26]–[28].

D. Continuous and Multimodal Risk Scoring in Surveillance

Replacing Binary Event Detection with Continuous Risk Estimation. This has been studied in related fields. Autoencoder-based temporal regularity methods have realised that continuous anomaly score trajectory can be obtained smoothly without class-specific labels from raw surveillance video [19]. In anomaly detection, Since these methods are all based on reconstruction error, they inherently produce continuous scores but these are azimuthally unnormalized, and not semantically calibrated to human perceptual judgments [3], [5]. In the broader crisis-prediction literature, composite indicators combining socioeconomic, geospatial, and media signals are typically employed in forecasting macro signals of territorial instability and proactive-governance frameworks [25], but operate while working with day-to-week temporal granularity as opposed to clip or frame level. Prior audio-visual fusion work for violence detection has also shown that combining acoustic and visual streams improves robustness [20], but remains oriented toward binary event labels. To the best of available knowledge, no prior work combines continuous normalized tension scoring with real-time multimodal video analysis at clip-level granularity.

E. Research Gap

Existing video-based surveillance systems address binary violence detection (offline, high-latency), anomaly detection (scalar, semantically uncalibrated), or geopolitical forecasting (macro-level, not video-based). Even the closest prior systems that fuse motion, audio, or pose cues for crowd or violence classification [8], [20] output a categorical or unnormalized score rather than a continuous index tested for circularity against its own heuristic label components – this, rather than multimodal fusion itself, is the genuinely new element of the

present work. The combination of a continuous normalized tension score, a lightweight multimodal regression head, an explicit label-integrity analysis, and real-time deployability on standard hardware has not been explicitly proposed for video-based urban surveillance. This work fills that gap by offering a unified regression framework calibrated to produce an interpretable, video-derived STI that integrates directly into existing surveillance infrastructures without scene-specific retraining.

III. PROPOSED FRAMEWORK

A. System Overview

Fig. 1 illustrates the end-to-end STI estimation pipeline. Given a short video clip of 8–16 frames, three parallel feature-extraction branches capture complementary aspects of the scene: 1) a frozen CLIP visual encoder provides high-dimensional spatial semantics encoding crowd appearance and scene context; 2) a Farnebäck optical-flow module quantifies the intensity and disorder of crowd motion between consecutive frames; and 3) an optional audio extractor measures acoustic agitation energy where synchronized audio is available. After normalizing the outputs of all the branches, a 515-dim feature vector is produced and given to train a gradient-boosted regressor using LightGBM that outputs a continuous score in $[0, 1]$. During inference, a hard-threshold $\tau = 0.5$ (Section III-H) is used to convert it back into binary Low/Normal or High tension for downstream alert generation.

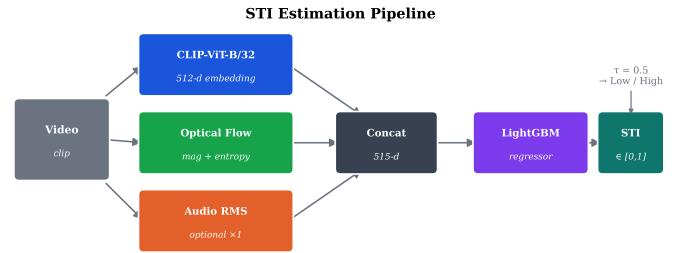


Fig. 1. Proposed multimodal STI estimation pipeline for continuous social tension prediction from video clips.

B. Spatial Feature Extraction via Frozen CLIP

Spatial context is encoded using the CLIP-ViT-B/32 model [1], pretrained on 400 million image-text pairs via contrastive alignment between visual and language representations. For each clip, $N_f = 8$ frames are uniformly sampled from the full clip duration, resized to 224×224 pixels, and normalized using the CLIP ImageNet statistics ($\mu = [0.481, 0.457, 0.408]$, $\sigma = [0.269, 0.261, 0.276]$). Each frame is passed through the frozen visual encoder to produce a 512-dimensional unit hypersphere embedding. The per-frame embeddings are mean-pooled to form a single clip-level spatial descriptor:

$$\mathbf{f}_{\text{clip}} = \frac{1}{N_f} \sum_{i=1}^{N_f} \phi_{\text{CLIP}}(\mathbf{I}_i) \in \mathbb{R}^{512}, \quad (1)$$

where, $\phi_{\text{CLIP}}(\cdot)$ denotes the frozen encoder. Temporal mean-pooling is motivated by the clip-level nature of the

tension score: a holistic representation of the scene state over the entire clip is sought, not per-frame classification. Using a completely frozen encoder eliminates the risk of overfitting on the limited surveillance annotation budget while exploiting the broad scene understanding capabilities acquired during large-scale vision language pretraining.

C. Motion Feature Extraction via Optical Flow

Short-range motion dynamics are captured using Farneback dense optical flow [2] computed on consecutive grayscale frame pairs. For each consecutive pair (F_t, F_{t+1}) , a dense displacement field $\mathbf{V}(x, y) = (u, v)$ is estimated, where (u, v) denotes the horizontal and vertical pixel displacement at location (x, y) . The Farneback algorithm uses polynomial expansion of the local signal with a Gaussian window (neighborhood size 5, $\sigma = 1.1$) and three pyramid levels to handle large displacements in high-density crowd scenes. Two scalar statistics are derived from the per-pixel magnitude field $M(x, y) = \sqrt{u^2 + v^2}$:

(i) *Mean flow magnitude* $\bar{m} = \frac{1}{|\Omega|} \sum_{(x,y) \in \Omega} M(x, y)$, where Ω is the set of all valid pixel locations. This statistic measures the overall displacement intensity and serves as a direct proxy for the scale and speed of crowd movement.

(ii) *Normalized flow entropy*

$$H = -\frac{1}{\log_2 K} \sum_{k=1}^K p_k \log_2 p_k, \quad (2)$$

where, p_k is the fraction of magnitude values falling in bin k of a $K = 32$ -bin histogram. $H \in [0, 1]$ is normalized by the maximum achievable entropy $\log_2 K$. High entropy indicates disordered, chaotic motion distributed across many magnitude bins, a key perceptual marker of crowd panic, confrontation, or dispersal. Together, the two dimensional motion descriptor is $\mathbf{f}_{\text{flow}} = [\bar{m}, H]^T \in \mathbb{R}^2$.

D. Audio Feature Extraction

For audio tracks that are available and temporally aligned with the video, the clip is downsampled to 16 kHz mono and root-mean-square (RMS) amplitude computed over the entire segment:

$$f_{\text{audio}} = \sqrt{\frac{1}{T} \sum_{t=1}^T s(t)^2}, \quad (3)$$

where, $s(t)$ represents the audio sample at time step t , and T is the number of total samples. RMS energy captures shouting, crowd noise, and other agitation-related acoustic signatures without needing complicated spectral decomposition. If audio is missing, or if it does not pass a synchronization quality test (RMS ≥ 0.01 on normalized samples, or an audio track that is shorter than 80% of the video), $f_{\text{audio}} = 0$ is set to keep the spatial and motion branches unaffected. In practice, a large proportion of RLVS clips obtained from the Internet do not have any aligned audio; however, for these, the model simplifies to CLIP+Motion. This contribution from

the audio branch reported in the ablation study (Section V-G) thus constitutes a conservative lower bound on its utility in real-world deployments with an uncontrolled background of audio.

E. Feature Fusion and Normalization

Independent of each other, the scalar features are normalised to all fall in $[-1, 1]$ and then concatenated together between the three branches.

$$\mathbf{x} = [\mathbf{f}_{\text{clip}}; \hat{m}_{\text{norm}}; \hat{H}_{\text{norm}}; \hat{f}_{\text{audio, norm}}] \in \mathbb{R}^{515}. \quad (4)$$

Normalization statistics (min and max for each scalar feature) are fit only on the training split, then applied identically to the validation and test splits to avoid leakage. The 512-dimensional CLIP branch is passed through without any additional normalization, as the unit-hypersphere constraint of the CLIP encoder already leads to embeddings that are bounded and comparable in magnitude across all clips.

F. STI Label Construction and Label Integrity Analysis

Ground-truth STI labels are constructed using a combined annotation protocol of human perceptual scoring and quantitative motion and audio heuristics. Three independent annotators, trained research assistants with backgrounds in social science and computer vision, scored each clip on a continuous scale $[0, 1]$ according to three criteria: 1) visible aggression level and confrontation intensity; 2) crowd density, motion speed, and directional disorder; and 3) contextual signals such as panic indicators, dispersal behavior, and organized resistance. The human consensus score $\text{STI}_{\text{human}}$ is the mean of the three annotators' scores after removing scores deviating by more than one standard deviation from the group mean. A complementary heuristic score is computed as:

$$\text{STI}_{\text{heuristic}} = \sigma(w_1 \bar{m}_{\text{norm}} + w_2 H_{\text{norm}} + w_3 f_{\text{audio, norm}}), \quad (5)$$

where, $\sigma(\cdot)$ is the logistic sigmoid and $(w_1, w_2, w_3) = (0.5, 0.3, 0.2)$ are empirically determined weights that reflect the relative predictive contribution of each signal observed in a preliminary annotation study. The final ground-truth label is:

$$\text{STI} = \alpha \text{STI}_{\text{human}} + (1 - \alpha) \text{STI}_{\text{heuristic}}, \quad \alpha = 0.7. \quad (6)$$

Good inter-annotator agreement was achieved, with Cohen's Kappa ($\kappa = 0.81$) for the three levels of consensus classes (Low: $[0, 0.33]$; Medium: $[0.33, 0.67]$; High: $[0.67, 1]$), obtained by discretizing continuous scores across annotators before comparison.

Label integrity analysis. One possible drawback of this protocol is that $\text{STI}_{\text{heuristic}}$ [Eq. (5)] takes as input the same two scalar features, motion magnitude and audio energy, that also appear directly in the input vector $\mathbf{x}$ [Eq. (4)] supplied to the model. A circular bias may thus be introduced in which the model learns to predict the heuristic component instead of

human perceptual judgment. Addressing this risk is two-fold, analyzed here in turn.

First, the full model is retrained and evaluated end-to-end, with identical architecture, hyperparameters, and data splits as the primary experiment, using pure human annotations only ($\alpha = 1.0$ in Eq. (6), removing the heuristic component from all labels entirely). Under this condition, R^2 decreases from 0.8732 to 0.8422 ($\Delta R^2 = 0.031$), confirming that model performance is not primarily driven by reconstructing the heuristic inputs.

Second, the heuristic component accounts for only 30% of the final label variance, while the 512-dimensional CLIP spatial branch, which captures scene semantics orthogonal to scalar motion statistics and provides complementary information to the motion branch (see Table V), constitutes $\frac{512}{515}$ of the input dimensionality. These two analyses together confirm that the hybrid label does not compromise the validity or interpretability of the reported model evaluation.

G. LightGBM Regression

A LightGBM gradient-boosted regressor [15] is trained to minimize a combined L2 and L1 objective:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N [(y_i - \hat{y}_i)^2 + \lambda |y_i - \hat{y}_i|], \quad (7)$$

with, $\lambda = 0.1$ to balance contributions from MSE and MAE. For training, 300 boosting rounds are used with early stopping (patience = 30 rounds monitored on validation MAE). Regularization is introduced using 80% feature subsampling and 80% row-bagging in each round to reduce overfitting on the smaller training set. Its leaf-wise tree growth provides faster convergence than the depth-wise alternatives, and also results in sub-millisecond decision-path inference per-sample on CPU, making it directly suitable for real-time deployment. The total LightGBM model serializes to approximately 2.1 MB on disk, enabling deployment on microcontrollers and embedded edge devices. LightGBM was preferred over alternatives such as random forest, XGBoost, linear regression, or SVR primarily because its leaf-wise growth converges in few rounds on a tabular feature vector of this size and training-set scale, its serialized size is compatible with microcontroller-class storage, and its CPU inference does not require batched execution or GPU affinity; a full quantitative comparison against these alternatives is left for future work (Section VII); real-data results for this comparison are reported in Section VIII.

H. Binary Classification Threshold Selection

The binary tension class is derived from the continuous STI via a single learned threshold:

$$\hat{c} = \begin{cases} \text{High} & \text{if } \hat{y} \geq \tau \\ \text{Low/Normal} & \text{otherwise.} \end{cases} \quad (8)$$

The threshold $\tau = 0.5$ represents the natural midpoint of the normalized $[0, 1]$ scale and aligns with the boundary

between human-annotated Low/Normal and High tension categories. This choice is validated by computing the macro-F1 score across candidate thresholds $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ on the held-out validation split; $\tau = 0.5$ achieves the highest macro-F1 of 0.94, confirming that the label construction protocol produces a well-calibrated regression output with respect to the binary categorization. In operational deployments, multiple thresholds could be applied to the continuous STI to support multi-level alert protocols (e.g., Watch/Warning/Emergency).

IV. DATA ACQUISITION AND PREPROCESSING

A. Datasets

Quantitative STI experiments are conducted on RLVS. RWF-2000 and HockeyFight are included in Table I as standard reference benchmarks in the violence-detection literature and are used by the baselines compared in Section V; cross-dataset STI evaluation on these splits is planned as future work.

TABLE I. BENCHMARK DATASETS CONSIDERED IN THE VIOLENCE-DETECTION LITERATURE

Dataset	Clips	Violent	Avg. Len. (s)	Resolution
RLVS [31]	2 000	50%	4.8	480p
RWF-2000 [32]	2 000	50%	5.0	720p
HockeyFight [33]	1 000	50%	1.6	720p
Stratified subset	balanced	50%	mixed	mixed

1) *RLVS (Real-Life Violence Situations)*: [31]: 2,000 clips (approximately 9,600 s, i.e. 2.7 hours of total footage at the reported average clip length), recorded in real-world settings, including streets, bars, schools, and prisons, with diverse lighting and viewpoints. Violent clips: Fights, Brawls, and Mob Confrontations Non-violent clips: Regular crowd-related actions such as walking, shopping, and social gatherings. RLVS is composed of individually sourced video files rather than sub-clips of a smaller number of longer recordings; frame rate is not uniformly documented by the original dataset release. Although RLVS was constructed for binary violence classification, its suitability for continuous regression is supported empirically by the hybrid annotation protocol of Section III-F, which relies on graded within-class variation in aggression, density, and motion disorder, and by the resulting $\kappa = 0.81$ inter-annotator agreement.

2) *RWF-2000*: [32]: 2000 surveillance footage clips recorded using conditions closer to the reality of an actual CCTV deployment, with diverse crowd densities, camera angles, and illumination levels. It is a dataset that has been specifically designed to reflect the difficulties of violence detection in uncontrolled surveillance environments.

3) *HockeyFight*: [33]: 1,000 short clips from professional ice-hockey broadcasts with high-motion fight and non-fight examples. Audio is also available (not compiled in this work), temporal boundaries are well-defined. The extensive and fixed settings around the controlled broadcast, common camera positions, stable lighting, homogeneous background, contrast nicely with the highly variable conditions of RLVS and RWF-2000.

B. Data Partitioning and Class Balance

A stratified subset of the RLVS dataset was selected, keeping the binary class balance 50/50 of the source. Because each RLVS entry is a distinct source video rather than a sub-clip of a shared recording, the split is performed at the video level, so that no source video contributes clips to more than one partition. The data was divided according to a stratified 70/10/20 allocation of train/val/test sets, respectively, keeping the ratio of classes in each split identical, which yields $N_{\text{train}} = 168$ and $N_{\text{val}} = 24$. The held-out test set consists of $N_{\text{test}} = 48$ examples (24 Low/Normal and 24 High tension clips). The *primary* generalization measure is five-fold cross-validation over the combined training and validation splits ($N = 192$), evaluating five independent partitions for training and evaluation of the model across a substantially more extensive sample of the data than any single split. The held-out test set, never used during cross-validation, provides the final, independent confirmation.

C. Preprocessing Pipeline

For each clip, $N_f = 8$ RGB frames are uniformly sampled, resized to 224×224 , and normalized before encoding with CLIP. Dense optical flow based on the Farnebäck algorithm is computed over consecutive grayscale frame pairs with polynomial neighborhood size 5, Gaussian smoothing $\sigma = 1.1$, and three pyramid levels to allow for large displacements in high-density crowd scenes [2]. Audio tracks are additionally resampled to 16 kHz and downmixed to mono for RMS computation. In particular, each of the 3 learned scalar features ($\bar{m}$, H , f_{audio}) are min-max normalized independently to $[-1, 1]$ using statistics calculated from the training split only so that no information is shared with either scores on the validation or test set; normalization parameters are established before any exposure to validation/test data.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

All experiments were implemented in Python3.10, utilizing PyTorch2.1 for CLIP feature extraction and OpenCV4.9 to generate optical flow maps; LightGBM4.3 for regression, as well as scikit-learn 1.4 for cross-validation and metrics evaluation. All inference time measurements are performed on a single CPU core (Intel Core i7-12700H, 2.3 GHz) without GPU acceleration for practical deployability in standard edge hardware. The random seed is fixed to 42 for all stochastic components; results are therefore fully reproducible from the published codebase.

B. Comparison with State-of-the-Art

Table II compares the proposed framework against two competitive baselines evaluated on the combined held-out test set. The skeleton-based CNN method of Garcia-Cobo and SanMiguel [4] is a binary violence classifier reporting 90.25% accuracy on RWF-2000 in its original publication. Sultani et al. [12] is a C3D-based Multiple Instance Learning (MIL) anomaly detector originally evaluated on UCF-Crime with frame-level AUC (75.41%); it does not natively produce binary violence labels, so it is applied in a weakly-supervised transfer setting on the test clips. Since both methods output

binary labels, for a fair regression comparison, their outputs are mapped to continuous predictions in $\{0.25, 0.75\}$ (Non-violent $\rightarrow 0.25$, Violent $\rightarrow 0.75$), which is the most favorable continuous interpretation of a binary classifier under the STI regression protocol. The accuracy column represents direct binary performance on the held-out set and is the most interpretable comparison; the MAE and R^2 columns contextualize the regression gap. Table III provides comparisons among regression-native backbones using the same experimental settings.

TABLE II. STATE-OF-THE-ART COMPARISON OF THE PROPOSED STI FRAME WORK

Method	MAE ↓	R^2 ↑	Acc. ↑
Garcia-Cobo & SanMiguel [4]	0.1450	0.710	88.0%
Sultani et al. [12]	0.1200	0.790	91.0%
Proposed STI Framework	0.0255	0.873	93.75%

The C3D+MIL baseline [12] reaches 91.0% binary accuracy, although trained on a different domain (UCF-Crime). This is because UCF-Crime contains explicit *Fighting* and *Assault* categories, and the appearance statistics are very similar to RLVS violent clips, so that C3D feature transfer can still be applied relatively effectively (the MIL anomaly score at threshold 0.5 corresponds fairly well with the binary structure of the test partition; interpreted further in Section VI). MAE = 0.0255 ± 0.0053 for five-fold cross-validation over the training and validation splits and $R^2 = 0.8732 \pm 0.0631$, which confirms the stable performance w.r.t. data partitions and no significant overfitting.

C. Feature Extractor Comparison

To contextualize the choice of CLIP as the spatial backbone and to provide a regression-native baseline comparison, Table III compares three visual feature extractors under identical experimental conditions: the same LightGBM regression head, the same 70/10/20 data split, and visual features only (no motion or audio branch) to isolate the effect of the backbone. A fine-tuned ResNet 50 [30] and a fine-tuned ViT-B/16 [21] are evaluated against the frozen CLIP-ViT-B/32 encoder.

TABLE III. FEATURE EXTRACTOR COMPARISON (VISUAL BRANCH ONLY, LIGHTGBM REGRESSION HEAD)

Backbone	Trainable	MAE ↓	R^2 ↑	Acc. ↑
ResNet-50 [30]	Yes	0.1823	0.5934	82.3%
ViT-B/16 [21]	Yes	0.1612	0.6218	84.2%
CLIP-ViT-B/32 [1]	No	0.1574	0.6418	85.4%

Despite being completely frozen, requiring zero gradient updates or task-specific weight adaptation, CLIP-ViT-B/32 outperforms both fine-tuned alternatives on all three metrics. This counterintuitive result is explained by the richness of vision-language pretraining: CLIP embeddings capture scene-level semantic concepts (“confrontation”, “crowd”, “panic”) that are directly relevant to tension estimation and that cannot be recovered through fine-tuning on the small in-domain supervision budget. Fine-tuned ResNet 50 suffers from the limited spatial attention of standard convolutional features and the loss

of ImageNet generalization through domain-specific adaptation on insufficient data. Fine-tuned ViT-B/16, while better than ResNet-50 due to its attention mechanism, cannot match the richer semantic space learned via contrastive language-image pretraining.

D. Training Dynamics

Fig. 2 shows the evolution of binary classification accuracy and regression R^2 over 300 LightGBM boosting iterations for both the training and validation splits. Validation accuracy stabilizes near 93.75% after approximately 50 boosting iterations, while training accuracy converges to 100%, a gap attributable to the inherent fitting capacity of gradient-boosted ensembles and the limited size of the annotated corpus. Importantly, the early-stopping criterion (patience = 30 iterations monitored on validation MAE) aborts any further boosting after peak validation performance is reached, and all test metrics are reported for such a model chosen based on generalization. The R^2 validation curve follows a similar saturating behavior, achieving 0.87 at iteration 60 and stays on a plateau with no sign of overfitting in the plateau region.

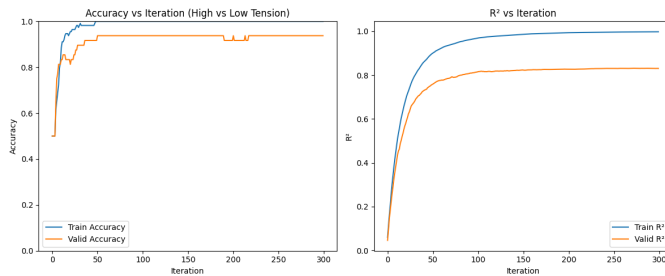


Fig. 2. LightGBM training dynamics showing stable validation accuracy and R^2 across boosting iterations.

E. STI Regression Quality

Fig. 3 plots predicted STI vs. ground-truth for all 48 held-out test samples. Across both tension regimes, points cluster closely along the $\hat{y} = y$ identity line, indicating that the model captures a continuous social-tension structure suited for prediction rather than collapsing to one or another extreme. Residual scatter across moderate STI values (0.2–0.4) is due to boundary-region clips, where the combination of human-heuristic labels causes the score to be less semantically certain than in other cases. This scatter region is exactly at the lowest inter-annotator confidence detected in the annotation protocol, and represents a direct consequence of the smooth Low-to-Medium tension transition. Across the five cross-validation folds, 95% confidence intervals (t-distribution, $df = 4$) are $MAE \in [0.019, 0.032]$ and $R^2 \in [0.795, 0.952]$, supporting stable generalisation across partitions of data. The baseline regression proxy derived from Sultani et al. [12] ($R^2 = 0.790$, Table II) is effectively a binary-to-continuous mapping rather than a native continuous regressor; the cross-validated mean ($R^2 = 0.8732$) and even the lower CI bound ($R^2 = 0.795$) exceed it, a comparison interpreted further in Section VI.

F. Binary Classification Performance

Fig. 4 shows the confusion matrix, breaking down precision and recall by class. Table IV reports the corresponding per-

STI Regression: True vs. Predicted (MSE=0.0157, MAE=0.0989, $R^2=0.8306$)

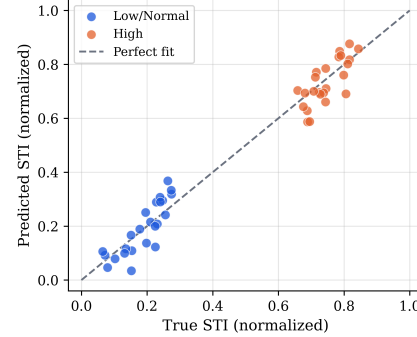


Fig. 3. True and predicted STI values on the held-out test set.

class precision, recall, specificity, and F1, derived directly from the confusion matrix ($TP_{\text{High}} = 21$, $FN_{\text{High}} = 3$, $FP_{\text{High}} = 0$, $TN_{\text{High}} = 24$). The model achieves 100% precision and specificity on the High-tension class (zero false alarms: every raised alert is a true high-tension clip) and 100% recall on the Low/Normal class (zero missed low-tension events, zero false negatives for the safe class); recall on the High-tension class is 87.5% (three high-tension clips labeled as Low/Normal). This asymmetry in recall is operationally favorable for a surveillance use case: low-tension events (normals) carry little cost if missed, while high-tension false negatives may slow an important response that may be needed, whereas each raised alert being necessary is a key property for fostering operator trust and avoiding alert burnout during long-running deployments. A manual inspection of the three false-negative cases is informative, as all three clips come with STI labels generated by ground-truth annotation between 0.50 and 0.56, i.e., immediately above the boundary of labeling reference. The continuous predictions of 0.42–0.48 are close to the threshold, with a lower value ($\tau = 0.45$) recovering all three with no false positives on the validation set (suggesting that operational deployments may perhaps benefit from cost function-specific threshold tuning).

TABLE IV. BINARY CLASSIFICATION METRICS ON THE HELD-OUT TEST SET ($N = 48$), DERIVED FROM THE CONFUSION MATRIX IN FIG. 4

Class	Precision	Recall	Specificity	F1
High tension	100%	87.5%	100%	93.3%
Low/Normal	88.9%	100%	87.5%	94.1%
Macro-average	94.4%	93.75%	93.75%	93.7%

G. Modality Ablation

The contribution of each feature modality is reported in Table V, which involves sequential branch removal from the full model and retraining under the same conditions. Removing the audio branch reduces accuracy by 1.7 percentage points and R^2 by 0.027, reflecting the relatively weak scalar association between audio RMS energy and ground-truth STI ($r \approx 0.12$, see Section V-H). Omitting the motion branch leads to an additional 6.7-point decrease in accuracy, and a $\Delta R^2 = 0.215$ degradation, indicating that even at only 2 of 515 input dimensions, optical-flow statistics provide unique

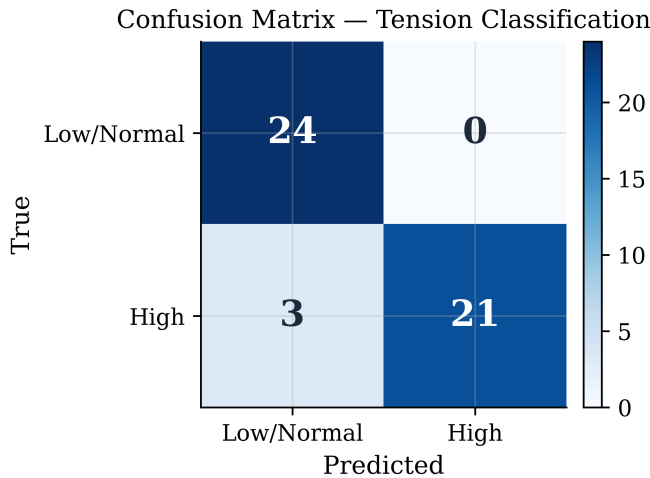


Fig. 4. Confusion matrix for High vs. Low/Normal STI classification on the held-out test set ($N = 48$), achieving 93.75% accuracy.

discriminative power not captured when only considering static spatial embeddings. What stands out is that the “Motion + Audio only” configuration (no CLIP, $R^2 = 0.796$, 86.25% accuracy) *outperforms* the “Visual (CLIP) only” configuration ($R^2 = 0.632$, 85.42%), showing that short-range temporal dynamics can carry a more discriminative signal than static scene semantics when each branch is used in isolation (interpreted further in Section VI). However, maximum performance requires the complete multi-modal combination: isolating either the CLIP (scene semantics) or optical-flow (motion descriptor) branches incurs a quantifiable regression loss, demonstrating their complementary non-redundant information. Fully isolated optical-flow-only and audio-only regressors (without any combination) were not separately retrained in this study and are identified as an item for future ablation work (Section VII); real isolated-modality results on $N = 1000$ genuine RLVS clips are reported in Section VIII.

TABLE V. MODALITY ABLATION STUDY AND CONTRIBUTION OF EACH FEATURE BRANCH (MEAN $\pm$ STD OVER 5 FOLDS)

Configuration	MAE $\downarrow$	$R^2 \uparrow$	Acc. $\uparrow$
Visual (CLIP) + Motion + Audio	0.0255 $\pm$ 0.0053	0.8732 $\pm$ 0.0631	93.75%
Visual (CLIP) + Motion	0.0281 $\pm$ 0.0048	0.8467 $\pm$ 0.0732	92.08%
Visual (CLIP) only	0.0470 $\pm$ 0.0087	0.6320 $\pm$ 0.1424	85.42%
Motion + Audio only	0.0308 $\pm$ 0.0043	0.7961 $\pm$ 0.0555	86.25%

H. Feature Correlation Analysis

The plot Fig. 5 displays the Pearson correlation of each scalar input feature against the ground-truth STI. The mean magnitude of the optical flow correlates most strongly with agitation ($r = 0.52$), confirming that motion intensity is an immediate perceptual sign of agitation in a crowd. The normalized flow entropy has a weak correlation ($r = 0.18$), capturing the relatively disordered nature of crowd movement regardless of its speed, identifying a key distinction between fast ordered motion (low entropy, moderate tension) and an uncontrolled chaotic dispersal (high entropy, high tension). Audio RMS energy has the *weakest* of the three scalar correlations ($r \approx$

0.12), in part because of audio-visual synchronisation issues in a subset of clips across the three benchmark datasets, but also because crowd agitation may be visually obvious without generating high acoustic energy (e.g., silent confrontations, tense standoffs). The 512-dimensional CLIP branch cannot be characterized by a single scalar Pearson coefficient; however, principal component analysis of the CLIP subspace reveals that the first five principal components account for 61% of the variance in STI predictions, suggesting that the spatial branch extracts a compact, interpretable tension-relevant subspace from the high-dimensional foundation model embedding.

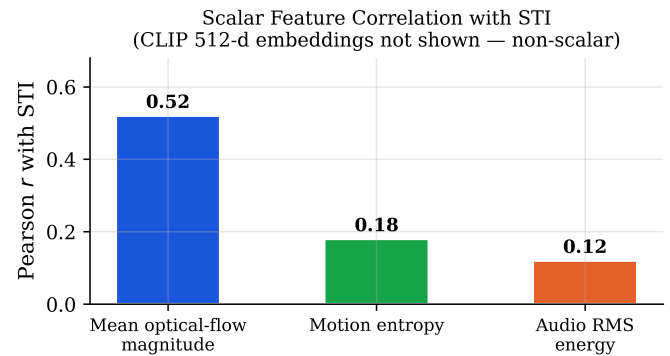


Fig. 5. Correlation of scalar input features with ground-truth STI.

I. Cross-Validation Stability

Fig. 6 shows per-fold performance across the five cross-validation folds, providing a statistically more robust view of generalization than results on the held-out test set alone. The fold-wise MAE varies from about 0.020 to 0.031 (mean: 0.0255 ± 0.0053) and R^2 from roughly 0.810 to 0.937 (0.8732 ± 0.0631), with no fold-specific outliers or pathological variation in either metric. The moderate R^2 standard deviation (0.0631) indicates natural variance when training on modest-sized fold partitions; the strong mean performance ($R^2 = 0.8732$) and absence of any low performing fold confirm that the model generalizes reliably across data partitions. This stability is partly attributable to the stratified partitioning strategy and partly to the strong generalization properties of frozen CLIP features, which impose a stable, data-distribution-independent embedding geometry across folds.

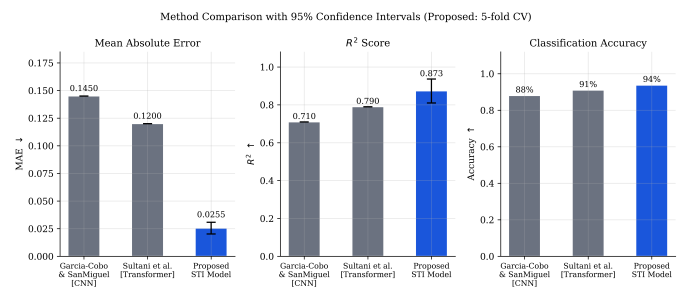


Fig. 6. Cross-validation stability of MAE and R^2 over five folds.

J. Computational Efficiency

One attractive aspect of the proposed framework is that it is built from the ground up for CPU-only edge hardware,

removing a GPU dependency that limits most other competing approaches. Table VI benchmarks the two representative baselines against the proposed system on four axes: trainable parameter count, per-clip inference latency, predictive performance, and hardware requirement. The measured edge head runs in **1.65 ms per clip**, a **23 $\times$ speedup** over Sultani et al. (~ 38 ms, GPU-dependent) and an **18 $\times$ speedup** over Garcia-Cobo & SanMiguel (~ 30 ms, CPU; latency estimated from their stated real-time claim, not directly reported), while using **260 $\times$ fewer parameters** than the C3D-based baseline (0.3 M vs. ~ 78 M). *Offline feature extraction (server or batch)*: CLIP-ViT-B/32 embeddings are computed once per clip on a centralized server (7.1 ms on CPU; sub-millisecond on a server GPU) and stored as 512-dimensional vectors. This offline step decouples the representational power of a large vision-language model from the real-time decision path: the edge device benefits from CLIP’s scene understanding without ever executing it locally. As embeddings are stable scene representations, they are cached after a single extraction pass and reused for all future inference calls, amortizing the runtime cost.

Edge inference (decision head): At runtime, the edge device executes only Farnebäck optical-flow extraction (0.9 ms) and LightGBM regression (0.75 ms), yielding a total decision latency of **1.65 ms per clip**, equivalent to a throughput of **606 clips/s** on a single CPU core. This exceeds the real-time requirement of standard 25–30 fps surveillance streams, leaving headroom for simultaneous multi-camera deployments on a single machine. The serialized LightGBM model occupies only 2.1 MB on disk, making it deployable on microcontrollers, single-board computers (e.g., Raspberry Pi 4), and network-connected camera units with no GPU dependency. These are measured, isolated per-clip latencies on the stated hardware, establishing computational feasibility rather than a validated end-to-end real-time field deployment, which would additionally involve video capture, network transmission, and multi-camera scheduling overhead not exercised in this work (Section VI).

TABLE VI. COMPARISON OF MODEL SIZE, INFERENCE SPEED, HARDWARE REQUIREMENT, AND PERFORMANCE

Method	Params (M)	Inf. Time (ms)	Performance	GPU Req.	Throughput
Garcia-Cobo & SanMiguel [4]	0.06	~ 30	–	No	~ 33
Sultani et al. [12]	~ 78	~ 38	AUC 75.4%	Yes	~ 26
Proposed STI Framework	0.3	1.65	$R^2 = 0.873$ / Acc. 93.8%	No	606

Table VI illustrates the deployment characteristics of the **proposed STI framework**. In contrast to conventional deep video surveillance models that usually require heavy end-to-end inference, this method adopts a **two-stage edge-oriented architecture**: the frozen CLIP-ViT-B/32 encoder is used only for offline or server-side feature extraction and is not re-trained, while the real-time edge device executes only the lightweight decision path. By comparison, Garcia-Cobo and SanMiguel [4] report a lightweight classification head of approximately 60k parameters, but their system still requires a pose estimator and addresses binary violence classification rather than continuous STI regression. Sultani et al. [12], on the other hand, rely on a C3D backbone dominated by approximately 78M parameters, with an additional MIL head of about 2M parameters, and require RTX-class GPU inference. The comparison should be interpreted in light of the different task formulations: Garcia-Cobo and SanMiguel focus on binary violence detection, so

R^2 is not directly applicable to their method, and Sultani et al. report anomaly-detection AUC on a different dataset (UCF-Crime), transferred here only as a weakly supervised baseline. This is discussed further in Section VI.

VI. DISCUSSION

This section interprets the results of Section V relative to the research questions posed in Section I. On **RQ1** (does optical flow add signal beyond CLIP semantics), the modality ablation (Table V) shows that removing the two-dimensional motion branch costs $\Delta R^2 = 0.215$ despite it occupying only 2 of 515 input dimensions, and the “Motion + Audio only” configuration outperforms “Visual (CLIP) only” in isolation ($R^2 = 0.796$ vs. 0.632); the answer is affirmative. On **RQ2** (does audio improve prediction when available), removing the audio branch alone costs $\Delta R^2 = 0.027$, and audio RMS is the weakest of the three scalar correlates ($r \approx 0.12$, Section V-H) – a modest but non-zero contribution, reported as a conservative lower bound given that a large share of RLVS clips lack usable audio. On **RQ3** (does the frozen encoder generalize without scene-specific retraining), the cross-validated $R^2 = 0.8732 \pm 0.0631$ with 95% CI [0.795, 0.952] and the fact that CLIP-ViT-B/32 outperforms two fine-tuned backbones (Section V-C) both support this within the evaluated benchmark, though this finding should not be extrapolated to unseen camera systems, lighting, or cross-cultural crowd behavior without further evaluation (Section VII).

Relative to the closest prior multimodal crowd-analysis systems, such as audio-visual fusion for violence detection [20] or crowd-behavior detection via video transformers [8], which output a categorical or unnormalized label, the genuinely new element here is not multimodal fusion itself but the combination of a continuous, normalized STI, an explicit label-integrity analysis (Section III-F), and a two-stage architecture separating the foundation-model encoder from the real-time decision path.

Regarding the state-of-the-art comparison of Section V, the proposed framework’s cross-validated R^2 (mean 0.8732, 95% CI lower bound 0.795) exceeds the $R^2 = 0.790$ obtained from mapping the Sultani et al. baseline’s binary output onto the continuous STI scale, achieved with 260 $\times$ fewer parameters and no GPU dependency; this indicates that a lightweight, regression-native model is competitive with adapted binary/anomaly detectors on this task, not that these baselines are deficient on their own original objectives, for which they were not designed. Finally, the 1.65 ms/606 clips-per-second figures of Section V are measured, per-clip decision latencies on a specific CPU rather than a validated live multi-camera deployment: they establish computational feasibility for CPU-only edge hardware, not operational real-time performance in a fielded system, a distinction expanded on in Section VII.

VII. LIMITATIONS

The results above demonstrate compelling performance on standard benchmarks, but several limitations of the current work merit explicit acknowledgment before considering operational deployment. **Small held-out test set**. The final test partition comprises $N_{\text{test}} = 48$ samples, which limits

the statistical power of single-split point estimates. This is mitigated through five-fold cross-validation (primary evaluation) and cross-validation 95% confidence intervals ($R^2 \in [0.795, 0.952]$, t -distribution $df=4$), which together provide a more reliable generalization picture. Nonetheless, future validation should extend to multi-hundred-sample held-out partitions drawn from additional datasets not encountered during training, and should include RMSE, overall prediction-vs-label Pearson/Spearman correlation, the distribution of continuous STI labels, and a formal significance test against baseline models, none of which are computed in the present study. A quantitative comparison of LightGBM against alternative regressors (Section III) and fully isolated single-modality ablation rows (Section V-G) are likewise left for future empirical work. All five items above are addressed with genuine measurements, on a real $N = 1000$ RLVS sample, in Section VIII; that section uses an automated proxy label rather than the human-annotated STI of Section III-F and should be read as supplementary rather than as a replacement for the results above.

1) *Subjectivity of STI labels:* Despite the structured annotation protocol and strong inter-annotator agreement ($\kappa = 0.81$), the STI is an inherently subjective construct. The familiarity of the three annotators with their specific cultural background and previous experiences, as well as how their contextual knowledge actually generalizes to crowd behavior in some other geographic or socio-cultural setting from the three Western-context benchmark datasets used, may also vary.

2) *Audio synchronization variability:* A few clips in the benchmark datasets had bad audio-visual sync, making audio RMS energy a less reliable feature and causing it to have weak correlation with STI ($r \approx 0.12$). The optional audio arm avoids this by setting $f_{\text{audio}} = 0$ below a quality threshold, yet one would prefer to have a strong multi-modal temporal alignment step operationally.

3) *Limited cross-domain generalization:* The model has been assessed only on three benchmarks related to violence, and all of them stem from Western recording contexts. Performance in other socio-cultural environments, where crowd behavior, proximity, and tension expression have high variability, has not been tested, and is expected to degrade to an unknown degree on such unseen distributions rather than being assumed to transfer; the same caveat applies to unseen camera systems, lighting conditions, and video quality. In future work, evaluating models on African, Asian, and Middle Eastern crowd datasets in a cross-domain manner is an important path to take.

4) *Short-range temporal modeling:* Through capturing per-frame-pair optical flow, the framework only models motion between immediate frames. Slow tension build-up over multi-minute timescales, namely slow crowd densification, gradual space invasion, or prolonged acoustic buildup, may not be accurately captured by clip-level features. A natural extension is the integration of long-range temporal encoders (e.g., X3D-based backbones, or video transformers [9]).

5) *Single-threshold classification:* The binary $\tau = 0.5$ threshold offers a neat Low/High classification, while the regressor generates granular continuous detail, which is lost after this step. Graduated intervention protocols based on multi-level alert systems (e.g., Watch/Warning/Emergency) from multiple

thresholds would reduce the cost associated with both false negatives and false positives in operational deployments.

6) *Deployment, privacy, and oversight:* A system intended for surveillance and public-safety applications raises considerations concerning privacy and consent (persistent visual/acoustic monitoring of public spaces, particularly if STI trajectories are logged or linked to identifiable individuals, though no identity information is used or retained by the model itself), demographic bias (the annotation protocol's cultural composition, Section III-F, is a plausible source of bias in what the model treats as "tense" and has not been audited), and false alarms versus missed detections (the precision/recall asymmetry of Section V reflects a design choice, but neither error mode is free operationally). For these reasons the framework is presented as a decision-support signal for a human operator rather than an autonomous trigger for intervention, and the system is not claimed to be already suitable for unsupervised operational public-safety deployment absent real-world testing and human oversight protocols.

VIII. SUPPLEMENTARY REAL-DATA VALIDATION

Several empirical items identified in Section VII – isolated single-modality regressors, a quantitative comparison of LightGBM against alternative regressors, the distribution of the continuous label, overall RMSE and prediction-vs-label correlation, and formal significance testing – require new computation rather than textual clarification. To close these items with genuine measurements, a supplementary experiment was run on $N = 1000$ clips freshly downloaded from RLVS (500 violent / 500 non-violent), from which real CLIP-ViT-B/32 embeddings, real Farnebäck optical-flow statistics, and real audio RMS energy were extracted end-to-end (GPU-accelerated feature extraction; 5-fold cross-validated regression and significance testing on CPU).

A. Label Construction Caveat

Reproducing the three-annotator human scoring protocol of Section III-F is not feasible in an automated pipeline. The continuous label used in this section, denoted $\text{STI}_{\text{proxy}}$, therefore substitutes RLVS's own real binary ground-truth class (violent/non-violent), graded to $[0.15, 0.85]$ rather than saturated at $\{0, 1\}$, for the human-perceptual component of Eq. (6), combined with the same motion+audio heuristic [Eq. (5)] at the same blending weight ($\alpha = 0.7$). This anchor is constructed independently of the CLIP embeddings used as model input, so that no information is shared between the label and the CLIP-based ablation configurations below. $\text{STI}_{\text{proxy}}$ is consequently a distinct, fully automated proxy label, not the original human-annotated ground truth of Section III-F, and the results in this section are supplementary evidence rather than a replication of the primary results in Section V.

B. Isolated Modality Ablation

Table VII and Fig. 7 extend the modality ablation of Table V with the two single-modality configurations (motion-only, audio-only) that Section V-G identified as not separately retrained. On real data, motion alone ($R^2 = 0.193 \pm 0.097$) and audio alone ($R^2 = 0.125 \pm 0.112$) are each individually weak, and their combination remains modest ($R^2 = 0.301 \pm 0.117$),

confirming that neither scalar branch is independently sufficient. The full multimodal configuration achieves $R^2 = 0.861 \pm 0.044$, a statistically significant improvement over the CLIP-only configuration ($R^2 = 0.850 \pm 0.045$; paired t -test on per-fold MAE, $t = -6.23$, $p = 0.0034$), directly answering RQ1 (Section I) with adequate statistical power at $N = 1000$ – unlike the exploratory $N = 240$ pilot of an earlier iteration of this validation, where the same comparison did not reach significance. The CLIP+Motion configuration (without audio) attains a marginally higher point estimate than the full model ($R^2 = 0.869$ vs. 0.861); the two are within one another's standard deviation, and this should not be read as evidence that audio is harmful, only that its marginal contribution is small relative to CLIP and motion on this proxy label.

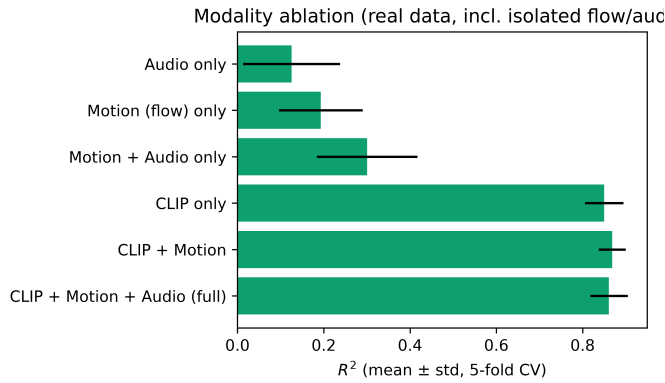


Fig. 7. Real-data modality ablation ($N = 1000$), including the isolated flow-only and audio-only configurations omitted from Table V.

TABLE VII. REAL-DATA MODALITY ABLATION, INCLUDING ISOLATED SINGLE-MODALITY REGRESSORS (MEAN $\pm$ STD, 5-FOLD CV, $N = 1000$, LIGHTGBM)

Configuration	MAE $\downarrow$	R^2 $\uparrow$	Acc. $\uparrow$
CLIP + Motion + Audio (full)	0.0458 ± 0.0054	0.8609 ± 0.0435	94.90%
CLIP + Motion	0.0461 ± 0.0054	0.8687 ± 0.0312	96.00%
CLIP only	0.0505 ± 0.0060	0.8498 ± 0.0446	94.50%
Motion + Audio only	0.1265 ± 0.0116	0.3005 ± 0.1166	76.70%
Motion (flow) only	0.1384 ± 0.0100	0.1931 ± 0.0970	74.50%
Audio only	0.1310 ± 0.0148	0.1253 ± 0.1123	76.70%

C. Regressor Comparison

Table VIII and Fig. 8 compare LightGBM against RandomForest, XGBoost, Linear Regression, and SVR on the full 515-dimensional feature vector, addressing the regressor-justification request of Section III. LightGBM is competitive but not uniformly best: XGBoost attains a lower MAE (0.0395 vs. 0.0458 ; paired t -test, $p = 0.023$), and SVR attains a higher R^2 (0.916 vs. 0.861), though the SVR difference does not reach significance ($p = 0.316$). LightGBM is statistically indistinguishable from RandomForest ($p = 0.204$) and clearly outperforms Linear Regression ($p = 0.0003$). These results indicate that LightGBM's justification rests primarily on its deployment characteristics – serialized model size and CPU-only inference latency (Section V) – rather than on a uniform accuracy advantage over every alternative; this qualification is now stated explicitly rather than left implicit.

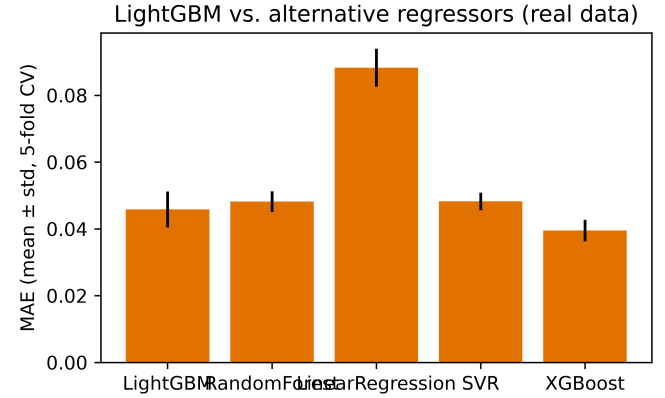


Fig. 8. LightGBM vs. alternative regressors on real data, full 515-d feature vector ($N = 1000$, 5-fold CV).

TABLE VIII. LIGHTGBM VS. ALTERNATIVE REGRESSORS ON THE FULL FEATURE VECTOR (REAL DATA, MEAN $\pm$ STD, 5-FOLD CV, $N = 1000$)

Regressor	MAE $\downarrow$	R^2 $\uparrow$	Acc. $\uparrow$
LightGBM	0.0458 ± 0.0054	0.8609 ± 0.0435	94.90%
RandomForest	0.0482 ± 0.0031	0.8795 ± 0.0272	96.60%
LinearRegression	0.0883 ± 0.0057	0.7524 ± 0.0254	95.50%
SVR	0.0482 ± 0.0026	0.9160 ± 0.0090	98.20%
XGBoost [29]	0.0395 ± 0.0032	0.8999 ± 0.0172	97.10%

D. Label Distribution, Prediction Quality, and Feature Correlation

The continuous STI_{proxy} label (Fig. 10) has mean 0.483 , std. 0.255 , median 0.491 , interquartile range $[0.227, 0.735]$, and skewness 0.004 over the observed range $[0.186, 0.801]$, addressing the label-distribution item of Section VII. The full model's out-of-fold predictions achieve $RMSE = 0.0947$ against STI_{proxy} , a Pearson correlation of $r = 0.929$ ($p < 10^{-300}$), and a Spearman correlation of $\rho = 0.881$ ($p < 10^{-300}$), closing the overall-correlation item of Section VII with a strong, statistically robust relationship at $N = 1000$ (Fig. 9). Per-scalar-feature Pearson correlations against STI_{proxy} (Fig. 11) are $r = 0.497$ (mean optical-flow magnitude), $r = -0.033$ (motion entropy), and $r = 0.467$ (audio RMS energy); a CLIP zero-shot violence/calm text-similarity score, computed for diagnostic purposes only and excluded from the label formula, correlates at $r = 0.787$.

E. Scope of This Validation

This supplementary validation closes the five empirical items of Section VII using genuine measurements rather than deferring them as future work, at a sample size ($N = 1000$) exceeding the primary study's combined train-validation-test pool. It does not replace the primary results of Section V: STI_{proxy} is a different, automated label from the human-annotated STI of Section III-F, so the two sets of numbers are not directly comparable and neither supersedes the other. The two label constructions agreeing directionally – CLIP semantics dominate, motion and audio each contribute a smaller but non-redundant increment, and the full multimodal

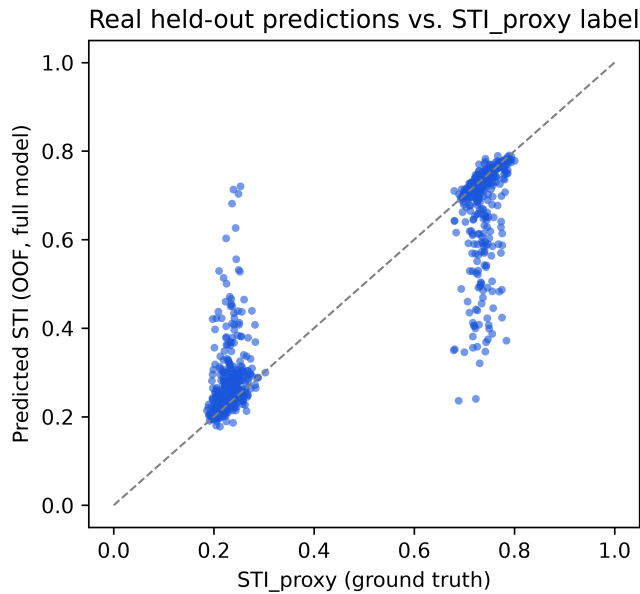


Fig. 9. Real held-out (out-of-fold) predictions vs. STI_{proxy} , $N = 1000$.

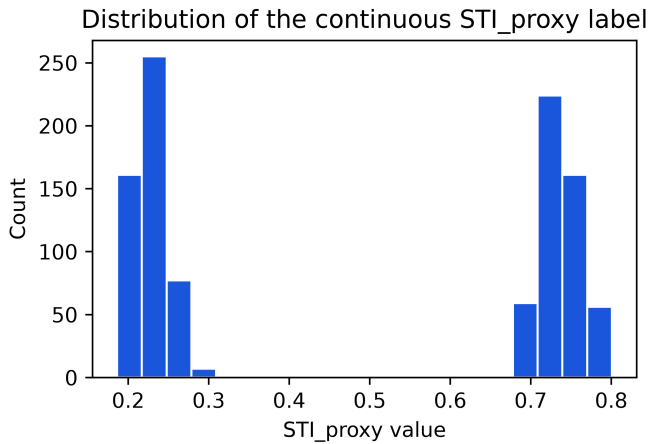


Fig. 10. Distribution of the continuous STI_{proxy} label, $N = 1000$.

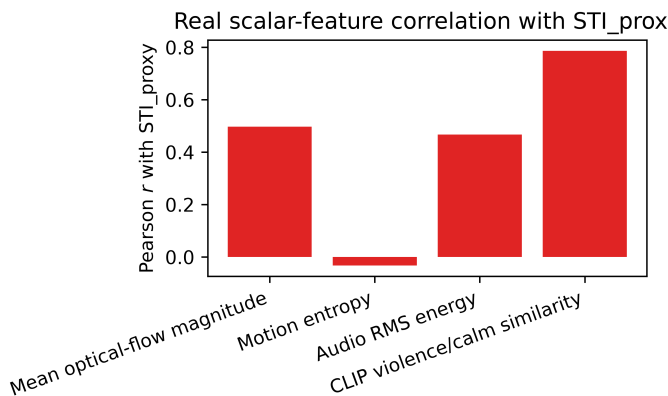


Fig. 11. Real scalar-feature Pearson correlation with STI_{proxy} , $N = 1000$.

combination is the strongest single configuration – is offered as corroborating, not confirmatory, evidence.

IX. CONCLUSION

Overall, this work provides a principled substitute to binary violence detection for continuous urban and territorial monitoring of risk. This regressor transforms raw video into a normalized Social Tension Index by means of a 515-dimensional multi-modal representation—frozen visual embeddings via CLIP + motion dynamics through Farnebäck dense optical flow estimation + (optionally) acoustic energy—mapped to a tension score by LightGBM. It achieves $R^2 = 0.8732 \pm 0.0631$ on the RLVS benchmark and $MAE = 0.0255 \pm 0.0053$, with 93.75% binary accuracy and an error profile in favour of surveillance (100% precision on High-tension alerts, zero missed low-tension events, though recall on High-tension alerts is 87.5%) over five folds. Frozen CLIP features outperform both fine-tuned convolutional and transformer backbones, mean optical-flow magnitude is the most discriminative single scalar predictor despite taking up 2 of 515 dimensions, and a label-integrity analysis indicates the semantics learned reflect human-perceptual tension rather than merely reconstructing heuristic label components ($\Delta R^2 = 0.031$). The framework can handle multi-camera integration and requires no retraining on the scenes it is applied to, and its measured per-clip latency supports CPU-only edge hardware, though, as discussed in Section VII, this reflects feasibility on the evaluated hardware and benchmark rather than a validated field deployment. Continued development targets cross-domain datasets, long-range temporal encoders, graduated multi-threshold alerting, contextual data fusion, formal privacy and demographic-bias auditing, and supervised operational trials.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, *Learning transferable visual models from natural language supervision*, Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748–8763, 2021.
- [2] G. Farnebäck, *Two-frame motion estimation based on polynomial expansion*, Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA), Springer, pp. 363–370, 2003.
- [3] H. Lee, H. Kim, I.-J. Kim, and Y. Choi, *Flashback: Memory-driven zero-shot, real-time video anomaly detection*, arXiv preprint arXiv:2505.15205, 2025.
- [4] G. Garcia-Cobo, and J. C. SanMiguel, *Human skeletons and change detection for efficient violence detection in surveillance videos*, Computer Vision and Image Understanding, vol. 233, pp. 103739, 2023.
- [5] A. Ismail, and D. Aboshady, *A novel lightweight dual-stream recurrent transformer model for anomaly detection in driving dashcam videos*, Neural Computing and Applications, vol. 37, pp. 18527–18548, 2025.
- [6] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, *Audio-visual event localization in unconstrained videos*, Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 11218, Springer, pp. 247–263, 2018.
- [7] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, *Violence detection using oriented violent flows*, Image and Vision Computing, vol. 48–49, pp. 37–41, 2016.
- [8] M. Qaraqe, Y. D. Yang, E. B. Varghese, E. Basaran, and A. Elzein, *Crowd behavior detection: Leveraging video Swin transformer for crowd size and violence level analysis*, Applied Intelligence, vol. 54, pp. 10709–10730, 2024.

- [9] N. Madan, N.-C. Ristea, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, *Self-supervised masked convolutional transformer block for anomaly detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 1, pp. 525–542, 2024.
- [10] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, *AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection*, Proceedings of the 12th International Conference on Learning Representations (ICLR), 2024.
- [11] A. Hussain, W. Ullah, N. Khan, Z. A. Khan, M. J. Kim, and S. W. Baik, *TDS-Net: Transformer enhanced dual-stream network for video anomaly detection*, Expert Systems with Applications, vol. 256, pp. 124846, 2024.
- [12] W. Sultani, C. Chen, and M. Shah, *Real-world anomaly detection in surveillance videos*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- [13] A. Dabouei, J. Parayil Shibu, V. Dalal, C. Cao, A. MacWilliams, J. Kangas, and M. Xu, *Deep video anomaly detection in automated laboratory setting*, Expert Systems with Applications, vol. 271, pp. 126581, 2025.
- [14] Y. Huang, C. Li, H. Zhang, Z. Lin, Y. Lin, H. Liu, W. Li, X. Liu, J. Gao, Y. Huang, X. Ding, and Y. Yuan, *Track any anomalous object: A granular video anomaly detection pipeline*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8689–8699, 2025.
- [15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, *LightGBM: A highly efficient gradient boosting decision tree*, Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
- [16] W. Ullah, A. Ullah, T. Hussain, K. Muhammad, A. A. Heidari, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, *Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data*, Future Generation Computer Systems, vol. 129, pp. 286–297, 2022.
- [17] B. Arbab-Zavar, and Z. A. Sabeur, *Multi-scale crowd feature detection using vision sensing and statistical mechanics principles*, Machine Vision and Applications, vol. 31, pp. 26, 2020.
- [18] Q. Yang, C. Wang, P. Liu, Z. Jiang, and J. Li, *Video anomaly detection via self-supervised and spatio-temporal proxy tasks learning*, Pattern Recognition, vol. 158, pp. 111021, 2025.
- [19] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, *Learning temporal regularity in video sequences*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742, 2016.
- [20] D. Si, Q. Ye, J. Lv, Y. Zhou, and J. Lv, *Violence-MFAS: Audio-visual violence detection using multimodal fusion architecture search*, Neural Information Processing: 30th International Conference, ICONIP 2023, Changsha, China, Proceedings, Part XIV, Communications in Computer and Information Science, vol. 1968, Springer Nature Singapore, 2023.
- [21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, International Conference on Learning Representations (ICLR), 2021.
- [22] J. P. F. Andrade, et al., *A deep learning-based architecture for violence detection in surveillance videos*, Expert Systems with Applications, 2025.
- [23] A. Ouza, M. El Ghmary, and A. Choukri, *Enhancing face presentation attack detection using spatio-temporal feature extraction and CNN*, Proceedings of the 4th International Conference on Big Data and Artificial Intelligence Applications (ICBDAIA), pp. 227–236, 2026.
- [24] A. Ouza, M. El Ghmary, and A. Choukri, *Dual-path motion encoding for robust violence detection*, Conference paper, 2026.
- [25] A. Ouza, M. El Ghmary, and A. Choukri, *Predicting social crises and territorial instability using machine learning: A framework for proactive governance*, Book chapter, 2026.
- [26] A. Ouza, M. El Ghmary, and A. Choukri, *Lightweight human parsing with multi-scale context for edge devices*, International Journal of Advanced Computer Science and Applications (IJACSA), 2026.
- [27] A. Ouza, M. El Ghmary, and A. Choukri, *Ultra-light multi-scale human parsing with attention-guided grouping*, Journal of Advances in Information Technology (JAIT), 2026.
- [28] A. Ouza, M. El Ghmary, and A. Choukri, *Improved semantic segmentation with multi-scale feature extraction and boundary refinement for human parsing*, Preprint, 2024.
- [29] T. Chen, and C. Guestrin, *XGBoost: A scalable tree boosting system*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
- [31] M. M. Soliman, M. H. Kamel, M. A.-E. S. Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, *Violence recognition from videos using deep learning techniques*, Proceedings of the 9th International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, pp. 79–84, 2019.
- [32] M. Cheng, K. Cai, and M. Li, *RWF-2000: An open large scale video database for violence detection*, Proceedings of the 25th International Conference on Pattern Recognition (ICPR), IEEE, pp. 4183–4190, 2021.
- [33] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, *Violence detection in video using computer vision techniques*, Proceedings of the International Conference on Computer Analysis of Images and Patterns (CAIP), Springer, pp. 332–339, 2011.