

Real-Time Suspicious-Event Monitoring for Exam Integrity Using Visual Behavior Monitoring and Lightweight Acoustic Assistance

Tianzhen Chen^{1,*}

¹EIT Data Science and Communication College, Zhejiang Yuexiu University, Shaoxing, Zhejiang, China

*Corresponding author: terrychen17@zyufl.edu.cn

Abstract: Online proctoring systems commonly make visual judgments frame by frame. In practice, however, changes in participant appearance or acquisition conditions can make ordinary posture shifts look suspicious, leading to brief false alarms or repeated warnings. This paper studies real-time suspicious-event monitoring with a role-separated design: the visual stream is kept as interpretable evidence for monitoring and review, whereas automatic risk estimation is handled by a lightweight causal acoustic path. YuNet, 13 pose-geometric features, 20 contour features, and a frozen Random Forest produce the visual score. On the audio side, 13 lightweight features are computed from a strictly causal trailing window, standardized, and evaluated by logistic regression; a persistence rule then converts the binary outputs into Suspicious Events. Experiments on the public OEP dataset use participant-level outer validation, with development, outer-test, and system validation evidence kept separate. In the clean outer test, Acoustic increases Macro Participant PR-AUC from 0.2987 to 0.5666 and Event F1 from 0.2003 to 0.4120, while FP/min decreases from 2.1604 to 1.0004 relative to Visual. The simple Visual + Acoustic setting reaches an Event F1 of 0.4050 and does not provide a stable additional gain. These findings motivate the final division of roles: acoustic cues drive automatic risk discrimination, while visual cues remain available for monitoring and human review. Across 250,932 streaming ticks, no binary, temporal-state, or alarm mismatches are observed; the core acoustic alarm engine has a P95 latency below 1 ms and does not require a GPU.

Keywords: Exam Integrity Monitoring, Computer Vision, Acoustic Activity Analysis, Suspicious-Event Detection, Causal Temporal Decision, Real-Time Monitoring

1. Introduction

Online examinations need integrity monitoring that works continuously, not only a classifier that labels one frame at a time. Existing systems already use facial behavior, gaze, object detection, audio, and temporal cues ^[1]. Yet normal actions can resemble suspicious behavior after transfer to new participants or recording conditions. Domain shift can reduce reliability ^[2], while video-anomaly research also stresses temporal context rather than isolated frames ^[3]. A brief frame error may therefore become a repeated alert and add unnecessary review work.

Accordingly, the study is organized around suspicious events rather than isolated frame labels. We first examine how the frozen visual model behaves after transfer to unseen participants, then test causal temporal stabilization, and finally compare Visual, Acoustic, and Visual + Acoustic at participant level. The resulting evidence points to a practical separation of duties: acoustic information is used to estimate automatic risk, while the visual stream is retained for observation and later review. A system alert is therefore reported only as a Suspicious Event; the reviewer, not the model, makes the final misconduct judgment.

The study makes three contributions: a controlled cross-participant modality comparison; a strictly causal persistence rule that converts frame-level risk into event-level alerts while exposing the trade-off among false alarms, delay, and complexity; and a lightweight streaming prototype evaluated for equivalence, runtime, long-run stability, real-time replay, and human-review functionality.

2. Related Work

2.1. Computer Vision-Based Online Exam Proctoring

Current proctoring systems generally rely on more than a single face or object check. OEP, for example, includes webcam, wearcam, microphone, identity, gaze, speech, active-window, and phone information^[1]. Proctor Net extends face recognition with eye/head-pose and mouth-opening cues^[4]. Singh et al. use another combination that includes object detection, facial orientation, liveness, person counting, and audio^[5]. More recent comparisons also show that recognition performance is only part of the deployment problem; inference cost matters when monitoring has to run continuously^[6].

2.2. Pose, Gaze, and Behavioral Representation

Pose and gaze are useful because they give interpretable clues about changes in attention. CHEESE combines gaze, head pose, facial appearance, body information, and background context^[7]. Human-in-the-loop test-security methods can further control which atypical cases are sent for expert review^[8]. At deployment time, cross-scene domain shift, acquisition differences, imbalance, and privacy remain practical concerns^[9]. These issues also arise in exam monitoring, where viewpoint, lighting, background, and participant appearance vary across sessions.

2.3. Multi-Source Monitoring and Cross-Modal Anomaly Detection

Multi-source monitoring can provide complementary evidence when visual behavior is ambiguous. Kaddoura and Gumaei combine front-camera, rear-camera, and speech modules through soft voting^[10]. Cross-modal anomaly studies further exploit visual-text integration^[11] and semantic consistency^[12]. Such results motivate multimodal evaluation, but do not guarantee that simple fusion will improve performance in a new domain.

2.4. Temporal Event Monitoring and Real-Time Deployment

Continuous monitoring also requires temporal stability and efficient inference. Huy et al. combine object detection with CNN, Bi-GRU, and CNN-LSTM models for online-exam behavior detection^[13]. Fast anomaly detection has been studied with high-throughput inference^[14], while multi-frame prediction error can reduce false positives at low computational cost^[15]. Accordingly, our evaluation includes event-level false alarms, delay, causality, latency, and stability.

3. Proposed Method

3.1. Overall Framework

Figure 1 presents the system in three parts: Visual Behavior Monitoring, Lightweight Causal Acoustic Event Detection, and Real-Time Monitoring and Human Review. In the visual part, YuNet locates the face and a 20%-padded ROI is formed. Pose13 and Contour20 are extracted from this region and passed to the frozen Random Forest, which outputs a Visual Suspicious Score. The score is displayed as behavioral evidence for monitoring and later review rather than used to trigger the automatic alarm.

The acoustic path runs in parallel. A strictly causal trailing buffer supplies Audio13, which is standardized before logistic regression estimates suspicious risk. After the fixed 0.5 threshold, the binary sequence enters the causal T1 persistence rule; a Suspicious Event is issued only when the required persistence condition is met. The monitoring interface combines the event output with auxiliary visual evidence through a risk curve, timeline, and event log. A proctor can mark the interval as Reviewed, Confirmed Concern, or Dismissed. Thus, visual information remains available for interpretation without changing the acoustic alarm decision.

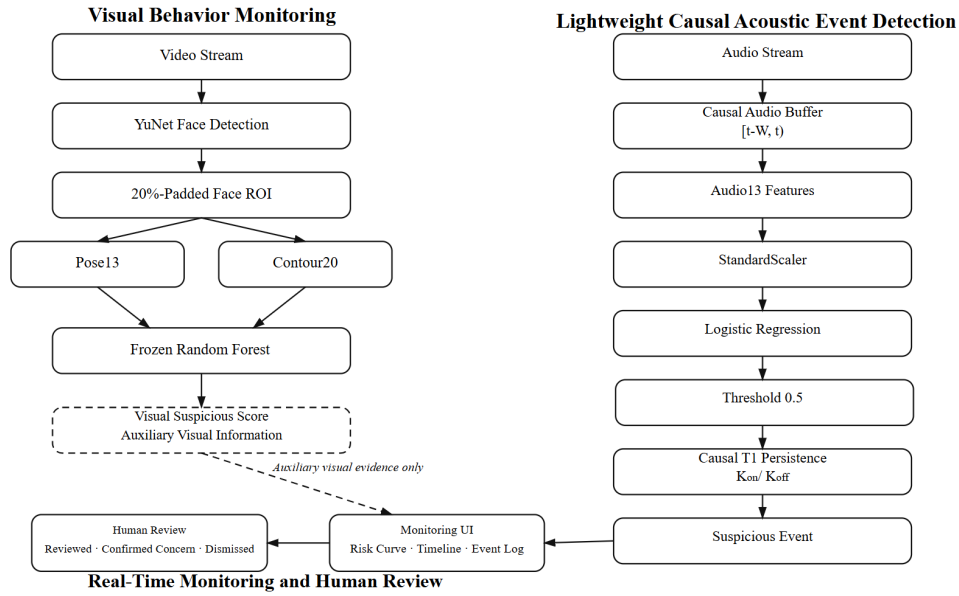


Figure 1. Overall framework of the proposed functionally decoupled real-time monitoring system.

3.2. Visual and Acoustic Representation

YuNet locates the face and returns five landmarks. The face box is enlarged by 20% on each side and clipped at the image boundary. From the resulting ROI, 13 pose-geometric features describe eye, nose, and mouth relationships, including slopes, distances, and ratios. Another 20 contour features describe edge density, largest-contour geometry, centroid, solidity, extent, seven Hu moments, and directional edge distributions. The frozen RF receives Contour20 followed by Pose13; their joint representation and visual suspicious score are given in Eq. (1):

$$v_i=[c_i;p_i] \in \mathbb{R}^{33}, s_i^v=P_{\text{RF}}(\text{head}_{\text{down}})+P_{\text{RF}}(\text{head}_{\text{tum}})=1-P_{\text{RF}}(\text{normal}) \quad (1)$$

The score s_i^v lies in $[0,1]$ but is not probability-calibrated; it is therefore used only as a Visual Suspicious Score. OEP audio is 8-kHz mono 8-bit unsigned PCM, normalized by $x[n]=(u[n]-128)/128$. At decision time t , the strictly causal trailing window is defined in Eq. (2), and only completed 25-ms frames before the current decision time are used:

$$X_i=x(\tau)|t-W \leq \tau < t \quad (2)$$

Each causal window yields 13 acoustic variables: RMS amplitude, log RMS, mean absolute amplitude, population standard deviation, within-frame zero-crossing rate, peak amplitude, peak-to-RMS ratio, spectral centroid, 85% rolloff, spectral flatness, 300-3400-Hz and 0-300-Hz power ratios, and a robust short-frame activity fraction. Frequency-domain quantities use a Hann window and 200-point RFFT. The activity threshold is $\text{median}(\text{RMS})+1.5 \cdot \text{MAD}$, with the frozen numerical safeguards retained. Eq. (3) summarizes this representation.

$$a_i=[a_1,a_2,\dots,a_{13}]^T \in \mathbb{R}^{13} \quad (3)$$

3.3. Risk Estimation and Causal Event Formation

Within each outer fold, StandardScaler is learned from the development participants only. The standardized Audio13 features are then passed to logistic regression for suspicious-risk estimation. For the formal A1 setting, we keep `class_weight=balanced`, the `liblinear` solver, L2 regularization, `random_state=42`, `max_iter=2000`, and a fixed decision threshold of 0.5. Eq. (4) gives the resulting probability and binary decision:

$$p_i=1/[1+\exp(-(b+w^T z_i))], \hat{y}_i=I(p_i \geq 0.5) \quad (4)$$

T1 Persistence uses two states, NORMAL and SUSPICIOUS. While the system is NORMAL, K_{on} consecutive positive predictions are required before the state changes to SUSPICIOUS. While it is SUSPICIOUS, K_{off} consecutive negative predictions are required before the state returns to NORMAL. An alarm is generated only when the system enters SUSPICIOUS, preventing repeated alerts during a

continuing event. The transition logic is strictly causal and relies only on the current and historical observations, without access to future information. If the current input is unavailable, the current state and both persistence counters are left unchanged. Eq. (5) defines the alarm indicator and predicted event interval, and Figure 2 summarizes the state-transition logic.

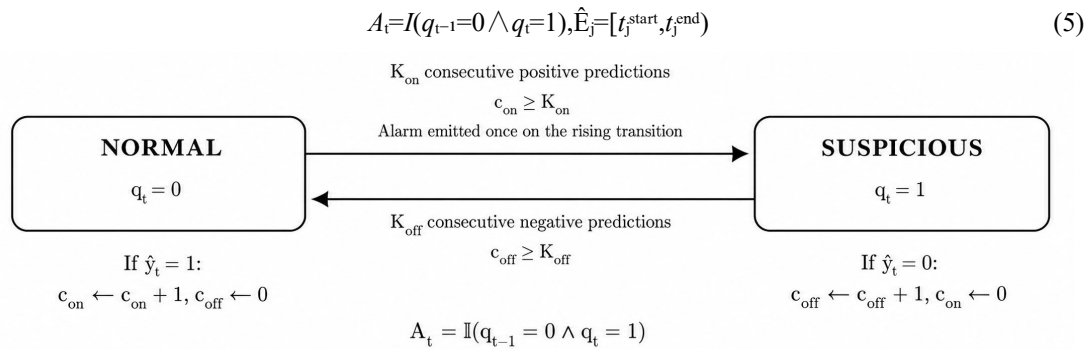


Figure 2. Strictly causal persistence-based state transition.

The GUI runs the acoustic alert chain and visual monitoring chain in parallel. Human review supports Reviewed, Confirmed Concern, and Dismissed; Confirmed Concern is a review state, not an automatic cheating determination. Raw-media recording is off by default.

4. Experimental Evaluation and Discussion

4.1. Data, Protocol, and Metrics

The OEP material used in the formal experiments covers 24 participants and provides 24 webcam recordings, 24 wearcam recordings, and 24 audio recordings [1]. We use the 6.983 h webcam view together with the corresponding audio and leave wearcam out of the final model. The locally available annotations contain 565 raw suspicious intervals, while the original paper reports 569. Because the four unobserved records cannot be recovered with confidence, we preserve this discrepancy instead of reconstructing them manually. Two intervals are merged only when they overlap for a positive duration; this leaves 475 GT events, represented as left-closed/right-open intervals $[t_{\text{start}}, t_{\text{end}})$. At 5 Hz, the data yield 125,706 webcam ticks, 122,044 valid visual outputs, and 121,804 paired Visual/Acoustic ticks.

Generalization is assessed with three participant-level outer folds, each containing 16 development participants and 8 outer-test participants. The development portion is used for scaler/model fitting and parameter selection. The clean outer test is accessed once for the generalization assessment; streaming and GUI experiments performed afterward are treated as system validation rather than new performance tests. A0 contains the frozen visual score, A1 contains Audio13, and A2 adds the frozen visual score to Audio13 as a deliberately simple controlled combination. Frame-level reporting includes Precision, Recall, F1, Balanced Accuracy, ROC-AUC, and PR-AUC. For event matching, the first alarm inside a GT interval counts as the detection, additional alarms in the same interval are duplicates, and alarms outside all GT intervals are false alarms. We report Event Precision, Event Recall, Event F1, FP/min, Median Delay, and P95 Delay.

Fold-specific A1 settings are $(W, C) = (2.0 \text{ s}, 10)$, $(2.0 \text{ s}, 0.1)$, and $(1.0 \text{ s}, 10)$, with T1 settings $(K_{\text{on}}, K_{\text{off}}) = (8, 3)$, $(8, 3)$, and $(10, 3)$. The separate GUI demonstration model uses all 24 participants with $W = 2.0 \text{ s}$, $C = 10$, $K_{\text{on}} = 8$, $K_{\text{off}} = 3$, and threshold 0.5; it is not used to report additional generalization performance.

Experimental Environment. All experiments and system benchmarks were carried out on a laptop equipped with a 12th Gen Intel Core i9-12900H processor (14 cores, 20 logical processors, 2.50 GHz base frequency), 16 GB of system memory, Intel Iris Xe graphics, an NVIDIA GeForce RTX 2050 with 4 GB of dedicated VRAM, and NVMe solid-state storage. GPU acceleration was not used by the evaluated pipeline. The latency, resource-use, streaming-equivalence, and long-run stability measurements therefore mainly describe CPU-based execution on this machine.

4.2. From Visual Frame Predictions to Stable Events

When the frozen visual model is transferred directly to OEP, it obtains frame Precision 0.2578,

Recall 0.7017, F1 0.3771, Balanced Accuracy 0.4886, ROC-AUC 0.5070, and PR-AUC 0.2920. Although the recall appears acceptable when considered alone, the resulting alert stream is not practical. With no temporal constraint, the model detects 320 of the 475 GT events while also producing 6,563 false alarms and 2,099 duplicates. This corresponds to Event Precision 0.0356, Event Recall 0.6737, Event F1 0.0677, and 21.386 FP/min. In addition, 72.45% of background timestamps receive a non-normal visual label, and edge_density has an absolute standardized mean difference of 3.862. Taken together, these observations point to substantial cross-domain shift and explain why high frame recall is not enough for continuous alerting. Table 1 reports the event-level comparison of the temporal strategies.

Table 1. Event-level comparison of temporal strategies.

Method	Event P	Event R	Event F1	FP/min	Duplicates	Median Delay
T0 Raw	0.0356	0.6737	0.0677	21.386	2099	1.0 s
T1 Persistence	0.1572	0.3263	0.2122	2.460	76	4.4 s
T2 Window	0.1771	0.2505	0.2075	1.584	67	4.2 s
T3 State Machine	0.1888	0.2421	0.2122	1.463	45	5.0 s

Table 1 shows that temporal constraints sharply reduce false and repeated alarms. T1 raises Event F1 from 0.0677 to 0.2122 and lowers FP/min from 21.386 to 2.460. T3 reduces FP/min further to 1.463 and duplicates to 45, but Event F1 remains 0.2122 while recall falls and median delay grows. Because T1 reaches the same Event F1 with a less involved state rule, it is used in the remaining experiments. The choice follows the measured event-level trade-off rather than a general preference for simpler models.

Before accessing the clean outer test, we used development-only analyses to decide whether visual adaptation or acoustic information merited formal evaluation. Visual adaptation moves mean PR-AUC from 0.329849 to 0.343840 (+0.013991) and Balanced Accuracy from 0.502893 to 0.552928; the corresponding importance split is 63.31% Pose and 36.69% Contour. In the acoustic feasibility analysis, development PR-AUC is 0.272385 for Visual, 0.565901 for Acoustic, and 0.564900 for Visual + Acoustic. With T1 kept frozen, Acoustic changes Event F1 from 0.190348 to 0.417023 and FP/min from 2.137793 to 1.168454. These numbers are development evidence only. They were used to justify the already specified clean outer-test comparison, not as a substitute for it.

4.3. Clean Outer-Test Modality Comparison

The clean participant-level outer-test results for A0, A1, and A2 are summarized in Table 2.

Table 2. Clean participant-level outer-test results.

Metric	A0 Visual	A1 Acoustic	A2 Visual + Acoustic
Frame Precision	0.235935	0.655673	0.654315
Frame Recall	0.376002	0.549226	0.542172
Frame F1	0.289939	0.597747	0.592988
Balanced Accuracy	0.469415	0.722837	0.719667
ROC-AUC	0.464254	0.782030	0.782232
Pooled PR-AUC	0.258030	0.673864	0.671634
Macro Participant PR-AUC	0.298686	0.566561	0.567493
Event Precision	0.154639	0.355183	0.351393
Event Recall	0.284211	0.490526	0.477895
Event F1	0.200297	0.412025	0.404996
FP/min	2.160425	1.000378	0.993861
Median Delay	3.60 s	2.40 s	2.40 s
P95 Delay	22.90 s	9.76 s	11.02 s

Table 2 contains the main generalization evidence. Against A0, A1 gains 0.267875 in Macro Participant PR-AUC, 0.415834 in Pooled PR-AUC, 0.207100 in Macro Balanced Accuracy, and 0.237302 in Macro Frame F1. At event level, F1 moves from 0.200297 to 0.412025 (+0.211728) and Recall increases by 0.206316, while FP/min decreases by 1.160047. P95 Delay also shortens, from 22.90 s to 9.76 s. Looking across participants rather than only at pooled metrics, A1 is better than A0 for 21 participants, similar for one, and worse for two.

The simple addition of the frozen visual score to Audio13 is not consistently helpful. Under A2, Macro PR-AUC changes by only +0.000932; Pooled PR-AUC, Macro Frame F1, and Event F1 change

by -0.002230, -0.004533, and -0.007029, respectively. The participant-level pattern is similarly mixed: A2 is better than A1 for three participants, similar for 19, and worse for two. In other words, having both sensing channels available does not by itself create useful complementarity. One likely reason is the label mismatch: the visual score was learned for normal/head_down/head_turn, whereas OEP is evaluated as background/suspicious-event. Under that shift, the frozen score can add noise to an Audio13 representation that is already discriminative. We did not introduce a new fusion design after observing the outer-test result.

4.4. Behavior-Dependent Modal Roles

Behavior-dependent Event Recall across the Official OEP types is reported in Table 3 for all three modality configurations.

Table 3. Event Recall across Official OEP types.

Type	Events	A0 Visual	A1 Acoustic	A2 Visual + Acoustic
Type 1	124	0.177419	0.145161	0.129032
Type 2	255	0.274510	0.666667	0.662745
Type 3	37	0.270270	0.378378	0.297297
Type 4	14	0.071429	0.357143	0.357143
Type 5	18	0.722222	0.222222	0.222222
Mixed	27	0.703704	0.814815	0.814815

Official-Type labels are used only for post-test diagnostics. Acoustic is especially effective for Type 2 (talking), where recall rises from 0.274510 to 0.666667, and Type 4 (telephone help), where it rises from 0.071429 to 0.357143. Type 5 shows the opposite pattern: Visual reaches 0.722222 recall, whereas Acoustic reaches 0.222222. This contrast is consistent with the sensing characteristics of the behaviors. Speech-related events produce salient acoustic activity, while silent device-related actions may remain primarily visual. The findings therefore support different functional roles for the two channels rather than replacement of Visual by Audio: Acoustic is used for automatic alerting, while Visual remains useful for observing silent behavior, interpreting posture, and supporting review.

4.5. Streaming and System Validation

Streaming equivalence, runtime, resource consumption, and stability measurements are summarized in Table 4.

Table 4. Streaming equivalence, runtime, and stability.

Category	Metric	Result
Equivalence	Compared ticks	250,932
Equivalence	Max feature error	0
Equivalence	Max probability error	1.088×10^{-14}
Equivalence	Binary/T1/alarm mismatch	0 / 0 / 0
Audio Engine	Mean / P95 / P99 latency	0.441 / 0.624 / 0.813 ms
Audio Engine	RTF / realtime capacity	0.00220 / 453.87×
Audio Engine	Process CPU / Peak RAM / GPU	4.85% / 152.05 MB / 0
Integrated	Mean / P95 latency	60.16 / 80.24 ms
Integrated	Effective FPS / Peak RAM / GPU	16.62 / 235.33 MB / 0
Long-run	1800 s, 9000 ticks	0 dropped; 0-ms backlog; +1.10 MB
Replay	3 x 90 s streams	mean 12.07 ms; P95 17.26 ms; max 19.17 ms; 0 dropped
GUI	Mean / P95 / P99 latency	42.83 / 47.40 / 53.71 ms
GUI	Effective FPS / Peak RAM / GPU	23.35 / 253.94 MB / 0

The streaming implementation reproduces the offline implementation to floating-point precision: the feature difference is zero, and no binary, T1-state, or alarm mismatch is observed. For the core acoustic path (Audio13, StandardScaler, Logistic Regression, and T1 Persistence), P95 latency is 0.624 ms. The integrated pipeline has a P95 latency of 80.24 ms, which remains below the 200-ms budget at 5 Hz; neither path relies on GPU acceleration. The auxiliary visual branch also fits within the same budget, with 24.51 ms for YuNet, 2.85 ms for feature extraction, and 30.23 ms for the complete visual score. In the tested long-run and 1x replay conditions, no dropped ticks or sustained backlog are

observed. Figure 3 shows the resulting interface.

The prototype brings the main monitoring outputs into one interface: visual evidence, acoustic risk, T1 state, a risk/event timeline, an event log, and controls for human review. Fourteen functional tests pass, including startup, camera-only and audio-only operation, integrated monitoring, device disconnection, restart, logging, review, and long-run stability. The automated test environment did not contain a physical camera, so camera handling was examined mainly with development/demo recordings and degraded-mode paths. For that reason, we do not claim prospective detection accuracy in a real examination room. Human Review is reported as an implemented interface capability only; the experiments do not show gains in accuracy, efficiency, or inter-rater agreement from that review step.

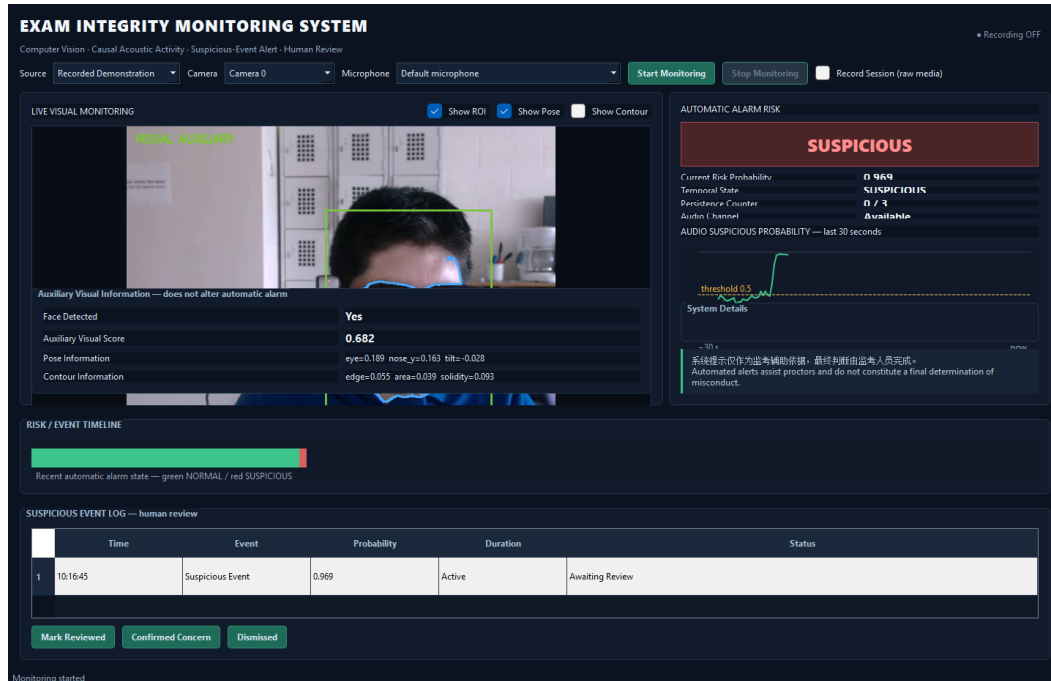


Figure 3. Real-time monitoring prototype in a Suspicious Event state.

5. Conclusion and Future Work

The results make clear that continuous exam-integrity monitoring cannot be judged from independent frame classification alone. Although the frozen visual model retains relatively high frame recall, its event-level false-alarm load is too large for practical alerting. Causal persistence cuts this burden, and the more elaborate temporal rules tested here do not improve Event F1. In the clean participant-level outer test, the lightweight acoustic setting increases Macro Participant PR-AUC from 0.2987 to 0.5666 and Event F1 from 0.2003 to 0.4120, while FP/min decreases from 2.1604 to 1.0004. Visual + Acoustic reaches an Event F1 of 0.4050 and does not exceed Acoustic-only. On this evidence, the final system uses lightweight acoustic risk with causal persistence to generate Suspicious Events and keeps the visual branch for behavioral interpretation and human review.

System tests further confirm offline-to-streaming agreement over 250,932 ticks, sub-millisecond P95 latency for the core acoustic alarm engine, operation within the 5-Hz decision budget, zero dropped ticks in the tested long-run and replay conditions, and no GPU requirement. These measurements come from the i9-12900H laptop described in Section 4.1, with GPU acceleration disabled. The integrated pipeline already meets the 5-Hz budget on this hardware, so a newer or faster general-purpose laptop may offer extra margin for a higher visual refresh rate, a higher decision frequency, or more concurrent monitoring tasks; cross-device benchmarking is still needed to verify that expectation. The present study is limited by the 24-participant OEP scale, behavior-dependent acoustic effectiveness, untested microphone and device robustness, the absence of independent probability calibration and hardware audio-video synchronization, prerecorded rather than prospective real-exam validation, and the lack of a proctor user study. Future work will therefore extend the evaluation to larger and more varied datasets, study behavior-aware modality selection or confidence gating, compare runtime across devices, and examine human-in-the-loop use with practicing proctors.

Acknowledgements

This work received support from the 2024 Scientific Research Project of Zhejiang Yuexiu University (Project No. N2024013), "Research on Examination Integrity Detection Based on Computer Vision Contour Extraction".

References

- [1] Y. Atoum, L. Chen, A. X. Liu, S. D. H. Hsu, and X. Liu. Automated Online Exam Proctoring[J]. *IEEE Transactions on Multimedia*, 2017, 19(7): 1609-1624. DOI: 10.1109/TMM.2017.2656064.
- [2] K. Chen, E. Gal, H. Yan, and H. Li. Domain Generalization with Small Data[J]. *International Journal of Computer Vision*, 2024, 132(8): 3172-3190. DOI: 10.1007/s11263-024-02028-4.
- [3] Q. Yang, C. Wang, P. Liu, Z. Jiang, and J. Li. Video Anomaly Detection via self-supervised and spatio-temporal proxy tasks learning[J]. *Pattern Recognition*, 2025, 158: 111021. DOI: 10.1016/j.patcog.2024.111021.
- [4] P. Tejaswi, S. Venkatramaphanikumar, and K. Venkata Krishna Kishore. Proctor Net: An AI framework for suspicious activity detection in online proctored examinations[J]. *Measurement*, 2023, 206: 112266. DOI: 10.1016/j.measurement.2022.112266.
- [5] T. Singh, R. R. Nair, T. Babu, and P. Duraisamy. Enhancing Academic Integrity in Online Assessments: Introducing an Effective Online Exam Proctoring Model using YOLO[C]//*Proceedings of the International Conference on Machine Learning and Data Engineering (ICMLDE 2023)*. *Procedia Computer Science*, 2024, 235: 1399-1408. DOI: 10.1016/j.procs.2024.04.131.
- [6] S. Essahraoui, I. Lamaakal, Y. Maleh, K. El Makkaoui, M. F. Bouami, I. Ouahbi, et al. Deep Learning Models for Detecting Cheating in Online Exams[J]. *Computers, Materials & Continua*, 2025, 85(2): 3151-3183. DOI: 10.32604/cmc.2025.067359.
- [7] Y. Liu, J. Ren, J. Xu, X. Bai, R. Kaur, and F. Xia. Multiple Instance Learning for Cheating Detection and Localization in Online Examinations[J]. *IEEE Transactions on Cognitive and Developmental Systems*, 2024, 16(4): 1315-1326. DOI: 10.1109/TCDS.2024.3349705.
- [8] M. Fauss, X. Liu, C. Li, I. Choi, and H. V. Poor. Bayesian Selection Policies for Human-in-the-Loop Anomaly Detectors with Applications in Test Security[J]. *Psychometrika*, 2026, 91(1): 279-311. DOI: 10.1017/psy.2025.10056.
- [9] R. Delussu, L. Putzu, and G. Fumera. Synthetic Data for Video Surveillance Applications of Computer Vision: A Review[J]. *International Journal of Computer Vision*, 2024, 132: 4473-4509. DOI: 10.1007/s11263-024-02102-x.
- [10] S. Kaddoura and A. Gumaei. Towards effective and efficient online exam systems using deep learning-based cheating detection approach[J]. *Intelligent Systems with Applications*, 2022, 16: 200153. DOI: 10.1016/j.iswa.2022.200153.
- [11] C. Li, H. Li, and G. Zhang. Cross-modality integration framework with prediction, perception and discrimination for video anomaly detection[J]. *Neural Networks*, 2024, 172: 106138. DOI: 10.1016/j.neunet.2024.106138.
- [12] Y. Su, Y. Tan, S. An, M. Xing, and Z. Feng. Semantic-driven dual consistency learning for weakly supervised video anomaly detection[J]. *Pattern Recognition*, 2025, 157: 110898. DOI: 10.1016/j.patcog.2024.110898.
- [13] D. P. M. Huy, G. N. Nguyen, and D.-N. Le. A Hybrid Deep Learning Approach for Real-Time Cheating Behaviour Detection in Online Exams Using Video Captured Analysis[J]. *Computers, Materials & Continua*, 2026, 86(3): 48. DOI: 10.32604/cmc.2025.070948.
- [14] C. Park, D. Kim, M. A. Cho, M. Kim, M. Lee, S. Park, and S. Lee. Fast video anomaly detection via context-aware shortcut exploration and abnormal feature distance learning[J]. *Pattern Recognition*, 2025, 157: 110877. DOI: 10.1016/j.patcog.2024.110877.
- [15] Y. Kim and Y.-G. Kim. MPE: Multi-frame prediction error-based video anomaly detection framework for robust anomaly inference[J]. *Pattern Recognition*, 2025, 164: 111595. DOI: 10.1016/j.patcog.2025.111595.