applied sciences

MDPI

Article

An Auditable Downstream Safety-Intelligence Framework for Manufacturing Workplaces: Reliability-Aware PPE Evidence, Regulation-Grounded Retrieval, and Role-Specific Decision Support

Dongkyu Yun [1], Wooseok Choi [1], Joonghun Cho [1] and Sanghyun Choi [1,2,*]

[1] Department of Big Data, Chungbuk National University, Cheongju 28644, Republic of Korea; dongkyu.yun@chungbuk.ac.kr (D.Y.); cdt3017@naver.com (W.C.); jojh87@chungbuk.ac.kr (J.C.)
[2] Department of Management Information Systems, Chungbuk National University, Cheongju 28644, Republic of Korea
* Correspondence: chois@cbnu.ac.kr

Featured Application

Auditable transformation of multimodal workplace-safety evidence into reliability-aware PPE information, regulation-grounded context, and role-specific decision support for manufacturing safety management.

Abstract

Vision-based industrial safety monitoring increasingly integrates heterogeneous accidents, personal protective equipment (PPE), and semantic evidence, yet preserving their reliability, uncertainty, and provenance throughout downstream decision support remains challenging. This study proposes an auditable downstream safety-intelligence framework that transforms frozen multimodal perception outputs into structured event evidence. Worker-centered temporal PPE observations are processed into Candidate and Confirmed operational evidence while accident-related uncertainty is preserved. In parallel, regulation-grounded retrieval maps the same evidence to candidate regulatory context from a frozen Korean industrial-safety corpus, and manager- and worker-oriented outputs are generated under distinct information boundaries and independently verified. On a fixed internal benchmark of 150 videos and 3000 frames, the Full-clean configuration achieved 98.23% frame-level accuracy (Macro-F1, 97.47%) and 98.67% video-level accuracy, although the video-level difference from the intermediate configuration was not statistically significant. PPE-violation detection based on the Raw temporal state achieved an F1-score of 99.31% on the ground-truth normal-scene stratum. Controlled Legal-RAG regression reproduced all 117 invariant cases and behaved as predefined across the 32 query-invariant evidence-change cases and the single query-changed case. All 234 notification–report links, 44 PPE semantic checks, and 82 verifier findings satisfied predefined validation criteria. These findings constitute an internal feasibility validation of provenance-preserving, human-review-oriented safety intelligence rather than independent external validation or evidence of field performance.

Keywords: industrial safety monitoring; multimodal safety evidence; personal protective equipment; retrieval-augmented generation; regulation-grounded retrieval; role-specific decision support; evidence provenance; auditability

Check for updates

Academic Editor: Dimitris Mourtzis

Received: 29 September 2026
Revised: 26 October 2026
Accepted: 2 November 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In high-risk industrial environments, including manufacturing and construction, considerable research has focused on the automated recognition of worker behaviors, accident situations such as falls and fires, personal protective equipment (PPE) use, and other unsafe conditions using closed-circuit television (CCTV) and computer vision [1,2]. Deep-learning-based object detection, pose estimation, tracking, and scene analysis have extended industrial safety surveillance toward continuous and quantitative visual monitoring, and recent studies have identified computer vision and artificial intelligence (AI) as key technologies for hazard detection, PPE monitoring, and proactive safety management [2–4]. However, the practical use of AI for industrial safety does not end with accurately detecting or classifying hazardous situations. Information generated by different perception sources must also be associated with the same safety event and transformed into safety information—including event and risk summaries, PPE evidence and its uncertainty, candidate regulatory context for managerial review, and immediate protective-action information for workers—while preserving the reliability and uncertainty of the underlying evidence. Accordingly, an important next challenge in vision-based safety monitoring is not only to improve perception performance, but also to determine how the resulting multimodal safety evidence can be transformed into reliable, traceable, and operationally usable safety intelligence.

Industrial safety monitoring has increasingly moved beyond single-frame object recognition toward contextual and temporal understanding. Approaches have been proposed that use Vision–Language Models (VLMs) to jointly interpret global-scene and region-level safety information [5,6], exploit temporal evidence to reduce frame-level false alarms [7], and associate worker identity with PPE status to construct continuous safety information [8]. Although these developments provide richer safety evidence, high accuracy or F1-score values, or contextual reasoning capability alone, do not ensure that the event identity, source provenance, and uncertainty of heterogeneous perception outputs remain consistent throughout downstream processing. Rather than collapsing heterogeneous perception outputs into a single final label, an event-level representation is therefore needed to structurally preserve accident-related, PPE-related, semantic, and routing evidence under a common event identity.

This reliability issue is particularly important in PPE monitoring. In real industrial scenes, PPE detection can become unstable because of occlusion, scale variation, worker overlap, movement, and posture changes [9]. Recent studies have therefore sought to address the limitations of single-frame detection by associating worker identity with PPE status or incorporating temporal evidence [7,8,10]. In safety-critical monitoring, however, an observation that "PPE was not detected in a particular frame" should not necessarily be treated as equivalent to the operational claim that "the worker was not wearing PPE." In particular, under conditions of reduced visual reliability, such as fall-like postures or partial occlusion, evidence reliability must be managed so that transient non-detection is not escalated into a more definitive claim [11]. PPE evidence should therefore be interpreted within a worker-centered temporal context, while the certainty and component specificity supported by the source evidence should be preserved in downstream information.

Industrial safety monitoring also cannot be considered independently of its regulatory context. Responsibility for industrial accidents has been addressed not only in terms of immediate workplace safety measures but also from the perspectives of organizational safety governance and management responsibility. For example, the United Kingdom's Corporate Manslaughter and Corporate Homicide Act 2007 treats the manner in which an organization's activities are managed or organized, together with the involvement of senior management, as important elements of organizational liability in cases involving serious fa-

talities [12]. In Korea, the implementation of the Serious Accidents Punishment Act (SAPA) has likewise strengthened corporate responsibility and the safety-management duties of responsible managing officers [13,14]. In light of these developments, AI-based safety monitoring should support not only hazard detection but also the review of regulatory information that may be relevant to an observed event. Retrieval-Augmented Generation (RAG) has emerged as an approach for grounding generated information by retrieving external knowledge sources [15], and studies have begun to connect regulatory knowledge with RAG in construction safety [16]. However, retrieving regulatory information must be distinguished from automatically determining the violation or applicability of a specific provision, legal responsibility, or penalties, and practical deployment requires attention to source traceability and human oversight [17,18]. Safety-critical decision support should therefore link observed safety evidence to potentially relevant candidate regulatory references in a repeatable and provenance-preserving manner under a fixed configuration, while leaving final legal judgment to human review.

Model-level explainability or interpretability and system-level auditability address different questions in safety-critical AI. The former concerns understanding or explaining the basis of a model prediction, whereas the latter concerns whether the evidence, processing steps, provenance, and downstream transformations can be reconstructed and independently checked. Trustworthy AI encompasses not only explainability but also robustness, transparency, accountability, traceability, and auditability [19,20]. Human-centered decision support likewise emphasizes maintaining human oversight by providing role-appropriate information rather than replacing human judgment [21]; in the present safety context, this includes event evidence and regulatory context for managers and immediate accident/PPE evidence and protective-action guidance for workers. Role-specific output should therefore be viewed not simply as producing longer or shorter versions of the same report, but as an information-boundary problem concerning which parts of the same evidence should be communicated, and to what extent, to each audience. When natural-language generation is used, it should also be possible to verify that the generated output does not strengthen claims beyond the source evidence, that retrieved references are preserved, and that information from different events or audiences is not mixed. The importance of such source-to-output provenance and auditability has also been emphasized in other safety-critical decision-support domains [22].

Taken together, prior research has advanced vision-based hazard recognition, PPE monitoring, temporal and contextual safety analysis, regulatory knowledge retrieval, and trustworthy decision support as important but largely distinct research directions [2,9,15,19]. However, placing these capabilities within a single application is different from explicitly managing the evidence lifecycle from perception to final decision-support output. In particular, comparatively less attention has been given to jointly structuring heterogeneous frame-level evidence under a common event identity, preserving PPE-related uncertainty throughout downstream processing, separating the provenance of visual evidence from that of regulatory references, applying different audience-specific information boundaries to the same event, and independently verifying this overall transformation. Accordingly, rather than proposing a new standalone perception architecture or a general-purpose legal reasoning model, this study uses previously trained and evaluated perception outputs as frozen multimodal safety evidence and proposes an auditable downstream safety-intelligence framework that transforms this evidence into operational safety decision support.

Accordingly, the technical contribution of this study lies in controlling and verifying the downstream evidence lifecycle across heterogeneous safety modules rather than in introducing a new standalone perception or retrieval model. The main contributions of this study are fourfold. First, we construct a structured event representation that associates

accident-related, PPE-related, semantic, and routing evidence with the same event identity, thereby avoiding the forced integration of heterogeneous evidence into a single classification label. Second, we establish a reliability-aware PPE evidence policy that distinguishes Candidate, Confirmed, and accident-related uncertainty based on worker-centered temporal PPE observations and model-derived accident conditions, and we apply semantic grounding so that downstream language does not exceed the component specificity or certainty supported by the source evidence. Third, we perform deterministic event-to-query routing from the same structured event evidence and independently retrieve event-specific Direct regulatory candidates and Governance context from a frozen regulatory corpus, without using the retrieved information for automated legal determination. Fourth, we transform the same source event into manager-oriented and worker-oriented safety information under distinct audience-specific information boundaries, and independently verify PPE semantic consistency, regulatory-reference preservation, audience/event integrity, and worker-facing legal-information isolation through end-to-end validation separated from report generation. Through these contributions, this study extends vision-based industrial safety monitoring from hazard recognition toward evidence-grounded, reliability-aware, regulation-grounded, role-specific, and auditable safety decision support.

2. Related Work

2.1. Vision-Based Industrial Safety Monitoring and Multimodal Safety Evidence

Computer-vision-based industrial safety monitoring has evolved toward the automated identification of workers, equipment, unsafe behaviors, and hazardous conditions from images and videos. Research initially focused on object detection, tracking, and action recognition, but has since expanded to PPE monitoring, fall-risk recognition, and worker–equipment interaction [23–25]. In manufacturing environments, multi-camera-based factory safety monitoring [26] and CCTV-based industrial accident detection [27] have also been proposed, along with approaches that address multiple safety tasks within a unified computer-vision monitoring framework [28]. These developments indicate that industrial CCTV can serve not only as a recording medium but also as a continuous source of safety-related visual evidence.

Recent research has increasingly moved beyond individual object detection to incorporate spatial, temporal, and semantic context. Representative examples include PPE interpretation using worker spatial context and depth information [29], spatiotemporal risk analysis based on multi-camera tracking and trajectory prediction [30], mitigation of frame-level false alarms through temporal windows [7], and the association of worker re-identification with PPE status [8]. Multimodal VLM-based approaches have further enabled context-aware safety interpretation [31,32], while semantic reasoning that combines global-scene and region-level evidence has extended safety evidence into natural-language information [6].

Through these developments, the analytical focus of industrial safety monitoring has expanded from individual objects and frames toward worker-, event-, and context-centered evidence. However, prior work has primarily emphasized detection or classification performance, temporal stability, or the quality of semantic outputs. How heterogeneous perception outputs should preserve a common event identity, reliability level, and provenance throughout downstream safety-management processes therefore remains a distinct evidence-management problem.

2.2. Reliability-Aware PPE Monitoring and Uncertainty Handling

PPE monitoring has evolved from the detection of individual items such as helmets and safety vests toward worker–PPE association and proper-use assessment [33,34]. Non-

compliance identification based on geometric relationships between workers and hardhats [35], together with systematic reviews of PPE monitoring [9], has highlighted the effects of posture, occlusion, camera conditions, and worker identification on practical PPE monitoring. Recent studies have also addressed challenging conditions such as small-scale targets and occlusion [36,37]. Approaches including pose-guided body-region anchoring [38] and worker-specific PPE assignment under occlusion [39] explicitly account for the relationship between a PPE object and the corresponding worker.

Temporal continuity is also important in video environments because movement, viewpoint changes, and temporary occlusion can cause the observed PPE state of the same worker to vary across frames. Previous studies have addressed this problem through PPE status tracking based on worker re-identification [8,40], helmet-state stabilization using preceding-frame information [41], and the mitigation of short-term detection dropout through temporal consistency [42]. Temporal stabilization, however, should be distinguished from preserving evidence uncertainty. In particular, immediately escalating transient PPE non-detection caused by fall-like postures or severe occlusion into persistent non-wearing or non-compliance can produce a claim stronger than the underlying source evidence supports.

Research has also sought to extend PPE monitoring into semantic information. For example, semantic PPE comparison using pose-based worker-region extraction and image captioning has been proposed [43]. Such capabilities can transform PPE evidence into human-readable information; however, when component identity or visual reliability is uncertain, the specificity and certainty of generated claims should not exceed what the source evidence supports. Accordingly, downstream safety decision support should preserve not only PPE detection results but also the reliability and uncertainty of worker-linked temporal evidence.

2.3. Regulation-Grounded Retrieval for Industrial Safety Decision Support

Industrial safety management requires the consideration of observed hazards together with relevant regulations, safety requirements, and management duties. Research has therefore explored the transformation of regulatory text into machine-readable representations for automated compliance checking, including semantic and ontology-based rule representation [44], NLP- and machine-learning-based compliance approaches [45], and relation extraction and knowledge-graph representation of safety regulations [46].

Even before the emergence of RAG-based approaches, NLP had been applied to automatically retrieve and associate safety requirements from project documents [47]. The development of RAG has enabled regulatory knowledge to be connected more directly with safety decision support. Approaches have included grounding generated outputs through evidence retrieved from external corpora [15], multi-hop retrieval over OSHA regulations for construction safety [16], and onsite safety assistants that connect observed hazards with safety regulations [48]. Regulatory information has consequently evolved from an independent document-search resource toward a knowledge-grounded resource that can be associated with visual or structured safety evidence.

Regulatory retrieval, however, should be distinguished from legal determination. In Legal RAG, source traceability and provision-level citation have been discussed as important requirements [18]. More recent work on legal-domain language models has also explored the use of RAG to support more trustworthy outputs [49]. Retrieval configuration has additionally been reported to affect factual consistency [50]. Furthermore, practical applications of automated regulatory compliance continue to emphasize the importance of human oversight rather than fully autonomous determination [17]. Industrial safety decision support should therefore provide relevant regulatory articles as traceable candidate context without

automatically determining the legal applicability of a provision, the existence of a violation, responsibility, or applicable penalties.

2.4. Auditable and Role-Specific AI Safety Decision Support

In safety-critical AI, human oversight, transparency, accountability, and traceability are treated as important system-level requirements alongside model accuracy and explainability [20]. Human-Centered AI research likewise emphasizes connecting data, models, and decision logic to human oversight rather than replacing human decision-makers [21]. Explanation alone has also been shown not to automatically guarantee appropriate user understanding or trust [51]. In human–AI collaboration, over-reliance on automated recommendations and automation bias have additionally been discussed as distinct human-oversight concerns [52].

In the present study, interpretability and auditability are treated as distinct properties. Interpretability concerns the extent to which the basis or internal reasoning of a model prediction can be understood or explained, for example through model structure, feature attribution, or other explanatory representations. Auditability, by contrast, concerns whether a system-level transformation can be reconstructed and verified from preserved inputs, event identities, model and configuration versions, decision rules, retrieved references, output linkages, and validation records. Accordingly, a pipeline may be auditable at the artifact and process level without making the internal operation of each neural component intrinsically interpretable. Conversely, model interpretability alone does not guarantee preservation of provenance or end-to-end output integrity. The present study focuses on the latter system-level auditability and does not claim intrinsic interpretability of its neural perception or language-model components.

The purpose and level of information required from the same AI evidence can differ across stakeholders. Research on stakeholder-tailored explanations has shown that user groups may require different levels of detail and supporting evidence [53]. In occupational safety, managers may require risk and management-related evidence, whereas frontline workers may benefit more from immediate feedback and actionable guidance [54]. Role-specific decision support should therefore consider audience-specific operational purposes and information boundaries rather than merely providing detailed and abbreviated versions of the same report.

When natural-language generation is involved, consistency between source evidence and generated claims is also important. LLM-based generation can produce information that conflicts with or is unsupported by the source [55], and in safety-critical contexts, temporary uncertainty should not be escalated into a definitive safety claim or transformed into a stronger legal interpretation than the retrieved regulation supports. Traceability approaches have therefore been proposed to record data, process, and artifact provenance [56,57], while RAG audit trails have been used to connect retrieved evidence with final recommendations [22]. These studies indicate that, independently of the linguistic quality of a final output, source–event correspondence, reference preservation, and evidence-to-output lineage should also be subject to verification.

2.5. Research Gap and Positioning of the Present Study

Taken together, the literature reviewed in Sections 2.1–2.4 shows that industrial safety AI has advanced in worker-centered temporal PPE monitoring [8,41], VLM-based contextual safety reasoning [6], RAG systems connecting visual evidence with regulatory knowledge [58–60], stakeholder-specific information design [53], and auditable RAG-based decision support [22], as summarized in Table 1. The research gap addressed in this study therefore does not arise from the individual absence of these capabilities or merely from whether VLM and RAG components can be combined.

Table 1. Compact comparison of representative studies related to the positioning of the present study.

Study	Primary Focus	Evidence/Regulatory Emphasis	Role-Specific/Provenance Emphasis
Cheng et al. [8]	Worker Re-ID and PPE monitoring	Worker identity and temporal PPE history	Role-specific policy and full downstream audit not primary
Zhang et al. [41]	Helmet state tracking	Temporal stabilization under occlusion and missed detection	Downstream uncertainty/provenance not primary
SafeVision [6]	Context-aware VLM safety reasoning	Scene- and region-level semantic grounding; regulatory retrieval not primary	Human-readable/VQA output; explicit manager–worker boundary and full downstream audit not primary
Guo et al. [58]	Visual-LLM and RAG for safety compliance	Context-aware visual reasoning with safety-rule retrieval and compliance analysis	Retrieval/compliance evaluation; explicit manager–worker boundary and full evidence-lifecycle audit not primary
Katsampiris et al. [59]	MLLM and RAG for manufacturing risk assessment	Structured multimodal inference with ISO/OSHA grounding	Citation verification and structured reporting; explicit manager–worker separation not primary
Orooje et al. [60]	Trustworthy VLM hazard detection with RAG	Visual grounding, hallucination mitigation, and regulation-aware verification	Traceable compliance support; cross-audience integrity not primary
Kim et al. [53]	Stakeholder-tailored XAI	Stakeholder-specific explanation rather than industrial perception/regulatory retrieval	Explicit stakeholder-specific information design
Alu and Oluwadare [22]	Auditable RAG-based decision support	Source-grounded domain-knowledge retrieval	Explicit source verification and audit trail
Present study	Downstream industrial safety-intelligence framework	Worker-centered PPE reliability and independent Direct/Governance regulatory retrieval without automated legal determination	Independent manager/worker information boundaries with frozen evidence/reference provenance and separated end-to-end verification

The central issue addressed in this study is the evidence lifecycle from perception to final decision-support output. More specifically, the problem involves structuring heterogeneous visual evidence under a common event identity, preserving accident-related ambiguity and PPE reliability in downstream language, separating the provenance of visual evidence from that of candidate regulatory references, applying distinct information boundaries for managers and workers to the same source event, and independently verifying these transformations. Prior studies have made important advances in each of these areas; the present study focuses on jointly managing event identity, evidence reliability, regulatory provenance, audience-specific information boundaries, and final-output integrity within a single frozen downstream analytical process.

Consistent with this positioning, the present study does not propose a new detector, VLM, or general-purpose RAG architecture. Its technical contribution instead lies in the downstream control architecture that manages how heterogeneous safety evidence is represented, transformed, routed, and verified after perception. Operationally, this positioning is realized through five design elements: the framework (i) preserves accident, PPE, semantic, and routing evidence under a common event identity rather than collapsing them into a single state; (ii) maintains worker-centered PPE reliability and accident-related uncertainty as explicit operational evidence; (iii) performs PPE processing and regulation-grounded retrieval as independent evidence pathways from the same structured event record; (iv) applies explicit manager/worker information-boundary policies to the same source event; and (v) verifies evidence/reference provenance and audience integrity through a validation

process separated from report generation. The novelty therefore concerns an auditable evidence-lifecycle architecture and its deterministic control/verification mechanisms rather than a new perception backbone or general-purpose retrieval model.

Accordingly, the research question is not whether perception accuracy can be further maximized or automated legal reasoning can be strengthened, but whether already generated multimodal safety evidence can be transformed into regulation-grounded, role-specific, and auditable safety information without compromising its reliability, uncertainty, or provenance. The validation in this study is therefore limited to evidence preservation, retrieval repeatability under the frozen configuration, and information integrity; it does not evaluate the final legal applicability of retrieved regulations or actual manager/worker comprehension, decision quality, or usability. In addition, because the study uses a fixed internal benchmark with historical provenance and a Korean industrial-safety corpus, the results should not be interpreted as demonstrating independent external generalization.

Accordingly, the present study is framed as an internal feasibility validation of the proposed downstream evidence lifecycle rather than as an independent external validation study.

3. Materials and Methods

3.1. Overall Study Framework

This study proposes an auditable downstream safety-intelligence framework that transforms frozen multimodal perception outputs into structured event evidence, reliability-aware PPE evidence, regulation-grounded candidate context, and role-specific decision support (Figure 1). The framework does not introduce a new perception architecture; instead, it controls how previously generated safety evidence is represented and propagated through downstream stages.

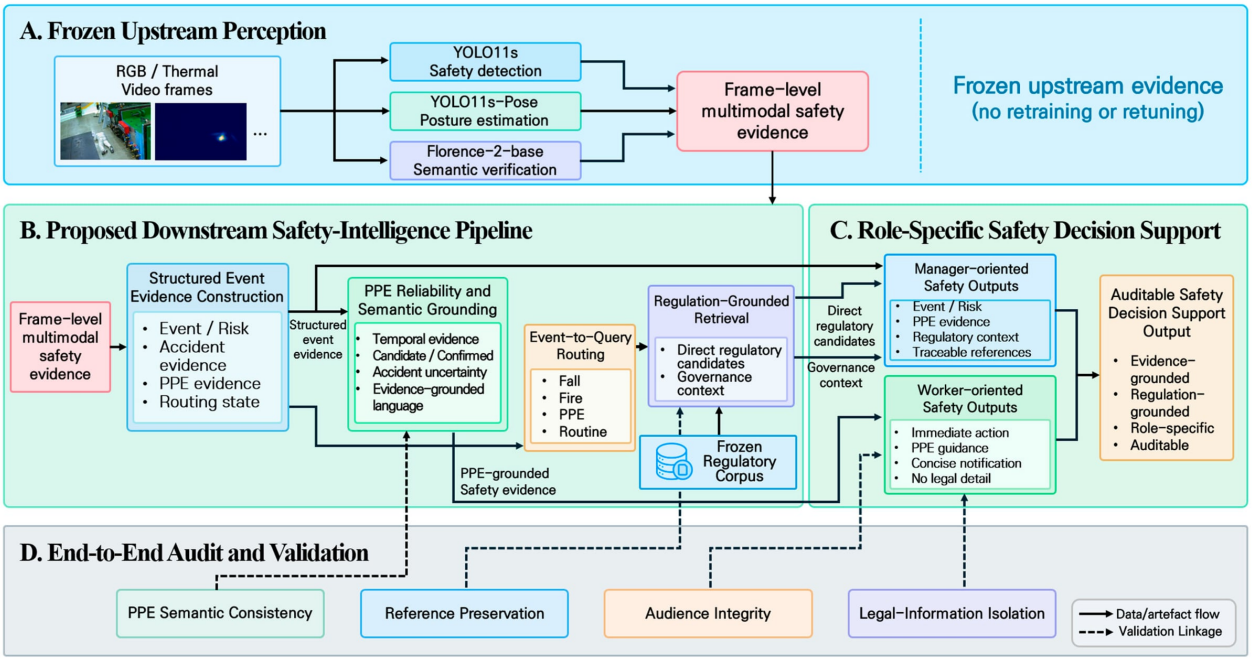


Figure 1. Overall architecture of the proposed downstream safety-intelligence framework. Frozen multimodal safety evidence is transformed into structured event evidence, which feeds independent PPE reliability/semantic-grounding and regulation-grounded retrieval pathways, followed by role-specific decision support and end-to-end audit and validation. Solid arrows indicate data/artefact flow, and dashed arrows indicate validation linkage; colors are used only to group the panels and modules.

Structured event evidence serves as the common input to two independent pathways: worker-centered PPE reliability/semantic grounding and deterministic regulation-grounded retrieval. The resulting evidence is then expressed under manager- and worker-specific information policies and checked by a separate end-to-end validation layer for PPE semantic consistency, regulatory-reference preservation, audience/event integrity, and worker-facing legal-information isolation. The frozen upstream evidence and downstream configurations were not readjusted on the basis of final evaluation results.

3.2. Dataset and Frozen Multimodal Safety Evidence

3.2.1. Fixed Internal Evaluation Benchmark

The video data used in this study were based on the AI-Hub dataset Data for Smart Manufacturing Facility Safety Monitoring [61]. The source dataset contains RGB and thermal CCTV videos together with object annotations and comprises 20,000 videos and approximately 400,000 extracted images.

For downstream evaluation, a fixed internal benchmark was constructed from Normal, Fall-down, Fall-off, Fire RGB, and Fire Thermal scenarios. The benchmark contained 150 videos and 3000 frames, with 20 evaluation frames selected from each video. For the final three-class evaluation, Fall-down and Fall-off were combined into the Fall Accident class, whereas RGB and thermal fire scenarios were combined into the Fire Accident class. RGB and thermal fire were assigned to the same event class because the downstream taxonomy was defined by the represented safety event rather than by sensing modality: both subsets represent the same report-oriented Fire Accident condition. The source modality remained preserved in the benchmark composition and provenance records but was not treated as a separate safety-event label, and this study did not evaluate RGB-versus-thermal domain generalization as an independent task. The resulting benchmark therefore contained 50 videos each for Normal, Fall Accident, and Fire Accident (Table 2).

Table 2. Composition of the fixed internal evaluation benchmark.

Class	Scenario	Modality	Videos	Frames
Normal	No accident	RGB	50	1000
Fall Accident	Fall-down	RGB	25	500
Fall Accident	Fall-off	RGB	25	500
Fire Accident	General fire	RGB	25	500
Fire Accident	General fire	Thermal	25	500
Total	-	-	150	3000

The evaluation scope was deliberately restricted to this compact event taxonomy because the primary objective of the study was to examine the downstream evidence lifecycle—structured event representation, reliability-aware PPE evidence, regulation-grounded retrieval, role-specific information boundaries, and audit integrity—under a controlled frozen benchmark rather than to establish a comprehensive hazard-recognition taxonomy. Fall and fire provided distinct accident conditions within the fixed benchmark, while helmet and safety vest were the PPE components explicitly labeled and evaluated in the 100-video human subset described in Section 3.4.1. Accordingly, the selected taxonomy should be interpreted as the controlled scope of the internal feasibility evaluation rather than as a complete representation of manufacturing hazards or PPE requirements.

Because this benchmark had been used in historical analyses, it was treated strictly as a fixed internal evaluation benchmark and not as a pristine blind test set or an independent external validation set. Its role in the present study was to support an internal feasibil-

ity evaluation of the frozen downstream safety-intelligence framework under controlled and traceable conditions. Once the final benchmark was fixed, no samples were added or removed and the class distribution was not altered on the basis of downstream results. The same 150-video benchmark was used for safety-event recognition, structured-event construction, and regulation-grounded downstream evaluation, whereas PPE reliability was evaluated separately on the 100-video human subset described in Section 3.4.1. Consequently, results obtained from this benchmark should not be interpreted as evidence of external-domain generalization or prospective field performance.

3.2.2. Frozen Upstream Perception Configuration

The frozen upstream perception component consisted of YOLO11s-based safety detection [62], YOLO11s-Pose-based worker posture estimation, and Florence-2-base-based semantic verification [63]. Because the objective of this study was not to further optimize the upstream perception architecture, the final model checkpoints and inference configurations were fixed before downstream analysis.

For safety detection, YOLO11s was retrained after removing direct training overlap with the fixed benchmark, while posture information was generated using pretrained YOLO11s-Pose. Florence-2-base was likewise retrained using training/internal-validation data that had been audited to exclude direct sample overlap with the benchmark. The fixed benchmark was not used for Florence-2 training or internal model selection. These corrections were intended to remove direct training exposure from the final perception models and do not redefine the historically used benchmark as an independent external validation set.

The effect of provenance correction was evaluated in Section 4.1 by comparing three configurations: the Historical Original legacy pair, an intermediate configuration combining clean YOLO with the legacy VLM, and the Full-clean configuration combining clean YOLO with overlap-audited clean Florence-2. The Historical and intermediate configurations were used only as references for provenance analysis. All downstream analyses from Section 3.3 onward used only frozen multimodal safety evidence generated by the Full-clean configuration.

Detailed operating parameters for safety detection, pose estimation, semantic verification, and accident-related evidence generation are provided in Supplementary Table S3. After the final configuration was fixed, model weights, decision thresholds, keyframe intervals, and accident-related decision criteria were not readjusted on the basis of the benchmark or downstream results. PPE-specific worker association and Candidate/Confirmed evidence generation were treated as a separate downstream reliability policy and are defined in Section 3.4.

3.2.3. Evaluation Metrics and Statistical Analysis

Safety-event recognition was evaluated at both the frame and video levels using the same fixed benchmark and evaluation units for the Historical Original, intermediate, and Full-clean configurations. The primary metrics were Accuracy and Macro-F1, while class-level prediction behavior and residual errors were examined through confusion-based analysis.

Uncertainty in paired performance differences was evaluated using a video-cluster bootstrap that accounted for clustering among frames extracted from the same video. Videos were resampled with replacement, and all evaluation frames belonging to each resampled video were included as a cluster. This procedure was repeated 10,000 times. The 2.5th and 97.5th percentiles of the resulting Accuracy and Macro-F1 differences were used to construct 95% confidence intervals. The same procedure was applied to video-level differences, with videos used as the resampling unit. A fixed random seed was used for all bootstrap analyses.

A two-sided exact McNemar test was applied to paired classification changes. Because frame-level observations within the same video are not independent, the frame-level McNemar result was used only as descriptive supporting analysis, whereas paired frame-level uncertainty was primarily characterized using the video-cluster bootstrap. For the video-level McNemar test, each video was treated as one paired observation. The video-level comparison was not interpreted as evidence of improvement, equivalence, or non-inferiority when the paired test and bootstrap interval did not support such a conclusion. Because multiple performance metrics and paired comparisons were examined and no multiplicity adjustment was applied, the inferential comparisons reported in this study were treated as exploratory. Accordingly, unadjusted *p*-values and bootstrap confidence intervals are reported to characterize the direction and uncertainty of the observed differences rather than as confirmatory evidence of superiority.

For PPE reliability, Accuracy, Precision, Recall, Specificity, F1-score, and Balanced Accuracy were calculated on ground-truth normal-scene human frames. Ninety-five percent confidence intervals for the primary metrics were estimated using 10,000 cluster-bootstrap resamples with the video as the resampling unit. Regulation-grounded retrieval regression and role-specific output integrity were evaluated using exact event/reference counts and satisfaction of a predefined deterministic validation contract.

A deployment-oriented computational profile was additionally reconstructed from the preserved logs of the frozen final runs without repeating model inference or report generation. The clean Florence-2 checkpoint used in the Full-clean configuration had been trained in a CUDA-enabled Linux/WSL2 environment using an NVIDIA GeForce RTX 5090 GPU with approximately 32 GB of device memory, CUDA 12.8, and PyTorch 2.6.0a0+ecf3bae40a.nv25.01. The subsequent frozen Full-clean evaluation was executed on CUDA device 0; however, because a separate hardware snapshot was not preserved for that evaluation run, the reported throughput was not interpreted as a hardware-normalized benchmark. For the Full-clean perception configuration, overall wall-clock throughput was calculated from the complete processing time of the 3000 evaluation frames, and the task-level Florence-2 latency fields recorded for triggered frames were summarized separately. For the API-based role-specific report-generation stage, computational demand was characterized using the preserved input/output token counters and the elapsed wall time of the uninterrupted resumed generation segment. The latter represents observed end-to-end report-generation throughput, including API communication and retry overhead, rather than isolated model-inference latency. Because the historical provider billing tariff was not preserved as part of the frozen experimental configuration, monetary API cost was not retrospectively estimated; token usage is reported instead as the recorded resource-consumption measure. These measurements are specific to the evaluated experimental configuration and were not treated as a controlled production-deployment benchmark.

3.3. Structured Event Evidence Construction

The frozen frame-level multimodal safety evidence generated in Section 3.2 includes worker/object location, posture, accident-related evidence, PPE observations, and semantic-verification information. To use this information consistently as an event-level input for downstream PPE processing, regulation-grounded retrieval, and role-specific safety-information generation, the frame-level evidence was transformed into structured event evidence at the video/event level, as shown in Panel B of Figure 1. This stage does not introduce a new classifier; rather, it structures and aggregates frozen upstream evidence according to a common event identity.

Each event record preserves the final video-level recognition class, report-oriented event type, fall- and fire-related evidence, PPE Candidate/Confirmed evidence and accident-

related uncertainty, supporting semantic evidence, and regulatory-routing information under the same event identity. PPE-related fields represent event-level aggregation of the operational evidence generated by the worker-centered temporal policy defined in Section 3.4; the corresponding decision rules are specified separately in that section.

The video-level recognition class was distinguished from the report-oriented event type. The recognition class represents the Normal, Fall Accident, or Fire Accident output generated by the frozen upstream perception component, whereas the report-oriented event type is an operational representation used for downstream safety-information generation. Consequently, a separate PPE-related state can be assigned even when the recognition class is Normal if PPE-related evidence is present. The report-oriented event type and operational risk level were determined using a deterministic priority rule that prioritizes Fire Accident and Fall Accident, followed sequentially by PPE Confirmed and Candidate evidence (Table 3).

Table 3. Deterministic mapping from structured safety evidence to report-oriented event type and operational risk level.

Evidence Condition	Report-Oriented Event Type	Operational Risk Level
Video-level class = Fire Accident	Fire Accident	Critical
Video-level class = Fall Accident	Fall Accident	High
Video-level class = Normal and Confirmed PPE evidence > 0	PPE non-compliance	Medium
Video-level class = Normal, no Confirmed PPE evidence, and Candidate PPE evidence > 0	PPE non-detection candidate	Low-Medium
None of the above	No reportable accident	Low

This priority rule is used only to select one representative event type and one operational risk level; it does not remove evidence that coexists within the same event. In particular, PPE non-compliance is a report-oriented operational state assigned on the basis of Confirmed PPE evidence. It does not imply the absence of unknown_during_accident evidence or establish a legal or workplace compliance violation. For example, when a Fall Accident coexists with PPE Candidate evidence or accident-related uncertainty, the representative event state is set to Fall Accident while the associated PPE evidence remains preserved in the structured event record.

Structured event evidence was also used as the common input to the deterministic event-to-query routing described in Section 3.5.2. Regulatory query categories generated from accident and PPE evidence were associated with and transmitted under the same event identity, while the detailed multi-query activation rules are defined in Section 3.5.2. Each video in the fixed benchmark was mapped to one event, producing a total of 150 structured event records. After the representation was frozen, event classification, operational risk level, PPE evidence, and routing state were not modified post hoc on the basis of downstream results.

3.4. PPE Reliability and Semantic Grounding

3.4.1. PPE Ground Truth and Observation

To evaluate the reliability of the PPE evidence used in the structured event representation defined in Section 3.3, this study distinguished among ground-truth PPE states, frame-level observations generated by the Full-clean perception configuration, and operational PPE evidence generated by the temporal policy described in Section 3.4.2. The overall

PPE evaluation set consisted of the 100-video human subset of the fixed internal benchmark, comprising 2000 frames: 50 Normal videos (1000 frames), 25 Fall-down videos (500 frames), and 25 Fall-off videos (500 frames). For evaluation purposes, these 2000 frames were further stratified using the ground-truth scene-status annotation into 1552 normal-scene frames and 448 accident-scene frames. This stratification was used only to define evaluation strata and was not used as an input to PPE inference or to the model-derived accident gate.

For each frame, helmet and safety-vest status were independently labeled as either worn or not worn. A frame was defined as PPE-violation positive when at least one component was labeled not worn. Of the 2000 frames, 1680 (84.0%) were PPE-positive, with the detailed distribution shown in Table 4.

Table 4. Distribution of frame-level PPE ground-truth states in the human subset.

PPE Ground-Truth State	Frames	Proportion
Helmet not worn only	321	16.05%
Safety vest not worn only	1059	52.95%
Both helmet and safety vest not worn	300	15.00%
Both helmet and safety vest worn	320	16.00%
Any PPE violation	1680	84.00%
Total	2000	100.00%

PPE ground truth was strongly clustered at the video level: all evaluation frames were PPE-positive in 84 videos, whereas all evaluation frames were PPE-negative in the remaining 16 videos. Accordingly, frame-level uncertainty was evaluated using the video-cluster bootstrap described in Section 3.2.3.

Frame-level PPE observations were generated using the Full-clean upstream configuration, and worker evidence was associated with helmet and safety-vest evidence at the component level. For evaluation alignment in frames containing multiple workers, a fixed primary-worker selection rule was applied. Detailed scoring and threshold rules for worker selection and worker–PPE association are provided in Supplementary Table S1. The 2000-frame human subset was not reduced by prediction quality, model confidence, or downstream PPE results. Instead, the predefined ground-truth scene-status annotation was used only to separate the evaluation into two strata: 1552 ground-truth normal-scene frames were used for the primary frame-level PPE reliability metrics, whereas the remaining 448 ground-truth accident-scene frames were retained for accident-aware uncertainty and condition-specific analyses. Thus, the 448 accident-scene frames were not excluded from the study. Ground-truth normal/accident scene-status annotations were not used as inputs to PPE inference or the accident-aware policy; the accident gate was determined solely from fall-related evidence generated by the frozen perception configuration.

3.4.2. Operational PPE Evidence Tiers

To avoid immediately interpreting single-frame PPE non-detection as non-compliance, observations from the same worker track were accumulated temporally. Candidate and Confirmed evidence were then generated after preserving uncertainty under model-derived accident conditions. These are not outputs of separate classifiers; rather, they are operational evidence generated by applying different reliability criteria to the same worker-centered temporal evidence.

Frame-level PPE observations were associated with individual workers through the worker–PPE association procedure defined in Section 3.4.1. Temporal evidence was con-

structed from the most recent available observations within the same worker track, up to a maximum of 10 observations, rather than from 10 consecutive video frames.

Let c denote the PPE component, r the worker track, and j an observation index within that track.

Let $X_{c,r,j}$ indicate whether PPE component c is validly detected for worker track r at observation j, and let $W_{r,k}$ denote the set of the most recent observations available at the k-th observation of worker track r. Present and missing counts were defined as follows:

$$P_{c,r}^{(k)} = \sum_{j \in W_{r,k}} X_{c,r,j},\ M_{c,r}^{(k)} = |W_{r,k}| - P_{c,r}^{(k)},\ |W_{r,k}| \leq 10 \quad (1)$$

Here, $X_{c,r,j}$ takes a value of 1 when PPE component c is validly detected at the j-th observation of worker track r, and 0 otherwise. $W_{r,k}$ is the index set of the most recent observations actually accumulated in the worker track at observation k, up to a maximum of 10 observations. $P_{c,r}(k)$ and $M_{c,r}^{(k)}$ represent the present and missing counts within this temporal window, respectively. When fewer than 10 observations were available during the early stage of a track, only the observations actually accumulated up to that point were used.

The Raw temporal state was categorized as worn, missing_candidate, not_detected_candidate, or uncertain based on the present and missing evidence. The state was defined as worn when the present count was at least 2. When the present count was less than 2 and the missing count was at least 5, the state was defined as missing_candidate, indicating persistent non-detection. When the present count was less than 2, the missing count was less than 5, and PPE was absent in the current observation, the state was retained as the preliminary not_detected_candidate state. All other cases were categorized as uncertain.

Because PPE observation reliability can temporarily deteriorate under fall-related posture changes or occlusion, a model-derived accident gate was applied. Ground-truth accident labels were not used to activate this gate. Instead, Broad-fall, Fall-down, and Fall-off evidence generated by the frozen perception configuration were combined as follows:

$$G_r^{(k)} = B_{\text{fall},r}^{(k)} \vee F_{\text{down},r}^{(k)} \vee F_{\text{off},r}^{(k)} \quad (2)$$

Here, $G_r^{(k)}$ is a binary model-derived accident gate for worker track r at observation k. $B_{\text{fall},r}^{(k)}$, $F_{\text{down},r}^{(k)}$, and $F_{\text{off},r}^{(k)}$ are the binary indicators of broad-fall, Fall-down, and Fall-off evidence, respectively, generated by the frozen perception configuration. OR denotes the logical OR operation. When the accident gate was active, the worn state was preserved, while all other Raw temporal states were converted to unknown_during_accident to retain accident-related uncertainty. When the gate was inactive, the Raw temporal state was left unchanged.

Candidate evidence was activated when, after accident-aware handling, the component state was different from worn and the missing count was at least 5. When the accident gate was inactive, this condition corresponded to persistent missing_candidate evidence. When the accident gate was active, persistent non-worn evidence could instead be represented as unknown_during_accident while still satisfying the same missing-count criterion. Therefore, unknown_during_accident alone was not sufficient to activate Candidate evidence; the missing-count threshold of at least 5 still had to be satisfied. The preliminary not_detected_candidate state, for which the missing count was below 5, did not activate Candidate evidence. Candidate evidence was treated as review-oriented evidence of persistent non-detection and was not interpreted as confirmed non-compliance.

Confirmed evidence was generated using stricter reliability criteria. It was activated only when, after accident-aware handling, the component state was not worn, the PPE-specific upright criterion was satisfied, the missing count was at least 8, and the present count was 0. The PPE-specific upright criterion was defined as either a valid torso angle of

at least 55° or a worker height-to-width ratio of at least 1.25. This criterion was treated as a PPE reliability condition distinct from the upstream fall-related posture criteria listed in Supplementary Table S3.

Candidate and Confirmed evidence were calculated separately for helmet and safety vest. If the corresponding flag was activated for at least one component, it was reflected in the frame-level PPE evidence. unknown_during_accident, Candidate, and Confirmed do not constitute a mutually exclusive multi-class classification. unknown_during_accident is a component state representing accident-related uncertainty, whereas Candidate and Confirmed are operational flags derived independently from the same temporal evidence under different reliability requirements. Because both Candidate and Confirmed require the post-gate component state to be different from worn, either flag may coexist with unknown_during_accident when its own persistence criteria are satisfied. Candidate requires a missing count of at least 5, whereas Confirmed additionally requires the PPE-specific upright criterion, a missing count of at least 8, and a present count of 0. Candidate and Confirmed may also coexist when both sets of criteria are satisfied. Such coexistence does not mean that one form of evidence replaces or invalidates another; each is preserved independently within the structured event record and is subsequently used according to the report-oriented mapping and regulatory-routing rules defined in Sections 3.3 and 3.5.2.

The overall worker-centered operational PPE evidence procedure is illustrated in Figure 2.

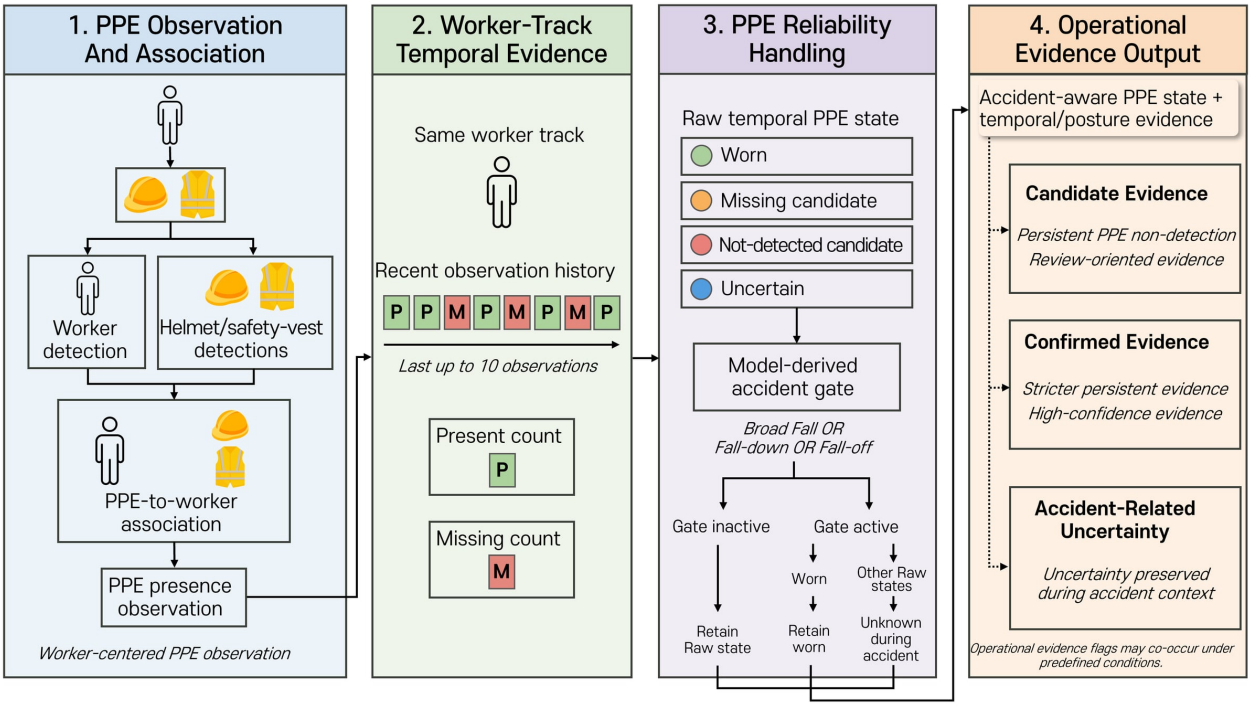


Figure 2. Worker-centered operational PPE evidence policy. Frame-level PPE observations are associated with individual workers and accumulated within each worker track to derive temporal PPE states. A model-derived accident gate preserves uncertainty under accident-like conditions, after which Candidate, Confirmed, and accident-related uncertainty evidence are retained as reliability-aware operational outputs. Solid arrows indicate the processing flow, and dotted arrows indicate the derivation of operational evidence flags from the accident-aware PPE state. In the observation history, P and M denote present and missing PPE observations, respectively. Operational evidence flags and uncertainty states are not globally mutually exclusive; their coexistence follows the criteria in Table 5.

The core frozen criteria for PPE evidence generation are summarized in Table 5. Detailed spatial and scoring rules for worker–PPE association are provided in Supplementary Table S1, while upstream accident-related parameters are provided in Supplementary Table S3.

Table 5. Core frozen criteria for worker-centered operational PPE evidence generation.

Processing Stage	Core Frozen Criterion
Temporal aggregation	Up to 10 recent observations within the same worker track
Raw temporal state	worn: present ≥ 2; missing_candidate: present < 2 and missing ≥ 5; not_detected_candidate: present < 2, missing < 5, and current PPE absent; otherwise, uncertain
Accident-aware handling	Broad Fall OR Fall-down OR Fall-off activates the gate; worn is retained and other Raw states become unknown_during_accident
Candidate evidence	Post-gate state $\neq$ worn (including unknown_during_accident) and missing count ≥ 5
Confirmed evidence	State $\neq$ worn (including unknown_during_accident), upright criterion satisfied, missing count ≥ 8, and present count $= 0$
Semantic-verification trigger	Confirmed evidence or the frozen long-sustained PPE condition activates Florence-2 verification; semantic verification does not modify temporal counts, accident state, or Candidate/Confirmed criteria

The criteria in Table 5 were retained as the frozen configuration used for the final Full-clean evaluation and were not readjusted on the basis of downstream PPE evaluation or report-generation results.

PPE-related Florence-2 semantic verification was selectively activated by either Confirmed evidence or a predefined long-sustained PPE condition. The long-sustained condition was defined as an upright worker with an inactive accident gate, a missing count of at least 8, and a present count of 0. Semantic-verification results did not modify temporal counts, the accident gate, the Raw state, or Candidate/Confirmed criteria. Florence-2 semantic information was therefore used only as auxiliary semantic evidence for connecting generated PPE evidence to downstream language, rather than as post-hoc decision evidence for redefining operational PPE evidence.

3.4.3. PPE Semantic Grounding

Candidate, Confirmed, and accident-related uncertainty were used only as grounding constraints on the component specificity and certainty of downstream PPE-related language; semantic grounding did not retrospectively modify PPE recognition or the operational evidence tiers defined in Section 3.4.2. Component-specific wording was permitted only when corresponding structured visual evidence was available, and Candidate or accident-related uncertainty alone was not escalated into confirmed non-compliance. Regulatory terminology was likewise not used to establish a visual PPE fact. When Confirmed evidence coexisted with uncertainty, both states remained independently preserved for downstream reporting and routing. These constraints were applied to role-specific outputs and evaluated through the semantic-integrity validation in Section 3.7.2.

3.5. Regulation-Grounded Retrieval

To connect the structured event evidence defined in Section 3.3 with industrial-safety regulatory information, an independent regulation-grounded retrieval pathway was constructed. This pathway does not use the semantic-grounding outputs described in Section 3.4 or downstream natural-language expressions as retrieval inputs. Instead, it uses only

the accident-related evidence and PPE Candidate/Confirmed evidence preserved in the structured event record. Accordingly, PPE reliability/semantic grounding and regulation-grounded retrieval are performed independently from the same structured event evidence.

Retrieval-Augmented Generation (RAG) is an approach that connects evidence retrieved from external knowledge sources with language generation to provide grounding and provenance [15]. In this study, retrieval outputs were not used to automatically determine the violation or legal applicability of a specific regulation, assign responsibility, or determine applicable penalties. Rather, candidate regulatory context potentially relevant to the observed safety evidence was provided to manager-oriented decision support. The overall pathway consists of a frozen regulatory corpus, deterministic event-to-query routing, Direct regulatory candidates, and Governance context.

3.5.1. Regulatory Corpus Construction and Freezing

The regulatory corpus was constructed from the official Korean-language texts of industrial-safety-related statutes and regulations provided by the Korean National Law Information Center [64]. The corpus included six regulatory instruments relevant to manufacturing-workplace safety management, and the original Korean texts were used for retrieval. To preserve temporal provenance, the statutory versions in force at the time the corpus was constructed were used (Table 6).

Table 6. Regulatory instruments and statutory versions used in the frozen regulatory corpus.

Regulatory Instrument	Statutory Version	Effective Date
Rules on Occupational Safety and Health Standards	Ministry of Employment and Labor Ordinance No. 450	2 March 2026
Occupational Safety and Health Act	Act No. 21374	1 August 2026
Enforcement Decree of the Occupational Safety and Health Act	Presidential Decree No. 36540	1 August 2026
Enforcement Rule of the Occupational Safety and Health Act	Ministry of Employment and Labor Ordinance No. 477	1 August 2026
Serious Accidents Punishment Act	Act No. 17907	27 January 2022
Enforcement Decree of the Serious Accidents Punishment Act	Presidential Decree No. 35805	1 October 2025

The retrieval unit was defined at the individual-article level. For each article, regulatory-instrument identity, article number, title, body text, and source/effective-date provenance were preserved. Fall-, fire-, PPE-, and routine-related metadata required for event routing were also maintained within the same corpus, together with the duty–penalty linkage used by the Governance branch.

A predefined runtime scope filter was applied to the 1251 article-level records in the frozen source corpus. This filter excluded 373 records, resulting in a final retrieval pool of 878 records. The final pool contained no duplicate law–article groups, future-effective marker records, or records with empty article titles (Table 7). The scope filter was applied as a predefined exclusion rule rather than as a post-hoc relevance adjustment based on retrieval results. After the corpus was frozen, article membership, query metadata, penalty linkage, and scope rules were not modified. Detailed scope-filter and retrieval configurations are provided in Supplementary Table S2.

Table 7. Composition and quality checks of the frozen regulatory corpus.

Corpus Property	Frozen Configuration
Regulatory instruments	6
Frozen article-level records	1251
Records excluded by predefined scope filter	373
Final retrieval pool	878
Duplicate law–article groups	0
Future-effective marker records	0
Empty article-title records	0

3.5.2. Event-to-Query Routing

Structured event evidence was converted into regulatory query categories through predefined deterministic routing rather than through queries freely generated by a language model. Only video-level accident recognition and PPE Candidate/Confirmed evidence were used for routing (Table 8).

Table 8. Deterministic mapping from structured event evidence to regulatory query categories.

Structured Evidence Condition	Routed Query Category
Video-level class = Fall Accident	Fall accident prevention and response
Video-level class = Fire Accident	Fire accident prevention and response
PPE Candidate > 0 or PPE Confirmed > 0	PPE compliance
No Fall, Fire, or PPE query activated	Routine monitoring

Routing was implemented as a multi-query structure rather than as a mutually exclusive single-class classification. Consequently, accident-related and PPE-related queries could be activated simultaneously. PPE Candidate/Confirmed evidence coexisting in the structured event record was therefore retained for regulatory routing regardless of the representative event type defined in Section 3.3.

By contrast, unknown_during_accident alone did not activate a PPE compliance query. PPE-related routing was performed only when Candidate or Confirmed evidence was present. Event-level trigger information was retained solely as provenance metadata and was not used for a learned query-generation mechanism or for determining legal applicability.

3.5.3. Direct Regulatory Candidates

The Direct branch retrieves technical and operational regulatory context directly related to the routed safety issue and is separated from the Governance branch, which provides broader management-duty and responsibility context. Candidates were ranked within the frozen 878-record pool using BM25 lexical relevance together with frozen query-type and rule-based signals. Only articles whose combined title and body Direct relevance scores satisfied the minimum criterion in Equation (3) were permitted as Direct candidates.

$$\text{DirectCandidate}(\text{a},\text{e}) = \begin{cases} 1, & S_{\text{title}}(a,e) + S_{\text{body}}(a,e) \geq 1.0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Here, a denotes a candidate regulatory article and e denotes a structured safety event. $S_{\text{title}}(a,e)$ and $S_{\text{body}}(a,e)$ represent the title and body Direct relevance scores, respectively,

for article a with respect to event e. Detailed scoring rules, BM25 parameters, and frozen retrieval settings are provided in Supplementary Table S2.

Direct retrieval did not forcibly include articles that failed to satisfy the threshold merely to fill a fixed candidate count. For multi-query events, references selected for each routed issue were retained within the same event, whereas Routine-only events produced no Direct references. Direct outputs were retained as candidate regulatory context for downstream manager-oriented decision support.

3.5.4. Governance Context

The Governance context was implemented as a branch independent of the Direct candidates and was designed to provide higher-level safety-management duty, responsibility, and penalty-linked regulatory context rather than event-specific technical measures. Although the two branches use the same frozen corpus, their retrieval purposes and provenance are kept separate.

Governance references were generated through two paths: a ranking-based path and a guaranteed worker-duty path. The ranking-based path selects governance-oriented articles according to event-to-query relevance. The guaranteed worker-duty path independently preserves relevant duty articles and linked penalty provenance to reduce the possibility that worker-related duty context is omitted by the general ranking process. These two paths do not represent different levels of legal certainty; rather, they indicate different reference-inclusion provenance.

Penalty linkage was retained only as provenance associated with duty/responsibility articles and did not determine penalty applicability. After retrieval was frozen, Governance membership, ordering, and ranking-based, guaranteed worker-duty, and penalty-linked provenance were preserved without downstream reranking, refiltering, truncation, or substitution.

No Direct or Governance references were generated for Routine monitoring events. The final Direct and Governance evidence was subsequently passed to the manager-oriented decision-support process described in Section 3.6.

3.6. Role-Specific Safety Decision Support

A role-specific safety decision-support layer was constructed to transform frozen structured event evidence into information that could be used in practical safety management. Manager- and worker-oriented outputs share the same source event but apply different audience policies. Managers are provided with structured event evidence together with Direct/Governance regulatory context, whereas workers receive accident/PPE evidence and immediate safety actions, with detailed legal, governance, and penalty-related information excluded. This separation reflects the different operational purposes of the two audiences: managers require traceable event and regulatory context for review and safety-management decisions, whereas worker-facing information prioritizes immediate, actionable protective guidance and avoids legal or penalty detail that is not required for the immediate safety response. Accordingly, PPE grounding and regulation-grounded retrieval are not combined as a serial process; instead, independent evidence generated from the same structured event evidence is selectively used according to audience purpose.

3.6.1. Audience-Specific Safety Reports

Manager-oriented reports were designed to include an event summary, operational risk information, accident/PPE evidence, and the Direct/Governance references described in Section 3.5. GPT-5.6 Terra (gpt-5.6-terra) was used for natural-language realization,

and the prompt/template configuration was fixed before final report generation. Detailed generation settings are provided in Supplementary Table S4.

The language model was used only to transform frozen structured evidence and regulatory context into audience-specific natural language. It was not allowed to newly determine or modify event classification, operational risk, PPE state, or regulatory-reference selection. Prompts were constructed deterministically from a predefined template, structured event evidence, and frozen regulatory references. Even when a generated output did not satisfy the validation conditions, the underlying evidence and references were left unchanged, and only the natural-language realization was retried, for a maximum of five attempts.

In manager-oriented outputs, accident evidence, PPE Candidate/Confirmed evidence, and accident-related uncertainty coexisting within the same event were retained as supporting evidence independently of the representative report state. Direct/Governance reference membership and ordering were also preserved during generation, with no reranking, refiltering, truncation, or reference substitution applied.

Worker-oriented reports were generated according to an independent audience policy rather than as abbreviated versions of manager-oriented reports. For accident-related events, the worker-oriented reports focused on immediate protective actions, while PPE-related information followed the same reliability and semantic-grounding policy described in Section 3.4. Specific helmet or safety-vest information was included only when directly supported by the corresponding structured evidence. Candidate evidence or accident-related uncertainty alone was not used to generate a stronger non-compliance claim. When separate Confirmed evidence coexisted with uncertainty evidence, the respective reliability states of both Confirmed and uncertainty evidence were preserved.

Worker-facing outputs did not include specific law names, article-level information, legal responsibility, or penalty-related details from Direct/Governance references. Regulatory terminology was also not used to infer a visual PPE state. The same information boundary was applied to detailed worker reports and notifications to prevent indirect legal-information exposure through linked reports. Role-specific generation was therefore limited to expressing frozen evidence according to different audience purposes and was not used to modify upstream safety, PPE, or legal decisions.

3.6.2. Audience-Specific Notification Routing

Detailed role-specific reports and notifications were treated as separate output layers. A manager-oriented report and a worker-oriented report were generated for every event so that each source event retained a complete pair of audience-specific outputs. Notification emission was then determined separately and deterministically from the frozen event/query state. Manager notifications were generated for reportable accident or PPE-related review events, whereas Routine monitoring events did not trigger a manager notification. Worker notifications were generated only when a PPE-related query was activated; thus, the worker-notification channel in this study was scoped to PPE-related immediate feedback rather than to every event for which a worker-oriented report existed. When accident and PPE evidence coexisted, a worker notification was generated because the PPE query was active. Consequently, the absence of a notification does not indicate a missing report or failed report generation, but reflects the predefined audience-specific notification-routing rule.

Each notification was linked only to a detailed report corresponding to the same source event and the same audience. Manager notifications were linked to manager-oriented reports, and worker notifications were linked to worker-oriented reports; cross-audience linkage was not permitted. Notification routing was limited to the final delivery layer and did not modify event classification, PPE interpretation, risk assessment, or regulatory-reference selection. Source event–report correspondence, audience consistency, and notification–

report integrity were evaluated through the independent validation procedure described in Section 3.7.2.

3.7. End-to-End Validation and Audit Protocol

An end-to-end validation protocol separated from report generation was applied to evaluate the provenance and integrity of the structured event evidence, PPE grounding, regulatory references, and role-specific outputs generated in Sections 3.3–3.6. Validation was not used as a feedback stage for modifying the model, thresholds, retrieval configuration, or report policy on the basis of downstream results. In this study, auditability refers to the ability to reconstruct and verify the transformation of preserved frozen artifacts through explicit event identities, configuration records, reference identities, validation contracts, and provenance fingerprints. This definition concerns process and artifact traceability rather than intrinsic model interpretability. It does not imply that the internal feature representations, decision mechanisms, or reasoning processes of YOLO, Florence-2, or the language-model component are themselves transparent or interpretable. This internal artifact-level auditability is distinguished from unrestricted public reproducibility, which additionally depends on external access to the source data, model artifacts, and derived analytical files. The validation protocol therefore evaluates repeatability under the frozen configuration, evidence preservation, and audience/information integrity, while the availability constraints affecting independent public reproduction are reported separately.

3.7.1. Regulatory Reference Preservation and Controlled Retrieval Regression

To verify that Direct and Governance references were transmitted unchanged from retrieval to manager-oriented outputs, reference identity and ordering were compared. For the Governance branch, preservation of ranking-based, guaranteed worker-duty, and penalty-linked provenance was evaluated together with whether any downstream reranking, refiltering, truncation, or reference substitution occurred.

To separate the repeatability of the frozen Legal-RAG process from the effects of upstream evidence correction, a controlled regression analysis was performed by inputting Historical and final Full-clean structured event evidence into the same regulatory corpus, runtime scope, routing rules, and retrieval configuration. Historical outputs were not treated as a legal gold standard; rather, they served as regression references for comparing the effects of evidence transitions within the same frozen retrieval process.

Events were divided into three transition groups. In the Strict invariant group, downstream-relevant evidence remained unchanged, and exact Direct/Governance retrieval was therefore required. In the Query-invariant evidence-change group, some structured evidence could change while the query scope was expected to remain unchanged; retrieval outcomes were then examined under the unchanged query scope. In the Query-changed group, a change in query scope caused by corrected evidence was defined as expected behavior, together with the corresponding retrieval change. This design was used to determine whether observed retrieval differences reflected changes in structured evidence and routing scope rather than corpus modification or post-hoc retuning.

3.7.2. Audience and End-to-End Integrity Verification

For role-specific outputs, audience/event consistency, PPE semantic integrity, and worker-facing legal-information isolation were evaluated. Notifications and linked reports were checked to verify that they preserved the same source event and the correct audience. Cross-audience linkage, a missing target report, or an event mismatch was defined as an integrity failure.

PPE semantic integrity was evaluated according to whether downstream language remained within the component specificity and certainty supported by the structured PPE evidence. Specific helmet or safety-vest claims were permitted only when corresponding supporting source evidence was available. The reliability states of Candidate, Confirmed, and accident-related uncertainty evidence were also checked for preservation throughout downstream processing. A semantic-integrity failure was defined as a case in which Candidate evidence or accident-related uncertainty alone was escalated into a stronger PPE claim, or regulatory terminology was used as evidence for a visual PPE fact. For PPE-primary events, consistency in component identity and certainty between the notification and its linked worker-oriented report were also evaluated.

Worker-facing legal-information isolation was evaluated by checking whether specific law names, article-level information, legal responsibility, or penalty-related details were exposed in either notifications or linked detailed reports. Indirect exposure of manager-oriented regulatory information through a linked report was likewise defined as a failure.

Final integrity was evaluated using a frozen end-to-end verifier that operated separately from the report-generation process. Under a predefined validation contract, the verifier evaluated source event–report correspondence, Direct/Governance reference preservation, manager/worker audience consistency, notification–report linkage, PPE semantic grounding, and worker-facing legal-information isolation on the fixed benchmark. Validation findings were not used as post-processing to automatically modify the outputs, nor were they used to readjust the perception model, PPE policy, Legal-RAG configuration, or report content.

Major analytical artifacts and execution configurations were fingerprinted with SHA-256 hashes and a manifest (Supplementary Table S4). Common event_id and video_id fields linked structured event evidence, Legal-RAG outputs, role-specific reports, notifications, and validation records, enabling reconstruction of the source-to-output lineage without recomputing the pipeline. Publication tables and figures were generated in a read-only stage, and representative images required exact event/frame matching. Section 4.6 provides a literal event-level audit trace, while the Supplementary Materials document the frozen PPE, retrieval, perception, report/validation configurations, provenance controls, and public/restricted resource map. These materials support independent inspection of the transformation logic but do not remove access restrictions on the derived benchmark and internal analytical artifacts.

OpenAI GPT Image 2 was used only as a visual reference during schematic design; all final schematics were manually recreated by the authors in Microsoft PowerPoint, and no AI-generated images were directly used. The tool was not used to generate, alter, or interpret source-event images, research data, quantitative results, or analytical outputs.

3.8. Design of the Controlled Downstream Component Ablation

To isolate the contribution of the major downstream components, a controlled post-freeze ablation analysis was performed using only the stored outputs of the final Full-clean configuration. The frozen perception outputs, structured event records, PPE observations, regulatory corpus and retrieval outputs, role-specific reports, and validation artifacts were reused without modifying the authoritative analytical results. No model retraining, new vision inference, threshold retuning, corpus modification, Legal-RAG reranking, or LLM/API generation was performed for the ablation analysis. Counterfactual conditions were applied only to the relevant downstream representation or policy layer and were not fed back into the primary pipeline.

The structured-event representation was evaluated by comparing the proposed multi-evidence deterministic routing with a single-state counterfactual in which only the query

associated with the representative report-oriented event state was retained. The comparison quantified the number of downstream query activations and secondary evidence routes preserved by the multi-evidence representation. PPE temporal aggregation was evaluated separately on the 1552 ground-truth normal-scene frames used for the primary PPE evaluation. The controlled baseline used only the current-frame PPE observation, defining a PPE-violation indication when either helmet or safety-vest evidence was absent for the primary worker. The proposed condition used the frozen worker-track temporal Raw-state policy with the original 10-observation window and frozen thresholds. Accuracy, Precision, Recall, Specificity, F1-score, and Balanced Accuracy were calculated for both conditions, and paired differences were evaluated with 10,000 video-cluster bootstrap resamples using the same video-level clustering procedure described in Section 3.2.3. No Candidate or Confirmed threshold was changed in this analysis.

Because multiple PPE performance metrics were compared in this post-freeze ablation and no multiplicity adjustment was applied, the corresponding bootstrap intervals were interpreted as exploratory estimates of the paired differences.

The accident-aware uncertainty policy was evaluated by comparing the pre-gate Raw temporal states with the post-gate states generated by the frozen policy. Because this policy was designed to preserve uncertainty rather than to improve classification accuracy, the analysis quantified the number of component states explicitly relabeled as unknown_during_accident, the number of accident-gated frames containing such uncertainty, and whether Candidate or Confirmed flags changed as a consequence of the semantic state conversion. For Governance retrieval, a ranking-only counterfactual was constructed by disabling the guaranteed worker-duty inclusion path while retaining the frozen ranking-derived Governance references. This condition was compared with the proposed combined ranking and guaranteed worker-duty policy using Governance-reference counts and worker-duty event coverage as provenance-preservation measures.

The role-specific information boundary was evaluated using a field-level counterfactual in which manager-oriented regulatory context was assumed to be exposed to the corresponding worker-facing output, without regenerating natural-language reports. This condition was compared with the actual worker-specific information boundary by counting worker outputs potentially exposed to detailed legal context. Finally, the independent validation contract was supplemented with a controlled integrity stress test performed only on non-authoritative copies of the final artifacts. The injected faults comprised Direct-reference identity substitution, Governance-reference identity substitution, worker-facing legal-information leakage, and cross-audience notification linkage. Detection of these deliberately introduced contract violations was evaluated separately from the historical frozen verifier results and from generation-time retry behavior. This stress test was intended to assess deterministic integrity checks and was not interpreted as an estimate of real-world failure prevalence, legal validity, human comprehension, or operational effectiveness.

3.9. Post-Freeze PPE Threshold Sensitivity Analysis

To examine whether the observed Candidate/Confirmed behavior depended strongly on the exact frozen temporal parameters, a read-only post-freeze sensitivity analysis was performed using the stored worker-track PPE observations without retraining the perception models, rerunning vision inference, regenerating reports, or modifying the authoritative pipeline. The frozen primary configuration used a temporal window of 10 observations, a worn-state minimum of 2 positive observations, a Candidate missing-observation threshold of 5, and a Confirmed missing-observation threshold of 8 together with zero positive observations and the frozen upright criterion.

Sensitivity was evaluated using one-factor-at-a-time perturbations around this frozen configuration. The temporal window was varied across 8, 10, and 12 observations; the Candidate missing threshold across 4, 5, and 6 observations; and the Confirmed missing threshold across 7, 8, and 9 observations. A secondary check varied the worn-state minimum across 1, 2, and 3 observations. Window values below 8 were not evaluated because they would make the frozen Confirmed missing-observation criterion structurally unattainable rather than representing a compatible local perturbation. PPE confidence/association thresholds, the model-derived accident gate, and the upright criterion were held fixed. The primary sensitivity metrics were calculated on the same 1552 ground-truth normal-scene frames used for the main PPE reliability evaluation, and event-level evidence activation was additionally summarized across the 100-video human subset. The sensitivity analysis was descriptive and was not used to retune or select a new primary operating point.

4. Results and Discussion

This section reports the frozen primary results for safety-event recognition, PPE reliability, regulation-grounded retrieval, role-specific outputs, and end-to-end integrity, followed by the revision-only controlled ablation and post-freeze PPE sensitivity analyses. The primary results use the configuration fixed before final evaluation, whereas the additional analyses reuse stored frozen artifacts without model retraining, new vision inference, threshold retuning, regulatory-corpus modification, or LLM/API regeneration.

4.1. Safety-Event Recognition and Provenance-Correction Effects

On the fixed internal benchmark, the Full-clean configuration achieved a frame-level Accuracy of 98.23% and Macro-F1 of 97.47%, together with a video-level Accuracy and Macro-F1 of 98.67%, correctly classifying 148 of 150 videos. The Historical Original configuration achieved frame-level Accuracy/Macro-F1 of 97.87%/96.98%, while the intermediate configuration achieved 97.83%/96.93%; both reference configurations achieved a video-level Accuracy of 98.00% (Figure 3). The Historical Original configuration was treated as a historical reference preceding provenance correction rather than as an independent performance baseline.

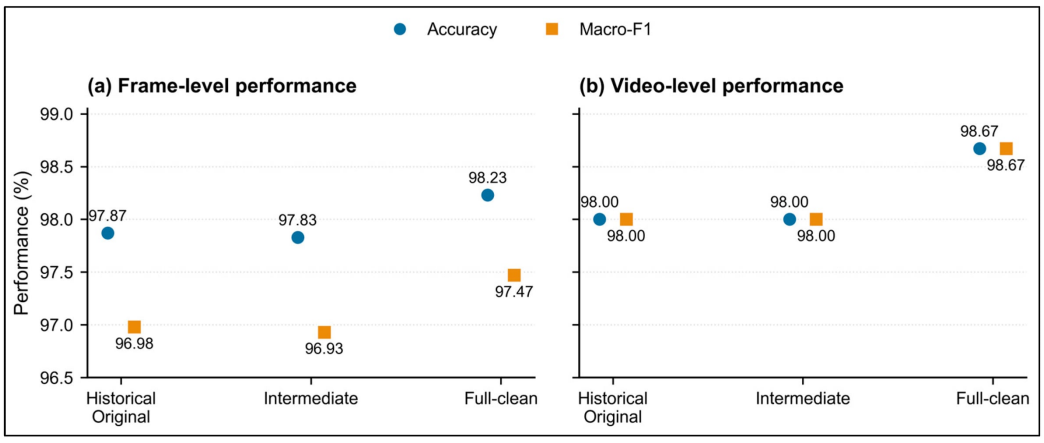


Figure 3. Frame-level and video-level safety-event recognition performance across the Historical Original legacy pair, YOLO-clean–legacy-VLM intermediate configuration, and Full-clean configuration on the fixed internal evaluation benchmark.

Compared with the intermediate configuration, the Full-clean configuration showed observed frame-level differences of +0.40 percentage points in Accuracy and +0.54 percentage points in Macro-F1. All 12 frames misclassified by the intermediate configuration were corrected without introducing any new frame-level errors. Importantly, all 12 corrected

frames had a frame-level ground-truth label of Normal and had been predicted as Fall Accident by the intermediate configuration. Their source-video scenarios comprised six Normal RGB cases, five Fall-down RGB cases, and one Fall-off RGB case; therefore, the source scenario should not be interpreted as the ground-truth label of the individual evaluation frame. All 12 cases had activated the broad-fall candidate and loose pose gate, whereas the clean Florence-2 semantic output indicated a normal event state and the final cascade rejected the fall-accident hypothesis. This pattern is consistent with semantic verification correcting broad-fall false positives triggered by geometric/postural evidence, although no more specific visual cause was assigned without dedicated image-level reannotation. The two-sided exact frame-level McNemar test yielded $p = 0.000488$; however, because frames within the same video are not independent, this result was treated only as descriptive supporting evidence. In the video-cluster bootstrap analysis, the 95% confidence intervals for the paired frame-level Accuracy and Macro-F1 differences were +0.13 to +0.77 and +0.17 to +1.07 percentage points, respectively (Figure 4). Because these comparisons were part of an unadjusted set of multiple performance analyses, they are interpreted as exploratory evidence of positive frame-level differences rather than as confirmatory evidence of model superiority.

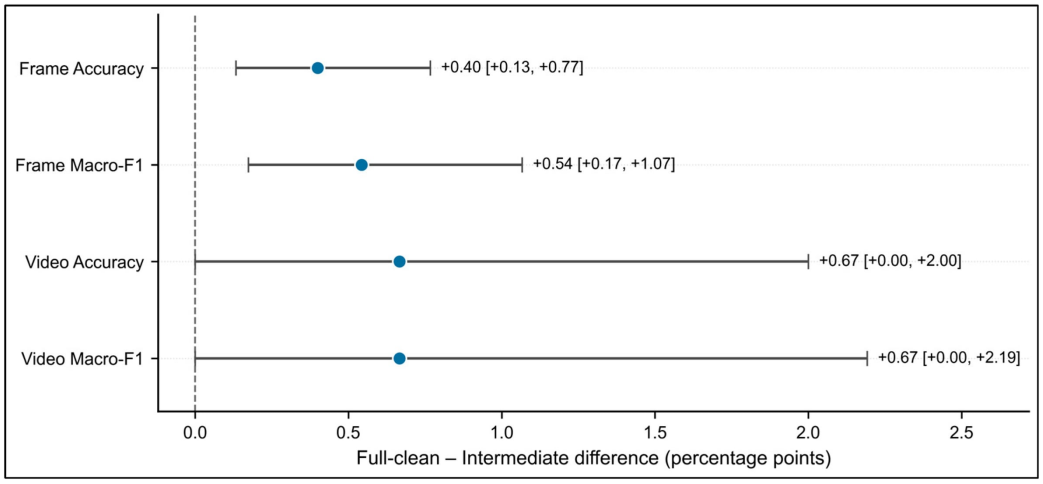


Figure 4. Video-cluster bootstrap 95% confidence intervals for paired performance differences between the Full-clean and YOLO-clean–legacy-VLM intermediate configurations at the frame and video levels.

At the video level, the observed number of correctly classified videos increased from 147/150 in the intermediate configuration to 148/150 in the Full-clean configuration because one previously misclassified Normal video was corrected and no new video-level errors were introduced. This numerical difference should not, however, be interpreted as evidence of a statistically significant video-level improvement. The paired exact McNemar test yielded $p = 1.0$, and the corresponding video-level bootstrap confidence interval included zero. The video-level result is therefore reported descriptively only and does not establish improvement, equivalence, or non-inferiority between the configurations. The two remaining Full-clean errors consisted of one Fall-off-to-Normal false negative and one Normal-to-Fall Accident false positive. These residual errors occurred on opposite sides of the fall-related decision boundary and therefore do not indicate a single directional bias. The false-negative case represents insufficient retained evidence for a ground-truth Fall-off event, whereas the false-positive case represents residual fall evidence sufficient to retain a Fall Accident decision in a ground-truth Normal video. Accordingly, the remaining video-level failures are interpreted as boundary cases of the fall-evidence aggregation process rather than as evidence of a common failure mechanism.

The frozen Full-clean perception run required 2654.66 s to process the 3000 evaluation frames, corresponding to an observed wall-clock throughput of approximately 1.13 processed frames/s under the evaluation configuration. Florence-2 semantic processing was activated for 2335 of the 3000 frames (77.8%). Among these triggered frames, the recorded mean task-level latencies were 122.97 ms for event semantic verification, 122.50 ms for PPE semantic verification, and 308.90 ms for safety-analysis generation. These task-level measurements do not include all detector, pose-estimation, data-loading, synchronization, and output-writing overhead; therefore, the overall 1.13 frames/s wall throughput provides a more conservative end-to-end characterization of the evaluated perception pipeline. This result does not demonstrate continuous-video real-time operation.

4.2. PPE Reliability and Operational Evidence Performance

The overall human PPE evaluation set comprised 2000 frames, of which 1552 ground-truth normal-scene frames formed the primary frame-level PPE reliability stratum defined in Section 3.4.1. The remaining 448 ground-truth accident-scene frames were retained separately for accident-aware uncertainty and condition-specific analyses rather than being excluded from evaluation.

For the combined PPE-violation decision derived from the frozen temporal Raw state, TP = 1296, TN = 238, FP = 16, and FN = 2. The corresponding performance was 98.84% Accuracy, 98.78% Precision, 99.85% Recall, 93.70% Specificity, 99.31% F1-score, and 96.77% Balanced Accuracy. The video-cluster bootstrap 95% confidence intervals were 97.66–99.67% for Accuracy, 97.32–99.72% for Precision, 99.50–100.00% for Recall, and 98.54–99.81% for F1-score.

At the component level, the helmet and safety-vest F1-scores on ground-truth normal-scene frames were 95.70% and 97.94%, respectively. The lowest condition-specific PPE performance was observed for helmet detection under the Fall-off condition, with a Precision of 65.65% and an F1-score of 77.83%. A focused audit of the 90 false-positive helmet observations in the Fall-off condition further characterized this boundary case. The current-frame helmet observation was absent in all 90 cases, and 87/90 had no matched helmet detection associated with the primary worker. In addition, the model-derived accident gate was active in 85/90 cases, and 67/90 did not satisfy the PPE-specific upright condition. Before accident-aware handling, the 90 observations were evenly divided between the not_detected_candidate and missing_candidate Raw states (45 each); after gating, 85 were explicitly represented as unknown_during_accident. These findings indicate that the principal failure pattern involved loss of reliable worker-linked helmet evidence during Fall-off-related observations rather than strong positive evidence that the helmet was not being worn. The audit therefore supports retaining an accident-specific uncertainty state, although it does not establish a specific visual cause such as occlusion or head rotation without additional image-level annotation.

For the 1372 frames in which the model-derived accident gate was inactive, the combined PPE F1-score was 99.79%. By contrast, among the 180 frames in which the gate was active, the F1-score decreased to 95.27%, with a Precision of 92.25% and a Specificity of 76.60%. These results indicate that PPE observation reliability can decrease when model-derived fall evidence is present even within the ground-truth normal-scene stratum, supporting the policy of preserving accident-related uncertainty separately from temporal Candidate/Confirmed evidence.

For the operational evidence tiers, Candidate achieved a Precision of 99.64%, Recall of 64.10%, and F1-score of 78.01%, whereas Confirmed achieved a Precision of 100.00%, Recall of 42.06%, and F1-score of 59.22% (Figure 5). Candidate and Confirmed are not separately trained classifiers; rather, they are operational evidence tiers generated by applying progres-

sively stricter reliability criteria to the same worker-track observations. The high Precision of Confirmed therefore reflects the intentionally conservative requirement for persistent non-detection together with the frozen posture and presence criteria, but its lower Recall has an important safety implication. Of the 1298 ground-truth PPE-violation-positive frames in the primary evaluation stratum, 546 satisfied the Confirmed criterion, whereas 752 did not; thus, Confirmed evidence used alone would fail to represent a substantial proportion of actual PPE violations at this high-confidence level. Confirmed should therefore be interpreted as a high-confidence escalation tier rather than as a stand-alone PPE detection or alert channel. Broader Candidate evidence remains available for review-oriented routing, while the underlying temporal and accident-related uncertainty states remain preserved in the structured record; consequently, absence of Confirmed evidence is not interpreted as evidence of PPE compliance.

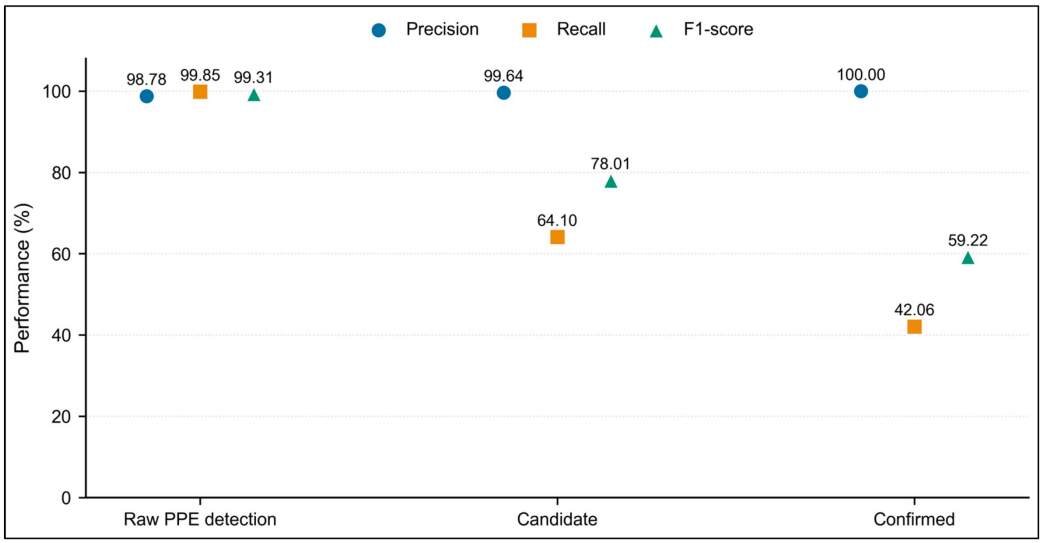


Figure 5. Precision, recall, and F1-score for temporal Raw-state PPE-violation detection and the Candidate and Confirmed operational evidence tiers on ground-truth normal-scene frames.

Before evaluating parameter perturbations, the post-freeze replay reproduced the frozen primary PPE outputs exactly, with zero mismatches for the Candidate, Confirmed, accident-uncertainty, and Raw temporal-violation flags. The temporal-window sensitivity was small over the compatible local range: windows of 8, 10, and 12 observations produced Candidate F1-scores of 77.98%, 78.01%, and 78.01%, respectively, while Confirmed performance remained unchanged at 100.00% Precision, 42.06% Recall, and 59.22% F1. Changing the Candidate missing-observation criterion produced the expected persistence–coverage trade-off. Thresholds of 4, 5, and 6 yielded Candidate Recall values of 71.57%, 64.10%, and 56.86%, respectively, while Precision remained between 99.57% and 99.73%. Similarly, varying the Confirmed missing threshold from 7 to 8 to 9 preserved 100.00% Precision but reduced Recall from 48.15% to 42.06% and 36.52%, respectively.

Within the evaluated local range, a Candidate missing-observation threshold of 4 produced a higher Candidate F1-score (83.28%) than the frozen threshold of 5 (78.01%), primarily because Recall increased from 64.10% to 71.57% with only a small change in Precision (99.64% to 99.57%). This post-freeze observation was not used to reselect the operating point because threshold selection on the evaluation benchmark would constitute post-hoc tuning; calibration of a more inclusive Candidate tier should instead be performed on independent data.

Secondary worn-state perturbations produced only small changes in frame-level Candidate F1 (77.78–78.01%), although the number of videos containing Candidate or Confirmed

evidence varied from 88 to 92 compared with 90 under the frozen baseline. Overall, the local sensitivity analysis preserved the qualitative distinction between Candidate as the broader review-oriented tier and Confirmed as the more conservative high-precision tier. The frozen primary thresholds were retained unchanged and were not reselected on the basis of this analysis. Detailed results are reported in Supplementary Table S7.

At the event level, PPE-related routing was activated for 90 of the 100 human-related videos, which comprised 84 ground-truth-positive and 16 ground-truth-negative videos. The routing results were TP = 84, TN = 10, FP = 6, and FN = 0, corresponding to an Accuracy of 94.00%, Precision of 93.33%, Recall of 100.00%, and F1-score of 96.55%. Neither the number of routing activations nor the video-level routing decisions changed between the Historical and Full-clean evidence. These results were interpreted as downstream routing performance for selecting events in which PPE-related regulatory candidates should be retrieved, rather than as a legal or compliance determination of PPE violations.

4.3. Regulation-Grounded Retrieval and Controlled Regression

Using the final Full-clean structured evidence, the report-oriented event types consisted of 50 Fall Accident events, 50 Fire Accident events, 43 PPE non-compliance events, 1 PPE non-detection candidate event, and 6 No reportable accident events. Deterministic routing activated 50 Fall-related, 50 Fire-related, 90 PPE-related, and 6 Routine monitoring queries. Because the query categories were not mutually exclusive, the 90 PPE-related events included multi-query events in which accident and PPE evidence coexisted.

Retrieval coverage and provenance results are summarized in Table 9. Direct and Governance references were generated for 144 of the 150 non-routine events, whereas no references were generated for the 6 Routine monitoring events. The Direct branch produced 1108 references corresponding to 23 unique articles, while the Governance branch produced 576 references corresponding to 6 unique articles. The Governance set comprised 288 ranked references and 288 guaranteed worker-duty references.

Table 9. Regulation-grounded retrieval coverage, reference provenance, preservation, and controlled-regression results.

Evaluation Item	Final Result
Direct retrieval event coverage	144/150 (96.0%)
Governance retrieval event coverage	144/150 (96.0%)
Direct regulatory references	1108 references; 23 unique articles
Governance references	576 references; 6 unique articles
Governance provenance	288 ranked references and 288 guaranteed worker-duty references
Guaranteed worker-duty event coverage	144/144
Penalty-linked Governance provenance	576/576
Strict invariant regression	117/117 exact retrieval
Query-invariant evidence-change regression	32/32 query scope preserved; 0 retrieval changes
Query-changed regression	1/1 expected retrieval change
Direct references preserved downstream	1108/1108
Governance references preserved downstream	576/576
Routine events with regulatory references	0

In the controlled regression analysis, all 117 Strict invariant events reproduced the exact retrieval results (117/117). Among the 32 evidence-change events for which the query scope remained unchanged, routing was preserved in all 32 cases and no retrieval changes occurred. In the single event for which the query scope changed, the expected retrieval change was observed. Specifically, the Historical evidence activated both Fall and PPE queries, whereas the Full-clean evidence activated only the PPE query; consequently, the number of Direct candidates decreased from 8 to 7, with only the fall-specific candidate removed. These results indicate that the frozen retrieval process reproduced the same retrieval outputs when the query scope remained unchanged and produced a corresponding retrieval change only when the upstream correction altered the routing scope.

All 1108 Direct and 576 Governance references were preserved downstream with their recorded provenance, and no reference loss attributable to report generation was observed. These results demonstrate repeatability and reference-transmission integrity, not legal applicability or retrieval completeness. Because no expert-derived relevance/completeness gold standard was available, semantically missing relevant provisions could not be quantified; Direct/Governance outputs therefore remain candidate regulatory context, and possible omissions are treated as an unresolved validation boundary.

4.4. Role-Specific Safety Outputs and End-to-End Integrity

Manager- and worker-oriented reports were generated for all 150 events. Manager-oriented natural-language generation satisfied the validation conditions on the first attempt for 135 of 150 events (90.0%). The remaining 15 events satisfied the validation conditions on Attempts 2–5, with 9, 2, 2, and 2 events succeeding at each respective attempt. No template fallback was required. During retries, the structured evidence and regulatory references remained unchanged, and only the natural-language realization was regenerated.

The report and notification counts reflect different output policies rather than missing outputs. Both manager- and worker-oriented reports were generated for all 150 events, providing a complete role-specific report pair for each source event. Manager notifications were expected for the 144 non-routine events because the 6 Routine monitoring events did not satisfy the manager-notification rule; all 144 expected notifications were generated. Worker notifications were expected for the 90 events in which the PPE-related query was activated, and all 90 were generated. Thus, the 150/150 report counts represent complete event-level report generation, whereas the 144 and 90 notification counts reflect the predefined audience-specific delivery rules. Across the resulting 234 notification–report links, both the source event and intended audience were consistent in all cases. No worker-facing legal-information leakage was observed, and PPE semantic consistency was satisfied in all 44 evaluated cases. The overall generation, routing, and integrity results are summarized in Table 10.

The independent frozen verifier evaluated source event–report correspondence, regulatory-reference preservation, audience consistency, notification linkage, PPE semantic grounding, and worker-facing legal-information isolation according to the predefined validation contract. All 82 verifier findings received a PASS outcome. Accordingly, the role-specific validation in this study demonstrates system-level information-boundary enforcement, source-event/audience consistency, and preservation of the predefined evidence constraints. It does not demonstrate that workers understood the worker-oriented information, that managers made better decisions, or that the reports improved usability or operational effectiveness, because no human-participant comprehension, usability, or decision-quality study was conducted. The legal applicability of the retrieved regulatory references was likewise outside the scope of this validation.

The API-based role-specific report-generation stage had a substantially different latency profile from the frozen perception stage. Across the final 150 event-level report

packages, the preserved generation records contained 779,701 input tokens and 427,252 output tokens, corresponding to 1206,953 tokens in total, or approximately 8046 tokens per event on average. During the uninterrupted resumed execution segment, 94 newly generated event packages consumed 745,299 tokens and required 51.58 min of wall time, corresponding to an observed throughput of approximately 32.9 s per newly generated event package. This value includes API communication and generation/retry overhead and should not be interpreted as isolated LLM inference latency. Together with the generation retries reported in Table 10, these measurements indicate that detailed LLM-based report generation is unsuitable for use as a blocking step in a time-critical safety-alert path in the evaluated configuration.

Table 10. Generation, audience-routing, semantic-grounding, and end-to-end integrity results for the final role-specific safety outputs.

Evaluation Item	Final Result
Manager-oriented reports generated	150/150
Worker-oriented reports generated	150/150
Manager-report first-attempt generation success	135/150 (90.0%)
Manager-report retry distribution (Attempts 2–5)	9/2/2/2
Template fallback	0
Manager notifications (non-routine events)	144/144 expected
Worker notifications (PPE-query events)	90/90 expected
Notification–report links	234/234
Audience/event mismatch	0
Worker-facing legal-information leakage	0 observed
PPE semantic consistency	44/44
Independent frozen verifier	PASS
Final validation findings	PASS = 82; WARN = 0; BLOCK = 0; FAIL = 0

4.5. Controlled Downstream Component Ablation

The six controlled contrasts have different evidential roles and were evaluated against the specific failure mode that each component was designed to address rather than as equivalent predictive-accuracy ablations. PPE temporal aggregation was evaluated empirically against ground truth; the structured-event contrast quantified the data-dependent amount of downstream evidence and query routing preserved by the multi-evidence representation; and accident-aware handling quantified uncertainty-state conversion while retaining the underlying persistence evidence. The Governance and audience-boundary contrasts evaluated whether their predefined policy guarantees were preserved on the actual event set, whereas the independent-verifier stress test evaluated detectability of deliberately injected integrity violations rather than their natural prevalence. The main results are summarized in Table 11.

The multi-evidence event representation preserved 196 deterministic query activations across the 150 events, compared with 150 activations under the single-state counterfactual. The difference comprised 46 secondary PPE routes, all associated with multi-query events in which fall-related and PPE-related evidence coexisted. Thus, 46 of the 90 PPE-related routes (51.1%) would have been discarded if each event had been reduced to only its representative report-oriented state. This result reflects downstream evidence preservation rather than an improvement in event-classification accuracy.

Table 11. Controlled downstream component ablation using frozen Full-clean artifacts. Counterfactual conditions were applied only to the corresponding downstream representation or policy layer and were not used to retrain, retune, or modify the primary pipeline.

Component	Controlled Contrast	Evaluation Target	Main Result
Structured event representation	Single representative state/query → multi-evidence deterministic routing	Downstream query/evidence preservation	Query activations: 150 → 196; 46 secondary PPE routes preserved
PPE temporal aggregation	Current-frame PPE observation → frozen worker-track temporal Raw state	Transient false-positive suppression	FP: 37 → 16 (−56.8%); Specificity: 85.43% → 93.70%; F1: 98.40% → 99.31%
Accident-aware uncertainty	Raw temporal state retained → accident-specific uncertainty state	Explicit uncertainty preservation	842 component states relabeled as unknown_during_accident; Candidate/Confirmed flag changes: 0/0
Governance worker-duty path	Ranking-only Governance → ranking + guaranteed worker-duty path	Worker-duty provenance coverage	Governance references: 288 → 576; worker-duty event coverage: 0/144 → 144/144
Role-specific information boundary	Shared manager-style legal context counterfactual → worker-specific boundary	Worker legal-information isolation	Potential worker legal exposure: 144 → 0 events
Independent validation contract	Deliberately corrupted artifact copies → deterministic contract checks	Detectability of predefined integrity violations	528/528 injected contract violations flagged in the post-freeze stress test

For PPE, the current-frame-only baseline yielded TP = 1293, TN = 217, FP = 37, and FN = 5 on the 1552 ground-truth normal-scene frames, compared with TP = 1296, TN = 238, FP = 16, and FN = 2 under the frozen worker-track temporal policy. The temporal policy therefore reduced false positives by 21 frames (56.8%) while also reducing false negatives from 5 to 2; Specificity increased from 85.43% to 93.70%, and F1-score increased from 98.40% to 99.31%. The paired difference in Specificity was +8.27 percentage points, with a 95% video-cluster bootstrap confidence interval of +2.39 to +16.58 percentage points; the F1-score difference was +0.91 percentage points, with a corresponding 95% interval of +0.26 to +1.91 percentage points. Because these paired metric comparisons were not adjusted for multiplicity, the reported bootstrap intervals are interpreted as exploratory rather than confirmatory. These results isolate the effect of temporal aggregation and do not involve Candidate/Confirmed threshold retuning.

Among the 2000 human-subset frames, the accident-aware policy was active in 606 frames, and 547 of these frames (90.3%) contained at least one component explicitly represented as unknown_during_accident. At the component level, 842 states were relabeled to accident-context uncertainty. Candidate and Confirmed flags were unchanged by this semantic state conversion, indicating that the policy altered the representation of certainty without discarding the separately retained persistence evidence. In the Governance ablation, disabling the guaranteed worker-duty path reduced Governance references from 576 to 288 and worker-duty event coverage from 144/144 to 0/144. Restoring the proposed guaranteed path recovered worker-duty provenance for all 144 reportable events. This comparison evaluates provenance preservation and does not establish legal relevance or applicability.

For the audience-boundary counterfactual, 144 events contained manager-oriented regulatory context that would have been exposed to worker-facing outputs if a shared information policy had been used. Under the proposed worker-specific boundary, no such legal-information exposure was observed. This comparison evaluates information isolation

and does not establish worker comprehension or usability. Separately, the historical frozen verifier retained its original PASS results under the predefined validation contract. In a post-freeze stress test performed only on non-authoritative copies, deterministic contract checks flagged all 528 deliberately introduced violations, comprising 144 Direct-reference identity substitutions, 144 Governance-reference identity substitutions, 150 worker-facing legal-information leakage cases, and 90 cross-audience notification links. These synthetic violations were used only to test contract enforcement and should not be interpreted as the prevalence of such failures in the original system. They were also evaluated separately from the 15 events that required generation-time retries.

4.6. Representative Event-Level Safety Evidence

Figure 6 presents qualitative Fall, Fire, PPE, and Routine examples using only exact source-event and target-frame matches; it complements, rather than re-estimates, the quantitative evaluations in Sections 4.1–4.5.

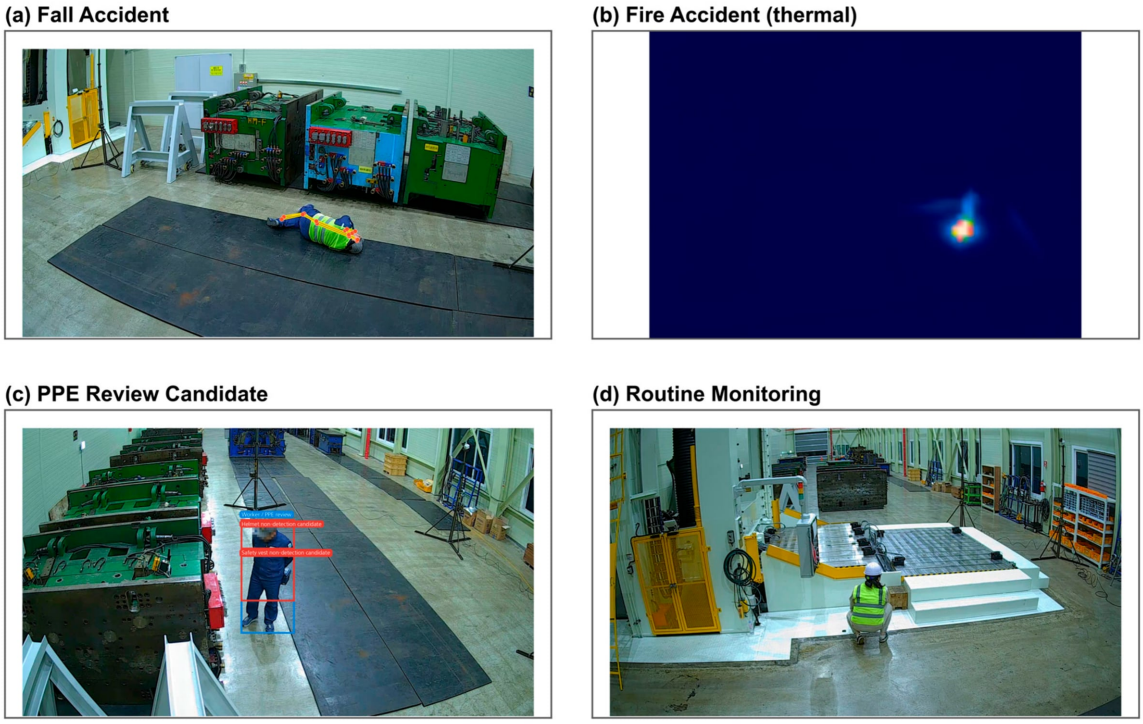


Figure 6. Representative event-level visual evidence for Fall, Fire, PPE, and Routine cases using exact event and target-frame matching. PPE examples represent review candidates and do not independently establish confirmed PPE non-compliance.

The representative frames do not independently determine video-level decisions, and PPE examples are not interpreted as confirmed non-compliance from a single frame; Candidate/Confirmed and accident-uncertainty evidence follow the temporal policy in Section 3.4. The Routine example illustrates a negative event for which no accident/PPE review condition, regulatory references, or alerts were activated.

Figure 7 summarizes the source-to-output evidence lineage; because a schematic does not itself constitute an audit record, a literal frozen-event trace is provided in Table 12.

Table 12 presents the read-only trace for event human-accident_fall-down_rgb_0016_cctv4 (video_000050), selected because fall/PPE evidence, Direct/Governance references, both audience outputs, and verifier records were available. The values are preserved artifact records rather than recomputed results.

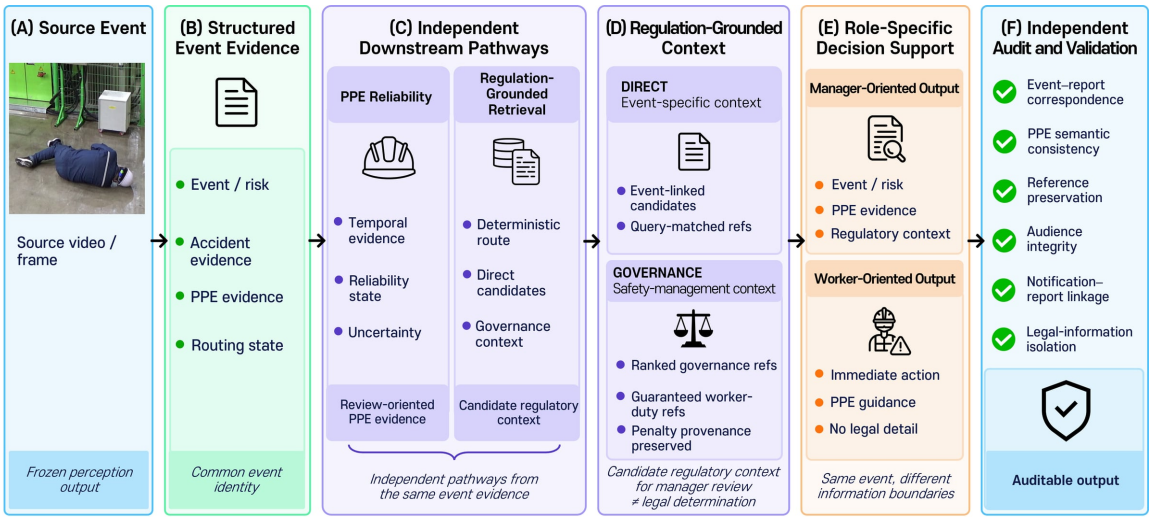


Figure 7. Schematic overview of the end-to-end evidence lineage from frozen multimodal safety evidence to role-specific decision support and independent audit and validation.

Table 12. Representative literal audit-trail fragment reconstructed from frozen artifacts for event human-accident_fall-down_rgb_0016_cctv4.

Audit Stage	Frozen Artifact-Level Trace
Source identity	event_id = human-accident_fall-down_rgb_0016_cctv4; video_id = video_000050; source video = human-accident_fall-down_rgb_0016_cctv4.mp4; exact representative source frame = frame 19 (human-accident_fall-down_rgb_0016_cctv4_f19.jpg)
Frozen perception identity	Clean YOLO11s SHA-256 = 5e2fb4a2...8f8e63a7; Clean Florence-2-base model-tree SHA-256 = 2d91b561...6afe96d6; frozen 05g pipeline SHA-256 = 7875bf09...b7e55722
Structured event evidence	Predicted video class = Fall Accident; report-oriented event type = Fall Accident; risk level = High; routed queries = fall_accident_prevention_and_response; ppe_compliance
PPE reliability evidence	Candidate evidence count = 16; Confirmed evidence count = 4; unknown_during_accident count = 10; PPE evidence remained preserved together with the Fall event state
Direct regulatory retrieval	8 Direct references; leading references included LAW_000033 (Article 32, protective-equipment provision), LAW_000045 (Article 44, safety-belt attachment facilities), and LAW_000044 (Article 43, opening/fall-protection measures)
Governance retrieval	4 Governance references: LAW_001242 Article 4 and LAW_001243 Article 5 through the ranked path; LAW_000716 Article 40 and LAW_000812 Article 133 through the guaranteed worker-duty/penalty provenance path
Role-specific reports	Manager and worker reports were both linked to the same event_id; manager output retained Direct/Governance context, whereas the worker output excluded detailed legal/penalty information
Notification routing	Manager notification generated and linked to the manager report; worker PPE notification generated and linked to the worker report; both retained event_id = human-accident_fall-down_rgb_0016_cctv4
Generation record	Manager report generation status = success_attempt_2; generation backend = template_skeleton_plus_gpt_audience_split_polish; no template fallback
Independent event-level verification	Direct-reference identity exact = True; Governance-reference identity exact = True; Governance count = 4/4 exact; manager and worker audience fields = exact; report paths = exact; linked report targets existed; worker-facing legal-information hits = 0
Frozen verifier identity	Verifier SHA-256 = fd4d7f41...b3852c52; final verifier execution returned PASS with no fail markers

Note. Abbreviated SHA-256 fingerprints are shown for readability; full fingerprints are retained in the frozen provenance manifest.

4.7. Discussion, Practical Implications, and Limitations

The primary significance of this study lies in extending manufacturing-workplace video analysis beyond accident classification or PPE detection by introducing a downstream

framework that transforms frozen perception evidence into structured and auditable safety information. Centered on structured event evidence, the framework performs PPE reliability/semantic grounding and regulation-grounded retrieval as independent downstream processes and applies different audience-specific information boundaries to the same source event. In this way, evidence reliability, provenance, and uncertainty preservation are treated as core design principles rather than relying solely on predictive performance.

The controlled ablations indicate that the proposed components address distinct downstream failure modes: multi-evidence routing preserves concurrent evidence, temporal PPE aggregation reduces transient false positives, accident-aware handling preserves uncertainty, the Governance path retains worker-duty provenance, role-specific policies enforce legal-information boundaries, and the independent verifier detects predefined integrity violations. These effects concern evidence preservation, reliability control, provenance retention, and information-boundary enforcement rather than a single common gain in predictive accuracy, and they do not establish external generalizability, legal applicability, human comprehension, or field-level operational effectiveness.

The perception and PPE results provide exploratory support for this reliability-centered design. The Full-clean configuration showed positive frame-level differences in the unadjusted exploratory analyses, whereas the observed video-level numerical difference was not supported as a statistically significant improvement. VLM verification should therefore be interpreted as an auxiliary semantic source for supporting ambiguous evidence rather than as an independently superior classifier. For PPE, progressively stricter reliability criteria produced the intended trade-off between evidence coverage and confidence. Confirmed achieved 100.00% Precision but only 42.06% Recall under the frozen primary configuration; therefore, using Confirmed as the sole intervention or alert criterion could leave a substantial proportion of actual PPE violations below the high-confidence tier. In the proposed framework, Confirmed is instead treated as conservative escalation evidence, while Candidate provides a broader review-oriented channel and the underlying temporal and accident-related uncertainty evidence remains preserved. The post-freeze sensitivity analysis showed the same pattern across nearby Confirmed persistence criteria: Precision remained 100.00% as the missing-observation threshold increased from 7 to 9, whereas Recall decreased from 48.15% to 36.52%. These findings indicate that the frozen threshold represents a predefined operating point in a precision–recall trade-off rather than a universally optimal setting. Practical deployment should therefore calibrate the Candidate/Confirmed operating points on independent data according to the relative safety costs of false alarms and missed PPE violations, while avoiding the interpretation of absent Confirmed evidence as evidence of compliance.

The Legal-RAG results support repeatable evidence linkage and provenance preservation under the frozen configuration, not automated legal determination. Direct and Governance outputs remain candidate context for manager review because regulatory relevance, completeness, and legal applicability were not independently validated.

For role-specific decision support, managers were provided with event evidence and regulatory context, whereas workers received accident/PPE evidence and immediate safety-oriented information, with detailed legal and penalty-related information kept separate. The present evaluation supports the technical enforcement of these information boundaries: source-event and audience consistency were preserved, no worker-facing legal-information leakage was observed, and the audience-specific notification rules were reproduced without missing expected notifications. However, these system-level integrity results should not be interpreted as evidence that the selected information was optimal for either audience. The study did not evaluate worker comprehension, manager decision quality, report usability, alert usefulness, trust, or behavioral response through human-participant testing. Accordingly, the role-specific outputs should presently be interpreted as a techni-

cally validated information-boundary design for human-review-oriented decision support; practical-deployment discussion in this study concerns technical pipeline feasibility and information-boundary enforcement only, not demonstrated worker comprehension, manager decision quality, usability, or operational effectiveness, which require prospective human-participant evaluation.

A cross-stage failure analysis further showed that the remaining limitations were not concentrated in a single module. The two residual video-level recognition errors occurred on opposite sides of the fall-decision boundary, while the principal PPE boundary case involved loss of reliable helmet evidence under Fall-off-related observations. In the regulatory pathway, no downstream reference loss was observed, but the absence of an expert-derived relevance gold standard means that semantically missing relevant provisions cannot yet be quantified. Likewise, no natural role-information leakage was observed; therefore, the corresponding boundary was assessed using controlled synthetic violations rather than natural failure prevalence. These findings emphasize that high aggregate performance or exact integrity checks do not eliminate module-specific failure modes and that the meaning of "failure" differs across perception, evidence reliability, retrieval completeness, generation availability, and information-boundary enforcement.

The computational profile also places an important boundary on deployment claims. The evaluated Full-clean configuration processed the fixed evaluation frames at approximately 1.13 frames/s, and the API-based detailed report-generation stage operated on a substantially slower, event-level timescale. The current implementation should therefore be interpreted as a proof-of-concept downstream safety-intelligence workflow rather than as a validated real-time manufacturing surveillance system. A practical deployment would need to decouple the time-critical perception/alert path from slower semantic, regulatory, and detailed-report processing so that an immediate safety response is not blocked by VLM or external LLM generation. Event- or keyframe-triggered semantic processing, model compression or quantization, optimized inference runtimes, caching of deterministic regulatory context, and asynchronous report generation are possible engineering strategies, but their performance was not evaluated in the present study. Production deployment would consequently require a separate target-hardware evaluation of end-to-end latency, tail latency, concurrent camera load, network/API dependence, and failure behavior before real-time or near-real-time operation could be claimed.

This study has six principal limitations. First, the evaluation was based on a fixed internal benchmark comprising 150 videos and 3000 frames with a history of prior use. The reported results should therefore be interpreted as internal feasibility evidence under the frozen evaluation setting and not as a pristine blind test, independent external validation, or evidence of generalizable field performance. Second, the safety-event taxonomy was limited to Normal, Fall Accident, and Fire Accident, while PPE evaluation was limited to helmets and safety vests. Validation across additional hazard types and PPE components is therefore required. Third, the regulatory evaluation was conducted using a frozen Korean industrial-safety corpus and predefined retrieval rules; no expert-derived legal gold standard or cross-jurisdictional applicability assessment was performed. Fourth, the integrity validation assessed evidence preservation and information boundaries rather than worker comprehension, manager decision quality, response time, alert fatigue, trust, or actual usability. Fifth, public reproducibility remains constrained by the access conditions of the source dataset and by restrictions on redistribution of the derived internal benchmark and analytical artifacts. Although the source-data access route, regulatory sources, frozen configurations, provenance fingerprints, validation rules, and a representative artifact-level audit trail are provided to support independent inspection, the complete frozen experiment cannot be reproduced from identical derived artifacts using the public materials alone.

Sixth, the computational measurements were obtained from the frozen offline evaluation and report-generation runs rather than from a controlled deployment benchmark on target manufacturing hardware. The observed throughput therefore does not characterize camera-to-alert latency, multi-camera concurrency, tail latency, network variability, or optimized edge/server execution, and the present study does not establish real-time or near-real-time deployment capability.

Future work should therefore evaluate generalizability and operational utility using independent external datasets, an expanded hazard/PPE taxonomy, safety/legal expert assessment, and human-centered prospective field validation.

5. Conclusions

This study proposed an auditable downstream safety-intelligence framework that converts frozen multimodal workplace-safety evidence into structured event records, reliability-aware PPE evidence, regulation-grounded candidate context, and role-specific outputs, with separate validation of PPE semantics, reference preservation, audience/event integrity, and worker-facing legal-information isolation.

On the fixed internal benchmark, the Full-clean configuration achieved a frame-level Accuracy of 98.23%, a frame-level Macro-F1 of 97.47%, and a video-level Accuracy and Macro-F1 of 98.67%. Relative to the intermediate configuration, all 12 of its frame-level errors were corrected without introducing new frame-level errors; however, the associated inferential comparisons were unadjusted for multiplicity and are therefore interpreted as exploratory. At the video level, the observed change from 147/150 to 148/150 correctly classified videos did not demonstrate a statistically significant improvement (paired exact McNemar $p = 1.0$; the bootstrap confidence interval included zero). On the ground-truth normal-scene PPE stratum, PPE-violation detection based on the Raw temporal state achieved an F1-score of 99.31%, while Candidate and Confirmed evidence exhibited higher Precision and lower Recall under progressively stricter reliability criteria. In the regulation-grounded controlled regression, all 117 Strict invariant cases reproduced the same retrieval results, all 32 Query-invariant cases preserved query scope without unexpected retrieval changes, and the single Query-changed event produced the expected retrieval change. All 1108 Direct references and 576 Governance references were preserved throughout downstream processing. No audience/event mismatch was observed across the 234 notification–report links, and all 44 PPE semantic checks and all 82 findings from the independent verifier satisfied the predefined validation contract.

Within the fixed internal benchmark used in this study, the results provide feasibility evidence that multimodal manufacturing-safety outputs can be transformed into evidence-grounded, reliability-aware, regulation-grounded, role-specific, and auditable decision-support information while preserving uncertainty and provenance across downstream processing. The findings should therefore be interpreted as an internal feasibility validation of evidence preservation, retrieval repeatability under the frozen configuration, and information integrity within the proposed evidence lifecycle and human-review support framework; they do not establish independent external generalization, legal applicability, human understanding or usability, or field-level operational effectiveness. The evaluated configuration also does not establish real-time deployment feasibility: the Full-clean perception run achieved approximately 1.13 processed frames/s, while detailed API-based report generation operated on an event-level timescale and should not be treated as a time-critical alert mechanism. Independent external validation, an expanded hazard and PPE taxonomy, safety/legal expert assessment, and human-centered prospective field evaluation are required before broader operational generalization can be claimed.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162110893/s1: Table S1: Frozen primary-worker selection and worker–PPE association rules used in PPE evaluation and evidence generation; Table S2: Key frozen configurations of the regulation-grounded retrieval pipeline; Table S3: Frozen upstream perception and accident-evidence configuration used to generate multimodal safety evidence; Table S4: Frozen configuration and provenance controls for role-specific report generation and end-to-end validation; Table S5: Reproducibility resources, access conditions, and their roles in independent inspection of the frozen pipeline; Table S6: Failure-case and boundary-condition analysis of the frozen final pipeline; Table S7: Post-freeze one-factor-at-a-time sensitivity analysis of the frozen PPE temporal evidence policy.

Author Contributions: Conceptualization, D.Y. and S.C.; methodology, D.Y., W.C., J.C. and S.C.; software, D.Y.; validation, D.Y., W.C. and J.C.; formal analysis, D.Y.; investigation, D.Y.; data curation, D.Y.; writing—original draft preparation, D.Y.; writing—review and editing, D.Y., W.C., J.C. and S.C.; visualization, D.Y.; supervision, S.C.; project administration, S.C.; funding acquisition, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Regional Innovation System & Education (RISE) program through the Chungbuk Regional Innovation System & Education Center, funded by the Ministry of Education (MOE) and Chungcheongbuk-do, Republic of Korea (2026-RISE-11-014-03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source industrial-safety video dataset used in this study, "Data for Smart Manufacturing Facility Safety Monitoring," is available through AI-Hub subject to the provider's terms of use and access conditions. The Korean industrial-safety regulatory source texts used to construct the frozen corpus are publicly accessible through the Korean National Law Information Center, and the statutory versions used in this study are identified in Table 6. To facilitate independent inspection of the analytical workflow, the accompanying Supplementary Materials provide the frozen worker–PPE rules, Legal-RAG configuration, upstream perception configuration, role-specific generation and validation controls, and a reproducibility-resource map; the main manuscript additionally provides an event-level audit-trail example and frozen provenance fingerprints. The derived 150-video internal benchmark, structured event records, intermediate processing artifacts, and downstream analytical outputs are not publicly redistributed because of source-data usage restrictions and project data-management constraints. Accordingly, the publicly available materials support inspection and partial reconstruction of the frozen analytical logic but do not permit exact independent reproduction from the identical derived artifacts.

Acknowledgments: This research used data from the Open AI Dataset Project (AI-Hub, Republic of Korea), specifically the "Data for Smart Manufacturing Facility Safety Monitoring" dataset. Dataset information is available through AI-Hub (www.aihub.or.kr (accessed on 28 September 2026)). During the preparation of this manuscript, the authors used OpenAI ChatGPT(GPT-5.6 Sol, OpenAI, San Francisco, CA, USA) for language polishing and OpenAI GPT Image 2 (OpenAI, San Francisco, CA, USA) only as a visual reference during schematic design. All final schematic figures were manually recreated and edited by the authors in Microsoft PowerPoint, and no AI-generated image was directly used in the manuscript. The authors reviewed and edited all AI-assisted outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fang, W.; Ding, L.; Love, P.E.D.; Luo, H.; Li, H.; Peña-Mora, F.; Zhong, B.; Zhou, C. Computer vision applications in construction safety assurance. *Autom. Constr.* **2020**, *110*, 103013. [CrossRef]
2. Guo, B.H.W.; Zou, Y.; Fang, Y.; Goh, Y.M.; Zou, P.X.W. Computer vision technologies for safety science and management in construction: A critical review and future research directions. *Saf. Sci.* **2021**, *135*, 105130. [CrossRef]

3. Badhan, S.J.; Samsami, R. Artificial Intelligence (AI) in Construction Safety: A Systematic Literature Review. *Buildings* **2025**, *15*, 4084. [CrossRef]
4. Damilos, S.; Saliakas, S.; Karasavvas, D.; Koumoulos, E.P. An Overview of Tools and Challenges for Safety Evaluation and Exposure Assessment in Industry 4.0. *Appl. Sci.* **2024**, *14*, 4207. [CrossRef]
5. Sivanraj, S.; Uduwage, D.N.L.S.; Tripathi, M. Real-time safety detection on construction sites using a vision-language and NLP-based model. *Adv. Eng. Inform.* **2026**, *69*, 103889. [CrossRef]
6. Abidi, S.M.H.; Raza, S.M.; Shin, S.Y. SafeVision: Vision–language reasoning for context-aware safety monitoring. *Neurocomputing* **2026**, *669*, 132479. [CrossRef]
7. Zaidi, S.F.A.; Yang, J.; Abbas, M.S.; Hussain, R.; Lee, D.; Park, C. Vision-Based Construction Safety Monitoring Utilizing Temporal Analysis to Reduce False Alarms. *Buildings* **2024**, *14*, 1878. [CrossRef]
8. Cheng, J.C.P.; Wong, P.K.Y.; Luo, H.; Wang, M.; Leung, P.H. Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification. *Autom. Constr.* **2022**, *139*, 104312. [CrossRef]
9. Vukicevic, A.M.; Petrovic, M.; Milosevic, P.; Peulic, A.; Jovanovic, K.; Novakovic, A. A systematic review of computer vision-based personal protective equipment compliance in industry practice: Advancements, challenges and future directions. *Artif. Intell. Rev.* **2024**, *57*, 319. [CrossRef]
10. Chen, Q.; Long, D.; Wang, S.; Chen, Q.; Yuan, B. Real-Time Detection of Personal Protective Equipment Violations for Construction Workers Using Semisupervised Learning and Video Clips. *J. Constr. Eng. Manag.* **2025**, *151*, 04024213. [CrossRef]
11. Meyers, C.; Löfstedt, T.; Elmroth, E. Safety-critical computer vision: An empirical survey of adversarial evasion attacks and defenses on computer vision systems. *Artif. Intell. Rev.* **2023**, *56*, 217–251. [CrossRef]
12. Price, L. Finding fault in organisations—Reconceptualising the role of senior managers in corporate manslaughter. *Leg. Stud.* **2015**, *35*, 385–407. [CrossRef]
13. Park, H.; Woo, M.; Lee, W. Early Effects of the Serious Accidents Punishment Act on Occupational Health Outcomes in Korea: A Nationwide Difference-in-Differences Study (2017–2023). *Am. J. Ind. Med.* **2026**, *69*, 657–665. [CrossRef] [PubMed]
14. Lee, S. The Concept of "Responsible Managing Officer, etc." of the Serious Accidents Punishment Act. *Contemp. Rev. Crim. Law* **2022**, *77*, 80–115. [CrossRef]
15. Karakurt, E.; Akbulut, A. Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review. *Appl. Sci.* **2026**, *16*, 368. [CrossRef]
16. Zhang, Y.; Wang, X.; Hu, M.; Zhang, Z. Bridging dual knowledge graphs for multi-hop question answering in construction safety. *Autom. Constr.* **2026**, *183*, 106794. [CrossRef]
17. Beach, T.H.; Hippolyte, J.-L.; Rezgui, Y. Towards the adoption of automated regulatory compliance checking in the built environment. *Autom. Constr.* **2020**, *118*, 103285. [CrossRef]
18. Qian, H.; Lin, J.; Tang, J.; Zhou, H.; Peng, Y.; Liu, L.; Liu, X. Lightweight Adapted Retrieval-Augmented Inference Network for Legal Issues. *Electronics* **2026**, *15*, 2996. [CrossRef]
19. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
20. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López de Prado, M.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
21. Fanfani, M.; Ipsaro Palesi, L.A.; Nesi, P. Human-Centered AI for Decision Support Systems: A Systematic Review of Application Domains, Architecture Designs, Current Trends and Future Directions. *Big Data Cogn. Comput.* **2026**, *10*, 186. [CrossRef]
22. Alu, F.F.; Oluwadare, S. An auditable and source-verified framework for clinical AI decision support: Integrating retrieval-augmented generation with data provenance. *Front. Artif. Intell.* **2026**, *9*, 1737532. [CrossRef] [PubMed]
23. Zhang, M.; Shi, R.; Yang, Z. A critical review of vision-based occupational health and safety monitoring of construction site workers. *Saf. Sci.* **2020**, *126*, 104658. [CrossRef]
24. Fang, W.; Love, P.E.D.; Luo, H.; Ding, L. Computer vision for behaviour-based safety in construction: A review and future directions. *Adv. Eng. Inform.* **2020**, *43*, 100980. [CrossRef]
25. Paneru, S.; Jeelani, I. Computer vision applications in construction: Current state, opportunities & challenges. *Autom. Constr.* **2021**, *132*, 103940. [CrossRef]
26. Ahn, J.; Park, J.Y.; Lee, S.S.; Lee, K.H.; Do, H.; Ko, J.G. SafeFac: Video-based smart safety monitoring for preventing industrial work accidents. *Expert Syst. Appl.* **2023**, *215*, 119397. [CrossRef]
27. Lee, T.; Woo, K.; Kim, P.; Jung, H. Design and Implementation of Industrial Accident Detection Model Based on YOLOv4. *Appl. Sci.* **2023**, *13*, 10163. [CrossRef]
28. Liu, L.; Guo, Z.; Liu, Z.; Zhang, Y.; Cai, R.; Hu, X.; Yang, R.; Wang, G. Multi-Task Intelligent Monitoring of Construction Safety Based on Computer Vision. *Buildings* **2024**, *14*, 2429. [CrossRef]

29. Chern, W.C.; Hyeon, J.; Nguyen, T.V.; Asari, V.K.; Kim, H. Context-aware safety assessment system for far-field monitoring. *Autom. Constr.* **2023**, *149*, 104779. [CrossRef]
30. Jeon, Y.; Tran, D.Q.; Kulinan, A.S.; Kim, T.; Park, M.; Park, S. Vision-based motion prediction for construction workers safety in real-time multi-camera system. *Adv. Eng. Inform.* **2024**, *62*, 102898. [CrossRef]
31. Chan, C.F.; Wong, P.K.Y.; Guo, X.; Cheng, J.C.P.; Chan, J.P.C.; Leung, P.H.; Tao, X. Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring. *Autom. Constr.* **2025**, *177*, 106305. [CrossRef]
32. Kim, T.; Kim, S.; Chern, W.C.; Park, S.; Kim, D.; Kim, H. Optimizing large vision-language models for context-aware construction safety assessment. *Autom. Constr.* **2025**, *180*, 106510. [CrossRef]
33. Chen, S.; Demachi, K. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Autom. Constr.* **2021**, *125*, 103619. [CrossRef]
34. Li, J.; Zhao, X.; Zhou, G.; Zhang, M. Standardized use inspection of workers' personal protective equipment based on deep learning. *Saf. Sci.* **2022**, *150*, 105689. [CrossRef]
35. Park, M.-W.; Elsafty, N.; Zhu, Z. Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers. *J. Constr. Eng. Manag.* **2015**, *141*, 04015024. [CrossRef]
36. Ma, L.; Chen, C.; Zhou, W.; Wang, B. A Real-Time High-Accuracy Framework for Small-Scale PPE Detection in Complex Construction Environments. *J. Constr. Eng. Manag.* **2026**, *152*, 04026107. [CrossRef]
37. Zhang, H.; Li, J.; Li, Z.; Kong, L.; Zhao, X. Real-time safety helmet detection using multi-scale feature-gated fusion. *Autom. Constr.* **2026**, *186*, 106891. [CrossRef]
38. Xiong, R.; Tang, P. Pose guided anchoring for detecting proper use of personal protective equipment. *Autom. Constr.* **2021**, *130*, 103828. [CrossRef]
39. López, L.; Suárez-Ramírez, J.; Alemán-Flores, M.; Monzón, N. Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation. *Autom. Constr.* **2025**, *176*, 106231. [CrossRef]
40. Li, F.; Chen, Y.; Hu, M.; Luo, M.; Wang, G. Helmet-Wearing Tracking Detection Based on StrongSORT. *Sensors* **2023**, *23*, 1682. [CrossRef] [PubMed]
41. Zhang, Y.; Huang, S.; Qin, J.; Li, X.; Zhang, Z.; Fan, Q.; Tan, Q. Detection of helmet use among construction workers via helmet-head region matching and state tracking. *Autom. Constr.* **2025**, *171*, 105987. [CrossRef]
42. Zhou, X.; Jia, X.; Bai, J.; Lv, X.; Lv, X.; Zhang, G. Optimizing Helmet Use Detection in Construction Sites via Fuzzy Logic-Based State Tracking. *Sensors* **2025**, *25*, 6487. [CrossRef] [PubMed]
43. Gil, D.; Lee, G. Zero-shot monitoring of construction workers' personal protective equipment based on image captioning. *Autom. Constr.* **2024**, *164*, 105470. [CrossRef]
44. Beach, T.H.; Rezgui, Y.; Li, H.; Kasim, T. A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Syst. Appl.* **2015**, *42*, 5219–5231. [CrossRef]
45. Lee, J.H.; Ji, S.Y.; Ostwald, M.J. Automated compliance checking across the building lifecycle: Systematic and semantic review integrating PRISMA and deep search. *Autom. Constr.* **2026**, *185*, 106859. [CrossRef]
46. Wang, X.; El-Gohary, N. Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements. *Autom. Constr.* **2023**, *147*, 104696. [CrossRef]
47. Wu, Z.; Ma, G. NLP-based approach for automated safety requirements information retrieval from project documents. *Expert Syst. Appl.* **2024**, *239*, 122401. [CrossRef]
48. Hussain, R.; Lee, D.; Abbas, M.S.; Zaidi, S.F.A.; Pedro, A.; Park, C. Vision-language model-based intelligent assistant for onsite construction safety inspection. *Autom. Constr.* **2026**, *182*, 106728. [CrossRef]
49. Ren, W.; Shen, L.; Zhu, J.; Li, S.; Hong, Y.; Wang, J.; Cao, D. Automated data synthesis and retrieval-augmented generation for legal large language models. *Knowl.-Based Syst.* **2026**, *347*, 116267. [CrossRef]
50. Seo, J.; Min, J. Optimization of a Hybrid RAG System for Korean Legal QA. *J. Korea Soc. Comput. Inf.* **2025**, *30*, 53–63. [CrossRef]
51. Naiseh, M.; Al-Thani, D.; Jiang, N.; Ali, R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int. J. Hum.-Comput. Stud.* **2023**, *169*, 102941. [CrossRef]
52. Romeo, G.; Conti, D. Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI Soc.* **2026**, *41*, 259–278. [CrossRef]
53. Kim, M.; Kim, S.; Kim, J.; Song, T.-J.; Kim, Y. Do stakeholder needs differ?—Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. *Int. J. Hum.-Comput. Stud.* **2024**, *181*, 103160. [CrossRef]
54. Hazarika, S.; Scott, M.; Fugate, B.; Sabherwal, R. When Safety Technologies Backfire: How Monitoring Affects Drivers' Safety Behavior. *J. Bus. Logist.* **2026**, *47*, e70057. [CrossRef]
55. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
56. Mora-Cantallops, M.; Sánchez-Alonso, S.; García-Barriocanal, E.; Sicilia, M.-A. Traceability for Trustworthy AI: A Review of Models and Tools. *Big Data Cogn. Comput.* **2021**, *5*, 20. [CrossRef]

57. Kalokyri, V.; Tachos, N.S.; Kalantzopoulos, C.N.; Sfakianakis, S.; Kondylakis, H.; Zaridis, D.I.; Colantonio, S.; Regge, D.; Papanikolaou, N.; ProCancer-I Consortium. AI Model Passport: Data and system traceability framework for transparent AI in health. *Comput. Struct. Biotechnol. J.* **2025**, *28*, 386–404. [CrossRef] [PubMed]
58. Guo, K.X.; Wong, P.K.Y.; Cheng, J.C.P.; Chan, C.F.; Leung, P.H.; Tao, X. Enhancing visual-LLM for construction site safety compliance via prompt engineering and Bi-stage retrieval-augmented generation. *Autom. Constr.* **2025**, *179*, 106490. [CrossRef]
59. Katsampiris-Salgado, K.; Dimitropoulos, N.; Michalos, G. Generative AI-based risk assessment and hazard identification towards safety compliant workplace design. *CIRP Ann.* **2026**, *75*, 173–177. [CrossRef]
60. Orooje, M.S.; Re Cecconi, F.; Lee, G.; Mei, Q.; Adil, M. Toward trustworthy construction safety hazard detection with visual-language models. *J. Inf. Technol. Constr.* **2026**, *31*, 1067–1095. [CrossRef]
61. AI-Hub. Data for Smart Manufacturing Facility Safety Monitoring. Available online: https://www.aihub.or.kr/aihub-data/data/view.do?dataSetSn=71679 (accessed on 28 September 2026).
62. Ultralytics. Ultralytics YOLO11. Available online: https://docs.ultralytics.com/models/yolo11/ (accessed on 28 September 2026).
63. Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; Yuan, L. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 4818–4829.
64. Korea Ministry of Government Legislation. Korean Law Information Center (National Law Information Center). Available online: https://www.law.go.kr/main.html (accessed on 28 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.