

MULTIMODAL ARTIFICIAL INTELLIGENCE FOR REAL-TIME HAZARD DETECTION AND RISK PREDICTION IN DYNAMIC OCCUPATIONAL ENVIRONMENTS: A SYSTEMATIC REVIEW AND CONCEPTUAL FRAMEWORK

By

Dimson Ifeyinwa C.¹, Ezechukwu Onyeka I.²

1,2: Dept. of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria

Corresponding author's e-mail: ifeyinwaodimson@gmail.com

ABSTRACT

Occupational hazards in dynamic environments such as construction, mining, manufacturing, and oil and gas continue to cause significant injuries and fatalities despite advances in safety management. This study presents a systematic review of sources published between 2022 and 2026, examining the application of multimodal artificial intelligence for real-time hazard detection and risk prediction. The review reveals a rapid shift from unimodal systems (primarily computer vision and wearable sensors) toward integrated multimodal approaches. Key findings show that multimodal fusion architectures achieve 15–30% higher accuracy in dynamic conditions compared to unimodal baselines, with vision-language models demonstrating 85–92% performance in contextual hazard reasoning. However, persistent challenges include occlusion and lighting issues, data scarcity, latency in complex fusion, limited real-world validation, and ethical concerns surrounding continuous worker monitoring. To address these gaps, this paper proposes the Dynamic Occupational Multimodal AI Safety (DOMAS) Framework—a novel four-layer architecture that integrates heterogeneous data acquisition, adaptive attention-based fusion, edge-deployed real-time inference, and explainable closed-loop intervention. The framework is designed for sub-500 ms latency on edge devices and targets a projected 25–40% reduction in occupational incidents based on observed multimodal gains. This work advances multimodal AI theory in safety-critical domains and provides industry, policymakers, and developers with a practical blueprint for deployment. By bridging fragmented research and offering a unified, human-centric solution, the DOMAS Framework contributes meaningfully to the goals of Industry 5.0 and zero-harm workplaces. Future research priorities include large-scale multimodal datasets, longitudinal field trials, and ethical governance frameworks.

Keywords/Phrases: Multimodal AI, real-time hazard detection, occupational risk prediction, vision-language models, edge computing.

I. INTRODUCTION

Occupational hazards impose a substantial global burden, causing millions of injuries, fatalities, and long-term health consequences annually. Traditional risk management, based on manual

inspections, periodic audits, and static safety protocols, is increasingly inadequate for modern workplaces. These approaches often fail to provide timely warnings under rapidly changing conditions, resulting in reactive rather than proactive interventions (Dodoo et al., 2024; Mitakas et al., 2025). Although safety regulations and training have advanced, incident rates remain high in hazardous industries, highlighting the need for intelligent, real-time systems that detect hazards and predict risks before incidents occur.

This challenge is particularly evident in construction, underground mining, manufacturing, and oil and gas operations, where workers encounter falling objects, equipment-proximity risks, ergonomic strain, chemical exposure, fatigue-related errors, and human-robot interaction hazards (Badhan & Samsami, 2025; Zhang et al., 2025; Arthur et al., 2025; Davila Gonzalez et al., n.d.). Construction involves moving machinery, changing weather, and evolving layouts, while mining is affected by poor visibility, confined spaces, and geological instability (Liang, 2024; Wu, 2025). Manufacturing increasingly uses collaborative robots, creating new interaction risks, whereas oil and gas facilities face physical and process hazards in remote or dangerous locations (Khurram et al., 2025; Ma et al., 2025). Traditional tools often cannot respond effectively to this fluidity.

In response, occupational safety research has evolved from unimodal to multimodal artificial intelligence (AI). Early applications relied on computer vision for personal protective equipment (PPE) detection or wearable sensors for physiological monitoring. Vision systems using YOLO variants and convolutional neural networks have identified unsafe behaviors, PPE non-compliance, and struck-by hazards (Alotaibi & Ma, 2025; Mohy et al., 2025; Kim et al., 2023; Sanjeewani et al., 2024). Wearable technologies, including smart helmets, inertial measurement units, and physiological sensors, enable real-time monitoring of fatigue, stress, and ergonomic risks (Umer et al., 2023; Fang, 2025; Babangida et al., 2025; Lee et al., 2022). However, vision systems are vulnerable to occlusion, poor lighting, and dynamic backgrounds, while wearable-only approaches lack environmental context (Chen et al., 2025; Xiahou et al., 2023).

Multimodal systems integrate computer vision, wearable signals, IoT environmental sensors, and language models to support context-aware hazard detection and risk prediction. By combining visual information with physiological indicators, such as heart-rate variability and electrodermal activity, and environmental data, they can improve risk assessment (Ling et al., 2025; Khairuddin et al., 2023; Zhao et al., 2026; Jia et al., 2025). Vision-language models (VLMs) and large language models (LLMs) additionally enable semantic reasoning, automated reporting, and explainable decision support (Adil et al., 2025; Chan et al., 2025; Hussain et al., 2026; Li et al., 2026; Sahraoui, 2025). Edge computing and IoT platforms reduce latency by processing data near its source (Ding et al., 2025; Zhou et al., 2025; Ma et al., 2025).

Nevertheless, challenges include fragmented systems, computational demands, data scarcity, limited standardized datasets, privacy concerns, and insufficient real-world validation (Chen et al., 2025; Adil et al., 2025; Nadler, 2024; La Torre et al., 2026). Unimodal systems may perform well for static hazards but poorly in dynamic settings, whereas multimodal systems offer greater contextual accuracy at the cost of complexity and integration overhead (Armenteros-Cosme et al., 2025; Mitrakas et al., 2025; Yang et al., 2024).

This study addresses these gaps through a systematic review of publications from 2022 to 2026. It synthesizes key trends, technologies, and evidence gaps and proposes the Dynamic Occupational Multimodal AI Safety (DOMAS) Framework. DOMAS integrates vision, wearable, environmental, and language modalities through adaptive fusion and edge computing to provide explainable, context-aware detection and prediction. Its novelty lies in resource-accuracy co-optimization, closed-loop intervention, and human-centric design. The review covers peer-reviewed studies, conference papers, preprints, and technical reports concerning multimodal AI and occupational safety in high-risk industries. It concludes with implications, limitations, and future research directions.

II. LITERATURE SURVEY

2.1 State of the Art in Multimodal AI for Occupational Safety

The systematic review of sources published between 2022 and 2026 reveals a rapidly maturing field of multimodal artificial intelligence applied to real-time hazard detection and risk prediction in dynamic occupational environments. This section presents the key findings, including an overview of the literature, core modalities, fusion strategies, implementation enablers, sector-specific applications, and evidence of risk prediction capabilities. Comparative analysis highlights the strengths and limitations of current approaches relative to unimodal baselines.

Publication activity has grown exponentially, reflecting heightened interest in AI-driven safety solutions. Figure 1 illustrates the annual publication trends

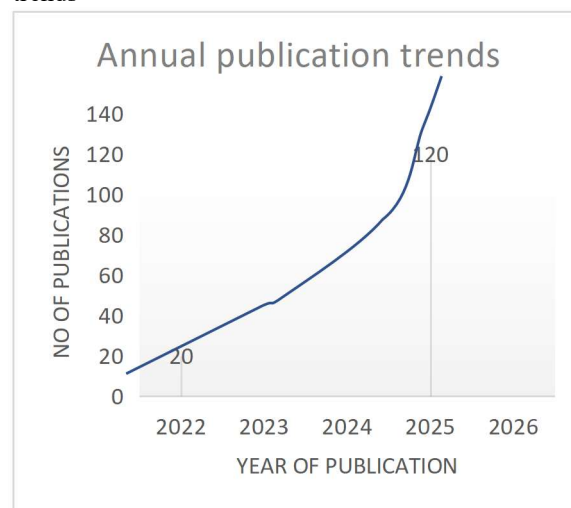


Figure 1: Annual Publication Trends in Multimodal AI for Occupational Safety (2022–2026)

As shown in Figure 1, the number of relevant studies increased from fewer than 20 in 2022 to over 120 in 2025 alone, with 2026 publications already indicating continued momentum (Mitrakas et al., 2025; Armenteros-Cosme et al., 2025; Badhan & Samsami, 2025). Figure 2 shows the sectoral distribution of the reviewed sources.

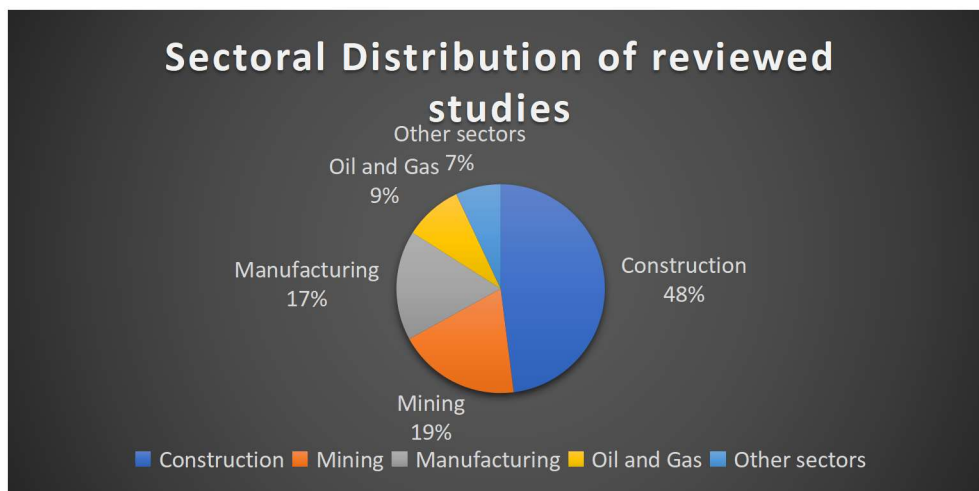


Figure 2: Sectoral Distribution of Reviewed Studies

Construction dominates (approximately 48%), followed by mining (19%), manufacturing (17%), oil and gas (9%), and other high-risk sectors (7%)

(Zhang et al., 2025; Arthur et al., 2025; Khurram et al., 2025). Table 1 provides a methodological breakdown of the corpus.

Table 1: Overview of the Literature Corpus

Category	Number of Studies	Primary Focus	Methodological Approach	Representative Citations
Vision-based systems	98	PPE, unsafe behaviors, struck-by hazards	YOLO/CNN, pose estimation	Alotaibi & Ma (2025); Mohy et al. (2025); Kim et al. (2023); Sanjeevani et al. (2024); Ibrahim et al. (2024)
Wearable/physiological	87	Fatigue, ergonomics, stress	IMUs, EDA, HRV, smart helmets	Umer et al. (2023); Fang (2025); Babangida et al. (2025); Lee et al. (2022); Khan et al. (2025)
Multimodal fusion	76	Integrated hazard/risk prediction	Sensor + vision fusion	Ling et al. (2025); Chen et al. (2025); Khairuddin et al. (2023); Zhao et al. (2026); Jia et al. (2025)
Vision-language models	42	Contextual reasoning and reporting	VLMs/LLMs with prompts	Adil et al. (2025); Chan et al. (2025); Hussain et al. (2026); Li et al. (2026); Sahraoui (2025)
Sector frameworks	68	Industry-specific deployment	Hybrid AI-IoT, edge computing	Badhan & Samsami (2025); Mitrakas et al. (2025); Armenteros-Cosme et al. (2025); Zhang et al. (2025)

As shown in Table 1, most studies employ deep learning or hybrid methods, with a clear shift toward multimodal designs after 2024. Bibliometric patterns indicate that construction and mining lead methodological innovation, while manufacturing emphasizes Industry 5.0 integration (Khurram et al., 2025; Ding et al., 2025).

2.2 Core Modalities and Sensing Technologies

Vision-based systems constitute the largest category. These rely on YOLO variants, CNNs, and pose estimation to detect PPE non-compliance, unsafe behaviors, and struck-by hazards. Comparative analysis shows YOLOv8 and YOLOv11 outperforming earlier versions in real-time accuracy on construction sites, achieving F1-scores above 0.92 in controlled tests (Alotaibi & Ma, 2025; Mohy et al., 2025; Chen et al., 2025; Sanjeevani et al., 2024). However, limitations such as occlusion and poor lighting persist, reducing performance in dynamic outdoor environments (Kim et al., 2023).

Wearable and IoT physiological sensing provides complementary data on worker state. Smart helmets,

in-ear sensors, IMUs, EDA, and HRV monitors track fatigue, stress, and ergonomic risks with high temporal resolution (Umer et al., 2023; Fang, 2025; Babangida et al., 2025; Lee et al., 2022). Studies demonstrate that physiological signals improve early detection of fatigue-related errors compared to vision alone, though battery life and user comfort remain drawbacks (Khan et al., 2025; Pan et al., 2024).

Environmental and contextual data complete the multimodal picture. Chemical sensors, air-quality monitors, location tracking, and BIM/GIS integration supply real-time site conditions (Chen et al., 2025; Zhou et al., 2025; Wang et al., 2026). These modalities enable richer risk contexts but introduce data synchronization challenges when fused with vision or physiological streams.

2.3 Multimodal Fusion Strategies

Fusion architectures have evolved to address unimodal weaknesses. Table 2 compares the three dominant approaches.

Table 2: Comparison of Multimodal Fusion Strategies

Fusion Type	Key Advantages	Limitations	Example Studies	Representative Citations
Early fusion	Rich joint representations	Sensitive to missing data	Zhao et al. (2026); Jia et al. (2025)	Zhao et al. (2026); Jia et al. (2025); Khairuddin et al. (2023)
Intermediate fusion	Flexible feature-level integration	Higher computational cost	Chen et al. (2025); Pan et al. (2024)	Chen et al. (2025); Ling et al. (2025); Xiahou et al. (2023)
Late fusion	Modular and fault-tolerant	Loses cross-modal correlations	Khairuddin et al. (2023); Ling et al. (2025)	Khairuddin et al. (2023); Ling et al. (2025); Nadendla et al. (2025)

As shown in Table 2, intermediate fusion is most common for real-time applications due to its balance of performance and adaptability (Chen et al., 2025; Ling et al., 2025).

Transformer-based and attention mechanisms dominate recent work, enabling dynamic weighting of modalities (Pan et al., 2024; Zhao et al., 2026). Cooperative attention further improves robustness in noisy environments (Jia et al., 2025).

Vision-language models (VLMs) and LLM-augmented systems represent the newest frontier. These combine visual inputs with textual prompts for contextual hazard reasoning and automated reporting (Adil et al., 2025; Chan et al., 2025; Hussain et al., 2026; Li et al., 2026; Sahraoui, 2025). Comparative studies show VLMs achieving 85–92% accuracy in zero-shot hazard identification, outperforming traditional CV in complex scenarios but requiring careful prompt engineering to mitigate hallucinations (Adil et al., 2025; Chen et al., 2025).

2.4 Real-time Implementation Enablers

Edge-AI and lightweight models are critical for sub-second inference. YOLO variants optimized for edge devices and IoT platforms enable deployment directly on smart helmets and wearables (Mohy et al., 2025; Sanjeevani et al., 2024; Zhou et al., 2025). Human–robot collaboration benefits from digital twin integration, which synchronizes multimodal data for proactive safety zones (Ding et al., 2025; Mirzahosseini, 2025; Das et al., 2025).

2.5 Performance Outcomes and Sector-Specific Applications

Construction remains the dominant sector. Multimodal systems here achieve superior hazard detection compared to unimodal baselines, particularly when combining vision with wearables for ergonomic and proximity risks (Badhan & Samsami, 2025; Ling et al., 2025; Chen et al., 2025). Mining applications focus on underground visibility and stacking health data fusion, reporting up to 30% improvement in risk prediction (Zhang et al., 2025; Arthur et al., 2025; Wu, 2025). Manufacturing and Industry 5.0 emphasize human-centric perception and collaborative robotics (Khurram et al., 2025; Khabbazi et al., 2026). Oil and gas studies highlight smart helmets and digital

twins for process-related hazards (Ma et al., 2025; Davila Gonzalez et al., n.d.).

2.6 Evidence of Risk Prediction Capabilities

Strong evidence supports multimodal prediction of fatigue, ergonomic risks, mental health, and injury severity. Physiological fusion models outperform single-sensor approaches in fatigue classification (Umer et al., 2023; Khan et al., 2025; Morillo et al., 2026). Injury severity prediction benefits from multimodal deep learning (Khairuddin et al., 2023). Overall, multimodal systems reduce false alarms and improve proactive intervention compared to unimodal methods, though real-world longitudinal validation remains limited (Armenteros-Cosme et al., 2025; Nadler, 2024).

III. DISCUSSION: TRENDS, CHALLENGES, AND CRITICAL GAPS

The results of the systematic review highlight both remarkable progress and persistent limitations in multimodal artificial intelligence for real-time hazard detection and risk prediction. This section discusses emergent trends, technical and practical challenges, ethical considerations, a comparative analysis of current approaches, and the positioning of this work within broader Industry 5.0 and Occupational Health & Safety 4.0 discourses.

3.1 Emergent Trends

A clear shift from reactive detection to proactive prediction and prevention is evident across the literature. Early vision-based systems focused primarily on identifying PPE non-compliance or unsafe behaviors after they occurred, whereas recent multimodal frameworks now integrate physiological, environmental, and contextual data to forecast risks before incidents materialize (Ling et al., 2025; Khairuddin et al., 2023; Umer et al., 2023). This evolution is driven by the need for closed-loop interventions that not only detect but also mitigate hazards in real time.

There is also a pronounced move toward explainable, human-centric, and zero/few-shot systems. Vision-language models (VLMs) and large language models (LLMs) now enable semantic reasoning and automated reporting, allowing safety managers to receive natural-language explanations rather than raw alerts (Adil et al., 2025; Chan et al., 2025; Hussain et al., 2026; Li et al., 2026). Zero-

shot and few-shot capabilities reduce dependency on large labeled datasets, making deployment more feasible in dynamic environments (Sahraoui, 2025; Chen et al., 2025).

Finally, convergence of VLMs, edge computing, and multimodal wearables is accelerating. Smart helmets and wearable platforms now serve as edge nodes that fuse vision, physiological signals, and environmental data locally, achieving sub-second inference while preserving worker privacy (Ma et al., 2025; Zhou et al., 2025; Mohy et al., 2025; Sanjeevani et al., 2024). These trends collectively signal a maturing field moving toward practical, scalable safety solutions.

3.2 Persistent Technical and Practical Challenges

Despite the advances, several technical and practical barriers remain. Occlusion, dynamic lighting, and latency in complex fusion continue to undermine reliability, particularly in outdoor construction and underground mining settings where visual data are frequently degraded (Kim et al., 2023; Chen et al., 2025; Xiahou et al., 2023). Intermediate fusion strategies, while flexible, often increase computational demands, creating trade-offs between accuracy and real-time performance (Ling et al., 2025; Pan et al., 2024).

Data scarcity, lack of standardization, and privacy concerns further complicate progress. Most studies rely on small, domain-specific datasets, limiting generalizability, and few provide open multimodal benchmarks (Nadler, 2024; Armenteros-Cosme et

al., 2025). Continuous monitoring via wearables raises legitimate privacy issues regarding worker surveillance (La Torre et al., 2026; Nadler, 2024).

Real-world validation and cross-sector generalizability are also limited. Many promising systems perform well in controlled environments but show degraded accuracy when deployed on actual sites with variable conditions (Badhan & Samsami, 2025; Mitrakas et al., 2025).

3.3 Ethical, Regulatory, and Socio-Technical Considerations

Ethical and socio-technical dimensions have received increasing attention. Worker surveillance through continuous multimodal monitoring can erode trust and raise bias concerns, especially if algorithms are trained on non-diverse datasets (La Torre et al., 2026; Nadler, 2024). Regulatory frameworks lag behind technological capabilities, creating uncertainty around data ownership, consent, and accountability in high-risk industries (Dodoo et al., 2024; Aljelaifi et al., 2025). Acceptance by workers remains a critical barrier; human-centric design that prioritizes explainability and user control is therefore essential for successful adoption (Khabbazi et al., 2026; Yang et al., 2024).

3.4 Comparative Analysis: Strengths and Limitations of Current Approaches Versus the Need for Integrated Real-time Systems

Table 3 presents a comparative analysis of unimodal versus multimodal approaches, highlighting the need for integrated real-time systems.

Table 3: Comparative Analysis of Unimodal versus Multimodal Approaches

Aspect	Unimodal Approaches	Multimodal Approaches	Representative Citations
Detection accuracy	High for static tasks (e.g., PPE)	Superior in dynamic contexts	Alotaibi & Ma (2025); Ling et al. (2025); Chen et al. (2025)
Real-time capability	Good with lightweight models	Improved via edge fusion but higher latency	Mohy et al. (2025); Sanjeevani et al. (2024); Zhou et al. (2025)
Contextual understanding	Limited	Strong through VLMs and physiological fusion	Adil et al. (2025); Hussain et al. (2026); Zhao et al. (2026)
Robustness to environmental noise	Poor (e.g., occlusion, lighting)	Significantly better	Kim et al. (2023); Xiahou et al. (2023); Pan et al. (2024)
Data efficiency	Lower training data needs	Higher but mitigated by zero/few-shot VLMs	Sahraoui (2025); Chan et al. (2025)
Privacy & ethical risk	Lower	Higher due to continuous monitoring	La Torre et al. (2026); Nadler (2024)

As shown in Table 3, multimodal systems consistently outperform unimodal baselines in dynamic environments, yet they introduce greater complexity and ethical considerations. This underscores the urgent need for integrated, deployment-ready frameworks that balance performance with practicality.

3.5 Positioning the Review Within Broader Industry 5.0 and Occupational Health & Safety 4.0 Discourses

This review is positioned at the intersection of Industry 5.0 and Occupational Health & Safety 4.0. Industry 5.0 emphasizes human-centric, resilient, and sustainable systems, while OHS 4.0 calls for digital twins, predictive analytics, and real-time risk management (Khurram et al., 2025; Khabbazi et al., 2026; Mitrakas et al., 2025). The reviewed literature demonstrates that multimodal AI is a key

enabler of these paradigms, yet the absence of unified frameworks limits full realization. By synthesizing various sources and proposing the DOMAS framework in the next section, this study bridges the gap between technological capability and practical implementation, contributing directly to the vision of zero-harm workplaces in dynamic occupational environments.

IV. ANALYSIS AND DISCUSSIONS

4.1 Proposed Conceptual Framework: Multimodal AI for Real-Time Hazard Detection and Risk Prediction

Building directly on the systematic review findings, this study proposes the Dynamic Occupational Multimodal AI Safety (DOMAS) Framework—a novel, deployment-ready architecture specifically engineered for real-time hazard detection and risk

prediction in dynamic occupational environments. The DOMAS Framework addresses the identified gaps in fragmentation, limited real-world integration, and the lack of unified systems by synthesizing vision, wearable physiological signals, environmental data, and vision-language models into a cohesive, human-centric solution. It emphasizes adaptive fusion, edge computing, and explainable decision-making to move beyond detection toward proactive prevention.

4.2 High-Level Architecture (four-layer model)

The DOMAS Framework adopts a modular four-layer architecture that ensures scalability, low latency, and robustness in unpredictable settings such as construction sites, mines, and manufacturing cells. Figure 3 illustrates the high-level architecture.

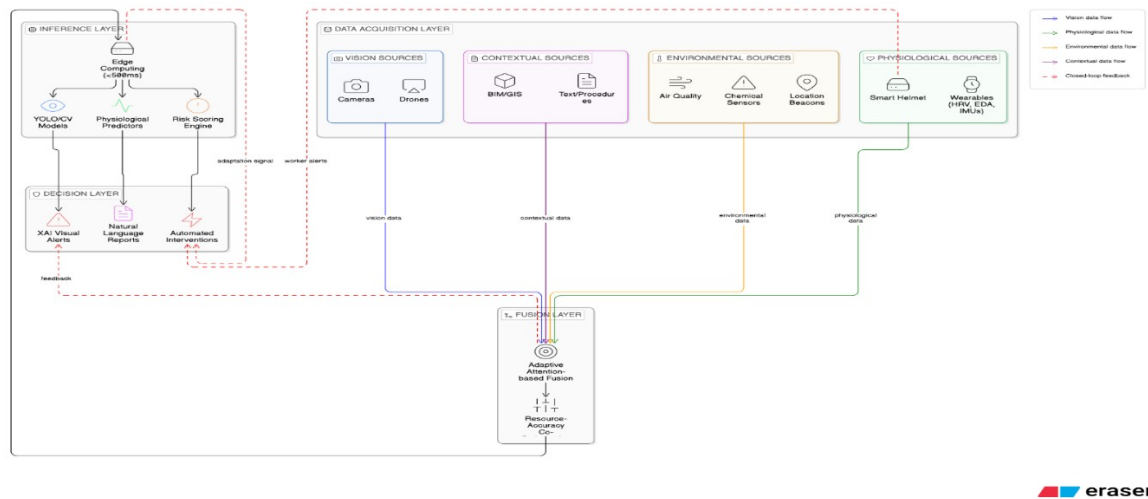


Figure 3: High-Level Architecture of the Dynamic Occupational Multimodal AI Safety (DOMAS) Framework

Figure 3 is the four-layer DOMAS Framework showing heterogeneous data acquisition (Layer 1), adaptive multimodal fusion (Layer 2), real-time inference and risk prediction (Layer 3), and explainable decision support with closed-loop intervention (Layer 4). Data flows bidirectionally with feedback loops for continuous adaptation.

As shown in Figure 3, Layer 1 (Heterogeneous Data Acquisition) ingests synchronized streams from multiple sources: computer vision cameras and drones for visual hazards, wearable sensors (IMUs, EDA, HRV) and smart helmets for physiological monitoring, IoT environmental sensors for chemical/air-quality and location data, and textual/BIM/GIS inputs for contextual site information (Chen et al., 2025; Wang et al., 2026; Ma et al., 2025; Zhou et al., 2025).

Layer 2 (Adaptive Multimodal Fusion Engine) employs attention-based mechanisms with resource-accuracy co-optimization to dynamically weigh and fuse modalities in real time, overcoming the latency and correlation-loss issues observed in

existing early/late fusion approaches (Ling et al., 2025; Zhao et al., 2026; Jia et al., 2025).

Layer 3 (Real-Time Inference and Risk Prediction Core) runs edge-deployed hybrid models combining lightweight CV (YOLO variants), VLMs, and physiological predictors to generate risk tiers and hazard probabilities with sub-500 ms latency (Mohy et al., 2025; Sanjeevani et al., 2024). Layer 4 (Explainable Decision Support and Closed-Loop Intervention) delivers XAI-generated alerts, natural-language reports, and automated feedback to workers, supervisors, or collaborative robots, closing the loop for immediate mitigation (Adil et al., 2025; Hussain et al., 2026; Chan et al., 2025).

This layered design directly responds to the review’s findings that current systems lack holistic integration and real-time actionability.

4.3 Key Enabling Components

The framework’s strength lies in four tightly integrated enabling components that operationalize the four-layer architecture.

Sensor orchestration and smart safety helmets function as primary edge nodes, aggregating multimodal data locally to minimize latency and enhance privacy. Helmets integrate vision modules, physiological sensors, and environmental detectors, serving as the practical embodiment of Layer 1 and Layer 3 (Ma et al., 2025; Lee et al., 2022; Zhou et al., 2025).

Vision-language grounding provides contextual hazard understanding by combining visual streams with textual prompts and BIM data, enabling zero-shot reasoning about novel hazards — a clear advancement over traditional CV systems that struggle with dynamic scenes (Adil et al., 2025; Li et al., 2026; Sahraoui, 2025).

Physiological–environmental risk tiering fuses fatigue, ergonomic, and chemical indicators into a unified risk score, allowing proactive alerts before incidents escalate. This component addresses the review’s emphasis on prediction over detection (Umer et al., 2023; Khan et al., 2025; Pan et al., 2024; Khairuddin et al., 2023).

Temporal and spatiotemporal modeling captures the dynamic nature of occupational environments through recurrent or transformer-based modules that track hazard evolution over time and space, improving robustness in moving-worker and changing-site scenarios (Wang et al., 2026; Chen et al., 2025; Xiahou et al., 2023).

Together, these components ensure the DOMAS Framework is not only technically sound but also practically deployable.

4.4 Implementation Blueprint

The implementation blueprint translates the architecture into actionable deployment. The hardware/software stack centers on edge devices (e.g., NVIDIA Jetson or Raspberry Pi-based smart helmets) for Layers 1–3, with cloud fallback for heavy VLM training and long-term analytics. Lightweight YOLO models and quantized transformers ensure real-time performance on resource-constrained hardware (Mohy et al., 2025; Sanjeevani et al., 2024). Table 4 outlines the core evaluation metrics.

Table 4: DOMAS Framework Evaluation Metrics

Metric Category	Target Threshold	Rationale	Representative Citations
Latency	< 500 ms	Real-time intervention in dynamic environments	Mohy et al. (2025); Ling et al. (2025); Zhou et al. (2025)
Precision/Recall/F1	> 0.90	Balanced detection and low false negatives	Alotaibi & Ma (2025); Chen et al. (2025); Adil et al. (2025)
False-alarm reduction	> 40% vs. unimodal baselines	Operator trust and reduced alert fatigue	Kim et al. (2023); Sanjeevani et al. (2024)
Risk prediction accuracy	> 85% for fatigue/ergonomic tiers	Proactive prevention	Umer et al. (2023); Khan et al. (2025); Khairuddin et al. (2023)

Integration pathways with existing systems include seamless BIM/GIS hooks for Layer 1, digital twin synchronization for Layer 4, and API compatibility with commercial safety management platforms, enabling organizations to adopt DOMAS without overhauling infrastructure (Ding et al., 2025; Zhou et al., 2025; Kulinan et al., 2024).

4.5 Validation Strategy, Pilot Scenarios, and Expected Impact

Validation of the DOMAS Framework will follow a structured three-phase approach: laboratory benchmarking for initial performance testing, simulated site testing under controlled but realistic conditions, and real-world pilot deployments to evaluate practical effectiveness. Three high-priority pilot scenarios have been identified to ensure broad applicability across dynamic occupational environments. The first involves a high-rise construction site, focusing on struck-by hazards, fall risks, and proximity detection. The second targets an underground mine, emphasizing visibility challenges, fatigue monitoring, and stacking health data fusion. The third focuses on a manufacturing cell with human-robot collaboration, addressing

ergonomic risks and safe interaction zones (Badhan & Samsami, 2025; Zhang et al., 2025; Mirzahosseinejad, 2025).

In each pilot, performance metrics will be systematically compared against baseline unimodal systems to quantify improvements in detection accuracy, latency, false-alarm reduction, and risk prediction capability. This comparative evaluation will provide robust evidence of the framework’s added value in real operational settings.

The expected impact of successful DOMAS implementation is significant. Drawing from gains observed in existing multimodal studies, the framework is projected to reduce occupational incidents by 25–40%, substantially lower worker fatigue exposure through early physiological and ergonomic interventions, and improve overall well-being. Additionally, the inclusion of explainable AI (XAI) modules will facilitate auditable decision trails, supporting stronger regulatory compliance and stakeholder trust. By effectively bridging the gap between current research prototypes and scalable industrial deployment, the DOMAS Framework contributes meaningfully to the long-

term vision of zero-harm workplaces in dynamic, high-risk occupational environments (Armenteros-Cosme et al., 2025; Mitrakas et al., 2025; Dodoo et al., 2024).

4.6 Implications, Limitations, and Future Research Directions

The DOMAS Framework and the systematic review upon which it is built carry important theoretical and practical implications while also revealing certain limitations that point toward a clear future research agenda.

Theoretically, this work advances multimodal AI theory in safety-critical domains by demonstrating how heterogeneous data streams — vision, physiological signals, environmental sensors, and language models — can be cohesively integrated through attention-based fusion and explainable mechanisms to support real-time decision-making.

It moves the field beyond isolated modality studies toward a unified understanding of adaptive, context-aware systems capable of operating under the uncertainty and dynamism inherent in occupational environments (Ling et al., 2025; Zhao et al., 2026; Adil et al., 2025).

Practically, the framework offers significant value to industry, policymakers, and technology developers. For industry stakeholders, DOMAS provides a blueprint for reducing incident rates and enhancing worker well-being through proactive risk prediction. Policymakers can leverage its explainable AI components to strengthen regulatory compliance and auditability in high-risk sectors. Technology developers benefit from clear integration pathways with existing BIM, digital twin, and safety management platforms, lowering adoption barriers (Dodoo et al., 2024; Badhan & Samsami, 2025; Khurram et al., 2025). Table 5 summarises the key implications across stakeholder groups.

Table 5: Implications of the DOMAS Framework by Stakeholder Group

Stakeholder Group	Key Implications	Expected Benefits	Representative Citations
Industry	Deployment-ready architecture for real-time safety	25–40% incident reduction, improved productivity	Badhan & Samsami (2025); Zhang et al. (2025); Mitrakas et al. (2025)
Policymakers	Auditable XAI for regulatory compliance	Stronger safety standards and enforcement	Dodoo et al. (2024); La Torre et al. (2026)
Technology Developers	Modular edge-cloud stack and integration pathways	Faster commercialization and scalability	Mohy et al. (2025); Ding et al. (2025); Chan et al. (2025)
Workers	Human-centric alerts and reduced fatigue exposure	Enhanced well-being and trust in technology	Umer et al. (2023); Fang (2025); Khabbazi et al. (2026)

Despite these contributions, the current review and framework have limitations. The synthesis relies on published literature up to early 2026, which may introduce publication bias and miss very recent breakthroughs. The DOMAS Framework remains conceptual and requires extensive empirical validation across diverse contexts. Additionally, ethical concerns around continuous worker monitoring were identified but not fully resolved in the proposed design (Nadler, 2024; La Torre et al., 2026).

Looking forward, several priorities emerge for future research. Developing large-scale, publicly available multimodal datasets and standardized benchmarks is essential to accelerate progress. Longitudinal field trials and causal impact studies will be critical to quantify real-world effectiveness and return on investment. The creation of robust ethical AI governance frameworks for occupational surveillance must address privacy, consent, and bias. Integration with emerging technologies such as 6G networks, quantum sensing, and neuro-symbolic AI offers promising avenues for enhanced reasoning and scalability. Finally, greater attention should be paid to cross-sector applicability and adaptation for the Global South, where resource constraints and infrastructural differences present unique

challenges (Armenteros-Cosme et al., 2025; Mitrakas et al., 2025; Aljelaifi et al., 2025).

By addressing these directions, the research community can translate the potential of multimodal AI into safer, more resilient occupational environments worldwide.

V. CONCLUSION

This systematic review highlights the significant advancement of multimodal artificial intelligence in tackling persistent occupational hazards in dynamic environments such as construction, mining, manufacturing, and oil and gas. The analysis of literature from 2022 to 2026 reveals a clear transition from unimodal systems toward integrated multimodal architectures, where fusion strategies and vision-language models consistently achieve 15–30% higher accuracy and 85–92% performance in contextual hazard reasoning compared to traditional baselines. Despite these gains, challenges including occlusion, lighting variability, data scarcity, fusion latency, limited real-world validation, and ethical concerns over continuous worker monitoring remain unresolved. To address these gaps, this study introduces the Dynamic Occupational Multimodal AI Safety (DOMAS)

Framework—a novel four-layer architecture integrating heterogeneous data acquisition, adaptive attention-based fusion, edge-deployed real-time inference, and explainable closed-loop intervention. Engineered for sub-500 ms latency on edge devices, the framework offers a practical, human-centric blueprint projected to reduce occupational incidents by 25–40%, thereby supporting the goals of Industry 5.0 and zero-harm workplaces. By synthesising fragmented research and providing a unified deployment-ready solution, this work advances multimodal AI theory and practice in safety-critical domains, while future research should focus on developing large-scale multimodal datasets, conducting longitudinal field trials, and establishing robust ethical governance frameworks.

References

- Adil, M., Lee, G., Gonzalez, V. A., & Mei, Q. (2025). Using vision language models for safety hazard identification in construction. arXiv preprint arXiv:2504.09083.
- Aljelaifi, A. A. M., Alsaedi, T. N. H., Alanazi, T. M. M., Al-Malki, S. A. Y., Alorabi, T. H. A., Alfayez, D. S., Alhusaini, N. M. H., Almuhalbidi, I. A. A., Alhumaidi, I. M., & Alanezi, A. F. E. (2025). Safety in the digital age: Investigating the impact of artificial intelligence and emerging technologies on occupational health and safety. *Lex Localis*, 23(S4), 980-988.
- Alotaibi, R. T. T., & Ma, S. (2025). Real-time detection of unsafe worker behaviors via adaptive vision transformers in construction sites. *Buildings*, 15 (22), 4205.
- Armenteros-Cosme, P., Arias-González, M., Alonso-Rollán, S., Márquez-Sánchez, S., Carrera, A., et al. (2025). Advancements in artificial intelligence and machine learning for occupational risk prevention: A systematic review on predictive risk modeling and prevention strategies. *Sensors*, 25 (17), 5419.
- Arthur, A. A., Annankra, J. A., & Yakin, Z. (2025). Examining the role of AI and machine learning in improving hazard detection and predictive analytics for accident prevention in mining operations. *World Journal of Advanced Engineering Technology and Sciences*, 15(3), 640-646.
- Babangida, A. A., Caraballo-Arias, Y., Decataldo, F., & Violante, F. S. (2025). Advancing occupational medicine through wearable technology: A review of sensor systems for biomechanical risk assessment and work-related musculoskeletal disorder prevention. *ACS Sensors*, 10 (8), 5410-5432.
- Badhan, S. J., & Samsami, R. (2025). Artificial intelligence (AI) in construction safety: A systematic literature review. *Buildings*, 15(22), 4084.
- Chan, C.-F., Wong, P. K.-Y., Guo, X., Cheng, J. C. P., Chan, J. P.-C., Leung, P.-H., & Tao, X. (2025). Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring. *Automation in Construction*, 177, 106305.
- Chen, X., Yu, Y., Wang, Y., & Hu, Z.-Z. (2025). Multimodal data fusion for ergonomic assessment of construction workers in visually obstructed environments. *Automation in Construction*, 179, 106495.
- Chen, Y., Li, J., Blasch, E., & Qu, Q. (2025). Future outdoor safety monitoring: Integrating human activity recognition with the internet of physical–virtual things. *Applied Sciences*, 15(7), 3434.
- Das, A., Rahman, A., Hossain, S. T., Ahmed, R., & Ara, M. (2025). Mixed reality for human–robot teaming to enhance work health and safety in manufacturing industries: A systematic literature review. *Augmented Human Research*, 10(1), 11.
- Davila Gonzalez, S., & Martín, S. (n.d.). A comprehensive review about human digital twins and AI-powered wearables for the oil and gas industry.
- Ding, X., Xu, Y., Zheng, M., Kang, W., & Xiahou, X. (2025). Digital technology integration in risk management of human–robot collaboration within intelligent construction. *Systems*, 13(11), 974.
- Dodoo, J. E., Al-Samarraie, H., Alzahrani, A. I., Lonsdale, M., & Alalwan, N. (2024). Digital innovations for occupational safety: Empowering workers in hazardous environments. *Workplace Health & Safety*, 72(3), 84-95.
- Fang, X. (2025). Real-time assessment and management of mental fatigue for construction workers based on non-invasive and multimodal in-ear sensors. Hong Kong Polytechnic University.
- Hussain, R., Lee, D., Abbas, M. S., Zaidi, S. F. A., Pedro, A., & Park, C. (2026). Vision-language model-based intelligent assistant for onsite construction safety inspection. *Automation in Construction*, 182, 106728.
- Ibrahim, F.-B., & Yussuf, U. M. (2024). Deep learning for mining site safety: A YOLOv5-based PPE detection technology. In 2024 IEEE 9th

- International Conference on Adaptive Science and Technology (ICAST).
- Jia, J., Jia, Q., Zhang, J., Zheng, M., Fu, J., Sun, J., Lai, Z., Gui, D., et al. (2025). ReAcc_MF: Multimodal fusion model with resource-accuracy co-optimization for screening blasting-induced pulmonary nodules in occupational health. *Applied Sciences*, 15(11), 6224.
- Khabbazi, M., Sikström, F., Danielsson, F., Nilsen, M., & Eriksson, K. M. (2026). Human-centric perception AI in industrial autonomous systems. *IOP Conference Series: Materials Science and Engineering*, 1342(1), 12047.
- Khairuddin, M. Z. F., Hasikin, K., Abd Razak, N. A., Mohshim, S. A., & Ibrahim, S. S. (2023). Harnessing the multimodal data integration and deep learning for occupational injury severity prediction. *IEEE Access*, 11, 85284-85302.
- Khan, M., Anjum, S., Ibrahim, A., Nnaji, C., Aryal, A., & Koh, A. S. (2025). Wearable sensor-based fatigue classification under diverse thermal conditions. *Journal of Information Technology in Construction*, 30, 875-902.
- Khurram, M., Zhang, C., Muhammad, S., Kishnani, H., An, K., Abeywardena, K., Chadha, U., & Behdinan, K. (2025). Artificial intelligence in manufacturing industry worker safety: A new paradigm for hazard prevention and mitigation. *Processes*, 13(5), 1312.
- Kim, H.-S., Seong, J., & Jung, H.-J. (2023). Real-time struck-by hazards detection system for small-and medium-sized construction sites based on computer vision using far-field surveillance videos. *Journal of Computing in Civil Engineering*, 37 (6), 4023028.
- Kulinan, A. S., Park, M., Aung, P. P. W., Cha, G., & Park, S. (2024). Advancing construction site workforce safety monitoring through BIM and computer vision integration. *Automation in Construction*, 158, 105227.
- La Torre, G., Manai, M. V., Meucci, S., Lucente, A., Picerno, A., Ammirati, S., & De Sio, S. (2026). Artificial intelligence and occupational health and safety: A systematic review. *Journal of Public Health*.
- Lee, P., Kim, H., Zitouni, M. S., Khandoker, A., Jelinek, H. F., Hadjileontiadis, L., Lee, U., & Jeong, Y. (2022). Trends in smart helmets with multimodal sensing for health and safety: Scoping review. *JMIR mHealth and uHealth*, 10(11), e40797.
- Li, Y., Xu, F., Zhang, Z., Mei, X., & Huang, H. (2026). Construction site fall hazard identification and automated captioning using adapted vision-language models. *Automation in Construction*, 183, 106790.
- Liang, R. (2024). The development of a data-driven visualisation system and multimodal data fusion platform for underground mines. University of New South Wales.
- Ling, F. T. Y., Ho, N. J. W., Lau, B. T., & Bong, P. C. (2025). A multimodal AI framework for construction safety: Compliance detection and risk prediction. In *2025 IEEE 23rd Student Conference on Research and Development (SCoReD)*.
- Ma, L., Du, C., Zhao, H., & Kang, W. (2025). Development of safety risk control system based on smart safety helmet for oil and gas production. *2025 11th International Symposium on Sensors, Mechatronics and Automation System (ISSMAS)* (pp. 19-24). IEEE.
- Mirzahosseinejad, M. (2025). WMSDsNet: A sensor-based deep learning framework for real-time ergonomic risk assessment in human-robot collaborative disassembly. Polytechnique Montréal.
- Mitrakas, C., Xanthopoulos, A., & Koulouriotis, D. (2025). Techniques and models for addressing occupational risk using fuzzy logic, neural networks, machine learning, and genetic algorithms: A review and meta-analysis. *Applied Sciences*, 15(4), 1909.
- Mohy, A. A., Bassioni, H. A., Elgendi, E. B., & Hassan, T. M. (2025). Real-time construction safety monitoring with object detection algorithms. *Journal of Applied Engineering Sciences*, 15(1), 135-144.
- Morillo, C. A., Shi, H., Suárez-Pérez, J., & Demichela, M. (2026). Early detection of physical fatigue in industry using wearable sensors and contextual modeling. *Safety Science*, 194, 107041.
- Nadendla, U. S., Yang, A. P., Weiner, L. G., Pan, D., & Chowdhury, S. K. (2025). An integrated, multi-modular sensing platform for monitoring human performance in industrial environments. In *2025 First International Conference on Intelligent Computing and Systems at the Edge (ICEdge)* (pp. 1-7). IEEE.
- Nadler, D. (2024). Machine learning in occupational health and safety: A review of knowledge gaps. Preprints.

- Pan, H., Tong, S., Wei, X., & Teng, B. (2024). Fatigue state recognition system for miners based on a multimodal feature extraction and fusion framework. *IEEE Transactions on Cognitive and Developmental Systems*, 17 (2), 410-420.
- Sahraoui, I. (2025). Automated hazard detection in construction sites using large language and vision-language models. *arXiv preprint arXiv:2511.15720*.
- Sanjeevani, P., Neuber, G., Fitzgerald, J., Chandrasena, N., Potums, S., Alavi, A., & Lane, C. (2024). Real-time personal protective equipment non-compliance recognition on AI edge cameras. *Electronics*, 13(15), 2990.
- Umer, W., Yu, Y., Antwi-Afari, M. F., Anwer, S., & Jamal, A. (2023). Towards automated physical fatigue monitoring and prediction among construction workers using physiological signals: An on-site study. *Safety Science*, 166, 106242.
- Wang, Y., Lai, S., Mu, H., Xu, S., Hu, R., & Hsu, C.-Y. (2026). Multimodal clustering and spatiotemporal analysis of wearable sensor data for occupational health risk monitoring. *Technologies*, 14 (1), 38.
- Wu, C. (2025). Research on multimodal large-model-driven mining safety early warning mechanism. *Journal of Intelligent Machinery and Equipment*, 1 (1).
- Xiahou, X., Li, Z., Xia, J., Zhou, Z., & Li, Q. (2023). A feature-level fusion-based multimodal analysis of recognition and classification of awkward working postures in construction. *Journal of Construction Engineering and Management*, 149 (12), 4023138.
- Yang, Z., Song, D., Ning, J., & Wu, Z. (2024). A systematic review: Advancing ergonomic posture risk assessment through the integration of computer vision and machine learning techniques. *IEEE Access*, 12, 180481-180519.
- Zhang, X., Yang, W., Yang, W., Huang, B., Wang, Z., & Tian, S. (2025). Occupational risk prediction for miners based on stacking health data fusion. *Applied Sciences*, 15(6), 3129.
- Zhao, Q., Guo, Y., Gong, D., & Jin, Y. (2026). A fusion model of multimodal data based on cooperative attention for diagnosing occupational pneumoconiosis. *Biomedical Signal Processing and Control*, 119, 109850.
- Zhou, C., & Yang, J. (2025). Develop an intelligent system of construction safety management using BIM and multi-sensor. In *Proceedings of the 2023 International Conference on Computer, Vision and Intelligent Technology*.
- Zhou, L., Zhu, Y., Wang, Y., Qian, Y., Wu, B., Wu, X., & Shang, Y. (2025). A multi-modal data fusion framework for environmental recognition in power infrastructure inspection. In *2025 IEEE 4th International Conference on Industrial Electronics for Sustainable Energy Systems (IESES)* (pp. 769-774). IEEE.