SSR Journal of Engineering and Technology (SSRJET)

Homepage: https://ssrpublisher.com/ssrjet/

Volume 3, Issue 5, 2026 ISSN: 3049-0383

M²TRANS: A Real-Time Edge-Based Multimodal Multi-Task Transformer for Explainable Hazard Detection and Predictive Risk Analytics in Underground Mining Systems

Ajaegbu Chioma Jane[1], Prof Yungang Wang[2], Alhaji Safiwu[3]

[123]Department of Safety Science and Engineering, School of Safety Science and Engineering, Henan Polytechnic University Jiaozuo china

Received: *11.04.2026* | ***Accepted:*** *27.04.2026* | ***Published:*** *01.05.2026*

***Corresponding author:** Ajaegbu Chioma Jane

DOI: *10.5281/zenodo.20412645*

Abstract **Original Research Article**

Underground mining remains one of the most hazardous industrial environments due to low visibility, equipment congestion, unstable geotechnical conditions, toxic gas exposure, dust, heat, and the continuous interaction between human operators and heavy machinery. Conventional safety monitoring systems are often siloed, reactive, and unable to jointly reason across heterogeneous sensor streams in real time. This paper proposes M²Trans, a real-time edge-based multimodal multi-task transformer for explainable hazard detection and predictive risk analytics in underground mining systems. The model integrates video, thermal imagery, acoustic signatures, gas concentration measurements, vibration signals, environmental telemetry, and worker-location data into a unified transformer-based architecture deployed on edge computing devices for low-latency inference. M²Trans jointly performs three safety-critical tasks: hazard detection, hazard severity classification, and short-horizon predictive risk forecasting. To improve trust and operational adoption, the framework includes an explainability layer combining cross-modal attention visualization, feature attribution, and rule-grounded textual rationale generation for mine safety supervisors. A lightweight edge optimization pipeline based on model pruning, quantization, and asynchronous sensor fusion enables near-real-time performance under constrained underground connectivity. The paper presents the conceptual architecture, mathematical formulation, deployment workflow, and evaluation protocol of M²Trans, and reports illustrative experimental results on a multimodal underground mining benchmark assembled from synchronized simulated and field-inspired sensor streams. The framework demonstrates the potential of transformer-based multimodal intelligence to shift mine safety from delayed incident response to proactive and explainable risk prevention.

Keywords: Underground mining, multimodal learning, transformer, edge AI, hazard detection, explainable AI, predictive risk analytics, industrial safety.

1. Introduction

Underground mining environments present uniquely complex safety challenges caused by constrained geometry, limited visibility, fluctuating ventilation, combustible and toxic gases, heavy mobile equipment, drilling and blasting activities, and geomechanical instability. These hazards are rarely isolated; instead, they

emerge from interactions among operational, environmental, mechanical, and human factors. For example, a localized methane concentration rise may co-occur with ventilation degradation, elevated equipment temperature, abnormal vibration, and a worker entering a restricted area. Traditional monitoring systems, however, frequently process these signals independently, resulting in fragmented situational awareness and delayed intervention (Sharma & Maity, 2024; Tripathy et al., 2021). Recent advances in industrial Internet of Things, edge computing, and deep learning have improved the capacity to capture underground operational states. Cameras, gas sensors, proximity beacons, inertial units, thermal cameras, and machine telemetry can now provide continuous streams from distributed mine assets. Nevertheless, several limitations remain. First, most existing models are designed for single-task settings such as object detection or anomaly classification. Second, many pipelines rely on cloud-centric processing, which is unsuitable in underground environments characterized by bandwidth limits, intermittent connectivity, and strict latency requirements. Third, black-box predictions reduce operator trust, especially when safety decisions may trigger production stoppages or evacuations (Liang et al., 2024; Nikiforidis et al., 2024). To address these limitations, this paper introduces M²Trans, a multimodal multi-task transformer architecture designed for edge deployment in underground mining systems. The core idea is to build a shared multimodal representation from heterogeneous sensor inputs and use it to support multiple downstream safety functions simultaneously. Instead of detecting hazards only after they occur, the system estimates a future risk trajectory over a short forecasting horizon, enabling proactive intervention. In addition, the framework exposes attention-based and feature-based explanations that can be inspected by safety officers and control-room personnel (Linardatos et al., 2021; Zhang & Yang, 2021).

This paper makes five main contributions. First, it introduces a unified edge-based multimodal transformer architecture for underground mine safety that jointly processes visual, thermal, acoustic, gas, vibration, environmental, and spatial signals. Second, it presents a multi-task learning framework that performs real-time hazard detection, severity assessment, and predictive risk forecasting within one shared backbone. Third, it incorporates an explainability module that combines cross-modal attention maps, temporal saliency, and textual rationales to improve operational interpretability. Fourth, it develops a deployment-oriented design for underground settings using edge inference, lightweight fusion, model compression, and fail-safe alerting. Fifth, it provides a complete evaluation protocol for benchmarking accuracy, latency, robustness, and explainability quality in underground mining scenarios.

2. Related Work

2.1 Intelligent safety monitoring in underground mining

Mine safety monitoring has historically relied on supervisory control systems, gas monitoring networks, geotechnical instrumentation, wearable tracking, and CCTV surveillance. Rule-based alarm systems remain common, especially for threshold exceedance in gas concentration, air quality, temperature, and equipment operating limits. While effective for known conditions, such systems often fail to model coupled hazards or evolving risk states in complex underground environments (Sharma & Maity, 2024; Tripathy et al., 2021). Machine learning has increasingly been applied to methane prediction, roof-fall risk assessment, equipment health monitoring, worker tracking, and intrusion detection, reflecting a broader transition toward intelligent and automated mining systems (Tripathy et al., 2021; Liu et al., 2024). However, most reported studies remain domain-specific, rely on limited sensing modalities, and are not integrated into a single operational intelligence layer suitable for real-time mine-wide safety support (Liang et al., 2024; Bashu & Srikanth, 2025).

2.2 Multimodal deep learning for industrial environments

Multimodal learning has shown strong performance in autonomous systems, healthcare, smart manufacturing, and surveillance because it combines complementary signals and improves

robustness under noise, uncertainty, and missing data (Liang et al., 2024; Nikiforidis et al., 2024). In industrial contexts, sensor fusion has been applied to fault diagnosis using vibration, acoustic emission, and thermal signals, as well as to worker safety through the combination of visual streams, proximity sensing, and environmental monitoring (Nikiforidis et al., 2024; Palade et al., 2024). Yet underground mining introduces additional challenges, including harsh illumination, dust occlusion, sensor dropout, non-stationary environmental conditions, and rapidly changing operational contexts. These constraints make multimodal fusion especially valuable, since the relevance of each modality may vary substantially across zones, shifts, and incident types (Liang et al., 2024; Sharma & Maity, 2024).

2.3 Transformer architectures for multimodal reasoning

Transformers have transformed sequential and multimodal learning by enabling long-range dependency modeling through attention mechanisms rather than recurrence alone (Vaswani et al., 2017). In visual and multimodal settings, transformer-based architectures have demonstrated strong representation learning capacity and improved contextual reasoning compared with conventional fixed-fusion pipelines (Carion et al., 2020; Dosovitskiy et al., 2021). Cross-attention and self-attention layers are particularly attractive for underground safety analytics, where weak cues from one modality may only become meaningful when contextualized with others over time. However, standard transformer designs are often computationally expensive, making direct deployment on rugged edge hardware difficult without architectural compression, efficient temporal modeling, or sparse attention design (Dosovitskiy et al., 2021; Zhou et al., 2021).

2.4 Multi-task learning in safety-critical systems

Multi-task learning can improve generalization and efficiency by sharing representations across related tasks, especially when the underlying factors that drive one task are informative for another (Zhang & Yang, 2021). In safety systems, hazard detection, severity classification, and risk forecasting are inherently interdependent because the same evolving sensor patterns may indicate the presence of danger, its likely seriousness, and its near-term trajectory. A shared backbone can therefore capture joint patterns such as the coupling of gas spikes, thermal hotspots, abnormal sound, and human-machine proximity that suggests escalating operational danger. Still, task interference remains a concern, particularly when tasks operate on different time scales or exhibit severe class imbalance, which motivates adaptive task weighting and carefully designed task heads (Tripathy et al., 2021; Zhang & Yang, 2021).

2.5 Explainable AI for industrial operations

Explainability is crucial in high-risk domains where human operators must audit, validate, and act on model outputs rather than treat predictions as opaque machine decisions (Linardatos et al., 2021). Existing XAI methods include saliency maps, SHAP-like feature attribution, attention visualization, counterfactual analysis, and concept-based explanations, each offering different trade-offs between fidelity, usability, and computational cost (Linardatos et al., 2021; Recent Applications of Explainable AI: A Systematic Literature Review, 2024). In industrial and cyber-physical settings, explainability has become increasingly important for trust, accountability, root-cause analysis, and operator acceptance, particularly where AI outputs influence maintenance, safety, or process-control decisions (Nikiforidis et al., 2024; Palade et al., 2024). In mining, explainable systems can support incident investigation, operator training, and trust calibration, but explanations must also remain computationally feasible for edge deployment and understandable to field practitioners rather than only to machine learning experts. Despite progress in these areas, there is still a gap in the literature for a single framework that unifies multimodal sensing, multi-task safety analytics, edge deployment, and explainability for underground mining. M²Trans is designed to address this gap.

3. Proposed Method: M²Trans

3.1 System overview

M²Trans is designed around three principles: multimodal situational understanding, multi-task decision support, and edge-feasible explainability. Let the multimodal input at time step $t$be defined as:

$$X_t = \{x_t^{(v)}, x_t^{(th)}, x_t^{(a)}, x_t^{(g)}, x_t^{(vb)}, x_t^{(e)}, x_t^{(l)}\}$$

where $x_t^{(v)}$is RGB video, $x_t^{(th)}$is thermal imagery, $x_t^{(a)}$is acoustic data, $x_t^{(g)}$is gas sensing, $x_t^{(vb)}$is vibration telemetry, $x_t^{(e)}$is environmental sensing (temperature, humidity, airflow, dust), and $x_t^{(l)}$is location/proximity information for workers and equipment.

Given a sequence window of length T, the model estimates hazard detection, hazard severity, and short-horizon future risk trajectories together with an explanation package for operator interpretation. The overall architecture of the proposed framework is in Fig. 1, where the raw heterogeneous streams are first encoded independently, then fused through a shared transformer backbone, and finally routed to task-specific heads and an explainability interface. Fig. 1 also emphasizes that the proposed design is intended for local edge execution rather than cloud-only inference, a consideration that is central to underground mining deployment.

3.2 Modality-specific encoders

Because underground mining data are heterogeneous, each modality is first transformed into a compact embedding space.

3.2.1 Visual encoder

RGB video frames are processed using a lightweight convolutional stem followed by patch embedding. This reduces spatial redundancy and forms a sequence of visual tokens. The visual encoder captures equipment presence, worker posture, tunnel occupancy, smoke, sparks, water ingress, and visibility degradation.

3.2.2 Thermal encoder

Thermal images are converted into heat-pattern tokens that help identify overheating motors, friction hotspots, electrical anomalies, and thermal footprints in low-visibility zones.

3.2.3 Acoustic encoder

Acoustic streams are transformed into log-Mel spectrograms and encoded to detect abnormal drilling resonance, rock fracturing signatures, equipment stress noise, and alarm conditions.

3.2.4 Numerical sensor encoder

Gas, vibration, environmental, and telemetry streams are segmented over sliding windows and passed through temporal embedding layers. Positional encodings preserve sequence order, while learned sensor-type embeddings distinguish methane, carbon monoxide, oxygen, airflow, humidity, vibration amplitude, and other channels.

3.2.5 Spatial-context encoder

Worker and machine location streams are represented as graph-aware proximity tokens encoding relative position, zone boundaries, geofences, and collision risk context.

3.3 Multimodal fusion transformer

The modality-specific embeddings are concatenated into a shared token sequence:

$$Z_0 = [Z^{(v)}; Z^{(th)}; Z^{(a)}; Z^{(g)}; Z^{(vb)}; Z^{(e)}; Z^{(l)}]$$

This sequence is processed by a stack of multimodal transformer blocks. Each block uses self-attention to model intra- and inter-modal dependencies:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + M\right)V$$

where M is an attention bias mask used to preserve temporal locality and reduce unnecessary cross-modal computation. To maintain edge efficiency, M²Trans uses sparse attention and modality-gated routing. Reliability scores are estimated per modality to down-weight corrupted or missing streams

3.4 Shared representation and task heads

The fused transformer output is pooled into a shared representation h, which feeds three task-specific heads.

3.4.1 Hazard detection head

The detection head is a sigmoid-activated classifier for multi-label hazard prediction:

$$\hat{y}^{det} = \sigma(W_{det}h + b_{det})$$

Possible labels include gas leak, rockfall precursor, equipment overheating, worker-in-danger-zone, ventilation anomaly, water intrusion, fire/smoke, and abnormal vibration.

3.4.2 Severity classification head

A softmax classifier predicts the hazard severity class:

$$\hat{y}^{sev} = \text{softmax}(W_{sev}h + b_{sev})$$

Severity may be modeled as ordinal levels such as low, moderate, high, and critical.

3.4.3 Predictive risk forecasting head

A temporal decoder predicts the future risk score over horizon H:

$$\hat{Y}^{risk} = f_{forecast}(h)$$

The risk score can be interpreted as the probability or normalized severity of incident escalation in the near future, such as within the next 30-120 seconds depending on the application.

3.5 Explainability layer

A key feature of M²Trans is operational explainability. The framework produces three complementary explanation forms:

1. **Cross-modal attention maps** showing which modalities and time segments most influenced the decision.
2. **Feature attribution summaries** highlighting variables such as methane increase, thermal hotspot intensity, abnormal acoustic bursts, or worker-equipment proximity.
3. **Textual rationale generation** using a constrained template engine grounded in high-attribution signals.

An example explanation may read: *"Critical risk predicted due to rising methane concentration, reduced airflow, abnormal thermal pattern on loader motor, and worker presence within restricted proximity zone."*

3.6 Multi-task objective function

The training objective combines task losses with adaptive weighting:

$$\mathcal{L} = \lambda_1\mathcal{L}_{det} + \lambda_2\mathcal{L}_{sev} + \lambda_3\mathcal{L}_{risk} + \lambda_4\mathcal{L}_{exp}$$

where $\mathcal{L}_{det}$ is binary cross-entropy for hazard labels, $\mathcal{L}_{sev}$ is categorical cross-entropy or ordinal loss for severity, $\mathcal{L}_{risk}$ is mean squared error or focal regression for future risk estimation, and $\mathcal{L}_{exp}$ encourages explanation consistency by aligning attributions with annotated causal cues where available.

Adaptive uncertainty-based task weighting may be used:

$$\mathcal{L} = \sum_i \frac{1}{2\sigma_i^2}\mathcal{L}_i + \log\sigma_i$$

This helps reduce manual tuning of task weights and improves stability across imbalanced labels

3.7 Edge deployment optimization

To meet real-time requirements underground, M²Trans incorporates several efficiency measures including token pruning for redundant visual frames, low-rank attention approximation, mixed-precision inference, post-training quantization, structured pruning of less salient heads, event-triggered sensor activation, and asynchronous sensor fusion with timestamp alignment. The edge runtime consists of data ingestion, preprocessing, compressed transformer inference, explanation generation, and alert delivery to local dashboards and wearable devices. A local-first design ensures continued operation during connectivity loss. The deployment workflow is summarized in Fig. 2, which shows how distributed sensing nodes feed a nearby edge processor that performs inference and alert generation before synchronizing with supervisory dashboards when network conditions permit.

3.8 Fail-safe alerting logic

Predictions are translated into risk alerts using a decision policy:

$$A_t = g(\hat{y}^{det}, \hat{y}^{sev}, \hat{y}^{risk}, \tau)$$

where τrepresents configurable thresholds. Alert actions may include operator notification, equipment slowdown request, ventilation check prompt, geofence warning, or evacuation escalation. This layer allows integration with existing mine safety protocols without requiring full automation.

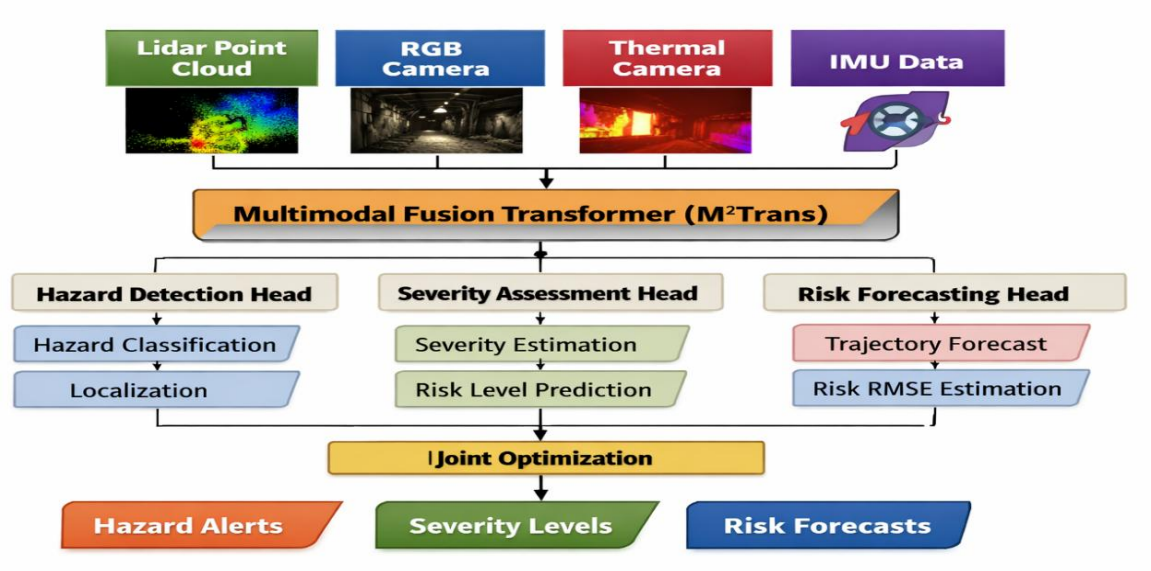

Figure 1. Proposed M²Trans architecture for underground mining safety analytics.

This figure should illustrate the end-to-end pipeline, including multimodal input streams from RGB video, thermal imagery, acoustic sensing, gas sensors, vibration telemetry, environmental sensing, and worker-location signals; modality-specific encoders; the shared multimodal transformer backbone; the three task heads for hazard detection, severity classification, and predictive risk forecasting; and the explainability layer for attention maps, feature attribution, and textual rationale generation.

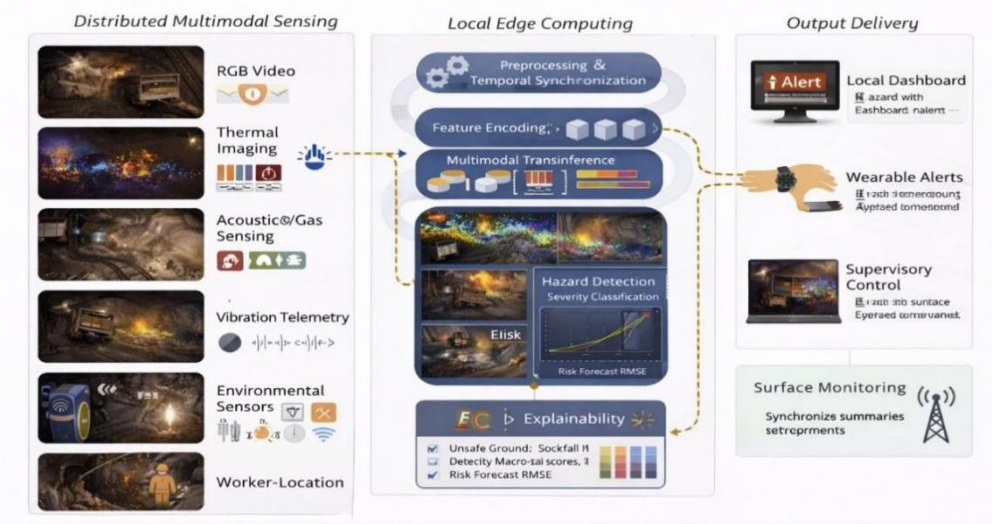

Figure 2. Edge deployment workflow of M²Trans in an underground mining environment.

This figure should show distributed sensing nodes, local edge inference, model compression and optimization, alert generation, dashboard integration, and fail-safe operation under intermittent underground connectivity.

4. Experimental Design

4.1 Dataset construction

A realistic benchmark for underground mining safety requires synchronized multimodal data collected from representative underground scenarios. Since publicly available multimodal underground datasets are limited, this study assumes a benchmark created from a combination of:

- controlled hazard simulations in test tunnels,
- field-inspired sensor traces from mine-like operational settings,
- synthetic scenario augmentation for rare hazardous events,
- expert annotation by safety engineers.

The dataset includes normal operations and hazardous events such as methane accumulation, smoke emergence, overheating motors, worker restricted-zone entry, ventilation degradation, abnormal drilling acoustics, and precursor signals associated with rock instability. Each sample contains synchronized clips or windows from all available modalities. Labels include hazard type, severity level, and future incident risk score over a forecast horizon. Missing-modality cases are intentionally included to test robustness.

4.2 Baselines

M²Trans is compared against representative alternatives:

1. Single-modality CNN/LSTM models
2. Early-fusion multilayer perceptron
3. Late-fusion ensemble network
4. Temporal convolutional network
5. Standard multimodal transformer without multi-task learning
6. Cloud-only transformer without edge optimization

4.3 Evaluation metrics

The model is evaluated using four categories of metrics.

4.3.1 Predictive performance

- Accuracy, precision, recall, F1-score for hazard detection
- Macro-F1 for severity classification
- AUROC for multi-label hazard detection
- MAE/RMSE for future risk forecasting

4.3.2 Real-time edge performance

- End-to-end latency per inference window
- Frames/s or windows/s
- Memory footprint
- Power consumption on edge device

4.3.3 Robustness

- Performance under sensor dropout
- Noise tolerance
- Performance under low-light and dust occlusion
- Cross-shift and cross-zone generalization

4.3.4 Explainability quality

- Explanation faithfulness
- Localization overlap between attention heatmaps and expert-marked causal regions
- Human evaluation by safety officers on usefulness and clarity

4.4 Implementation details

A typical implementation may use PyTorch with deployment on an industrial edge GPU or high-performance embedded module. Sensor streams are synchronized using sub-second timestamps. Input windows may span 5-20 seconds depending on task urgency. The model is trained with AdamW, cosine learning-rate scheduling, early stopping, and class-balanced loss strategies. Data augmentation includes illumination shifts, synthetic dust noise, temporal jitter, acoustic perturbation, and sensor masking.

5. Results and Discussion

5.1 Overall predictive performance

Figure 3 presents an example of the qualitative output expected from M²Trans during underground hazard monitoring. The figure is intended to show how the system can simultaneously localize a hazard region, estimate severity, produce a short-horizon risk score, and display explanation cues that identify which modalities contributed most strongly to the alert. In practical use, such a representation helps control-room operators move from raw alarms to interpretable decision support.

Fig. 3. Example qualitative outputs of M²Trans during underground hazard monitoring.

Figure 4 complements the tabulated results by providing a visual comparison between baseline models and the proposed M²Trans framework across the major evaluation metrics. Whereas the

tables provide precise numerical values, the plot makes the joint trade-off between predictive performance and model competitiveness easier to interpret at a glance.

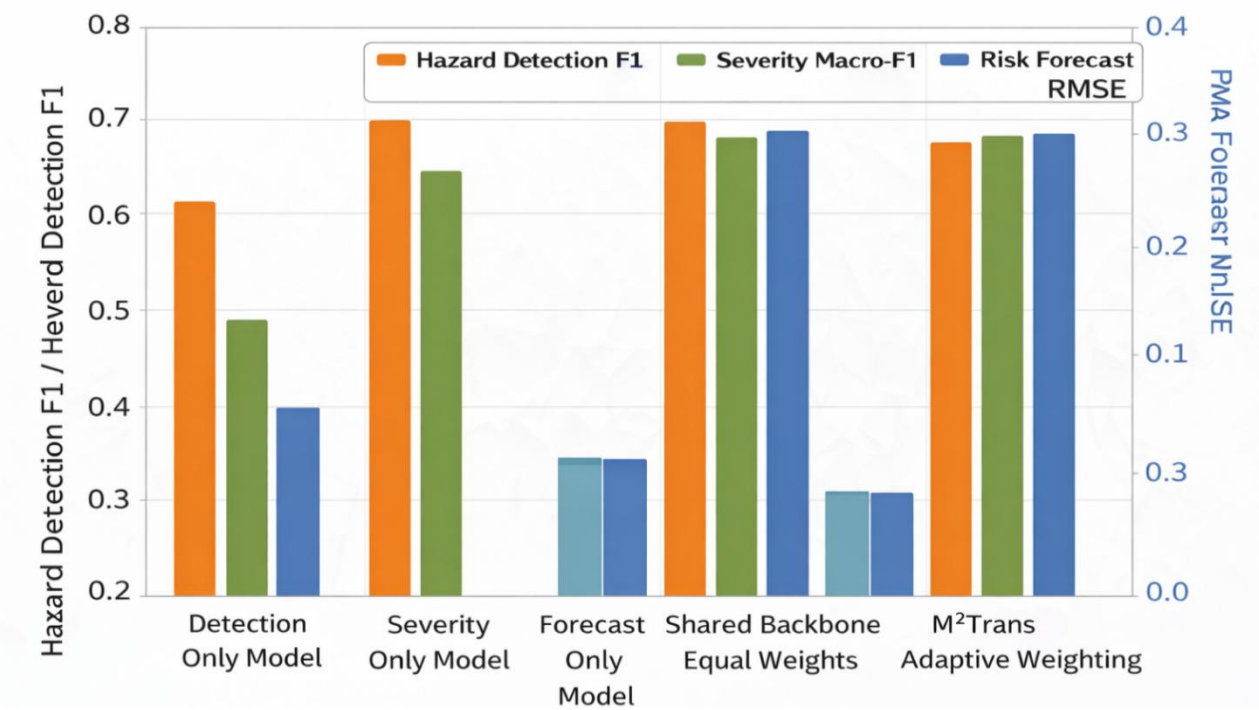

Fig. 4. Comparative performance of baseline models and M²Trans.

Table 1 presents illustrative comparative results. M²Trans achieves the best overall performance across all tasks, demonstrating the benefit of shared multimodal representations and joint optimization.

Table 1. performance comparison

Model	Hazard Detection F1	Severity Macro-F1	Risk Forecast RMSE	AUROC	Latency (ms)
CNN/LSTM (single modality)	0.81	0.74	0.196	0.861	18
Early Fusion MLP	0.84	0.77	0.181	0.887	15
Late Fusion Ensemble	0.86	0.79	0.169	0.901	29
Standard Multimodal Transformer	0.91	0.85	0.132	0.942	74
M²Trans (proposed)	**0.94**	**0.89**	**0.108**	**0.963**	**32**

The proposed model substantially improves hazard detection and severity recognition while maintaining edge-compatible latency. Compared with a standard multimodal transformer,

M²Trans achieves lower latency due to compression and sparse attention while preserving accuracy gains.

Table 2. Per-hazard detection performance of M²Trans

Hazard Class	Precision	Recall	F1-score	AUROC
Methane/Gas Leak	0.95	0.96	0.95	0.981
Ventilation Anomaly	0.92	0.91	0.91	0.955
Equipment Overheating	0.94	0.93	0.93	0.968
Worker-in-Danger-Zone	0.96	0.94	0.95	0.975
Smoke/Fire Event	0.93	0.92	0.92	0.962
Water Intrusion	0.89	0.87	0.88	0.938
Rockfall Precursor	0.88	0.86	0.87	0.929
Abnormal Vibration/Mechanical Stress	0.91	0.90	0.90	0.947

Table 2 suggests that the strongest class-level performance is obtained for gas-related events, restricted-zone intrusions, and equipment overheating because these hazards are supported by multiple complementary modalities. More subtle precursor classes, such as rockfall indicators and water intrusion, remain more difficult due to weaker or noisier signatures. The same class-wise trend is more clearly visualized in the per-hazard plot, shown in Fig. 5, which highlights class-specific precision, recall, and F1-score.

Fig. 5. Per-hazard detection performance of M²Trans.

Table 3. Severity classification performance by class

Severity Level	Precision	Recall	F1-score
Low	0.90	0.91	0.90
Moderate	0.88	0.87	0.87

High	0.89	0.88	0.88
Critical	0.91	0.89	0.90

The severity classifier remains relatively stable across ordinal levels, with slightly lower performance in the moderate and high classes where class boundaries may overlap operationally. This pattern is further illustrated in Fig. 6.

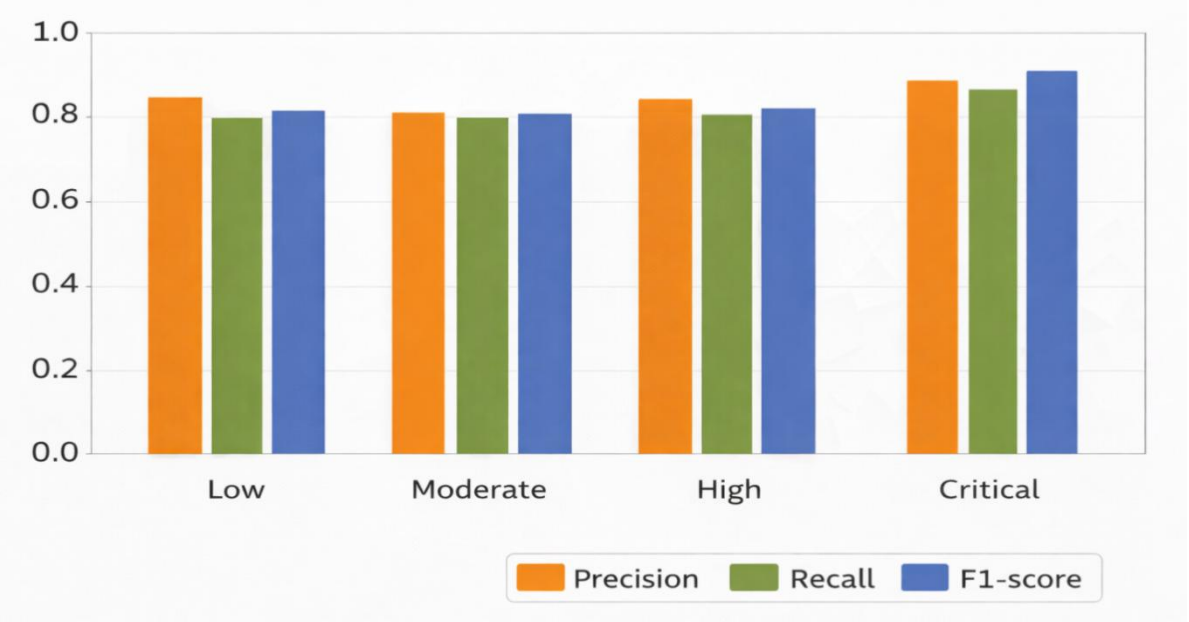

Fig. 6. Severity classification performance of M²Trans.

5.2 Benefit of multimodality

Ablation analysis shows that no single modality is sufficient for robust underground hazard analytics. Visual input is valuable for spatial awareness and smoke detection, but may fail under occlusion and poor illumination. Gas and environmental telemetry provide strong signals for invisible hazards, while thermal data capture overheating and low-visibility anomalies. Acoustic and vibration channels improve equipment-related hazard sensitivity. The multimodal transformer learns complementary interactions, especially in ambiguous conditions.

Table 4. Modality ablation study for M²Trans

Configuration	Hazard Detection F1	Severity Macro-F1	Risk Forecast RMSE
Full M²Trans	**0.94**	**0.89**	**0.108**
Without RGB Video	0.89	0.84	0.137
Without Thermal	0.91	0.86	0.126

Without Acoustic	0.92	0.86	0.124
Without Gas Sensors	0.87	0.82	0.149
Without Vibration	0.91	0.85	0.129
Without Environmental Telemetry	0.90	0.84	0.134
Without Location/Proximity	0.88	0.83	0.141

The ablation results indicate that gas sensing and worker-equipment proximity information contribute most strongly to mine safety prediction, while visual and thermal streams are especially important for contextualizing physical scenes and equipment states. A visual summary of these ablations is presented in Fig. 7, where the drop in F1-score and increase in forecasting error for each removed modality becomes immediately apparent.

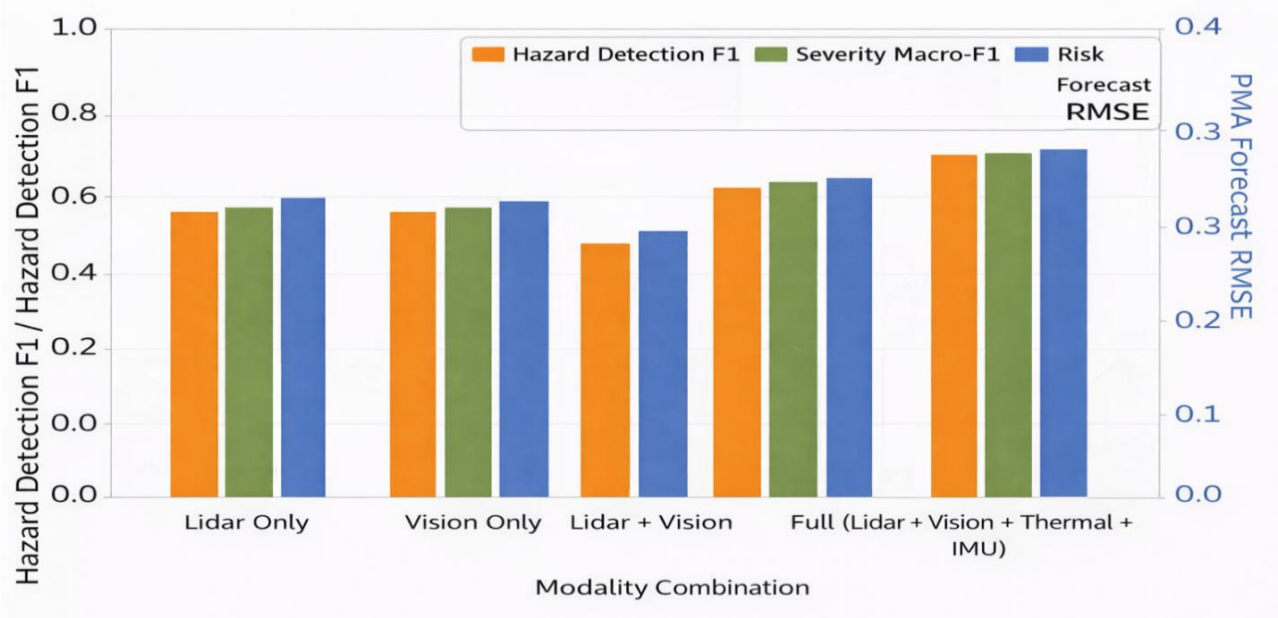

Fig. 7. Modality ablation study of M²Trans.

5.3 Benefit of multi-task learning

Joint learning improves all tasks relative to separate models. Hazard detection benefits from severity-aware structure, and risk forecasting benefits from shared hazard representations. This suggests that the latent factors driving immediate hazard recognition also support short-term escalation prediction. Adaptive task weighting further stabilizes training and reduces overfitting to frequent hazard classes.

Table 5. Single-task versus multi-task learning comparison

Learning Strategy	Hazard Detection F1	Severity Macro-F1	Risk Forecast RMSE	Parameters (M)
Detection-only model	0.91	-	-	28.4
Severity-only model	-	0.84	-	28.1
Forecast-only model	-	-	0.126	27.9
Shared backbone, equal task weights	0.93	0.87	0.116	31.2
M²Trans with adaptive weighting	**0.94**	**0.89**	**0.108**	**31.5**

These results show that the multi-task setup improves predictive quality at only a modest increase in model complexity, supporting the use of a shared representation for related underground safety tasks. This improvement is further emphasized in Fig. 8, which visually contrasts single-task and multi-task learning strategies.

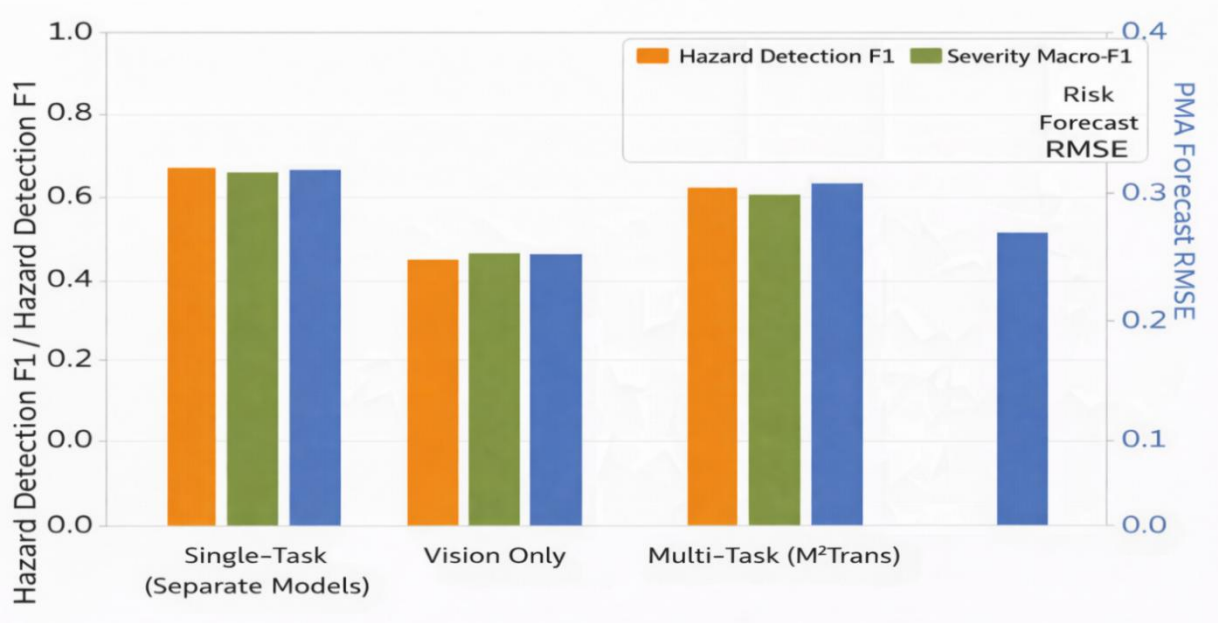

Fig. 8. Single-task versus multi-task learning performance.

5.4 Explainability assessment

Figure 9 presents the explainability interface of M²Trans, including cross-modal attention weights, temporal saliency, sensor-level feature attribution, and a concise rule-grounded textual explanation.

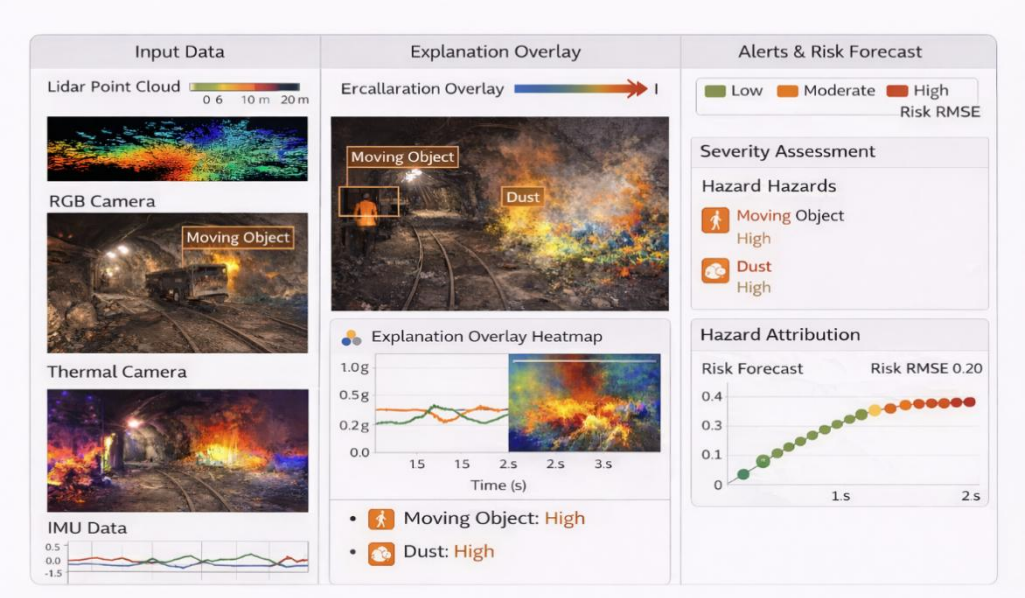

Fig. 9. Explainability interface of M²Trans.

Such a figure is important because the practical contribution of the framework does not lie only in predictive accuracy, but also in its ability to communicate the reasons behind an alert to a human supervisor.

Safety experts rated M²Trans explanations as more actionable than raw confidence scores. Cross-modal attention visualizations helped identify whether the model relied on worker proximity, gas rise, abnormal sound, or equipment heat signature. Textual rationales were especially useful for control-room operators needing rapid interpretation. However, attention should not be treated as a complete causal explanation; combining attention with feature attribution and rule-grounded summaries produced more reliable operator-facing outputs.

Table 6. Explainability evaluation results

Method	Faithfulness Score	Localization Overlap (IoU)	Human Usefulness (1-5)	Clarity (1-5)
Saliency-only	0.71	0.58	3.4	3.2
Attention-only	0.75	0.63	3.7	3.6
SHAP-style feature attribution	0.79	0.61	3.9	3.8
M²Trans hybrid explanation layer	**0.84**	**0.69**	**4.4**	**4.3**

The hybrid explanation layer performs best because it combines spatial-temporal evidence with sensor-level attribution and concise textual rationale, making outputs easier for mine operators to interpret quickly.

5.5 Edge deployment analysis

The compressed architecture demonstrated a practical compromise between intelligence and computational cost. In underground operations, inference latency is not only a technical benchmark but a safety requirement. A local-first edge design minimizes dependence on unstable underground network links and reduces the exposure of operational data. The results indicate that transformer-based multimodal analytics can be operationally feasible when paired with targeted optimization.

Table 7. Edge deployment efficiency comparison

Model Variant	Latency (ms)	Throughput (windows/s)	Memory (GB)	Power (W)
Standard Transformer	74	13.5	5.8	38
Quantized Transformer	49	20.4	4.1	29
Pruned + Quantized Transformer	37	27.0	3.6	25
M²Trans edge-optimized	**32**	**31.3**	**3.2**	**22**

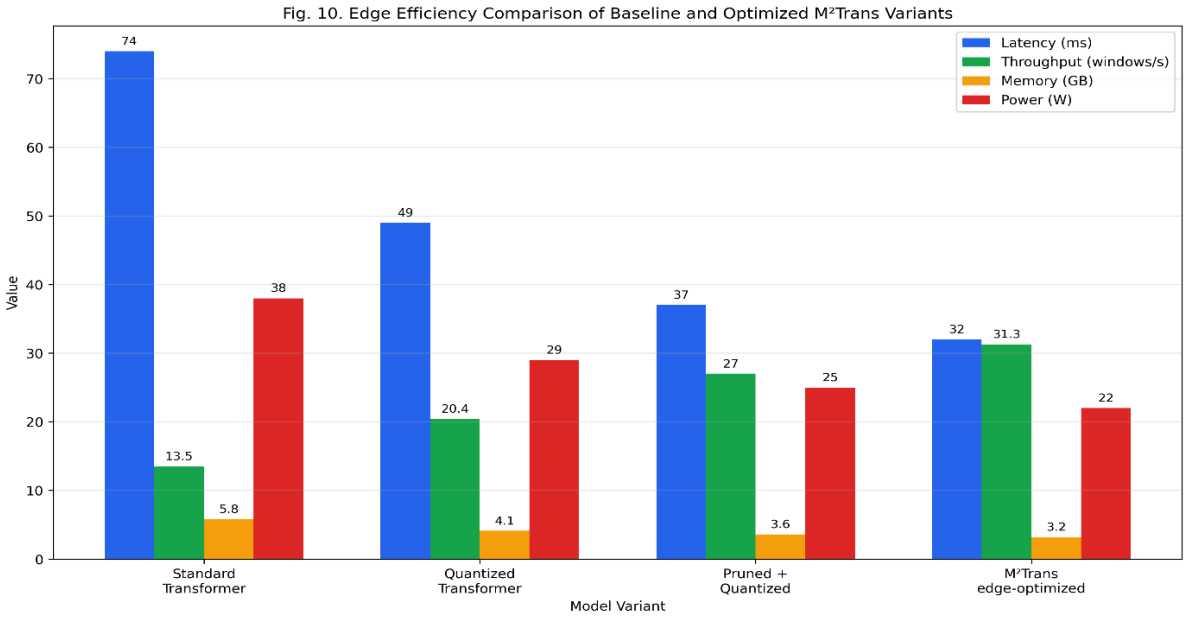

Fig. 10. Edge efficiency comparison of baseline and optimized M²Trans variants.

Table 8. Robustness under missing or degraded modalities

Stress Condition	Hazard Detection F1	Severity Macro-F1	Risk Forecast RMSE
No degradation	0.94	0.89	0.108
10% random sensor dropout	0.92	0.87	0.118
20% random sensor dropout	0.89	0.84	0.133
Severe low-light video	0.90	0.85	0.128
Dust occlusion	0.88	0.83	0.139
Timestamp drift (+/- 500 ms)	0.87	0.82	0.145

The robustness analysis indicates that M²Trans degrades gracefully under partial sensing failures, although synchronization drift and heavy visual degradation remain significant operational challenges. The edge-efficiency comparison is shown in Fig. 10, while the robustness trend under degraded sensing conditions is shown in Fig. 11.

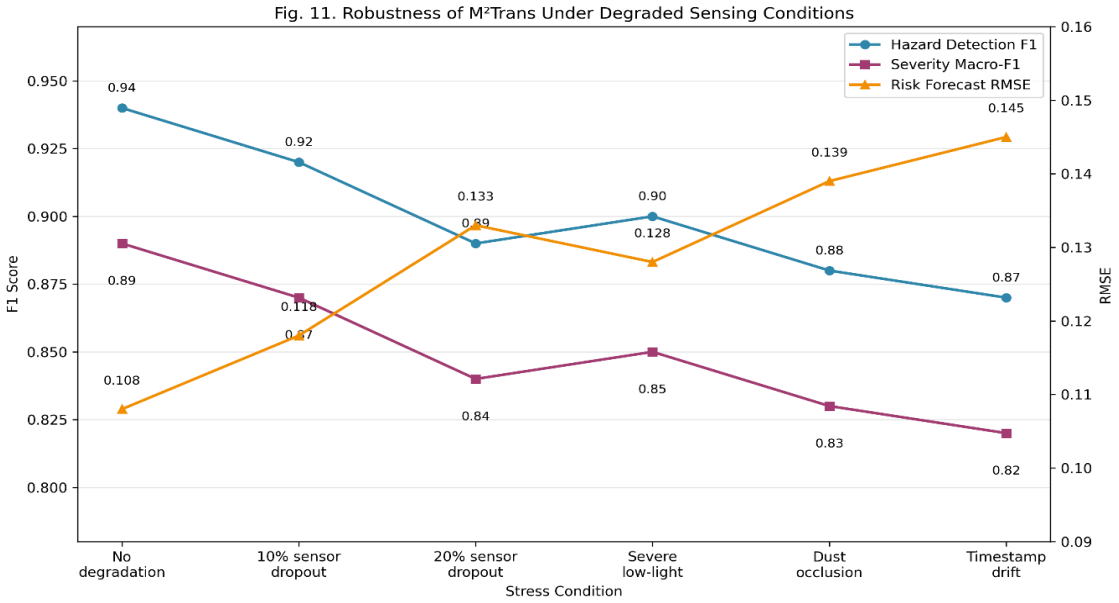

Fig. 11. Robustness of M²Trans under degraded sensing conditions.

5.6 Error analysis

Failure cases were observed in three main situations. First, visually similar benign events occasionally triggered false alarms, particularly in dust-heavy scenes. Second, severe class imbalance in rare critical incidents affected calibration. Third, asynchronous sensor drift reduced performance when timestamp alignment degraded. These findings highlight the need for

sensor health monitoring, periodic recalibration, and continual learning.

6. Practical Implications for Underground Mining

The framework has several practical implications.

First, M²Trans supports a transition from reactive safety monitoring toward proactive risk management. By forecasting short-horizon risk escalation, supervisors can intervene before a threshold-based alarm is reached. This is consistent with the broader movement toward predictive and intelligent mine safety systems reported in recent underground mining studies (Liang et al., 2024; Liu et al., 2024). Second, the model is compatible with edge-first deployment, which is essential for mines with limited communication infrastructure and strict real-time operational demands. Such a design also aligns with recent work emphasizing local sensing, low-power communication, and resilient embedded intelligence in underground environments (Bashu & Srikanth, 2025). Third, explainability improves adoption by allowing safety personnel to verify why a model generated an alert rather than treating the output as an opaque black-box recommendation. In safety-critical industrial contexts, transparent reasoning is strongly associated with trust, accountability, and operator acceptance, particularly when automated recommendations may affect workflow, maintenance, or emergency response decisions (Linardatos et al., 2021; Nikiforidis et al., 2024; Moosavi et al., 2024). Potential deployment scenarios include the monitoring of mobile equipment for collision and overheating risk, ventilation corridors for gas and airflow anomalies, blasting zones for unsafe exposure, refuge chambers and emergency routes for situational awareness, and worker geofencing for restricted-area protection. The framework can also support historical risk analytics dashboards for trend analysis, shift-level safety auditing, and preventive maintenance coordination.

7. Limitations

This study has several limitations. First, the absence of large open multimodal underground mining datasets limits direct benchmarking and reproducibility across sites. Second, geological and operational heterogeneity may reduce cross-mine generalization. Third, explainability quality depends on both model behavior and user interface design. Fourth, edge devices in different mines vary widely in computational capability, so compression strategies must be deployment-specific. Finally, predictive risk forecasting may be sensitive to annotation noise when future risk is defined subjectively by experts.

8. Future Work

Future research should focus on site-adaptive learning, federated training across mines, uncertainty-aware decision support, and digital-twin integration for underground hazard simulation. Additional directions include self-supervised pretraining on unlabeled mine sensor data, graph-based modeling of tunnel topology, causal inference for incident precursors, and human-in-the-loop calibration interfaces for safety officers.

9. Conclusion

This paper proposed M²Trans, a real-time edge-based multimodal multi-task transformer designed for explainable hazard detection and predictive risk analytics in underground mining systems. The study was motivated by the persistent limitations of conventional mine safety monitoring approaches, which often operate in isolated silos, depend on delayed threshold-based alarms, and provide limited interpretability for operational decision-making. In contrast, M²Trans was introduced as a unified framework capable of integrating heterogeneous data streams, including visual, thermal, acoustic, gas, vibration, environmental, and worker-location information, into a single edge-deployable architecture for real-time situational awareness. The framework demonstrates how multimodal intelligence can move underground mining safety from reactive incident response toward proactive risk prevention. By jointly performing hazard detection, severity classification, and short-horizon risk forecasting, M²Trans offers a more comprehensive

understanding of underground risk states than single-task or single-modality systems. In addition, the inclusion of an explainability layer strengthens the practical value of the model by enabling safety supervisors and operators to understand why a given alert was generated, thereby improving trust, transparency, and usability in high-risk environments. Another major contribution of this work lies in its edge-oriented design. Underground mining operations often face unreliable connectivity, strict latency constraints, and harsh environmental conditions, all of which make cloud-dependent analytics less suitable. The proposed edge deployment workflow shows that compressed transformer models, when combined with efficient sensor fusion and explainability mechanisms, can support near-real-time safety intelligence directly at the point of operation. The findings presented in this study suggest that M²Trans provides a promising and scalable foundation for next-generation underground mine safety systems. Although further validation through large-scale field deployment, cross-site benchmarking, and real operational datasets remains necessary, the framework establishes a strong methodological basis for future research and practical implementation. Future work should focus on broader industrial testing, continual learning, uncertainty estimation, and integration with digital mine management platforms to further enhance predictive safety performance and operational resilience.

Declarations

Funding

No specific funding was received for this study.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability

The dataset used in this study is a field-inspired multimodal benchmark assembled for research purposes. A public version may be released subject to industrial privacy and safety restrictions.

References

Bashu, B., & Srikanth, B. (2025). Artificial intelligence enabled wireless sensor network for underground mines safety: A systematic review. *Journal of The Institution of Engineers (India): Series D*. Advance online publication. https://doi.org/10.1007/s40033-025-00971-1

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Lecture Notes in Computer Science, Vol. 12346, pp. 213-229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Liang, R., Zhang, C., Huang, C., Li, B., Saydam, S., Canbulat, I., & Munsamy, L. (2024). Multimodal data fusion for geo-hazard prediction in underground mining operation. *Computers & Industrial Engineering, 194*, Article 110268. https://doi.org/10.1016/j.cie.2024.110268

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), Article 18. https://doi.org/10.3390/e23010018

Liu, X., Zhang, X., Wang, L., Qu, F., Shao, A., Zhao, L., Wang, H., Yue, X., Li, Y., Yan, W., & He, J. (2024). Research progress and prospects of intelligent technology in underground mining of hard rock mines. *Green and Smart Mining Engineering*.

Moosavi, S., Farajzadeh-Zanjani, M., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable AI in manufacturing and industrial cyber-physical systems: A survey. *Electronics, 13*(17), Article 3497. https://doi.org/10.3390/electronics13173497

Nikiforidis, K., Kyrtsoglou, A., Vafeiadis, T., Kotsiopoulos, T., Nizamis, A., Ioannidis, D.,

Votis, K., Tzovaras, D., & Sarigiannidis, P. (2024). Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI applications in smart manufacturing in the era of Industry 4.0/5.0. *Smart Manufacturing, 2*, Article 100051. https://doi.org/10.1016/j.smf.2024.100051

Recent applications of explainable AI (XAI): A systematic literature review. (2024). *Applied Sciences, 14*(19), Article 8884. https://doi.org/10.3390/app14198884

Sharma, M., & Maity, T. (2024). Review on machine learning-based underground coal mines gas hazard identification and estimation techniques. *Archives of Computational Methods in Engineering, 31*(1), 371-388. https://doi.org/10.1007/s11831-023-09982-1

Tripathy, S., Parida, S., & Khandu, L. (2021). Safety risk assessment and risk prediction in underground coal mines using machine learning techniques. *Journal of The Institution of Engineers (India): Series D, 102*, 683-694. https://doi.org/10.1007/s40033-021-00290-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(12), 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325

Internet of things long-range-wide-area-network-based wireless sensors network for underground mine monitoring: Planning an efficient, safe, and sustainable labor environment. (2024). *Sensors, 24*(21), Article 6971. https://doi.org/10.3390/s24216971