https://doi.org/10.37420/j.jime.2025.003

Research on Multimodal Large-Model-Driven Mining Safety Early Warning Mechanism

Changwen Wu*

Wuhan Yishikong Technology Co., Ltd., Wuhan, Hubei 430000, China

***Corresponding author, E-mail: woochangwen@163.com**

Abstract:

Mining safety production is a high-risk sector within the industrial system, facing severe challenges such as complex geological conditions, dynamic environmental changes, and intertwined human operation risks. Traditional single-modal monitoring technologies exhibit significant limitations in real-time perception and intelligent decision-making, failing to meet the demands of modern mining safety management. To address these issues, this paper proposes a multimodal large-model-driven mining safety early warning mechanism. By leveraging multimodal data integration and cross-modal learning, the mechanism achieves comprehensive perception and accurate early warnings for mining operations. The research encompasses multimodal data collection and processing, large-model design and implementation, as well as the construction and application of the safety warning mechanism. This approach significantly enhances the intelligence level of mining safety management and accident prevention, providing essential technical support for the development of smart mining.

Keywords:

Mining Safety; Multimodal Technology; Large Model; Safety Early Warning; Smart Mining

1 INTRODUCTION

Mine safety has consistently been a critical area of focus within industrial production. As mining operations expand in scale and working environments grow increasingly complex, frequent mining accidents pose a severe threat to the lives of personnel and production efficiency. Traditional mine safety early warning technologies primarily rely on single data sources and rule-driven systems. For instance, environmental parameters such as methane concentration, temperature, and humidity are monitored via sensors, with alarms triggered by predefined thresholds. Alternatively, statistical analysis methods based on historical data are employed to predict potential hazardous situations. However, these approaches exhibit significant limitations. They suffer from narrow data sources that fail to comprehensively reflect the complex mining environment, lack flexibility in rule-based systems to address sudden and diverse safety risks, and insufficient capability in traditional models to handle non-linear data relationships and multimodal information. Consequently, the accuracy and timeliness of early warnings remain suboptimal. In recent years, the rapid advancement of artificial intelligence has seen machine learning and deep learning technologies progressively integrated into mine safety early warning systems. For instance, time series analysis techniques are employed to predict equipment failures, while image recognition technologies monitor operational environments within mining areas. Nevertheless, these approaches remain predominantly reliant on unimodal data, struggling to fully leverage the multimodal data resources inherent in mining operations and thus failing to comprehensively reflect the intricate characteristics of complex working environments. Consequently, harnessing multimodal

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

technologies to enhance the efficacy of mine safety early warning systems has emerged as a critical research focus.

2 CURRENT SITUATION ANALYSIS

In recent years, large language models have emerged as a significant breakthrough in artificial intelligence, demonstrating formidable potential particularly in natural language processing and computer vision applications. Models such as GPT (Generative Pre-trained Transformer) for natural language processing and visual Transformers for computer vision possess robust feature extraction capabilities, cross-modal learning abilities, as well as versatility and transferability. Through training on vast datasets, these models can extract deep-level features from multimodal data and adapt to specific domain tasks with minimal fine-tuning. Within industrial safety, large models are primarily applied in fault detection, risk assessment, and intelligent monitoring. However, in the field of mining safety, their application remains in the early exploratory stages, with no mature solutions yet established. Multimodal technology achieves more comprehensive and precise analysis by integrating information from diverse data sources or modalities (such as text, images, audio, and video). Its core lies in data fusion and complementarity, where different modalities provide multidimensional insights into the same event, thereby overcoming the limitations of single-modality data. Multimodal technology has demonstrated significant application potential across numerous domains. Within mine safety, the introduction of multimodal technology can effectively integrate environmental monitoring data (such as gas concentration, temperature, humidity), video surveillance footage, textual records (like operational logs), and other data sources, thereby providing more comprehensive support for safety alerts. Despite achievements in industrial safety, applying multimodal technology and large models to mine safety presents several challenges. Firstly, multimodal data in mining environments exhibits heterogeneity and spatio-temporal inconsistency, making effective fusion a critical issue. Secondly, mine safety alerts demand real-time responsiveness, whereas current large models' computational complexity may compromise system timeliness. Furthermore, existing large models are predominantly general-purpose systems, necessitating further research into tailored designs addressing the specific requirements of mine safety. Consequently, developing a mine safety early warning mechanism based on multimodal large models—by integrating multimodal data and leveraging the cross-modal learning capabilities of large models—will enhance the accuracy and real-time responsiveness of mine safety alerts, thereby providing more reliable safety assurance for mining operations.

3 KEY TECHNOLOGY DESIGN AND IMPLEMENTATION

This study proposes a mine safety early warning mechanism based on multimodal large language models. By integrating multimodal data fusion with the cross-modal learning capabilities of large models, it achieves comprehensive perception and efficient early warning of mine operational environments. The mechanism’s design encompasses multimodal data acquisition and processing, large model design and implementation, and the construction of a model-based safety early warning system.

3.1 Multimodal Data Acquisition and Fusion

The mining environment is complex and dynamic, with data originating from diverse sources including environmental monitoring, video surveillance, and textual records. These data exhibit modal diversity and spatio-temporal heterogeneity, making their effective collection and processing fundamental to designing early warning mechanisms.

Firstly, during the data collection phase, environmental monitoring data is primarily obtained through sensor networks deployed across the mining area, capturing parameters such as methane concentration, temperature,

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

humidity, and wind speed. Video surveillance data is gathered in real-time via cameras positioned throughout the mine, documenting worker behaviour and equipment operational status. Textual data originates from mining operation logs, equipment maintenance records, and historical accident reports. This multimodal data reflects the safety conditions of mining operations from multiple dimensions.

Secondly, during the data processing stage, a unified preprocessing workflow addresses the heterogeneity of multimodal data. This includes operations such as data cleansing, format conversion, and temporal alignment. During data cleansing, outliers and noise in sensor data are removed, while invalid frames in video data are discarded. Format conversion standardises disparate data formats into a unified input format. Temporal alignment synchronises multimodal data via timestamps, ensuring spatio-temporal consistency. Furthermore, to enhance data usability and information content, feature engineering techniques are employed for feature extraction and fusion across multimodal data.

The multimodal data fusion mechanism constructs a multi-level collaborative modelling framework by integrating heterogeneous modal information such as images, text, and audio. Its core lies in cross-modal feature alignment and dynamic semantic complementarity. This mechanism is typically organised across three levels: data-level and feature-level. Data-level fusion directly processes raw data, such as aligning LiDAR point clouds with camera images through calibration techniques to resolve spatio-temporal heterogeneity; Feature-level fusion employs deep learning models to extract high-dimensional abstract features from each modality, utilising attention mechanisms or graph neural networks to enable cross-modal interaction. Spatial features from video data are extracted via convolutional neural networks, semantic information from textual data is derived through natural language processing techniques, and these are combined with the temporal characteristics of sensor data to form multimodal feature vectors, thereby achieving multimodal data fusion.

3.2 Design and Implementation of Large Models

The large model design is grounded in the core principles of multimodal fusion and cross-modal learning, employing the Transformer architecture as its foundational framework. The Transformer has gained widespread application in natural language processing and computer vision due to its formidable capabilities in feature extraction and cross-modal learning.

In the model architecture design, a multimodal feature encoder was first constructed to address the characteristics of multimodal data. This encoder processes environmental monitoring data, video data, and text data independently, extracting high-level features from each modality.

During the feature fusion stage, a Cross-Modal Attention Mechanism is employed to integrate multimodal features. This mechanism enables complementary information exchange between modalities, yielding more expressive joint feature representations. Furthermore, to enhance the model's generalisation capability and robustness, a self-supervised learning strategy is adopted for model pre-training, leveraging unlabelled data to uncover latent intra- and inter-modal associations.

During model training and optimisation, a multi-task learning framework divides the mine safety early warning task into three sub-tasks: hazard factor identification, risk prediction, and early warning information generation. Shared model parameters enable joint optimisation across these sub-tasks. Concurrently, to address data distribution imbalances in mining environments, data augmentation techniques and weighted loss function strategies are employed to enhance the model's recognition capability for minority class samples.

3.3 Implementation of the Safety Early Warning Mechanism

Leveraging the multimodal learning capabilities of large language models, a mine safety early warning mechanism has been established to enable real-time hazard identification, risk prediction, and the generation of alert notifications.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

During the hazard identification phase, the model analyses features across multimodal data streams to detect potential risk factors that could precipitate safety incidents. For instance, video data is scrutinised to identify instances of non-compliant operator behaviour, sensor readings are examined to detect environmental parameters exceeding safety thresholds, and textual data is mined to uncover latent risk points within historical records.

During the risk prediction phase, the model combines historical and real-time data, employing time series analysis and deep learning techniques to forecast risk trends within the mining environment. Examples include predicting methane concentration trajectories to assess explosion risks, and analysing equipment operational states to anticipate potential failures.

During the early warning generation phase, the model produces specific alerts based on identification and prediction outcomes, presenting these via a visualised interface to mine management personnel. For instance, upon detecting an abnormal rise in methane concentration, the system generates an alert stating: 'Methane concentration exceeds permissible limits; personnel should evacuate immediately.' When identifying non-compliant operations by personnel, the system issues a prompt: 'Worker not wearing safety helmet; rectify immediately.'

4 APPLICATION SCENARIO ANALYSIS

The multi-modal large-model-based mine safety early warning mechanism demonstrates significant practical value across multiple operational scenarios within mining environments. By deeply integrating multi-modal information such as sensor data, video surveillance footage, and textual records, this mechanism enables efficient risk perception and early warning within complex, dynamic mining conditions. Below, we conduct a detailed analysis of this early warning mechanism's application within typical mining scenarios, demonstrating its potential to enhance mine safety management efficiency and accident prevention capabilities.

4.1 Early Warning System for Coal Mine Gas Leakage and Explosion Risks

Gas leakage represents one of the most critical risks in mine safety management, particularly during coal mining operations. Abnormal increases in methane concentrations can directly endanger the lives of personnel and result in significant economic losses. Traditional gas monitoring methods primarily rely on real-time monitoring via single sensors, which, while providing basic alarm functions, lack the comprehensive analytical capability to assess the interplay of multiple factors within complex environments.

An early warning mechanism based on multimodal large models achieves precise identification and prediction of gas leakage risks by integrating data from methane concentration sensors, video surveillance, and historical accident records. For instance, when sensors detect abnormal concentration spikes, the system analyses video footage to assess ventilation conditions, equipment operation, and worker distribution within the mine, thereby determining potential escalation risks. Simultaneously, the system extracts analogous cases from historical accident records and generates tailored warning messages based on current environmental characteristics. For instance, during elevated methane concentrations coupled with abnormal ventilation equipment operation, the system issues alerts such as: 'Methane concentration exceeds permissible limits with inadequate ventilation; immediate personnel evacuation recommended.' This provides mine management with timely decision-making support.

4.2 Monitoring of Operational Personnel Behaviour Violations

Worker non-compliance constitutes a significant contributing factor to mining safety incidents.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

Conventional behavioural monitoring methods typically rely on manual inspections or rudimentary video surveillance, proving inadequate for achieving comprehensive coverage and real-time analysis across large-scale operational areas. In contrast, an early warning mechanism based on multimodal large language models can automatically identify worker violations and issue alerts through real-time analysis of video data, integrated with environmental sensor readings and operational log text data.

In practical application, the system can utilise video surveillance data to capture in real time whether personnel are wearing safety helmets, entering hazardous zones, or engaging in other non-compliant operations. Upon detecting abnormal behaviour, the system integrates environmental monitoring data to assess the current operational hazard level and generates corresponding warning messages. For instance, should a worker be observed without a safety helmet while approaching a high-risk zone, the system will generate an alert stating: 'Worker not wearing safety helmet and approaching hazardous area. Correct immediately.' This notification is disseminated via voice broadcast or SMS to relevant supervisors, thereby effectively mitigating safety incidents arising from non-compliance.

4.3 Equipment Operational Status Monitoring and Fault Prediction

The operational status of mining equipment directly impacts production efficiency and operational safety. Equipment failures may not only cause production interruptions but also trigger severe secondary incidents. Traditional equipment monitoring methods typically rely on single-sensor data, lacking comprehensive analytical capabilities regarding equipment operational status. The early warning mechanism developed in this study achieves multidimensional monitoring and fault prediction by integrating equipment operational sensor data, video data, and maintenance log text data.

For instance, when sensors detect abnormal vibrations or elevated temperatures, the system can analyse visual data to assess external changes (such as oil leaks or unusual shaking) alongside historical fault records in maintenance logs to determine potential failure risks. Should the system detect multiple overlapping fault indicators—such as abnormal vibration, excessive temperature, and maintenance records indicating recent similar issues—it generates an alert stating: 'Equipment operation abnormal, potential severe fault risk present. Immediate inspection required.' This enables management to take proactive measures, thereby preventing safety incidents caused by equipment failure.

4.4 Prediction and Early Warning of Mine Collapse Risks

Mine collapses represent another prevalent major safety hazard in mining operations, often arising from multiple factors including geological conditions, equipment vibrations, and operational practices. Conventional collapse risk prediction methods typically rely on geological sensor data, lacking the capacity for comprehensive analysis of other pertinent factors. The early warning mechanism developed in this study achieves multidimensional prediction and alerting of collapse risks by integrating geological sensor data, video surveillance footage, and textual operational log data. In practical application, when geological sensors detect abnormal increases in rock pressure, the system analyses equipment operational status and worker distribution within the mine using video data. It then evaluates historical records from operational logs to assess whether other collapse triggers exist under current working conditions. For instance, upon detecting both elevated rock pressure and intense equipment vibration, the system generates an alert stating: 'Abnormally elevated rock pressure coupled with severe equipment vibration indicates collapse risk. Cease operations immediately.' This warning is presented via a visual interface to mine management personnel, thereby providing timely evacuation guidance for workers.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

4.5 Comprehensive Application Value

Analysis of the aforementioned application scenarios demonstrates that the multimodal large-model-based mine safety early warning mechanism holds significant practical value across multiple critical stages of mining operations. Its core advantage lies in comprehensively perceiving the safety conditions of the mining environment, accurately predicting potential risks, and generating specific warning information through the deep integration and intelligent analysis of multimodal data. This not only substantially enhances the efficiency of mine safety management but also effectively reduces casualties and economic losses resulting from safety incidents. Moving forward, the widespread adoption of this mechanism within mining operations will provide robust technical support for the safe production of mining enterprises. Simultaneously, it offers valuable reference points for safety management across other industrial sectors.

5 TECHNICAL CHALLENGES AND OUTLOOK

The design and implementation of a mine safety early warning mechanism based on multimodal large language models has demonstrated its potential and practical value within complex industrial settings. Nevertheless, as the system undergoes progressive development and experimental validation, several issues and challenges warrant further examination. This chapter will explore its technical advantages, limitations in real-world applications, potential avenues for improvement, and future development trends, thereby providing reference for subsequent research and practical implementation.

5.1 Analysis of Technical Advantages

The multimodal large-model-driven mine safety early warning mechanism proposed in this study possesses the following significant advantages:

1) Multimodal data fusion capability

By integrating environmental sensor data, video surveillance footage, textual records and other multimodal information, the early warning mechanism comprehensively perceives dynamic changes within the mining operational environment. Multimodal data fusion not only enhances the ability to supplement deficiencies in single data sources but also identifies potential risk correlations through cross-modal interactive analysis.

2) Cross-modal learning and feature extraction

Large models based on the Transformer architecture efficiently extract features from diverse modalities and achieve deep feature fusion through cross-modal attention mechanisms. This approach effectively overcomes the limitations of traditional single-modal feature extraction and fusion, enabling the model to demonstrate greater robustness and generalisation capabilities in complex scenarios.

5.2 Limitations in practical application

Although this mechanism demonstrated favourable performance in experimental validation, it still faces certain challenges and limitations in practical application:

1) Data Quality and Multimodal Data Imbalance

Data quality in mining environments is frequently compromised by sensor failures, network latency, and harsh environmental conditions, resulting in missing or noisy data. Furthermore, variations exist in the sampling frequency and coverage of different modal data types—for instance, sensor data may be sampled at high frequency while textual records are updated infrequently. Achieving efficient multimodal fusion under such imbalanced data conditions remains a significant challenge.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

2) Computational Complexity and Resource Requirements

Training and inference of multimodal large models demand substantial computational resources, particularly when processing high-resolution video data and extensive historical text records, necessitating high hardware performance. Resource-constrained mining enterprises may face significant cost pressures in deploying and maintaining such complex models.

3) Scenario Adaptability and Robustness

Mining environments are complex and variable, with significant differences in geological conditions, equipment types, and operational workflows across distinct mining areas. Although multimodal large models possess a degree of generalisation capability, their robustness requires further enhancement when confronting extreme scenarios within specific contexts—such as complete sensor data loss or video surveillance blind spots.

5.3 Future Development Trends

Research and application of mine safety early warning mechanisms based on multimodal large models remain in their infancy. Future development trends are primarily reflected in the following aspects:

1) Intelligence and Automation

With continuous advancements in artificial intelligence technology, mine safety early warning systems will progressively transition from passive monitoring to active intervention. For instance, upon detecting high-risk events, the system can automatically trigger emergency response plans or control equipment operations, thereby further enhancing safety management efficiency.

2) Integration with IoT and Blockchain Technologies

By integrating with IoT technologies, systems can achieve more efficient data collection and equipment coordination. The incorporation of blockchain technology enhances data credibility and security, providing safeguards for cross-organisational collaboration.

3) Continuous Dynamic Optimisation

As mining operational environments and safety management requirements evolve, systems must possess capabilities for continuous learning and dynamic optimisation. For instance, online learning techniques can be employed to update model parameters in real time, ensuring the system maintains high predictive accuracy and adaptability.

6 CONCLUSION

Mine safety management constitutes a pivotal component in safeguarding the lives of mining personnel and ensuring uninterrupted production operations. Conventional single-modal monitoring approaches frequently struggle to achieve comprehensive perception and precise early warning when confronted with the complex and dynamic conditions inherent in mining environments. The multimodal large-model-based mine safety early warning mechanism leverages the advantages of artificial intelligence in data processing and risk prediction by integrating multimodal information such as sensor data, video surveillance footage, and textual records, thereby providing an innovative solution for mine safety management.

By integrating diverse data sources, this mechanism enables dynamic perception of mining environments within complex scenarios, addressing the limitations of single-modal approaches. Utilising a large model based on the Transformer architecture, it achieves deep fusion and analysis of multimodal features through cross-modal attention mechanisms, delivering efficient risk identification and decision support for safety management in intricate industrial settings.

Whilst this research has yielded significant outcomes, certain limitations persist, including uneven quality of

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

multimodal data, high computational complexity of the model, and room for improvement in scenario adaptability. Future research directions encompass: optimising data acquisition and processing workflows; developing lightweight models to reduce hardware requirements; enhancing system robustness and adaptive capabilities; and further improving user experience alongside the interpretability of warning information.

Through continuous technological innovation and engineering practice, multimodal large-model-based mine safety early warning mechanisms hold promise to play an increasingly significant role in mine safety management, providing robust safeguards for the safe production of mining enterprises and the health and safety of their personnel.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.

[2] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A review and classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[3] Zhang, Z. Q., Han, G., Xu, L., et al. (2021). Deep learning-based safety monitoring and early warning system for underground mines. Safety Science, 139, 105248.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[5] Zhou, B., & Pan, J. (2020). Intelligent Prediction of Coal Mine Risks Based on IoT and Machine Learning. Journal of Cleaner Production, 262, 121233.

[6] Li, Y., Wang, X., Liu, J., et al. (2022). Multimodal Fusion in Industrial Safety Monitoring: Challenges and Opportunities. IEEE Journal of IoT, 9(5), 3391–3406.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), 785–794.

[8] He, K. M., Zhang, X., Ren, S. Q., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[9] Zhu, J. G., & Wu, X. M. (2023). Applications of artificial intelligence-driven multimodal systems in high-risk industries: The case of mine safety. Industrial Safety and Risk Management, 45(3), 211–225.

[10] Wang Qiang, Li Ming, Zhang Xiaodong. (2021). Research on Mine Safety Monitoring Systems Based on Multimodal Data Fusion. Journal of Safety Science, 31(8), 15–22.

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]. International Conference on Learning Representations (ICLR), 2021.

[12] Wu, T., & Chen, Y. (2020). Key technologies and applications for intelligent development in mining[J]. Mining Research and Development, 40(5), 1–6.

© 2025 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative CommonsAttribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)