Journal of Zhejiang University-SCIENCE A
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Correspondence
https://doi.org/10.1631/jzus.A2500399

Utility-grade safety intelligence for thermal power: knowledge graph-augmented multimodal monitoring

Xinrong YAN[1,2], Chongbo ZHOU[1✉], Jiaxu QIAN[3], Zhengtao DING[1]

[1]*Huadian Electric Power Research Institute Co., Ltd., Hangzhou 310030, China*
[2]*College of Energy Engineering, Zhejiang University, Hangzhou 310027, China*
[3]*School of Computer Science, Peking University, Beijing 100871, China*

1 Introduction

Coal conveying and unloading areas in thermal power plants combine high dust concentration, vibration, noise, occlusion, and frequent changes in lighting and operating conditions. Workers, belts, rollers, motors, and sensors interact in a confined space, so hazards such as missing personal protective equipment, belt blockage, overheating, smoke, or unauthorized entry require rapid detection and contextual judgment. Traditional monitoring relies on manual patrols, camera feeds, and isolated alarms, which are useful but often fragmented. A video detector may identify a worker without a helmet, but it may not reason that the same worker is close to an overheated motor. Conversely, a sensor alarm may lack the visual context needed to judge personnel exposure.

Recent foundation and vision-language models provide strong visual and semantic priors (Radford et al., 2021; Bai et al., 2025). However, direct deployment in safety-critical industrial environments remains difficult because plant data are noisy, low-light, partially occluded, and highly site-specific (Wang and Chung, 2022). Existing vision-based detectors, including faster region-based convolutional neural network (Faster R-CNN) and You Only Look Once (YOLO) variants, have been used for personal protective equipment and anomaly detection (Girshick, 2015; Jiang et al., 2022; Lee et al., 2023), but many systems remain single-task pipelines. Multimodal learning offers a route to fuse visual, textual, and sensor evidence (Baltrusaitis et al., 2019; Kim et al., 2021), while knowledge graphs (KGs) can encode safety rules and support traceable reasoning (Bordes et al., 2013; Huang et al., 2025). Edge-cloud deployment further enables low-latency screening with high-capacity reasoning when needed (Iwanicki, 2018).

Despite these advances, three gaps remain. First, many industrial safety systems are vision-only or task-specific and cannot form a coherent multimodal situation view. Second, outputs are often insufficiently traceable for post-incident inspection and compliance review. Third, edge-cloud systems may allow an edge false negative to suppress downstream cloud reasoning, which is undesirable under glare, dust, or camera occlusion. We address these gaps with a KG-augmented multimodal framework that combines explicit safety rules, multimodal fusion, and hybrid triggering.

In this paper, "utility-grade safety intelligence" means a deployable monitoring system that simultaneously provides high recall for severe hazards, controlled false alarms to avoid unnecessary operational disruption, second-scale alert response, and evidence packages that are auditable after an incident. The main contributions are as follows: (1) a KG-augmented multimodal monitoring framework that aligns video, sensor, and log evidence with explicit safety rules; (2) an adaptation pipeline combining low-rank adaptation (LoRA), safety-aware proximal policy optimization (PPO), domain adaptation, semisupervised learning, and long-tail-aware distillation; (3) a two-tier edge-cloud deployment validated on a multiyear plant dataset and in live coal-conveyor operation.

✉ Chongbo ZHOU, chongbo-zhou@chder.com
Chongbo ZHOU, https://orcid.org/0000-0002-8433-9620

Received Oct. 24, 2025; Revision accepted Apr. 22, 2026;
Crosschecked July 20, 2026

© Zhejiang University Press 2026

2 Materials and methods

2.1 System architecture and hybrid triggering

The monitoring system follows a two-tier edge-cloud architecture (Fig. 1). Edge devices perform lightweight screening on incoming camera streams and local sensor signals. Routine frames are processed locally and discarded or logged, while suspicious event packages are escalated to the cloud for multimodal fusion and KG-guided reasoning. The cloud stage is not triggered solely by edge visual detection. Instead, it can be activated by three independent sources: (1) edge-detected visual anomalies, (2) sensor threshold violations such as abnormal temperature, vibration, or gas concentration, and (3) KG-derived rule triggers, such as a restricted-zone rule becoming active during a maintenance operation.

This hybrid trigger reduces the risk that poor visibility or domain shift makes the edge model a single point of failure. For cloud-side fusion, the system uses a trigger-centered sliding time window. Sensor readings within the window are aggregated into synchronized features, while key frames or short clips around the trigger timestamp are selected as visual input. The multimodal model then applies confidence-aware fusion: when visual evidence is degraded by dust or glare, sensor and rule evidence receive higher weight; when a sensor spike is isolated and unsupported by visual or KG evidence, its contribution is down-weighted to reduce false alarms.

2.2 Safety knowledge graph and explainable reasoning

A safety KG was constructed for coal conveying and unloading operations. Its ontology includes entities such as Person, Helmet, Harness, ConveyorBelt, Roller, Motor, DustSensor, GasSensor, and RestrictedZone, and relations such as “wears”, “isNear”, “monitors”, “hasComponent”, “indicates”, and “requires”. Candidate triples and rules were extracted from safety standards, plant operation manuals, incident reports, and expert interviews, followed by manual validation by safety engineers. Examples include (Person, mustWear, Helmet), (ConveyorBelt, hasComponent, Roller), and (GasSensor, indicates, HighGasLevel). Supplementary examples are provided in Table S1 of the electronic supplementary materials (ESM).

KG embeddings are produced with TransE (Bordes et al., 2013) and injected into the multimodal model as structured semantic tokens. TransE is computationally efficient and suitable for many directional and rule-centric relations used in this deployment, such as personal protective equipment requirements, proximity constraints, and sensor-to-risk mappings. Its limitations for symmetric, one-to-many, and highly compositional relations are discussed in Section 4 and will be addressed by more expressive graph-neural or rotational embedding methods in future work.

For each alert, the system records the activated observations, matched safety rules, and KG reasoning path that links evidence to the final hazard label and risk level. For example, if the vision module detects a worker without a helmet and the sensor stream reports abnormal motor temperature nearby, the explanation path may include Person→not_wearing→Helmet, Person→near→Motor, and Motor→status→Overheating. These links support the alert “personal protective equipment (PPE) violation near equipment anomaly” and provide evidence for inspection and post-event review (Fig. 2).

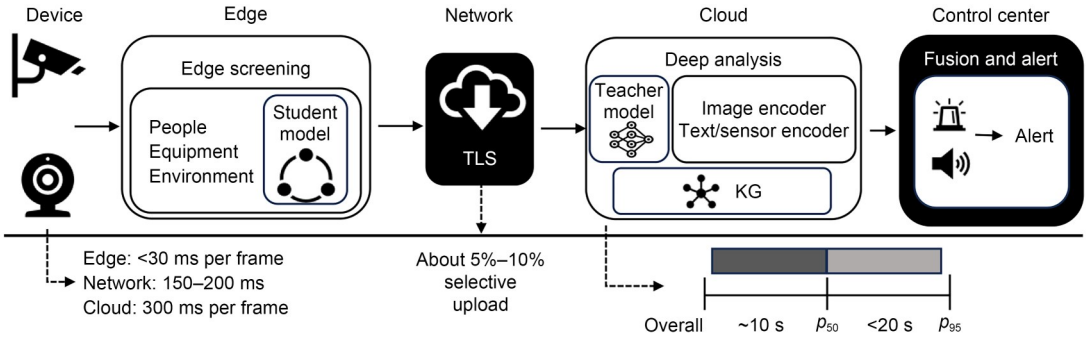


Fig. 1 Two-tier edge-cloud monitoring architecture with hybrid triggering and KG-guided cloud reasoning. TLS denotes the transport layer security; p_{50} and p_{95} denote the 50th percentile (median) and the 95th percentile of the end-to-end alert response time (ART), respectively

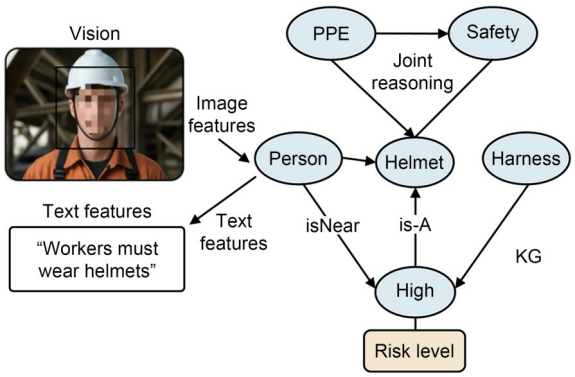


Fig. 2 KG excerpt for the person–PPE–safety domain and multimodal alignment

2.3 Multisource dataset and annotation

A multimodal dataset was assembled from coal feeding, conveying, and unloading operations during 2018–2024. The raw corpus includes more than 50000 h of fixed closed-circuit television (CCTV) video, extracted images and short clips, time-stamped sensor readings, maintenance logs, incident reports, and operator notes. After alignment, filtering, and preprocessing, the final dataset contained 235000 safety-relevant labeled samples (Table 1). Approximately 200000 samples were used for training, with the remainder used for validation and testing.

Each sample was annotated by three certified safety engineers in multiple passes. Labels cover human factors, equipment states, and environmental conditions, with severity levels such as warning, high, and severe. The dataset split is approximately 85% training, 5% validation, and 10% testing. Class imbalance is addressed by class-balanced sampling and loss reweighting, especially for rare but severe events.

Targeted augmentation simulated low-light, glare, dust, blur, and noise conditions calibrated against real plant footage. The augmentation ranges were kept conservative because overly aggressive degradation can suppress fine visual details and create unrealistic artifacts. Data augmentation and KG examples are summarized further in Section S1 of the ESM.

2.4 Model adaptation and long-tail protection

The base model is Qwen2.5-VL-72B (Bai et al., 2025), adapted for industrial monitoring through a KG-guided training pipeline. Detected visual entities are mapped to KG embeddings and serialized as semantic tokens so that the transformer can attend to visual, sensor, text, and rule-derived evidence. A lightweight mixture-of-experts module is used to expand task capacity efficiently, following the motivation of sparse expert models and multimodal mixture of experts (MoE) scaling (Fedus et al., 2022; Li et al., 2024).

LoRA is applied to the attention and feed-forward modules while keeping the backbone weights frozen (Hu et al., 2022). Instruction-style prompts and risk-level annotations supervise structured hazard descriptions. After supervised fine-tuning, safety-aware PPO is used to reflect the asymmetric cost of industrial safety (Schulman et al., 2017). The reward R is defined as

$$R=\alpha\cdot T_{\text{sev}}-\beta\cdot M_{\text{sev}}-\gamma\cdot F, \tag{1}$$

where T_{sev} denotes the number of correctly identified severe hazards, M_{sev} denotes the number of missed severe hazards, and F denotes the number of false alarms. The coefficient α specifies the reward assigned to each correctly identified severe hazard, thereby encouraging timely and accurate hazard detection. The coefficient β specifies the penalty for each missed severe hazard and is set substantially larger than γ, because a

Table 1 Summary of the curated multimodal safety dataset

Category	Approximate sample size	Description
Person-centric	108000	Helmet and harness status, worker activity, falling, climbing, and restricted-zone location
Equipment-centric	60000	Conveyor-belt jam, stuck roller, motor overheating, smoke or fire, and damaged safety facilities
Environment-centric	40000	Dust, visibility, material spillage, abnormal temperature, and other environmental conditions
Critical hazards	27000	High-severity samples such as no-helmet events, person falling, visible fire, and major equipment anomalies
Total labeled samples	235000	Aligned multimodal annotations, including image/video evidence, labels, and severity levels

missed severe hazard may lead to injury or shutdown-level consequences, whereas a false alarm mainly causes operational interruption. The coefficient γ specifies the penalty for each false alarm and remains non-trivial to prevent the monitoring system from generating excessive high-frequency alarms.

Domain adaptation uses a domain classifier with gradient reversal to encourage site-invariant features (Liu et al., 2019). High-confidence predictions above τ=0.95 on unlabeled plant imagery are used as pseudo-labels with perturbation consistency. Few-shot adaptation is supported by model-agnostic meta-learning (Arnold et al., 2021). The 72B teacher is distilled into an 8-bit 7B edge model with class-balanced sampling and hazard-weighted distillation so that rare but severe events are not suppressed by common normal samples. Additional training and distillation details are provided in Section S2 of the ESM.

2.5 Deployment and evaluation metrics

The cloud back end runs on Kunpeng CPUs (Xia et al., 2021) and Ascend AI accelerators (Liao et al., 2021) with MindSpore. Each edge unit uses a Huawei Ascend neural processing unit (NPU) module for 8-bit quantized inference and is connected to local cameras and sensors. Edge screening runs at millisecond-level latency per incoming frame. Full student-model analysis is executed only for triggered event packages, such as key frames or short clips, and typically completes within several seconds depending on clip length and batching conditions.

The evaluation uses the alert response time (ART), detection rate, miss rate, false alarm rate, and expert-verified precision. ART is defined as the interval from hazard onset to alert generation. Hazard onset was determined from synchronized sensor logs, video review, or manual observation. For implementation measurements, we distinguish frame-level edge screening throughput from full triggered-package processing time. Detailed latency decomposition is provided in Table S2 of the ESM.

3 Results

3.1 Field responsiveness

During a two-week field trial on the coal conveyor line, every triggered event was logged and reviewed. ART ranged from approximately 5 to 20 s depending on hazard complexity. Minor events such as unauthorized entry were typically detected within 5–10 s. Multistage faults such as conveyor-belt jams were detected during fault progression before full escalation.

The legacy monitoring procedure, based on human watchkeeping and basic sensor alarms, showed average ART values of approximately 25–30 s. In a controlled simulation of the legacy small-model pipeline, 13 independent detectors were run sequentially, resulting in approximately 30 s latency. Under comparable conditions, the proposed dual-tier system achieved an average ART of 9.97 s, corresponding to an approximately 67.1% faster response.

3.2 Hazard detection performance

Detection performance was evaluated on seven representative hazard categories covering personnel, equipment, and environmental risks. The proposed system outperformed the legacy baseline in all categories (Table 3). For severe human-related hazards, the miss rate was reduced to 1.3% for no-helmet violations and 4.8% for no-harness violations. For equipment faults, detection rates exceeded 95% for conveyor-belt jams and motor faults. False alarm rates remained between 1.6% and 4.0%, aided by multimodal conflict arbitration and KG rule constraints.

A random sample of 130 alerts from 3652 field alerts was reviewed by three senior safety engineers. Expert-confirmed precision reached 90%–100% for

Table 2 Comparison of ART and processing latency

System or case	ART/latency	Note
AI-enabled monitoring (field trial)	5–20 s	Two-week live deployment; latency depending on hazard complexity
Minor hazards	5–10 s	Unauthorized entry and similar single-stage events
Multistage faults	10–20 s	Conveyor jam progression detected before full escalation
Legacy monitoring	25–30 s	Human watchkeeping with basic sensor alarms
Legacy sequential models	About 30 s	13 specialized detectors executed serially
Proposed dual-tier system	9.97 s	Edge screening plus parallel cloud fusion and KG reasoning

Table 3 Detection performance across critical hazard categories

Hazard	Severity	Proposed system			Baseline	
		Detection rate	Miss rate	False alarm rate	Detection rate	Miss rate
No helmet	Severe	98.7%	1.3%	2.1%	76.4%	23.6%
No safety harness	Severe	95.2%	4.8%	1.9%	70.1%	29.9%
Conveyor-belt jam	High	96.8%	3.2%	3.5%	84.3%	15.7%
Motor fault	High	95.1%	4.9%	2.8%	85.6%	14.4%
Smoke/flame	Medium	91.7%	8.3%	3.1%	74.2%	25.8%
Unauthorized entry	High	94.3%	5.7%	1.6%	79.9%	20.1%
Environmental hazard	Medium	88.5%	11.5%	4.0%	71.0%	29.0%

severe “must-fix” hazards and above 95% for well-represented common hazards such as conveyor jams. For rare or broadly defined hazards with less training coverage, the precision was approximately 70%–80%. The expert consensus corresponded to an average F1 score of approximately 85.3%, consistent with the test-set results.

Compared with the legacy ensemble, the unified multimodal model also improves maintainability and few-shot adaptability. With 20 labeled examples of a novel hazard, the multimodal model achieved approximately 90% detection accuracy, whereas a new small convolutional neural network (CNN) trained from scratch achieved approximately 50%. At the system level, the proposed framework replaces 13 specialized detectors with one shared model, improves the average detection accuracy by approximately 16 percentage points, improves the average recall by approximately 25 percentage points, and produces KG-grounded evidence packages for review.

4 Discussion

The field results indicate that multimodal evidence, explicit rule grounding, and hybrid triggering improve safety monitoring beyond isolated vision pipelines. The hybrid trigger is important in harsh industrial scenes: when dust, glare, or occlusion reduces visual reliability, sensor violations and rule triggers can still activate cloud reasoning. Conversely, confidence-aware fusion reduces false alarms by checking whether isolated sensor spikes are supported by visual context or safety rules.

The KG contributes not only to risk inference but also to auditability. Each alert stores timestamps, key frames or clips, sensor snapshots, matched rules, and a reasoning path. This evidence package supports post-incident review, responsibility analysis, and compliance-oriented inspection. In practice, the system should be regarded as decision support and alarm prioritization rather than a replacement for safety governance because final responsibility remains with plant operators and safety managers.

Several limitations remain. First, KG coverage and freshness may affect judgments when plant-specific standard operating procedures change. Second, near-miss labels inevitably involve expert subjectivity, especially when historical logs lack complete context. Third, the reported latency was obtained with an on-premises deployment and a dedicated 10 Gigabit Ethernet (GbE) link, and performance may vary under constrained networks. Fourth, the current implementation uses a fixed edge hardware and quantization configuration, so the throughput-accuracy trade-off may shift on other platforms. Finally, pseudolabeling and domain adaptation can introduce confirmation bias if not periodically checked by human reviewers.

5 Conclusions and future work

This study presented a KG-augmented multimodal safety monitoring system for coal conveying and unloading operations in thermal power plants. The framework combines hybrid triggering, multimodal fusion, KG reasoning, and a two-tier edge-cloud architecture to provide fast and traceable alerts under noisy industrial conditions. In live deployment, the system reduced median ART to approximately 10 s with $p_{95} \leqslant 20$ s and achieved an approximately 67.1% faster response than a sequential legacy pipeline.

The system generated 3652 field alerts, and expert review confirmed 90%–100% precision for severe categories. The detection accuracy reached 98.7% for no-helmet violations, 96.8% for conveyor-belt jams, 95.1% for motor faults, 94.3% for unauthorized entry, and 91.7% for smoke/flame events. These results show that KG-augmented multimodal monitoring can improve both operational responsiveness and auditability in safety-critical power-plant environments.

Future work will focus on five directions. First, broader validation across additional plants and operating areas is needed to test cross-site generalization under different layouts, lighting, dust levels, and safety policies. Second, compound hazards and causal chains should be modeled more explicitly, especially when personnel, equipment, and environmental risks co-occur over time. Third, the KG should be expanded with schema versioning, provenance tracking, and collaborative updating to reflect evolving regulations and site-specific procedures. Fourth, lighter edge models should be developed through sparsity, distillation, mixed precision, and dynamic early-exit strategies while preserving recall for severe hazards. Finally, privacy-preserving continual learning, degraded-mode inference during network or sensor outages, adversarial robustness, secure audit trails, and digital-twin stress testing should be investigated before wider deployment.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2023YFB3307105). The authors sincerely thank the China Huadian Yong'an Power Plant, especially Xinghui CAO, for providing both experimental guidance and facilities. The authors also thank Guangshui ZHANG from the Electric Power Industrial Product Quality Standards Research Institute Co., Ltd., China for providing safety documentation.

Author contributions

Xinrong YAN, Chongbo ZHOU, and Zhengtao DING designed the research. Chongbo ZHOU wrote the first draft of the manuscript. Jiaxu QIAN helped organize the manuscript and revised and edited the final version.

Conflict of interest

Xinrong YAN, Chongbo ZHOU, Jiaxu QIAN, and Zhengtao DING declare that they have no conflicts of interest.

Declaration on the use of generative AI tools

During revision of this manuscript, Doubao was used only for language polishing and formatting assistance, and Qwen was used as a base model for finetuning. The authors reviewed and edited all AI-assisted content and take full responsibility for the final manuscript.

Data availability

The plant operation data used in this study are not publicly available because of safety, privacy, and industrial confidentiality restrictions. Deidentified data or implementation details may be made available from the corresponding author upon reasonable request and with approval from the participating plant.

References

Arnold SMR, Iqbal S, Sha F, 2021. When MAML can adapt fast and how to assist when it cannot. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, p.244-252.

Bai S, Chen KQ, Liu XJ, et al., 2025. Qwen2.5-VL technical report. arXiv:2502.13923. https://doi.org/10.48550/arXiv.2502.13923

Baltrusaitis T, Ahuja C, Morency LP, 2019. Multimodal machine learning: a survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Bordes A, Usunier N, Garcia-Duran A, et al., 2013. Translating embeddings for modeling multi-relational data. Proceedings of the 27th International Conference on Neural Information Processing Systems, p.2787-2795.

Fedus W, Zoph B, Shazeer N, 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(1):120.

Girshick R, 2015. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, p.1440-1448. https://doi.org/10.1109/ICCV.2015.169

Hu EJ, Shen YL, Wallis P, et al., 2022. LoRA: low-rank adaptation of large language models. Proceedings of the Tenth International Conference on Learning Representations.

Huang X, Li P, Wang Y, et al., 2025. Knowledge graph-augmented ERNIE-CNN method for risk assessment in secondary power system operations. *Energies*, 18(8):2104. https://doi.org/10.3390/en18082104

Iwanicki K, 2018. A distributed systems perspective on industrial IoT. Proceedings of the IEEE 38th International Conference on Distributed Computing Systems, p.1164-1170. https://doi.org/10.1109/ICDCS.2018.00116

Jiang PY, Ergu D, Liu FY, et al., 2022. A review of Yolo algorithm developments. *Procedia Computer Science*, 199:1066-1073. https://doi.org/10.1016/j.procs.2022.01.135

Kim W, Son B, Kim I, 2021. ViLT: vision-and-language transformer without convolution or region supervision. Proceedings of the 38th International Conference on Machine Learning, p.5583-5594.

Lee JY, Choi WS, Choi SH, 2023. Verification and performance comparison of CNN-based algorithms for two-step helmet-wearing detection. *Expert Systems with Applications*, 225:120096. https://doi.org/10.1016/j.eswa.2023.120096

Li JC, Wang XY, Zhu SJ, et al., 2024. CuMo: scaling multimodal LLM with co-upcycled mixture-of-experts. Proceedings of the 38th International Conference on Neural Information Processing Systems, p.4170.

Liao H, Tu JJ, Xia J, et al., 2021. Ascend: a scalable and unified architecture for ubiquitous deep neural network computing: industry track paper. Proceedings of the 2021 IEEE International Symposium on High-Performance Computer Architecture, p.789-801. https://doi.org/10.1109/HPCA51647.2021.00071

Liu AA, Xu N, Nie WZ, et al., 2019. Multi-domain and multi-task learning for human action recognition. *IEEE Transactions on Image Processing*, 28(2):853-867. https://doi.org/10.1109/TIP.2018.2872879

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, p.8748-8763.

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. arXiv:1707.06347. https://doi.org/10.48550/arXiv.1707.06347

Wang Y, Chung SH, 2022. Artificial intelligence in safety-critical systems: a systematic review. *Industrial Management & Data Systems*, 122(2):442-470. https://doi.org/10.1108/IMDS-07-2021-0419

Xia J, Cheng CN, Zhou XP, et al., 2021. Kunpeng 920: the first 7-nm chiplet-based 64-core ARM SoC for cloud services. *IEEE Micro*, 41(5):67-75. https://doi.org/10.1109/MM.2021.3085578

Electronic supplementary materials

Sections S1 and S2, Tables S1 and S2