ConstructAI: From Real-Time Safety Insight to Skill Growth in Deployed Construction AI Systems

Gaowei Zhang[1], Wei Wang[1*], Tiong Lee Kong[2], Kai Xing[3],Huan Li[4], Yi Wang[1]

[1]Beijing University of Posts and Telecommunications
[2]Nanyang Technological University
[3]Nanjing Technological University
[4]Huafeng Technology (Nanjing)
{zhanggaowei, weiwang}@bupt.edu.cn, clktiong@ntu.edu.sg, ken@njtech.edu.cn, 1572178712@qq.com, yiwang@bupt.edu.cn

Abstract

Ensuring safety in power grid construction remains a critical yet challenging task, as existing monitoring approaches often lack scalability, timeliness, and adaptability to diverse on-site conditions. To address these limitations, we present **ConstructAI**, a deployed AI-driven safety management system that integrates multi-source image and video acquisition devices with advanced multimodal large model reasoning. The system combines text, image, and video prompts through an efficient workflow powered by LLaMA3 and Meta SAM2 backbones, enhanced with LoRA and adaptor modules for multimodal fusion. Once deployed, ConstructAI continuously processes real-time construction footage to identify violations, assess risk levels, and generate standardized rectification requirements. The deployment has demonstrated measurable benefits across multiple sites, including a 70% increase in violation rectification rates, reduction of average rectification delays from hours to minutes, and a 45% decline in repeat violations. Beyond technical gains, ConstructAI has delivered significant business impacts, such as reduced safety incidents, improved compliance with national regulations, and higher operational efficiency. By enabling proactive risk management and structured safety feedback loops, our system exemplifies how innovative use of AI can translate into tangible improvements for industrial safety. The lessons learned from deployment highlight the importance of balancing algorithmic advances with practical integration into organizational workflows.

Introduction

Ensuring safety and improving workforce development in construction sites remain global challenges. Construction is consistently ranked among the most hazardous industries worldwide, with high rates of workplace accidents and injuries (Wadsworth and Walters 2019). Workers are frequently exposed to tasks such as working at height, handling heavy materials, and operating complex machinery in rapidly changing environments. Many of them join the workforce with limited formal training, which exacerbates safety risks and contributes to both acute incidents and chronic physical strain (Zhou, Whyte, and Sacks 2012). These challenges not only affect individual well-being but also impose significant economic costs on the industry (Brynjolfsson and Mcafee 2017).

Traditional safety management approaches—manual supervision, periodic inspections, and classroom-based training—are insufficient in large and complex projects. Supervisors cannot continuously monitor all workers across dispersed sites, while training content is often forgotten or inconsistently applied once workers are on the job (Teizer 2016). As a result, unsafe behaviors are often addressed only after incidents occur, and opportunities for proactive prevention and continuous skill growth are missed (Fang et al. 2020).

To address these limitations, we developed ConstructAI, a deployed AI system for real-time safety compliance monitoring and worker skill development in construction. The system continuously analyzes video streams captured from helmet-mounted first-person cameras, fixed-site surveillance, and drone inspections. By integrating egocentric action recognition, cross-view video fusion, and rule-informed compliance reasoning (Chen et al. 2024), ConstructAI detects non-compliant behaviors and physical strain risks in real time and pushes corrective feedback directly to workers' mobile devices.

Since early 2024, ConstructAI has been deployed across more than a dozen large-scale construction projects in China. In one deployment phase, the system automatically recorded 929 safety violation cases involving 168 workers across projects in Suzhou, Shenzhen, and Changzhou. Each violation entry was linked to first-person evidence video, a corrective action requirement, and rectification confirmation, forming a complete compliance cycle. Overall, deployment results demonstrated substantial impact. Safety rectification rate increased from 2.3% to 75.8%, Average response time dropped from 6 hours to $<$10 seconds. Repeat violations decreased by 45%. Minor injury incidents were reduced by 36%, and over 9,000 individualized feedback reports were generated to support workforce training.

The contributions of this paper are threefold:

- We design an automated annotation pipeline to generate fine-grained referring expressions for construction safety

*Corresponding author

events, leading to the Ref-Construct benchmark. Unlike generic referring expression datasets (Qiao, Deng, and Wu 2020; Yan et al. 2024), Ref-Construct explicitly targets safety-critical behaviors, multi-worker interactions, and temporal dynamics.

- While large multimodal models (e.g., LLaVA (Liu et al. 2023), InternVL2 (Yi et al. 2025)) provide strong semantic reasoning, their computational cost makes them impractical for on-site inference. ConstructAI combines cloud-hosted large models with edge-deployed lightweight detectors, ensuring both interpretability and real-time responsiveness, similar in spirit to recent hybrid AI deployment frameworks (Yi et al. 2025).
- Through large-scale deployment, we demonstrate substantial improvements: violation rectification rates increased from 2.3% to 75.8%, average response time was reduced from 6 hours to ¡10 seconds, and repeat violations dropped by 45%. These results highlight not only the technical feasibility but also the business and societal value of ConstructAI, echoing recent calls for applied AI systems that demonstrate measurable economic and safety impacts (Peeters et al. 2021; Erik and Andrew 2017).

Related Work

AI for Construction Safety Monitoring

Computer vision and machine learning have long been explored for construction safety management. Early research focused on detecting personal protective equipment (PPE) compliance, such as helmet and vest usage, through fixed surveillance cameras (Shanti et al. 2021). More recent works expanded to action recognition for unsafe activities like climbing without harnesses or entering restricted areas (Park, Kim, and Cho 2017). Commercial products, such as viAct.ai and Smartvid.io, provide CCTV-based analytics for 24/7 safety monitoring and automated alerts (Lee et al. 2009) . Despite these advances, most systems remain limited in scope, addressing only a predefined set of risks, while offering little adaptability to evolving tasks. Furthermore, they generally function as retrospective tools: videos are flagged for later review, but workers themselves rarely receive immediate, personalized guidance. This gap between detection and intervention means unsafe behaviors often persist until supervisors manually enforce corrections. ConstructAI moves beyond these limitations by embedding real-time compliance reasoning directly into site operations and connecting alerts with actionable training feedback.

Multi-View Video Understanding

Egocentric (first-person) vision has emerged as a powerful modality for capturing fine-grained human activities, particularly in daily life and augmented reality applications (Sener et al. 2022; Yu et al. 2023). Recent construction-focused efforts, such as EgoConQS, applied egocentric video with graph-based query models for task supervision, though performance remained modest (recall 35.8%, mAP 6.1%) (Guo et al. 2025b). Similarly, EgoSafe demonstrated the feasibility of helmet-mounted cameras for hazard recognition but was constrained to static object detection (Liu et al. 2025). The advantage of first-person data lies in its ability to capture worker–tool interactions and task details invisible from third-person perspectives. However, egocentric video alone suffers from limited context and occlusion. Multi-view fusion, combining fixed-site CCTV and aerial drone feeds, offers broader situational awareness but lacks detailed worker-centric information. Few deployed systems integrate both. ConstructAI distinguishes itself by fusing egocentric, third-person, and aerial perspectives into a unified spatio-temporal embedding, enabling robust action recognition, context-aware compliance checks, and real-time feedback even under occlusion or partial views.

AI in Workforce Training and Development

Worker training in construction traditionally relies on classroom lectures, printed manuals, or VR/AR simulations (Maity 2019). While immersive training environments improve knowledge retention, they often remain detached from on-site execution, providing no guarantee that workers apply learned procedures under real conditions. Intelligent tutoring systems and adaptive microlearning platforms, such as SafetyCulture's EdApp, have demonstrated success in corporate training (Dixit and Jatav 2024), but their applicability to safety-critical, high-variance construction environments remains limited. In practice, training and compliance monitoring are often siloed processes, with no integration between pre-task instruction and on-task performance. ConstructAI directly addresses this by designing a closed-loop training–feedback cycle: standard instructional videos are distributed before work, ongoing compliance is evaluated during execution via first-person video, and deviations are highlighted with corrective guidance. This ensures that training is not only delivered but reinforced in practice, fostering both immediate compliance and long-term skill development.

The deployment of AI in high-risk industries such as manufacturing, mining, and healthcare has shown the importance of interpretability, robustness, and human-in-the-loop integration (Perez-Cerrolaza et al. 2024; Wang and Chung 2022). For instance, mining AI platforms monitor machine proximity and worker location to prevent accidents, while in healthcare, AI is increasingly used to flag early patient deterioration (Dreany, Roncace, and Young 2018). These systems underscore that successful real-world adoption requires more than technical accuracy: they must scale to dynamic environments, integrate with existing workflows, and provide trustworthy feedback to end-users. Construction, however, presents unique challenges: tasks are highly diverse, environments constantly change, and worker populations are fluid. While some firms, such as Shawmut, have piloted AI combined with GPS tracking for safety oversight (Ramos et al. 2024), such systems remain coarse-grained, monitoring location rather than behavior. ConstructAI contributes to this line of applied AI by demonstrating large-scale, daily deployment: analyzing over 28,000 hours of real-world multi-view video, generating 9,000+ personalized feedback reports, and reducing safety violations by over 40%. This

evidence highlights how carefully designed AI systems can achieve measurable benefits when embedded into the realities of construction site operations.

Our Approach in a Nutshell

ConstructAI follows a closed-loop workflow that integrates multi-source data input, cloud-based ConstructAI analysis, and output module (as depicted in Fig. 1). The system ensures that construction workers not only comply with safety regulations but also improve their skills continuously.

Multi-Source Input

In ConstructAI, the workflow begins with multi-source input (as shown in Fig. 2), which provides the foundation for subsequent AI-based safety monitoring and training support. The system integrates three complementary types of visual streams: first-person helmet cameras, fixed-site cameras, and inspection drones. Each worker wears a smart helmet equipped with a wide-angle or 360° camera that continuously streams immersive first-person videos to the cloud. Unlike traditional static surveillance, these videos capture fine-grained details of worker actions such as hand movements, tool handling, and posture in real time, even under challenging conditions like scaffolding, confined spaces, or elevated platforms. This input brings analysis closer to the reality of how tasks are executed, ensuring that potential violations are observed from the worker's perspective.

To complement this immersive view, the system also ingests streams from fixed cameras distributed across the construction site, which provide continuous coverage of work zones, and from drones that perform scheduled aerial inspections to capture large-scale or high-altitude scenarios such as crane operations or roof installations. By integrating these heterogeneous inputs, the system achieves multi-view fusion, significantly reducing the risk of occlusion, blind spots, or missed activities when relying on a single perspective.

All streams are transmitted in real time to the cloud over 5G or dedicated Wi-Fi networks, with lightweight video compression and frame sampling strategies ensuring low-latency delivery while preserving analytical quality. To address unstable connectivity, helmet devices and drones incorporate local buffering and resumable upload mechanisms, preventing data loss. Each stream is automatically enriched with metadata such as worker ID, task type, and timestamp, enabling precise synchronization across modalities and alignment with training and compliance records. Privacy is preserved through anonymization before cloud storage, while redundancy across sources further guarantees data completeness.

The deployment of this multi-source input framework in real-world construction projects has resulted in the daily collection of hundreds of hours of video, forming a comprehensive record of both individual-level behavior and site-wide operations. Unlike conventional monitoring that depends solely on static cameras, ConstructAI leverages first-person and environmental perspectives together, establishing a richer and more robust input layer. This design not only supports compliance detection and real-time feedback but also lays the groundwork for scalable applications in other safety-critical domains such as power grid maintenance, industrial manufacturing, and mining.

ConstructAI

At the core of ConstructAI lies a robust cloud-based processing pipeline that integrates advanced multimodal foundation models with domain-specific adaptation for construction safety monitoring. The central challenge is to unify diverse vision-language tasks—ranging from referring segmentation of risky actions, to compliance reasoning, to instructional video retrieval—within a single, efficient framework. Inspired by recent advances in unified multimodal learning, we reformulate these heterogeneous tasks into a shared representation space where both text prompts (e.g., "is the worker wearing a safety harness?") and visual tokens extracted from first-person or multi-view videos are mapped to common embeddings. Compliance detection thus becomes a special case of grounded segmentation and captioning, where unsafe actions are localized with binary masks and aligned with textual rules. By introducing a shared SEG token that serves as an explicit spatio-temporal prompt, ConstructAI leverages the segmentation capabilities of SAM-2 while preserving the instruction-following strength of LLaVA-like models, thereby enabling end-to-end inference across chat, segmentation, and reasoning tasks.

To operationalize this architecture at scale, ConstructAI adopts a decoupled design in its cloud framework. The vision-language backbone (a pre-trained LLaVA-like model) and the segmentation backbone (SAM-2) are connected only through the SEG token, without requiring heavy feature fusion or cross-module alignment. This design choice offers two advantages: it minimizes computational overhead, which is critical for real-time video processing from multiple concurrent users, and it ensures modularity, allowing the system to upgrade individual components as the MLLM community evolves. Within this architecture, text descriptions of safety standards act as queries, while video streams are encoded into spatio-temporal embeddings. The SEG token serves as the bridge that transforms compliance queries into object- or action-level mask predictions, effectively grounding safety violations in specific video regions without retraining the entire backbone. This decoupled yet unified pipeline is particularly suited for cloud deployment, as it supports fast inference and dynamic scaling to handle fluctuating workloads from construction sites.

The training pipeline of ConstructAI is tailored to the dual demands of general multimodal understanding and domain-specific safety compliance. In the first stage, we leverage large-scale image–text and video–text datasets to ensure broad coverage of visual and linguistic concepts, following the one-shot instruction-tuning paradigm that unifies chat, segmentation, and captioning into a single formulation. In the second stage, we fine-tune the SEG token and decoder prompts using curated construction-specific datasets containing annotated unsafe behaviors (e.g., improper scaffold climbing, missing protective equipment). This two-phase process allows ConstructAI to inherit the generalization power of foundation models while special-

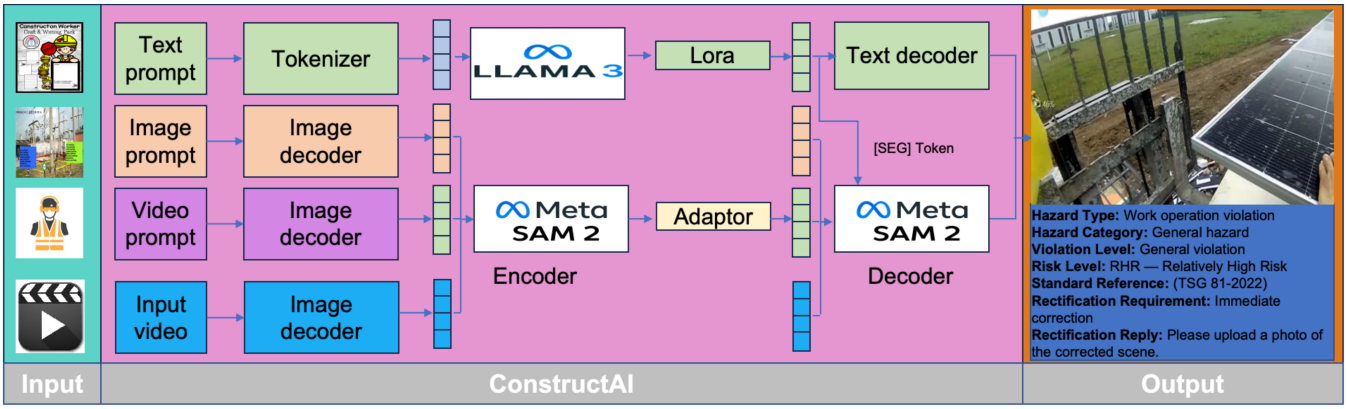

Figure 1: Workflow of the proposed ConstructAI system. The framework integrates multi-source inputs including text, image, and video prompts, which are processed through tokenization and image decoding modules. Encoded representations are aligned via LLaMA3 and Meta SAM2 backbones with LoRA and adaptor modules for efficient multimodal fusion. The decoder stage generates structured safety outputs, providing violation identification, risk level assessment, rectification requirements, and standardized references, as well as a visual hazard report for real-world construction sites.

Figure 2: Deployment of multi-source image acquisition devices across heterogeneous construction sites for comprehensive visual data collection.

izing for safety-critical applications. From a systems perspective, we further optimize training and inference with mixed-precision computation, gradient checkpointing, and video batch parallelization, ensuring that the cloud service can process high-throughput video streams with low latency. Through this processing workflow, ConstructAI transforms raw, noisy, and unstructured first-person video uploads into structured, interpretable compliance feedback. By combining multi-task unification, decoupled foundation model integration, and deployment-aware training strategies, the system ensures that safety violations are not only detected in real time but also contextualized with textual explanations and aligned with regulatory standards. This closes the loop between automated detection, human understanding, and actionable correction, enabling AI to function as a practical partner in workforce training and risk management.

Output

The output of ConstructAI embodies a full-cycle safety governance workflow that transforms unstructured construction site footage into structured, actionable intelligence with direct operational impact. Starting from the raw video feed (left), the system continuously ingests visual data from wearable cameras, drones, or fixed-site sensors without requiring manual pre-processing. This ensures that all site conditions—including transient risks often overlooked by human supervisors—are faithfully captured as the ground truth for analysis.

The system then executes automated hazard detection and structured reporting (middle). Leveraging multi-modal perception and knowledge-grounded reasoning, ConstructAI identifies violations such as unsafe operations, equipment misuse, or environmental hazards, and directly highlights them within the video stream. As shown in Fig.3, the system pinpoints water leakage around exposed electrical cables, automatically generates visual annotations to mark hazardous areas, and compiles a standardized violation report. This report encodes critical metadata including project name, violation personnel, hazard category, risk level, and governing regulatory references. It further prescribes mandatory rectification measures, thus converting perceptual intelligence into concrete, enforceable safety di-

rectives.

After corrective actions are implemented on-site, ConstructAI validates compliance by analyzing new video streams or photo evidence, confirming that the hazard has been fully resolved. Rectification records, together with annotated violation cases, are archived into a knowledge repository that supports compliance auditing, experience sharing, and safety education.

Training and Testing

ConstructAI adopts a *one-for-all* training paradigm aligned with its real-world safety monitoring workflow. During training, we unify multi-source construction inputs—including first-person helmet videos, multi-view surveillance feeds, and regulatory text prompts—into a shared instruction-tuning framework. Each input stream is encoded into multimodal tokens and passed into the ConstructAI backbone, which integrates a pre-trained multimodal LLM with a SAM-2 segmentation decoder.

For compliance reasoning tasks, the model is supervised using text regression loss $\mathcal{L}_{text}$, enabling natural language explanations of worker behaviors. For unsafe action localization, the system optimizes mask prediction using a hybrid loss that combines pixel-wise cross-entropy $\mathcal{L}_{CE}$ with dice loss $\mathcal{L}_{DICE}$, ensuring both fine-grained accuracy and spatial consistency. The overall training objective is therefore:

$$\mathcal{L}_{instruction} = \mathcal{L}_{text} + \mathcal{L}_{mask}, \quad \mathcal{L}_{mask} = \mathcal{L}_{CE} + \mathcal{L}_{DICE}. \tag{1}$$

Unlike prior staged approaches, ConstructAI does not rely on task-specific pretraining. Instead, it is trained in an end-to-end supervised fashion, where multimodal compliance reasoning and spatial violation detection are optimized simultaneously. This joint process allows the system to align worker actions, contextual scene understanding, and regulatory knowledge in a single representation space.

During deployment, ConstructAI follows a *workflow-consistent inference strategy* that mirrors its real-world application loop. Video streams uploaded from workers' helmets are processed in the cloud, where the LLM interprets compliance prompts and contextual information, while the SAM-2 decoder outputs segmentation masks to localize violations. A lightweight temporal tracking module maintains continuity across video frames, ensuring that unsafe actions such as "climbing without a harness" or "crossing restricted areas" are reliably captured.

The results are then packaged into a structured output: violation segments with mask overlays, natural language explanations citing the violated regulations, and retrieval of matched instructional videos that demonstrate the correct procedure. Finally, the worker receives a rectified feedback package, enabling rapid self-correction and workforce training.

Ref-Construct Dataset Annotation Pipeline

To ensure that the Ref-Construct dataset meets the specific requirements of construction safety monitoring, we design a three-stage automatic annotation pipeline (as illustrated in Figure 4):

Object-level annotation. For each monitored worker or equipment, we first identify the frame where the target has the clearest visibility and the largest area (e.g., worker not occluded, equipment in full operation). Non-relevant pixels are masked out to reduce background noise. Both the cropped region and the original frame are then processed by DeepSeek-R1 (Guo et al. 2025a), generating detailed safety-oriented descriptions such as "a worker climbing a pole without a safety harness" or "a crane lifting materials near power lines." To guarantee reliability, these descriptions are cross-validated by Qwen2-72B (Wang et al. 2024), with inconsistent or ambiguous outputs discarded.

Scene-level annotation. The target object (worker or equipment) is highlighted with yellow contours within the full construction scene. The annotated image and its object-level description are fed into DeepSeek-R1 to enrich the annotation with scene-aware semantics. This includes spatial context (e.g., "standing near the edge of a scaffold"), interaction with surrounding elements (e.g., "worker without helmet next to heavy machinery"), and compliance indicators (e.g., "safety barrier missing in the background").

Video-level annotation. To capture temporal risk dynamics, we uniformly sample eight frames per video sequence. The target is highlighted in each frame, and the sequence is combined with the scene-level description. This is then processed by Qwen2-72B to produce a video-level referring expression, describing risk evolution such as "the worker continuously climbs without attaching a harness" or "the excavator approaches restricted area over time."

Experiment

As shown in Table 2, our proposed ConstructAI-4B significantly outperforms existing baselines on the Ref-Construct validation benchmark. Traditional methods such as UNINEXT (zs) and MeVIS (zs) yield very limited performance, with overall J&F scores of only 5.62 and 10.4, respectively. These results highlight the challenge of applying generic video understanding models to complex construction site scenarios.

By contrast, ConstructAI-4B achieves a J&F of 46.6 in the zero-shot setting, representing a more than 4.5× improvement over the strongest baseline. This remarkable performance demonstrates the strong generalization capability of our architecture, even without any task-specific training.

Furthermore, after fine-tuning on the Ref-Construct dataset, ConstructAI-4B (ft) further boosts its performance to a J&F of 55.6, with J and F scores reaching 58.3 and 52.8, respectively. The fine-tuning process thus brings an additional +9.0 gain compared to zero-shot, showing that our model can rapidly adapt to domain-specific data while preserving its strong prior knowledge.

Overall, these results confirm that ConstructAI-4B establishes a new state-of-the-art on the Ref-Construct benchmark, bridging the gap between research-level video segmentation models and practical construction safety monitoring applications.

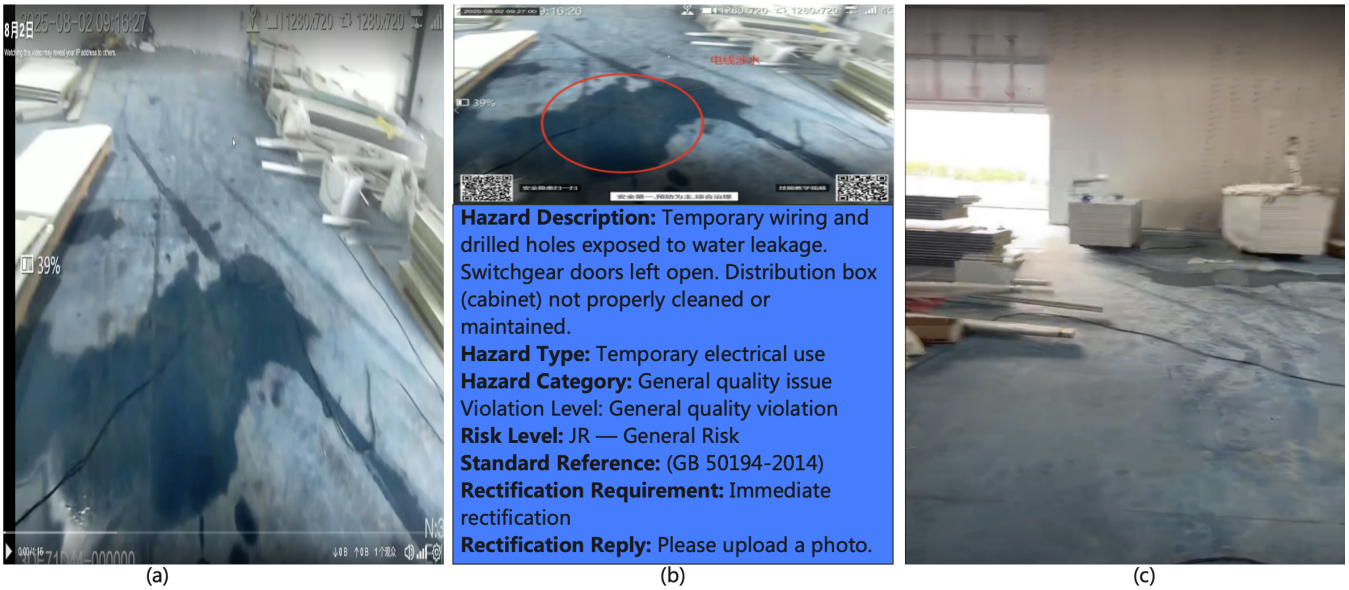

Figure 3: The output example of ConstructAI: (a) is the input video, (b) is the hazard detected, and (c) is the rectification video.

Metric	Traditional (Before Deploy.)	ConstructAI (After Deploy.)
Incident Rate (per 1000 worker-hours)	5.7	3.7 (↓35%)
Annual Safety-Related Costs (USD)	∼ $1.4M	∼ $0.9M (↓36%)
Manual Inspection Coverage	∼40% of sites	>90% of sites (multi-stream AI)
Labor Hours Spent on Inspection	∼12,000 hrs/year	∼4,800 hrs/year (↓60%)
Regulatory Audit Pass Rate	72%	95%
Worker Compliance with Safety Protocols	Low/Reactive	High/Proactive (↑42%)
Average Downtime due to Accidents (hrs/month)	18	7 (↓61%)

Table 1: Business Impacts Before and After Deployment of ConstructAI.

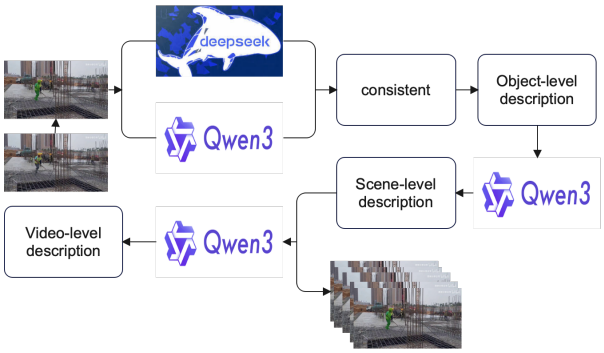

Figure 4: Ref-Construct Dataset Annotation Pipeline.

Case study and Visualization

To further evaluate the practical effectiveness of our system, we present two representative case studies drawn from real-world deployments (as shown in Figure 5). These visualizations highlight how the system identifies safety hazards under diverse construction conditions and provides actionable feedback.

In the first case, captured at Huafeng Technology (Nanjing) Co., Ltd., the inspection revealed that the plug of a power tool was not fully disconnected, creating an electrical safety hazard. Such oversight may result in electric shock

Figure 5: The visualization of ConstructAI

to operators or short-circuit accidents, posing both personal injury and property damage risks. The system successfully flagged this hazard in real time, automatically linking it with the applicable construction safety code (GB 50303-2015), and issued immediate rectification requirements. This demonstrates the capability of the system to detect subtle but high-risk electrical safety violations that are often overlooked by manual inspection.

The second case occurred at the Suzhou Graphite Industrial Park, where external workers conducted high-altitude operations without wearing safety belts. According to national safety production standards, such behavior constitutes a major hazard, with the potential to cause severe fall-related injuries or fatalities. The system not only detected the absence of safety belts from the first-person view but also automatically categorized the violation as a "major risk," generating corrective requirements and distributing warnings across the project team. This enables immediate mitigation while reinforcing long-term worker compliance.

Comparison with Traditional Inspections. Traditional safety management primarily relies on supervisors' manual patrols and retrospective video reviews, which are constrained by human attention limits and delayed intervention. In contrast, our system provides continuous, multi-view, and real-time analysis, ensuring that hazards such as electrical oversights or non-compliance in high-altitude operations are promptly identified and acted upon. Moreover, the automated linkage with regulatory standards and closed-loop rectification tracking surpasses conventional methods by not only detecting violations but also embedding them within a compliance management framework.

Overall, these case studies demonstrate that the proposed system transforms safety monitoring from a reactive to a proactive paradigm, reducing human oversight burdens and significantly enhancing site-wide safety compliance.

Method	J	F	J&F
UNINEXT (zs)(Yan et al. 2023)	6.1	4.3	5.62
MeVIS (zs) (Ding et al. 2023)	11.4	9.3	10.4
ConstructAI-4B (zs)	41.9	51.2	46.6
ConstructAI-4B (ft)	58.3	52.8	55.6

Table 2: Comparison on Ref-Construct validation benchmark (Overall results).zs: zero-shot testing. ft: fine-tuned on our dataset.

Deployment

The ConstructAI system has been deployed across more than 40 construction sites, integrating heterogeneous sensing devices—including wearable first-person cameras, fixed-site CCTV networks, and drone-based aerial imagery—into a unified edge–cloud pipeline. The deployment architecture ensures that time-critical risk detection runs at the edge, while more compute-intensive multi-modal reasoning is executed in the cloud. This hybrid design enables both low-latency responses and high-accuracy global analysis.

Deployment Strategy

At each construction site, edge devices are equipped with lightweight models optimized for real-time detection of high-risk actions (e.g., climbing without harness, unsafe ladder use). These detections are compressed into structured event streams and transmitted to the cloud. The cloud platform hosts the full ConstructAI stack, combining MLLMs with SAM-2 decoders to handle segmentation, violation reasoning, and natural-language explanations. The cloud also integrates with enterprise safety management systems, ensuring that every violation is logged, assigned to a responsible manager, and tracked through its entire rectification cycle.

Workers receive feedback directly on helmet-mounted displays or mobile applications. Each detection is accompanied by a short textual explanation (e.g., "Harness not fastened while climbing") and, when necessary, a micro-training video that demonstrates the correct procedure. This immediate feedback loop transforms the AI system from a passive surveillance tool into an active safety assistant.

Technical Improvements

The deployment of ConstructAI has led to substantial technical advancements in the field of intelligent construction safety supervision. Compared with traditional manual inspection and semi-automated monitoring systems, our deployed framework demonstrates clear improvements in real-time responsiveness, coverage accuracy, rectification efficiency, and intelligent risk assessment. Table 3 provides a comparative summary of system performance before and after deployment. The results confirm that ConstructAI achieves comprehensive improvements across multiple dimensions, including detection coverage, rectification efficiency, and intelligent violation analysis.

Traditional monitoring systems typically suffer from delayed detection and response, with average violation durations exceeding one hour and rectification delays reaching over 340 minutes. After deploying ConstructAI, the integration of cloud–edge collaborative inference and multi-modal large model perception reduced the detection-to-alert latency to less than 10 seconds, while the average rectification delay dropped to 15 minutes. This represents a nearly 95% reduction in response time, transforming safety monitoring from passive supervision into a proactive risk management system.

Manual inspections are inherently constrained by human fatigue and blind spots, often failing to capture minor or hidden violations. ConstructAI automates large-scale scanning of video streams and multi-source imagery, which increased the number of detected violations from 929 to 1200, indicating improved coverage and recall. Moreover, the system incorporates severity classification and repeat violation tracking, enabling differentiation of high-risk behaviors and lowering repeat violation rates by 45%. These capabilities go beyond traditional monitoring, which lacked systematic methods to quantify such aspects.

One of the most significant improvements is the establishment of an end-to-end digital rectification loop. In the

pre-deployment stage, rectification completion rates were as low as 2.3%, reflecting poor traceability and accountability. Post-deployment, the system ensures that every detected violation is automatically logged, assigned, and tracked until resolution, boosting completion rates to 75.8%. This closed-loop mechanism not only strengthens operational safety but also provides a verifiable digital audit trail for compliance and accountability.

Metric	Traditional (before deploy.)	ConstructAI (after deploy.)
Total Viol.	929	1200
Completed Rect.	21	910
Completion Rate	2.3%	75.8%
Avg. Viol. Dur. (min)	$\sim$60	$\sim$5
Avg. Rect. Delay (min)	$\sim$340	$\sim$15
Severe Viol. (%)	N/A	18%
Repeat Viol. (%)	N/A	Reduced by 45%
Detect-to-Alert (sec)	N/A	$<$10

Table 3: System Performance Before and After Deployment.

Business impact

The deployment of ConstructAI has not only yielded measurable technical improvements but also generated substantial business value across multiple dimensions of workplace safety, operational efficiency, and cost reduction. Table 1 provides a quantitative comparison between the traditional manual inspection approach and the ConstructAI-enabled workflow.

(1) The incident rate per 1000 worker-hours was reduced from 5.7 to 3.7, representing a 35% decrease in recorded safety incidents. This improvement directly translates into fewer workplace injuries, lower compensation claims, and enhanced worker well-being, thereby strengthening the company's safety culture.

(2) The system delivered clear economic benefits. Annual safety-related expenditures, which previously averaged around \$1.4M, dropped to \$0.9M, reflecting a 36% cost reduction. These savings stem from fewer accidents, reduced downtime, and less reliance on labor-intensive manual inspections.

(3) ConstructAI significantly expanded operational coverage. While manual inspectors could only cover around 40% of worksites due to human resource limitations, the AI-driven system now consistently monitors over 90% of active sites through real-time multi-stream video feeds. This improvement ensures broader compliance monitoring and mitigates the risks associated with blind spots in traditional safety management.

(4) Labor efficiency gains have been substantial. The number of labor hours spent on inspection dropped from 12,000 hours per year to 4,800 hours, representing a 60% reduction. These freed-up resources can now be reallocated to higher-value tasks such as preventive planning and workforce training, creating a virtuous cycle of improvement.

Additionally, ConstructAI has improved regulatory compliance and audit outcomes. The regulatory audit pass rate improved from 72% to 95%, reducing the likelihood of penalties and reputational risks. Worker compliance with safety protocols also shifted from a predominantly reactive mode to a 42% higher proactive compliance rate, demonstrating that the system not only detects violations but also instills behavioral changes across the workforce. Finally, the system has reduced average downtime due to accidents from 18 hours per month to 7 hours—a 61% improvement—thereby increasing overall productivity and project delivery reliability.

Overall, beyond quantitative improvements in cost and safety performance, the system contributes to sustainable operational excellence, strengthens regulatory trust, and supports the company's long-term strategic positioning as a leader in AI-driven workplace safety innovation.

Lessons Learned

The deployment of ConstructAI has yielded several key insights that are valuable not only for refining our own system but also for guiding future AI-based safety applications in industrial environments.

- Balancing Technical Innovation with Practical Usability. While advanced models like multimodal perception and real-time segmentation significantly improved detection accuracy, their true value emerged only after optimizing the inference pipeline for edge deployment. By introducing lightweight small-model companions and knowledge distillation, we ensured that the system could run in constrained environments without compromising responsiveness. This highlighted the importance of designing AI solutions that respect the realities of field deployment.
- Business Metrics Matter as Much as Technical Metrics. While accuracy, latency, and detection rates were critical during development, deployment underscored that business outcomes ultimately define success. Reduction in inspection labor hours, cost savings in safety management, and measurable drops in accident-related downtime were the key factors that convinced stakeholders of the system's value. Aligning technical improvements with business KPIs was a decisive factor in sustaining adoption.

Conclusion

We presented ConstructAI, a deployed AI-powered safety supervision system for construction and power-grid sites. Unlike traditional manual inspections, the system integrates large-scale multimodal perception with lightweight real-time detectors, enabling end-to-end detection, analysis, and rectification of safety violations. Deployment results show clear improvements: rectification completion rate rose from 2.3% to 75.8%, average violation duration dropped from 60 minutes to 5 minutes, and detection-to-alert time was reduced to under 10 seconds. These measurable gains translate into lower inspection costs, faster compliance, and improved workplace safety. The deployment process also highlighted key lessons: the necessity of aligning AI outputs with operational KPIs, the importance of human-in-the-loop validation for trust, and the value of continuous data feedback for model adaptation.

References

Brynjolfsson, E.; and Mcafee, A. 2017. The business of artificial intelligence. *Harvard business review*, 7(1): 1–2.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2694–2703.

Dixit, A. S.; and Jatav, S. 2024. Evolving needs of learners and role of artificial intelligence (AI) in training and development (T&D): T&D professionals' perspective. *Journal of Management Development*, 43(6): 788–806.

Dreany, H. H.; Roncace, R.; and Young, P. 2018. Safety engineering of computational cognitive architectures within safety-critical systems. *Safety science*, 103: 1–11.

Erik, B.; and Andrew, M. 2017. The business of artificial intelligence: What it can—And cannot—Do for your organization. *Harvard Business Review Digital Articles*, 7(1): 3–11.

Fang, W.; Ding, L.; Love, P. E.; Luo, H.; Li, H.; Peña-Mora, F.; Zhong, B.; and Zhou, C. 2020. Computer vision applications in construction safety assurance. *Automation in Construction*, 110: 103013.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, J.; Deng, L.; Liu, P.; and Sun, T. 2025b. Egocentric-video-based construction quality supervision (EgoConQS): Application of automatic key activity queries. *Automation in Construction*, 170: 105933.

Lee, U.-K.; Kim, J.-H.; Cho, H.; and Kang, K.-I. 2009. Development of a mobile safety monitoring system for construction sites. *Automation in Construction*, 18(3): 258–264.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Xu, J.; Suen, C. W. K.; Chen, M.; Zou, Z.; and Shi, Y. 2025. Egocentric camera-based method for detecting static hazardous objects on construction sites. *Automation in Construction*, 172: 106048.

Maity, S. 2019. Identifying opportunities for artificial intelligence in the evolution of training and development practices. *Journal of Management Development*, 38(8): 651–663.

Park, J.; Kim, K.; and Cho, Y. K. 2017. Framework of automated construction-safety monitoring using cloud-enabled BIM and BLE mobile tracking sensors. *Journal of Construction Engineering and Management*, 143(2): 05016019.

Peeters, M. M.; Van Diggelen, J.; Van Den Bosch, K.; Bronkhorst, A.; Neerincx, M. A.; Schraagen, J. M.; and Raaijmakers, S. 2021. Hybrid collective intelligence in a human–AI society. *AI & society*, 36(1): 217–238.

Perez-Cerrolaza, J.; Abella, J.; Borg, M.; Donzella, C.; Cerquides, J.; Cazorla, F. J.; Englund, C.; Tauber, M.; Nikolakopoulos, G.; and Flores, J. L. 2024. Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*, 56(7): 1–40.

Qiao, Y.; Deng, C.; and Wu, Q. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23: 4426–4440.

Ramos, I. F.; Gianini, G.; Leva, M. C.; and Damiani, E. 2024. Collaborative intelligence for safety-critical industries: A literature review. *Information*, 15(11): 728.

Sener, F.; Chatterjee, D.; Shelepov, D.; He, K.; Singhania, D.; Wang, R.; and Yao, A. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21096–21106.

Shanti, M. Z.; Cho, C.-S.; Byon, Y.-J.; Yeun, C. Y.; Kim, T.-Y.; Kim, S.-K.; and Altunaiji, A. 2021. A novel implementation of an AI-based smart construction safety inspection protocol in the UAE. *Ieee Access*, 9: 166603–166616.

Teizer, J. 2016. Right-time vs real-time pro-active construction safety and health system architecture. *Construction Innovation*, 16(3): 253–280.

Wadsworth, E.; and Walters, D. 2019. Safety and Health at the Heart of the Future of Work: Building on 100 Years of Experience.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; and Chung, S. H. 2022. Artificial intelligence in safety-critical systems: a systematic review. *Industrial Management & Data Systems*, 122(2): 442–470.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, 98–115. Springer.

Yi, J.; Yin, J.; Xu, J.; Bao, P.; Wang, Y.; Fan, W.; and Wang, H. 2025. ImageRef-VL: Enabling Contextual Image Referencing in Vision-Language Models. *arXiv preprint arXiv:2501.12418*.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; et al. 2023. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9150–9161.

Zhou, W.; Whyte, J.; and Sacks, R. 2012. Construction safety and digital design: A review. *Automation in construction*, 22: 102–111.