DOI - 10.68648/ijte-q-2026-49-03-008

A Preference-Aligned Vision Transformer Framework for Camera-Based Safety Monitoring in Building Construction

Vishal Kumar Chaudhary
Assistant Professor
Dept. of Civil Engineering
Government Engineering College Khagaria
Khagaria, Bihar
✉ vishal@geckhagaria.org.in

Pankaj Kumar
Assistant Professor
Dept. of Civil Engineering
BP Mandal College of Engineering, Madhepura
Madhepura, Bihar
✉ pankajkumarpk8888@gmail.com

ABSTRACT

Construction remains among the most hazardous sectors in India, where workers bear a share of global occupational hazards far exceeding their share of the labour force. Automated visual monitoring offers continuous supervision, but existing systems rely on convolutional detectors demanding large annotated datasets and specialist expertise, placing them beyond the reach of small and medium contractors. Recent multimodal language models have been evaluated only as benchmarks on static images, without a deployable architecture or any means of calibrating alerts to site conditions. This paper presents a four-layer framework integrating commodity Wi-Fi cameras with a compact hybrid convolutional-vision transformer backbone, a rule-mapping module that encodes site safety requirements as visually verifiable checks, and a preference-alignment loop in which routine confirm-or-dismiss responses from the site engineer are converted into preference pairs and used to calibrate alert severity through direct preference optimization. Three algorithms specify adaptive frame sampling, hazard severity ranking and on-site alignment. Compared with manual supervision, wearable sensing, convolutional detectors and cloud inference, the design requires no site-specific labelled dataset from the adopting contractor and is reconfigured by editing a table rather than by retraining. Limitations and field validation requirements are stated.

KEYWORDS *: Construction safety, Vision transformer, Preference alignment, Site monitoring, Personal protective equipment.*

NOTATIONS

Symbol	Description
H, W	Frame height and width
P	Patch edge length
N	Number of patch tokens
d	Embedding dimension
L	Number of encoder blocks
h	Number of attention heads
d_k, d_v	Key and value dimensions
C	Convolutional stem output channels
Δt	Sampling interval
p_{conf}	Detection confidence
w_r	Rule weight
g_z	Zone exposure factor
o	Observation record for one frame
r	Rule index
s(o,r)	Initial severity score
v(o,a)	Latent score of alert a
a_w, a_l	Preferred and rejected alert
τ	Escalation threshold
π_θ, π_r	Aligned and reference policy
β	Divergence coefficient
σ	Logistic function
N_{min}	Minimum preference-pair count

INTRODUCTION

The construction industry sustains a disproportionate share of the world's occupational injury burden. Reviewing the Indian position, Samanta and Gochhayat

report that 16.4 per cent of global occupational hazards are borne by workers in the Indian construction sector, which constitutes only 7.5 per cent of the world labour force, and identify non-use of personal protective equipment (PPE) among the leading contributory challenges [1]. Studying safety performance in the small-scale Indian construction industry, Singh and Misra found, through a multi-criteria ranking based on the dominance-based rough set approach, that the deployment of dedicated staff to verify that workers follow documented safety protocols ranks among the highest-weighted determinants of firm safety performance [2].

That finding frames the practical difficulty. Verification of compliance is valued but is labour-intensive and intermittent, and it competes for the same supervisory resource as production. Nor does training alone close the gap. In a study conducted in Hong Kong, Chung et al. found that poor worker knowledge of PPE remained the principal weakness even after mandatory basic safety training, and proposed continuous monitoring as the remedy [3].

Automated visual monitoring has therefore attracted sustained attention. Systems based on convolutional detectors have demonstrated reliable PPE compliance classification [4], simultaneous detection of fire and protective equipment [5], aerial inspection of unsafe conditions [6] and real-time alerting over consumer messaging platforms [7]. Each, however, depends on a purpose-built annotated dataset and on machine learning expertise to train and periodically retrain the model. Surveying practitioners in a developing-country setting, Yap et al. identified technological limitation, weak organisational commitment, and privacy and data security concerns among six dimensions of barriers to adopting new safety technologies [8]. For Indian construction specifically, Singh et al. applied a Grey-DEMATEL analysis and found real-time data processing, model accuracy and domain validity, and computational resource requirements to be the most influential blockades to generative artificial intelligence adoption [9].

Multimodal language models offer a route past the annotation requirement, since they interpret visual scenes from natural-language instruction without domain-specific training. Chaudhary et al. evaluated five such models on construction hazard recognition and concluded that structured prompt design offers a practical pathway for small firms; among their directions for future research they identified integrating model outputs into interactive workflows that close the loop between machine recognition and human decision-making [10]. No published architecture yet addresses that integration for continuous site operation, and none addresses a second problem that governs whether any monitoring system is used at all: the calibration of alert behaviour. Work on video anomaly detection notes that reducing false alarm rates is the dominant concern of that field [11], yet in construction monitoring the escalation threshold remains a fixed constant, insensitive to the work being performed or to the judgement of the engineer receiving the alerts.

This paper addresses both gaps. It presents the design of a framework in which commodity Wi-Fi cameras feed a compact hybrid convolutional-vision transformer backbone, site safety rules are expressed as visually verifiable checks, and the severity ranking that governs alert escalation is aligned to the preferences of the site engineer through routine feedback. Put plainly, the system watches the work through ordinary cameras and learns, from the engineer's own handling of the alerts it sends, which situations are worth flagging. The objectives are to specify the architecture at a level permitting implementation, to state the algorithms governing sampling, ranking and alignment, and to evaluate the design against existing approaches. The work is a design and feasibility study.

LITERATURE REVIEW

Vision-based compliance monitoring

Delhi et al. established the reference approach, applying transfer learning to a YOLOv3 base network to classify compliance into four categories from a dataset of 2,509 images collected at several construction sites, with the framework intended for integration into an organisation's safety workflow [4]. Kumar, Gupta and co-workers extended the scope to fire and trapped persons using YOLOv4, reporting a maximum mean average precision of 76.86 per cent on a self-constructed dataset [5]. Aerial platforms have been applied to the same problem by a separate group, who

deployed a UAV-mounted Faster R-CNN system that detects workers without helmets and issues immediate alerts, reducing the inspection workload placed on site supervisors [6]. More recently, Geetha et al. combined YOLOv8 with consumer messaging for automated notification of violations [7]. The paradigm has been extended to quality management, where Mahamivanan et al. addressed material recognition using test-time augmentation and outlier training samples to improve robustness against varying illumination [12].

Across this body of work the dependency is common: a labelled dataset must be assembled for the classes of interest, and any change to the rule set requires re-annotation and retraining. This is the mechanism by which the dataset and expertise barriers identified in [8] and [9] operate in practice.

Instrumented and sensor-based alternatives

Wearable sensing offers a different route. Arun Kumar et al. proposed a LoRa and Zigbee based architecture with helmet, shoe and glove detection motes coordinated by a local server and gateway [13]. Such systems avoid image processing but require an instrumented device per worker with attendant battery, ruggedisation and issue-and-return overheads, and they detect the presence of the instrumented article rather than its correct use. IoT-based environmental and personnel monitoring has been shown to work at the system level [3].

Transformer architectures and preference alignment

The transformer architecture [14] was adapted to images by Dosovitskiy et al., who showed that a sequence of linearly embedded fixed-size image patches processed by a standard transformer encoder attains classification performance competitive with convolutional networks [15]. Hybrid designs have since proved effective in civil engineering field conditions: Li et al. developed CrackTrNet, a lightweight U-shaped model in which a convolutional component extracts low-level features and a transformer component learns global context, reaching 97.60 per cent segmentation accuracy at a model size of 34.86 MB, between 66.12 and 87.85 per cent smaller than mainstream comparators, and deployed it in a working mobile inspection application [16]. Paulraj and Vairavasundaram combined shifted-window attention with convolutional recurrent layers for weakly supervised anomaly detection in industrial surveillance, noting explicitly that reduction of false alarm rates is the field's governing objective [11].

Alignment to human preference is the idea that a system can be taught what is wanted by being shown which of two outcomes a person prefers, rather than by being given a formula to optimise. It originates with Christiano et al., who showed that complex behaviour can be learned from such pairwise comparisons where no explicit reward function can be specified [17]. Ouyang et al. established the canonical pipeline of supervised fine-tuning, reward modelling and policy optimisation [18]. Rafailov et al. subsequently showed that the same objective can be solved with a simple classification loss on preference pairs, yielding an algorithm that is stable, computationally lightweight and free of extensive hyperparameter tuning [19]. Retzlaff et al. argue that human-in-the-loop methods are particularly advantageous precisely where the reward function is difficult or impossible to define, and set out the explanation requirements for users who are neither AI specialists nor machine learning practitioners [20]. Most directly, Zhang et al. applied supervised fine-tuning followed by direct preference optimization to multimodal disaster severity assessment. The fine-tuning stage raised accuracy from 73.64 to 78.29 per cent and improved explanation quality by approximately 25 per cent; the preference-optimization stage that followed improved interpretability and the agreement between model reasoning and human judgement, and the framework as a whole reduced high-risk misclassification [21]. The authors are careful to note that the gain attributable to preference optimization alone is modest, obtained on a preference set of 229 training instances.

That result is significant for the present work because it demonstrates, in an adjacent visual safety-assessment domain, the mechanism proposed here.

METHODOLOGY

System architecture

The proposed framework is organised as four layers, shown in Fig. 1. The separation is functional: each

layer exposes a defined interface to the next, so that a component may be substituted without redesign of the remainder. All inference is performed on site; no worker imagery leaves the premises, which addresses the privacy and data security barrier identified in [8] and removes the dependence on site connectivity implied by cloud inference.

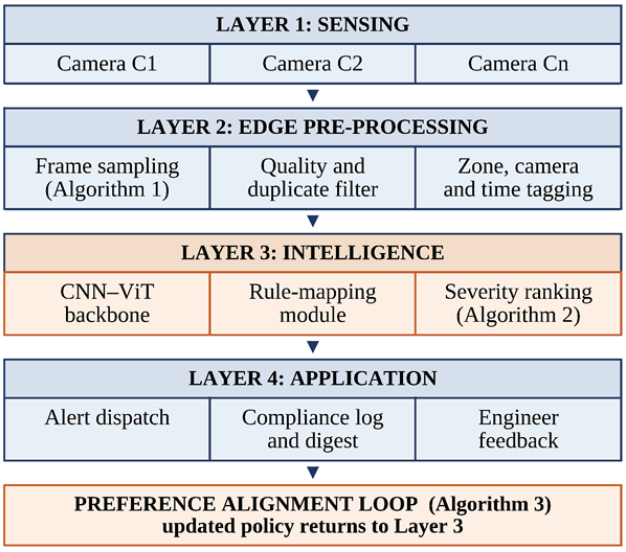


Fig. 1 Four-layer architecture of the proposed monitoring framework

Sensing layer

The sensing layer comprises commodity Wi-Fi or IP cameras of 1080p resolution supporting the real-time streaming protocol. Cameras cover defined work zones rather than the site as a whole, and are mounted at a height of 3.5 to 4.5 m with a downward tilt that keeps the working plane in the central part of the frame, where perspective distortion is least. Frames are cropped to the monitored zone and resized to the square input the backbone expects, so that the sides of the wider 16:9 sensor frame, which lie outside the zone, are discarded rather than compressed. Fig. 2 shows a typical arrangement. Two cameras give overlapping cover of the zone, so that a worker hidden from one view is usually visible in the other. Each camera carries a zone identifier and its own subset of rules, so the same installation applies different checks to a reinforcement yard and to an edge-protection zone.

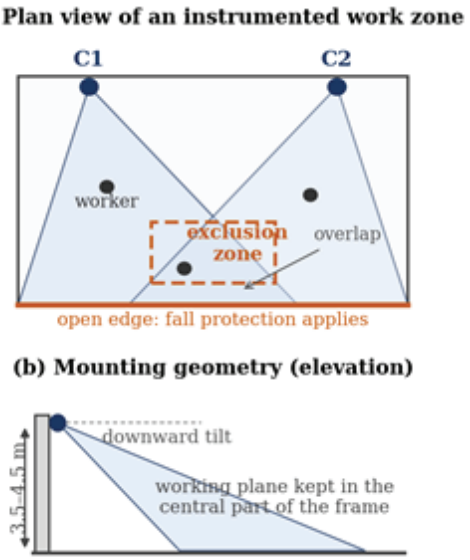


Fig. 2 Camera placement and zone coverage

Edge pre-processing layer

Continuous video is unnecessary and wasteful for the analysis performed here. The edge layer, hosted on a single-board computer, applies scheduled sampling at interval Δt together with motion-triggered capture, then discards frames that are blurred or that duplicate the preceding retained frame. Blur is assessed by the variance of the Laplacian; redundancy by the normalised absolute difference between successive frames. Algorithm 1 states the procedure. For a deployment of n cameras at $\Delta t = 300$ s over a ten-hour working day, the nominal load is 120n frames per day before filtering, which is within the capacity of modest edge hardware.

Algorithm 1: Selecting frames worth analysing

```
Input : the cameras, the sampling
        interval, and the limits for
        sharpness and change
Output: a queue of frames to analyse

1.  For each camera in turn:
2.    Capture a frame, on schedule or on
      detected movement.
3.    Measure how sharp it is. If too
      blurred, discard it and go to the
      next camera.
4.    Compare it with the last frame kept
      from that camera. If almost nothing
      has changed, discard it and go to
      the next camera.
5.    Otherwise label it with work zone,
      camera and time.
6.    Add it to the queue and keep it as
      the new reference frame.
7.  Return the queue of retained frames.
```

Intelligence layer: perception

The perception module is a compact hybrid backbone, illustrated in Fig. 3. A shallow convolutional stem extracts local features before tokenisation, following the division of labour validated for civil engineering imagery in [16]; the transformer encoder then relates every part of the frame to every other part. This matters because safety compliance is a judgement about relationships rather than about isolated objects, such as a worker standing next to an unguarded edge, or a helmet that is present but not being worn.

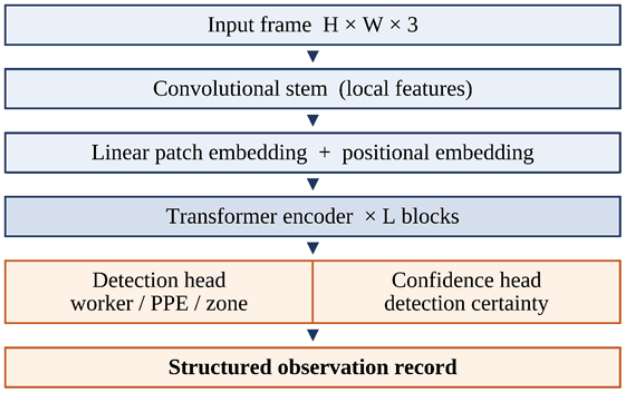


Fig. 3 Perception pipeline of the intelligence layer

Following [15], the stem output, which retains the spatial resolution H × W of the input frame, is divided into N non-overlapping patches of size P × P, where

$$N = H\,W / P^2 \qquad (1)$$

Each patch is flattened and linearly projected to dimension d, and a learnable positional embedding is added, giving the token sequence

$$z_0 = [\, x^1_p\, E\, ;\, \ldots\, ;\, x^N_p\, E\,] + E_{pos} \qquad (2)$$

The classification token of [15] is omitted, since the heads described below operate on the patch tokens themselves. Each encoder block applies multi-head self-attention, in which queries and keys have dimension d_k and values dimension d_v [14],

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k})\, V \qquad (3)$$

The backbone is not trained from scratch by the adopting contractor. It is initialised from transformer weights pretrained on general image collections and then fine-tuned once, by the developer, on publicly available construction safety image sets. What the framework removes is therefore not model training as such, but the requirement that each contractor label a dataset of its own before use, and repeat that whenever a rule changes. Table 1 gives the proposed configuration; the parameter budget keeps the compiled model within the size range shown to be deployable on constrained hardware in [16].

Table. 1. Proposed configuration of the perception backbone

Parameter	Value	Basis
Input H × W	640 × 640	Retains worker-scale detail across the zone
Patch size P	16	Standard granularity [15]
Tokens N	1,600	From Eq. (1)
Dimension d	384	Capacity vs. size
Blocks L	8	Reduced depth for edge
Heads h	6	d/h = 64 per head [14]
Conv. stem	3 layers	Local features [16]
Stem channels C	32	Feeds the patch embedding
Parameters	≈ 18 M	From the above configuration
Target size	< 40 MB	At half precision; cf. [16]

Intelligence layer: rule mapping and ranking

The rule-mapping module translates site safety requirements into visually verifiable conditions, each carrying a rule weight w_r reflecting the consequence of the violation. This structure is what permits reconfiguration without retraining: adding a rule adds an entry to the mapping table, not a new training cycle. Table 2 gives representative entries. Evidence that language models can be aligned to formal construction regulatory text supports extension of this table by retrieval from codified sources rather than by manual entry [22].

Table. 2. Representative rule-to-visual-check mapping

Requirement	Visually verifiable condition	w_r	Default escalation
Head protection	Helmet worn by each detected worker	0.9	Immediate
High-visibility clothing	Reflective vest on torso region	0.6	Immediate

Fall protection	Harness visible in edge zone	1.0	Immediate
Zone restriction	No worker in exclusion polygon	1.0	Immediate
Housekeeping	Access route free of material	0.4	Digest
Plant separation	No worker in plant swing radius	0.9	Immediate

For an observation o at time t comprising a detected violation of rule r with confidence p_{conf} in zone z, an initial severity score is formed as

$$s(o,r) = w_r \cdot p_{conf} \cdot g_z,\ w_r, g_z \in (0,1] \quad (4)$$

where g_z is a zone exposure factor. Both w_r and g_z lie in (0,1], so s is bounded by unity and τ may be read directly as a fraction of the maximum severity. Alerts are escalated when s(o,r) exceeds a threshold τ. In a conventional system τ is a fixed constant. This is, on the present argument, one source of the false alarm rates reported as the governing difficulty in [11]: a threshold that suppresses nuisance alerts in one work context suppresses genuine ones in another. In the proposed framework the ranking is produced by a learned policy π_θ aligned to site preference, as set out below. Algorithm 2 states the inference and ranking procedure.

Algorithm 2: Deciding which findings to report

```
Input : the frame queue, the rules for
        each zone, the ranking policy,
        the escalation threshold
Output: alerts to send, and a full log

1.  For each frame in the queue:
2.    Identify workers, protective
      equipment and zones.
3.    For each rule applying to the zone:
4.      Check whether it is broken.
5.      If so, rate the seriousness from
        the rule, the confidence of the
        analysis and the location.
6.      Adjust the rating using the
        current ranking policy.
7.      Record the finding in the log.
8.      If the rating is high enough,
        prepare an alert carrying the
        image and a short reason.
9.  Return the alerts and the log.
```

Preference alignment loop

The correct escalation for a given observation cannot be written down in advance, because it depends on the activity in progress, the stage of works and the tolerance of the site. This is the condition under which human-in-the-loop methods are held to be particularly advantageous [20]. The framework therefore treats the site engineer's routine handling of alerts as its teacher, much as a new assistant learns which matters to raise by seeing which reports are acted upon and which are waved away. Each dispatched alert carries three responses, namely confirm, dismiss and defer, and each requires a single interaction. Because a response is elicited on an alert the engineer must address in any case, the marginal burden is close to zero.

Responses are converted into preference pairs. Where one alert, a w, is confirmed and a comparable alert, a_l, dismissed, the pair is written to a buffer, $a_w > a_l$ denoting the preferred alert. Under the Bradley–Terry model, writing v(o,a) for the latent score expressing how strongly alert a merits escalation given observation o, the probability of the observed preference is

$$P(a_w > a_l \mid o) = \sigma[\, v(o,a\,w) - v(o, a_l)\,] \quad (5)$$

Following [19], the policy is obtained without fitting an explicit reward model by minimising

$$L = -E\{\log \sigma[\, \beta \log(\pi_\theta(a_w \mid o)/\pi_r(a_w \mid o)) - \beta \log(\pi_\theta(a\,l \mid o)/\pi_r(a_l \mid o))\,]\,\} \quad (6)$$

where π_r is the initial policy and β controls divergence from it. The update executes on site when the buffer reaches a minimum size N_{min}, taken as 200 pairs. Since a pair requires two comparable observations handled differently, pairs accumulate more slowly than responses; at a nominal 15 to 25 responses per working day, a first alignment cycle may be expected after roughly four to six working weeks. Retaining π_r as a reference and bounding divergence prevents the aligned policy from drifting arbitrarily far from the initial rule-derived behaviour, which is itself set from the mapping of Table 2. Fig. 4 shows the loop and Algorithm 3 states the procedure.

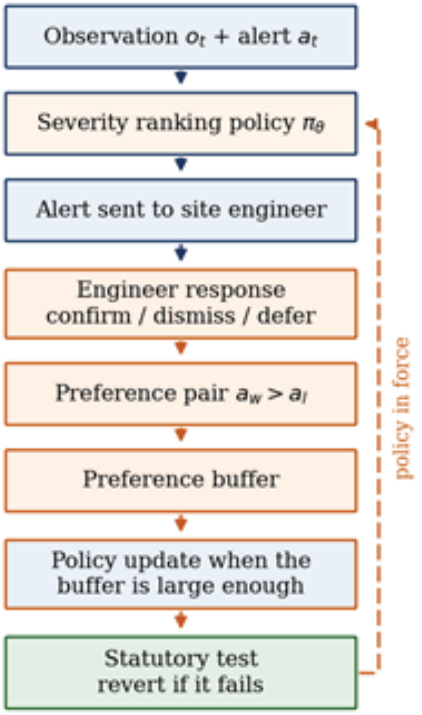


Fig. 4 Site-specific preference alignment loop

Algorithm 3: Learning the site's preferences

```
Input : the alerts sent, the reference
        policy, the preference buffer,
        the minimum buffer size
Output: an updated ranking policy

1.  For each alert sent, note whether the
    engineer confirmed, dismissed or
    deferred it. Ignore the deferred.
2.  Pair similar situations where one
    alert was confirmed and another
    dismissed; each pair records which
    the engineer preferred.
3.  Add the pairs to the buffer.
4.  Once the buffer reaches the minimum
    size:
5.    Update the policy to favour the
      alerts the engineer confirms.
6.    Test it against the statutory rules
      that must always be enforced.
7.    If the test fails, keep the
      previous policy.
8.    Discard the oldest pairs so the
      policy tracks current work.
9.  Return the policy in force.
```

Application layer

Escalated alerts are dispatched over a consumer messaging service, following the practicability demonstrated in [7], each accompanied by the annotated frame and a short rationale naming the rule and the observed condition. The explanation requirement follows [20]: a site engineer is neither an AI specialist nor a machine learning practitioner, and an alert without an intelligible basis will not be acted upon. Lower-severity observations accumulate into a daily compliance digest, and the full log supports periodic review and toolbox talks, a secondary use also identified in [10].

RESULTS AND DISCUSSION

Comparative evaluation

Table 3 evaluates the proposed design against the four alternatives available to a contractor. The comparison uses characteristics that determine adoption rather than accuracy alone, since the barriers documented in [8] and [9] are organisational and practical as much as technical.

Table. 3. Comparative analysis of construction safety monitoring approaches

Approach	Labelling by adopter	Adopter ML skill	Reconfig-uration	Alert calibration
Manual supervision	None	None	Instruction	Human judgement
Wearable sensing [13]	None	Low	Hardware change	Fixed rule
CNN detector [4,5]	≈ 2,500	High	Retraining	Fixed threshold
Cloud VLM [10]	None	Low	Prompt edit	Not addressed
Proposed framework	None by adopter	Low	Mapping-table edit	Aligned to site

Three characteristics omitted from Table 3 for space reinforce the comparison. Wearable schemes alone require a device on every worker, with the attendant issue-and-return burden; manual supervision alone gives intermittent rather than continuous coverage; and cloud inference alone requires site imagery to leave the premises.

Two differences are decisive. The first is where the labelling burden falls. The reference convolutional system had to be taught from 2,509 images hand-

labelled for that deployment [4], and every new rule calls for further labelling by the adopter. In the proposed framework that burden is borne once by the developer, using public datasets, and the rule set is thereafter extended by adding a line to a table. The second is alert calibration. No comparator adapts its alerting to the site: the deployed systems apply a fixed threshold or a fixed rule, while the benchmarking study defines no alerting mechanism at all. The surveillance literature identifies false alarm rate as its governing problem [11], and the proposed framework is the only one of the five in which escalation adapts to the judgement of the person receiving the alerts.

Expected effect of alignment

No claim of measured performance is made here. The expectation rests on the closest available evidence, and that evidence must be read precisely. In [21] the large measured gains, accuracy from 73.64 to 78.29 per cent and explanation quality by approximately 25 per cent, belong to the supervised fine-tuning stage rather than to preference optimization. What preference optimization contributed there was improved interpretability and closer agreement between model reasoning and human judgement, which its authors describe as modest and attribute to a preference set of only 229 training instances. That is nonetheless the quantity this framework needs. The objective here is not higher detection accuracy but alerting that matches the judgement of the person receiving it, which is precisely what preference optimization was found to improve. The structural correspondence is close, since both are vision-language pipelines producing a graded severity judgement aligned by human preference through a single annotation workflow, and the proposed loop accumulates preference pairs continuously rather than from a fixed set of a few hundred. The transfer is therefore a reasoned expectation, not a result.

Worker consent and acceptance

Continuous camera observation of a workforce is not a neutral act, and a monitoring system that workers regard as a disciplinary instrument will be resisted or defeated. The framework keeps imagery on site and raises alerts about conditions rather than named individuals, but neither measure substitutes for prior consultation with workers and their representatives, plain notification that the zone is monitored, and a written undertaking that the compliance log is used to correct hazards and not to penalise individuals. Where the workforce is largely unorganised, as it is for the tens of millions employed in Indian construction [1], that undertaking carries particular weight and should be treated as a condition of deployment rather than as good practice.

Limitations

Five limitations are material. First, environmental variability: illumination change, dust, rain and partial occlusion degrade visual inference, and the augmentation strategies used in [12] would form part of any implementation without eliminating the problem. Second, view stability: a construction site reconfigures as work proceeds, and formwork, scaffolding and stacked material will in time obstruct or invalidate a fixed viewpoint, so camera positions must be reviewed at each stage change, which is an operational burden the framework reduces but does not remove. Third, feedback quality: a policy aligned to one engineer inherits that engineer's inconsistencies and biases, a limitation examined generally in [23] and specifically in relation to reward misspecification and sparse feedback in [24]. No step of Algorithm 3 corrects this fully; the statutory test at step 6 catches only those biases that would breach statutory duty. Fourth, incentive alignment: the engineer who teaches the system is also the person whose site its output reflects upon, so dismissals made to reduce workload would train the system to under-report. The statutory test is the safeguard, and periodic independent audit of the compliance log should accompany deployment; alignment optimising for local preference may otherwise drift from statutory duty. Fifth, and governing, the framework is unvalidated. It is a specification supported by evidence from adjacent domains, not a measured system.

CONCLUSIONS

This paper has specified a four-layer framework for camera-based construction safety monitoring that addresses the two barriers keeping such systems out of small and medium contracting practice, namely the need for a site-specific annotated dataset and the need for machine learning expertise. By combining a compact hybrid convolutional-vision transformer backbone with a rule-mapping module, the design removes the

requirement for each contractor to assemble a purpose-built annotated dataset, and permits the rule set to be extended by editing a table rather than by retraining, answering both barriers as they are documented for developing-country construction. By converting the site engineer's routine confirm-or-dismiss handling of alerts into preference pairs and aligning the severity ranking through direct preference optimization, it addresses fixed-threshold alerting, which the present work identifies as a source of the false alarm rates reported as the governing difficulty in the surveillance literature, and it does so at close to zero marginal burden on the engineer. All inference is performed on site, which answers the privacy and data security barrier and removes dependence on connectivity.

The novelty lies in the combination: existing work either benchmarks multimodal models on static images without a deployable architecture, or deploys convolutional detectors with fixed thresholds and heavy annotation requirements.

The framework has not been validated on site data. Field implementation on an operating building construction project, measurement of detection precision and recall against expert annotation, and quantification of the change in false alarm rate before and after alignment constitute the necessary next stage. Extension of the perception layer from independent frames to short temporal sequences, and of the mapping table from safety to quality monitoring [12] and to progress monitoring [25], are identified as further directions.

REFERENCES

1. Sasmita Samanta and Jyotiranjan Gochhayat (2023), "Critique on Occupational Safety and Health in Construction Sector: An Indian Perspective", Materials Today: Proceedings, Vol. 80, Pp. 3016-3021. DOI: 10.1016/j.matpr.2021.05.707
2. Arpit Singh and Subhas C. Misra (2021), "Safety Performance & Evaluation Framework in Indian Construction Industry", Safety Science, Vol. 134, Art. 105023. DOI: 10.1016/j.ssci.2020.105023
3. William Wong Shiu Chung, Salman Tariq, Saeed Reza Mohandes and Tarek Zayed (2023), "IoT-based Application for Construction Site Safety Monitoring", International Journal of Construction Management, Vol. 23, No. 01, Pp. 58-74. DOI: 10.1080/15623599.2020.1847405
4. Venkata Santosh Kumar Delhi, R. Sankarlal and Albert Thomas (2020), "Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques", Frontiers in Built Environment, Vol. 06, Art. 136. DOI: 10.3389/fbuil.2020.00136
5. Saurav Kumar, Himanshu Gupta, Drishti Yadav, Irshad Ahmad Ansari and Om Prakash Verma (2022), "YOLOv4 Algorithm for the Real-time Detection of Fire and Personal Protective Equipments at Construction Sites", Multimedia Tools and Applications, Vol. 81, No. 16, Pp. 22163-22183. DOI: 10.1007/s11042-021-11280-6
6. Sourav Kumar, Mukilan Poyyamozhi, Balasubramanian Murugesan, Narayanamoorthi Rajamanickam, Roobaea Alroobaea and Waleed Nureldeen (2024), "Investigation of Unsafe Construction Site Conditions Using Deep Learning Algorithms Using Unmanned Aerial Vehicles", Sensors, Vol. 24, No. 20, Art. 6737. DOI: 10.3390/s24206737
7. G. Geetha, J. N. Swaminathan, K. Vijayalakshmi and M. Vinothkumar (2026), "Intelligent Worksite Protective Equipment Compliance Detection Using Deep Neural Networks", Journal of Engineering and Technology for Industrial Applications, Vol. 12, No. 58, Pp. 981-986. DOI: 10.5935/jetia.v12i58.3224
8. Jeffrey Boon Hui Yap, Canwin Guan Ying Lam, Martin Skitmore and Nima Talebian (2022), "Barriers to the Adoption of New Safety Technologies in Construction: A Developing Country Context", Journal of Civil Engineering and Management, Vol. 28, No. 02, Pp. 120-133. DOI: 10.3846/jcem.2022.16014
9. Arpit Singh, Anchal Patil and Shri Krishna Pandey (2026), "Unravelling the Blockades in Implementing Large Language Models in Construction Sector", International Journal of Mathematical, Engineering and Management Sciences, Vol. 11, No. 02, Pp. 991-1022. DOI: 10.33889/IJMEMS.2026.11.2.041
10. Nishi Chaudhary, S. M. Jamil Uddin, Sathvik Sharath Chandra, Anto Ovid and Alex Albert (2026), "Prompt to Protection: A Comparative Study of Multimodal LLMs in Construction Hazard Recognition", IEEE Access, Vol. 14, Pp. 70565-70580. DOI: 10.1109/ACCESS.2026.3691685
11. Shalmiya Paulraj and Subramaniyaswamy Vairavasundaram (2025), "Transformer-enabled

Weakly Supervised Abnormal Event Detection in Intelligent Video Surveillance Systems", Engineering Applications of Artificial Intelligence, Vol. 139, Art. 109496. DOI: 10.1016/j.engappai.2024.109496

12. Hadi Mahamivanan, Navid Ghassemi, Mohammad Tayarani Darbandy, Afshin Shoeibi et al. (2025), "Material Recognition for Construction Quality Monitoring Using Deep Learning Methods", Construction Innovation, Vol. 25, No. 03, Pp. 732-760. DOI: 10.1108/CI-04-2022-0074

13. G. S. Arun Kumar, Rajesh Singh, Anita Gehlot and Shaik Vaseem Akram (2022), "LoRa Enabled Real-time Monitoring of Workers in Building Construction Site", International Journal of Electrical and Electronics Research, Vol. 10, No. 01, Pp. 41-50. DOI: 10.37391/ijeer.100106

14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit et al. (2017), "Attention is All You Need", Advances in Neural Information Processing Systems, Vol. 30, Pp. 5998-6008. DOI: 10.48550/arXiv.1706.03762

15. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn et al. (2021), "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale", International Conference on Learning Representations. DOI: 10.48550/arXiv.2010.11929

16. Mingchao Li, Jingyue Yuan, Qiubing Ren, Qiling Luo, Junen Fu and Zhitang Li (2024), "CNN-Transformer Hybrid Network for Concrete Dam Crack Patrol Inspection", Automation in Construction, Vol. 163, Art. 105440. DOI: 10.1016/j.autcon.2024.105440

17. Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic et al. (2017), "Deep Reinforcement Learning from Human Preferences", Advances in Neural Information Processing Systems, Vol. 30, Pp. 4299-4307. DOI: 10.48550/arXiv.1706.03741

18. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida et al. (2022), "Training Language Models to Follow Instructions with Human Feedback", Advances in Neural Information Processing Systems, Vol. 35, Pp. 27730-27744. DOI: 10.48550/arXiv.2203.02155

19. Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon et al. (2023), "Direct Preference Optimization: Your Language Model is Secretly a Reward Model", Advances in Neural Information Processing Systems, Vol. 36. DOI: 10.48550/arXiv.2305.18290

20. Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi et al. (2024), "Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities", Journal of Artificial Intelligence Research, Vol. 79, Pp. 359-415. DOI: 10.1613/jair.1.15348

21. Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid and Mourad Oussalah (2026), "Towards Reliable Multimodal Disaster Severity Assessment through Preference Optimization and Explainable Vision-Language Reasoning", Reliability Engineering and System Safety, Vol. 275, Art. 112674. DOI: 10.1016/j.ress.2026.112674

22. Mohammad Aqib, Mohd Hamza, Qipei Mei and Ying Hei Chui (2025), "Fine-tuning Large Language Models and Evaluating Retrieval Methods for Improved Question Answering on Building Codes", Smart Construction, Vol. 02, No. 03. DOI: 10.55092/sc20250021

23. Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert et al. (2023), "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback", Transactions on Machine Learning Research, Vol. 2023. DOI: 10.48550/arXiv.2307.15217

24. Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit et al. (2025), "RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs", ACM Computing Surveys, Vol. 58, No. 02. DOI: 10.1145/3743127

25. Aritra Pal, Jacob J. Lin, Fouad Amer, Shang-Hsien Hsieh and Mani Golparvar-Fard (2026), "Automatic Mapping of Schedule Activities and Reality Models for Tracking Construction Progress", Engineering, Construction and Architectural Management, Vol. 33, No. 03, Pp. 1932-1958. DOI: 10.1108/ECAM-08-2024-1166