

Article

PowerXGraph++: Regulation-Constrained Spatiotemporal Evidence Graphs with Counterfactual Explanations for Power-Worksite Safety Monitoring

Qizhe Zhang ^{1,*}, Yi Zhang ¹, Zhengwei Chang ², Zheng Wang ¹ and Guozheng Peng ¹¹ Beijing Key Laboratory of Intelligent Cognition and Analysis for Power Systems, China Electric Power Research Institute, Beijing 100192, China² Power System Security and Operation Key Laboratory of Sichuan Province, State Grid Sichuan Electric Power Research Institute, Chengdu 610041, China

* Correspondence: sxjczqz@163.com or zhangqizhe@epri.sgcc.com.cn

Abstract

Automated power-worksite monitoring requires accurate event recognition and auditable explanations. This paper presents PowerXGraph++, a regulation-constrained spatiotemporal evidence-graph framework for missing PPE, boundary crossing, unsafe proximity, falls, and fire/smoke. Typed evidence tokens are encoded by a transformer; full-context prediction is separated from explanation selection, and explanations are tested for sufficiency, necessity, sparsity, and regulation consistency. Distillation, event-boundary counterfactuals, background-invariance training, and non-inferiority checkpoint selection preserve predictive performance. On annotation-derived construction-PPE graphs, three-seed Macro-F1, violation recall, and rationale F1 were 0.989, 0.965, and 0.901. Under the same frozen predictor, the supervised, grammar-aware selector attained rationale F1 of 0.983 and regulation consistency of 1.000, although the post hoc baselines lacked equivalent supervision. A single YOLO reproduction run using seed 17 on 197 matched person graphs yielded Macro-F1 of 0.844, missing-PPE recall of 0.591, and rationale F1 of 0.660. Expert evaluation produced Fleiss' kappa of 0.740 and method-to-majority rationale F1 of 0.870. Held-out-camera reasoning and a two-camera fence-crossing pilot provide additional evidence under supplied-evidence conditions; the results do not establish automatic generation of every semantic predicate or unrestricted raw video performance.

Keywords: explainable artificial intelligence; evidence-graph reasoning; power-worksite safety; counterfactual explanation; personal protective equipment; fence crossing; industrial surveillance; visual reasoning



Academic Editor: Wangpeng He

Received: 27 July 2026

Revised: 31 August 2026

Accepted: 2 September 2026

Published: 11 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Power-worksite video analytics has expanded from PPE detection and pose-based safety assessment to broader behavior analysis and geometry-aware risk perception [1–4]. The original Detection Transformer (DETR) was introduced by Carion et al. in 2020 [5]. However, a detection score or action label does not show which observations satisfy a safety rule or whether removing them would change the alarm.

This distinction is important in power-worksite supervision. A raised leg is not necessarily a fence-crossing event; a crouching worker is not necessarily falling; steam is not necessarily smoke; and being near de-energized equipment is not equivalent to violating the clearance of energized equipment. Reliable alarms, therefore, require relationships that

evolve over time, such as a person approaching, intersecting, and crossing a boundary. They also require explanations that expose those relationships in an auditable form.

Graphs explicitly represent safety entities, relations, and time. However, the proposed predictor is not a message-passing GNN: it applies a transformer to graph-derived evidence tokens, while graph topology supports construction, provenance, grammar checking, and valid intervention. Thus, the GNNExplainer-, PGExplainer-, and CF-GNNExplainer-style baselines are evidence-token adaptations [6–8], not unmodified GNN methods.

A second challenge is the common accuracy-interpretability trade-off. If a predictor is forced to classify an event using only a small rationale during ordinary inference, evidence omitted by the selector may reduce accuracy. Conversely, if explanations are extracted only after prediction, they may be plausible without being causally faithful. We address this conflict by decoupling the full-context prediction path from the explanation selector, while coupling them again through counterfactual interventions on the same predictor. Following the general teacher–student knowledge-distillation paradigm [9], we additionally apply a validation-set non-inferiority constraint to prevent explanation optimization from selecting a substantially weaker checkpoint.

The contributions of this study are as follows:

- (1) A regulation-constrained spatiotemporal safety evidence graph is introduced to unify PPE compliance, boundary crossing, unsafe proximity, falls, and fire or smoke through event-specific evidence grammars.
- (2) A decoupled prediction-explanation architecture is proposed, in which selected evidence is verified by replaying the same predictor on the selected and complementary subgraphs. Directly measurable sufficiency and necessity are thereby obtained rather than inferred from attention weights alone.
- (3) Knowledge distillation, minimally changed event-boundary counterfactual pairs, background-invariance training, and non-inferiority checkpoint selection are combined to improve explanation quality without sacrificing event accuracy.
- (4) An event-specific evidence-completion request mechanism is formulated to associate incomplete regulation grammars with compatible perception modules. For viewpoint-sensitive fence crossing, a geometry-aware routing strategy is further instantiated and evaluated using a strict leave-one-camera-out protocol.
- (5) A tiered evaluation protocol is established to cover controlled evidence reasoning, matched-input explanation comparison, predicted-evidence error propagation, camera-disjoint real-site testing, and independent expert assessment. Model development and testing are separated by a camera to evaluate transfer under changes in viewpoint and background.

2. Related Work

2.1. Visual Safety Monitoring in Power and Construction Sites

Detector-based methods primarily address PPE compliance and on-site unsafe behavior [1,10,11]. Pose estimators and skeleton-based action models extend recognition to posture and fall cues [2,12–15]. Tracking and video-behavior models introduce temporal continuity [3,16–18], whereas point-cloud, binocular, monocular, and intrusion-detection approaches provide spatial and geometric context [4,19–21]. These studies provide the entities, motion cues, and geometric observations used by PowerXGraph++, which further organizes the available evidence into event graphs and evaluates selected facts through regulation grammars and counterfactual replay.

2.2. Graph Explanations and Counterfactual Faithfulness

Message-passing GNNs use neighborhood aggregation [22,23], whereas PowerX-Graph++ contextualizes typed evidence tokens with a transformer. Post hoc explainers identify influential subgraphs or probabilistic dependencies [6,7,24]. Counterfactual and evaluation frameworks assess prediction-changing perturbations and explanation faithfulness [8,25,26], while high-stakes work motivates directly auditable reasoning [27]. Here, these ideas are adapted to evidence-token masking or deletion rather than original edge-based GNN implementations.

2.3. Selective and Performance-Constrained Reasoning

Selective prediction rejects inadequate-evidence cases to trade coverage for lower risk [28]. In our framework, an incomplete event grammar is converted into an explicit evidence-completion request associated with a compatible perception module. Geometry-aware selection between alternative evidence sources is instantiated for viewpoint-sensitive fence crossing, where projection conditions substantially affect pose and motion reliability.

3. Materials and Methods

3.1. Problem Formulation

Let a video window contain frames $I_{t_0:t_1}$. Perception modules generate tracked entities and pairwise relations. These observations are transformed into an evidence set:

$$\mathcal{E} = \{e_i = (\tau_i, z_i, t_i, k_i)\}_{i=1}^N \quad (1)$$

where τ_i is a discrete evidence token, $z_i \in \mathbb{R}^7$ contains normalized numeric attributes, t_i is the frame index, and k_i is a unique evidence key. The event classifier estimates $p_\theta(y | \mathcal{E})$, where $y \in \mathcal{Y}$ and $\mathcal{Y}$ denotes the safety-event label set:

$$\mathcal{Y} = \{y_s, y_{pp}, y_c, y_{pr}, y_f, y_{fs}\} \quad (2)$$

Here, the six labels correspond to safe operation, missing PPE, safety-boundary crossing, unsafe proximity to energized equipment, worker fall, and fire/smoke, respectively. An explanation selector produces a subset $\mathcal{R} \subseteq \mathcal{E}$. A valid explanation should preserve the event prediction when only $\mathcal{R}$ is retained, reduce the event probability when $\mathcal{R}$ is removed, be sparse, and satisfy the event's regulation grammar.

3.2. Spatio-Temporal Safety Evidence Graph

Each frame is represented by a typed graph $G_t = (V_t, E_t)$. Nodes denote people, PPE items, safety fences, work zones, energized or de-energized equipment, fire, smoke, and scene context. Edges encode relations such as wearing, missing, approaching, intersects, crosses, too-close-to, falls-down, on-ground, and persists. Track edges link the same entity across frames. Evidence extraction converts graph entities and relations into tokens while retaining confidence, time, subject, and object identifiers.

The graph differs from a generic scene graph in two ways. First, relations correspond to safety-relevant predicates rather than all visually possible relations. Second, event definitions specify mandatory groups of alternative tokens. For example, the crossing grammar requires (i) a person, (ii) a safety fence, (iii) approach evidence, (iv) intersection evidence, and (v) a confirmed crossing relation. The PPE grammar requires a person, a missing-equipment entity, and a corresponding missing-equipment relation. The rule engine selects the highest-confidence item in each required group and reports missing groups explicitly.

Predicate provenance differs by experiment. Construction-PPE person, worn-PPE, and missing-equipment labels are converted into person-level facts; an item is assigned when its box center lies inside the person box expanded by 8% in width and height. The same rule is applied without correction to YOLO detections; this PPE task is image-level. The fence pilot derives crossing evidence from pose, motion, and fixed-camera geometry. The real-site experiment organizes site-derived person, object, spatial, motion, and temporal observations into four-frame windows with predicates including missing, approaching, intersects, crosses, too-close-to, falls-down, on-ground, and persists. It tests reasoning over those supplied predicates, not their complete raw video extraction.

Entity and relation evidence use different, but dimensionally aligned, numeric encodings. For an entity with a normalized box center (x, y) , width w , height h , confidence c , and relative time $\bar{t}$, the numeric vector is

$$z^{ent} = [c, x, y, w, h, \bar{t}, 0] \quad (3)$$

For a relation, the corresponding vector is

$$z^{rel} = [c, d, o, m, \Delta, \bar{t}, 1] \quad (4)$$

where d , o , m , and Δ denote normalized distance, overlap, motion, and duration when defined. Unused attributes are zero. The final binary component distinguishes entity from relation evidence without requiring a second encoder. Every evidence key identifies its frame, graph element type, and participating entities, so an intervention can remove a fact deterministically.

Graph intervention also preserves structural validity. When an entity is removed, relations whose subject or object no longer exists are removed with it. This prevents a counterfactual subgraph from retaining a semantically impossible relation, such as crosses-safety boundary, without either a person or a fence. In contrast, background context entities may be removed independently, which makes it possible to test whether a prediction depends on a causal safety relation or on a site-specific visual correlate.

Figure 1 summarizes four stages: (1) available perception or annotation inputs form typed evidence with confidence, time, and subject-object identifiers; (2) a transformer encodes tokens for separate prediction and rationale queries; (3) the same predictor re-plays selected and complementary evidence to measure sufficiency and necessity; and (4) a grammar gate checks mandatory groups and supports completion requests, routing, and audit records. The figure is architectural; not every proposed perception route was experimentally validated.

3.3. Decoupled Predictor and Explanation Selector

The graph and encoder serve distinct roles. Typed entities, subject-object links, track edges, and dependencies govern evidence construction and structurally valid deletion; no adjacency matrix or message-passing operator enters the predictor. Entity or relation identity, attributes, time, and identifiers are instead contextualized by transformer self-attention. PowerXGraph++ is, therefore, graph-structured evidence reasoning, not a message-passing GNN.

Evidence tokens are embedded and added to a learned projection of the seven numeric attributes. A two-layer transformer encoder with hidden dimension 96 and four attention heads contextualizes the evidence sequence. The predictor and selector use different class queries:

$$H = \text{Enc}_{\theta}(\text{Emb}(\tau) + \text{MLP}(z)) \quad (5)$$

$$a_y^p = \text{softmax}(Hq_y^p / \sqrt{d}) \quad (6)$$

$$a_y^r = \text{softmax}(Hq_y^r / \sqrt{d}) \quad (7)$$

Let $v_{iy} = w_y^\top h_i$ be the class-specific evidence value. The event logit is

$$\ell_y = \sum_{i=1}^N a_{iy}^p v_{iy} + b_y \quad (8)$$

In Equations (5)–(8), Enc_θ is the transformer encoder; $\text{Emb}(\tau)$, and $\text{MLP}(z)$ encode the discrete token and seven numeric attributes; $H = [h_1, \dots, h_N]$ is the contextualized sequence; q_y^p and q_y^r are prediction and rationale queries; d is the query dimension; a_y^p and a_y^r are normalized attention weights; and w_y and b_y parameterize class evidence values. In Equations (9)–(12), $\mathcal{R}^*$ is the target rationale; $\mathcal{R} \subseteq \mathcal{E}$ is the explanation selected from the full evidence set $\mathcal{E}$; $\mathcal{E} \setminus \mathcal{R}$ is its complement; and $p_y(S)$ is the probability of class y when the predictor is evaluated on evidence set S .

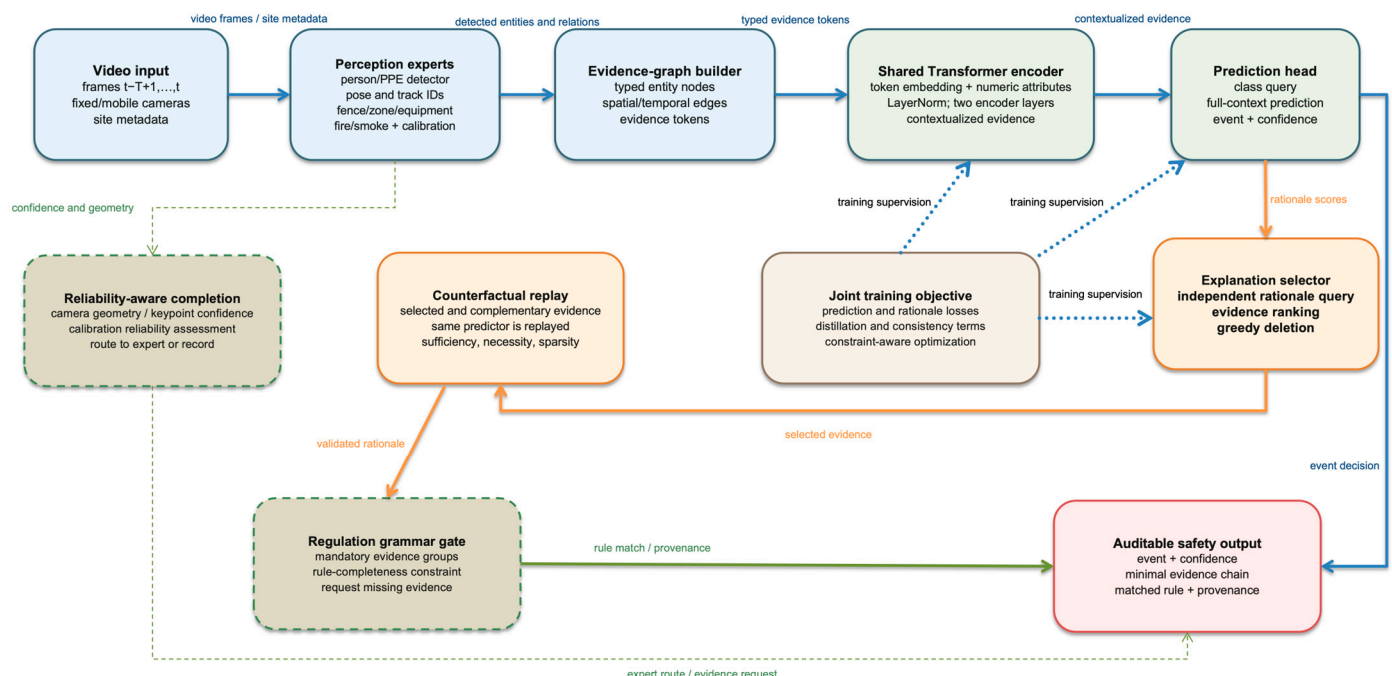


Figure 1. PowerXGraph++ architecture. The upper horizontal path constructs typed evidence tokens from video inputs and perception outputs and performs full-context event prediction. The lower branches show reliability-aware evidence completion and routing, rationale selection, counterfactual replay on selected and complementary evidence, and regulation-grammar checking before auditable output. The joint training objective is connected to the encoder, prediction head, and explanation selector through auxiliary training links. Solid arrows denote inference or verification flow, whereas dashed arrows denote training or reliability-routing dependencies.

Only prediction attention a_y^p contributes to ℓ_y . Explanation attention a_y^r ranks evidence but cannot replace the full-context prediction path. This separation is important because the smallest human-readable rationale is not necessarily the complete feature set needed for robust classification. The shared encoder still allows rationale supervision to shape representations, while different queries prevent the selector from becoming an information bottleneck for the predictor.

For samples with rule-grounded rationale indices $\mathcal{R}^*$, the normalized target distribution is uniform over the annotated rationale. The selector is trained by

$$\mathcal{L}_{\text{rationale}} = -\frac{1}{|\mathcal{R}^*|} \sum_{i \in \mathcal{R}^*} \log a_{iy}^r \quad (9)$$

Safe samples have no positive rationale target and therefore do not contribute to this term. This prevents an arbitrary “safe explanation” from being learned when no violation grammar is complete.

For inference-time explanation, a greedy deletion procedure starts from all evidence and removes low-ranked items while the predicted class remains unchanged and its probability deviates from the full-graph value by no more than 0.08. Mandatory regulation evidence cannot be deleted. This yields a sparse selected set $\mathcal{R}$.

3.4. Counterfactual Verification

The same predictor is evaluated on the full graph, the selected evidence, and the complement. Let $p_y(\mathcal{E})$ be the probability of the predicted event. We measure

$$\text{Gap}_{\text{suff}} = |p_y(\mathcal{E}) - p_y(\mathcal{R})| \quad (10)$$

$$\text{Drop}_{\text{nec}} = p_y(\mathcal{E}) - p_y(\mathcal{E} \setminus \mathcal{R}) \quad (11)$$

$$\text{Sparsity} = 1 - \frac{|\mathcal{R}|}{|\mathcal{E}|} \quad (12)$$

Sufficiency gap measures probability loss when only the explanation is retained; necessity drop measures loss when it is removed; regulation consistency is mandatory group coverage. These measures assess the supplied graph, not the independent correctness of upstream predicates.

The deletion search is prediction-preserving rather than score-maximizing. Starting from $\mathcal{R}_0 = \mathcal{E}$, candidates are processed from low to high explanation attention. Candidate e_j is removed only if

$$\arg \max_{y'} p(y' | \mathcal{R}_s \setminus e_j) = y \text{ and } p_y(\mathcal{R}_s \setminus e_j) \geq p_y(\mathcal{E}) - \delta \quad (13)$$

with $\delta = 0.08$. Mandatory evidence keys supplied by the rule engine are protected. After deletion terminates, the complement $\mathcal{E} \setminus \mathcal{R}$ is rescored. The output record contains the full, selected, and complement probabilities; selected keys and tokens; rule consistency; and missing evidence groups. Consequently, every reported explanation metric can be recomputed from the stored prediction record.

3.5. Performance Preservation and Event-Boundary Counterfactuals

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{event}} + \lambda_{\text{KD}} \mathcal{L}_{\text{KD}} + \lambda_r \mathcal{L}_{\text{rationale}} + \lambda_{\text{cf}} (\mathcal{L}_{\text{selected}} + \mathcal{L}_{\text{suff}} + \mathcal{L}_{\text{nec}}) + \lambda_p \mathcal{L}_{\text{pair}} + \lambda_b \mathcal{L}_{\text{background}} \quad (14)$$

The event loss trains the full-context predictor. Knowledge distillation uses temperature $T = 2$ and minimizes the Kullback–Leibler divergence between student and frozen teacher distributions. Rationale supervision aligns the selector with rule-grounded evidence. During training, rule-grounded selected and complementary masks provide a stable counterfactual target before greedy inference is used. The selected mask is trained to

classify the correct event and match the full-graph probability; the complement is required to reduce the event probability by a margin of 0.2.

The pair loss uses minimally changed examples located at a safety-event boundary: approaching versus crossing a fence, correctly worn versus missing PPE, crouching versus falling, safe versus unsafe equipment distance, and steam or transient light versus persistent smoke/fire. Removing the decisive boundary fact creates a safe target and enforces a probability separation of at least 0.25 from the original event. Background invariance changes site-specific distractors without changing the event label and minimizes the divergence between the original and intervened predictions. These two losses have different purposes: the pair loss makes the predictor sensitive to causal boundary evidence, whereas background invariance makes it insensitive to non-causal context.

Checkpoint selection is subject to a non-inferiority constraint. If F_{val}^T is the validation Macro-F1 of the teacher, candidate checkpoint k is eligible only if

$$F_{val}^{(k)} \geq F_{val}^T - \epsilon \quad (15)$$

where $\epsilon = 0.005$. Among eligible checkpoints, the model with the best explanation objective is retained. This makes the accuracy requirement explicit rather than hoping that an explanation regularizer does not degrade the classifier.

3.6. Regulation Grammar and Evidence-Completion Requests

The grammar is represented as mandatory groups of alternative tokens rather than a single rigid string. An event is grounded only when every group contains at least one observed fact. Alternatives allow, for example, any supported missing-PPE category or either fire or smoke to instantiate the same higher-level rule. Table 1 lists the implemented engineering grammar and the minimally changed negative used during counterfactual training.

Table 1. Event-specific evidence grammar and minimally changed negative cases. The grammar is an engineering event definition and must be reviewed against the applicable site procedure before deployment.

Event	Mandatory Evidence Groups	Boundary Negative
Missing PPE	person; a missing-PPE entity; corresponding missing relation	same person with worn PPE or without a grounded missing relation
Boundary crossing	person; safety fence; approaching; intersection; confirmed side transition	approaching or touching the fence without a confirmed transition
Unsafe proximity	person; energized equipment; approach; below-threshold distance relation	proximity to de-energized equipment or distance above the threshold
Worker fall	person; rapid fall relation; persistent on-ground relation	crouching, kneeling, or a transient low posture
Fire or smoke	person/site context; fire or smoke entity; temporal persistence	steam, glare, or a non-persistent plume

Table 1 is an author-defined engineering grammar, not a certified safety standard. Training rationale indices are evidence keys instantiating their mandatory groups. Separately, three power-safety or field-operation experts who did not develop the model selected evidence keys independently while blinded to method identity and importance scores; their majority labels were not used for training, checkpoint selection, or budget tuning.

For example, a crossing graph may contain person (P), fence (F), approaching (P, F, t1), intersects (P, F, t2), crosses (P, F, t3), and background. A complete rationale covers the first five groups; deleting P also removes its incident relations. Counterfactual replay tests the

retained and complementary sets, and expert evaluation compares selected keys with the independent majority reference.

When a predicted event lacks a mandatory evidence group, the missing group is converted into an explicit evidence-completion request rather than being inferred implicitly. Requests for body-state evidence are associated with pose estimation, crossing evidence with boundary geometry, proximity evidence with calibrated distance estimation, and persistence evidence with temporal segmentation. Each request records the missing group, the candidate perception module, and the available reliability metadata. When additional evidence is returned, its confidence and provenance are appended to the evidence graph. This representation distinguishes incomplete observation from a rule-grounded negative decision.

3.7. Geometry-Aware Expert Routing

Some evidence is intrinsically viewpoint-dependent. We instantiate the evidence-completion routing mechanism for fence-crossing analysis, where the reliability of pose and geometric evidence varies substantially with camera viewpoint. In a frontal view, two legs may overlap in the image, making 2D angle or ankle-asymmetry evidence unreliable. In a side view, the same cues are more informative. We therefore associate each fixed camera with auditable reliability metadata. The router chooses either a conservative temporal expert or the learned evidence-fusion expert before test labels are observed. The route and its reason are stored with the explanation.

In the private crossing pilot, frontal footage uses the conservative stabilized leg-lifting rule, while side-view footage uses evidence fusion over leg angle, ankle asymmetry, fence proximity, and temporal consistency. This routing policy provides a controlled implementation of the request-and-routing mechanism and examines whether camera-specific reliability metadata can prevent the known cross-view failure of unrestricted evidence fusion.

4. Experimental Setup

4.1. Datasets and Evidence Levels

Table 2 separates four evidence levels: annotation-derived PPE reasoning, matched detector-derived PPE evaluation, held-out-camera reasoning over supplied multi-hazard evidence, and a two-camera crossing pilot. Their claims are not interchangeable.

The construction-PPE dataset contains 11 classes, including people, worn PPE, and explicit missing-equipment categories [29]. The detector is evaluated on the official test split. For the reasoning experiment, people and PPE ground-truth boxes are converted into person-level graphs. Before balancing, the training split contains 1370 safe-person graphs and 400 missing-PPE graphs; training uses 400 graphs per class. The official test split contains 141 images and 236 annotation-derived person graphs, comprising 189 safe graphs and 47 missing-PPE graphs.

The real-site dataset contains six classes of four-frame windows: safe, missing PPE, boundary crossing, unsafe proximity, worker fall, and fire/smoke. Each class has 80/20/40 training/validation/test windows (480/120/240 total). Cameras a and b supply development data; unseen camera c supplies all test data and is excluded from fitting, hyperparameter selection, and checkpoint selection.

Available aggregates do not separate camera-a and camera-b counts or enumerate source videos, workers, physical events, continuous intervals, stride, or overlap. Thus, camera-c separation is documented, but event or sequence independence between training and validation is not claimed.

Table 2. Datasets and the claim supported by each experiment.

Dataset	Size and Split	Evidence Source	Purpose
Construction-PPE [29]	1416 images; 1132/143/141 train/validation/test	YOLOv11n predictions for perception; annotation-derived person/PPE evidence for controlled reasoning	Public PPE detection, annotation-derived reasoning, and matched predicted-box error propagation
Real-site camera-disjoint multi-hazard dataset	480/120/240 train/validation/test four-frame evidence windows; six classes; camera-a/b to unseen camera-c	Site-derived person, object, spatial, and temporal evidence	Held-out-camera reasoning over supplied multi-hazard evidence graphs
Private fence-crossing pilot	Two cameras, 1032 video frames, 1549 person-frame labels, six action intervals	Local pose metrics and fixed-camera fence annotations	Leave-one-camera-out crossing pilot; raw media kept local
Independent expert rationale set	180 stratified cases; 30 per event class; three independent raters	Minimal evidence chains marked directly on stored evidence windows	Inter-rater agreement and method-to-expert rationale alignment

The confidential pilot comprises a 20.47 s side-view video at 1280×720 and a 13.93 s frontal video at 1920×1080 , both at 30 frames/s. It contains 126 positive and 1423 negative person-frame labels forming six contiguous action intervals. Raw videos were neither uploaded nor copied into the manuscript project. Evaluation holds out an entire camera/video, and all thresholds are selected on the other video.

Three model-independent power-safety or field-operation experts assessed 180 cases (30 per class), blinded to method identity and importance scores. Evidence-key majority votes form the evaluation reference and are unused in training or selection.

4.2. Leakage Controls and Evaluation Protocol

Leakage controls are the official construction-PPE train/test partitions; validation-only reasoner and non-inferiority checkpoint selection; and leave-one-camera-out crossing-pilot tuning, with routing fixed by viewpoint metadata rather than held-out labels. These controls do not prove target independence of semantic predicates.

Accordingly, the real-site claim is transferred from development cameras a and b to unseen camera c under a common ontology and supplied evidence, not event-disjoint raw video generalization.

In controlled PPE reasoning, official annotations provide person and worn or missing-equipment entities and rule-grounded rationale keys; image pixels are not input. Because missing-equipment labels are close to the target, results are conditional on annotated evidence. The predicted-box experiment instead applies the same graph rules to uncorrected YOLO outputs. Training is class-balanced; the test distribution is unchanged.

Semantic predicates may encode useful domain knowledge and target-correlated information. No lower-level-only ablation was available, so the manuscript does not claim that missing, crosses, too-close-to, falls-down, on-ground, or persists are generated label-independently for every class.

The explanation baselines apply evidence-token adaptations of GNNExplainer, PGExplainer, and CF-GNNExplainer to the same frozen predictor, candidate graphs, and correctly classified violations. GNNExplainer optimizes instance masks; PGExplainer learns a training-set mask generator; CF-GNNExplainer searches structurally valid prediction-changing deletions. Removing an entity also removes incident relations. Unlike PowerXGraph++, none receives rationale labels, mandatory-evidence protection, or grammar constraints; the comparison is matched on inputs and predictor, not supervision.

4.3. Implementation Details

Experiments were executed on Windows with Python 3.12, PyTorch 2.13.0+cu130, Ultralytics 8.4.95, and an NVIDIA GeForce RTX 3080 Ti Laptop GPU with 16 GB of memory. The YOLOv11n construction-PPE detector used 640×640 inputs, a batch size of 16, random seed 17, deterministic execution, weight decay of 5×10^{-44} , and 45 training epochs; the 10-epoch checkpoint is retained as an optimization reference. Predicted-box evaluation used a detection-confidence threshold of 0.25, a nonmaximum-suppression IoU threshold of 0.70, and one-to-one person-box matching at $\text{IoU} \geq 0.50$. Evidence-reasoning experiments use the fixed public split and do not tune on the test set.

The evidence model was trained for 25 epochs with batch size 64 and learning rate 3×10^{-4} . The rationale, counterfactual, distillation, event-boundary, and synthetic-background weights were 1.0, 0.2, 0.5, 0.5, and 0.2, respectively; background loss was disabled for construction-PPE. Results are reported for seeds 17, 23, and 41 as mean $\pm$ standard deviation. Up to 100 test explanations per run were counterfactually evaluated.

The PGExplainer-style mask generator uses a two-layer multilayer perceptron to produce instance-specific evidence-selection probabilities from contextualized evidence representations. It is trained for 140 epochs using Adam with a learning rate of 3×10^{-3} , weight decay of 10^{-5} , a mask-size coefficient of 0.035, and an entropy coefficient of 0.005. Five initializations are trained for each frozen checkpoint, and model selection is based on validation prediction fidelity without access to validation or test rationale identities.

The CF-GNNExplainer-style baseline adapts the counterfactual principle of CF-GNNExplainer [8] to predictor-specific evidence tokens rather than its original edge-perturbation representation. Per-instance continuous deletion masks are optimized for 100 steps with a learning rate of 0.12, followed by structurally valid discrete deletion and redundancy pruning. Schematically, $L_{CF} = L_{\text{flip}} + 0.04L_{\text{size}} + 0.002L_{\text{entropy}}$, where L_{flip} promotes a change in the original predicted class, L_{size} controls evidence-deletion size, and L_{entropy} regularizes mask entropy. These coefficients and optimization settings are implementation choices, not universal constants of the original method. Removing an entity also removes its incident relations; a counterfactual is successful only when the retained evidence changes the original predicted class.

The private pilot used monotonic evidence fusion, a training-selected decision threshold, and predeclared viewpoint routing. Scene-level event post-processing merged gaps of at most eight frames and removed alarms supported by fewer than three frames. This protocol reduces sensitivity to short track-ID fragmentation.

All evidence models use a 96-dimensional hidden representation, four attention heads, two transformer encoder layers, feed-forward width 192, dropout 0.1, AdamW optimization with a learning rate of 3×10^{-4} and weight decay of 1×10^{-4} , and gradient-norm clipping at 5.0. Evidence sequences are padded within a mini-batch, and masking prevents padded facts from contributing to either query. The public PPE test graphs contain 10.05 evidence items on average (median 9, range 1–29); real-site held-out-camera test graphs contain 14.76 items on average (median 14, range 12–18).

For runtime measurement, 40 warm-up iterations were followed by 200 synchronized CUDA forward passes using sequence length 16. Greedy explanation latency was measured separately on 100 real-site cross-camera test events because it requires repeated prediction-preserving deletion and cannot be represented by one forward pass. Perception, tracking, and graph-construction latency are excluded from the reasoner benchmark and are reported separately where available.

4.4. Evaluation Metrics

Perception is measured by precision, recall, mAP_{50} , $mAP_{50:95}$, and inference time. Event prediction is measured by Macro-F1 and violation recall. The private pilot additionally reports frame-level precision, recall, and F1, together with event-level precision, recall, and F1.

Explanation selection is compared with rule-grounded rationale labels using rationale F1. Regulation consistency measures the completion of required evidence groups. Sufficiency gap and necessity drop follow Equations (10) and (11). Counterfactual rule consistency measures whether minimally changed test cases receive explanations that satisfy the corresponding alternative event grammar. For independent human assessment, Fleiss' κ quantifies three-rater agreement, and method-to-expert rationale F1 measures overlap with the majority-vote evidence chain.

For CF-GNNExplainer-style, the structurally valid deleted evidence set is treated as the returned counterfactual explanation. Necessity drop and sparsity are directly aligned with its prediction-changing objective, whereas rationale F1, regulation consistency, and sufficiency gap are additionally reported to maintain a uniform comparison with factual explanation selectors. Counterfactual success rate measures the proportion of evaluated samples for which the deletion changes the original predicted class.

Sparsity is reported with faithfulness because a small but incomplete explanation is not desirable; The explanation comparisons therefore pair sparsity with sufficiency, necessity, and regulation consistency. In the result tables, the column headers are annotated with directional arrows to indicate the preferred trend for each metric: $\downarrow$ denotes that lower values are better, whereas $\uparrow$ denotes that higher values are better.

4.5. Statistical Reporting and Reproducibility

Primary reasoner results use seeds 17, 23, and 41; means and population standard deviations are reported, with paired seed-wise directions serving as a small-sample check. The same frozen checkpoints are reused in the explanation comparison. Expert agreement uses the common 180-case set. The detector-derived PPE results are from a single reproduction run using seed 17 on 197 matched person graphs. Because raw site videos are restricted, reproducibility is limited to the reported aggregate definitions, configurations, split totals, and de-identified outputs.

5. Results

5.1. Optimization Behavior

Figure 2 reports optimization traces from the retained experiment logs. For YOLOv11n, the sum of training box, classification, and distribution-focal losses decreases from 7.514 to 4.390 over 10 epochs; the corresponding validation sum decreases from 7.159 to 4.783. Validation mAP_{50} rises from 0.202 to 0.559 and $mAP_{50:95}$ from 0.102 to 0.284. Precision is less monotonic because several rare negative-PPE categories contain few examples, but recall and both mAP measures improve overall.

For the complete evidence reasoner, the three-seed mean composite loss falls from 2.976 to 1.295 on public PPE graphs. Validation Macro-F1 is already 1.000 in the validation subset, while rationale F1 increases from 0.832 to 0.965. On the real-site cross-camera dataset, the mean loss falls from 6.578 to 2.137, the validation Macro-F1 rises from 0.048 to 1.000, and the rationale F1 rises from 0.293 to 0.687. The nonzero final loss is expected because rationale, distillation, sufficiency, necessity, pair-margin, and background-invariance terms remain active after classification convergence.

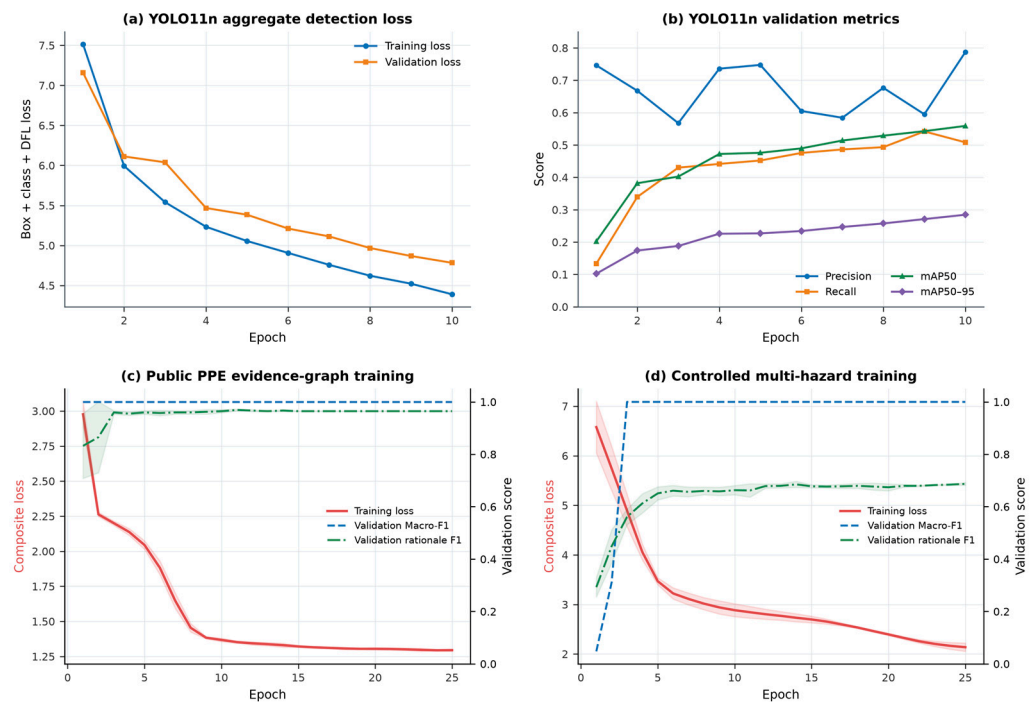


Figure 2. Training dynamics from retained logs. (a) Aggregate YOLOv11n training and validation losses. (b) YOLOv11n validation detection metrics. (c) PowerXGraph++ results on public PPE evidence graphs, averaged over three random seeds. (d) The corresponding real-site multi-hazard results, averaged over three random seeds. In panels (c,d), each curve represents the mean across the three random seeds, and the matching light-colored shaded band represents the mean $\pm$ one standard deviation. The graph-model loss is a composite objective and is not on the same scale as the detection loss.

5.2. Public PPE Perception Baseline

Table 3 reports both the 10-epoch reference and the retained 45-epoch checkpoint. Longer training increases recall from 0.492 to 0.560 and mAP50 from 0.521 to 0.565, while mAP50:95 changes from 0.266 to 0.280. This modest aggregate gain masks larger relative improvements in rare missing-PPE categories.

Table 3. Construction-PPE YOLOv11n perception results at 10 and 45 epochs.

Metric	10 Epochs	45 Epochs
Precision	0.648	0.627
Recall	0.492	0.560
mAP50	0.521	0.565
mAP50:95	0.266	0.280
Inference time	2.9 ms/image	2.6 ms/image

Table 4 reports the complete 10-epoch per-class baseline. For the four rare missing-PPE categories discussed here, the corresponding 45-epoch mAP50:95 values are also stated in the text: no helmet, no goggles, no gloves, and no boots improve from 0.061, 0.038, 0.018, and 0.005 to 0.092, 0.079, 0.050, and 0.018, respectively. Although longer training improves these categories, their absolute values remain low, motivating the separation between perception-derived and annotation-derived reasoning experiments.

Table 4. Per-class construction-PPE test mAP50:95 for the complete 10-epoch YOLOv11n baseline; selected 45-epoch rare-class values are reported in the accompanying text.

Class	mAP	Class	mAP	Class	mAP
Helmet	0.515	Gloves	0.374	Vest	0.563
Boots	0.386	Goggles	0.304	Person	0.508
No helmet	0.061	No goggles	0.038	No gloves	0.018
No boots	0.005	None/other	0.150		

5.3. Construction-PPE Explanation Ablation

The experiments in this subsection use annotation-derived evidence graphs to control upstream perception variables and isolate the contribution of each reasoning component. They are not intended to represent the end-to-end performance of the complete video-monitoring pipeline. The propagation of realistic detector errors is evaluated separately in Section 5.10. Table 5 compares five variants. The classification teacher is accuracy-oriented and provides attention as an unconstrained explanation. Decoupled-rule separates prediction and explanation queries. Decoupled-counterfactual adds intervention loss. Performance-constrained training adds distillation and the validation Macro-F1 floor. The complete PowerXGraph++ additionally uses event-boundary pairs.

Table 5. Construction-PPE person-level graph results over three seeds (mean $\pm$ standard deviation). Evidence graphs in this table are derived from ground-truth annotations.

Variant	Macro-F1	Violation Recall	Rationale F1	CF Rule Consistency	Suff. Gap $\downarrow$	Nec. Drop $\uparrow$
Classification teacher	0.975 $\pm$ 0.008	0.922 $\pm$ 0.025	0.213 $\pm$ 0.125	0.360 $\pm$ 0.046	0.070 $\pm$ 0.023	0.048 $\pm$ 0.048
Decoupled + rule	0.977 $\pm$ 0.008	0.929 $\pm$ 0.025	0.856 $\pm$ 0.022	0.338 $\pm$ 0.004	0.075 $\pm$ 0.057	0.037 $\pm$ 0.014
Decoupled + counterfactual	0.975 $\pm$ 0.004	0.922 $\pm$ 0.012	0.868 $\pm$ 0.015	0.341 $\pm$ 0.008	0.089 $\pm$ 0.073	0.039 $\pm$ 0.028
+performance constraint	0.982 $\pm$ 0.004	0.943 $\pm$ 0.012	0.877 $\pm$ 0.015	1.000 $\pm$ 0.000	0.039 $\pm$ 0.008	0.211 $\pm$ 0.093
PowerXGraph++	0.989 $\pm$ 0.010	0.965 $\pm$ 0.033	0.901 $\pm$ 0.033	1.000 $\pm$ 0.000	0.037 $\pm$ 0.016	0.272 $\pm$ 0.034

Relative to the teacher, PowerXGraph++ improves Macro-F1 by 1.4 percentage points and violation recall by 4.3 points. Rationale F1 increases by 68.8 points. The sufficiency gap falls from 0.070 to 0.037, while the necessity drop increases from 0.048 to 0.272. The counterfactual and rule-only variants improve rationale matching but do not make the selected evidence strongly necessary. Most of the necessity improvements appear only after distillation, the performance floor, and constrained selection are used together.

The direction of the final accuracy change is consistent in all three paired runs: teacher-to-PowerXGraph++ Macro-F1 changes are +0.020, +0.014, and +0.007 for seeds 17, 23, and 41. Rationale F1 changes are +0.858, +0.603, and +0.603, and necessity drop changes are +0.282, +0.154, and +0.237. Thus, the mean gain is not produced by one unusually favorable seed. Rule-only and unconstrained counterfactual variants remain highly sparse (approximately 0.88) but have low necessity, illustrating why sparsity alone is insufficient evidence of faithfulness.

Figure 3 reorganizes the ablation into two complementary views. The heatmap direction aligns all metrics so that brighter cells indicate a more desirable outcome, while retaining the original means as cell annotations. The Pareto view then isolates the prediction-faithfulness trade-off. Only the performance-constrained and complete variants simultaneously occupy the high-Macro-F1 and high-necessity region, and the event-boundary

pairs move the complete model beyond the performance-constrained checkpoint without violating the non-inferiority floor.

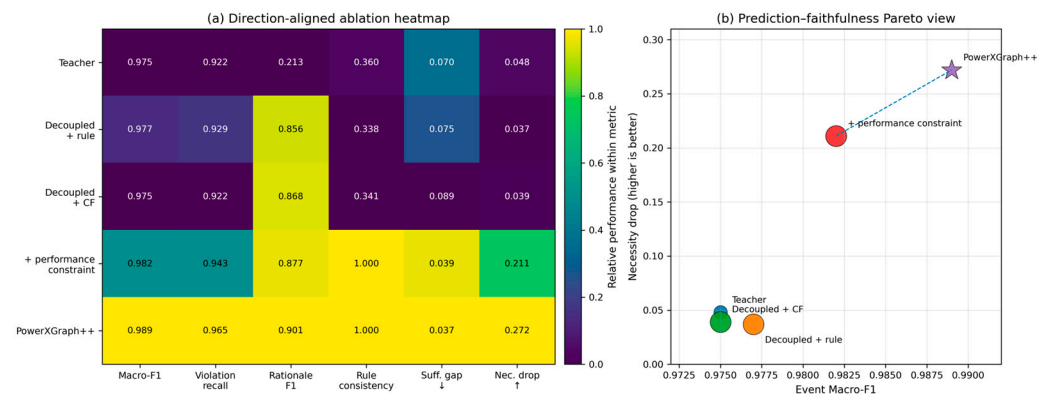


Figure 3. Multi-view ablation analysis on construction-PPE evidence graphs. (a) Direction-aligned heatmap across prediction and explanation metrics; colors are normalized within each metric, whereas the annotated values are the original three-seed means. (b) Prediction-faithfulness Pareto view; the marker area is proportional to rationale F1. The upper-right direction is preferred.

5.4. Same-Predictor Explanation Baselines

Table 6 replays identical retained predictors on identical correctly classified violation graphs. Fixed-budget methods use grammar-group counts; for construction-PPE, $k = 3$, and random- k is averaged over 20 draws. GNNExplainer-style uses instance masks and PGExplainer-style a shared training set generator, both truncated to this budget.

Table 6. Matched-input explanation comparison under the same frozen predictor. Values are mean $\pm$ standard deviation over the three retained checkpoints.

Method	Rationale F1	Rule Consistency	Suff. Gap ↓	Nec. Drop ↑	Sparsity ↑
Random-k	0.330 $\pm$ 0.004	0.652 $\pm$ 0.006	0.140 $\pm$ 0.013	0.031 $\pm$ 0.014	0.675 $\pm$ 0.005
Attention top-k	0.916 $\pm$ 0.014	0.929 $\pm$ 0.006	0.038 $\pm$ 0.012	0.270 $\pm$ 0.037	0.675 $\pm$ 0.005
GNNExplainer-style mask [6]	0.780 $\pm$ 0.005	0.750 $\pm$ 0.011	0.050 $\pm$ 0.008	0.205 $\pm$ 0.003	0.675 $\pm$ 0.012
PGExplainer-style mask [7]	0.909 $\pm$ 0.007	0.921 $\pm$ 0.007	0.044 $\pm$ 0.012	0.265 $\pm$ 0.046	0.675 $\pm$ 0.005
CF-GNNExplainer-style deletion [8]	0.236 $\pm$ 0.202	0.593 $\pm$ 0.133	0.228 $\pm$ 0.216	0.256 $\pm$ 0.288	0.505 $\pm$ 0.146
Occlusion top-k	0.702 $\pm$ 0.072	0.741 $\pm$ 0.078	0.047 $\pm$ 0.010	0.222 $\pm$ 0.035	0.675 $\pm$ 0.005
Greedy-Preserve	0.385 $\pm$ 0.027	0.343 $\pm$ 0.008	0.047 $\pm$ 0.009	0.038 $\pm$ 0.011	0.886 $\pm$ 0.004
PowerXGraph++	0.983 $\pm$ 0.018	1.000 $\pm$ 0.000	0.037 $\pm$ 0.016	0.272 $\pm$ 0.034	0.666 $\pm$ 0.007

Across events, the budgets are $k = 3$ for missing PPE, $k = 5$ for crossing, $k = 4$ for unsafe proximity, $k = 3$ for fall, and $k = 3$ for fire/smoke; safe cases are excluded. PGExplainer selection uses validation fidelity without rationale labels. Equal budgets do not remove PowerXGraph++'s additional rationale supervision and grammar access.

CF-GNNExplainer-style instead searches for a minimal structurally valid deletion that changes the original predicted class. Occlusion top- k ranks evidence according to the probability decrease caused by deleting one item, while Greedy-Preserve performs prediction-preserving deletion within the 0.08 tolerance. Thus, the comparison covers random, attention-based, mask-based, perturbation-based, prediction-preserving, and prediction-changing explanation strategies while keeping the predictor and candidate evidence fixed.

Attention top- k obtains a necessity drop of 0.270, close to the 0.272 of PowerXGraph++, but its regulation consistency remains 0.929. PGExplainer-style also identifies highly predictive evidence, reaching a rationale F1 of 0.909, regulation consistency of 0.921, and

necessity drop of 0.265. These results indicate that parameterized prediction-fidelity masks can recover much of the decision-relevant evidence but do not consistently produce a complete regulation chain.

CF-GNNExplainer-style obtains a necessity drop of 0.256 but shows substantially larger variation across checkpoints. Its rationale F1 of 0.236 and regulation consistency of 0.593 indicate that a minimal prediction-changing deletion is not necessarily equivalent to a complete factual rationale. The large standard deviations further reflect sensitivity to the predictor checkpoint and the discrete perturbation path.

PowerXGraph++ reaches rationale F1 of 0.983, regulation consistency of 1.000, sufficiency gap of 0.037, and necessity drop of 0.272 under its supervised, grammar-constrained setting. Table 6, therefore, measures the contribution of this domain information, not supervision-matched universal superiority over original graph explainers.

5.5. Camera-Disjoint Real-Site Multi-Hazard Evaluation

On the six-class camera-c test evidence, PowerXGraph++ attains Macro-F1 of 1.000 and rationale F1 of 0.666 (Table 7). Because the inputs include site-derived semantic predicates, these are held-out-camera reasoning results over supplied evidence, not unrestricted raw video recognition.

Table 7. Real-site camera-a/b to unseen camera-c multi-hazard validation over three seeds.

Variant	Macro-F1	Rationale F1	Rule Consistency	Suff. Gap ↓	Nec. Drop ↑
Classification teacher	0.980 ± 0.017	0.255 ± 0.025	0.316 ± 0.003	0.184 ± 0.079	0.234 ± 0.084
Decoupled + rule	0.999 ± 0.002	0.626 ± 0.015	0.294 ± 0.012	0.086 ± 0.021	0.087 ± 0.015
Decoupled + counterfactual	0.999 ± 0.002	0.594 ± 0.040	0.286 ± 0.006	0.115 ± 0.025	0.103 ± 0.008
+performance constraint	1.000 ± 0.000	0.666 ± 0.061	1.000 ± 0.000	0.090 ± 0.067	0.438 ± 0.066
PowerXGraph++	1.000 ± 0.000	0.666 ± 0.020	1.000 ± 0.000	0.082 ± 0.083	0.726 ± 0.074

Event-wise decomposition in Table 8 uses the 300 stored PowerXGraph++ explanations across three seeds. Rule consistency is 1.000 for every hazard. Boundary crossing and unsafe proximity have necessity drops above 0.90 because their decisive transitions or distance relations are difficult to reconstruct from the complement. Fire/smoke and missing-PPE necessity is lower, approximately 0.46 and 0.41, because repeated entity observations or correlated context may retain partial event probability after the minimal grounded chain is removed. This identifies an actionable limitation that is hidden by the aggregate score.

Table 8. Event-wise real-site cross-camera explanation analysis pooled over 300 explanations from three seeds.

Event	Samples	Suff. Gap ↓	Nec. Drop ↑	Sparsity ↑	Rule Consist.
Missing PPE	56	0.077	0.410	0.786	1.000
Boundary crossing	67	0.079	0.908	0.691	1.000
Unsafe proximity	64	0.078	0.920	0.765	1.000
Worker fall	56	0.086	0.875	0.751	1.000
Fire or smoke	57	0.089	0.460	0.786	1.000

Figure 4 complements the aggregate cross-camera result with an event-wise diagnostic profile. Boundary crossing, unsafe proximity, and worker fall retain necessity drops above 0.87, whereas missing-PPE and fire/smoke remain below 0.50 despite complete rule consistency. The profile, therefore, distinguishes events whose decisive relation is difficult to reconstruct from complements of events with repeated entities or correlated contexts that can partially preserve the target probability.

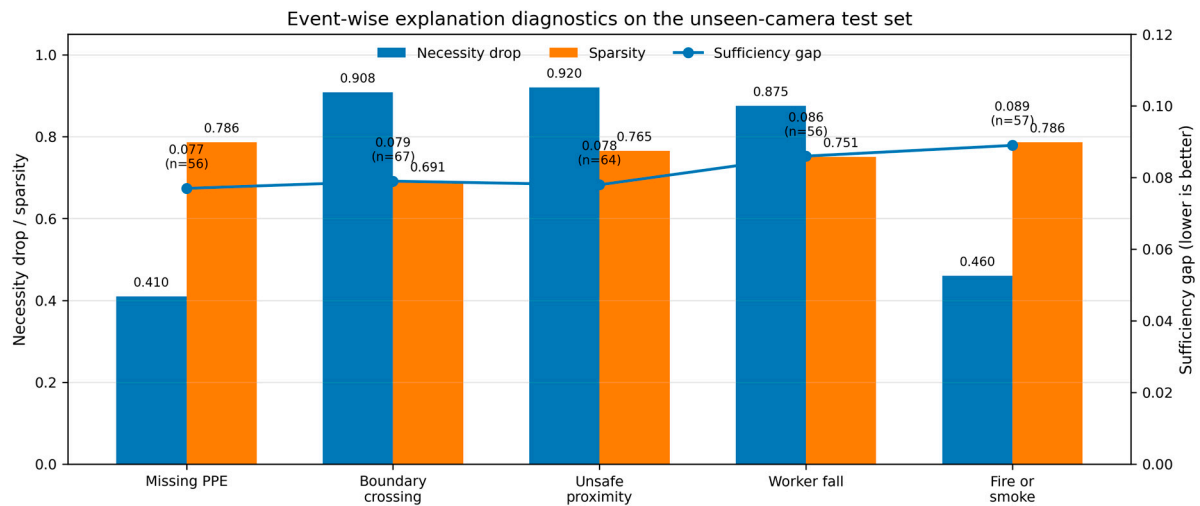


Figure 4. Event-wise explanation diagnostics on the unseen-camera test set. Bars report necessity drop and sparsity, and the line reports sufficiency gap. Sample counts are pooled over 300 stored explanations from three seeds; regulation consistency is 1.000 for every event.

5.6. Geometry-Aware Routing Under Viewpoint Shift: A Local Fence-Crossing Pilot

Table 9 reports strict leave-one-camera-out results. The unconstrained evidence-fusion model improves the side view but transfers poorly to the frontal view, causing its Macro-F1 to fall below the baseline. This negative result is retained because it motivates geometry-aware routing. The routed system uses evidence fusion only where 2D leg geometry is reliable and otherwise preserves the conservative temporal rule. No test label is used to select the expert.

Table 9. Local-only fence-crossing pilot, macro-averaged over two held-out cameras.

Method	Frame Precision	Frame Recall	Frame F1	Event Precision	Event Recall	Event F1
Stabilized pose rule	0.721	0.608	0.638	1.000	0.700	0.786
Evidence fusion without routing	0.378	0.490	0.425	0.500	0.900	0.614
Geometry-aware routing	0.757	0.840	0.795	0.833	0.900	0.864

In the six-interval two-camera pilot, routed predictions attain necessity of 1.000, top-feature sufficiency of 0.625, and mean primary-evidence probability drop of 0.752. These local results do not validate a general multi-expert router.

The folds show opposite transfer: side-view event F1 rises from 0.571 for the rule to 0.727 for fusion, whereas frontal-view fusion falls from 1.000 to 0.500. Predeclared routing retains the suitable expert, but the two-camera sample limits generalization.

5.7. Faithfulness Trade-Off and Viewpoint Sensitivity

Figure 5 jointly visualizes sufficiency and necessity. The desired region is toward the upper left: a selected rationale should reproduce the decision with a small gap, and its removal should strongly reduce event probability. On public PPE graphs, the performance constraint moves the model from low-necessity solutions toward this region, while the complete model provides an additional necessity gain. On the real-site cross-camera dataset, PowerXGraph++ is clearly separated from the teacher and unconstrained selectors. The lower panels show why the private macro average must be interpreted with the two camera folds visible: fusion is beneficial in the side view but harmful in the frontal view.

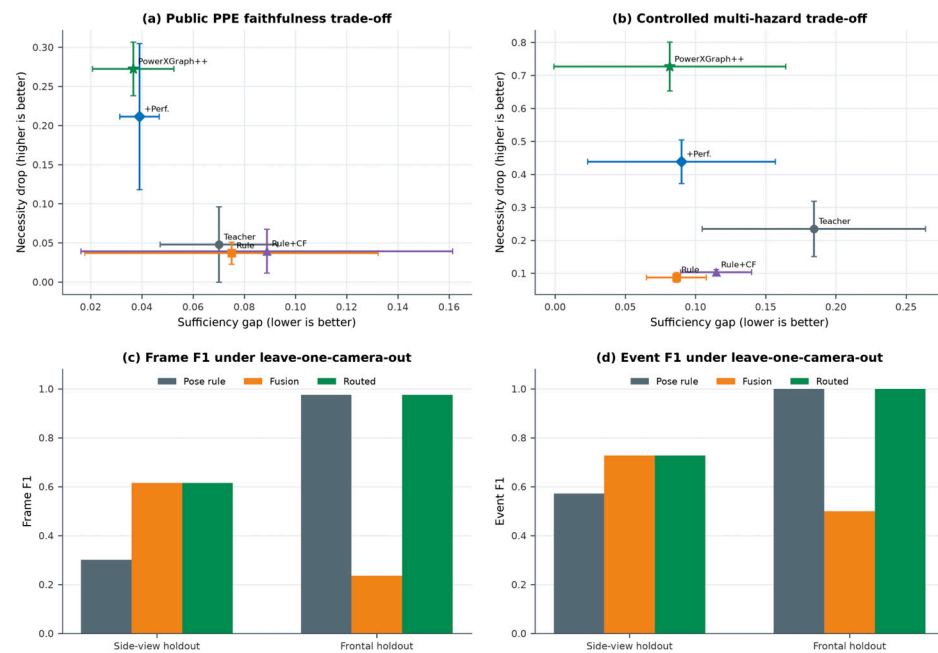


Figure 5. Explanation and viewpoint analysis. (a,b) Mean $\pm$ standard deviation of sufficiency gap and necessity drop across three seeds. (c,d) Frame- and event-level F1 for each held-out camera in the local fence-crossing pilot. The routed model selects the evidence-fusion expert for the side view and the conservative temporal expert for the frontal view.

5.8. Model Size and Reasoning Runtime

Table 10 reports the evidence reasoner separately from perception. Decoupling adds only 192 parameters to the two-class public model because it introduces one additional 96-dimensional query per class. The final reasoner remains below 0.17 million parameters, and its checkpoint is approximately 0.66 MB. A single length-16 graph forward pass takes 0.731 ms on the RTX 3080 Ti Laptop GPU; batch-64 throughput corresponds to 0.012 ms per sample. The real-site six-class reasoner contains 164,652 parameters. Its complete greedy explanation replay averages 16.11 ms over 100 events and selects 4.08 evidence items on average.

Table 10. Evidence-reasoner complexity and synchronized CUDA latency. Perception, tracking, and graph construction are excluded.

Item	Teacher	PowerXGraph++
Public two-class parameters	162,244	162,436
Checkpoint size	0.664 MB	0.664 MB
Length-16, batch-1 forward	-	0.731 ± 0.136 ms
Length-16, batch-64 forward	-	0.778 ± 0.065 ms total
Controlled greedy replay (100 events)	-	16.11 ± 2.29 ms/event
Mean evidence retained in controlled replay	-	4.08 items

The detector remains the dominant learned perception component at 2.6 ms/image, but this number cannot simply be added to the reasoner time: tracking, pose, relation extraction, multi-frame buffering, and optional expert invocation introduce additional latency. The transformer forward cost is $O(N^2d + Nd^2)$, where N denotes the number of evidence items and d is the hidden dimension. Greedy deletion may require up to N replay calls and therefore exhibits a worst-case cubic dependence on N when progressively shorter transformer sequences are rescored.

5.9. Bootstrap Uncertainty and Selective Coverage

In total, 10,000 explanation-record bootstrap resamples gave necessity drop 95% CIs of 0.273 [0.244, 0.302] for PowerXGraph++, 0.212 [0.184, 0.241] for the constrained variant, 0.047 [0.037, 0.057] for the teacher, and 0.726 [0.696, 0.756] on real-site evidence. Because the same cases recur across three seeds, records are dependent, and intervals may be narrower than case-clustered estimates; they are descriptive. Paired seed-wise changes in Section 5.3 provide a run-level check.

Requiring regulation consistency of 1.000 and confidence above threshold gave coverage of 1.000/0.860 on construction-PPE and 1.000/0.643 on real-site evidence at thresholds 0.50/0.90. Accepted-set Macro-F1 and selective risk were not computed, so these values do not prove abstention benefit. Low-confidence complete alarms go to manual review; incomplete chains request additional evidence; rejection is not treated as safe.

Figure 6, therefore, reports descriptive record-level uncertainty and two measured coverage points, not independent-case significance, an interpolated calibration curve, or risk-coverage improvement.

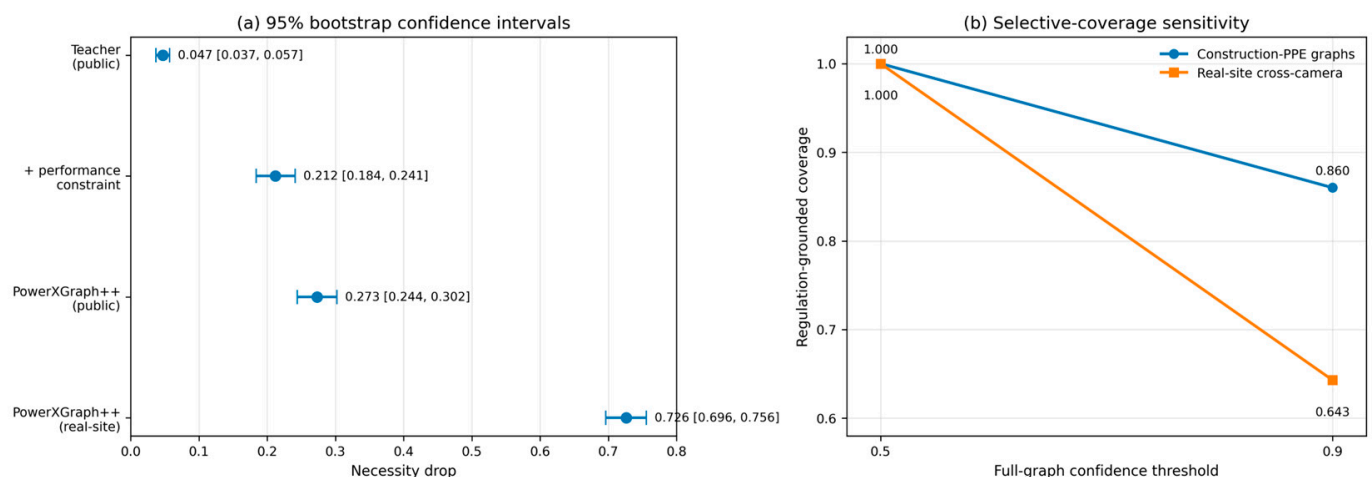


Figure 6. Descriptive uncertainty and selective-coverage sensitivity. (a) Mean necessity drops with 95% record-level bootstrap intervals from 10,000 resamples; repeated cases across seeds are not independent. (b) Regulation-grounded explanation coverage at the two measured full-graph confidence thresholds. Coverage requires regulation consistency of 1.000 and confidence above the threshold; accepted-set risk was not evaluated.

5.10. Matched Predicted-Box PPE Evaluation

Table 11 compares annotation-derived and matched YOLO-derived PPE evidence. In the single seed-17 reproduction run with YOLO inputs, Macro-F1 falls from 0.989 to 0.844, missing-PPE recall from 0.965 to 0.591, rationale F1 to 0.660, and regulation consistency to 0.760. Unmatched detections are disclosed as perception attrition rather than downstream classification cases; this is a matched PPE pathway, not a full video-monitoring pipeline.

Annotation-derived entries are mean $\pm$ standard deviation over reasoner seeds 17, 23, and 41. YOLO entries are from a single reproduction run using seed 17 with the 45-epoch detector. The 141-image test set contains 236 annotated persons (189 safe and 47 missing PPE). At confidence = 0.25 and NMS IoU = 0.70, YOLO produces 244 person graphs; one-to-one matching at $\text{IoU} \geq 0.50$ yields 197, leaving 47 unmatched predictions and 39 unmatched references. The 197 matches contain 153 reference-safe and 44 reference-missing-PPE cases; predictions are 171 safe and 26 missing PPE. With missing PPE as the positive class, the confusion counts are TP = 26, FP = 0, FN = 18, and TN = 153, giving Macro-F1 = 0.844 and missing-PPE recall = 0.591. PPE association uses the same 8-expanded-box center rule.

With the same YOLO evidence and frozen predictor, event metrics are identical across explainers; PowerXGraph++ raises rationale F1 from 0.600 to 0.660 and regulation consistency from 0.680 to 0.760. This supervision-asymmetric comparison illustrates how perception errors limit explanation completeness.

Table 11. Construction-PPE results with annotation-derived evidence (three-seed mean $\pm$ standard deviation) and YOLO-predicted evidence (single seed-17 reproduction run; 197 matched person graphs). Event metrics are shared by explanation methods using the same frozen predictor.

Evidence Source	Explanation Method	Macro-F1	Violation Recall	Rationale F1	Rule Consistency
Annotation-derived boxes (three seeds)	PowerXGraph++	0.989 ± 0.010	0.965 ± 0.033	0.983 ± 0.018	1.000
YOLO-predicted boxes (seed = 17; $n = 197$)	PowerXGraph++	0.844	0.591	0.660	0.760
YOLO-predicted boxes (seed = 17; $n = 197$)	GNNExplainer-style mask	0.844	0.591	0.600	0.680

5.11. Independent Expert Rationale Assessment

Figure 7 visualizes the matched predicted-box performance drop (panel a) and the agreement between each explanation method and the expert-majority rationales (panel b). Three independent, blinded power-safety or field-operation experts assessed 180 stratified cases. Fleiss' kappa was 0.740; method-to-majority rationale F1 was 0.870 for PowerXGraph++, 0.750 for GNNExplainer-style, and 0.700 for attention top-k. The advantage is consistent with PowerXGraph++'s rationale supervision and grammar constraints, not a supervision-matched algorithm ranking.

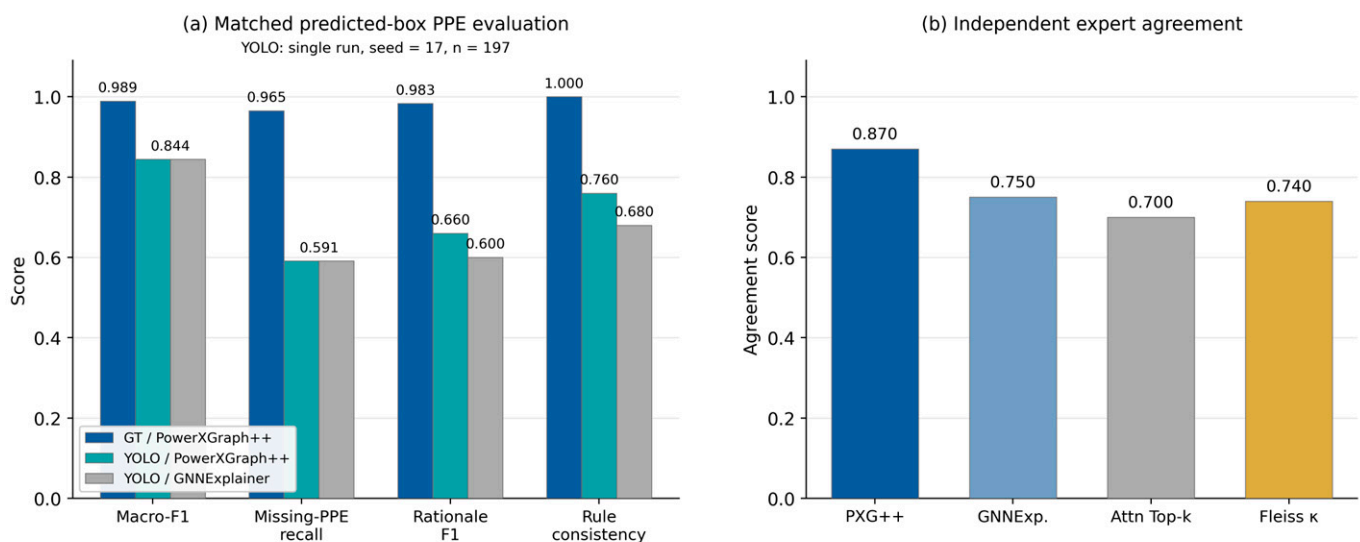


Figure 7. Matched predicted-box and independent expert-evaluation results. (a) Annotation-derived results and a single YOLOv11n reproduction run using seed 17 on 197 matched person graphs (reference safe/missing PPE: 153/44; predicted: 171/26; Macro-F1: 0.844; missing-PPE recall: 0.591). (b) Method agreement with expert-majority rationales and Fleiss' κ among the three experts.

Expert similarity and counterfactual faithfulness are complementary. The 180-case result supports agreement under the supplied representation but does not establish broader usability or remove the influence of the author-defined grammar. Table 12 summarizes independent expert agreement and method-to-expert rationale alignment on the 180 stratified cases.

Table 12. Independent expert agreement and method-to-expert rationale alignment on 180 stratified cases.

Evaluated Quantity	Metric	Result
Agreement among three experts	Fleiss' κ	0.740
PowerXGraph++ vs. expert majority	Rationale F1	0.870
GNNExplainer-style mask vs. expert majority	Rationale F1	0.750
Attention top-k vs. expert majority	Rationale F1	0.700

6. Discussion

6.1. Does Explainability Reduce Accuracy?

The public PPE ablation shows improved explanations without reduced event accuracy under annotation-derived evidence. Rationale supervision, counterfactual replay, grammar constraints, distillation, and the non-inferiority gate serve complementary roles. The single seed-17 YOLO reproduction run on 197 matched person graphs yields Macro-F1 of 0.844 and missing-PPE recall of 0.591, highlighting upstream perception limits rather than full-pipeline performance.

The rule checker requires complete event-specific groups and flags missing evidence for review or completion. Its grammar is editable engineering logic, not certified law; site professionals must adapt operational requirements. Grammar completeness does not prove label-independent predicate extraction.

Table 13 compares the conceptual scope of the methods, whereas the same-predictor explanation comparison evaluates evidence-token adaptations under one frozen Transformer predictor. The baselines lack PowerXGraph++'s rationale and grammar inputs, so results quantify unequal supervision and inductive bias rather than a fair universal ranking.

Table 13. Conceptual comparison with representative graph-explanation approaches. “Both” denotes explicit selected-subgraph sufficiency and complementary-subgraph necessity evaluation.

Method	Explanation Form	Temporal Graph	Domain Grammar	Faithfulness Target	Prediction Floor
GNNExplainer [6]	post hoc subgraph/feature mask	not explicit	no	mutual-information fidelity	no
PGExplainer [7]	parameterized post hoc subgraph	not explicit	no	prediction fidelity	no
CF-GNNExplainer [8]	counterfactual edge perturbation	no	no	prediction change	no
GraphFramEx [25]	evaluation framework	general	no	sufficiency/necessity views	no
CoDy [26]	dynamic counterfactual	yes	no	prediction changes over time	no
PowerXGraph++	typed minimal evidence chain	yes	yes	both + rule completion	yes

6.2. Viewpoint Robustness, Failure Modes, and Operational Oversight

The pilot shows viewpoint-dependent transfer: side-view geometry helps, frontal projection can mislead, and routing matters. Future routing should use keypoint, homography, and calibration uncertainty with abstention; the current rule is only a two-camera demonstration.

Five failure layers emerge: perception omission, incomplete relations, context confounding, projection ambiguity, and tracking-induced event fragmentation. Their diagnostics are respectively class-wise detector metrics, grammar completion, complementary-

graph probability, routing provenance, and event interval statistics. Unlike a heatmap, this layered record indicates whether remediation requires data, another perception expert, a relation rule, calibration, or temporal smoothing.

An audit record may include event, confidence, interval, subject, grammar, selected keys, counterfactual scores, and routing. Low-confidence complete cases go to human review, and incomplete chains request additional evidence; these are proposed controls, not deployment outcomes.

6.3. Privacy and Data Availability

Private videos were read in place, and only aggregate metrics were entered into the project; no raw frames were included. This protects the site but limits independent pilot reproduction. Evidence graphs remove appearance texture but can still disclose trajectories, timestamps, or camera identity, so releases should pseudonymize identifiers, coarsen time, remove absolute coordinates, and audit rare metadata combinations.

6.4. Limitations

Controlled results condition on annotation-derived or supplied evidence; semantic predicates are close to targets, no lower-level-only ablation was run, and automatic label-independent extraction was not established for all classes. The single detector reproduction run using seed 17 covers 197 matched graphs and excludes 47 unmatched predictions and 39 missed references from downstream denominators. Camera c is held out, but worker, event, or sequence separation within training and validation is undocumented. Multi-site, event-level, and end-to-end evaluation remains needed.

Routing was tested only for fence crossing on two cameras and six intervals using fixed viewpoint metadata. Other hazards require explicit extractors, reliability estimates, and independent event-level evaluation.

The expert study measures evidence-key agreement, not usability or review burden; baseline supervision is asymmetric; record bootstrap ignores case clustering; coverage lacks selective risk. Future work should address these items, multi-seed detection, lower-level ablations, and multi-site evaluation.

7. Conclusions

PowerXGraph++ combines transformer-encoded graph facts, full-context prediction, counterfactual replay, and engineering grammar for auditable evidence selection. The completion interface identifies missing observations for additional perception or human review.

Across three seeds, constrained training improved annotation-derived PPE reasoning and explanation. Supervision-asymmetric matched-input comparisons favored regulation-complete selection; a single YOLO reproduction run using seed 17 on 197 matches exposed perception limits; expert agreement reached kappa of 0.740 and method-to-majority F1 of 0.870. Camera-c reasoning and the two-camera crossing pilot support transfer only under supplied evidence. The results do not establish label-independent predicates, event-disjoint development data, full raw video performance, or broad deployment.

Author Contributions: Conceptualization, G.P. and Q.Z.; methodology, Q.Z. and G.P.; software, Q.Z., Y.Z. and Z.W.; validation, Z.W.; formal analysis, Z.C.; investigation, Z.W.; resources, Z.C.; data curation, Q.Z.; writing-original draft preparation, Q.Z.; writing-review and editing, G.P., Z.C. and Y.Z.; visualization, Q.Z. and Z.W.; supervision, G.P.; project administration, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the State Grid Corporation of China Headquarters Project, which is “Research and Application of Key Technologies for Explainable Multi-object

Recognition of On-site Operational Violations Based on Image-Text Models,” under Grant 5700-202426249A-1-1-ZN.

Data Availability Statement: Construction-PPE is publicly available from Ultralytics at Construction-PPE is publicly available from Ultralytics at <https://docs.ultralytics.com/datasets/detect/construction-ppe/> (accessed on 1 September 2026). The cross-camera multi-hazard and fence-crossing data were collected in real electric-power worksite environments. They are not publicly available because the recordings contain operational surveillance imagery and are subject to enterprise data security, privacy, and ownership restrictions. De-identified derived evidence graphs, camera-wise split statistics, expert annotation guidelines, per-expert evidence-key labels, majority-vote rationales, code, configurations, and aggregate outputs may be made available from the corresponding author upon reasonable request and with permission from the data owner.

Conflicts of Interest: Authors Qizhe Zhang, Yi Zhang, Zheng Wang, and Guozheng Peng were employed by the company China Electric Power Research Institute. Author Zhengwei Chang was employed by the company State Grid Sichuan Electric Power Research Institute. The authors declare that this study received funding from the State Grid Corporation of China. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

References

1. Chang, R.; Li, B.; Dang, J.; Yang, C.; Pan, A.; Yang, Y. Real-Time Intelligent Detection System for Illegal Wearing of On-Site Power Construction Worker Based on Edge-YOLO and Low-Cost Edge Devices. *Appl. Sci.* **2023**, *13*, 8287. [CrossRef]
2. Park, M.; Kulinan, A.S.; Tran, D.Q.; Bak, J.; Park, S. Preventing Falls from Floor Openings Using Quadrilateral Detection and Construction Worker Pose-Estimation. *Autom. Constr.* **2024**, *165*, 105536. [CrossRef]
3. Yan, Y.; Chen, X.; Jin, H.; Qi, D.; Chu, H.; Wang, J. Research Status and Prospects of Computer-Vision-Based Behavior Analysis for Electric-Power Workers. *High Volt. Eng.* **2024**, *50*, 1842–1854.
4. Yang, S.; Wang, Y.; Guo, S.; Feng, S. An Operation Site Security Detection Method Based on Point Cloud Data and Improved YOLO Algorithm under the Architecture of the Power Internet of Things. *J. Eng.* **2024**, *2024*, e12344. [CrossRef]
5. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision-ECCV 2020*; Springer: Cham, Switzerland, 2020; Volume 12346, pp. 213–229. [CrossRef]
6. Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems 32*; Curran Associates: Red Hook, NY, USA, 2019; pp. 9240–9251.
7. Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; Zhang, X. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems 33*; Curran Associates: Red Hook, NY, USA, 2020; pp. 19620–19631.
8. Lucic, A.; Ter Hoeve, M.A.; Tolomei, G.; De Rijke, M.; Silvestri, F. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual*, 28–30 March 2022; Volume 151, pp. 4499–4511.
9. Beyer, L.; Zhai, X.; Royer, A.; Markeeva, L.; Anil, R.; Kolesnikov, A. Knowledge Distillation: A Good Teacher Is Patient and Consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022.
10. Yang, L.; Guo, Y.; Huo, Y.; Ren, X.; Lin, P.; Zhang, Z. Application of Improved YOLOv5 to Non-Compliant Wearing Detection in Electric Power Production Sites. *Power Syst. Prot. Control* **2023**, *51*, 160–168.
11. Guo, P.; Li, L.; Wang, Y.; Zheng, B.; Wang, X.; Chen, Y. Safety-Behavior Recognition at Power Worksites Based on Improved DETR. *Electr. Power Inf. Commun. Technol.* **2024**, *22*, 63–70.
12. Su, Q.; Zhang, J.; Chen, M.; Peng, H. PW-YOLO-Pose: A Novel Algorithm for Pose Estimation of Power Workers. *IEEE Access* **2024**, *12*, 116841–116860. [CrossRef]
13. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 19–20 June 2022; pp. 2637–2646.
14. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2–7 February 2018; Volume 32, pp. 7444–7452. [CrossRef]

15. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
16. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *Computer Vision-ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 1–21. [\[CrossRef\]](#)
17. Zhou, Z.; Song, Y.; He, Y.; Wang, L.; Huang, H.; Zhu, Z.; Yan, Y. Extensible Classification Method for Power Personnel Behavior Based on Grouped-Query Attention. *Electr. Power* **2023**, *56*, 77–85. [\[CrossRef\]](#)
18. Wu, T.; Wan, Y.; Wang, S.; Xiao, B.; Fang, C.; Li, P. Video Abnormal Behavior Recognition Method for 10 kV Live-Line Workers Based on Improved ST-GCN. *South. Power Syst. Technol.* **2024**, *18*, 31–39. [\[CrossRef\]](#)
19. Qian, Y.; Zhu, T.; Cao, X.; Li, R.; Gong, E.; Wang, J. Intelligent Personnel-Intrusion Detection System for Substations Based on Hi3559A. *Power Syst. Technol.* **2021**, *45*, 4181–4188.
20. Cai, C.; Nie, J.; Shi, M.; Zhang, X.; Zhang, Z.; He, Z. Complex Violation Recognition for Electric-Power Workers Based on Binocular Vision and Point-Cloud Fusion. *Electr. Power Inf. Commun. Technol.* **2025**, *23*, 35–42. [\[CrossRef\]](#)
21. Zhang, Y.; Tang, S.; Yu, Z.; Feng, J. Three-Dimensional Localization of Personnel in Power-Worksite Scenes Based on Monocular Vision. *Electr. Power Inf. Commun. Technol.* **2025**, *23*, 53–59. [\[CrossRef\]](#)
22. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
23. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
24. Vu, M.N.; Thai, M.T. PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems 33*; Curran Associates: Red Hook, NY, USA, 2020; pp. 12225–12235.
25. Amara, K.; Ying, Z.; Zhang, Z.; Han, Z.; Zhao, Y.; Shan, Y.; Brandes, U.; Schemm, S.; Zhang, C. GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks. In Proceedings of the First Learning on Graphs Conference, Virtual, 9–12 December 2022; Volume 198, pp. 44:1–44:23.
26. Qu, Z.; Gomm, D.; Färber, M. CoDy: Counterfactual Explainers for Dynamic Graphs. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025; Volume 267, pp. 50762–50785.
27. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 2151–2159.
29. Dalvi, M.; Singh, N.; Bhingarde, S.; Chalke, K. *Construction-PPE: Personal Protective Equipment Detection Dataset, Version 1.0.0*; Ultralytics: Frederick, MD, USA, 2025. Available online: <https://docs.ultralytics.com/datasets/detect/construction-ppe/> (accessed on 15 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.