PRISM: Deployable Counterfactual Recourse for Metro Maintenance via Physics-Constrained Search and RL Distillation

Xing Su[1,2], Fei Teng[1], Yifan Zhuang[2], Jiasheng Fang[2], Zixuan Xu[2] and Dong Zou[2]
1 *School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China*
2 *Zhejiang Rail Transit Operation Management Group Co., Ltd., Hangzhou, China*
Email: {suxing, zhuangyifan, fangjiasheng, xuzixuan, zoudong}@rtom.com.cn, fteng@swjtu.edu.cn

***Abstract*—Urban metro safety affects millions of daily journeys. Abnormal contact-wire wear is a major threat to reliable metro current collection. Severe wear can trigger arcing and overheating, accelerate component degradation, and in extreme cases lead to service disruptions and safety hazards. Preventing such events requires more than early warning: operators also need actionable and physically feasible mitigation suggestions for the rare but high-risk severe cases. To address this gap, we propose PRISM, an industrial framework that links reliable ordinal wear grading with physics-constrained mitigation. PRISM first performs imbalance-aware multi-head grading classification, and then formulates mitigation as a constrained counterfactual recourse problem. PRISM instantiates a prototype-anchored search teacher, guided by physical priors derived from a differentiable surrogate to accelerate optimization and enable audit-friendly explanations. To meet online constraints, PRISM distills the teacher into an offline actor–critic student with verification and teacher fallback in reinforcement learning. Experiments on three real-world datasets demonstrate that PRISM consistently improves wear grading and achieves the best severe-wear mitigation success over strong baselines.**

***Index Terms*—predictive maintenance, imbalance-aware multi-class classification, counterfactual recourse, physics-informed optimization, offline reinforcement learning, policy distillation**

I. Introduction

Urban metro systems have become a backbone of modern cities, carrying massive daily passenger flows and directly shaping urban mobility, economic efficiency, and public safety. Ensuring high availability and safety therefore hinges on reliable operation and maintenance (O&M) practices [1]. Among critical subsystems, the pantograph–catenary interface plays a pivotal role in traction power supply, where the contact wire serves as the primary current-collection medium [2]. As shown in Fig. 1, abnormal contact-wire wear can degrade current collection quality, trigger arcing and overheating, accelerate component degradation, and in severe cases lead to service disruptions, emergency repairs, or safety hazards. Consequently, accurate wear assessment and timely mitigation are central to metro O&M, with tangible impacts on safety assurance, lifecycle cost, and operational continuity.

Despite extensive sensing infrastructure and growing adoption of data-driven analytics, addressing contact-wire wear in real-world metro maintenance remains challenging. First, wear

Corresponding Author: Zixuan Xu

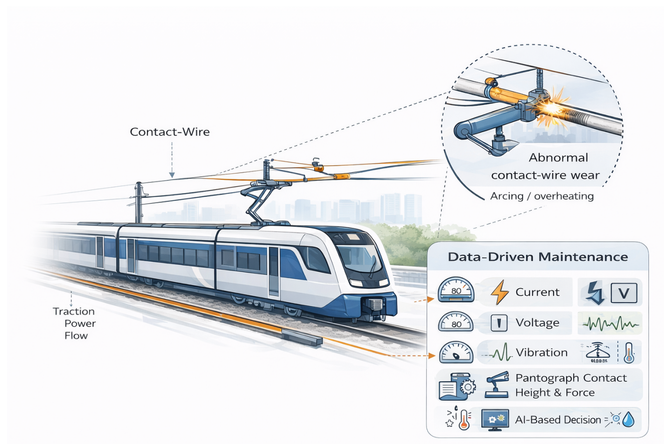


Fig. 1. Urban metro contact-wire wear and data-driven maintenance scenario.

patterns are influenced by complex and non-stationary operating conditions (*e.g.*, speed profiles, environmental factors, load variations), and multi-sensor signals are heterogeneous and noisy. Second, wear-grade distribution is heavily imbalanced: severe wear is rare but critical, while most samples concentrate in mild-to-moderate grades. This imbalance makes multi-class learning prone to severe-class performance collapse and unreliable probability calibration [3], [4]. Third, metro O&M demands strict latency requirements for online decision support [5]. These constraints jointly render "train a classifier and raise alarms" insufficient for practical deployment.

Crucially, wear grading alone provides limited operational value: it tells *what* the system believes the current wear state is, but not *how* to act. Maintenance engineers require *actionable recourse* [6], a minimal, feasible, and auditable intervention plan indicating which signal-related factors should be adjusted (and in which direction) to reduce predicted wear severity under operational constraints and physical coherence across related signals. Naively perturbing features may yield unrealistic or contradictory suggestions, while direct optimization-based mitigation can be computationally expensive. Therefore, the practical objective in metro O&M is not only imbalance-aware multi-class wear grading, but also physics-consistent counterfactual mitigation that is fast, reliable, and interpretable.

To address these needs, we propose **PRISM** (**P**hysics-informed **R**ecourse with **I**mbalance-aware learning and **S**tudent distillation for **M**etro maintenance), an industrial AI framework that unifies decision-grade wear assessment, actionable mitigation, and deployable acceleration. PRISM follows a three-stage technical pipeline: **(i)** an imbalance-aware multi-class grader built on Gradient-Boosted Decision Trees (GBDT) [7] with a gated multi-head architecture and calibrated fusion for reliable probabilistic grading; **(ii)** a physics-informed counterfactual recourse engine that searches for feasible group-level interventions via prototype-anchored cross-entropy optimization [8], where a surrogate model accelerates candidate screening and instantiates our domain-encoded direction and influence priors from sensitivity signals to guide the search; and **(iii)** an industrial deployment module that distills the optimization-based teacher into an offline Reinforcement Learning (RL) student policy (with verification/fallback) to meet millisecond-level latency constraints [6], [9], [10].

Our main contributions are summarized as follows:

- **Imbalanced multi-class wear grading.** To detect rare high-severity wear, we proposed an imbalance-aware multi-head GBDT grader with calibrated fusion and conservative gating.
- **Physics-informed and efficient recourse formulation.** To provide actionable maintenance guidance, we formulate wear mitigation as *constrained counterfactual recourse* over a prototype-anchored, feasibility-constrained counterfactual teacher guided by surrogate-assisted screening and two domain-encoded priors to improve efficiency and auditability.
- **Low-latency industrial deployment.** To overcome the latency bottleneck of recourse in online metro O&M, we distill the teacher into an offline entropy-regularized actor–critic student with imitation regularization, coupled with verification/fallback for reliability.

II. Related Work

Predictive Maintenance and Wear Assessment in Railway Electrification Systems: Condition-based maintenance for railway electrification increasingly relies on multi-sensor monitoring and data-driven diagnosis to complement high-fidelity physical simulations that are costly to calibrate under field variability. Recent surveys summarize the progress of intelligent inspection and defect detection for catenary infrastructure [11]. For pantograph–catenary interaction, prior work has analyzed wear mechanisms and dynamic coupling, providing the engineering backdrop for data-driven assessment and intervention [12], [13]. In practice, wear assessment for metro O&M is often performed on heterogeneous tabular features, where tree ensembles remain competitive due to robustness under noise and missingness [7], [14].

Class-Imbalanced Multi-Class Learning for Tabular Data Safety-critical railway monitoring exhibits a long-tailed label distribution, making severe conditions rare yet decision-critical. TabR integrates retrieval to enhance tabular generalization [15], while LOS targets long-tailed recognition by mitigating over-smoothing effects [16]. CLIMB further provides a benchmark and protocol for class-imbalanced tabular learning, highlighting the gap between overall accuracy and minority reliability [17]. From a representation-theoretic perspective, imbalanced training can induce distinct collapse behaviors that correlate with minority failure modes [18]. These findings motivate our design that explicitly decouples severe detection from fine-grained grading and emphasizes calibration/reliability rather than only aggregate metrics.

Actionable Recourse and Constrained Counterfactuals Counterfactual explanations have evolved from post-hoc perturbations [19] to actionable recourse that respects intervention feasibility [6], [20]. DiCE promotes diversity but may violate operational constraints [21], whereas FACE improves feasibility by constraining counterfactuals to manifold-consistent paths [22]. Comprehensive reviews further stress that feasibility, cost, and domain constraints are central for deploying recourse in high-stakes settings [23]. Our work aligns with this line by formulating mitigation as constrained recourse over signal$\times$family groups with explicit budgets and coherence, matching the semantics of metro maintenance actions.

Physics-Informed Priors and Structured Interventions Physics-informed machine learning provides a principled way to inject domain structure into learning [24]. In railway engineering, physics-informed modeling for moving-load/rail structures has been increasingly adopted, underscoring the value of encoding constraints and interpretable structure [25]. For constrained, non-convex recourse optimization, the cross-entropy method is a classical derivative-free tool that remains effective when combined with good anchors and constraint handling [26]. In PRISM, the direction/influence matrices serve as domain-encoded priors for guidance and audit, while the optimization operates in a structured action space aligned with maintainable signal groups.

Offline Reinforcement Learning and Policy Distillation for Low-Latency Deployment Optimization-based recourse can be too slow for online decision loops. Offline RL enables learning policies from logged data without unsafe online exploration [27], with conservative algorithms addressing out-of-distribution actions [10]. Soft Actor-Critic provides an actor–critic backbone widely used for stable off-policy learning [9]. Offline behavior distillation further studies expert-like behaviors for efficient deployment [28], while diffusion-based trust-region style offline RL has been explored as a robustification mechanism [29]. For safety-critical deployment, recent offline safe RL studies explicitly model reward–constraint trade-offs under fixed datasets and support deployment-time adaptation to different safety budgets [30]–[32]. However, standard formulations do not encode the physical and semantic constraints needed for actionable maintenance recourse, and purely learned policies may fail on rare, high-risk instances.

III. Problem Statement

A. Notation and semantic group structure

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote a metro monitoring dataset, where $\mathbf{x} \in \mathbb{R}^F$ is an engineered feature vector extracted from

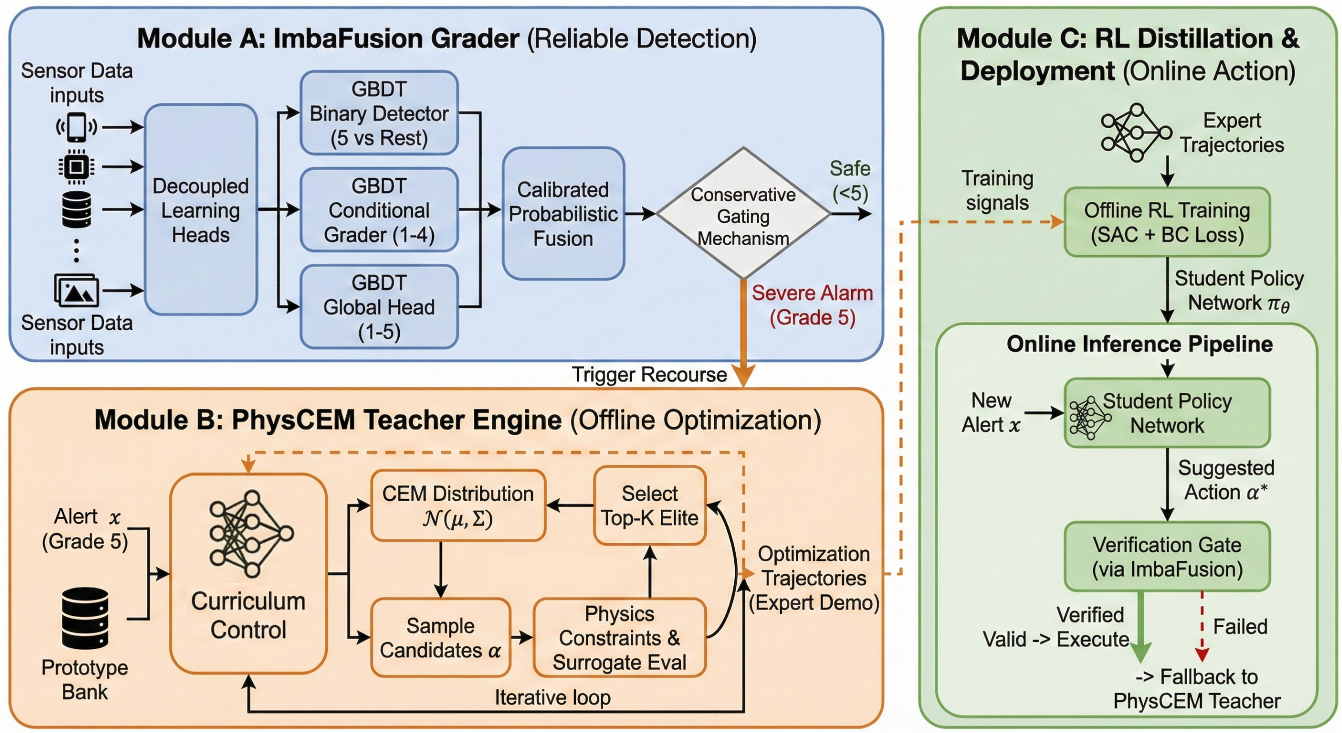


Fig. 2. Overall framework of PRISM, coupling reliable wear grading, physics-constrained recourse, and deployment-time distillation with verification/fallback.

multi-sensor streams and $y \in \{1,2,3,4,5\}$ is an *ordinal* wear grade (larger indicates more severe wear). The class distribution is heavily imbalanced, where high-severity grades are rare but safety-critical.

a) Semantic decomposition of features: Each feature dimension $j \in \{1,\ldots,F\}$ is associated with a feature name that admits deterministic semantic parsing consistent:

$$\pi(j) \triangleq \big(s(j),\, k(j)\big), \tag{1}$$

where $s(j) \in \mathcal{S}$ is a *signal token* (sensor/channel identifier) and $k(j) \in \mathcal{K}$ is a *feature family* (*e.g.*, amplitude/diff/spectral/entropy/shape) parsed from the suffix pattern. This induces *semantic intervention groups* defined by signal$\times$family:

$$\begin{aligned} \mathcal{G} &\triangleq \{g=(s,k) : s\in\mathcal{S},\, k\in\mathcal{K},\, \mathcal{I}(g)\neq\emptyset\},\\ \mathcal{I}(g) &\triangleq \{j\in\{1,\ldots,F\} : \pi(j)=g\}, \end{aligned} \tag{2}$$

where $\mathcal{I}(g)$ denotes the feature-index set of group g, and $G \triangleq |\mathcal{G}|$. We further define the group-to-signal map $\sigma(g)=s$ and group-to-family map $\phi(g)=k$ for $g=(s,k)$, and the signal-wise neighborhood

$$\mathcal{G}(s) \triangleq \{g\in\mathcal{G} : \sigma(g)=s\}. \tag{3}$$

This structure aligns algorithmic interventions with operational semantics (signal-level adjustability and feature-family coherence), enabling auditable recourse with per-signal budgeting.

B. Task 1: imbalance-aware multi-class grading

We learn a probabilistic grader $f_{\text{cls}} : \mathbb{R}^F \to \Delta^5$ that outputs $\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}),\ldots,p_5(\mathbf{x})]$, where Δ^5 denotes the 5-simplex. The predicted grade is $\hat{y}(\mathbf{x}) = \arg\max_{c\in\{1,\ldots,5\}} p_c(\mathbf{x})$. The goal is to maximize overall grading quality while remaining *reliable* on rare high-severity conditions under heavy imbalance (*e.g.*, avoiding severe-class recall collapse and miscalibration).

C. Task 2: physics-informed constrained recourse (mitigation)

Given a high-severity instance $\mathbf{x}$ (predicted or labeled), we seek an actionable mitigation plan that produces a counterfactual $\mathbf{x}'$ whose predicted severity is reduced to a target level y_{tar}. Actions are defined over semantic groups. Let $\boldsymbol{\alpha} \in \mathbb{R}^G$ be a group-level action (one coefficient per $g \in \mathcal{G}$), and let $\mathbf{x}^{\text{p}} \in \mathbb{R}^F$ denote a *prototype anchor* selected from a low-severity prototype bank $\mathcal{P}$. We define a structured edit operator $\mathcal{T}$ as group-wise interpolation:

$$\mathbf{x}' \triangleq \mathcal{T}(\mathbf{x},\mathbf{x}^{\text{p}};\boldsymbol{\alpha}),\, \mathbf{x}'_{\mathcal{I}(g)} = (1-\alpha_g)\,\mathbf{x}_{\mathcal{I}(g)} + \alpha_g\,\mathbf{x}^{\text{p}}_{\mathcal{I}(g)},\ \forall g\in\mathcal{G}. \tag{4}$$

The feasible set $\mathcal{C}$ encodes operational constraints (sparsity, budgets, and coherence):

$$\mathcal{C} \triangleq \left\{ \boldsymbol{\alpha} \,\middle|\, \begin{array}{ll} 0 \le \alpha_g \le \alpha_{\max}, & \forall g\in\mathcal{G}, \\ \|\boldsymbol{\alpha}\|_0 \le K, & \\ \displaystyle\sum_{g\in\mathcal{G}(s)} \alpha_g \le B_{\text{sig}}, & \forall s\in\mathcal{S}, \\ \displaystyle\sum_{g\in\mathcal{G}} \alpha_g \le B_{\text{tot}}, & \\ \Omega_{\text{coh}}(\boldsymbol{\alpha}; s) \le \kappa_{\text{coh}}, & \forall s\in\mathcal{S}, \end{array} \right\}, \tag{5}$$

where $\Omega_{\text{coh}}(\boldsymbol{\alpha}; s)$ measures within-signal incoherence (*e.g.*, the variance of $\{\alpha_g : g \in \mathcal{G}(s)\}$).

The recourse objective is to find $(\mathbf{x}^{\text{p}}, \boldsymbol{\alpha})$ such that

$$\hat{y}(\mathbf{x}') \le y_{\text{tar}}, \quad \mathbf{x}' = \mathcal{T}(\mathbf{x}, \mathbf{x}^{\text{p}}; \boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \mathcal{C}, \tag{6}$$

while minimizing a cost that balances severity reduction and intervention economy (formalized in Sec. IV).

IV. Methodology

PRISM casts metro wear decision support as a staged pipeline composed of three modules, as illustrated in Fig. 2. First, *ImbaFusion Grader* performs multi-grade classification built on gradient-boosted decision trees with a gated multi-head architecture and calibrated fusion for reliable probabilistic grading. Second, for cases of severe wear, *PhysCEM Teacher Engine* searches for an auditable mitigation plan via prototype-based curriculum control with physics constraints and surrogate evaluation to propose physically coherent recourse trajectories. Finally, *RL Distillation and Deployment* distills these expert trajectories into a low-latency student policy and wraps it with verification and teacher fallback, so that online decisions achieve millisecond response while retaining the robustness of the offline teacher. The following subsections detail the design of each module, from imbalance-aware grading, through physics-informed recourse search, to deployment-time distillation.

A. ImbaFusion-GBDT: Imbalance-Aware Multi-Class Wear Grading

Wear grading in metro O&M exhibits a pronounced long-tail distribution, *i.e.*, high-severity wear is sparse yet dominates safety risk. In this regime, a single 5-way classifier optimized for aggregate accuracy often suffers from (i) severe-class under-recognition, and (ii) probability miscalibration, which in turn destabilizes downstream recourse optimization. We therefore adopt a decomposition-and-fusion principle that explicitly separates *severe detection* from *non-severe grading*, and then reconstructs a coherent 5-class distribution.

We train three gradient-boosted decision tree (GBDT) heads with imbalance-aware objectives: a binary detector f_{bin} for 5-vs-rest, a full multi-class head f_{full} for global class structure, and a conditional head $f_{1:4}$ trained on non-severe samples to refine discrimination among grades 1–4. Denote their probabilistic outputs by

$$\begin{aligned}
\mathbf{p}^{\text{bin}}(\mathbf{x}) &= [p^{\text{bin}}_{\neg 5}(\mathbf{x}),\, p^{\text{bin}}_5(\mathbf{x})] \in \Delta^2, \\
\mathbf{p}^{\text{full}}(\mathbf{x}) &= [p^{\text{full}}_1(\mathbf{x}), \dots, p^{\text{full}}_5(\mathbf{x})] \in \Delta^5, \\
\mathbf{p}^{1:4}(\mathbf{x}) &= [p^{1:4}_1(\mathbf{x}), \dots, p^{1:4}_4(\mathbf{x})] \in \Delta^4.
\end{aligned} \tag{7}$$

We construct the final distribution $\mathbf{p}(\mathbf{x}) \in \Delta^5$ by stabilizing the severe probability via a calibrated mixture and allocating the residual mass through conditional normalization. Specifically, the severe mass is

$$p_5(\mathbf{x}) = \eta\, p^{\text{bin}}_5(\mathbf{x}) + (1-\eta)\, p^{\text{full}}_5(\mathbf{x}), \qquad \eta \in [0,1], \tag{8}$$

and the remaining classes are defined as

$$p_c(\mathbf{x}) = \big(1 - p_5(\mathbf{x})\big)\, p^{1:4}_c(\mathbf{x}), \quad c \in \{1,2,3,4\}. \tag{9}$$

By construction, $\mathbf{p}(\mathbf{x}) \in \Delta^5$ holds for all $\mathbf{x}$, *i.e.*, $p_c(\mathbf{x}) \ge 0$ and $\sum_{c=1}^{5} p_c(\mathbf{x}) = 1$, ensuring well-defined risk signals for the controller in Sec. IV-B.

In safety-critical O&M, a severe prediction triggers costly maintenance actions, hence the decision rule should prioritize reliability rather than purely $\arg\max$. We adopt a conservative gate that activates grade 5 only under joint absolute confidence and relative dominance:

$$\hat{y}(\mathbf{x}) = \begin{cases} 5, & p_5(\mathbf{x}) \ge \tau \ \wedge\ p_5(\mathbf{x}) \ge \\ & \rho \cdot \max_{c \in \{1,2,3,4\}} p_c(\mathbf{x}), \\ \arg\max\limits_{c \in \{1,2,3,4\}} p_c(\mathbf{x}), & \text{otherwise}, \end{cases} \tag{10}$$

where (η, τ, ρ) are selected on validation data with imbalance-aware criteria. This gate mitigates long-tail failure modes and yields decision-grade probabilities for recourse generation.

B. PhysCEM-Recourse: Physics-Informed Counterfactual Mitigation

Operational mitigation in metro O&M requires more than "flipping" a prediction: the intervention must be feasible under engineering constraints, plausible with respect to signal semantics, and auditable for root-cause analysis. Accordingly, we instantiate mitigation as a constrained counterfactual recourse problem over semantic groups. Given an input $\mathbf{x}$, we seek a prototype anchor $\mathbf{x}^{\text{p}} \in \mathcal{P}$ and a group-level action $\boldsymbol{\alpha} \in \mathcal{C}$ so that the counterfactual $\mathbf{x}' = \mathcal{T}(\mathbf{x}, \mathbf{x}^{\text{p}}; \boldsymbol{\alpha})$ (Eq. (4)) reduces predicted severity while satisfying feasibility constraints $\boldsymbol{\alpha} \in \mathcal{C}$ (Eq. (5)). This parameterization ensures that each intervention is expressed in terms of interpretable signal$\times$family groups, and that operational budgets and coherence can be enforced directly in the action space.

To maintain plausibility, we anchor edits to prototypes drawn from low-severity regimes. For each query instance $\mathbf{x}$, we retrieve a small candidate set of anchors from $\mathcal{P}$ and solve a constrained optimization for each anchor. Anchoring acts as a structural prior that prevents arbitrary feature perturbations and stabilizes optimization under tight budgets.

Let $\mathbf{p}(\mathbf{x}') = f_{\text{cls}}(\mathbf{x}') = [p_1(\mathbf{x}'), \dots, p_5(\mathbf{x}')]$ be the grader output on the counterfactual. We quantify the mitigation goal by the expected severity $\text{sev}(\mathbf{x}') \triangleq \sum_{c=1}^{5} c\, p_c(\mathbf{x}')$, and penalize residual mass on high-wear regimes via $r_{\text{high}}(\mathbf{x}')$ (*e.g.*, $p_4(\mathbf{x}') + p_5(\mathbf{x}')$). We then define a risk-aware energy that trades off severity reduction, intervention economy, and coherence:

$$\begin{aligned}
\min_{\mathbf{x}^{\text{p}} \in \mathcal{P},\ \boldsymbol{\alpha} \in \mathcal{C}}\ \mathcal{E}(\mathbf{x}; \mathbf{x}^{\text{p}}, \boldsymbol{\alpha}) &\triangleq \underbrace{\text{sev}(\mathbf{x}') + \lambda_r\, r_{\text{high}}(\mathbf{x}')}_{\text{risk-aware severity}} \\
&+ \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 \|\mathbf{x}' - \mathbf{x}\|_2^2 + \lambda_c\, \Omega_{\text{coh}}(\boldsymbol{\alpha}; s) + \mathcal{R}_{\text{prior}}(\boldsymbol{\alpha}),
\end{aligned} \tag{11}$$

where $\mathbf{x}' = \mathcal{T}(\mathbf{x}, \mathbf{x}^{\text{p}}; \boldsymbol{\alpha})$ and Ω_{coh} is the within-signal coherence term defined in Sec. III. The ℓ_1 and ℓ_2 terms promote compact interventions and limit unnecessary deviation from the original state, complementing the hard constraints already encoded in $\mathcal{C}$.

We optimize Eq. (11) using a cross-entropy method (CEM) [26] in the group-action space. Each iteration samples a batch of actions from a parametric distribution, projects samples onto $\mathcal{C}$, evaluates $\mathcal{E}$, and updates the sampling distribution using elite samples. This produces a search-based teacher that is robust in nonconvex setting and naturally accommodates structured feasibility constraints induced by industrial semantics.

CEM can be bottlenecked by repeated calls to f_{cls}. We therefore train a lightweight surrogate $\tilde{f}_{\text{cls}}$ to approximate the grader's probabilistic output and use it for batched screening: all candidates are ranked by $\tilde{f}_{\text{cls}}$, and only the top-K are re-scored by f_{cls} to preserve decision fidelity. Moreover, we introduce two signal-level priors, a direction matrix $\mathbf{D}$ and an influence matrix $\mathbf{M}$ whose structure and audit semantics are designed from domain knowledge and the semantic signal mapping. The surrogate is used as a differentiable proxy to efficiently estimate sensitivity statistics required to instantiate these matrices. We inject the priors into mitigation via a regularizer that biases search toward direction-consistent and coupling-coherent interventions:

$$\mathcal{R}_{\text{prior}}(\boldsymbol{\alpha}) = \lambda_d\,\mathcal{R}_{\text{dir}}(\boldsymbol{\alpha};\mathbf{D}) + \lambda_m\,\mathcal{R}_{\text{inf}}(\boldsymbol{\alpha};\mathbf{M}). \tag{12}$$

This simultaneously improves optimization efficiency and enables post-hoc auditability.

C. Industrial Deployment: Offline Policy Distillation

The PhysCEM recourse teacher (Sec. IV-B) is decision-optimal under constraints, but its iterative sampling-and-selection procedure is computationally intensive for latency-critical online O&M. Therefore, we distill the teacher into a lightweight student policy that amortizes the recourse computation, while retaining a reliability mechanism that reverts to the teacher on rare hard cases.

We collect an offline dataset from teacher rollouts,

$$\mathcal{D}_T = \{(\mathbf{x}_t, \tilde{\boldsymbol{\alpha}}_t^T, \mathbf{x}_{t+1})\}_{t=1}^{N_T}, \tag{13}$$

where $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ are the pre-/post-edit feature states produced by the deterministic transition $\mathbf{x}_{t+1} = \mathcal{T}(\mathbf{x}_t, \mathbf{x}_t^{\text{p}}; \boldsymbol{\alpha}_t^T)$ (Eq. (4)) under the teacher's feasible actions, and $\tilde{\boldsymbol{\alpha}}_t^T \in [-1,1]^G$ denotes a normalized action representation. Teacher actions are naturally bounded ($0 \le \alpha_g \le \alpha_{\max}$); we map them into the policy output space via an affine scaling,

$$\tilde{\boldsymbol{\alpha}} \;=\; 2\frac{\boldsymbol{\alpha}}{\alpha_{\text{scale}}} - \mathbf{1}, \qquad \boldsymbol{\alpha} \;=\; \frac{\alpha_{\text{scale}}}{2}\,(\tilde{\boldsymbol{\alpha}} + \mathbf{1}), \tag{14}$$

followed by projection onto $\mathcal{C}$ when enforcing strict feasibility (Eq. (5)). Here α_{scale} is a conservative scale chosen from the allowed range and observed teacher magnitudes to avoid saturation at the boundaries of $\tanh(\cdot)$.

We train a student policy $\pi_\theta(\tilde{\boldsymbol{\alpha}} \mid \mathbf{x})$ using an entropy-regularized actor–critic objective on $\mathcal{D}_T$, augmented by imitation regularization toward the teacher action [9]. The actor is trained by minimizing

$$\begin{aligned} \min_\theta\; & \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_T}\big[\lambda_{\text{ent}} \log \pi_\theta(\tilde{\boldsymbol{\alpha}} \mid \mathbf{x}) \;-\; Q_\phi(\mathbf{x}, \tilde{\boldsymbol{\alpha}})\big] \\ & + \beta\,\mathbb{E}_{(\mathbf{x},\tilde{\boldsymbol{\alpha}}^T)\sim\mathcal{D}_T}\big[\|\tilde{\boldsymbol{\alpha}} - \tilde{\boldsymbol{\alpha}}^T\|_2^2\big], \end{aligned} \tag{15}$$

where λ_{ent} is a fixed entropy temperature for stable offline training and β controls imitation strength. Rewards used to learn Q_ϕ are computed from the grader outputs on the resulting counterfactual, consistent with the risk-aware shaping in Eq. (11).

At inference time, the student outputs $\tilde{\boldsymbol{\alpha}}$, which is mapped back to $\boldsymbol{\alpha}$ via Eq. (14) and projected onto $\mathcal{C}$. We then form a candidate counterfactual $\mathbf{x}' = \mathcal{T}(\mathbf{x}, \mathbf{x}^{\text{p}}; \boldsymbol{\alpha})$ using the same prototype retrieval module as the teacher. To ensure decision-grade reliability, we verify feasibility $\boldsymbol{\alpha} \in \mathcal{C}$ and attainment of the mitigation target under the original grader, *i.e.*, $\hat{y}(\mathbf{x}') \le y_{\text{tar}}$ (Eq. (6)). If verification fails, typically due to rare hard instances or distribution shift, we invoke the PhysCEM teacher to re-optimize $(\mathbf{x}^{\text{p}}, \boldsymbol{\alpha})$ under the exact objective and constraints. This student-first, teacher-on-demand strategy yields millisecond-level average latency while preserving high success rates in safety-critical deployment.

D. Time Complexity

Grader inference is $O(Td)$ per head for T trees of depth d, with a constant number of heads and $O(1)$ fusion/gating. For the teacher, each CEM iteration evaluates M candidates in G-dimensional group-action space; surrogate screening reduces expensive grader calls to top-K re-scoring, yielding $O(M \cdot C_{\text{surr}} + K \cdot C_{\text{cls}} + MF)$ per iteration (linear in the number of iterations). The student performs a single forward pass plus projection and one grader verification; teacher fallback is triggered only on a small fraction of cases in hybrid deployment.

V. Experiments

A. Experimental Setup

1) Datasets: All datasets are collected from real-world urban metro O&M systems and labeled into five ordinal wear grades. Table I summarizes the class distributions, showing a pronounced long tail for grade 5. Each sample corresponds to a monitored operational window/segment with synchronized multi-sensor measurements (*e.g.*, current, voltage, temperature, vibration, and other onboard/wayside channels). Raw signals are segmented into fixed-length windows and transformed into engineered features spanning multiple families (*e.g.*, statistics, differential, spectral, entropy).

TABLE I
Summary of three real-world metro wear datasets.

Dataset	Grade 1 (Healthy)	Grade 2 (Minor)	Grade 3 (Moderate)	Grade 4 (Significant)	Grade 5 (Severe)
MetroWear-A	2480	1976	2273	978	293
MetroWear-B	4998	3905	3565	2942	590
MetroWear-C	15003	11700	10304	9194	1799

2) Metrics: For wear grading, we report Acc, Macro-F1, QWK, and Sev-AUPRC (grade 5 vs rest). For severe wear recourse, we report Success (target achieved within constraints), Steps (lower is better), and Sparsity (fraction of edited groups; lower is better).

TABLE II
OVERALL GRADING PERFORMANCE ACROSS DATASETS (MEAN ± STD OVER 5 FOLDS).

Method	METROWEAR-A				METROWEAR-B				METROWEAR-C			
	Acc	Macro-F1	QWK	Sev-AUPRC	Acc	Macro-F1	QWK	Sev-AUPRC	Acc	Macro-F1	QWK	Sev-AUPRC
LGBM-1H	$0.58_{\pm0.003}$	$0.57_{\pm0.004}$	$0.57_{\pm0.005}$	$0.59_{\pm0.004}$	$0.62_{\pm0.003}$	$0.60_{\pm0.004}$	$0.63_{\pm0.003}$	$0.62_{\pm0.004}$	$0.67_{\pm0.004}$	$0.68_{\pm0.004}$	$0.67_{\pm0.004}$	$0.69_{\pm0.004}$
XGBoost	$0.59_{\pm0.003}$	$0.57_{\pm0.003}$	$0.58_{\pm0.002}$	$0.59_{\pm0.003}$	$0.62_{\pm0.003}$	$0.60_{\pm0.003}$	$0.63_{\pm0.005}$	$0.63_{\pm0.004}$	$0.67_{\pm0.003}$	$0.68_{\pm0.003}$	$0.67_{\pm0.003}$	$0.69_{\pm0.005}$
CatBoost	$0.60_{\pm0.002}$	$0.58_{\pm0.005}$	$0.59_{\pm0.003}$	$0.61_{\pm0.003}$	$0.63_{\pm0.005}$	$0.61_{\pm0.004}$	$0.63_{\pm0.004}$	$0.63_{\pm0.004}$	$0.68_{\pm0.003}$	$0.68_{\pm0.004}$	$0.68_{\pm0.002}$	$0.70_{\pm0.004}$
TabNet	$0.65_{\pm0.003}$	$0.64_{\pm0.004}$	$0.64_{\pm0.004}$	$0.66_{\pm0.003}$	$0.81_{\pm0.003}$	$0.79_{\pm0.003}$	$0.81_{\pm0.004}$	$0.81_{\pm0.004}$	$0.87_{\pm0.003}$	$0.87_{\pm0.003}$	$0.87_{\pm0.002}$	$0.88_{\pm0.002}$
FT-Transformer	$0.77_{\pm0.005}$	$0.76_{\pm0.004}$	$0.76_{\pm0.003}$	$0.78_{\pm0.003}$	$0.88_{\pm0.003}$	$0.86_{\pm0.004}$	$0.88_{\pm0.002}$	$0.88_{\pm0.005}$	$0.91_{\pm0.003}$	$0.91_{\pm0.004}$	$0.91_{\pm0.004}$	$0.93_{\pm0.003}$
TabR	$0.70_{\pm0.004}$	$0.69_{\pm0.005}$	$0.69_{\pm0.003}$	$0.71_{\pm0.004}$	$0.90_{\pm0.005}$	$0.89_{\pm0.004}$	$0.91_{\pm0.004}$	$0.90_{\pm0.002}$	$0.93_{\pm0.003}$	$0.93_{\pm0.004}$	$0.93_{\pm0.004}$	$0.94_{\pm0.002}$
LOS	$0.80_{\pm0.004}$	$0.80_{\pm0.003}$	$0.80_{\pm0.004}$	$0.83_{\pm0.002}$	$0.93_{\pm0.003}$	$0.92_{\pm0.005}$	$0.94_{\pm0.004}$	$0.93_{\pm0.002}$	$0.98_{\pm0.005}$	$0.97_{\pm0.002}$	$0.98_{\pm0.004}$	$0.98_{\pm0.002}$
PRISM (ours)	$\mathbf{0.83}_{\pm0.002}$	$\mathbf{0.82}_{\pm0.002}$	$\mathbf{0.82}_{\pm0.002}$	$\mathbf{0.84}_{\pm0.002}$	$\mathbf{0.94}_{\pm0.002}$	$\mathbf{0.92}_{\pm0.002}$	$\mathbf{0.95}_{\pm0.001}$	$\mathbf{0.94}_{\pm0.002}$	$\mathbf{0.98}_{\pm0.002}$	$\mathbf{0.98}_{\pm0.001}$	$\mathbf{0.98}_{\pm0.001}$	$\mathbf{0.99}_{\pm0.002}$

TABLE III
OVERALL MITIGATION PERFORMANCE ACROSS DATASETS (MEAN ± STD OVER 5 FOLDS). 'OOT' INDICATES TIMEOUT (>24H).

Method	METROWEAR-A			METROWEAR-B			METROWEAR-C		
	Success ↑	Steps ↓	Sparsity ↓	Success ↑	Steps ↓	Sparsity ↓	Success ↑	Steps ↓	Sparsity ↓
CEM-Plain	$91.30\%_{\pm2.62}$	$21.39_{\pm0.62}$	$62.52_{\pm1.65}$	OOT	OOT	OOT	OOT	OOT	OOT
DiCE-Actionable	$79.05\%_{\pm2.01}$	$8.97_{\pm0.79}$	$35.69_{\pm1.99}$	$81.92\%_{\pm2.88}$	$9.35_{\pm0.98}$	$\mathbf{32.72}_{\pm2.83}$	$77.89\%_{\pm2.93}$	$11.44_{\pm0.98}$	$\mathbf{35.76}_{\pm2.71}$
FACE	$64.75\%_{\pm1.34}$	$9.28_{\pm0.78}$	$107.28_{\pm2.87}$	$68.15\%_{\pm1.19}$	$12.05_{\pm0.81}$	$112.47_{\pm2.98}$	$50.80\%_{\pm2.39}$	$13.23_{\pm0.85}$	$107.84_{\pm1.72}$
IP-Recourse	$42.07\%_{\pm3.12}$	$\mathbf{1.57}_{\pm0.13}$	$\mathbf{16.10}_{\pm2.83}$	OOT	OOT	OOT	OOT	OOT	OOT
PRISM (Teacher)	$\mathbf{96.61}\%_{\pm0.97}$	$5.58_{\pm0.21}$	$43.27_{\pm0.79}$	$94.91\%_{\pm1.09}$	$6.52_{\pm0.11}$	$43.58_{\pm0.54}$	$\mathbf{94.50}\%_{\pm1.32}$	$7.67_{\pm0.17}$	$42.85_{\pm0.63}$
PRISM (Student)	$94.52\%_{\pm1.02}$	$5.03_{\pm0.25}$	$37.55_{\pm0.72}$	$91.26\%_{\pm1.12}$	$\mathbf{5.69}_{\pm0.12}$	$41.81_{\pm0.55}$	$92.51\%_{\pm1.03}$	$\mathbf{5.19}_{\pm0.21}$	$38.90_{\pm1.14}$
PRISM (Hybrid)	$\mathbf{96.61}\%_{\pm1.23}$	$5.92_{\pm0.30}$	$41.87_{\pm1.02}$	$\mathbf{94.92}\%_{\pm0.82}$	$6.85_{\pm0.26}$	$42.61_{\pm0.77}$	$\mathbf{94.50}\%_{\pm0.94}$	$7.99_{\pm0.12}$	$40.67_{\pm0.53}$

3) Baselines: Grading baselines include representative tree models and novel tabular deep models:

- **LGBM-1H** [7]: single-head LightGBM multi-class classifier with class weighting.
- **XGBoost** [33]: multi-class XGBoost with imbalance-aware weighting.
- **CatBoost** [14]: multi-class CatBoost model.
- **TabNet** [34]: an attentive deep tabular model.
- **FT-Transformer** [35]: transformer for tabular data.
- **TabR** [15]: retrieval-augmented tabular model.
- **LOS** [16] aims to address class imbalance; here, we combine this strategy with FT-Transformer for classification.

Mitigation baselines cover the following optimization-based and explanation-style recourse methods, evaluated under the same feasibility interface where applicable:

- **CEM-Plain** [8]: cross-entropy search without prototypes/priors and without surrogate top-K refinement.
- **DiCE-Actionable** [21]: diverse counterfactual recourse adapted to group/budget constraints.
- **FACE** [22]: sequential-based recourse.
- **IP-Recourse** [6]: integer-programming-based actionable recourse.

4) Implementation Details: For all models, the train-validation-test split of three datasets is 70%-10%-20%. The experimental results are averaged over 5–fold cross-validation. We implement our method in PyTorch and conduct experiments in Python 3.9 on one NVIDIA GPU with 30GB DRAM. In the parameter settings of PRISM, we set the fusion weight η to 0.45 and the gate thresholds τ to 0.25 in ImbaFusion-GBDT. In PhysCEM-Recourse, we set top-K = 250 to control the surrogate refinement, and balance the controller budgets through setting $B_{\text{tot}} = 5.5$.

B. Overall Performance: Wear Grading and Mitigation

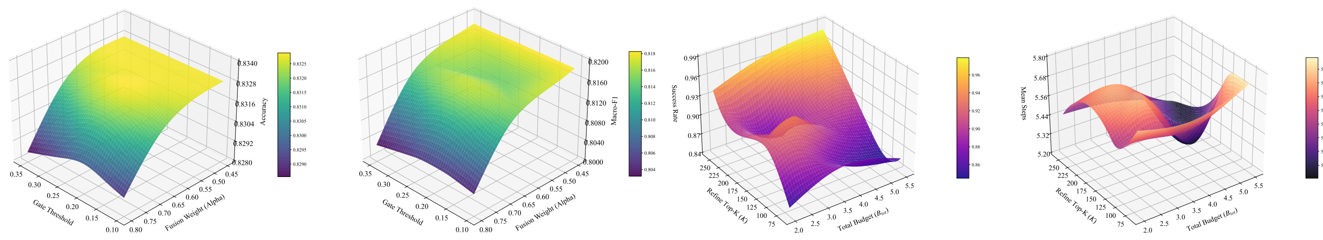
Table II shows that PRISM achieves the best grading performance across all datasets, with consistent gains on both ordinal-aware metrics (QWK) and the severe-focused Sev-AUPRC. On METROWEAR-A, PRISM improves Acc/Macro-F1/QWK over all baselines, indicating that multi-head fusion and conservative gating alleviate severe-class failure modes. On METROWEAR-B and METROWEAR-C, PRISM matches or exceeds the best-performing methods, while achieving the highest Sev-AUPRC, which is critical for safety-trigger decisions.

Table III reports mitigation results on severe cases. PRISM (Teacher) attains the highest success rates (94.5%–96.6%) with moderate steps and controlled sparsity. PRISM (Student) preserves high success (91.3%–94.5%) with fewer steps and lower sparsity, reflecting effective policy distillation. Compared with baselines, DiCE and FACE exhibit substantially lower success and much larger edits, while CEM-Plain and IP-Recourse become computationally impractical under our constraints on some datasets (OOT), underscoring the need for guided, constraint-aware optimization.

C. Ablation Studies

We perform ablations to attribute PRISM's gains to individual grading and mitigation components.

a) ImbaFusion-GBDT ablation: Figure 4 shows that incrementally adding *Multi-head*, *Fusion*, and *Gate* consistently improves all metrics on METROWEAR-B. Multi-head enhances ordinal consistency and tail sensitivity under extreme imbalance, while calibrated fusion reduces over-confident errors by consolidating complementary heads. The conservative gate further improves the reliability of high-stakes predictions by filtering uncertain severe outputs, mitigating severe-class collapse without hurting overall grading.



(a) (b) (c) (d)

Fig. 3. Parameter analysis of the proposed PRISM.

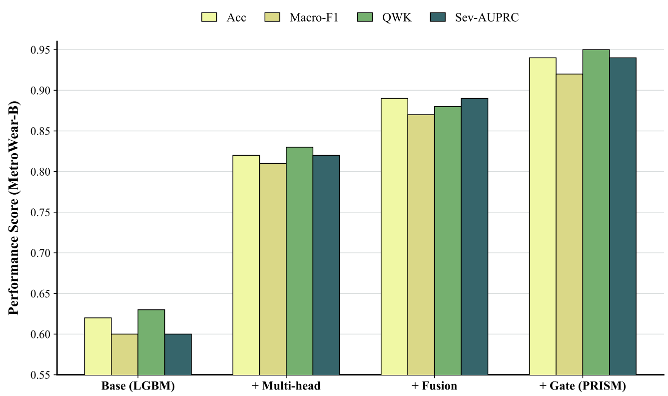


Fig. 4. Classifier ablation on METROWEAR-B.

b) PhysCEM-Recourse ablation: Table IV validates the necessity of PhysCEM components on METROWEAR-A. Removing prototype anchoring sharply degrades success and increases steps/edits, confirming anchoring as a strong plausibility prior for sparse, realistic interventions. Removing physics priors lowers success and increases steps, indicating that direction/influence guidance improves search efficiency and solution quality. Removing surrogate screening makes runtime prohibitive, whereas PRISM achieves the best accuracy–efficiency trade-off, which is essential for online operation.

TABLE IV
ABLATION STUDY OF THE PHYSCEM CONTROLLER COMPONENTS ON METROWEAR-A.

Method Variant	Success (↑)	Steps (↓)	Sparsity (↓)	Time_Mean (s) ↓
w/o Anchoring	85.4%$_{\pm 1.2}$	12.4$_{\pm 0.5}$	68.4$_{\pm 1.5}$	19.5$_{\pm 1.5}$
w/o Physics Priors	91.9%$_{\pm 1.2}$	11.4$_{\pm 0.5}$	62.4$_{\pm 1.5}$	15.8$_{\pm 1.5}$
w/o Surrogate	96.0%$_{\pm 1.2}$	5.7$_{\pm 0.5}$	43.7$_{\pm 1.5}$	> 120
PRISM	**96.6%$_{\pm 0.9}$**	**5.6$_{\pm 0.2}$**	**43.3$_{\pm 0.8}$**	**11.8$_{\pm 1.0}$**

D. Parameter Sensitivity and Robustness

We evaluate PRISM's robustness to key hyperparameters in the grader (fusion–gating) and the controller (refinement–budget). Fig. 3 shows a broad stable operating region: the grader's Accuracy/Macro-F1 varies smoothly with fusion and gating, indicating that calibrated fusion with conservative gating is not brittle to tuning. For mitigation, increasing the edit budget yields a predictable success–efficiency trade-off, while the refinement setting primarily affects runtime and steps with limited impact on success once in a reasonable range. Overall, these trends justify the default parameter ranges used in the main experiments and confirm that PRISM behaves stably under practical operating variations.

E. Case Study: Interpretability and Physical Priors

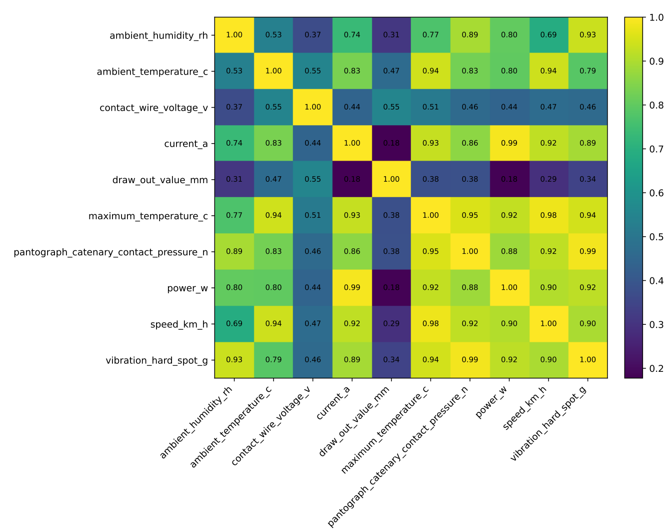


Fig. 5. Case study on METROWEAR-A: the influence-prior heatmap ($\mathbf{M}$) highlighting physically consistent cross-signal couplings for mitigation.

We conduct a case study on METROWEAR-A (Fig. 5) to verify that PRISM's recourse actions follow physically plausible cross-signal relations rather than arbitrary feature perturbations. The influence-prior heatmap $\mathbf{M}$ explicitly summarizes which sensor signals should move together in traction current collection (*e.g.*, `current_a`, `power_w`, `contact pressure_n`, and `maximum temperature_c` form a tightly coupled block, while ambient channels show weaker links). If an action tries to change one signal group but ignores its strongly coupled partners, the objective receives an extra penalty; in contrast, adjusting coupled signals in a coordinated way is promoted. As a result, the mitigation is physically coherent: it reduces current-related groups while simultaneously adjusting contact pressure/power groups (instead of unrelated features), which matches the engineering intuition that overheating/arcing risk is driven by the joint state of current draw, contact condition, and power response. This explains why the severe-risk score decreases step-by-step while keeping the edited *signal*×*family* set small and interpretable.

F. Deployment Efficiency and Distillation

Table V quantifies the success–latency trade-off. Teacher-only deployment achieves the best success but requires 11.82s on average due to iterative CEM. Student-only deployment reduces runtime to 0.61s with a small success drop. Hybrid deployment combines both: verification keeps the success at the teacher level, while the fallback rate is only 9% and the mean time is reduced to 4.18s.

TABLE V
DEPLOYMENT EFFICIENCY AND DISTILLATION EFFECTS ON METROWEAR-A.

Variant	Success ↑	Time_Mean(s) ↓	Fallback rate ↓
Teacher-only	96.61%	11.82	/
Student-only	94.52%	0.61	/
Hybrid	96.61%	4.18	9%

VI. Conclusion

We presented PRISM, a deployable AI framework for urban metro contact-wire wear that couples imbalance-aware multi-class grading with physics-informed counterfactual mitigation and offline policy RL distillation. PRISM unifies calibrated multi-head grading, prototype-anchored PhysCEM search guided by interpretable direction and influence priors, and a student policy with verification and teacher fallback. Experiments on real-world metro data demonstrate strong grading performance, high mitigation success, and millisecond-level student inference with verification/fallback reliability.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62272398) and the Sichuan Science and Technology Program (Grant No. 2024NS-FJQ0019).

References

[1] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, 2006.

[2] S. Bruni, G. Bucca, M. Carnevale, A. Collina, and A. Facchinetti, "Pantograph–catenary interaction: recent achievements and future research challenges," *Int. J. Rail Transp.*, vol. 6, no. 2, 2018.

[3] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017, pp. 1321–1330.

[5] T. S. I. M. LEARNING, "Dataset shift in machine learning," 2009.

[6] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *FAccT*, 2019, pp. 10–19.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *NIPS*, vol. 30, 2017.

[8] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodol. Comput. Appl. Probab.*, vol. 1, no. 2, pp. 127–190, 1999.

[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *ICML*, 2018, pp. 1861–1870.

[10] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *NIPS*, vol. 33, pp. 1179–1191, 2020.

[11] J. He, J. He, Z. Gou, and F. Lv, "Defect detection in railway catenary systems: a survey," *ITI*, p. liaf027, 2025.

[12] G. Bucca and A. Collina, "A procedure for the wear prediction of collector strip and contact wire in pantograph–catenary system," *Wear*, vol. 266, no. 1-2, pp. 46–59, 2009.

[13] G. Wu, K. Dong, Z. Xu, S. Xiao, W. Wei, H. Chen, J. Li, Z. Huang, J. Li, G. Gao *et al.*, "Pantograph–catenary electrical contact system of high-speed railways: recent progress, challenges, and outlooks," *Railw. Eng. Sci.*, vol. 30, no. 4, pp. 437–467, 2022.

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *NIPS*, vol. 31, 2018.

[15] Y. Gorishniy, I. Rubachev, N. Kartashev, D. Shlenskii, A. Kotelnikov, and A. Babenko, "TabR: Unlocking the power of retrieval-augmented tabular deep learning," *arXiv*, vol. 2307.14338, 2023.

[16] S. Sun, H. Lu, J. Li, Y. Xie, T. Li, X. Yang, L. Zhang, and J. Yan, "Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance," in *ICLR*.

[17] Z. Liu, Z. Li, Z. Yang, T. Wei, J. Kang, Y. Zhu, H. Hamann, J. He, and H. Tong, "CLIMB: Class-imbalanced learning benchmark on tabular data," *arXiv*, vol. 2505.17451, 2025.

[18] H. Yan, Y. Qian, F. Peng, J. Luo, F. Li *et al.*, "Neural collapse to multiple centers for imbalanced data," *NIPS*, vol. 37, pp. 65 583–65 617, 2024.

[19] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[20] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: from counterfactual explanations to interventions," in *FAccT*, 2021, pp. 353–362.

[21] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAccT*, 2020, pp. 607–617.

[22] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: feasible and actionable counterfactual explanations," in *AIES*, 2020, pp. 344–350.

[23] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *DMKD*, vol. 38, no. 5, 2024.

[24] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nat. Rev. Phys.*, vol. 3, no. 6, 2021.

[25] T. Kapoor, H. Wang, A. Núñez, and R. Dollevoet, "Physics-informed machine learning for moving load problems," in *J. Phys. Conf. Ser.*, vol. 2647, no. 15, 2024.

[26] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*, 2004.

[27] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv*, vol. 2005.01643, 2020.

[28] S. Lei, S. Zhang, and D. Tao, "Offline behavior distillation," *NIPS*, vol. 37, pp. 115 098–115 118, 2024.

[29] T. Chen, Z. Wang, and M. Zhou, "Diffusion policies creating a trust region for offline reinforcement learning," *NIPS*, vol. 37, pp. 50 098–50 125, 2024.

[30] Z. Liu, Z. Guo, Y. Yao, Z. Cen, W. Yu, T. Zhang, and D. Zhao, "Constrained decision transformer for offline safe reinforcement learning," in *ICML*, 2023, pp. 21 611–21 630.

[31] Q. Lin, B. Tang, Z. Wu, C. Yu, S. Mao, Q. Xie, X. Wang, and D. Wang, "Safe offline reinforcement learning with real-time budget constraints," in *ICML*, 2023, pp. 21 127–21 152.

[32] Z. Guo, W. Zhou, and W. Li, "Temporal logic specification-conditioned decision transformer for offline safe reinforcement learning," *arXiv*, vol. 2402.17217, 2024.

[33] T. Chen, "XGBoost: A scalable tree boosting system," *Cornell University*, 2016.

[34] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *AAAI*, vol. 35, no. 8, 2021, pp. 6679–6687.

[35] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *NIPS*, vol. 34, pp. 18 932–18 943, 2021.