PRISM: Asymmetric Precision-Recall Optimization for Million-Scale Root Cause Analysis

Wei Zhang[1*] **Rongyao Cai**[1,2*] **Xiaosheng Li**[1] **Hang Yu**[1†] **Rui Wang**[1†]

[1]**Ant Group** [2]**Zhejiang University**

[*]**Equal contribution** [†]**Corresponding authors**

Abstract

Multi-dimensional Root Cause Analysis (RCA) requires searching an exponential $\mathcal{O}(2^M)$ attribute-subset space over millions of records within seconds to isolate the root-cause combination. Existing methods fail at this industrial scale because heuristic-search approaches scan and score the massive normal-dominated dataset symmetrically ($\mathcal{O}(|\mathcal{U}| \cdot M)$ with $\mathcal{O}(2^M)$ search), while pairwise optimization methods incur prohibitive $\mathcal{O}(|\mathcal{U}|^2)$ normal-normal comparisons. To overcome both bottlenecks, we propose **PRISM** (Precision-Recall Informed Subspace Mining), which reformulates RCA as an asymmetric Precision-Recall optimization. Precision is defined over anomaly-normal pairs and Recall over anomaly-anomaly pairs, eliminating uninformative normal-normal comparisons entirely. Coupled with one-pass tensor compression and a two-stage solver—combining Bernoulli straight-through subspace pruning with exact combinatorial verification—PRISM reduces iterative pairwise complexity to an anomaly-bounded $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ and exact search to $\mathcal{O}(2^m)$ ($m \leq 5$). Evaluations on four public benchmarks and a new 5.13M-record industrial dataset show that PRISM achieves the highest public-benchmark average F1 (0.92 vs. 0.91 for RiskLoc and 0.81 for PSqueeze), outperforms all baselines on a tractable 20K-record production control (0.82 vs. 0.53 F1), and is the only method sustaining high accuracy (0.82 F1) and sub-second latency (0.29 s) across a 2^{110} attribute-value combination space at full 5.13M-record scale. We publicly release our code and this dataset as **NFC**, establishing a production-sourced, million-scale benchmark for multi-dimensional RCA. [Code] [Dataset]

1 Introduction

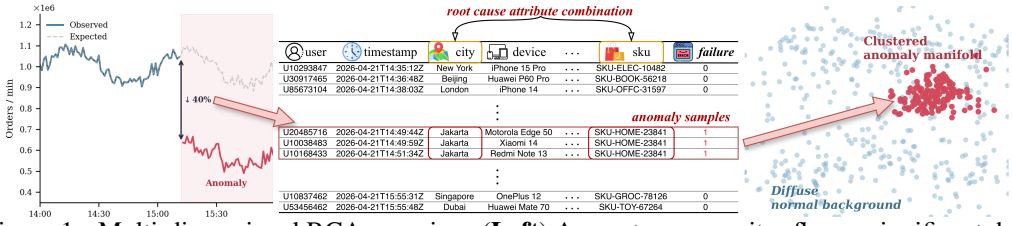


Figure 1: Multi-dimensional RCA overview. (**Left**) An upstream monitor flags a significant drop (e.g., 40%) in a business metric. (**Center**) Diagnostic telemetry contains M categorical attributes, where the root cause resides in a specific attribute-value combination (highlighted). (**Right**) Projected onto the root-cause subspace, anomalous samples cluster tightly while normal samples remain diffuse, motivating asymmetric optimization.

Consider an e-commerce platform where the order success rate abruptly plummets by 40%. Such incidents trigger urgent investigations into massive telemetry datasets: often millions of records, each tagged with M categorical attributes such as Region, Device, and OS. Operators must rapidly identify the *root cause combination* (*e.g.*, Region=Beijing $\wedge$ Device=Mobile) that explains the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

anomaly. In production, every minute of unresolved diagnosis delay translates directly into revenue loss, degraded user experience, and cascading failures across dependent services. This task, known as *multi-dimensional Root Cause Analysis* (RCA) [1–5], requires searching an exponential $\mathcal{O}(2^M)$ attribute-subset space to isolate a compact subset $\mathcal{C}^* \subseteq \{1, \ldots, M\}$ where anomalous samples exhibit a coherent cluster (Figure 1, Center & Right).

Modern operational scales are staggering: platforms like Google Pay and Alipay process hundreds of millions of transactions daily [6–8]. Within a standard diagnostic window, telemetry can exceed millions of records across 10–15 dimensions, yielding attribute-value combinations surpassing 2^{100}. Beyond industrial incident response, isolating specific attribute combinations under system interventions connects closely to causal mechanism shift detection and explainable attribution [9–12]. At million-record scale, resolving diagnoses within seconds is an operational requirement for system availability.

Over the past decade, attribute-combination RCA methods have progressed from hand-crafted scoring with heuristic cuboid search [2, 3, 13–16] to learnable attribute weighting [17]. As summarized in Sec. 2 and Table 1, published evaluations typically operate on $\sim 10^4$–10^5 records and 3–6 dimensions. Algorithms tailored for that regime break down when confronted with the 10–100$\times$ simultaneous growth in record volume, attribute dimensionality, and value cardinality common in modern production telemetry.

We argue that this scalability bottleneck stems from a fundamental **modeling oversight** (detailed in Sec. 2 and Sec. 4.1). RCA data exhibits an inherent structural asymmetry: *anomalous samples ($\mathcal{P}$) aggregate tightly on the root-cause attributes, whereas normal samples ($\mathcal{N}$) remain stochastically diffused* (Figure 1, Right). Yet existing paradigms fail to exploit this role asymmetry in two distinct ways (Table 1): (i) **Heuristic-search methods** (*e.g.*, Squeeze [3], HotSpot [2], RiskLoc [16]) treat $\mathcal{P}$ and $\mathcal{N}$ symmetrically when scanning and aggregating cuboids over the full, normal-dominated dataset ($\mathcal{O}(|\mathcal{U}| \cdot M)$ with $|\mathcal{N}| \gg |\mathcal{P}|$), while still facing $\mathcal{O}(2^M)$ subset or value-grid traversal; and (ii) **Pairwise and metric-learning methods** (*e.g.*, MID [18], RC-LIR [17]) incur $\mathcal{O}(|\mathcal{U}|^2)$ symmetric comparisons in which uninformative normal-normal (N-N) pairs carry zero root-cause signal yet consume the vast majority of computation. We ask: ***Can multi-dimensional RCA be deployed genuinely at million-scale under real-time constraints?***

Yes! To address both bottlenecks, we propose **PRISM** (Precision-Recall Informed Subspace Mining), which reformulates RCA as an *asymmetric Precision-Recall joint optimization*. By assigning $\mathcal{P}$ and $\mathcal{N}$ distinct relational roles—focusing Precision on inter-class separation (P-N) and Recall on intra-class cohesion (P-P)—PRISM eliminates uninformative N-N comparisons entirely and enables gradient-guided subspace pruning. After a single linear compression pass, PRISM reduces iterative pairwise complexity from $\mathcal{O}(|\mathcal{U}|^2 \cdot M)$ to an anomaly-bounded $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ and collapses combinatorial search from $\mathcal{O}(2^M)$ to $\mathcal{O}(2^m)$ ($m \leq 5$). In summary, our contributions are:

- **Asymmetric Problem Formulation.** We reformulate multi-dimensional RCA as an asymmetric Precision-Recall optimization that decouples inter-class separation (P-N) from intra-class cohesion (P-P), excluding all uninformative N-N pairs. Unlike prior methods that fold normal/anomalous partitions into a single heuristic risk score [16] or symmetric pairwise margins [17], PRISM embeds role asymmetry directly into a differentiable subspace-selection objective.
- **Scalable Optimization Framework.** We combine one-pass tensor compression with a two-stage solver integrating Bernoulli straight-through pruning and exact combinatorial verification. We establish an $\mathcal{O}(1/\sqrt{T})$ convergence guarantee for the idealized proximal relaxation and exact optimality within the retained subspace, bounding post-compression pairwise cost to $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ for sub-second million-scale diagnosis.
- **Million-Scale Evaluation and Open Industrial Benchmark.** Across four public benchmarks, a tractable 20K-record production control (**NFC-20K**), and the full 5.13M-record industrial dataset (**NFC**), PRISM achieves the highest public-benchmark average F1 (0.92 vs. 0.91 for RiskLoc and 0.81 for PSqueeze), leads all evaluated baselines on NFC-20K (0.82 vs. 0.53 F1), and uniquely sustains 0.82 F1 in 0.29 s on full NFC where prior baselines time out (>2 h) or collapse. We publicly release **NFC** as a production-sourced, million-scale RCA benchmark.

Table 1: Complexity comparison across representative RCA regimes. $|\mathcal{U}|=|\mathcal{P}|+|\mathcal{N}|$: total records; $|\mathcal{P}|$: anomalous records; M: attributes; C_j: cardinality of attribute j; $|\text{agg}|$: materialized leaf aggregates; m: pruned subspace size ($m \leq 5$). PRISM performs one linear compression scan before $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ optimization.

Paradigm	Representatives	Scan / Prep. Cost	Pairwise Cost	Search / Enumeration Cost	Asym. Obj.
Marginal / Cuboid heuristic	Adtributor, HotSpot, Squeeze, PSqueeze, RiskLoc	$\mathcal{O}(\|\mathcal{U}\| \cdot M)$	—	$\mathcal{O}(\sum_j C_j)$ to $\mathcal{O}(2^M \prod_j C_j)$	✗
Graph propagation	SinkFlow	$\mathcal{O}(\|\mathcal{U}\| \cdot M)$	—	$\mathcal{O}(\|\text{agg}\| \cdot 2^M)$	✗
Meta-heuristic / Learnable	MID, RC-LIR (+ solver)	$\mathcal{O}(\|\mathcal{U}\| \cdot M)$	$\mathcal{O}(\|\mathcal{U}\|^2 \cdot M)$	Heuristic / Downstream $\mathcal{O}(2^M)$	✗
PRISM (ours)	—	$\mathcal{O}(\|\mathcal{U}\| \cdot M)$	$\mathcal{O}(\|\mathcal{P}\|^2 \cdot M)$	$\mathcal{O}(2^m),\ m \leq 5$	✓

2 Related Works

Multi-Dimensional Attribute-Combination RCA. Multi-dimensional attribute-combination RCA aims to isolate specific attribute-value patterns associated with an upstream alert [1, 5]. Most prior methods follow a **score-then-search** paradigm. **Heuristic search methods** rely on hand-crafted deviation or risk scores: Adtributor [1] and R-Adtributor [19] rank candidates via KL-divergence explanatory power; iDice [14] measures isolation power via entropy; HotSpot [2] and the Squeeze family [3, 13, 20] exploit the *ripple effect* via Generalized Potential Scores and clustering-based pruning; RiskLoc [16] partitions leaf elements into normal and abnormal sets to compute weighted risk scores; FTM-RCA [15] filters anomaly propagation; SinkFlow [21] propagates scores over an event aggregation graph (EAG); and MID [18] applies meta-heuristic search over entropy-based pairwise structures. We separately highlight **learnable-weighting methods** because they represent the closest transition from manual heuristics toward data-driven optimization: RC-LIR [17] adapts I-Relief [22] to learn margin-based attribute weights via same-class (hit) and different-class (miss) neighbors before invoking a downstream heuristic solver, while CMMD [23] models cross-metric dependencies via graph attention. However, heuristic methods still scan normal-dominated cuboids ($\mathcal{O}(|\mathcal{U}| \cdot M)$) and suffer exponential $\mathcal{O}(2^M)$ or value-grid search under high cardinality—even RiskLoc [16], despite using a two-way normal/abnormal leaf split, collapses them into a heuristic cuboid search rather than differentiable subspace pruning—while pairwise formulations [17, 18] incur $\mathcal{O}(|\mathcal{U}|^2)$ N-N overhead (Table 1). PRISM instead formulates RCA itself as an asymmetric differentiable optimization with anomaly-bounded $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ iterative cost.

Service-Level and Causal/LLM Diagnosis vs. Attribute-Level Localization. A complementary body of research studies RCA at the *service*, *variable*, or *incident* level [5], including dependency-graph traversal [24–26], causal structure and mechanism-shift discovery [9–12], and LLM-driven semantic triage such as RCACopilot [27] and OpenRCA [28]. These methods primarily answer *where* a failure originates across system components, whereas PRISM pinpoints *what* exact multi-dimensional attribute signature (*e.g.*, user segment, device batch, firmware) drives the failure within a component's telemetry. Because LLMs cannot natively traverse 2^{100} structured combinations within sub-second budgets, PRISM complements high-level semantic or causal triage as a fast, exact combinatorial localizer. Further discussion connecting PRISM to *differentiable subset selection* and *asymmetric loss design* is provided in Appendix A.

3 Preliminary

Let $\mathcal{U} = \{x_1, \ldots, x_N\}$ denote a dataset of $N = |\mathcal{U}|$ telemetry records (or leaf aggregates) characterized by M categorical attributes. Following standard multi-dimensional RCA protocols [2, 3, 13, 16], localization is triggered downstream of an anomaly-detection module that flags an incident window and assigns each record x_i a binary status $L_i \in \{0, 1\}$, where $L_i = 1$ signifies an anomalous record and $L_i = 0$ a normal record. We denote the realized value of the j-th attribute for sample x_i as $A_j^{(i)}$. An *attribute combination* $\mathcal{C} \subseteq \{1, \ldots, M\}$ defines a specific dimensional subspace of the feature space. For any sample x_i, let $A^{(i)}(\mathcal{C})$ denote the projected tuple (or signature) of values in this subspace (*e.g.*, (Beijing, Mobile)). Furthermore, for any subset $\mathcal{S} \subseteq \mathcal{U}$, we define $A^{\mathcal{S}}(\mathcal{C}) = \{A^{(i)}(\mathcal{C}) \mid x_i \in \mathcal{S}\}$ as the set of attribute-value signatures observed in $\mathcal{S}$ under subspace $\mathcal{C}$.

Based on the upstream labels, we partition $\mathcal{U}$ into the positive anomalous set $\mathcal{P} = \{x_i \in \mathcal{U} \mid L_i = 1\}$ and the negative normal set $\mathcal{N} = \{x_i \in \mathcal{U} \mid L_i = 0\}$, where $\mathcal{U} = \mathcal{P} \cup \mathcal{N}$. Following standard convention, for any logical predicate or property ϕ, we denote $\mathcal{N}(\phi) \subseteq \mathcal{N}$ and $\mathcal{P}(\phi) \subseteq \mathcal{P}$ as the subsets of normal and anomalous samples satisfying ϕ, respectively.

Problem Statement. Given $\mathcal{U} = \mathcal{P} \cup \mathcal{N}$ with M categorical attributes $\{\mathcal{A}_1, \ldots, \mathcal{A}_M\}$, the goal of multi-dimensional RCA is to identify a compact *root cause attribute combination* $\mathcal{C}^* \subseteq \{1, \ldots, M\}$ and its corresponding value signature V^* under which anomalous samples $\mathcal{P}$ collapse onto a coherent pattern while normal samples $\mathcal{N}$ remain diffuse. When an incident involves multiple concurrent root causes, localization proceeds via sequential peeling—isolating the dominant signature, filtering its explained anomalous records, and repeating on the residual anomalies.

4 Methodology

PRISM achieves million-scale, sub-second diagnosis through a three-level complexity reduction: (1) **Asymmetric Formulation:** An objective over P-N and P-P pairs that eliminates N-N comparisons. (2) **Tensor Compression:** A one-pass reduction that decouples iterative tensor cost from background volume $|\mathcal{N}|$. (3) **Two-Stage Optimization:** A solver where *Stage 1* prunes $\mathcal{O}(2^M)$ to a compact candidate support $\mathcal{A}_{\text{cand}}$ ($|\mathcal{A}_{\text{cand}}| \leq m$) via Bernoulli relaxation and *Stage 2* selects the exact F1-optimal subset within $\mathcal{A}_{\text{cand}}$.

4.1 Asymmetric Precision-Recall Objective

(1) Structural Asymmetry in Attribute Subspaces. The foundation of PRISM lies in exploiting the inherent topological asymmetry of telemetry data. We formalize this through two key observations:

Observation 4.1 (Asymmetric Aggregation of Anomalies). Under the ground-truth root cause combination $\mathcal{C}^*$, anomalous samples ($\mathcal{P}$) aggregate into a dense subspace, sharing a coherent attribute-value signature. Conversely, normal samples ($\mathcal{N}$) remain stochastically diffused, exhibiting no persistent clustering patterns under any attribute subset $\mathcal{C}$.

Observation 4.2 (Label Contamination). In production environments, $\mathcal{P}$ is rarely clean or exhaustive. Due to limited monitoring coverage, asynchronous logging, or upstream alert noise, true anomalies may "bleed" into $\mathcal{N}$ (and vice versa), creating label noise that undermines strict margin classifiers.

Observation 4.1 reflects a structural consensus implicit in prior RCA principles—such as *isolation power* in iDice [14] and the *ripple effect* in HotSpot and Squeeze [2, 3]—and is empirically validated across both NFC (Figure 7) and all four public benchmarks via t-SNE projections (Appendix H.2, Figure 9). Together, Observation 4.1 renders $\mathcal{N}$-$\mathcal{N}$ modeling redundant, while Observation 4.2 cautions against rigid sample-wise margins. Consequently, we replace dense $\mathcal{O}(|\mathcal{U}|^2)$ symmetric comparisons with an asymmetric dual objective: (1) **Precision (Inter-Class Separation)**: Minimizes signature overlap between $\mathcal{N}$ and $\mathcal{P}$ to isolate the root cause. (2) **Recall (Intra-Class Cohesion)**: Enforces structural aggregation within $\mathcal{P}$ under the selected dimensions.

(2) Formalizing Precision and Recall on Attribute Manifolds. To operationalize these, we first define an exact binary selection vector $\mathbf{w} \in \{0,1\}^M$, where $w_j = 1$ denotes the inclusion of attribute $\mathcal{A}_j$. In the continuous relaxation, we reuse $\mathbf{w} \in [0,1]^M$ as a selection-weight vector.

Definition 4.3 (Pairwise Attribute Distance)**.** For $x_1, x_2 \in \mathcal{U}$ and an exact binary selection vector $\mathbf{w} \in \{0,1\}^M$, the weighted pairwise attribute distance is:

$$d(x_1, x_2;\, \mathbf{w}) = \text{Sign}\Big[\sum_{j=1}^{M} w_j \cdot \mathbf{1}\big[A_j^{(1)} \neq A_j^{(2)}\big]\Big] \tag{1}$$

where $\text{Sign}(x) = \mathbf{1}[x > 0]$ is the Heaviside step function. $d(x_1, x_2; \mathbf{w}) = 0$ if and only if x_1 and x_2 share identical attribute values across all attributes selected by $\mathbf{w}$, and equals 1 otherwise.

To efficiently compute this over $\mathcal{U}$, we define $D^{\mathcal{N},\mathcal{P}} \in \{0,1\}^{|\mathcal{N}|\times|\mathcal{P}|\times M}$ as the attribute-wise difference tensor between $\mathcal{N}$ and $\mathcal{P}$, where $D^{\mathcal{N},\mathcal{P}}_{i,j,k} = \mathbf{1}[A_k^{(i)} \neq A_k^{(j)}]$ for $x_i \in \mathcal{N}$, $x_j \in \mathcal{P}$, and attribute $\mathcal{A}_k$. We also define the intra-anomaly difference tensor $D^{\mathcal{P},\mathcal{P}} \in \{0,1\}^{|\mathcal{P}|\times|\mathcal{P}|\times M}$. For each pair (i,j), the slice $D_{i,j} \in \{0,1\}^M$ gives $d(x_i, x_j; \mathbf{w}) = \mathbf{1}[\mathbf{w}^\top D_{i,j} \neq 0]$.

Definition 4.4 (Precision & Recall)**.** We define Precision as the fraction of normal samples distinct from the anomalous cluster, and Recall as the internal cohesion of that cluster:

$$\text{Precision}(\mathbf{w}) = 1 - \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \mathbf{1}\Big[\min_{j\in\mathcal{P}} \mathbf{w}^\top D_{i,j}^{\mathcal{N},\mathcal{P}}=0\Big],\ \text{Recall}(\mathbf{w}) = 1 - \frac{1}{|\mathcal{P}|^2}\sum_{i,j\in\mathcal{P}} \text{Sign}\big(\mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}}\big) \tag{2}$$

Remark. Unlike standard classification metrics, our Precision and Recall are defined directly on attribute manifolds. Precision measures inter-class separation, while Recall measures intra-class cohesion. This formulation inherently avoids uninformative $\mathcal{N}$-$\mathcal{N}$ comparisons, reducing the pairwise workload from $\mathcal{O}(|\mathcal{U}|^2)$ to $\mathcal{O}(|\mathcal{N}||\mathcal{P}| + |\mathcal{P}|^2)$.

(3) Continuous Relaxation and Tractable Optimization. Before exact verification, Stage 1 uses the following constrained Precision–Recall criterion to identify a compact candidate support:

Problem 4.5 (Stage 1 Pruning Criterion). Find a pruning-oriented binary selection vector $\mathbf{w}_{\text{prune}} \in \{0,1\}^M$ that favors inter-class separation while preserving anomaly cohesion:

$$\mathbf{w}_{\text{prune}} = \underset{\mathbf{w}\in\{0,1\}^M}{\arg\max}\ \text{Precision}(\mathbf{w}), \quad \text{s.t.} \quad \text{Recall}(\mathbf{w}) \geq \tau \tag{3}$$

where $\tau \in (0,1)$ is the tolerance for intra-class cohesion.

Stage 1 therefore outputs only a candidate support; Stage 2 makes the final RCA selection by exact F1 verification over that support. Directly solving Eq. (3) is intractable due to the exponential search space $\{0,1\}^M$ and the non-differentiable nature of the Sign and $\min$ operators. To enable efficient gradient-based optimization, we derive a surrogate objective through three relaxation steps.

Differentiable Surrogate for Precision. Maximizing precision is equivalent to minimizing the fraction of negative samples that collapse into the anomalous cluster. Approximating the zero-distance indicator $\mathbf{1}[X=0]$ for $X = \mathbf{w}^\top D_{i,j}^{\mathcal{N},\mathcal{P}} \geq 0$ via the smooth exponential kernel $\exp(-X/\sigma)$ ($\sigma > 0$) and averaging across $\mathcal{N}\times\mathcal{P}$ yields the smooth Precision Loss:

$$q(\mathbf{w}) = \frac{1}{|\mathcal{N}||\mathcal{P}|}\sum_{i\in\mathcal{N},\, j\in\mathcal{P}} \exp(-\mathbf{w}^\top D_{i,j}^{\mathcal{N},\mathcal{P}}/\sigma) \tag{4}$$

Minimizing $q(\mathbf{w})$ effectively drives $\mathbf{w}^\top D_{i,j}^{\mathcal{N},\mathcal{P}}$ away from zero for $i \in \mathcal{N}$, thereby maximizing inter-class separation. A formal scaled sandwich bound relating $q(\mathbf{w})$ to the exact precision error is provided in Proposition I.14 (Appendix).

Differentiable Surrogate for Anomaly Cohesion. The $\text{Sign}(\cdot)$ function in the Recall definition (Eq. (2)) inhibits gradient flow. Given that $\mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}}$ is inherently non-negative for $\mathbf{w} \in [0,1]^M$, we replace $\text{Sign}(\cdot)$ with a linear surrogate. This yields a differentiable cohesion loss $h(\mathbf{w})$:

$$h(\mathbf{w}) \triangleq \frac{1}{|\mathcal{P}|^2}\sum_{i,j\in\mathcal{P}} \text{ReLU}\left(\mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}}\right) \tag{5}$$

This term supplies the continuous descent signal for intra-anomaly cohesion. The recall threshold τ is used separately for deterministic recall pruning and feasibility interpretation through $h(\mathbf{w}) \leq 1-\tau$.

Final Unified Objective. To bridge the discrete-continuous gap, we relax the domain of $\mathbf{w}$ to $[0,1]^M$ and introduce an ℓ_1 regularization term $\eta\|\mathbf{w}\|_1$ to promote sparsity. The final objective is:

$$\min_{\mathbf{w}\in[0,1]^M} \mathcal{L}(\mathbf{w}), \quad \mathcal{L}(\mathbf{w}) \triangleq q(\mathbf{w}) + \lambda h(\mathbf{w}) + \eta\|\mathbf{w}\|_1 \tag{6}$$

where $\lambda > 0$ enforces intra-anomaly cohesion and $\eta > 0$ regularizes support size in the idealized relaxation. For later use, denote $f(\mathbf{w}) \triangleq q(\mathbf{w}) + \lambda h(\mathbf{w})$ and $g(\mathbf{w}) \triangleq \eta\|\mathbf{w}\|_1$, so $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + g(\mathbf{w})$.

By construction, $\mathcal{L}(\mathbf{w})$ exclusively involves P-N and P-P pairs, successfully eliminating the uninformative $\mathcal{N}$-$\mathcal{N}$ background. However, $q(\mathbf{w})$ still formally scales with $|\mathcal{N}|$. In the next subsection, we introduce Tensor Compression to further decouple the optimization cost from the raw record count.

Gradient Dynamics. The final objective admits a natural push-pull decomposition governing attribute selection:

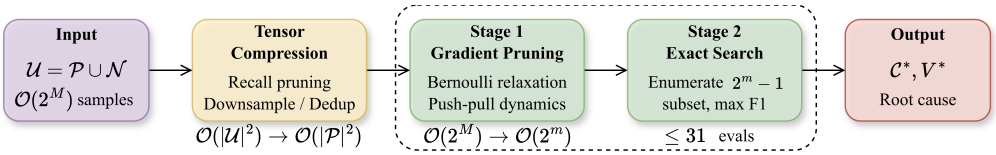


Figure 2: PRISM pipeline. Tensor compression reduces pairwise complexity from $\mathcal{O}(|\mathcal{U}|^2 \cdot M)$ to $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$. Stage 1 prunes the $\mathcal{O}(2^M)$ search space to $\mathcal{O}(2^m)$ via Bernoulli relaxation. Stage 2 guarantees global optimality within the pruned subspace by exhaustive verification (≤ 31 evaluations).

(i) **The Discriminative Push ($q(\mathbf{w})$):** The smooth precision loss has gradient

$$\mathbf{g}_q = -\frac{1}{|\mathcal{N}||\mathcal{P}|\sigma} \sum_{i \in \mathcal{N},\, j \in \mathcal{P}} \exp(-\mathbf{w}^\top D_{i,j}^{\mathcal{N},\mathcal{P}}/\sigma) \cdot D_{i,j}^{\mathcal{N},\mathcal{P}} \preceq \mathbf{0} \tag{7}$$

which is elementwise non-positive and therefore **amplifies attributes that separate $\mathcal{P}$ from $\mathcal{N}$**.

(ii) **The Cohesive Pull ($h(\mathbf{w})$):** A subgradient $\mathbf{g}_h \in \partial h(\mathbf{w})$ of the recall penalty is:

$$\mathbf{g}_h = \frac{1}{|\mathcal{P}|^2} \sum_{i,j \in \mathcal{P}} \mathbb{I}(i,j \mid \mathbf{w}) \cdot D_{i,j}^{\mathcal{P},\mathcal{P}} \succeq \mathbf{0}, \quad \mathbb{I}(i,j \mid \mathbf{w}) = \begin{cases} 1, & \text{if } \mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

This term is elementwise non-negative and **suppresses attributes fragmenting the anomaly cluster**.

4.2 Tensor Compression

The second level of complexity reduction operates on the data itself, decoupling computational cost from the large background volume $|\mathcal{N}|$, where typically $|\mathcal{N}| \gg |\mathcal{P}|$. After the asymmetric objective removes all $\mathcal{N}$-$\mathcal{N}$ terms, the remaining bottleneck is purely physical: materializing the sample-level tensors $D^{\mathcal{N},\mathcal{P}}$ and $D^{\mathcal{P},\mathcal{P}}$ still costs $\mathcal{O}(|\mathcal{N}| \cdot |\mathcal{P}| \cdot M)$. PRISM therefore compresses the data, replacing raw samples with a smaller representative set while preserving the anomaly-relevant topology.

(i) **Negative downsampling**: since $\mathcal{N}$ is diffuse (Observation 4.1), its gradient signal saturates early once enough normal variability has been observed. We therefore cap $|\mathcal{N}|' = \min(|\mathcal{N}|, \gamma \cdot |\mathcal{P}|)$, which immediately reduces subsequent tensor cost to $\mathcal{O}((\gamma + 1) \cdot |\mathcal{P}|^2 \cdot M)$.

(ii) **Long-tail anonymization**: attribute values appearing only in $\mathcal{N}$ are replaced with a universal `<other>` token. From the perspective of anomaly matching, these values are all equivalent mismatch states, so collapsing them substantially amplifies subsequent deduplication.

(iii) **Isomorphic deduplication**: samples sharing identical attribute vectors induce identical tensor contributions and can be merged into weighted representatives, yielding zero information loss.

Additionally, attributes failing the isolated recall bound ($\mathrm{Recall}(\mathbf{e}_k) < \tau$) are deterministically pruned before tensor construction, since they cannot satisfy the recall requirement even in isolation.

After compression, the effective set sizes satisfy $|\mathcal{N}|'' \leq \gamma \cdot |\mathcal{P}|$ and $|\mathcal{P}|'' \leq |\mathcal{P}|$. After the linear preprocessing pass, the resulting pairwise tensors are anomaly-bounded rather than background-bounded, which makes iterative Stage 1 optimization practical on million-scale windows. A detailed complexity derivation is given in Appendix C.

4.3 Two-Stage Optimization

The third level of complexity reduction addresses the exponential search space: 2^M possible attribute combinations must be narrowed to an exactly solvable scale.

4.3.1 Probabilistic Relaxation and Stochastic Gradient Estimation

Although the relaxed objective $f(\mathbf{w})$ is convex on $[0,1]^M$, solving it directly via continuous proximal subgradient descent yields fractional weights $\mathbf{w}^* \in [0,1]^M$ that need not be sparse or aligned with discrete subset boundaries. Because Stage 1 aims specifically to *prune* the discrete hypercube $\{0,1\}^M$, we introduce a **probabilistic sampling relaxation**: each $w_j \in [0,1]$ parameterizes $v_j \sim \mathrm{Bernoulli}(w_j)$, directly probing binary masks and driving irrelevant weights toward zero.

At each iteration t, we draw B independent binary samples $\{\mathbf{v}^{(k)}\}_{k=1}^{B}$ from $\mathbf{w}^t$ and update via straight-through subgradient estimation (STE) [29] with z-score normalization to prevent gradient saturation:

$$\mathbf{w}^{t+1} \leftarrow \mathrm{clip}\left(\mathbf{w}^t - \alpha \cdot \mathrm{Normalize}\left[\frac{1}{B}\sum_{k=1}^{B}\mathbf{g}_f\left(\mathbf{v}^{(k)}\right)\right],\, 0,\, 1\right) \tag{9}$$

where $\mathbf{g}_f(\mathbf{v}) \in \partial_{\mathbf{v}} f(\mathbf{v})$ is a subgradient of f at $\mathbf{v}$ ($\mathrm{ReLU}'(0)$=0). While $\eta\|\mathbf{w}\|_1$ models sparsity in the idealized proximal analysis of $\mathcal{L}(\mathbf{w})$ ($\mathcal{O}(1/\sqrt{T})$, Appendix I), the deployed normalized STE update needs no explicit η step: zero-mean normalization maps weaker-than-average push signals to positive steps (Proposition I.19).

Combined with $[0, 1]$ clipping, irrelevant attribute weights decay toward zero across iterations, yielding a compact candidate subspace for Stage 2 (Section 4.3.2).

4.3.2 Two-Stage Subspace Projection and Combinatorial Search

Rather than relying on continuous relaxation to output the final combination directly, we use it as a **subspace projection** and delegate final selection to exact combinatorial verification:

Stage 1 — Gradient-Driven Subspace Pruning. We execute the stochastic updates (Eq. (9)) over the compressed tensors until $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_1 < \delta$ for a small tolerance $\delta > 0$. After termination at iteration T, PRISM defines $\mathcal{A}_{\mathrm{cand}} = \mathrm{Top}_m^+(\mathbf{w}^T)$ as the at most m attributes with the largest positive weights.

Stage 2 — Exact Combinatorial Verification. Having reduced the attribute universe from M attributes to $\mathcal{A}_{\mathrm{cand}}$, we exhaustively evaluate all $2^{|\mathcal{A}_{\mathrm{cand}}|} - 1$ non-empty subsets and select $\mathcal{C}^*$ maximizing the exact F1 score, *i.e.*, the harmonic mean of Precision and Recall (Eq. (2)).

Stage 1 achieves $\mathcal{O}(2^M) \to \mathcal{O}(2^{|\mathcal{A}_{\mathrm{cand}}|})$ pruning, while Stage 2 guarantees global optimality within $\mathcal{A}_{\mathrm{cand}}$. We set the cap $m \leq 5$ for two practical reasons: (i) in production experience, actionable root causes rarely exceed five intersecting dimensions (empirically, Stage 1 retains a median of 3 candidates while preserving all true root-cause attributes), and (ii) bounding $m \leq 5$ restricts Stage 2 to at most $2^5 - 1 = 31$ subset evaluations. Each Stage 1 step costs $\mathcal{O}\big(B(\gamma|\mathcal{P}||\overline{\mathcal{P}}|'' + |\mathcal{P}|'^2)M\big)$.

Materializing the Root Cause Signature. Given $\mathcal{C}^*$, we extract the root-cause value signature V^* by projecting $\mathcal{P}$ onto $\mathcal{C}^*$:

$$V^* = \underset{v \in A^{\mathcal{P}}(\mathcal{C}^*)}{\arg\max} |\mathcal{P}(v)| \tag{10}$$

where $\mathcal{P}(v) = \{x_i \in \mathcal{P} \mid A^{(i)}(\mathcal{C}^*) = v\}$ is the subset of anomalies sharing signature v under $\mathcal{C}^*$. For concurrent multi-cause incidents, records in $\mathcal{P}(V^*)$ are filtered and the two-stage solver is re-invoked sequentially on residual anomalies.

Figure 2 summarizes the PRISM pipeline; pseudo-code is provided in Algorithm 1 (Appendix B). Furthermore, we prove the $\mathcal{O}(1/\sqrt{T})$ convergence rate for the idealized proximal relaxation (Theorem I.1) and exact optimality within the retained subspace (Proposition I.3) in Appendix I.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate PRISM across five real-world datasets (statistics and anomaly rates in Appendix D, Tables 5–6): four public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$) [3], each containing 8 subsets of varying root-cause granularity, and our industrial **NFC** benchmark comprising 5.13M production transaction records across M=15 attributes (2^{110} value combinations) over 100 SRE-verified incidents, plus a root-cause-preserving **NFC-20K** control (20K records, M=8).[1]

Baselines. We compare PRISM against 11 baselines (details in Appendix E): ten **heuristic/search-based** methods—Adtributor [1], R-Adtributor [19], HotSpot [2], Squeeze [3], PSqueeze [13], RiskLoc [16], AprioriRCA [30], ImpAPTr [31], SinkFlow [21], and MID [18]—plus the **learnable-weighting** method RC-LIR [17] (paired with PSqueeze).

[1] Code: `https://github.com/goaheand/PRISM`; NFC dataset: `https://huggingface.co/datasets/Ysqq/NFC`.

Table 2: Performance (exact-match F1) and diagnostic latency (seconds) averaged over 5 random seeds (standard deviations in Table 9). Public-benchmark averages (**Public Avg.**) are reported separately from the 5.13M-record **NFC** benchmark. **Bold** denotes the best result (including ties) and underlining the second best. 'T/O' indicates no output within the extended 2-hour timeout cap.

Method	Dataset $\mathcal{A}$		Dataset $\mathcal{D}$		Dataset $\mathcal{B}_3$		Dataset $\mathcal{B}_4$		Public Avg.		NFC (5.13M, Ours)	
	F1	Lat.	F1	Lat.	F1	Lat.	F1	Lat.	F1	Lat.	F1	Lat.
Adtributor	0.12	0.04	0.10	**0.02**	0.13	**0.02**	0.13	**0.02**	0.12	0.03	0.01	17.36
R-Adtributor	0.01	1.38	0.09	2.78	0.00	3.86	0.00	3.12	0.03	2.79	0.01	627.10
HotSpot	0.40	11.20	0.51	4.25	0.43	18.10	0.42	7.91	0.44	10.37	T/O	T/O
Squeeze	0.60	2.58	0.88	5.12	0.70	6.27	0.62	8.44	0.70	5.60	T/O	T/O
PSqueeze	0.68	4.88	0.94	4.95	0.83	9.42	0.77	9.49	0.81	7.19	T/O	T/O
RiskLoc	0.77	0.20	**0.96**	0.25	**0.97**	0.49	**0.95**	0.49	0.91	0.36	T/O	T/O
RC-LIR	0.63	2.89	0.92	3.12	0.57	4.34	0.57	4.13	0.67	3.62	0.01	1336.00
AprioriRCA	0.06	1.38	0.92	6.94	0.75	58.91	0.61	44.50	0.59	27.93	T/O	T/O
ImpAPTr	0.39	0.42	0.48	0.28	0.50	1.76	0.51	0.63	0.47	0.77	T/O	T/O
MID	0.33	61.20	0.53	27.81	0.29	31.72	0.31	40.50	0.37	40.31	0.00	62.57
SinkFlow	0.15	1.20	0.45	1.20	0.47	1.20	0.44	1.20	0.38	1.20	T/O	T/O
PRISM (Ours)	**0.87**	**0.02**	**0.96**	**0.02**	0.93	**0.02**	0.91	**0.02**	**0.92**	**0.02**	**0.82**	**0.29**

Unified Upstream Label & Input Protocol. All methods operate downstream of an upstream anomaly partition, and *root-cause annotations are never accessed during label generation or metric reconstruction* (Table 7, Appendix D): (1) On public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$), which lack preset record labels, an unsupervised leaf-level MAPE-plus-elbow detector derives binary labels solely from the empirical observed/predicted residual distribution; measure-based baselines retain the original observed/predicted measures, whereas PRISM and RC-LIR consume only the thresholded binary labels (merged losslessly via isomorphic deduplication). (2) On NFC, binary labels come natively from an independent production detector operating for years (retaining real false-positive/false-negative noise). Because measure-based baselines require observed/forecast pairs, we deterministically map anomalous attribute-value frequencies to observed/forecast deviations via a fixed monotonic piecewise transformation (seed 42), supplying identical reconstructed values to every measure-based baseline.

Implementation, Metrics & Timeout Protocol. All experiments run in Python 3.10 on an Apple M3 Pro (12-core CPU, 18 GB memory). We report Precision, Recall, and F1 under a *strict exact-match* criterion (partial attribute-value matches count as incorrect) and end-to-end Latency (seconds) after data loading, including tensor compression and Stages 1–2 for PRISM. On full NFC (Table 2), we extended the per-case timeout fourfold from 30 minutes to **2 hours** ('T/O'); no additional baseline completed. Controlled scaling sweeps (Figures 4–5) use a 10-minute cap, while MID uses its internal 60 s no-improvement early-stopping guard.

5.2 Performance Evaluation

We evaluate accuracy and runtime across public benchmarks, NFC-20K, and full-scale NFC.

Separated Public Accuracy vs. Industrial NFC Feasibility. Table 2 separates shared public-benchmark accuracy from full-scale NFC deployability. On the four small-scale, semi-synthetic public benchmarks, PRISM achieves the highest average F1 (**0.92**) at **0.02 s** latency—outperforming the risk-partitioning heuristic RiskLoc (0.91 F1, 0.36 s; 18× speedup) and PSqueeze (0.81 F1, 7.19 s; +11 points, 350× speedup). On the 5.13M-record industrial NFC benchmark, however, seven baselines (including RiskLoc, PSqueeze, and SinkFlow) exceed the 2-hour timeout, while completed baselines (Adtributor, R-Adtributor, RC-LIR, MID) collapse to ≤ 0.01 F1. PRISM alone sustains **0.82 F1** ($\sigma_{\mathrm{F1}} = 0.019$, Table 9) in **0.29 s** across the 2^{110} combination space.

Table 3: NFC-20K controllable diagnostic control (20K records, M=8, 100 incidents; all methods complete).

Method	Micro F1	Lat. (s/case)
PRISM (Ours)	**0.82**	**0.04**
SinkFlow	0.53	8.03
R-Adtributor	0.05	9.58
Adtributor	0.02	0.45
MID / RC-LIR / AprioriRCA	0.01	60.10 / 25.04 / 13.39
RiskLoc	0.00	0.46
ImpAPTr / PSqueeze	0.00	420.00 / 450.23

NFC-20K Control: Disentangling Localization from Scale. To verify that baseline failures on NFC are not merely timeout artifacts and that NFC contains non-PRISM-recoverable signal, we construct **NFC-20K** (Appendix D.1), reducing each of the 100 incidents to 20K records and

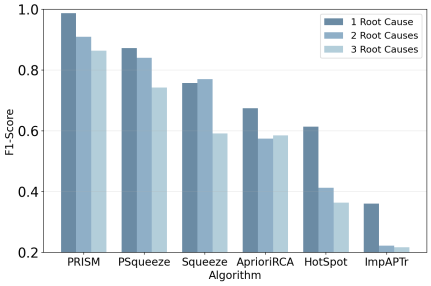


Figure 3: Exact-match F1 for 1–3 concurrent causes on public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$).

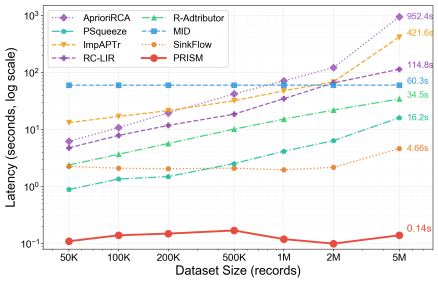


Figure 4: Diagnostic latency (log) vs. records $|\mathcal{U}|$ (50K $\rightarrow$ 5M, M=3, 10-min cap) on NFC.

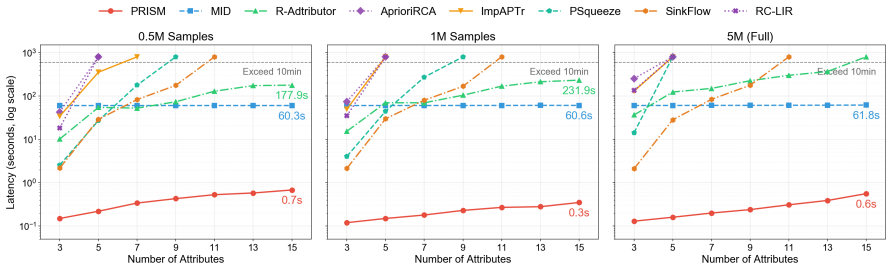


Figure 5: Diagnostic latency (log scale) vs. attribute count M ($3 \rightarrow 15$) across 0.5M, 1M, and 5M NFC records (dashed line: 10-min cap; MID plateaus at its 60 s early-stopping guard).

8 attributes while preserving root-cause records. As shown in Table 3, all evaluated methods complete: **SinkFlow reaches 0.53 F1** in 8.03 s using the reconstructed input, confirming genuine, independently recoverable root-cause signal, yet times out (>2 h) on full NFC due to $\mathcal{O}(|\text{agg}| \cdot 2^M)$ graph construction. Meanwhile, Adtributor (0.45 s) and RiskLoc (0.46 s) finish sub-second with 0.02 and 0.00 F1, and PSqueeze/ImpAPTr spend $\sim$7 minutes for 0.00 F1, because high-cardinality identifiers and sparse anomalies cause noisy low-level cuboids to outrank true causes. PRISM achieves **0.82 F1** in **0.04 s** on NFC-20K and preserves **0.82 F1** in **0.29 s** on full NFC, demonstrating both superior localization fidelity and industrial-scale deployability.

Impact of Root Cause Complexity. As concurrent root causes increase from 1 to 3 (Figure 3), heuristic baselines degrade sharply because sequential candidate expansion propagates early misrankings. PRISM sustains higher F1 by jointly pruning subspaces before sequential peeling.

5.3 Scalability and Latency

Total Volume ($|\mathcal{U}|$) vs. Anomaly-Bounded ($|\mathcal{P}|$) Scaling. Figure 4 scales total input volume $|\mathcal{U}|$ from 50K to 5M records on NFC (M=3), where PRISM stays within 0.10–0.17 s while cuboid/pattern baselines climb to hundreds of seconds and MID flatlines at its 60 s early-stopping guard. To directly isolate post-compression $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$ scaling, we additionally hold $|\mathcal{U}| = 5{,}135{,}932$ (M=15) fixed and vary $|\mathcal{P}|$ from 100 to 20,000 (200×; Appendix H.1, Table 10): PRISM's latency grows only 38× (0.09 s to 3.45 s), remaining under 3.5 s even at $|\mathcal{P}|$=20,000 (over 6× NFC's median $|\mathcal{P}|$=3,144), whereas MID ($\sim$61–65 s) and Adtributor ($\sim$17.5 s) remain invariant as their scans depend on $|\mathcal{U}|$.

Dimensionality Robustness. Figure 5 varies M from 3 to 15 across three data volumes on NFC. Baselines exceed 60 s or hit the 10-minute cap once $M > 9$ due to $\mathcal{O}(2^M)$ search growth. PRISM remains sub-second throughout (0.12–0.70 s), exhibiting linear $\mathcal{O}(M)$ growth (appearing concave on the logarithmic y-axis) as Stage 1 caps Stage 2 verification at $2^5 - 1 = 31$ subsets for any M.

5.4 Diagnostic Analysis

Ablation Study. Table 4 ablates PRISM's core designs on NFC. Removing the Recall penalty ($h(\mathbf{w})$) lowers F1 by 7.6% (0.818 $\rightarrow$ 0.756), as precision-only optimization over-fragments the anomaly cluster. Replacing the asymmetric (q, h) objective with a single symmetric pairwise cross-entropy loss (`w/o Asym. Loss`) omits the separate P-P cohesion pass (slightly reducing runtime to 0.225 s) but collapses Recall to 0.473 (-29.8% F1), confirming the necessity of role asymmetry. Removing Stage 2 exact verification (`w/o Comb. Verif.`, thresholding $\mathbf{w}^T$ directly) drops F1 by 37.4%, confirming that continuous pruning requires exact subset verification.

Configuration	P	R	F1	Lat. (s)
Full PRISM	**0.824**	**0.813**	**0.818**	0.290
w/o Recall Const.	0.761	0.751	0.756	0.295
w/o Asym. Loss	0.731	0.473	0.574	0.225
w/o Comb. Verif.	0.690	0.407	0.512	0.264

Table 4: Component ablation on full NFC (100 incidents, 5-seed mean).

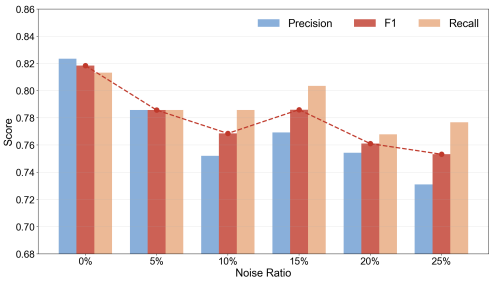


Figure 6: Exact-match F1 under random label flips (0%–25%) on NFC.

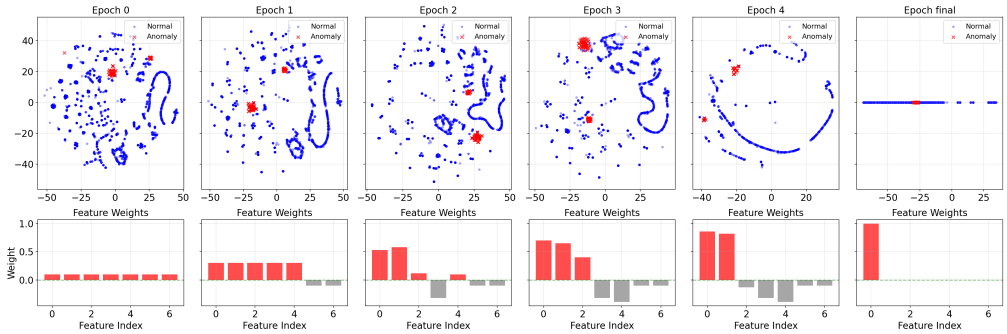


Figure 7: t-SNE evolution of weighted representations across six Stage 1 epochs on NFC: anomalous samples (red) collapse into a dense cluster while normal samples (blue) remain diffuse.

Robustness to Label Noise & Validation of Asymmetry. Under 0%–25% random label flips on NFC (Figure 6), PRISM's F1 degrades by only $\sim$7%, demonstrating practical resilience beyond the noiseless identifiability regime by aggregating collective P-N and P-P statistics. Figure 7 further shows anomalous samples collapsing into a tight cluster over six Stage 1 epochs on NFC while normal samples remain diffuse; Appendix H.2 (Figure 9) confirms this "anomalies collapse, normals diffuse" geometry across all public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$). Sensitivity to τ, σ, γ is in Appendix F.

6 Limitations

Our method and benchmark have several scope boundaries. First, to guarantee sub-second exact verification in Stage 2, PRISM caps the candidate subspace at $m \leq 5$ (≤ 31 subsets), which risks truncating rare faults involving more than five intersecting dimensions; replacing exhaustive enumeration with scalable combinatorial search for $|\mathcal{A}_{\text{cand}}| > 5$ is a natural extension. Second, our Recall objective targets a single cohesive cluster and handles concurrent multi-modal faults via sequential peeling rather than joint mixture disentanglement. Third, PRISM assumes upstream record-level anomaly labels, and our formal guarantees—noiseless identifiability ($\text{Recall}(\mathcal{C}^*)=1$, Theorem I.12) and idealized proximal convergence ($\mathcal{O}(1/\sqrt{T})$, Theorem I.1)—do not formally cover stochastic label noise or the normalized straight-through heuristic in Stage 1. Finally, NFC comes from a single mobile-payment platform, retains asynchronous label noise unsuitable for clean-margin methods, and focuses on flat categorical telemetry without trace graphs or service topologies.

7 Conclusion

PRISM addresses million-scale multi-dimensional RCA by coupling an asymmetric Precision-Recall objective with one-pass tensor compression and two-stage Bernoulli-pruned exact verification. Experiments across four public benchmarks, the tractable NFC-20K control, and the 5.13M-record industrial NFC benchmark confirm that PRISM achieves the highest public-benchmark average F1 (0.92 vs. 0.91) while uniquely sustaining 0.82 F1 and 0.29 s latency across a 2^{110} combination space. Extending asymmetric subspace optimization to noisy identifiability guarantees and parallel multi-cause mixture models remains promising future work.

Acknowledgments and Disclosure of Funding

We thank the Site Reliability Engineering (SRE) team at Ant Group for verifying the production incident tickets in the NFC benchmark, and the anonymous reviewers and Area Chair for constructive feedback—including the controllable NFC-20K diagnostic design—that substantially strengthened this work. This work was supported by Ant Group Research Intern Program. The authors declare no competing financial interests beyond their stated affiliations.

References

[1] Ranjita Bhagwan, Rahul Kumar, Ramachandran Ramjee, George Varghese, Surjyakanta Mohapatra, Hemanth Manoharan, and Piyush Shah. Adtributor: Revenue debugging in advertising systems. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, pages 43–55, USA, 2014.

[2] Yongqian Sun, Youjian Zhao, Ya Su, Dapeng Liu, Xiaohui Nie, Yuan Meng, Shiwen Cheng, Dan Pei, Shenglin Zhang, Xianping Qu, and Xuanyou Guo. HotSpot: Anomaly localization for additive KPIs with multi-dimensional attributes. *IEEE Access*, 6:10909–10923, 2018.

[3] Zeyan Li, Chengyang Luo, Yiwei Zhao, Yongqian Sun, Kaixin Sui, Xiping Wang, Dapeng Liu, Xing Jin, Qi Wang, and Dan Pei. Generic and robust localization of multi-dimensional root causes. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, pages 47–57, 2019.

[4] Chaoyu Chen, Hang Yu, Zhichao Lei, Jianguo Li, Shaokang Ren, Tingkai Zhang, Silin Hu, Jianchao Wang, and Wenhui Shi. BALANCE: Bayesian linear attribution for root cause localization. *Proc. ACM Manag. Data*, 1(1), May 2023.

[5] Nan Fu, Guang Cheng, Yue Teng, Guangye Dai, Shui Yu, and Zihan Chen. Intelligent root cause localization in microservice systems: A survey and new perspectives. *ACM Computing Surveys*, 57(12), 2025.

[6] PaySpace Magazine. Alipay Tap! surpasses 100 million daily transactions milestone. `https://payspacemagazine.com/news/alipay-tap-surpasses-100-million-daily-transactions-milestone/`, 2026. Accessed: 2026-05-03.

[7] Moneycontrol. PhonePe tops 10 billion monthly UPI transactions; Google Pay nears Rs 10 lakh crore value in March. `https://www.moneycontrol.com/technology/phonepe-tops-10-billion-monthly-upi-transactions-google-pay-nears-rs-10-lakh-crore-value-in-march-article-13896848.html`, 2026. Accessed: 2026-05-03.

[8] Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. RCAEval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 777–780, 2025.

[9] Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 2022.

[10] Tianyu Chen, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. iSCAN: Identifying causal mechanism shifts among nonlinear additive noise models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[11] Hongyuan Tao, Hang Yu, and Jianguo Li. DeepITE: Designing variational graph autoencoders for intervention target estimation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[12] Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through Granger causal discovery. In *The Thirteenth International Conference on Learning Representations*, 2025.

[13] Zeyan Li, Junjie Chen, Yihao Chen, Chengyang Luo, Yiwei Zhao, Yongqian Sun, Kaixin Sui, Xiping Wang, Dapeng Liu, Xing Jin, Qi Wang, and Dan Pei. Generic and robust root cause localization for multi-dimensional data in online service systems. *Journal of Systems and Software*, 203:111748, 2023.

[14] Qingwei Lin, Jian-Guang Lou, Hongyu Zhang, and Dongmei Zhang. iDice: Problem identification for emerging issues. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, pages 214–224, 2016.

[15] Yeqing Meng, Qianli Zhang, Xiangyu Tang, Wanhao Zhang, and Jilong Wang. FTM-RCA: A fast two-stage multi-dimensional root-cause analysis of network anomalies. In *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*, pages 1–10, 2023.

[16] Marcus Kalander. RiskLoc: Localization of multi-dimensional root causes by weighted risk. *arXiv preprint arXiv:2205.10004*, 2022.

[17] Yiran Cheng, Bo Cheng, Pengxiang Jin, Yongqian Sun, Xiaohui Nie, Nengwen Zhao, Shenglin Zhang, and Dan Pei. Effective attribute selection for multi-dimensional root cause analysis. In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, pages 321–331, 2022.

[18] Jiazhen Gu, Chuan Luo, Si Qin, Bo Qiao, Qingwei Lin, Hongyu Zhang, Ze Li, Yingnong Dang, Shaowei Cai, Wei Wu, Yangfan Zhou, Murali Chintalapati, and Dongmei Zhang. Efficient incident identification from multi-dimensional issue reports via meta-heuristic search. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pages 292–303, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Linnea Rudenius and Moa Persson. Anomaly detection and fault localization: An automated process for advertising systems. Master's thesis, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden, 2018.

[20] Pengkun Jing, Yanni Han, Jiyan Sun, Tao Lin, and Yanjie Hu. AutoRoot: A novel fault localization schema of multi-dimensional root causes. In *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7, 2021.

[21] Zhichao Hu, Likun Liu, Lina Ma, and Xiangzhan Yu. SinkFlow: Fast and traceable root-cause localization for multidimensional anomaly events. *Engineering Applications of Artificial Intelligence*, 139:109582, 2025.

[22] Baige Tang and Li Zhang. Local preserving logistic I-Relief for semi-supervised feature selection. *Neurocomputing*, 399:48–64, 2020.

[23] Shifu Yan, Caihua Shan, Wenyi Yang, Bixiong Xu, Dongsheng Li, Lili Qiu, Jie Tong, and Qi Zhang. CMMD: Cross-metric multi-dimensional root cause analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 4310–4320, 2022.

[24] Ping Wang, Jingmin Xu, Meng Ma, Weilan Lin, Disheng Pan, Yuan Wang, and Pengfei Chen. CloudRanger: Root cause identification for cloud native systems. In *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 492–502, 2018.

[25] Yingying Zhang, Zhengxiong Guan, Huajie Qian, Leili Xu, Hengbo Liu, Qingsong Wen, Liang Sun, Junwei Jiang, Lunting Fan, and Min Ke. CloudRCA: A root cause analysis framework for cloud computing platforms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 4373–4382, 2021.

[26] Dewei Liu, Chuan He, Xin Peng, Fan Lin, Chenxi Zhang, Shengfang Gong, Ziang Li, Jiayu Ou, and Zheshun Wu. MicroHECL: High-efficient root cause localization in large-scale microservice systems. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 338–347, 2021.

[27] Alexander Shan, Jasleen Kaur, Rahul Singh, Tarun Banka, Raj Yavatkar, and T. Sridhar. RCA Copilot: Transforming network data into actionable insights via large language models. In *ICC 2025 - IEEE International Conference on Communications*, pages 1566–1571, 2025.

[28] Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. OpenRCA: Can large language models locate the root cause of software failures? In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[30] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA, USA, 3rd edition, 2011.

[31] Hao Wang, Guoping Rong, Yangchen Xu, and Yong You. ImpAPTr: A tool for identifying the clues to online service anomalies. In *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1307–1311, 2020.

[32] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[33] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L0 regularization. In *International Conference on Learning Representations*, 2018.

[34] Muhammed Fatih Balin, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 444–453. PMLR, 2019.

[35] Nikolaos Karalias and Andreas Loukas. Erdős goes neural: an unsupervised learning framework for combinatorial optimization on graphs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, 2020.

[36] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In Qian Wang, Yinghuan Shi, Heung-Il Suk, and Kenji Suzuki, editors, *Machine Learning in Medical Imaging*, pages 379–387, Cham, 2017. Springer International Publishing.

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[38] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with AP-loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5114–5122, 2019.

[39] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, August 2009.

A Extended Related Works

Synergy with Service-Level and LLM-Based RCA. While Section 2 contrasts service-level/LLM triage (*where*) with attribute-level localization (*what*), the two paradigms naturally form a two-tier production pipeline. Graph- and causal-discovery frameworks [9–12, 24, 25] as well as LLM agents [27, 28] excel at correlating heterogeneous logs, traces, and service topologies to isolate the failing microservice, but incur multi-second to multi-minute inference latency and cannot enumerate 2^{100} structured attribute combinations. Once high-level triage narrows an incident to a target service or metric table, PRISM can be invoked as a sub-second combinatorial engine to pinpoint the exact anomalous attribute-value signature.

Differentiable Selection and Asymmetric Loss. Methodologically, our work intersects two broader ML paradigms. **Differentiable subset selection** relaxes discrete choices into continuous optimization using sparsity penalties (*e.g.*, Lasso [32]) or stochastic gates (*e.g.*, Hard Concrete [33], Gumbel-Softmax [34], and Bernoulli relaxations [35]). However, these are generally designed for generic feature selection or graph combinatorial problems rather than asymmetric telemetry subspaces. Similarly, **asymmetric loss design** addresses class imbalance via reweighting (*e.g.*, Tversky loss [36], Focal loss [37], or AP loss [38]), typically over continuous per-sample predictions in vision or ranking. PRISM bridges both paradigms by coupling stochastic Boolean subspace pruning with an asymmetric relational objective tailored to categorical RCA.

B Algorithm Pseudo-Code

Algorithm 1 PRISM: Precision-Recall Informed Subspace Mining

Require: Sample set $\mathcal{U} = \mathcal{P} \cup \mathcal{N}$, hyperparameters $\sigma, \tau, \lambda, \alpha, B, \gamma, \delta, m$
Ensure: Selected root cause attribute combination $\mathcal{C}^*$ and attribute value signature V^*

// Tractable Scaling and Tensor Compression
Subsample $\mathcal{N}$ to capacity $|\mathcal{N}|' = \min(|\mathcal{N}|, \gamma \cdot |\mathcal{P}|)$
Prune attributes $\mathcal{A}_k$ where $\mathrm{Recall}(\mathbf{e}_k) < \tau$
Apply long-tail anonymization with token `<other>`
Execute isomorphic deduplication in $\mathcal{N}$ and $\mathcal{P}$; aggregate scalar weights
Initialize compressed difference tensors $D^{\mathcal{N},\mathcal{P}}$ and $D^{\mathcal{P},\mathcal{P}}$
// Stage 1: Gradient-Driven Subspace Pruning
Initialize selection probabilities $\mathbf{w}^0 \leftarrow \mathbf{1}_M / M$
repeat
  for $k = 1$ **to** B **do**
    Draw $v_j^{(k)} \sim \mathrm{Bernoulli}(w_j^t)$ for all $j \in [M]$
  end for
  Estimate normalized stochastic subgradient: $\mathbf{g}^t \leftarrow \mathrm{Normalize}\left[\frac{1}{B}\sum_{k=1}^{B} \mathbf{g}_f(\mathbf{v}^{(k)})\right]$ ▷ $\mathbf{g}_f(\mathbf{v}) \in \partial_{\mathbf{v}} f(\mathbf{v})$
  Update weights with the practical normalized step: $\mathbf{w}^{t+1} \leftarrow \mathrm{clip}(\mathbf{w}^t - \alpha \cdot \mathbf{g}^t,\, 0,\, 1)$
until $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_1 < \delta$
// Stage 2: Exact Combinatorial Verification
Rank attributes by final weights w_j^{t+1}
Define $\mathcal{A}_{\mathrm{cand}} \leftarrow \mathrm{Top}_m^+(\mathbf{w}^{t+1})$, the top-$m$ positive-weight attributes
Enumerate all $2^{|\mathcal{A}_{\mathrm{cand}}|} - 1$ non-empty subsets of $\mathcal{A}_{\mathrm{cand}}$
$\mathcal{C}^* \leftarrow \arg\max_{\mathcal{C} \subseteq \mathcal{A}_{\mathrm{cand}},\, \mathcal{C} \neq \emptyset} \mathrm{F1}(\mathrm{Precision}(\mathcal{C}),\, \mathrm{Recall}(\mathcal{C}))$
// Materializing the Root Cause Signature
$V^* \leftarrow \arg\max_{v \in A^{\mathcal{P}}(\mathcal{C}^*)} |\mathcal{P}(v)|$
return $\mathcal{C}^*$, V^*

C Computational Complexity Analysis

One-Pass Preprocessing and Tensor Compression. Reading the input window and subsampling $\mathcal{N}$ to $|\mathcal{N}|' = \min(|\mathcal{N}|, \gamma|\mathcal{P}|)$ takes a single linear scan of $\mathcal{O}(|\mathcal{U}| \cdot M)$ (or $\mathcal{O}(|\mathcal{N}|)$ index sampling once

loaded). All subsequent compression operations run on the subsampled working set $\mathcal{U}' = \mathcal{P} \cup \mathcal{N}'$ ($|\mathcal{U}'| \leq (\gamma+1)|\mathcal{P}|$): deterministic isolated-recall pruning costs $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$, long-tail anonymization costs $\mathcal{O}(|\mathcal{N}|' \cdot M)$, and hash-based isomorphic deduplication costs $\mathcal{O}(|\mathcal{U}|' \cdot M)$. Following these operations, let $|\mathcal{N}|''$ and $|\mathcal{P}|''$ denote the compressed representative counts, satisfying $|\mathcal{N}|'' \leq |\mathcal{N}|' \leq \gamma \cdot |\mathcal{P}|$ and $|\mathcal{P}|'' \leq |\mathcal{P}|$. Thus, after a single linear pass over $\mathcal{U}$, all iterative optimization stages are strictly decoupled from the raw normal volume $|\mathcal{N}|$.

Difference Tensor Construction. Populating the compressed tensors incurs a spatial and temporal cost of

$$\mathcal{O}\big((|\mathcal{N}|''|\mathcal{P}|'' + |\mathcal{P}|''^2)M\big)\,.$$

Using $|\mathcal{N}|'' \leq \gamma|\mathcal{P}|$ gives the conservative bound

$$\mathcal{O}\big((\gamma|\mathcal{P}||\mathcal{P}|'' + |\mathcal{P}|''^2)M\big)\,,$$

which is further bounded by $\mathcal{O}((\gamma+1)|\mathcal{P}|^2 M)$ since $|\mathcal{P}|'' \leq |\mathcal{P}|$. This decouples the pairwise tensor memory and construction time from the massive raw background volume $|\mathcal{N}|$.

Stage 1: Stochastic Gradient Projection. At each iteration, PRISM draws B binary vectors and evaluates the gradient over the compressed tensors. The cost per step is

$$\mathcal{O}\big(B(\gamma|\mathcal{P}||\mathcal{P}|'' + |\mathcal{P}|''^2)M\big)\,,$$

and is therefore bounded by $\mathcal{O}(B(\gamma+1)|\mathcal{P}|^2 M)$. Over T iterations, the cumulative complexity multiplies this bound by T. Because the push-pull gradients drive rapid structural convergence, T remains a small constant ($T \ll |\mathcal{U}|$).

Stage 2: Exact Combinatorial Search. Given the retained candidate set $\mathcal{A}_{\text{cand}}$ with $|\mathcal{A}_{\text{cand}}| \leq m$, exhaustive verification costs

$$\mathcal{O}\Big(2^{|\mathcal{A}_{\text{cand}}|}(\gamma|\mathcal{P}||\mathcal{P}|'' + |\mathcal{P}|''^2)M\Big)\,.$$

Under the operational cap $m \leq 5$, this translates to at most 31 evaluations, rendering the exact search overhead mathematically negligible.

Root Cause Materialization. Extracting the modal signature via frequency counts requires traversing the projected subspace, incurring an insignificant $\mathcal{O}(|\mathcal{P}| \cdot m)$ operations.

Stage 1 Convergence and Pruning Statistics. Across all datasets and anomaly configurations, Stage 1 converges within $T \leq 15$ iterations and, under the operational cap $m \leq 5$, retains a candidate set with median size 3. In every case, the final candidate set contains all ground-truth root-cause attributes, empirically confirming the support-retention condition used in the theoretical analysis.

D Dataset Descriptions and Input Protocols

Table 5: Dataset overview. M: number of attributes; **Rec.**: records; **Value Combos**: total attribute-value combination space; **Src.**: domain; **Meas.**: monitored signal.

	M	Rec.	Value Combos	Src.	Meas.		M	Rec.	Value Combos	Src.	Meas.
$\mathcal{A}$	5	15K	2^{22}	Shop.	orders	$\mathcal{B}_4$	4	22K	2^{15}	Internet	page views
$\mathcal{D}$	4	22K	2^{15}	Internet	succ. rate	NFC-20K	8	20K	$\ll 2^{110}$	Transact.	binary alert
$\mathcal{B}_3$	4	22K	2^{15}	Internet	page views	NFC (Full)	15	5.13M	2^{110}	Transact.	binary alert

Table 6: Per-case record counts ($|\mathcal{U}|$), anomalous sample statistics ($|\mathcal{P}|$), and anomaly rates across all benchmarks.

Dataset	$\lvert\mathcal{U}\rvert$ per case	$\lvert\mathcal{P}\rvert$ mean	$\lvert\mathcal{P}\rvert$ median	$\lvert\mathcal{P}\rvert$ range	Anomaly rate (mean)	Anomaly rate (median)
$\mathcal{A}$	∼3,388	471.8	135.5	10–1,734	14.06%	4.00%
$\mathcal{D}$	∼13,806	1,028.2	298.5	10–6,649	7.45%	2.16%
$\mathcal{B}_3$	∼13,688	993.3	232.5	10–5,275	7.26%	1.70%
$\mathcal{B}_4$	∼13,691	993.9	217.5	10–5,180	7.26%	1.59%
NFC-20K	20,000	87.1	64.5	11–709	0.436%	0.323%
NFC (Full)	5,135,932	2,859.7	3,144.0	30–5,461	0.056%	0.061%

Public Benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$). These semi-synthetic benchmarks originate from an online shopping platform and an internet company [3, 13]. Following established protocols [3], each dataset

is partitioned into 8 subsets (x, y) denoting x concurrent root-cause combinations of y attributes each, with 100 injected fault cases per subset.

NFC Benchmark Construction, Anonymization, and SRE Verification. We construct **NFC** from the "Tap-to-Pay" transaction service of a large-scale mobile payment platform, spanning $5{,}135{,}932$ records across $M{=}15$ categorical dimensions (2^{110} attribute-value combinations) and 100 real production incidents:

1. **Column-Wise Anonymization:** Each non-empty categorical string within a column is bijectively mapped to a unique integer ID, while missing/empty entries remain empty, preserving exact combinatorial cardinality and sparsity without exposing proprietary metadata.
2. **Four-Step SRE Ground-Truth Verification:** Every incident follows a strict production verification chain: (i) the deployed online anomaly-detection system triggers an alert and emits record-level binary labels over the preceding 15-minute diagnostic window; (ii) the online diagnostic pipeline logs a suspected combination (`topCombination`); (iii) an on-call Site Reliability Engineer (SRE) inspects the corresponding incident ticket and manually confirms the true root cause (`risk_event.dimensions`); and (iv) the SRE-confirmed cause is mapped to the ground-truth attribute-value combination. Thus, while upstream record-level anomaly labels retain genuine production false positives and false negatives, *every root-cause annotation across all 100 cases is human-verified by an SRE*.

Upstream Label Generation and Measure Reconstruction Protocol. Table 7 summarizes how upstream labels and reconstructed measures are prepared without ground-truth leakage:

- **Public Benchmarks (Unsupervised MAPE-plus-Elbow):** Because $\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$ provide leaf-level observed/forecast KPI pairs without binary labels, we compute leaf-level MAPE residuals and select an unsupervised elbow threshold solely from the empirical MAPE distribution (mirroring the residual-thresholding step used inside HotSpot [2], Squeeze [3], PSqueeze [13], and RiskLoc [16]). Measure-based baselines receive the original observed/forecast pairs; PRISM and RC-LIR receive only the thresholded binary labels.
- **NFC (Native Binary Interface & Controlled Reconstruction):** In production NFC telemetry, device models, batches, and firmware have incompatible metric baselines, so the deployed detector outputs unified binary alert labels natively. To evaluate measure-based baselines that require observed/forecast KPIs, we reconstruct observed/forecast pairs *solely from the binary anomaly labels* (never accessing root-cause annotations): we count categorical-value frequencies among $\mathcal{P}$, sum constituent frequencies into a leaf-combination intensity normalized to $[0, 1]$, and apply a fixed monotonic piecewise map (seed 42) from intensity to observed/forecast deviation. Every measure-based baseline receives the exact same reconstructed values.

Table 7: Unified upstream anomaly-partition and measure-reconstruction protocol (zero root-cause leakage).

Dataset	**Upstream Label Source**	$\mathcal{P}/\mathcal{N}$ **Partition**	**Method Inputs**
$\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$	Unsupervised leaf-level MAPE + elbow detector on observed/forecast residuals	Dynamic per-case threshold on empirical MAPE	**PRISM / RC-LIR:** binary labels only. **Measure baselines:** original observed/forecast pairs.
NFC / NFC-20K	Independent production anomaly detector (retaining real FP/FN noise); root causes verified separately by SREs	Native upstream binary alert labels	**PRISM / RC-LIR:** native binary labels. **Measure baselines:** identical seed-42 frequency-mapped observed/forecast pairs.

D.1 NFC-20K Diagnostic Control and Complexity Breakdown

To disentangle intrinsic localization capability from million-scale computational infeasibility, **NFC-20K** converts each of the 100 NFC incidents into a $20{,}000$-record, 8-attribute task: (1) we retain 8 core attributes and project ground-truth root causes onto them; (2) all records in $\mathcal{P}$ matching at least one projected root cause are retained to preserve evaluable signal; (3) remaining rows are sampled without replacement (seed 42) from the background pool (normal records plus non-matching anomalous records) to reach exactly $20{,}000$ records; and (4) all methods receive identical sampled

records, remapped anomaly indices, and reconstructed measures (root causes are used only for subset construction and evaluation, never as method inputs).

Table 8 details the complexity regimes and measured latencies across all 10 methods on NFC-20K. Methods free of the attribute-value grid product $\prod_j C_j$ finish within 0.04–60.10 s, whereas ImpAPTr (420.00 s) and PSqueeze (450.23 s) require $\sim$7 minutes per case to traverse high-cardinality value grids. Crucially, low baseline accuracy is *not* a timeout artifact: RiskLoc obtains zero true positives across `risk_threshold` $\in [0.5, 0.95]$ because sparse, noisy low-level cuboids outrank true causes; Adtributor and R-Adtributor emit 604 and 566 false positives; AprioriRCA emits $57{,}341$ false positives; and RC-LIR retains excessive attributes (11 of 15 on full NFC at τ_w=1.0), still yielding 0.01 F1 at M=8. **SinkFlow** avoids exhaustive value grids via EAG propagation to reach **0.53 F1** on NFC-20K—confirming independently recoverable non-PRISM signal—but its $\mathcal{O}(|\text{agg}| \cdot 2^M)$ graph construction exceeds 2 hours at M=15 on full NFC. PRISM sustains **0.82 F1** at both scales (0.04 s on NFC-20K; 0.29 s on full NFC).

Table 8: Dominant complexity regimes and measured per-case F1 and latency on NFC-20K ($|\mathcal{U}|$=20,000, M=8).

Method	Dominant Computation	Key Dependency	Micro F1	Lat. (s)
PRISM (Ours)	$\mathcal{O}(\lvert\mathcal{P}\rvert^2 \cdot M)$ post-compression	$\lvert\mathcal{P}\rvert$, linear in M	**0.82**	**0.04**
Adtributor	$\mathcal{O}(\lvert\mathcal{U}\rvert M + \sum_j C_j)$	records & marginal cardinalities	0.02	0.45
RiskLoc	$\mathcal{O}(\lvert\mathcal{U}\rvert M + 2^M \lvert\text{agg}\rvert M)$	cuboids & leaf aggregates	0.00	0.46
SinkFlow	$\mathcal{O}(\lvert\text{agg}\rvert \cdot 2^M)$	event aggregates & cuboids	<u>0.53</u>	8.03
R-Adtributor	$\mathcal{O}(2^M \lvert\text{agg}\rvert M)$	recursive cuboid search	0.05	9.58
AprioriRCA	$\mathcal{O}(C^k_{\max})$ candidate growth	max cardinality & itemset size	0.01	13.39
RC-LIR	$\mathcal{O}(\lvert\mathcal{U}\rvert^2 M)$ prep. + solver	records & downstream search	0.01	25.04
MID	Iterative meta-heuristic search	early-stopped iterations	0.01	60.10
ImpAPTr	$\mathcal{O}(\prod_j C_j)$	high-cardinality value grid	0.00	420.00
PSqueeze	$\mathcal{O}(2^M \prod_j C_j)$	cuboids & value grid	0.00	450.23

E Baseline Methods

We summarize the 11 compared baselines below. All methods use their originally reported configurations unless dataset-specific adaptation is required.

- **Adtributor** [1] ranks single-dimension attribute values by surprise and explanatory power (KL divergence).
- **R-Adtributor** [19] extends Adtributor recursively across dimensions to explore multi-attribute combinations.
- **HotSpot** [2] localizes additive KPI anomalies via ripple-effect potential scores and Monte Carlo tree search.
- **Squeeze** [3] combines Generalized Potential Scores with density-based clustering and top-down cuboid pruning.
- **PSqueeze** [13] enhances Squeeze via probabilistic scoring and robust partition pruning for noisy KPIs.
- **RiskLoc** [16] partitions leaf elements into normal and abnormal sets via deviation scores and searches cuboids using weighted risk; it excels on low-dimensional synthetic KPIs ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$) but breaks down under high-cardinality, 15-dimensional telemetry. Note that its follow-up variant PRiskLoc has no publicly available source code, and RiskLoc's additional proprietary datasets ($\mathcal{A}, \mathcal{L}, \mathcal{H}$) are closed-source, whereas its public benchmarks match $\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$ evaluated here.
- **AprioriRCA** [30] mines frequent anomalous attribute-value itemsets via Apriori-style candidate generation.
- **ImpAPTr** [31] extracts implication rules over attribute-value grids to rank anomaly clues.
- **SinkFlow** [21] builds an event aggregation graph (EAG) and propagates anomaly scores via random walks.
- **MID** [18] searches abnormal attribute combinations via entropy-guided meta-heuristic perturbation with a 60 s no-improvement guard.

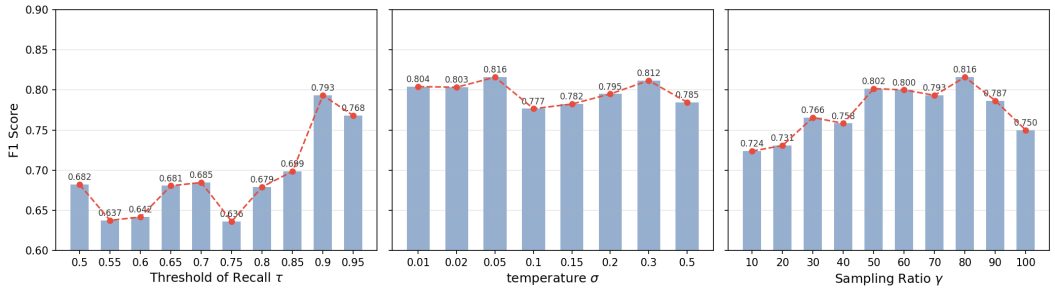


Figure 8: Hyperparameter sensitivity on NFC across isolated-recall threshold τ (left), Precision kernel temperature σ (center), and negative downsampling ratio γ (right).

- **RC-LIR** [17] learns margin-based dimension weights via Logistic I-Relief [22] before invoking downstream PSqueeze.

F Parameter Sensitivity

Figure 8 evaluates F1 on NFC across three hyperparameters without per-case tuning: (i) for the isolated-recall pruning threshold $\tau \in [0.50, 0.95]$, overly permissive thresholds ($\tau \leq 0.80$, F1 0.636–0.685) admit noisy attributes into Stage 1, whereas stricter cohesion filtering peaks at $\tau = 0.90$ (0.793 F1) and remains strong at $\tau = 0.95$ (0.768); (ii) for the Precision kernel temperature $\sigma \in [0.01, 0.50]$, performance stays consistently high between 0.777 and 0.816 (peaking at $\sigma = 0.05$ with 0.816), confirming insensitivity to kernel bandwidth; and (iii) for the negative sampling ratio $\gamma \in [10, 100]$, F1 rises from 0.724 (γ=10) to a broad plateau over $\gamma \in [50, 80]$ (0.793–0.816, peaking at γ=80), before mildly softening at γ=100 (0.750) as excessive noisy background samples enter the tensor. This empirically confirms that capping $|\mathcal{N}|'$ at a moderate multiple of $|\mathcal{P}|$ captures the diffuse normal distribution without full-dataset scans.

G Experimental Statistical Significance

To assess statistical reliability, we evaluate sensitivity across 5 random seeds. Among all 12 evaluated methods (11 baselines + PRISM), **HotSpot**, **MID**, **RC-LIR**, and **PRISM** involve internal stochasticity (*e.g.*, MCTS rollouts, meta-heuristic perturbation, neighbor sampling, or Bernoulli STE sampling), whereas the remaining 8 methods are deterministic in F1 given fixed input data (minor latency variance reflects OS scheduling). In particular, PRISM achieves $\sigma_{\text{F1}} = 0.019$ on NFC (0.82 ± 0.02), confirming tight seed-to-seed stability at million-record scale.

Table 9: Standard deviations of F1 (σ_{F1}) and latency in seconds (σ_{Lat}) across 5 random seeds. Deterministic methods have zero F1 variance by definition; '–' denotes timeout.

Method	**Dataset $\mathcal{A}$**		**Dataset $\mathcal{D}$**		**Dataset $\mathcal{B}_3$**		**Dataset $\mathcal{B}_4$**		**NFC (5.13M)**	
	σ_{F1}	σ_{Lat}	σ_{F1}	σ_{Lat}	σ_{F1}	σ_{Lat}	σ_{F1}	σ_{Lat}	σ_{F1}	σ_{Lat}
Adtributor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.56
R-Adtributor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	65.31
HotSpot	0.04	0.32	0.00	0.01	0.00	0.02	0.00	0.22	–	–
Squeeze	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
PSqueeze	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
RiskLoc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
RC-LIR	0.00	0.02	0.01	7.78	0.00	0.25	0.00	0.44	0.00	135.25
AprioriRCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
ImpAPTr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
MID	0.02	11.64	0.12	13.65	0.19	9.80	0.09	10.50	0.00	12.32
SinkFlow	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–	–
PRISM (Ours)	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.01	**0.019**	0.10

H Additional Scalability and Asymmetry Validation

H.1 Controlled $|\mathcal{P}|$-Bounded Scaling at Fixed $|\mathcal{U}|$

Because negative downsampling caps $|\mathcal{N}|' = \min(|\mathcal{N}|, \gamma|\mathcal{P}|)$, PRISM's post-compression complexity scales as $\mathcal{O}(|\mathcal{P}|^2 \cdot M)$. To directly isolate $|\mathcal{P}|$-scaling independently of total volume $|\mathcal{U}|$, we fix the full NFC telemetry table ($|\mathcal{U}| = 5{,}135{,}932$, $M = 15$) and vary the anomalous sample size $|\mathcal{P}|$ from 100 to 20,000 (sampled from a 212,802-record fault pool with 10% injected false-positive noise; 600 s timeout cap). As shown in Table 10, expanding $|\mathcal{P}|$ by 200× (100 → 20,000) increases PRISM's latency by only 38× (0.09 s → 3.45 s) thanks to isomorphic deduplication ($|\mathcal{P}|'' \ll |\mathcal{P}|$), resolving even $|\mathcal{P}|$=20,000 in under 3.5 s. By contrast, MID (∼61–65 s) and Adtributor (∼17.5–18.3 s) remain invariant across $|\mathcal{P}|$ because their computational bottlenecks depend on scanning the full 5.13M-record dataset $|\mathcal{U}|$, while R-Adtributor times out (> 600 s) everywhere.

Table 10: Diagnostic latency (seconds) when holding $|\mathcal{U}| = 5{,}135{,}932$ (M=15) fixed and varying anomalous count $|\mathcal{P}|$ from 100 to 20,000 ('T/O': > 600 s).

Method	$\|\mathcal{P}\|$=100	$\|\mathcal{P}\|$=500	$\|\mathcal{P}\|$=1k	$\|\mathcal{P}\|$=2k	$\|\mathcal{P}\|$=5k	$\|\mathcal{P}\|$=10k	$\|\mathcal{P}\|$=15k	$\|\mathcal{P}\|$=20k
PRISM (Ours)	**0.09 s**	**0.14 s**	**0.16 s**	**0.95 s**	**1.75 s**	**1.59 s**	**2.55 s**	**3.45 s**
Adtributor	18.34 s	17.69 s	17.68 s	17.45 s	17.49 s	17.56 s	17.67 s	17.52 s
MID	60.65 s	60.75 s	61.90 s	65.16 s	64.64 s	64.07 s	64.31 s	65.19 s
R-Adtributor	T/O	T/O	T/O	T/O	T/O	T/O	T/O	T/O

H.2 Cross-Dataset Validation of Observation 4.1 on Public Benchmarks

To verify that Observation 4.1 reflects an intrinsic geometric property of multi-dimensional RCA rather than an NFC-specific phenomenon, Figure 9 visualizes t-SNE embeddings (using uniform preprocessing and t-SNE hyperparameters) across all four public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$). In the **full attribute space** (top row), anomalous records (red crosses) and normal records (blue circles) overlap heavily due to uninformative dimensions. Once projected onto the **root-cause subspace** (bottom row), anomalous records consistently collapse into compact, isolated clusters while normal records remain stochastically dispersed across the space, corroborating Observation 4.1 across diverse domains.

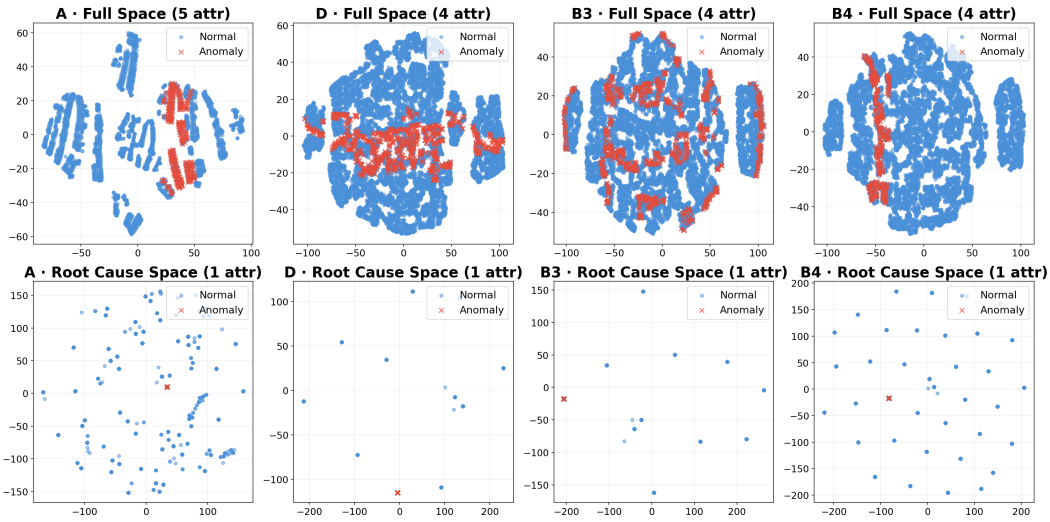


Figure 9: Cross-dataset t-SNE validation of Observation 4.1 on public benchmarks ($\mathcal{A}, \mathcal{D}, \mathcal{B}_3, \mathcal{B}_4$). **Top:** in the full attribute space, normal (blue) and anomalous (red) samples intermingle. **Bottom:** in the root-cause subspace, anomalies collapse into tight clusters while normal samples remain diffuse.

I Theoretical Analysis: Full Proofs

For the post-compression objective, we state the main guarantees:

Theorem I.1 (Idealized Stage 1 Convergence)**.** *Let* $R_0 = \|\mathbf{w}^0 - \mathbf{w}^*\|$ *and* $G_f = \sqrt{M}(1/\sigma + \lambda)$. *Under exact subgradients with step size* $\alpha = R_0/(G_f\sqrt{T})$*, the proximal subgradient iterates satisfy* $\min_{0\le t<T} \mathcal{L}(\mathbf{w}^t) - \mathcal{L}(\mathbf{w}^*) = \mathcal{O}(R_0 G_f/\sqrt{T})$.

Remark I.2 (Fast Pruning in Practice). This rate explains the pruning behavior of the idealized relaxation; the implemented normalized straight-through update is evaluated empirically.

The straight-through estimator used in practice is discussed separately as a low-variance heuristic (Remark I.7).

Proposition I.3 (Stage 2 Global Optimality)**.** *Let* $\mathcal{A}_{\text{cand}} = \text{Top}_m^+(\mathbf{w}^T)$ *with* $|\mathcal{A}_{\text{cand}}| \le m$*. Stage 2 returns* $\mathcal{C}^* = \arg\max_{\mathcal{C}\subseteq\mathcal{A}_{\text{cand}},\,\mathcal{C}\neq\emptyset} \text{F1}(\text{Precision}(\mathcal{C}), \text{Recall}(\mathcal{C}))$ *via exhaustive enumeration over* $2^{|\mathcal{A}_{\text{cand}}|} - 1$ *non-empty subsets (*≤ 31 *under the operational prior* $m \le 5$*).*

Remark I.4 (Exactness after Pruning). Once Stage 1 reduces the search to $\mathcal{A}_{\text{cand}}$, Stage 2 removes heuristic search error by solving the remaining subspace exactly.

I.1 Convergence Analysis of Stage 1

We establish the convergence guarantee for the idealized proximal relaxation of the gradient-based attribute pruning stage (Stage 1). Recall the decomposition $\mathcal{L}(\mathbf{w}) = f(\mathbf{w}) + g(\mathbf{w})$ from Section 4.1:

$$\mathcal{L}(\mathbf{w}) = \underbrace{q(\mathbf{w}) + \lambda h(\mathbf{w})}_{f(\mathbf{w})} + \underbrace{\eta\|\mathbf{w}\|_1}_{g(\mathbf{w})}, \tag{11}$$

where $f(\mathbf{w}) = q(\mathbf{w}) + \lambda h(\mathbf{w})$ is convex (q is smooth; h is piecewise linear due to ReLU and hence convex but non-smooth), and $g(\mathbf{w})$ is convex but non-smooth.

This decomposition naturally motivates a proximal (sub)gradient framework, where each iteration takes the form:

$$\mathbf{w}^{t+1} = \text{prox}_{\alpha g}\left(\mathbf{w}^t - \alpha \mathbf{g}_f^t\right), \quad \mathbf{g}_f^t \in \partial f(\mathbf{w}^t), \tag{12}$$

with the proximal operator admitting a closed-form solution. Since $\mathbf{w} \in [0,1]^M$ implies $g(\mathbf{w}) = \eta\sum_j w_j$ (no absolute value needed), the proximal step combined with box projection gives:

$$\text{prox}_{\alpha g}(\mathbf{u})_j = \text{clip}(u_j - \alpha\eta,\ 0,\ 1) = \min\big(\max(u_j - \alpha\eta,\ 0),\ 1\big). \tag{13}$$

We begin with two preparatory lemmas.

Lemma I.5 (Smoothness of q and Bounded Subgradient of f)**.** *The following properties hold:*

(i) The smooth component $q(\mathbf{w})$ *has Lipschitz-continuous gradient with constant* $L_s = M/\sigma^2$*:*

$$\|\nabla q(\mathbf{w}_1) - \nabla q(\mathbf{w}_2)\| \le \frac{M}{\sigma^2}\|\mathbf{w}_1 - \mathbf{w}_2\|, \quad \forall\, \mathbf{w}_1, \mathbf{w}_2 \in [0,1]^M. \tag{14}$$

(ii) The piecewise-linear component $h(\mathbf{w})$ *is convex with bounded subgradients* $\|\mathbf{g}_h\| \le \sqrt{M}$ *for all* $\mathbf{g}_h \in \partial h(\mathbf{w})$ *and all* $\mathbf{w} \in [0,1]^M$*.*

(iii) For any $\mathbf{w} \in [0,1]^M$ *and any subgradient* $\mathbf{g}_f = \nabla q(\mathbf{w}) + \lambda \mathbf{g}_h \in \partial f(\mathbf{w})$ *with* $\mathbf{g}_h \in \partial h(\mathbf{w})$*:* $\|\mathbf{g}_f\| \le G_f \triangleq \sqrt{M}/\sigma + \lambda\sqrt{M}$*.*

Proof. **(i)** The Hessian of q satisfies $\nabla^2 q(\mathbf{w}) = \frac{1}{|\mathcal{N}||\mathcal{P}|\sigma^2}\sum_{i,j}\exp(-\mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}}/\sigma)\cdot D_{ij}^{\mathcal{N},\mathcal{P}}(D_{ij}^{\mathcal{N},\mathcal{P}})^\top$. Since $D_{ij}^{\mathcal{N},\mathcal{P}} \in \{0,1\}^M$, $\|D_{ij}^{\mathcal{N},\mathcal{P}}\|^2 \le M$, and $\exp(\cdot) \le 1$, we have $\|\nabla^2 q(\mathbf{w})\|_2 \le M/\sigma^2$.

(ii) Since $h(\mathbf{w}) = \frac{1}{|\mathcal{P}|^2}\sum_{i,j\in\mathcal{P}}\text{ReLU}(\mathbf{w}^\top D_{ij}^{\mathcal{P},\mathcal{P}})$, any subgradient satisfies $\mathbf{g}_h \in \frac{1}{|\mathcal{P}|^2}\sum_{i,j}[0,1]\cdot D_{ij}^{\mathcal{P},\mathcal{P}}$, giving $\|\mathbf{g}_h\|_\infty \le 1$ and $\|\mathbf{g}_h\| \le \sqrt{M}$.

(iii) By the triangle inequality: $\|\mathbf{g}_f\| \le \|\nabla q\| + \lambda\|\mathbf{g}_h\| \le \sqrt{M}/\sigma + \lambda\sqrt{M}$. □

Lemma I.6 (Straight-Through Gradient Estimator)**.** *Let* $\mathbf{v}^{(k)} \sim \text{Bernoulli}(\mathbf{w})$ *component-wise. In the proximal framework, the sparsity term* $g(\mathbf{w}) = \eta\|\mathbf{w}\|_1$ *is handled by the proximal operator,*

so the gradient estimator targets $f = q + \lambda h$ only. The straight-through estimator *[29]* $\hat{\mathbf{g}} = \frac{1}{B}\sum_{k=1}^{B} \mathbf{g}_f(\mathbf{v}^{(k)})$*, where $\mathbf{g}_f(\mathbf{v}) \in \partial_{\mathbf{v}} f(\mathbf{v})$ is a fixed subgradient selection (with the convention $\mathrm{ReLU}'(0) = 0$), evaluates a subgradient of f at the discrete samples $\mathbf{v}^{(k)} \in \{0,1\}^M$.*

(i) ***Unbiased for the sampled landscape:***

$$\mathbb{E}\left[\hat{\mathbf{g}}\right] = \mathbb{E}_{\mathbf{v}\sim\mathrm{Bernoulli}(\mathbf{w})}\left[\mathbf{g}_f(\mathbf{v})\right]. \tag{15}$$

(ii) ***Bounded second moment:***

$$\mathbb{E}\left\|\hat{\mathbf{g}}\right\|^2 \le G^2 \triangleq M\left(\frac{1}{\sigma} + \lambda\right)^2, \tag{16}$$

uniformly over $\mathbf{w} \in [0,1]^M$ and batch size B.

(iii) ***Relationship to the true policy gradient:*** *$\mathbb{E}[\hat{\mathbf{g}}] \neq \nabla_{\mathbf{w}}\mathbb{E}[f(\mathbf{v})]$ in general. The unbiased policy gradient (REINFORCE) estimator would be $\mathbb{E}[f(\mathbf{v})\nabla_{\mathbf{w}} \log p(\mathbf{v}|\mathbf{w})]$, which has higher variance. The straight-through estimator trades unbiasedness for lower variance, and is a standard choice in discrete optimization [29].*

Proof. (i) is immediate from the i.i.d. sampling: $\mathbb{E}[\hat{\mathbf{g}}] = \mathbb{E}[\mathbf{g}_f(\mathbf{v})]$ by linearity of expectation.

(ii): At $\mathbf{v} \in \{0,1\}^M$, the function f may be non-differentiable (the ReLU kinks $\mathbf{v}^\top D_{ij}^{\mathcal{P},\mathcal{P}} = 0$ can occur with positive probability under discrete Bernoulli sampling). Under the convention $\mathrm{ReLU}'(0) = 0$, a subgradient $\mathbf{g}_f(\mathbf{v}) \in \partial_{\mathbf{v}} f(\mathbf{v})$ is well-defined everywhere. For q: each component of $\nabla q(\mathbf{v})$ is bounded by $1/\sigma$ in absolute value, so $\|\nabla q\| \le \sqrt{M}/\sigma$. For h: each component of the selected subgradient $\mathbf{g}_h(\mathbf{v})$ lies in $[0,1]$ (it is an average of $\{0,1\}$-valued terms), so $\|\mathbf{g}_h\| \le \sqrt{M}$. By the triangle inequality: $\|\mathbf{g}_f(\mathbf{v})\| \le \sqrt{M}/\sigma + \lambda\sqrt{M}$, hence $\|\mathbf{g}_f(\mathbf{v})\|^2 \le M(1/\sigma + \lambda)^2$. Since $\mathbb{E}\|\hat{\mathbf{g}}\|^2 \le \mathbb{E}\|\mathbf{g}_f(\mathbf{v}^{(1)})\|^2$ by Jensen, the bound follows. □

We now prove Theorem I.1 with explicit constants. Let $\mathbf{w}^* \in [0,1]^M$ denote a global minimizer of $\mathcal{L}(\mathbf{w})$, $R_0 = \|\mathbf{w}^0 - \mathbf{w}^*\|$, and $G_f = \sqrt{M}(1/\sigma + \lambda)$, an upper bound on $\|\mathbf{g}_f^t\|$ for any subgradient $\mathbf{g}_f^t \in \partial f(\mathbf{w}^t)$ with $\mathbf{w}^t \in [0,1]^M$ (cf. Lemma I.5(iii)).

Exact subgradient. Consider the proximal update (12) using any $\mathbf{g}_f^t \in \partial f(\mathbf{w}^t)$, with constant step size $\alpha = R_0/(G_f\sqrt{T})$. After T iterations:

$$\min_{0\le t<T} \mathcal{L}(\mathbf{w}^t) - \mathcal{L}(\mathbf{w}^*) = \mathcal{O}\left(\frac{R_0\, G_f}{\sqrt{T}}\right). \tag{17}$$

Proof of Theorem I.1. The objective decomposes as $\mathcal{L} = f + \tilde{g}$ where $f = q + \lambda h$ is convex (but non-smooth, since h is piecewise linear) and $\tilde{g}(\mathbf{w}) = \eta\|\mathbf{w}\|_1 + \iota_{[0,1]^M}(\mathbf{w})$ is convex and closed, with $\iota_{[0,1]^M}$ the indicator of the box constraint (value 0 on $[0,1]^M$, $+\infty$ elsewhere). The update (12) computes $\mathbf{w}^{t+1} = \mathrm{prox}_{\alpha\tilde{g}}(\mathbf{w}^t - \alpha\mathbf{g}_f^t)$ with $\mathbf{g}_f^t \in \partial f(\mathbf{w}^t)$; the proximal operator of $\tilde{g}$ is the clip formula $\mathrm{clip}(u_j - \alpha\eta,\, 0,\, 1)$. Let $\Delta = \mathbf{w}^{t+1} - \mathbf{w}^t$.

Step 1: Proximal Optimality. The first-order optimality condition gives $\tilde{\boldsymbol{\zeta}}^{t+1} \in \partial\tilde{g}(\mathbf{w}^{t+1})$ satisfying:

$$\mathbf{g}_f^t + \tilde{\boldsymbol{\zeta}}^{t+1} + \tfrac{1}{\alpha}\Delta = \mathbf{0}, \tag{18}$$

where $\partial\tilde{g}(\mathbf{w}^{t+1}) = \partial(\eta\|\cdot\|_1)(\mathbf{w}^{t+1}) + N_{[0,1]^M}(\mathbf{w}^{t+1})$ includes both the ℓ_1 subdifferential and the normal cone of the box constraint.

Step 2: One-Step Progress Bound. Combining two convexity inequalities—$f(\mathbf{w}^*) \ge f(\mathbf{w}^t) + \mathbf{g}_f^{t\top}(\mathbf{w}^* - \mathbf{w}^t)$ (applied at $\mathbf{w}^t$) and $\tilde{g}(\mathbf{w}^*) \ge \tilde{g}(\mathbf{w}^{t+1}) + \tilde{\boldsymbol{\zeta}}^{t+1\top}(\mathbf{w}^* - \mathbf{w}^{t+1})$ (applied at $\mathbf{w}^{t+1}$; valid since both points lie in $[0,1]^M$, so $\tilde{g}$ reduces to g)—with the identity $\langle\Delta,\, \mathbf{w}^* - \mathbf{w}^{t+1}\rangle = \frac{1}{2}(\|\mathbf{w}^t - \mathbf{w}^*\|^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|^2 - \|\Delta\|^2)$ and Young's inequality $-\langle\mathbf{g}_f^t, \Delta\rangle \le \frac{\alpha}{2}\|\mathbf{g}_f^t\|^2 + \frac{1}{2\alpha}\|\Delta\|^2$:

$$f(\mathbf{w}^t) + g(\mathbf{w}^{t+1}) - \mathcal{L}(\mathbf{w}^*) \le \frac{1}{2\alpha}\left(\|\mathbf{w}^t - \mathbf{w}^*\|^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|^2\right) + \frac{\alpha}{2}\|\mathbf{g}_f^t\|^2. \tag{19}$$

Note that convexity of f is applied at $\mathbf{w}^t$ and convexity of $\tilde{g}$ at $\mathbf{w}^{t+1}$—both at the *correct* points where the respective subgradients are available. The normal cone contribution from $\tilde{\boldsymbol{\zeta}}^{t+1}$ is absorbed into the convexity inequality for $\tilde{g}$ and does not appear explicitly.

Step 3: Telescoping. Summing (19) over $t = 0, \ldots, T-1$ with $\|\mathbf{g}_f^t\| \le G_f$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\big[f(\mathbf{w}^t) + g(\mathbf{w}^{t+1}) - \mathcal{L}(\mathbf{w}^*)\big] \le \frac{R_0^2}{2\alpha T} + \frac{\alpha G_f^2}{2}. \tag{20}$$

Setting $\alpha = R_0/(G_f\sqrt{T})$ yields a right-hand side of $R_0 G_f/\sqrt{T}$. To convert to $\mathcal{L}(\mathbf{w}^t)$: rewriting $\sum_t [g(\mathbf{w}^{t+1}) - g(\mathbf{w}^t)] = g(\mathbf{w}^T) - g(\mathbf{w}^0)$, which is bounded in absolute value by ηM on $[0,1]^M$, gives $\frac{1}{T}\sum_t \mathcal{L}(\mathbf{w}^t) \le \mathcal{L}(\mathbf{w}^*) + R_0 G_f/\sqrt{T} + \mathcal{O}(1/T)$, hence $\min_t \mathcal{L}(\mathbf{w}^t) - \mathcal{L}(\mathbf{w}^*) = \mathcal{O}(R_0 G_f/\sqrt{T})$, establishing (17).

□

Remark I.7 (Stochastic Extension — Informal). When the exact subgradient $\mathbf{g}_f^t$ in Theorem I.1 is replaced by the straight-through estimator $\hat{\mathbf{g}}^t$ (Lemma I.6), the one-step bound (19) acquires an additional cross-term $(\hat{\mathbf{g}}^t - \mathbf{g}_f^t)^\top(\mathbf{w}^* - \mathbf{w}^{t+1})$. Decomposing into bias $\mathbf{b}^t = \mathbb{E}[\hat{\mathbf{g}}^t] - \mathbf{g}_f^t$ and zero-mean noise: the variance contributes $\mathcal{O}(\alpha G^2/B)$ per step (Lemma I.6(ii)), and the bias contributes δ_{ST}. This suggests the heuristic bound:

$$\mathbb{E}\left[\min_{0\le t<T}\ \mathcal{L}(\mathbf{w}^t)\right] - \mathcal{L}(\mathbf{w}^*) \lesssim \mathcal{O}\left(\frac{R_0\, G_f}{\sqrt{T}}\right) + \frac{\alpha\, G^2}{2B} + \delta_{\mathrm{ST}}, \tag{21}$$

where $G^2 = G_f^2$ and δ_{ST} denotes the accumulated straight-through bias. A rigorous proof would require conditioning on the filtration $\mathcal{F}_t = \sigma(\hat{\mathbf{g}}^0, \ldots, \hat{\mathbf{g}}^{t-1})$ to properly handle the dependence of $\mathbf{w}^{t+1}$ on $\hat{\mathbf{g}}^t$; we leave the complete conditional-expectation derivation for future work. Setting $\alpha = \mathcal{O}(1/\sqrt{T})$ yields an informal rate of $\mathcal{O}(1/\sqrt{T} + \delta_{\mathrm{ST}})$.

Remark I.8. Several remarks are in order.

(i) **Rate.** The $\mathcal{O}(1/\sqrt{T})$ rate in Theorem I.1 is the standard guarantee for proximal subgradient methods on general convex objectives [39]. Although $\mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}} \ge 0$ reduces $h(\mathbf{w})$ to a linear form on $[0,1]^M$, PRISM adopts the boundary subgradient selection $\mathrm{ReLU}'(0) = 0$ (Eq. (8)) so that matched anomaly pairs ($\mathbf{w}^\top D_{i,j}^{\mathcal{P},\mathcal{P}} = 0$) incur zero pull penalty on discrete masks; because this oracle is discontinuous across boundary faces, Theorem I.1 uses the general subgradient analysis and yields the conservative $\mathcal{O}(1/\sqrt{T})$ bound.

(ii) **Straight-through bias in practice.** The straight-through estimator is a standard technique in discrete optimization [29]. Empirically, the bias δ_{ST} is small when the smooth component q varies slowly relative to the Bernoulli sampling scale (controlled by σ), so evaluating subgradients at nearby binary points introduces only mild bias. In practice, we find $B = 32$ sufficient for reliable convergence.

(iii) **Initialization.** We initialize $\mathbf{w}^0 = \mathbf{1}_M/M$. Since $\mathbf{w}^*, \mathbf{w}^0 \in [0,1]^M$, the worst-case bound is $\|\mathbf{w}^0 - \mathbf{w}^*\|^2 \le M$. In practice, $\mathbf{w}^*$ is sparse with $|\mathcal{C}^*| \ll M$ nonzero components, so $\|\mathbf{w}^0 - \mathbf{w}^*\|^2 \le |\mathcal{C}^*| + (M - |\mathcal{C}^*|)/M^2 \le |\mathcal{C}^*| + 1$, which is a small constant independent of M.

(iv) **Scope of the theorem.** Theorem I.1 analyzes the exact proximal subgradient update (12). The straight-through estimator and z-score normalization in Eq. (9) are practical pruning heuristics rather than part of the formal rate.

I.2 Optimality of Stage 2

Having established convergence for the idealized Stage 1 relaxation, we now prove Proposition I.3. Recall that for $\mathcal{A}_{\mathrm{cand}} = \mathrm{Top}_m^+(\mathbf{w}^T)$ (selecting the at most m attributes with the largest positive final weights), Stage 2 selects $\mathcal{C}^*$ maximizing exact F1 over $\mathcal{A}_{\mathrm{cand}}$, *i.e.*,

$$\mathcal{C}^* = \operatorname*{arg\,max}_{\mathcal{C}\subseteq\mathcal{A}_{\mathrm{cand}},\, \mathcal{C}\neq\emptyset} \mathrm{F1}(\mathrm{Precision}(\mathcal{C}), \mathrm{Recall}(\mathcal{C})). \tag{22}$$

Proof of Proposition I.3. Stage 2 exhaustively enumerates all $2^{|\mathcal{A}_{\text{cand}}|} - 1$ non-empty subsets of $\mathcal{A}_{\text{cand}}$ and evaluates the F1 score for each. Since the search is exhaustive over the entire feasible set, the returned solution is by definition globally optimal within $\mathcal{A}_{\text{cand}}$. With the operational cap $m \leq 5$, the number of subsets is at most $2^5 - 1 = 31$, making the enumeration computationally negligible. □

Remark I.9. The quality of the Stage 2 solution depends on whether the true root cause attributes are retained in $\mathcal{A}_{\text{cand}}$ after Stage 1 pruning. In Section I.5, we formalize the *discriminative margin condition* (DMC) as a sufficient support-recovery certificate for the idealized relaxation. Combined with Stage 2 exhaustive search, this yields a conditional end-to-end guarantee when $\mathcal{C}^* \subseteq \mathcal{A}_{\text{cand}}$ (Corollary I.21); empirically, Stage 1 retains the true support across our evaluation suite.

I.3 Root Cause Identifiability

Before analyzing the optimization algorithm, we establish that the F1 objective itself correctly identifies the true root cause. That is, we prove that $\mathcal{C}^*$ is the unique maximizer of $\text{F1}(\text{Precision}(\mathcal{C}), \text{Recall}(\mathcal{C}))$ under a formal data model that instantiates Observations 4.1–4.2.

Assumption I.10 (Structural Data Model)**.** The data $(\mathcal{P}, \mathcal{N})$ with true root cause $\mathcal{C}^* \subseteq \mathcal{M}$ satisfies:

(A1) **Anomaly Aggregation.** All anomalous samples share an identical attribute value signature under $\mathcal{C}^*$: for all $x_i, x_j \in \mathcal{P}$ and all $k \in \mathcal{C}^*$, $A_k^{(i)} = A_k^{(j)}$. Consequently, $\text{Recall}(\mathcal{C}^*) = 1$.

(A2) **Normal Diffusion.** Precision is monotonically non-decreasing in the attribute set: for any $\mathcal{C} \subseteq \mathcal{C}' \subseteq \mathcal{M}$, $|\mathcal{N}(A^{\mathcal{P}}(\mathcal{C}'))| \leq |\mathcal{N}(A^{\mathcal{P}}(\mathcal{C}))|$, *i.e.*, selecting more attributes can only reduce the number of normal samples that match $\mathcal{P}$'s signature.

(A3) **Irreducibility.** Every attribute in $\mathcal{C}^*$ is necessary for discrimination: for any $\mathcal{C} \subsetneq \mathcal{C}^*$, $|\mathcal{N}(A^{\mathcal{P}}(\mathcal{C}))| > |\mathcal{N}(A^{\mathcal{P}}(\mathcal{C}^*))|$, *i.e.*, dropping any root cause attribute strictly increases the false positive count.

(A4) **Non-Degeneracy.** Every attribute outside $\mathcal{C}^*$ fragments the anomaly cluster: for any $j \notin \mathcal{C}^*$, there exist $x_a, x_b \in \mathcal{P}$ with $A_j^{(a)} \neq A_j^{(b)}$, *i.e.*, including any non-root-cause attribute strictly reduces Recall.

(A5) **Discrimination Dominance.** The root cause signature is sufficiently distinctive that Precision under $\mathcal{C}^*$ exceeds the Recall of any competitor involving non-root-cause attributes:

$$\text{Precision}(\mathcal{C}^*) > \max_{\mathcal{C}' \not\subseteq \mathcal{C}^*} \text{Recall}(\mathcal{C}'). \tag{23}$$

Remark I.11. (A1)–(A4) formalize Observations 4.1–4.2: anomalies share a common fault signature, while normal samples are diffuse. (A5) is a quantitative condition requiring the root cause to be "strong enough": Precision$(\mathcal{C}^*)$ must exceed all competitors' Recall. In production RCA, the fault signature typically covers only a small fraction of normal traffic (yielding Precision ≈ 1), while any competitor involving irrelevant attributes has Recall < 1, making (A5) easily satisfied.

Theorem I.12 (Root Cause Identifiability)**.** *Under Assumption I.10, the true root cause $\mathcal{C}^*$ is the unique maximizer of* $\text{F1}(\text{Precision}(\mathcal{C}), \text{Recall}(\mathcal{C}))$ *over all non-empty $\mathcal{C} \subseteq \mathcal{M}$.*

Proof. We show that any $\mathcal{C}' \neq \mathcal{C}^*$ achieves strictly lower F1. Let $P = \text{Precision}(\mathcal{C}^*)$, $P' = \text{Precision}(\mathcal{C}')$, and $R' = \text{Recall}(\mathcal{C}')$. By (A1), $\text{Recall}(\mathcal{C}^*) = 1$, so:

$$\text{F1}(\mathcal{C}^*) = \frac{2P}{1+P}. \tag{24}$$

Case 1: $\mathcal{C}' \subsetneq \mathcal{C}^*$ **(under-selection).** Since $\mathcal{C}' \subsetneq \mathcal{C}^*$, by (A1) $\mathcal{P}$ remains aggregated under $\mathcal{C}'$, so $R' = 1$. By (A3), $P' < P$. Thus $\text{F1}(\mathcal{C}') = 2P'/(1+P') < 2P/(1+P) = \text{F1}(\mathcal{C}^*)$, since $x \mapsto 2x/(1+x)$ is strictly increasing.

Case 2: $\mathcal{C}' \supsetneq \mathcal{C}^*$ **(over-selection).** By (A4), $R' < 1$. Since $P' \leq 1$:

$$\text{F1}(\mathcal{C}') = \frac{2P'R'}{P'+R'} \leq \frac{2R'}{1+R'}, \tag{25}$$

where the inequality uses $P' \leq 1$ and the fact that $\frac{2ab}{a+b}$ is increasing in a for fixed b, maximized at $a = 1$. By (A5), $P > R'$, so $\frac{2R'}{1+R'} < \frac{2P}{1+P} = \mathrm{F1}(\mathcal{C}^*)$, since $x \mapsto 2x/(1+x)$ is strictly increasing.

Case 3: $\mathcal{C}' \not\subseteq \mathcal{C}^*$ and $\mathcal{C}' \not\supseteq \mathcal{C}^*$ (partial overlap). Since $\mathcal{C}' \setminus \mathcal{C}^* \neq \emptyset$, by (A4): $R' < 1$. The same bound as Case 2 applies: $\mathrm{F1}(\mathcal{C}') \leq 2R'/(1+R') < 2P/(1+P) = \mathrm{F1}(\mathcal{C}^*)$, where the last inequality uses (A5). □

Remark I.13. Theorem I.12 closes the first link in the proof chain: the F1 objective is a *faithful proxy* for the true root cause under the structural data model. The subsequent sections establish that PRISM's optimization algorithm recovers the F1 maximizer, completing the end-to-end guarantee. The five assumptions are verifiable on any given dataset: (A1) and (A4) can be checked from the ground-truth labels, (A2) holds by construction (adding attributes only refines the matching criterion), (A3) is testable by ablation, and (A5) reduces to checking whether the root cause signature is rare enough in the normal population. Note that Assumption I.10 characterizes an idealized noiseless regime ($\mathrm{Recall}(\mathcal{C}^*) = 1$, analogous to hard-margin separability in SVMs or noiseless support recovery in Lasso) to isolate the structural identifiability of the F1 objective; empirical robustness under realistic upstream label contamination ($\mathrm{Recall}(\mathcal{C}^*) < 1$, Observation 4.2) is governed by the relaxed cohesion penalty and validated in Figure 6.

I.4 Surrogate Tightness

The continuous surrogate $\mathcal{L}(\mathbf{w})$ (Eq. 6) approximates the discrete Precision–Recall objective (Problem 4.5). We quantify the approximation quality for each component.

Proposition I.14 (Precision Surrogate Bound)**.** *For any $\mathbf{w} \in \{0,1\}^M$ and $\sigma > 0$, let $\ell(\mathbf{w}) = 1 - \mathrm{Precision}(\mathbf{w})$ denote the true precision loss. Then:*

$$\frac{\ell(\mathbf{w})}{|\mathcal{P}|} \;\leq\; q(\mathbf{w}) \;\leq\; \ell(\mathbf{w}) + e^{-1/\sigma}. \tag{26}$$

Proof. For each $i \in \mathcal{N}$, define $S_i = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \exp(-\mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}}/\sigma)$, so that $q = \frac{1}{|\mathcal{N}|} \sum_i S_i$. Since $\mathbf{w} \in \{0,1\}^M$, each $\mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}} \in \{0, 1, \ldots, M\}$.

Case 1 ($\min_j \mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}} = 0$, sample i is a false positive): at least one summand equals $\exp(0) = 1$, so $S_i \geq 1/|\mathcal{P}|$; trivially $S_i \leq 1$.

Case 2 ($\min_j \mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}} \geq 1$, sample i is correctly separated): every summand satisfies $\exp(-\mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}}/\sigma) \leq e^{-1/\sigma}$, so $S_i \leq e^{-1/\sigma}$.

Let $n_{\mathrm{FP}} = |\{i \in \mathcal{N} : \min_j \mathbf{w}^\top D_{ij}^{\mathcal{N},\mathcal{P}} = 0\}|$, so $\ell(\mathbf{w}) = n_{\mathrm{FP}}/|\mathcal{N}|$. Summing:

$$q(\mathbf{w}) \geq \frac{n_{\mathrm{FP}}}{|\mathcal{N}||\mathcal{P}|} = \frac{\ell(\mathbf{w})}{|\mathcal{P}|}, \tag{27}$$

$$q(\mathbf{w}) \leq \frac{n_{\mathrm{FP}} + (|\mathcal{N}| - n_{\mathrm{FP}}) \cdot e^{-1/\sigma}}{|\mathcal{N}|} = \ell(\mathbf{w}) + (1 - \ell(\mathbf{w}))\, e^{-1/\sigma} \leq \ell(\mathbf{w}) + e^{-1/\sigma}. \qquad \square$$

The bound gives a scaled sandwich relation: false-positive samples contribute between $1/|\mathcal{P}|$ and 1 to q, while correctly separated samples contribute at most $e^{-1/\sigma}$.

Proposition I.15 (Recall Constraint Conservatism)**.** *For any $\mathbf{w} \in \{0,1\}^M$, if $h(\mathbf{w}) \leq 1 - \tau$ then $\mathrm{Recall}(\mathbf{w}) \geq \tau$. More precisely:*

$$1 - \mathrm{Recall}(\mathbf{w}) \;\leq\; h(\mathbf{w}) \;\leq\; M\big(1 - \mathrm{Recall}(\mathbf{w})\big). \tag{28}$$

Proof. For $\mathbf{w} \in \{0,1\}^M$, each $\mathbf{w}^\top D_{ij}^{\mathcal{P},\mathcal{P}} \in \{0, 1, \ldots, M\}$. For any non-negative integer k:

$$\mathrm{Sign}(k) \;\leq\; k = \mathrm{ReLU}(k) \;\leq\; M \cdot \mathrm{Sign}(k), \tag{29}$$

where the upper bound uses $k \leq M$ when $\mathrm{Sign}(k) = 1$. Summing over all $(i, j) \in \mathcal{P} \times \mathcal{P}$ and dividing by $|\mathcal{P}|^2$:

$$1 - \mathrm{Recall}(\mathbf{w}) = \frac{1}{|\mathcal{P}|^2} \sum_{i,j} \mathrm{Sign}(\mathbf{w}^\top D_{ij}^{\mathcal{P},\mathcal{P}}) \leq \frac{1}{|\mathcal{P}|^2} \sum_{i,j} \mathrm{ReLU}(\mathbf{w}^\top D_{ij}^{\mathcal{P},\mathcal{P}}) = h(\mathbf{w}).$$

If $h(\mathbf{w}) \leq 1 - \tau$, then $1 - \mathrm{Recall}(\mathbf{w}) \leq 1 - \tau$, *i.e.*, $\mathrm{Recall}(\mathbf{w}) \geq \tau$. The upper bound follows analogously from $\mathrm{ReLU}(k) \leq M \cdot \mathrm{Sign}(k)$. □

Together, Propositions I.14–I.15 establish that (i) the surrogate q has a scaled sandwich relation with the true precision loss, and (ii) the cohesion loss gives a conservative recall-feasibility test—any $\mathbf{w}$ with $h(\mathbf{w}) \leq 1 - \tau$ automatically satisfies $\mathrm{Recall}(\mathbf{w}) \geq \tau$.

I.5 Support Identification Guarantee

Stage 2 optimality (Proposition I.3) hinges on Stage 1 retaining the true root cause attributes in $\mathcal{A}_{\mathrm{cand}}$. We first establish an idealized support-recovery certificate for the proximal relaxation.

Definition I.16 (Discriminative Margin Condition)**.** The idealized relaxation satisfies the *discriminative margin condition* (DMC) with respect to the true root cause $\mathcal{C}^*$ and sparsity penalty $\eta > 0$ if, at the minimizer $\mathbf{w}^*$ of $f(\mathbf{w}) + \eta\|\mathbf{w}\|_1$ over $[0, 1]^M$:

(i) **Retention:** For every $j \in \mathcal{C}^*$ and every $g_{h,j} \in \partial_{w_j} h(\mathbf{w}^*)$,

$$\underbrace{\left|\nabla_{w_j} q(\mathbf{w}^*)\right|}_{\text{discrimination (push)}} > \underbrace{\lambda\, g_{h,j}}_{\text{fragmentation (pull)}} + \eta. \tag{30}$$

(ii) **Exclusion:** For every $j \notin \mathcal{C}^*$,

$$\left|\nabla_{w_j} q(\mathbf{w}^*)\right| < \eta. \tag{31}$$

Retention states that for each root cause attribute, the inter-class discrimination signal exceeds the intra-class fragmentation cost plus the sparsity threshold. Exclusion states that non-root-cause attributes carry discrimination signal below the sparsity threshold.

Theorem I.17 (Idealized Support Recovery)**.** *If the data satisfies the discriminative margin condition (Definition I.16), then the minimizer* $\mathbf{w}^*$ *of* $f(\mathbf{w}) + \eta\|\mathbf{w}\|_1$ *over* $[0, 1]^M$ *satisfies* $\mathrm{supp}(\mathbf{w}^*) = \mathcal{C}^*$*, i.e.,* $w_j^* > 0$ *for all* $j \in \mathcal{C}^*$ *and* $w_j^* = 0$ *for all* $j \notin \mathcal{C}^*$*.*

Proof. Since the feasible set is $[0, 1]^M$ and $\|\mathbf{w}\|_1 = \sum_j w_j$ on this domain, the objective $F(\mathbf{w}) = f(\mathbf{w}) + \eta \sum_j w_j$ is convex. The optimality conditions for the box-constrained minimizer $\mathbf{w}^*$ require that there exists a subgradient $g_{f,j} \in \partial_{w_j} f(\mathbf{w}^*)$ such that, for each component j:

- If $0 < w_j^* < 1$ (interior): $g_{f,j} + \eta = 0$.
- If $w_j^* = 0$ (lower boundary): $g_{f,j} + \eta \geq 0$, *i.e.*, $g_{f,j} \geq -\eta$.
- If $w_j^* = 1$ (upper boundary): $g_{f,j} + \eta \leq 0$, *i.e.*, $g_{f,j} \leq -\eta$.

For any subgradient $g_{f,j} \in \partial_{w_j} f(\mathbf{w}^*)$, we have $g_{f,j} = \nabla_j q(\mathbf{w}^*) + \lambda g_{h,j}$ for some $g_{h,j} \in \partial_{w_j} h(\mathbf{w}^*)$, where $\nabla_j q \leq 0$ (Eq. 7) and $g_{h,j} \geq 0$ (Eq. 8).

Retention ($j \in \mathcal{C}^*$). By condition (30), $-\nabla_j q(\mathbf{w}^*) > \lambda g_{h,j} + \eta$ holds for *every* $g_{h,j} \in \partial_{w_j} h(\mathbf{w}^*)$, so $g_{f,j} = \nabla_j q + \lambda g_{h,j} < -\eta$ for all subgradients. The KKT condition for $w_j^* = 0$ requires $g_{f,j} \geq -\eta$ for some subgradient, which is impossible. Hence $w_j^* > 0$.

Exclusion ($j \notin \mathcal{C}^*$). We show $w_j^* > 0$ leads to a contradiction. By condition (31), $|\nabla_j q(\mathbf{w}^*)| < \eta$, so $\nabla_j q(\mathbf{w}^*) > -\eta$. Since $g_{h,j} \geq 0$ for *every* subgradient selection, we have $g_{f,j} = \nabla_j q + \lambda g_{h,j} > -\eta$ for all $g_{h,j} \in \partial_{w_j} h(\mathbf{w}^*)$. Now:

- If $0 < w_j^* < 1$: optimality requires $g_{f,j} = -\eta$ for some subgradient. But $g_{f,j} > -\eta$ for all subgradients. Contradiction.

- If $w_j^* = 1$: optimality requires $g_{f,j} \leq -\eta$ for some subgradient. But $g_{f,j} > -\eta$ for all subgradients. Contradiction.

Hence $w_j^* = 0$ for all $j \notin \mathcal{C}^*$. □

Remark I.18 (Physical Interpretation of DMC). The discriminative margin condition has a natural interpretation through the lens of Observations 4.1–4.2:

(i) **Retention is grounded in Observation 4.1:** Under the true root cause $\mathcal{C}^*$, anomalous samples $\mathcal{P}$ share identical attribute values on dimensions $j \in \mathcal{C}^*$, so $D_{ij,k}^{\mathcal{P},\mathcal{P}} = 0$ for most (i, j) pairs when $k \in \mathcal{C}^*$. This drives all subgradients $g_{h,k} \in \partial_{w_k} h(\mathbf{w}^*)$ close to 0 (minimal fragmentation cost). Simultaneously, the diffuse distribution of $\mathcal{N}$ ensures substantial inter-class divergence on root cause dimensions, yielding large $|\nabla_{w_k} q|$ (strong discrimination signal).

(ii) **Exclusion is grounded in Observation 4.2:** Non-root-cause attributes exhibit similar value distributions between $\mathcal{P}$ and $\mathcal{N}$: label contamination mixes the two populations on irrelevant dimensions. Consequently, $D_{ij,k}^{\mathcal{N},\mathcal{P}}$ for $k \notin \mathcal{C}^*$ carries no systematic signal, yielding $|\nabla_{w_k} q| \approx 0$, well below η.

Empirically, the deployed top-m candidate set retains all true root-cause attributes across all tested datasets (Section 5, Figure 7), matching the support-retention behavior certified by DMC for the idealized relaxation.

I.6 Normalized Gradient Update

The practical PRISM algorithm applies z-score normalization to the stochastic gradient (Eq. 9). We record the pointwise order-preservation property used to interpret the normalized pruning signal.

Proposition I.19 (Order Preservation under Normalization)**.** *Let $\mathbf{g} \in \mathbb{R}^M$ be a gradient vector with non-degenerate variance ($\exists\, j, k : g_j \neq g_k$). The z-score normalized gradient $\tilde{\mathbf{g}}$, defined by $\tilde{g}_j = (g_j - \bar{g})/s_g$ where $\bar{g} = \frac{1}{M}\sum_k g_k$ and $s_g = \sqrt{\frac{1}{M}\sum_k (g_k - \bar{g})^2}$, satisfies:*

(i) ***Order preservation:*** *$g_j < g_k \iff \tilde{g}_j < \tilde{g}_k$ for all j, k.*

(ii) ***Sign semantics:*** *$\tilde{g}_j < 0 \iff g_j < \bar{g}$, i.e., attribute j has a stronger-than-average push signal.*

Proof. (i)–(ii) follow directly from the fact that z-score normalization is an affine map $\tilde{g}_j = (g_j - \bar{g})/s_g$ with positive scaling $s_g > 0$. Affine maps with positive slope preserve order and map the mean to zero. □

Remark I.20. Because z-score normalization is invariant to uniform shifts ($\mathrm{Normalize}(\mathbf{g} + \eta\mathbf{1}) = \mathrm{Normalize}(\mathbf{g})$) and maps weaker-than-average push signals ($g_j > \bar{g}$) to $\tilde{g}_j > 0$, combining $\tilde{\mathbf{g}}$ with $\mathrm{clip}(\cdot, 0, 1)$ inherently shrinks irrelevant weights toward zero without an explicit η parameter in Eq. (9). This proposition is pointwise: it preserves the ordering of a fixed gradient vector, but does not by itself imply convergence of the normalized iteration.

I.7 End-to-End Optimality Guarantee

Combining the preceding results, we establish the overall guarantee for PRISM.

Corollary I.21 (Conditional End-to-End Optimality under Candidate Retention)**.** *Suppose the following conditions hold:*

(C1) The data satisfies Assumption I.10 (structural data model).

(C2) Stage 1 outputs $\mathcal{A}_{\mathrm{cand}} = \mathrm{Top}_m^+(\mathbf{w}^T)$ with $\mathcal{C}^* \subseteq \mathcal{A}_{\mathrm{cand}}$.

Then the two-stage pipeline returns the true root cause $\mathcal{C}^*$, which is the unique global maximizer of F1 *over all $2^M - 1$ attribute combinations.*

Proof. The proof proceeds as follows:

Step 1: Identifiability. By (C1) and Theorem I.12, $\mathcal{C}^*$ is the unique maximizer of F1 over all $\mathcal{C} \subseteq \mathcal{M}$.

Step 2: Candidate retention. By (C2), the retained top-m candidate set contains the true root cause attributes: $\mathcal{C}^* \subseteq \mathcal{A}_{\text{cand}}$.

Step 3: Exact search. By Proposition I.3, Stage 2 exhaustively evaluates the *exact* F1 over all $2^{|\mathcal{A}_{\text{cand}}|} - 1$ non-empty subsets of $\mathcal{A}_{\text{cand}}$, including $\mathcal{C}^*$.

Step 4: Global optimality. Since $\mathcal{C}^*$ is feasible in the Stage 2 search space and is the unique F1 maximizer among all $2^M - 1$ combinations (Step 1), Stage 2 returns $\mathcal{C}^*$. □

Remark I.22 (Scope and Gaps). The proof chain from data model to algorithm output is:

$$\underbrace{\text{Data Model}}_{(C1)} \rightarrow \underbrace{\mathcal{C}^* = \arg\max \text{F1}}_{\text{Thm. I.12}} \rightarrow \underbrace{\mathcal{C}^* \subseteq \mathcal{A}_{\text{cand}}}_{(C2)} \rightarrow \underbrace{\text{Exact search}}_{\text{Prop. I.3}} \rightarrow \mathcal{C}^*.$$

Two gaps remain between the idealized analysis and the practical algorithm:

(i) **Candidate retention (C2).** Theorem I.1 guarantees $\min_t \mathcal{L}(\mathbf{w}^t) \rightarrow \mathcal{L}(\mathbf{w}^*)$ in objective value, but does not directly imply that the deployed $\text{Top}_m^+(\mathbf{w}^T)$ contains $\mathcal{C}^*$ in finite time. We therefore state the end-to-end guarantee conditional on $\mathcal{C}^* \subseteq \mathcal{A}_{\text{cand}}$ and verify this retention empirically across all datasets (Section 5).

(ii) **Normalized gradient vs. proximal subgradient.** The practical update (9) applies z-score normalization, which differs from the proximal subgradient analyzed in Theorem I.1. Proposition I.19 only preserves the ordering of a fixed gradient vector; a formal convergence rate for the normalized update is not established. The convergence theorem thus provides qualitative justification rather than a quantitative guarantee for the exact algorithm deployed.

Accordingly, the formal rate applies to the idealized relaxation, and end-to-end correctness is conditional on candidate retention; the deployed solver is validated empirically.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main contributions, separated public/industrial benchmark claims, and scope boundaries are accurately reflected in the abstract and introduction (Section 1).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Methodological and benchmark limitations are explicitly discussed in a dedicated Limitations section (Section 6).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theorems and propositions presented in the paper are supported by proofs in the corresponding appendix sections.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full experimental protocols, upstream label generation, measure-based reconstruction rules, and dataset statistics are disclosed in Section 5.1 and Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and reproduction scripts are publicly available at `https://github.com/goaheand/PRISM`, and the NFC dataset is publicly released at `https://huggingface.co/datasets/Ysqq/NFC`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Experimental settings and protocols are detailed in Section 5.1 and Appendix D, with hyperparameter sensitivity analyzed in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Standard deviations of F1 and latency across random seeds are reported in Appendix G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resource specifications and per-benchmark execution times are provided in Section 5.1 and Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to it in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work studies algorithmic root cause localization on anonymized operational telemetry to improve service reliability and poses no foreseeable negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released optimization code and fully anonymized NFC telemetry dataset pose no high risk for misuse or dual-use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The existing public datasets and baseline implementations used in the paper are open-source, properly cited, and used in accordance with their licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Anonymization, SRE verification protocols, and dataset statistics for the NFC benchmark are documented in Appendix D and alongside the dataset release at `https://huggingface.co/datasets/Ysqq/NFC`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human-subjects research requiring IRB approval.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used solely for writing and editing polish and did *not* impact the core methodology, experiments, or scientific rigor.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.