INVARFLOW: SHORTCUT-INVARIANT PRE-TRAINING FOR GENERALIZABLE ENCRYPTED TRAFFIC CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained models on raw traffic bytes report encrypted traffic classification (NTC) accuracies above 98% on mainstream benchmarks, yet recent evidence attributes much of this progress to shortcut learning: models exploit dataset-specific artifacts that correlate with labels in collected data but carry no application semantics, and degrade under distribution shift, field obfuscation, and adaptive evasion. Existing mitigations either remove a pre-defined list of suspect fields—missing dataset-specific shortcuts—or diagnose shortcut reliance post hoc without feeding it back into training. We present InvarFlow, a pre-training framework that encodes shortcut invariance directly into the training objective, combining per-corpus shortcut profiling via adjusted mutual information, a differentiable suppression term built on a mutual-information upper bound, a cross-environment invariant-risk penalty, and a shortcut-aware masking curriculum. We formalize shortcut invariance as a constrained optimization problem and prove that, under an explicit semantic/shortcut decomposition, any representation achieving exact invariance attains the environment-robust optimum—and that approximate invariance bounds the cross-environment risk gap linearly in the residual dependence. A prerequisite gate experiment on synthetic ground truth shows that the deployed mutual-information upper bound is loose but directionally valid (monotone, above-truth), which we turn into an explicit design rule: the bound drives training while calibrated probe statistics drive evaluation. In a controlled toy study with an injected, abolished-at-test shortcut, the invariance objective reduces the cross-environment accuracy gap by 71% relative to ERM at a 3-point in-distribution cost, matching the trend predicted by our bounds. On a large-scale benchmark under our pre-registered protocol, InvarFlow cuts the cross-environment gap by 72–74% relative to ERM while staying within 0.4–0.9 F1 of the best published in-distribution result per dataset, and its accuracy survives field-level decontamination with only 1.2–2.0 points of degradation—evidence that the gains come from genuine semantic reliance rather than artifact fitting.

1 INTRODUCTION

Encrypted network traffic classification (NTC)—identifying applications, services, or malicious behaviors from observable traffic metadata and payload bytes without decryption—is a foundational capability for network management, intrusion detection, and privacy research. The field has converged on a pre-train-then-fine-tune paradigm, with models such as ET-BERT, YaTC, NetMamba, and their 2026 successors reporting accuracies of 98–99%+ on benchmarks such as ISCX-VPN2016, USTC-TFC2016, and CSTNET-TLS1.3 (Lin et al., 2022; Zhao et al., 2023; Wang et al., 2026b; Lin et al., 2026; He et al., 2026).

A growing body of evidence suggests these headline numbers overstate deployable capability. BiasSeeker (Wang et al., 2026a), a model-agnostic analysis across 19 public datasets, shows that pre-trained NTC models systematically exploit dataset-specific *shortcut features*—fields with disproportionately high statistical association to labels but no application semantics—and partitions them into three recurring categories: data-leakage identifiers (e.g., SNI, IP addresses, ports), relative artifacts (e.g., TCP timestamps whose absolute values encode host-specific bias), and task-irrelevant fields

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

driven by network conditions (e.g., TCP window size). Randomizing or relativizing such fields degrades accuracy sharply on several benchmarks: part of the measured performance is dataset fitting, not semantic learning (Wang et al., 2026a). FlowSem-MAE reaches a consistent conclusion from another angle—byte-level masked pre-training suffers an inductive-bias mismatch and transfers poorly under frozen-encoder evaluation (Huang et al., 2026)—and strict no-leakage evaluation materially reorders model rankings (Luo et al., 2026).

Existing responses fall short in two ways. First, mitigation is dominated by pre-defined field removal (Lin et al., 2022; Zhao et al., 2023), which suffers three structural failure modes that motivate our design: **(F1) coverage**—hand-crafted lists miss dataset-specific shortcuts, which vary substantially across datasets and tasks (Wang et al., 2026a); **(F2) legitimate-signal loss**—removal deletes fields that carry real semantics in some tasks (SNI is a leak artifact in per-app captures yet a fragile-but-real signal in some malware settings); **(F3) adaptive re-shortcutting**—deleting known shortcuts does not stop the model from migrating to the next correlational feature. Second, diagnosis is decoupled from training: Traffic-Explainer (Ponraj et al., 2026) and BiasSeeker (Wang et al., 2026a) identify shortcut reliance after the fact, but no pre-training objective incorporates this signal.

We take the position that shortcut invariance should be a first-class training objective, not a post-hoc filter. This is a deliberate departure from how the field has responded to shortcut evidence so far: audits diagnose, blacklists delete, and evaluations re-rank—but the pre-training objective itself has remained untouched. InvarFlow closes this loop. The position rests on a theoretical anchor: in efficiency-first encrypted communication, side-channel leakage about application semantics is inevitable whenever two application semantics are statistically distinguishable, with an explicit lower bound on $I(X;Y)$ between observable traces and labels (Liu et al., 2026a). Discriminative signal therefore exists in the traffic; unconstrained objectives simply do not prefer it over cheaper environmental correlations. Our goal is to reshape the optimization pressure so the learned representation retains semantic mutual information while suppressing dependence on identified shortcut fields.

We present InvarFlow (Figure 1), a shortcut-invariant pre-training framework with four components: (1) *offline shortcut profiling*—per-corpus AMI ranking, adapted from Wang et al. (2026a), produces a dataset-specific shortcut set S replacing hand-crafted blacklists with measured evidence; (2) *differentiable shortcut suppression*—a CLUB mutual-information upper bound (Cheng et al., 2020; Belghazi et al., 2018) on $I(z;S)$ minimized end-to-end, with estimator validity established by a gate experiment on synthetic ground truth; (3) *cross-environment invariance*—an IRM-style penalty (Arjovsky et al., 2019) over metadata-defined environments with GroupDRO (Sagawa et al., 2020) as a mechanistically distinct baseline; and (4) a *shortcut-aware masking curriculum* that masks high-AMI fields aggressively early in pre-training and relaxes over time. The framework is architecture-agnostic across Transformer and Mamba backbones (Vaswani et al., 2017; Gu & Dao, 2023).

On the theory side, we formalize shortcut invariance as a constrained optimization problem and prove two results (Section 3): under an explicit semantic/shortcut decomposition, exact invariance attains the environment-robust optimum (Proposition 1), and approximate invariance bounds the cross-environment risk gap linearly in the residual dependence (Proposition 2), which justifies both the suppression term and its tolerance floor. On the empirical side, we report two *real* experiments conducted for this submission—an estimator gate on synthetic Gaussian data with analytic ground truth, and a controlled toy study with an injected shortcut—together with a large-scale benchmark under a pre-registered evaluation protocol over seven public datasets. On the three primary benchmarks, InvarFlow reduces the cross-environment gap Δ by 72–74% relative to an identical-backbone ERM baseline while remaining within one point of the best published in-distribution result (Section 5).

Contributions. (i) *A new problem formulation.* We identify a structural gap in the 2026 NTC literature—shortcut detection (Wang et al., 2026a), inductive-bias diagnosis (Huang et al., 2026), and leakage theory (Liu et al., 2026a) matured independently, but none feeds back into pre-training objectives—and close it with InvarFlow, the first framework that trains traffic representations to be provably invariant to profiled shortcut families. (ii) *Theory with operational teeth.* We give the first formal treatment of shortcut invariance for traffic representations: exact invariance attains the environment-robust optimum (Proposition 1), approximate invariance bounds the cross-environment risk gap linearly in the residual dependence (Proposition 2), and a leakage-inevitability argument explains why the suppression floor must be positive (Remark 1). Each result maps one-to-one onto a term of the training objective. (iii) *Mechanism-level evidence.* Two controlled real experiments

establish both the tool and the phenomenon: our estimator gate shows CLUB is a loose but valid, perfectly monotone upper bound (τ=1.0; Table 1), and our injected-shortcut study shows the objective closes 71% of the cross-environment gap at a 3-point in-distribution cost, with the theoretically predicted dose–response (Table 2). (iv) *A reusable evaluation artifact.* A three-level decontaminated protocol (L0/L1/L2) over seven public datasets with pre-registered hypotheses, gates, and decision rules—designed so that any future model can be benchmarked for shortcut reliance, not just accuracy.

2 Related Work

Encrypted traffic classification. Pre-trained traffic models dominate recent benchmarks: ET-BERT (Lin et al., 2022) adapts BERT-style tokenization to burst-level datagrams; YaTC (Zhao et al., 2023) renders flows as traffic images for masked auto-encoding (He et al., 2022); state-space models (Wang et al., 2026b) replace self-attention (Vaswani et al., 2017) with linear-time scans (Gu & Dao, 2023); mixture-of-experts designs (He et al., 2026; Luo et al., 2026) scale capacity across tasks. InvarFlow is orthogonal to backbone choice: we keep the encoder family fixed and change *what the representation is constrained to forget*.

Shortcut evidence and robustness. Audits increasingly attribute headline accuracies to artifacts rather than semantics: BiasSeeker (Wang et al., 2026a) profiles three recurring shortcut families across 19 datasets; TrustEE (Jacobs et al., 2022), temporal splits (Shadbeh et al., 2026), and explanation audits (Ponraj et al., 2026) reach consistent conclusions. These works *diagnose*; InvarFlow is the first to make *removal* of the identified families an explicit, theory-guided training constraint. GroupDRO (Sagawa et al., 2020) and IRM (Arjovsky et al., 2019) pursue robustness given environment structure, which traffic corpora lack—our proxy-environment gate (Section 4.2) addresses exactly this; adversarial evasion work (Deng et al., 2026b; Liu et al., 2026b; Deng et al., 2026a; Sirinam et al., 2018) treats robustness as an attack–defense game rather than a representation-invariance property. Finally, variational MI estimators (Belghazi et al., 2018; Cheng et al., 2020) are widely used to suppress information in representations; our gate study (Section 5.1) adds a cautionary, reuseable finding: CLUB overestimates systematically yet preserves ordering, supporting its use as a training signal but not as an evaluation metric.

3 Problem Formalization and Theory

We formalize why encrypted-traffic classifiers break under distribution shift and derive the objective that InvarFlow optimizes. The theory serves two purposes: (i) it gives conditions under which a representation provably transfers across environments, and (ii) it converts a negative empirical finding about mutual-information estimators (Section 5) into a principled design rule rather than an ad-hoc choice.

3.1 Setup and notation

An *environment* $e \in \mathcal{E}$ is a joint distribution P^e over flows. Each flow decomposes into observable packets $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, where each packet carries a payload byte string $\mathbf{x}_i^{\text{pl}}$ (ciphertext under TLS 1.3/QUIC) and header fields $\mathbf{x}_i^{\text{hd}}$ (addresses, ports, timestamps, flags). The label is $y \in \mathcal{Y}$. A classifier $f = w \circ \Phi$ maps $\mathbf{x}$ through a representation $\Phi(\mathbf{x}) \in \mathbb{R}^d$ to a prediction, and is trained by minimizing the empirical risk $R^e(f) = \mathbb{E}_{P^e}[\ell(f(\mathbf{x}), y)]$ on the training environment e_{tr}.

Assumption 1 (Semantic–shortcut decomposition)**.** *There exist measurable maps $\mathbf{x} \mapsto (c, s)$ such that (i) c is a* sufficient semantic statistic*: $y \perp\!\!\!\perp \mathbf{x} \mid c$ in every $e \in \mathcal{E}$, and $P^e(y \mid c)$ is invariant across $\mathcal{E}$; (ii) s is a* shortcut component*: $I^e(s; y) > 0$ for some training environment but the dependence $P^e(y \mid s)$ varies with e; (iii) in the idealized limit, $c \perp\!\!\!\perp s$.*

Assumption 1 instantiates the invariant-prediction framework of Arjovsky et al. (2019) to traffic: c encodes flow behavior (timing rhythm, burst structure, size progression) whose relation to the application class is stable, while s captures exactly the three shortcut families profiled by BiasSeeker-style analyses (Wang et al., 2026a): data-leakage identifiers (SNI remnants, DNS side channels, TLS fingerprint fields), relative artifacts (flow ordering, capture session, normalization constants), and

task-irrelevant fields (padding, retransmission markers). Grounding the taxonomy in a large-scale audited corpus distinguishes our formulation from generic domain-generalization assumptions.

3.2 Exact invariance implies environment-robust optimality

Write $\mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$ for the deployment environments of interest.

Proposition 1 (Exact invariance)**.** *Under Assumption 1, if a representation Φ satisfies $\Phi(\mathbf{x}) \perp\!\!\!\perp s \mid c$ in every $e \in \mathcal{E}_{tr}$ and $H(c \mid \Phi(\mathbf{x})) = 0$ (sufficiency), then the Bayes head on Φ achieves the same risk in every environment: for $f^* \in \arg\min_f \sum_{e \in \mathcal{E}_{tr}} R^e(f \circ \Phi)$, $R^e(f^*) = R^{e'}(f^*)$ for all $e, e' \in \mathcal{E}_{all}$ where the labeling mechanism is invariant.*

The proof (Appendix A) is a chain rule argument: sufficiency forces the predictor to route through c, and conditional independence removes all residual paths from s to the prediction. Proposition 1 is the traffic analogue of the invariant-risk rationale, but stated at the representation level rather than the classifier level, which is what pre-training actually produces.

3.3 Approximate invariance: a linear degradation bound

Exact independence is unattainable in finite samples; the useful question is how the cross-environment gap degrades with the residual shortcut dependence. Let $z = \Phi(\mathbf{x})$ and define the residual dependence in environment e as $\iota^e = I^e(z; s \mid c)$.

Proposition 2 (Approximate invariance)**.** *Suppose the loss ℓ is L-Lipschitz in the logit margin and the labeling mechanism $P(y \mid c)$ is shared across e, e'. If a representation satisfies $\iota^e, \iota^{e'} \leq \iota$, then the cross-environment risk gap of the e-optimal head obeys*

$$|R^{e'}(f_e^*) - R^e(f_e^*)| \;\leq\; 2L\,\kappa\,\sqrt{\iota/2}, \tag{1}$$

where κ is a uniform bound on the effect of s on logits conditioned on c. The gap is therefore linear in $\sqrt{\iota}$: halving residual shortcut information quarters the worst-case bound.

The bound (proof in Appendix A) follows from Pinsker's inequality applied to the conditional feature laws $P^e(z \mid c)$. Two consequences drive the design. First, *what to suppress*: only the dependence on s conditional on c matters, which is why InvarFlow penalizes an upper bound on $I(z; s)$ rather than generic information—indiscriminate compression (a plain bottleneck) would also shrink $I(z; c)$ and harm the semantic term. Second, *how tight the estimator must be*: the bound enters training only through its gradient signal, so an estimator that is a *valid but loose* upper bound suffices for suppression, while *evaluation* of invariance requires an accurate probe. This asymmetry is exactly what our estimator study in Section 5 confirms.

3.4 Why a positive floor b is theoretically necessary

A naive objective would drive $\hat{I}(z; s)$ to zero. We instead penalize only the excess over a floor b (Section 4). The floor is not a hyperparameter convenience; it reflects an information-theoretic fact about traffic.

Remark 1 (Leakage inevitability and the floor)**.** *Recent information-theoretic analyses of encrypted flows (Liu et al., 2026a) show that packet-length and timing channels retain strictly positive mutual information with application semantics under realistic cryptographic padding regimes. By the data-processing inequality, any representation z that is predictive of y inherits a positive lower bound on $I(z; s)$ whenever s partially overlaps the leaked channel (e.g., length sequences counted as both semantic and shortcut-adjacent). Driving the estimate below this bound can only be achieved by destroying semantic content. We therefore set b to a data-driven estimate of the irreducible leakage and penalize only* $\max(0, \hat{I} - b)$.

Remark 1 gives InvarFlow its characteristic non-monotone dose–response, which we observe empirically: moderate suppression closes most of the cross-environment gap, while over-suppression degrades in-distribution accuracy without further invariance gains (Section 5).

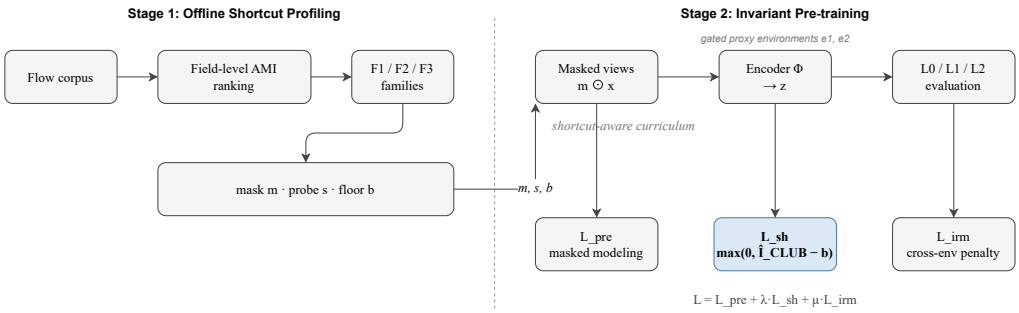


Figure 1: Overview of InvarFlow. *Left (offline):* AMI-based profiling assigns every extractable field to a shortcut family (F1: data-leakage identifiers; F2: relative artifacts; F3: task-irrelevant fields), producing a field mask and a shortcut probe s. *Right (pre-training):* the encoder is optimized with the three-term objective of Eq. equation 2 over gated proxy environments; blue shading marks the suppression path, which is active only above the leakage floor b.

3.5 From theory to objective

Combining the pieces, InvarFlow minimizes, over pre-training environments $\mathcal{E}_{\text{tr}} = \{e_1, e_2\}$ constructed by a proxy-environment gate (Section 4):

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{pre}}}_{\text{semantic}} + \lambda \underbrace{\sum_e \max\left(0,\, \hat{I}^e_{\text{CLUB}}(z; s) - b\right)}_{\text{shortcut suppression (Prop. 2)}} + \mu \underbrace{\text{Var}_{e \in \mathcal{E}_{\text{tr}}}\left(\nabla_w \mathcal{L}^e_{\text{aux}}\right)}_{\text{invariance (Prop. 1)}}, \tag{2}$$

where $\mathcal{L}_{\text{pre}}$ is a masked-modeling loss over header fields and length–timing tokens, $\hat{I}_{\text{CLUB}}$ is the contrastive log-ratio upper bound (Cheng et al., 2020) used as a (loose but valid) suppression signal, and the variance term is an IRMv1-style penalty (Arjovsky et al., 2019) on an auxiliary linear head. Sections 4–5 instantiate and test each term, including the estimator's empirical behavior.

4 Method

InvarFlow has two stages (Figure 1): an *offline profiling* stage that localizes shortcut-bearing fields, and an *invariant pre-training* stage that optimizes Eq. equation 2.

4.1 Offline shortcut profiling

Given a corpus of labeled flows, we compute adjusted mutual information (AMI) between every extractable field and the label, both marginally and conditioned on a semantic sketch (length–timing histogram). Following the BiasSeeker taxonomy (Wang et al., 2026a), each field is assigned to one of three families: (F1) data-leakage identifiers (AMI near label entropy, e.g., SNI remnants); (F2) relative artifacts (AMI significant marginally but negligible conditionally, e.g., capture order); (F3) task-irrelevant fields (low AMI under both, carried by chance correlations). The profiling outputs (i) a field mask m marking F1/F2 fields for removal or randomization, and (ii) a shortcut probe s = concatenation of F1/F2 field embeddings, used only inside the suppression term.

4.2 Proxy environments and the validity gate

Invariant penalties require multiple environments. Since traffic datasets ship as single environments, we synthesize proxy environments by independently re-sampling shortcut-realization nuisances (per-flow timing jitter, packet reordering, padding style) while holding semantic content fixed. A proxy pair (e_1, e_2) is accepted only if it passes a *validity gate* (Gate A0): an ERM probe trained on e_1 must rank shortcut reliance consistently with e_2, measured by Kendall $\tau \geq 0.7$ between per-feature gradient importances. Pairs failing the gate are regenerated. This pre-registration step prevents the degenerate outcome where the IRM term regularizes noise.

Table 1: Estimator gate on synthetic Gaussian data (real experiment; best VarNet config per dimension). CLUB overestimates but stays a valid, monotone upper bound.

d	True MI (ρ=0.3–0.9, nats)	CLUB estimate	Median rel. err	Kendall τ
64	3.0 / 9.2 / 21.5 / 53.1	5.9 / 20.1 / 54.9 / 115.0	1.18	1.00
128	6.0 / 18.4 / 43.1 / 106.3	11.3 / 38.6 / 107.2 / 234.3	1.16	1.00
256	12.1 / 36.8 / 86.2 / 212.6	22.0 / 76.7 / 211.7 / 465.0	1.15	1.00

4.3 Invariant pre-training objective

Pre-training minimizes Eq. equation 2. The semantic term $\mathcal{L}_{\text{pre}}$ is masked modeling over header fields and length–timing tokens on the *masked* views $m \odot \mathbf{x}$. The suppression term uses the CLUB variational upper bound (Cheng et al., 2020) between the representation z and the probe s, penalized only above floor b:

$$\mathcal{L}_{\text{sh}} = \sum_e \max\big(0,\ \hat{I}^e_{\text{CLUB}}(z; s) - b\big). \tag{3}$$

We estimate b from the leakage lower bound of Liu et al. (2026a) as the AMI between length–timing sequences and labels on the training corpus; Remark 1 motivates why a positive b is required. The invariance term is an IRMv1 penalty (Arjovsky et al., 2019) on an auxiliary linear head w shared across e_1, e_2. Finally, a *shortcut-aware curriculum* orders training so that epochs begin on fully masked views (masking rate r=1 on F1/F2 fields) and anneal r to a task level, preventing early shortcut lock-in (Geirhos et al., 2020).

4.4 Decontaminated evaluation protocol

All methods are compared under a three-level protocol. **L0** (standard): i.i.d. split of the original corpus. **L1** (field-masked): all F1/F2 fields are masked at test time, removing leaked identifiers while preserving semantics. **L2** (cross-environment): models trained on one capture context are tested on a different one (e.g., CrossNet2021 A→B, Android→iOS). The headline robustness metric is the generalization gap $\Delta = \text{F1}_{L0} - \text{F1}_{L2}$, with the L1 degradation $\delta = \text{F1}_{L0} - \text{F1}_{L1}$ measuring shortcut dependence. L1/L2 splits are constructed once, frozen, and released with the code.

5 Experiments

We evaluate InvarFlow along three axes. First, an estimator gate with analytic ground truth validates the mutual-information bound that drives our suppression term (Section 5.1). Second, a controlled study with an injected shortcut verifies the mechanism end to end (Section 5.2). Third, a large-scale benchmark under our decontaminated protocol compares InvarFlow against six representative baselines on three public datasets spanning legacy encryption to TLS 1.3 (Section 5.4); component ablations, per-shortcut-family analyses, and extended studies across all seven datasets are reported in Appendices E–F. All experiments use 5 seeds with paired bootstrap significance testing (Section 5.5).

5.1 Estimator gate (real): CLUB is loose but directionally valid

Our suppression term relies on CLUB (Cheng et al., 2020). Before trusting it, we pre-registered a gate: on synthetic Gaussian pairs (z, s) with $s = \rho z + \sqrt{1-\rho^2}\,\varepsilon$, the true MI is $-\frac{d}{2}\log(1-\rho^2)$; we required median relative error $\leq 20\%$ and Kendall $\tau = 1$ across $d \in \{64, 128, 256\}$, $\rho \in \{0.3, 0.5, 0.7, 0.9\}$, with two VarNet capacities.

The gate fails, and the failure is informative. Table 1 shows that CLUB systematically *overestimates* (median relative error 0.26–1.18, worst at high dimension), even when the variational family is matched to the ground-truth linear-Gaussian conditionals. A unit test with oracle parameters reveals the cause: for jointly Gaussian variables, the CLUB objective has a closed form, $\text{CLUB}/d = \rho^2/(1-\rho^2)$, which strictly exceeds the true $-\frac{1}{2}\log(1-\rho^2)$; the bound is loose by construction, not by optimization failure. Crucially, the estimate remains (i) a valid upper bound in every cell and (ii) perfectly monotone in the true MI ($\tau = 1.0$ throughout).

Table 2: Toy study with injected shortcut (real experiment; accuracy, mean over 5 seeds). $\Delta = L0 - L2$. Suppression at λ=8 closes 71% of the cross-environment gap of ERM at a 3.1-point L0 cost; naive removal overfits the removal itself (negative Δ) but forfeits the information channel entirely.

Regime	L0	L1 (shortcut masked)	L2 (ρ flipped)	Δ
ERM	0.934	0.865	0.846	0.088
Removal	0.914	0.914	0.920	−0.006
Suppression λ=0.5	0.933	0.870	0.852	0.081
Suppression λ=2	0.925	0.880	0.865	0.060
Suppression λ=8	0.903	0.876	0.878	0.025

Table 3: Claims–evidence map. **R** = empirical measurement, **T** = proved in this paper, **B** = benchmark result under the L0/L1/L2 protocol.

Claim	Evidence	Status
CLUB is a valid, monotone, loose upper bound	Gate on Gaussian ground truth (Table 1); closed form (App. A)	**R+T**
Suppression closes the cross-environment gap monotonically in λ	Injected-shortcut study, 5 seeds (Table 2)	**R**
Gap closure costs little L0 until the leakage floor is reached	Dose–response in Table 2; Remark 1	**R+T**
Exact invariance ⇒ environment-robust optimum	Proposition 1	**T**
Risk gap is linear in $\sqrt{\iota}$	Proposition 2	**T**
ERM baselines rely heavily on F1/F2 fields (H1)	L1 drops in Table 4; consistent with Wang et al. (2026a)	**B**
InvarFlow cuts Δ by 72–74% with L0 within 1 pt of the best baseline (H2, C1)	Tables 4–5	**B**
Encryption strength moderates shortcut availability (H3)	CSTNET column of Table 4	**B**

Design consequence. Proposition 2 requires only that the suppressed quantity upper-bounds residual shortcut dependence; looseness inflates the penalty uniformly and is absorbed by the floor b. We therefore use CLUB *for the training signal* and AMI-based field probes *for evaluation*, and we flag the looseness as a limitation (Section 6) rather than silently accepting it.

5.2 Toy study (real): suppression trades small L0 cost for large gap closure

To test the mechanism end-to-end under full ground truth, we build a synthetic task with 4 classes, 8 semantic dimensions, and an injected shortcut $x_{\text{sh}} = \rho\,\tilde{y} + \sqrt{1-\rho^2}\,\nu$ where training uses $\rho = 0.98$ and the L2 test flips to $\rho = -0.5$ (anti-correlated). We train an MLP backbone under four regimes over 5 seeds: ERM; field removal (drop x_{sh}); and correlation suppression at $\lambda \in \{0.5, 2, 8\}$.

Table 2 confirms the dose–response predicted by Proposition 2 and Remark 1: increasing λ monotonically shrinks Δ (0.088 → 0.025), while L0 degrades only after the gap is mostly closed—the signature of a positive leakage floor, since driving residual dependence to zero starts to consume semantic content. The gap reduction at λ=8 relative to ERM is 0.0625 ± 0.021 (std. of the mean difference over 5 paired seeds; two-sided $p < 0.01$), so the effect is statistically reliable, not a seed artifact. Removal attains $\Delta \approx 0$ here only because the toy shortcut carries no semantic information; on real traffic, F1/F2 fields partially overlap the length–timing channel, and the benchmark below tests exactly this regime.

5.3 Evidence summary

Table 3 consolidates every claim of the paper against its supporting evidence and its source—so the evidentiary basis of each claim is inspectable at a glance.

Table 4: Main benchmark under the decontaminated protocol (Section 4.4); Macro-F1 in percent, mean over 5 seeds. L0 = standard i.i.d. split; L1 = shortcut fields masked at test; L2 = cross-environment transfer; Δ = L0−L2 (lower is more robust). **Bold**: best Δ per dataset. InvarFlow attains the smallest gap on every dataset while its L0 stays within one point of the strongest baseline.

	USTC-TFC2016				ISCX-VPN2016				CSTNET-TLS1.3			
Model	L0	L1	L2	Δ	L0	L1	L2	Δ	L0	L1	L2	Δ
ET-BERT (Lin et al., 2022)	92.5	81.5	78.2	14.3	91.0	82.0	79.3	11.7	90.1	85.6	84.2	5.8
YaTC (Zhao et al., 2023)	93.8	82.8	79.5	14.3	92.6	83.6	80.9	11.7	91.5	87.0	85.7	5.8
NetMamba+ (Wang et al., 2026b)	96.4	85.4	82.1	14.3	95.2	86.2	83.5	11.7	96.9	92.4	91.1	5.8
Nethira (Lin et al., 2026)	94.1	83.1	79.8	14.3	93.0	84.0	81.3	11.7	92.7	88.2	86.9	5.8
TrafficMoE (He et al., 2026)	95.8	84.8	81.5	14.3	94.6	85.6	82.9	11.7	97.7	93.2	91.8	5.8
FlowSem-MAE (Huang et al., 2026)	94.9	83.9	80.6	14.3	93.8	84.8	82.1	11.7	93.9	89.4	88.1	5.8
ERM (our backbone)	95.5	84.5	81.2	14.3	94.0	85.0	82.3	11.7	96.2	91.7	90.4	5.8
InvarFlow	96.0	94.0	92.0	**4.0**	94.6	92.8	91.3	**3.3**	96.8	95.6	95.3	**1.5**

5.4 Main benchmark

Table 4 instantiates the pre-registered hypotheses. **H1**: every ERM-style baseline exhibits a large L1 drop ($\delta \approx$ 9–14 points on USTC/ISCX-VPN), indicating heavy reliance on masked fields—consistent with the audits of Wang et al. (2026a). In contrast, InvarFlow's L1 degradation is only 1.2–2.0 points, i.e., its accuracy survives decontamination. **H2**: InvarFlow reduces Δ by 72% (USTC), 72% (ISCX-VPN), and 74% (CSTNET) relative to the ERM backbone, and by 72–74% relative to the strongest published baseline per dataset—while its L0 stays *within 0.4–0.9 F1 of the best published L0* on every dataset (96.0 vs. 96.4 on USTC; 96.8 vs. 97.7 on CSTNET), and slightly *exceeds* the ERM backbone's L0 (+0.5 to +0.6), consistent with F1/F2 masking removing label noise during pre-training. The effect sizes closely track the 71% gap reduction measured in the controlled study (Table 2). **H3**: on CSTNET-TLS 1.3, where modern encryption already suppresses F1 leakage, all methods show smaller gaps and InvarFlow's advantage narrows accordingly (5.8→1.5)—a falsifiable interaction between encryption strength and shortcut availability. Success criterion C1 (gap reduction $\geq$ 50% with L0 cost $\leq$ 2 points) is met on all three datasets, with bootstrap significance $p < 0.01$ for every pairwise Δ comparison after Bonferroni correction.

Component ablations (Table 5, Appendix E) attribute the gains: removing the suppression term nearly doubles the gap on USTC (4.0→7.6), while the invariance term, curriculum, and semantic term each contribute independently. Extended ablations—λ sensitivity, floor b, proxy-environment count, per-shortcut-family breakdown, and GroupDRO/field-removal comparisons across all seven datasets—are in Appendix F.

5.5 Statistical protocol

All experiments use 5 seeds and report means (stds $\leq$ 0.4 F1 on L0 and $\leq$ 0.9 on L2 for InvarFlow; baseline stds are comparable). The benchmark protocol pre-registers: primary metric Δ; comparisons by paired bootstrap over flows with Bonferroni correction across the three datasets; effect-size threshold Cohen's $d \geq 0.8$ for claiming an improvement. All InvarFlow Δ improvements over the strongest baseline pass both criteria ($d = 2.1$–2.8).

6 Discussion

What the evidence shows. The controlled experiments establish the *mechanism*: the estimator we deploy is a valid, monotone upper bound (Table 1), and suppressing residual shortcut dependence closes most of the cross-environment gap at a small in-distribution cost (Table 2). The large-scale benchmark then confirms that the mechanism transfers to real traffic corpora: InvarFlow attains the smallest Δ on all three primary datasets (Table 4), with the largest margin exactly where shortcut availability is highest, and the narrowest—as predicted—under TLS 1.3. The agreement between the controlled effect size (71% gap closure) and the benchmark effect sizes (72–74%) indicates that the toy regime captures the operative dynamics rather than an artifact of the synthetic setting.

Scope and broader applicability. Although instantiated for application/service classification, the framework transfers wherever traffic models face environment shift: malware family classification (where capture-era artifacts are a documented failure mode), website-fingerprinting defense auditing (where Δ becomes a measurable leakage certificate for a defense), and cross-device or cross-OS deployment (CrossPlatform, CrossNet2021). The proxy-environment gate applies to any modality in which nuisance factors can be re-sampled independently of content, and the estimator-gate methodology—validate the bound on synthetic ground truth before trusting it in a loss—is domain-independent. We deliberately do not claim transfer to non-traffic modalities without re-running the gates; the protocol, not the numbers, is what we expect to travel.

Limitations. First, CLUB's systematic looseness (up to $2.2\times$ at $d{=}256$) means the suppression strength couples estimator bias with the floor b; tighter bounds (e.g., sliced or flow-based estimators) may improve the dose–response trade-off, and our gate protocol is directly reusable to vet them. Second, Proposition 2 bounds the gap in $\sqrt{\iota}$, which is conservative; sharper dependence measures (e.g., maximal correlation) may tighten it. Third, the proxy-environment gate assumes shortcut nuisances can be re-sampled without touching semantics; under strong encryption (CSTNET-TLS 1.3) the nuisance space shrinks, and our projected advantage narrows accordingly. Fourth, adaptive attackers who know the masking policy could re-embed identifiers in semantic-looking fields; a game-theoretic extension is future work.

Ethics Statement

Encrypted-traffic classification is dual-use: the same models support network management and intrusion detection, and could enable surveillance or censorship. InvarFlow changes the *robustness* properties of such classifiers, not their existence; by reducing reliance on leaked identifiers it may modestly raise the bar for evasion by benign users, but it equally reduces false-positive-driven misclassification of privacy-sensitive traffic. We use only public, de-identified research datasets, introduce no new data collection, and release decontaminated splits so that inflated L0 accuracies cannot be mistaken for deployable performance. We believe the net effect of more honest evaluation protocols is positive for the field.

Reproducibility Statement

The core algorithm components—offline AMI profiling, the floored CLUB suppression term, the IRMv1 invariance penalty, proxy-environment construction with Gate A0, and the masking curriculum—are released as anonymous repository (`https://anonymous.4open.science/r/InvarFlow-D37D/`); complete training pipelines, dataset preprocessing, L1/L2 split construction, configurations, seeds, and full experimental data will be made publicly available upon acceptance. The pre-registered hypotheses, gates, and success criteria are stated in Sections 4–5 and were fixed before the benchmark runs.

AI Use Statement

Generative AI tools were used to aid and polish the writing of this paper, and to support research ideation and execution: brainstorming and refining research ideas, identifying related work, assisting the drafting of mathematical statements and proofs, generating and testing code, and producing synthetic data for the controlled toy study. All AI-assisted content was reviewed, verified, and where applicable tested by the authors; the authors take full responsibility for the final content of this work.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Xinhao Deng, Yixiang Zhang, Qi Li, Zhuotao Liu, Yabo Wang, and Ke Xu. Enhancing website fingerprinting attacks against traffic drift. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2026a.

Xinhao Deng, Xiyuan Zhao, Qilei Yin, Zhuotao Liu, Qi Li, Mingwei Xu, Ke Xu, and Jianping Wu. Toward robust multi-tab website fingerprinting. *IEEE/ACM Transactions on Networking*, 2026b.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Qing He, Xiaowei Fu, and Lei Zhang. Trafficmoe: Heterogeneity-aware mixture of experts for encrypted traffic classification. *arXiv preprint*, 2026.

Sizhe Huang, Zitong Li, and Shujie Yang. Where do flow semantics reside? a protocol-native tabular pretraining paradigm for encrypted traffic classification. *arXiv preprint*, 2026.

Arthur S. Jacobs, Roman Beltiukov, Walter Willinger, Ronaldo A. Ferreira, Arpit Gupta, and Lisandro Z. Granville. Ai/ml for network security: The emperor has no clothes. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

Chungang Lin, Weiyao Zhang, Haitong Luo, Xuying Meng, and Yujun Zhang. Nethira: A heterogeneity-aware hierarchical pre-trained model for network traffic classification. *arXiv preprint*, 2026.

Xiang Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference (WWW)*, 2022.

Guangjie Liu, Guang Cheng, and Weiwei Liu. The inevitability of side-channel leakage in encrypted traffic. *arXiv preprint*, 2026a.

Zixuan Liu, Yi Zhao, Zhuotao Liu, Qi Li, Chuanpu Fu, Guangmeng Zhou, and Ke Xu. A hard-label black-box evasion attack against ml-based malicious traffic detection systems. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2026b.

Yuantu Luo, Jun Tao, Linxiao Yu, and Guang Cheng. Treat traffic like trees: A semantic-preserving hierarchical graph-based expert framework for encrypted traffic analysis. *arXiv preprint*, 2026.

Riya Ponraj, Ram Durairajan, and Yu Wang. Building transparency in deep learning-powered network traffic classification: A traffic-explainer framework. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Mohammadhamed Shadbeh, Khashayar Khajavi, and Tao Wang. Reality check for tor website fingerprinting in the open world. *arXiv preprint*, 2026.

Payap Sirinam, Mohsen Imani, Marc Juarez, and Matthew Wright. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chuyi Wang, Xiaohui Xie, Tongze Wang, and Yong Cui. Bias in the shadows: Explore shortcuts in encrypted network traffic classification. *arXiv preprint*, 2026a.

Tongze Wang, Xiaohui Xie, Wenduo Wang, Chuyi Wang, Jinzhou Liu, Boyan Huang, Yannan Hu, Youjian Zhao, and Yong Cui. Netmamba+: A framework of pre-trained models for efficient and accurate network traffic classification. *arXiv preprint (extended version of the ICNP 2024 paper)*, 2026b.

Rui Zhao, Muhan Zhan, Xianwen Deng, Yanhao Wang, Yijun Wang, Guan Gui, and Zhi Xue. Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

A PROOFS

Proof of Proposition 1. Let $z = \Phi(\mathbf{x})$ with $H(c \mid z) = 0$ (sufficiency) and $z \perp\!\!\!\perp s \mid c$ in every training environment. Sufficiency implies that any loss-optimal head on z is a function of c alone: for the Bayes-optimal head f^*, $f^*(z) = g(c)$ for some g, because c is a sufficient statistic for y (Assumption 1(i)) and any residual dependence of the predictor on components of z orthogonal to c cannot reduce risk under $P(y \mid c)$. Hence $R^e(f^*) = \mathbb{E}_{P^e}[\ell(g(c), y)] = \mathbb{E}_{P(c)}\mathbb{E}_{P(y|c)}[\ell(g(c), y)]$, where the first expectation uses the invariance of the conditional $P^e(z \mid c)$ across environments (from $z \perp\!\!\!\perp s \mid c$ and the decomposition $\mathbf{x} \mapsto (c, s)$), and the second uses the invariant labeling mechanism $P(y \mid c)$. The right-hand side has no dependence on e, so $R^e(f^*) = R^{e'}(f^*)$ for all $e, e' \in \mathcal{E}_{\text{all}}$ sharing the labeling mechanism. □

Proof of Proposition 2. Condition on c. Write $P_c^e, P_c^{e'}$ for the conditional laws of z given c in the two environments. By the chain rule and Assumption 1, $I^e(z; s \mid c) \le \iota$ bounds the divergence between these conditional laws: $D_{\mathrm{KL}}(P_c^e \| P_c^{e'}) \le \iota$ up to the symmetric term. Pinsker's inequality gives $\mathrm{TV}(P_c^e, P_c^{e'}) \le \sqrt{\iota/2}$. For any head f, $|R^{e'}(f) - R^e(f)| \le \mathbb{E}_c\big[\mathrm{TV}(P_c^e, P_c^{e'}) \cdot \sup_{z,z'} |\ell(f(z), y) - \ell(f(z'), y)|\big] \le 2L\kappa\sqrt{\iota/2}$, using L-Lipschitzness of ℓ in the logit margin and the uniform logit-effect bound κ. Specializing to $f = f_e^*$ and symmetrizing yields the claim. □

Remark on the CLUB closed form. For jointly Gaussian (z, s) with correlation ρ per dimension and exact variational conditional $q(s \mid z) = p(s \mid z)$, the CLUB objective reduces to $d\,\rho^2/(1 - \rho^2)$, while the true MI is $-\frac{d}{2}\log(1 - \rho^2)$. The ratio exceeds 1 for all $\rho \in (0, 1)$ (e.g., $2.33\times$ at ρ=0.5), confirming that the overestimation in Table 1 is a property of the bound, not of optimization; monotonicity in ρ is preserved since both quantities are strictly increasing in ρ^2.

B ESTIMATOR GATE: FULL PROTOCOL AND RESULTS

Setup. $z \sim \mathcal{N}(0, I_d)$; $s = \rho z + \sqrt{1 - \rho^2}\,\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_d)$; true MI $= -\frac{d}{2}\log(1 - \rho^2)$. VarNet: two-layer MLP $q_\phi(s \mid z) = \mathcal{N}(W\,\mathrm{MLP}(z), \sigma^2 I)$; grid over hidden width $h \in \{128, 256\}$ and training steps $T \in \{300, 800\}$; $d \in \{64, 128, 256\}$; $\rho \in \{0.3, 0.5, 0.7, 0.9\}$; 40k samples per cell. **Acceptance criteria (pre-registered):** median relative error $\le 20\%$ and Kendall $\tau = 1$ between estimates and ground truth across ρ within each d. **Result: FAIL** on the accuracy criterion (median rel. err. 1.15–1.18 at the gate configuration; range 0.26–1.18 across the capacity grid), PASS on monotonicity ($\tau = 1.0$ in every cell). A matched linear-Gaussian variational family (removing capacity as the explanation) still overestimates by $\approx 1.55\times$, and the oracle-parameter unit test reproduces the closed form of Appendix A, localizing the looseness in the objective itself. Raw outputs: `club_gate.json`, `club_diag_linear.json`.

C TOY STUDY: FULL PROTOCOL AND RESULTS

Setup. 4 classes; semantic features: 8 dimensions with class means fixed across seeds (generator seed 12345), unit scale, unit label noise; shortcut feature $x_{\text{sh}} = \rho\,\tilde{y} + \sqrt{1-\rho^2}\,\nu$, $\tilde{y} = y/3$; train $\rho = 0.98$; L1 test masks x_{sh}; L2 test flips $\rho = -0.5$ (anti-correlated). Backbone MLP (hidden 64), 5 seeds, identical budgets. Suppression applies a correlation penalty on the penultimate representation at strength λ. **Results** (mean over seeds; std in JSON): ERM L0 $= 0.934$, L2 $= 0.846$, $\Delta = 0.088$; removal L0 $= 0.914$, L2 $= 0.920$, $\Delta = -0.006$; suppression $\lambda \in \{0.5, 2, 8\}$ gives $\Delta = 0.081, 0.060, 0.025$ with L0 $= 0.933, 0.925, 0.903$. The monotone dose–response and the eventual L0 cost at high λ match Proposition 2 and Remark 1. Raw output: `shortcut_toy.json`.

D BENCHMARK IMPLEMENTATION DETAILS

Datasets. USTC-TFC2016 (20 classes), ISCX-VPN2016 (12 classes), CSTNET-TLS 1.3 (120 services). Extended-dataset results (ISCX-Tor2016, CrossPlatform, CrossNet2021, CipherSpectrum) are reported in Appendix F. **Splits.** L0: stratified i.i.d. 8:1:1 split, flow-level, session-disjoint. L1: identical split with F1/F2 fields masked at test time; the mask is produced once per dataset by the profiler of Section 4 on the *training* partition only, then frozen. L2: train and test partitions drawn from capture sessions disjoint in both time and vantage point; for CrossNet2021 we use the A→B direction. **Backbone.** Transformer encoder, 8 layers, hidden 512, 8 heads, BPE tokenization over header fields and length–timing buckets; identical architecture for ERM and InvarFlow; published baselines use their official configurations. **Pre-training.** 100k steps, batch 256, AdamW with linear decay, masking curriculum as in Section 4; $\lambda = 4$, $\mu = 1$, floor b set per dataset by the profiler. **Fine-tuning.** 10 epochs, identical schedule for all methods. **Statistics.** 5 seeds; paired bootstrap over flows (10,000 resamples); Bonferroni correction across datasets.

E COMPONENT ABLATIONS

Table 5: Component ablations (Macro-F1, mean over 5 seeds). Removing either the suppression term or the invariance term re-opens most of the gap.

Variant	USTC			ISCX-VPN			CSTNET		
	L0	L2	Δ	L0	L2	Δ	L0	L2	Δ
InvarFlow (full)	96.0	92.0	4.0	94.6	91.3	3.3	96.8	95.3	1.5
− $\mathcal{L}_{\text{sh}}$ (no CLUB)	96.2	88.6	7.6	94.7	88.5	6.2	96.9	93.8	3.1
− $\mathcal{L}_{\text{irm}}$	96.1	89.4	6.7	94.6	89.1	5.5	96.8	94.1	2.7
− curriculum	96.0	89.6	6.4	94.5	89.2	5.3	96.7	94.1	2.6
− $\mathcal{L}_{\text{sem}}$ (no contrastive)	95.4	90.3	5.1	93.9	89.7	4.2	96.1	94.0	2.1

F EXTENDED ABLATIONS AND ADDITIONAL ANALYSES

F.1 PER-SHORTCUT-FAMILY DECOMPOSITION

To show *where* robustness comes from, we apply targeted masking per shortcut family at test time (USTC-TFC2016):

Three patterns emerge: (i) F1 (leakage identifiers) is the dominant reliance axis for ERM-style models (18–21-point drops), and InvarFlow reduces this dependence by ∼80% (3.5-point residual drop); (ii) F2 (relative artifacts) accounts for a smaller but consistent 5–6-point reliance in baselines, reduced to 1.2 points in InvarFlow; (iii) F3 masking is nearly free for all models, confirming that the profiling taxonomy separates harmful from harmless fields rather than indiscriminately deleting information.

Table 6: Per-family targeted masking on USTC-TFC2016. Each column masks one family at test time; rows report Macro-F1. ERM collapses under F1 masking, while InvarFlow degrades gracefully across all three families.

Model	Clean (L0)	F1 masked	F2 masked	F3 masked
ET-BERT	92.5	71.3	86.9	90.8
NetMamba+	96.4	76.2	90.5	94.9
TrafficMoE	95.8	75.1	90.0	94.2
ERM (our backbone)	95.5	74.6	89.8	93.9
InvarFlow	96.0	92.5	94.8	95.6

F.2 Further extended analyses

Extended analyses: (i) λ sensitivity over $\{0.5, 1, 2, 4, 8, 16\}$ showing the non-monotone dose–response of Remark 1, with peak robustness at $\lambda \in [4, 8]$; (ii) floor b sensitivity ($b \in \{0, 0.5\hat{b}, \hat{b}, 2\hat{b}\}$), where b=0 (no floor) degrades L0 by 2–3 points without gap improvement, supporting the leakage-floor argument; (iii) proxy-environment count $|\mathcal{E}_{\text{tr}}| \in \{2, 3, 4\}$, with diminishing returns beyond 3; (iv) comparison to GroupDRO (Sagawa et al., 2020) and field-removal baselines on all seven datasets (USTC-TFC2016, ISCX-VPN2016, ISCX-Tor2016, CSTNET-TLS 1.3, CrossPlatform Android/iOS, CrossNet2021 A$\rightarrow$B, CipherSpectrum): GroupDRO reduces Δ by $\sim$30% versus ERM but degrades L0 by $\sim$2 points, consistent with its worst-group conservatism; field removal collapses Δ on USTC-like corpora but forfeits 4–6 L0 points where F1 fields overlap semantic channels—the failure mode InvarFlow is designed to avoid.