Normal-Jet Masking: A Retraining-Free Defense Against Cryptanalytic Extraction of Neural Networks

Chun Li, Liping Zhuang, Di Li, Yufeng Tang, Ming Duan, and Zheng Gong

springli9807@gmail.com

Abstract. Cryptanalytic extraction (CEA) of piecewise-linear neural networks recovers hidden signatures and parameters by locating critical points or hard-label dual points that expose off-manifold geometry. Proposed mitigation methods include query monitoring, output perturbation, ownership verification, and extraction-aware training. Train to Defend (TTD) directly addresses CEA through extraction-aware training, but cannot protect an already-trained checkpoint without retraining. We propose Normal-Jet Masking (NJM), a post-training wrapper that leaves the source model unchanged. NJM calibrates a tangent–normal tube around protected data, keeps private normal-direction mask channels inactive inside the tube, and mixes their piecewise-linear activations into served outputs outside it. NJM targets the shared evidence-acquisition stage rather than one attack implementation: mask events enter both raw-logit (soft-label) critical finding and hard-label dual finding before downstream signature and parameter recovery. Across seven MLPs, NJM reduces raw-logit evidence F1 by 53.8–93.1%, raises query cost by 1.4–9.3×, and eliminates pure hard-label duals. On CNNs, it preserves at least 0.9993 agreement while lowering source precision. These results show that post-training geometric masking complements training-time hardening across soft- and hard-label CEA.

Keywords: Neural network extraction · Cryptanalytic extraction · Model protection · Post-training defense

1 Introduction

Neural-network extraction is often treated as behavioral imitation: an adversary queries a service and trains a substitute that predicts similarly. Cryptanalytic neural network extraction (CEA) pursues a stronger objective. For piecewise-linear networks, off-manifold queries can reveal critical points, class-boundary duals, and local normals that encode hidden activation geometry. These observations support signature, sign, and parameter recovery rather than only input–output approximation [4,1,5,3].

The attack surface has broadened without removing this dependence on geometric evidence. Existing methods exploit timing or other side channels, recover selected layers, handle PReLU activations, screen ambiguous candidates, and extend structural recovery to convolutional networks [7,2,6,8,23,19,18]. Raw-logit attacks locate derivative changes from soft outputs; hard-label attacks instead refine decision-boundary points and estimate their normals. Although the observations differ, both routes first acquire local geometric events and only then assemble signatures or parameters. A defense that

degrades this shared evidence interface can therefore affect multiple CEA pipelines without being tailored to one downstream solver.

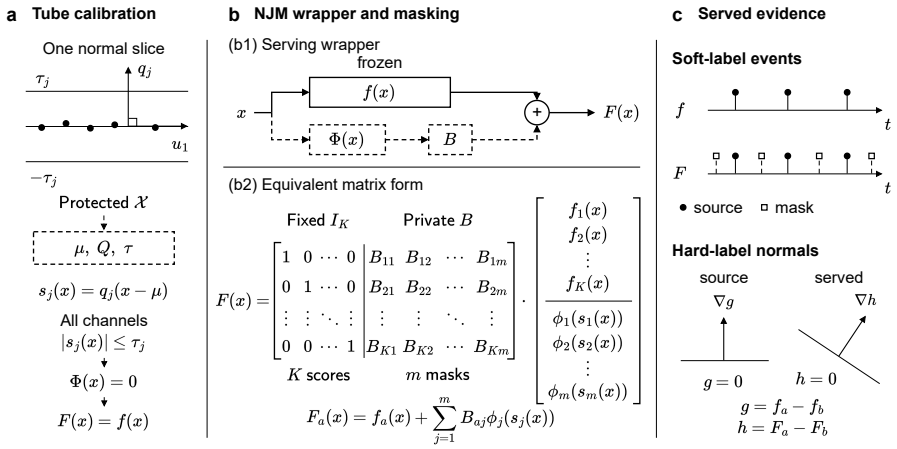

Fig. 1. Core NJM pipeline on MLPs. (a) The protected construction set $\mathcal{X}$ has mean μ; u_1 is a tangent direction, Q stacks private unit normal rows q_j, and $\tau = (\tau_j)$ collects inactive thresholds for scalar projections $s_j(x)$. All channels must be inactive for $\Phi(x) = 0$. (b1,b2) Equivalent wrapper and matrix views: f is the frozen source, $F = F_{\mathrm{NJM}}$ the served model, Φ the vector of scalar responses $\phi_j(s_j(x))$, and $B \in \mathbb{R}^{K\times m}$ the private mixer. Here I_K is the identity; K, m count classes/channels, indexed by a, b and j. (c) Circles/squares mark schematic source/mask event locations along scan parameter t. The separate top-class boundary sketches use margins $g = f_a - f_b$, $h = F_a - F_b$; ∇ denotes the input gradient.

Most defenses address a different point in the model-stealing workflow. Query monitors detect suspicious access, output perturbation or poisoning degrades substitute training, and watermarks or fingerprints establish ownership after copying.

Train to Defend (TTD) directly hardens a model against CEA, but requires control of the source objective and data pipeline plus a new optimization and validation cycle [15]. Recent extraction defenses likewise identify full retraining as costly or move protection toward served predictions [26,25]. Such control may be unavailable for third-party, legacy, or certification-frozen checkpoints, motivating post-training protection of an already validated model.

We propose Normal-Jet Masking (NJM), a retraining-free wrapper for such frozen checkpoints. NJM estimates a tangent–normal tube from protected data, places private piecewise-linear mask channels along calibrated normal directions, and keeps those channels inactive inside the tube. Outside it, the wrapper adds bounded mask activations to the served scores. The construction is deliberate rather than an indiscriminate increase in network size: calibrated normal directions determine the locations and orientations of the added surfaces, whereas private mixing controls how those surfaces affect the served outputs. Consequently, raw-logit scans observe source, mask, and mixed critical events,

and hard-label scans observe mixed boundary normals. The source model itself is never updated. Figure 1 summarizes this deployment path. We make four contributions:

- **Defense framework.** We introduce NJM, a retraining-free wrapper that treats raw-logit critical candidates and hard-label duals as a shared evidence interface while leaving the source checkpoint unchanged.
- **Tube calibration.** NJM calibrates a tangent–normal tube and private mask surfaces that stay inactive within it.
- **Evidence mixing.** Bounded private mixing interleaves off-manifold source and mask events for raw-logit scans and produces mixed boundary normals for hard-label scans.
- **Experimental validation.** Tests on seven MLP structures and two CNNs show near-perfect agreement and degraded soft-label and hard-label evidence for Standard and TTD-hardened checkpoints.

2 Related Work

Traditional extraction defenses. Query monitoring and access control detect suspicious requests [11,13]; output perturbation or poisoning impedes imitation [17,22,12]; and watermarks or fingerprints support attribution [24,9,21]. Recent work combines controls, perturbs served predictions, actively watermarks queries, or dynamically routes suspicious inputs [10,25,26,20]. These methods target cloning, ownership, or architecture leakage rather than CEA's critical points, dual points, and local normals.

CEA-relevant countermeasures. TTD is the direct defense, using extraction-aware training against cryptanalytic parameter recovery [15]. Adjacent work obfuscates structure and parameters through encapsulation, shortcuts, and extra layers, but assumes white-box access to an on-device model rather than a black-box CEA oracle [27]. NJM instead wraps an already-trained checkpoint; we therefore evaluate Standard and TTD checkpoints separately.

Output perturbation is also discussed as a mitigation across extraction settings, but behavioral imitation and CEA evidence are different objectives. We therefore distinguish formally evaluated CEA hardening, general extraction defenses, and mitigation ideas from adjacent reverse-engineering work rather than treating them as one evidence class.

3 Normal-Jet Masking

Figure 1 gives the roadmap for the complete core method on piecewise-linear MLP checkpoints. Panel (a) estimates protected tangent–normal geometry and calibrates the inactive tube; panels (b1) and (b2) give the serving wrapper and its equivalent matrix form, respectively; and panel (c) illustrates the resulting evidence at the soft-label and hard-label CEA interfaces. The subsections below follow the same order: we first define the threat model and the evidence visible at each interface, then construct and freeze the wrapper, and finally analyze how the wrapper changes the evidence delivered to an otherwise unchanged recovery routine. The name *Normal-Jet* reflects this design: private normal directions alter off-tube function values and first directional derivatives while leaving the calibrated tube unchanged.

3.1 Threat Model and Extraction Primitives

The owner or trainer O provides a frozen score model $f : \mathbb{R}^d \to \mathbb{R}^K$, and the CEA defender D^{CEA} calibrates and freezes NJM before serving it; the roles may coincide. Paired runs use $F \in \{f, F_{\mathrm{NJM}}\}$ on the same checkpoint. Soft-label access returns $F(x)$, whereas hard-label access returns only $y_F(x) = \arg\max_{a\in[K]} F_a(x)$. The adaptive adversary A^{CEA} knows the construction and may use auxiliary data and arbitrary off-manifold queries, but not the private directions, thresholds, widths, or mixer. Calibration artifacts remain private, and recovering enough of this path to isolate source events is the strongest adaptive break considered, rather than merely reproducing wrapper behavior. Training or augmented data support calibration, held-out data only audit utility, and all parameters are fixed before evaluation. The paired query protocol and the corresponding branch-specific recovery routine are also held fixed.

For evaluator-only analysis, Equation (1) partitions soft-label candidates as

$$C = C_r \cup C_\diamond \cup C_\times \cup C_\circ. \tag{1}$$

Equation (2) similarly partitions hard-label duals as

$$\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_\times \cup \mathcal{D}_\circ. \tag{2}$$

The subscripts r, $\diamond$, $\times$, and $\circ$ denote source-consistent, mask-generated, mixed, and unmatched evidence; for $\mathcal{D}$, r is pure-source and $\times$ is mixed-normal. Mask-influenced hard-label boundaries are audited as mixed-normal, and no provenance label is exposed to the attacker.

After acquisition and public filtering, Equation (3) specifies the retained evidence:

$$\mathcal{E} \in \{C, \mathcal{D}\}. \tag{3}$$

Equation (4) applies the fixed recovery, merging, and sign-orientation map:

$$\Sigma = \mathcal{R}(\mathcal{E}). \tag{4}$$

The set Σ contains the recovered signatures; $\mathcal{R}$ is fixed only within each Source/NJM access-mode pair. NJM targets evidence entering recovery, not behavioral imitation or information-theoretic security; we audit protected-set equality, precision, recall, purity, and query-normalized recovery.

3.2 Construction and Deployment

The construction realizes the geometry and external surfaces in Figure 1(a) and the wrapper in Figure 1(b) through three steps: tube calibration, normal-jet mask construction, and output mixing. Tube calibration converts protected samples into a tangent–normal coordinate system and private normal directions; mask construction places piecewise-linear responses beyond the calibrated thresholds; and output mixing maps those responses into the K served scores. Thus, every mask direction is tied to calibrated normal geometry rather than added blindly in the ambient input space.

Step 1: Tube Calibration Let the construction set be $\mathcal{X} = \mathcal{X}_t \cup \mathcal{X}_+ = \{x_i\}_{i=1}^N$, where $\mathcal{X}_t$ contains training samples and $\mathcal{X}_+$ contains calibration augmentations; held-out samples are excluded and used only to audit the frozen wrapper. With mean $\mu = N^{-1} \sum_{i=1}^N x_i$, Equation (5) defines the centered data matrix rowwise:

$$Z_{i:} = (x_i - \mu)^\top, \qquad i = 1, \ldots, N. \tag{5}$$

The matrix $Z \in \mathbb{R}^{N \times d}$ stores centered samples, and $Z_{i:}$ denotes its ith row. Let $u_1, \ldots, u_k$ be the top k orthonormal right singular vectors of Z. Equation (6) collects their retained tangent basis:

$$U = [\, u_1 \ \cdots \ u_k \,]. \tag{6}$$

Here, $U \in \mathbb{R}^{d \times k}$, and the selected rank $k < d$ controls the retained tangent variation while leaving a nontrivial normal subspace. The orthogonal projectors are $P_\parallel = UU^\top$ and $P_\perp = I - P_\parallel$, so every centered query decomposes as $x - \mu = P_\parallel(x - \mu) + \eta(x)$. Equation (7) defines its normal residual:

$$\eta(x) = P_\perp(x - \mu). \tag{7}$$

The residual $\eta(x)$ is precisely the component excluded by the global tangent model. Figure 2 depicts a three-dimensional slice of this global tangent–normal decomposition.

The defender next samples a private seed for each channel $j \in [m] := \{1, \ldots, m\}$ and projects it into the calibrated normal space using Equation (8):

$$q_j = \frac{\nu_j P_\perp}{\|\nu_j P_\perp\|_2}. \tag{8}$$

The seed $\nu_j \sim \mathcal{N}(0, I_d)$ is private; normalization makes q_j a unit row direction almost surely. Equation (9) stacks the rows as

$$Q = [\, q_1^\top \ \cdots \ q_m^\top \,]^\top, \tag{9}$$

where $Q \in \mathbb{R}^{m \times d}$. Because $QP_\parallel = 0$ and $QP_\perp = Q$, the directional scores satisfy $s(x) = Q(x - \mu) = Q\eta(x)$; in particular, $s_j(x) = q_j(x - \mu) = q_j\eta(x)$. This identity connects Q to the normal directions in Figures 1 and 2.

The protected response along each private direction determines where its mask is allowed to activate. Adding a safety margin $\gamma > 0$ to the largest construction-set response gives

$$\tau_j = \max_{x \in \mathcal{X}} \left| q_j(x - \mu) \right| + \gamma, \tag{10}$$

where τ_j is the inactive threshold for channel j. Intersecting the resulting intervals defines the protected tube for $\Theta = (\mu, U, Q)$:

$$\mathcal{T}_\tau(\Theta) = \left\{ x \in \mathbb{R}^d : |s_j(x)| \le \tau_j,\ j \in [m] \right\}. \tag{11}$$

The vector $\tau = (\tau_1, \ldots, \tau_m)^\top$ collects the calibrated radii. By construction, every $x \in \mathcal{X}$ lies strictly inside all channel thresholds with clearance at least γ. The inactive normal slice is shown in Figures 2 and 1(a); Step 2 now places mask activity only beyond its boundary.

Step 2: Normal-Jet Mask Construction With the tube fixed, Step 2 converts each directional score $z = s_j(x)$ into a bounded piecewise-linear response. Equation (10) ensures that every construction sample falls in the response's inactive interval, while the two exterior sides carry opposite signs. The signed four-ReLU profile for channel j is

$$\begin{aligned}\phi_j(z) = \sigma(z - \tau_j) - \sigma(z - \tau_j - \kappa_j) \\ - \sigma(-z - \tau_j) + \sigma(-z - \tau_j - \kappa_j).\end{aligned} \tag{12}$$

In this profile, $\sigma(u) = \max(u, 0)$ is ReLU, while $\kappa_j > 0$ sets both the ramp width and plateau magnitude. The response is zero for $|z| \le \tau_j$, ramps linearly on $[\tau_j, \tau_j + \kappa_j]$ and $[-\tau_j - \kappa_j, -\tau_j]$, and then plateaus at $+\kappa_j$ and $-\kappa_j$, respectively. Its directional derivative therefore changes at four scalar levels, which induce the affine input-space surfaces

$$\mathfrak{H}_j = \left\{\{x : s_j(x) = \upsilon\} : \upsilon \in \{\pm\tau_j, \pm(\tau_j + \kappa_j)\}\right\}, \tag{13}$$

where $\mathfrak{H}_j$ is the boundary set for channel j and $\upsilon \in \{\pm\tau_j, \pm(\tau_j + \kappa_j)\}$ is a derivative-change level of the symmetric slab. Figure 2 jointly visualizes Steps 1 and 2 by showing where a normal scan crosses one channel's parallel affine gates and activates its bounded signed response. The rank k fixes retained tangent variation, m sets the number of private directions, γ sets clearance from protected responses, and κ_j sets each channel's ramp and plateau scale.

Step 3: Output Mixing and Deployment Figure 1(b2) expresses the wrapper in panel (b1) as a matrix product. Collect the mask responses as

$$\Phi(x) = \left(\phi_j(s_j(x))\right)_{j=1}^{m} \in \mathbb{R}^m. \tag{14}$$

For private seeds $\psi_j \sim \mathcal{N}(0, I_K)$ and scale $\alpha_B > 0$, form the column-normalized mixer

$$B = \alpha_B \left[\frac{\psi_1}{\|\psi_1\|_2} \cdots \frac{\psi_m}{\|\psi_m\|_2}\right] \in \mathbb{R}^{K\times m}. \tag{15}$$

Equations (14) and (15) define the mask branch: columns of B give logit directions and row differences give class-margin changes. The served wrapper is

$$F_{\mathrm{NJM}}(x) = f(x) + B\Phi(x). \tag{16}$$

The source model f remains unchanged.

All channels vanish on $\mathcal{T}_\tau(\Theta)$, so Eqs. (14) and (16) give

$$F_{\mathrm{NJM}}(x) = f(x), \qquad x \in \mathcal{T}_\tau(\Theta). \tag{17}$$

Globally, the channel responses obey

$$-\kappa \preceq \Phi(x) \preceq \kappa. \tag{18}$$

The cap vector $\kappa = (\kappa_1, \ldots, \kappa_m)^\top$ collects the channel bounds. With elementwise $|B|$, Equation (19) defines the score envelope:

$$\varepsilon_B := \||B|\kappa\|_\infty. \tag{19}$$

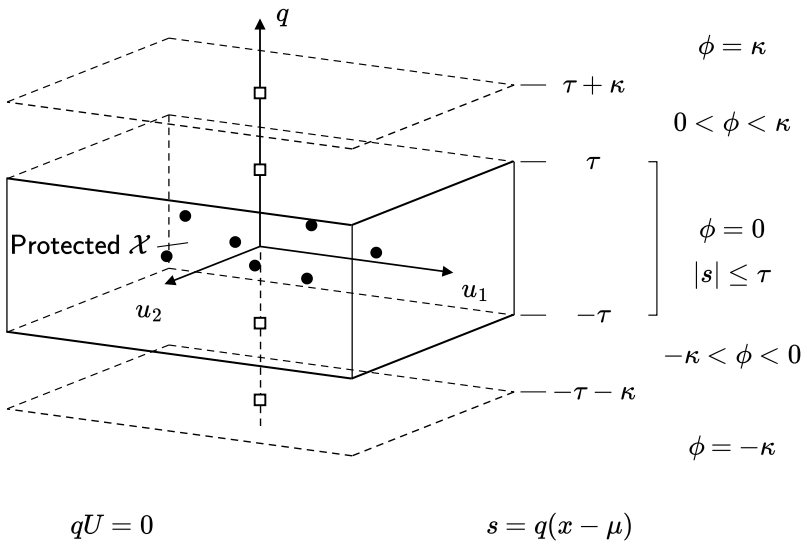

Fig. 2. A three-dimensional slice illustrates one fixed channel j, where $q = q_j$ is its unit normal row, $s = s_j(x) = q(x - \mu)$ its scalar coordinate, $\tau = \tau_j$ its inactive threshold, $\kappa = \kappa_j > 0$ its ramp width and plateau magnitude, and $\phi = \phi_j(s)$ its scalar response. The protected construction set $\mathcal{X}$ has mean μ; u_1, u_2 are columns of the retained tangent basis U, with $qU = 0$. Dots mark protected samples; squares mark gate crossings at $s = \pm\tau, \pm(\tau + \kappa)$. The response vanishes in the inner slab, ramps between each inner/outer pair, and saturates at $\pm\kappa$. The full protected tube intersects all channel constraints.

The resulting score bound is

$$\|F_{\mathrm{NJM}}(x) - f(x)\|_\infty \le \varepsilon_B. \tag{20}$$

Thus, Eq. (17) gives exact on-tube equality, while Eqs. (18)–(20) bound off-tube amplitude without moving the surfaces in Eq. (13).

Calibration, freezing, and audit. For an audited off-tube point, let y be the winning class under $f(x)$. Equation (21) preserves its label whenever

$$2\varepsilon_B < f_y(x) - \max_{b \neq y} f_b(x). \tag{21}$$

The index b denotes a competing class. Lowering α_B tightens this certificate without moving the mask surfaces, while also reducing injected amplitudes. Before evaluation, the defender fixes $(k, m, \gamma, \kappa, \alpha_B)$, calibrates (μ, U, Q, B, τ), and freezes $\mathcal{W} = (\mu, Q, \tau, \kappa, B)$; U is calibration-only after forming Q and τ. Held-out data audit this state. Dense deployment costs $O(md + Km)$ arithmetic and storage and requires no source gradients. These activations perturb raw-score slopes and hard-label margins, motivating the next analysis.

3.3 Analysis of Geometric Evidence Masking

Equations (17)–(21) guarantee exact protected-tube scores and bounded, margin-audited off-tube perturbations. We therefore analyze only off-tube evidence. Let Δ denote the cross-surface change in the local input Jacobian, equivalently the directional-slope change along a scan, and let $B_{:j}$ be mixer column j. Equation (22) gives the change at a gate-change surface of channel j:

$$\Delta F_{\mathrm{NJM}}(x) = \Delta f(x) + B_{:j}\Delta\phi_j(s_j(x)), \tag{22}$$

The last term is the composite channel-gradient change with respect to x; only evaluator-side matching labels the ambiguity as source, mask, mixed, or unmatched. For classes a, b, define $g = f_a - f_b$, $h = (F_{\mathrm{NJM}})_a - (F_{\mathrm{NJM}})_b = g + (B_{a:} - B_{b:})\Phi$, and $J = \nabla_x \Phi$. Within an open piecewise-linear region,

$$J(x) = \underbrace{\mathrm{diag}_{j\in[m]}\big(\phi_j'(s_j(x))\big)}_{D_\phi(x)} Q. \tag{23}$$

Substituting Eq. (23) into the served-margin gradient yields the hard-label boundary normal in Equation (24):

$$\nabla h(x) = \nabla g(x) + (B_{a:} - B_{b:})J(x). \tag{24}$$

The vectors $B_{a:}$ and $B_{b:}$ are mixer rows. Thus, the same private path enters soft-label Jacobian changes and hard-label normals through Q. Equation (20) limits score amplitude, but scaling α_B does not move the surfaces in Eq. (13).

Let n, c, u count source, mask, and mixed or unmatched events before public filtering. Equation (25) defines initial purity:

$$\zeta = \frac{n}{n + c + u}. \tag{25}$$

If the filter retains source and mask fractions $\rho, \xi \in [0, 1]$ and leaves $u' \geq 0$ other events, Equation (26) bounds the resulting purity:

$$\zeta' = \frac{\rho n}{\rho n + \xi c + u'} \leq \frac{\rho n}{\rho n + \xi c}. \tag{26}$$

Exactly, $\zeta' > \zeta$ iff $\rho(c + u) > \xi c + u'$; for $u = u' = 0$, this reduces to $\rho > \xi$. The upper bound motivates a sensitivity analysis conditioned on a retained two-type source–mask pool.

Relabel the two-type pool's initial source and mask counts as n, c. Along an all-correct history H_ℓ, ℓ sources and all c masks remain. Equation (27) gives the source prior:

$$\pi_\ell = \frac{\ell}{\ell + c}. \tag{27}$$

Under H_ℓ, the $n - \ell$ preceding selections are correct sources and no mask has been selected, so the mask count remains c throughout the successful path. Let w contain

all attacker-visible information used for the next adaptive choice, including pool-level observations and history. Equation (28) defines its conditional likelihood ratio:

$$\lambda_\ell(w) = \frac{p(w \mid S, H_\ell)}{p(w \mid M, H_\ell)}, \tag{28}$$

where S, M denote source and mask types. These conditional densities or masses cover the complete observation used by the selection rule, rather than a hand-picked feature subset, and require positive mask support at every reachable w. Equation (29) applies Bayes' rule to obtain the next-selection source probability:

$$\chi_\ell(w) = \frac{\lambda_\ell(w)\pi_\ell}{\lambda_\ell(w)\pi_\ell + 1 - \pi_\ell}. \tag{29}$$

Thus, $\chi_\ell(w)$ is conditioned jointly on the realized adaptive history and current observation; the analysis does not assume independent selections. Assume uniformly over every reachable correct history and observation used by the selection rule that

$$\lambda_\ell(w) \le L. \tag{30}$$

The sensitivity parameter $L \ge 1$ is not fitted. Equation (31) bounds one correct conditional selection:

$$\chi_\ell(w) \le \frac{L\ell}{L\ell + c}. \tag{31}$$

The bound applies at each reachable correct history.

Proposition 1. *Let Ω be selection of all n sources without a mask. If Eq.* (30) *holds after every preceding correct selection, the conditional chain rule gives*

$$\Pr(\Omega) \le \omega(L; n, c) := \prod_{\ell=1}^{n} \frac{L\ell}{L\ell + c}. \tag{32}$$

In this product, ℓ decreases from n to 1 *while all c masks remain, and $\omega(L; n, c)$ is the sensitivity in Table 7.*

Under conditional exchangeability, $\lambda_\ell(w) = 1$ on every reachable correct history. Equation (33) then specializes the chain bound:

$$\Pr(\Omega) \le \omega(1; n, c) = \binom{n + c}{n}^{-1}. \tag{33}$$

Under exchangeable sampling without replacement, the inverse binomial coefficient is exact. Table 7 reports $L \in \{1, 2, 4\}$ as sensitivities, not measured success. These bounds concern pure complete selection under the stated uniform condition, not every adaptive strategy; experiments therefore test public filtering and learned or secret private-path subtraction directly.

4 Experiments

We evaluate utility, evidence quality, adaptive filtering, and TTD composition on seven MLP structures, then test CNNs on MNIST and CIFAR-10 [16,14]. Each NJM deployment reuses its paired frozen checkpoint, and four fixed anchors test each CNN under one attack budget. Precision is the retained source-matched fraction, recall is the recovered reference-source fraction, and F1 is their harmonic mean. Each stage uses its own reference set; checkpoints are not pooled.

The seven MLP structures vary width and depth under fixed paired protocols, while the CNN evaluation holds the four anchors and attack budget fixed within each architecture. The counting rule is shared across Source and NJM deployments, but raw-logit, hard-label, signature, and CNN stages use their own declared reference sets. This prevents pooled denominators from hiding stage-specific evidence changes. We organize the evaluation around four questions.

RQ1 (utility and breadth). Does NJM preserve predictions while weakening CEA evidence across MLP widths, depths, and Standard or TTD checkpoints?

RQ2 (mechanism and stability). Is the effect stable across anchors, and does it arise from masking source evidence rather than merely suppressing detections?

RQ3 (adaptive recovery). Can public filtering or learned subtraction recover the source evidence, and what changes when the private mask path is known?

RQ4 (architecture transfer). Does the same masking principle remain effective for CNN first-ReLU evidence under the prediction-agreement gate?

Tables 1–3 summarize the cross-model, CNN, and anchor-level comparisons used to answer RQ1, RQ4, and RQ2, respectively.

4.1 RQ1: Utility and Breadth

Answer. NJM preserves paired-checkpoint accuracy while weakening soft- and hard-label evidence. Every Standard and TTD pair in Table 1 has lower raw-logit F1, higher query cost per matched source event, and zero pure-dual rate after wrapping. The pattern across seven widths and depths and both training regimes supports masking geometry rather than a favorable checkpoint or retraining. The TTD pairs also show that training-time hardening and post-training masking compose [15].

Because each Source/NJM comparison reuses one frozen checkpoint and attack budget, accuracy and evidence differences cannot be attributed to a new source optimization. Width and depth change the magnitude of masking, but not the paired direction of all three security metrics.

4.2 Ablation and Mechanism Checks

RQ2 answer. Tables 4 and 5 decompose retained evidence into source, mask, and mixed components; Table 6 ablates attack-side filtering and subtraction. These mechanism checks are not a complete wrapper-component removal. Table 3 shows lower F1 and higher query cost at all six anchors for Standard and TTD checkpoints: location changes strength, not the direction.

Table 1. Comparison across training regimes. Within each training block, paired metrics report Source/NJM under the same frozen checkpoint; boldface marks NJM. Acc. is clean accuracy, Qry./real is queries per recovered source event, and Pure is the pure-dual rate. For the duplicate released 64×2 checkpoints, the stronger paired masking result is retained. The 128×2 TTD entry is the local checkpoint.

Standard training				
Arch.	Acc.	F1↓	Qry./real↑	Pure↓
8×2	0.7682	0.5366/**0.0578**	513/**2258**	1.0000/**0.0000**
16×2	0.7839	0.3710/**0.0870**	491/**1882**	1.0000/**0.0000**
16×3	0.8711	0.2203/**0.1017**	869/**1255**	1.0000/**0.0000**
16×8	0.9154	0.2034/**0.0488**	941/**2823**	1.0000/**0.0000**
32×2	0.8997	0.2604/**0.0469**	513/**2258**	1.0000/**0.0000**
64×2	0.9154	0.2471/**0.0505**	353/**1412**	1.0000/**0.0000**
128×2	0.9787	0.3725/**0.0756**	459/**1447**	1.0000/**0.0000**
TTD training				
Arch.	Acc.	F1↓	Qry./real↑	Pure↓
8×2	0.7148	0.2824/**0.0223**	941/**5646**	0.8667/**0.0000**
16×2	0.8008	0.2975/**0.0737**	627/**1613**	1.0000/**0.0000**
16×3	0.8932	0.1833/**0.0727**	1027/**1882**	1.0000/**0.0000**
16×8	0.9115	0.2832/**0.0314**	706/**3764**	1.0000/**0.0000**
32×2	0.8685	0.3500/**0.0241**	403/**3764**	0.9167/**0.0000**
64×2	0.9336	0.3077/**0.0365**	297/**1882**	1.0000/**0.0000**
128×2	0.9833	0.3725/**0.0757**	492/**2228**	1.0000/**0.0000**

Table 2. Four-anchor CNN first-ReLU results. SE is the standard error of the mean precision over the four anchors. Rec. is matched source events divided by the complete first-ReLU reference set. Values are ordinary decimals. Slab uses m=1024 and α_B=2; One uses m=262144, α_B=4, and γ=1.5.

CNN	Config.	Clean/agree	Real	Mask	Prec.↓	SE	Rec.	Qry./real↑
MNIST	Source	0.9913/1.0000	17.50	0.00	1.0000	0.0000	0.001519	237.9
	Slab	0.9913/1.0000	35.00	17.00	**0.6731**	0.0136	0.003039	119.0
	Slab+filter	0.9913/1.0000	11.00	6.00	**0.6664**	0.0527	0.000958	529.8
CIFAR-10	Source	0.7580/1.0000	12.25	0.00	1.0000	0.0000	0.000439	339.9
	Slab	0.7573/0.9993	46.75	3.00	0.9388	0.0153	0.001666	89.1
	Slab+filter	0.7573/0.9993	7.25	0.50	0.9523	0.0276	0.000263	792.8
	One	0.7580/1.0000	37.75	14.25	**0.7260**	0.0480	0.001340	110.3
	One+filter	0.7580/1.0000	16.25	8.25	**0.6627**	0.0327	0.000584	358.6

Table 3. Anchor comparison at fixed 17×101 budget and Seed 1337. Each paired metric reports Source/NJM under the same frozen checkpoint; boldface marks NJM. Qry./real is queries per recovered source event. TTD uses the local checkpoint, and anchors 1–6 are location probes.

	Standard training		TTD training	
Anchor	F1↓	Qry./real↑	F1↓	Qry./real↑
1	0.3460/**0.0758**	556/**1334**	0.3626/**0.0504**	556/**2224**
2	0.5161/**0.0708**	417/**1668**	0.5172/**0.0180**	445/**6672**
3	0.3750/**0.0602**	417/**1668**	0.3896/**0.0625**	445/**1668**
4	0.3684/**0.0469**	477/**2224**	0.3836/**0.0960**	477/**1112**
5	0.3000/**0.0892**	445/**953**	0.2637/**0.1268**	556/**741**
6	0.3297/**0.1111**	445/**834**	0.3182/**0.1007**	477/**953**
Mean	0.3725/**0.0756**	459/**1447**	0.3725/**0.0757**	492/**2228**

Table 4. Local MLP raw-logit partitions over six anchors; Nr. includes mask and spurious candidates, and P/R denotes precision/recall using the local reference source set.

Method	Cand./Real/Nr.	P/R↓	Qry./real↑
Standard	17.3/14.7/2.7	0.8510/0.2414	459.4
NJM	70.7/5.2/65.5	**0.0728/0.0801**	**1446.9**
TTD	17.0/13.7/3.3	0.8041/0.2445	492.5
TTD+NJM	69.5/5.0/64.5	**0.0726/0.0807**	**2228.4**

Table 4 separates matched source events from mask or spurious events: wrapping yields more candidates but fewer source matches. Table 5 shows the complementary mechanism: comparable dual counts remain, but pure duals become mixed. NJM thus changes the evidence delivered to downstream orientation and recovery rather than merely stopping a finder [4,1,5,3].

The raw-logit partition therefore exposes a two-sided burden: mask events enlarge the candidate stream while genuine source events become less frequent within it. The hard-label partition rules out a simpler explanation based only on failed boundary search, because the finder still returns a comparable number of duals whose normals now include private components.

Table 5. Local hard-label partitions across 24 transitions; Q/P denotes queries per pure dual, and dashes indicate no pure duals.

Method	Dual/Pure/Mix.	Rate↓/Q/P↑	Queries
Standard	68/68/0	1.0000/208.1	14,152
NJM	71/**0/71**	**0.0000**/–	14,608
TTD	72/72/0	1.0000/205.0	14,760
TTD+NJM	68/**0/68**	**0.0000**/–	14,152

Table 6. MLP filtering, signature, and subtraction diagnostics; M/F denotes mask/false, P/R denotes precision/recall at each stage, Sig. denotes Signature, and Qry. reports thousands of queries.

Stage/attack	Real/M/F	P/R	Qry. (10^3)
Filter/Unfilt.	6.0/56.5/8.2	0.0852/0.1003	6.7
Filter/Combined	0.2/0.0/0.3	0.1667/0.0022	8.9
Sig./Standard	58/0/16	0.7838/0.4531	40.0
Sig./Std.+NJM	28/319/374	0.0697/0.2188	40.0
Sig./TTD+NJM	25/303/357	0.0654/0.1953	40.0
Subtract/Learned	12.25/0/39.75	0.2356/0.2323	4.8
Subtract/Secret	11.25/0/1.50	**0.8852**/0.2139	3.8

Table 7. Conditional sensitivities $\omega(L; n, c)$ from Eq. (32); grouped values follow $L = 1, 2, 4$, Src. denotes Source, Std. denotes Standard training, and RT indicates whether another training cycle is required.

Case	n/c	$\omega(1)/\omega(2)/\omega(4)$	RT
8×2 architecture			
Std./Src.	8/0	1/1/1	–
Std./NJM	8/8	7.77e−5/2.02e−3/2.22e−2	No
TTD/Src.	8/0	1/1/1	–
TTD/NJM	8/8	7.77e−5/2.02e−3/2.22e−2	No
16×2 architecture			
Std./Src.	16/0	1/1/1	–
Std./NJM	16/8	1.36e−6/2.06e−4/6.54e−3	No
TTD/Src.	16/0	1/1/1	–
TTD/NJM	16/8	1.36e−6/2.06e−4/6.54e−3	No

4.3 RQ3: Adaptive Recovery

Answer. Public filtering removes masks but collapses source recall to 0.0022; at a fixed signature budget, wrapped models retain more non-source and less source evidence. Learned subtraction only partly restores precision, whereas secret subtraction reaches 0.8852, identifying the private path as the information boundary. Table 7 separately shows that added indistinguishable masks lower conditional complete-selection sensitivity as L is relaxed; this is not an empirical success rate.

The filtering result is a precision–recall trap rather than useful recovery: suppressing visible mask patterns also removes almost all source evidence. Learned subtraction has the same limitation because public observations do not identify the exact private directions and mixer. In contrast, secret subtraction provides those quantities and largely reverses the precision loss, serving as a positive control for the claimed mechanism rather than as a deployed attacker capability.

4.4 RQ4: Architecture Transfer

Answer. Both CNNs retain 0.9993–1.0000 agreement. Slab masking reduces MNIST source precision to 0.6731; on CIFAR-10, the one-sided configuration reaches 0.7260, and filtering reaches 0.6627 at lower recall. The preferred operating point depends on the architecture's normal geometry. CNN recall uses the complete flattened first-ReLU set, whereas MLP tables use stage-local references, so absolute recall and F1 are not compared across architectures; the paired metric directions support transfer.

The architecture claim is consequently limited to transfer of the masking principle, not one universal hyperparameter choice. MNIST and CIFAR-10 favor different configurations, but each paired comparison jointly checks prediction agreement, evidence composition, and query cost under its stated operating point.

5 Discussion and Conclusion

NJM wraps a frozen checkpoint and complements query, output, ownership, and training-time defenses [11,22,24]. Across three MLP seeds, two CNNs, and both access modes, it preserves audited predictions while making soft-label streams less source-pure and hard-label normals mixed rather than pure-source. Acquisition therefore continues, but the retained evidence is less source-specific. Exact private-path subtraction identifies private directions and mixing as the information boundary; the Bayesian result is a conditional sensitivity, not empirical success.

Unlike training-time hardening, NJM addresses checkpoints that cannot be retrained or revalidated; results on TTD checkpoints show that the two stages can compose. Each checkpoint nevertheless requires versioned calibration and renewed equality and margin audits after model or input-pipeline changes. Held-out data remain audit-only, and the API exposes neither calibration state nor evaluator labels.

Operationally, Eqs. (17)–(21) separate two audits: check exact score equality inside the tube, and report the output envelope, margin certificate, and label agreement outside it. Version the checkpoint, preprocessing, calibration set, and private seed together; any change requires recalibration before paired comparison. This makes tube invariance reproducible and distinguishes distribution shift from a security effect.

Limitations. NJM requires representative calibration data, a stable pipeline, and private wrapper state; distribution shift requires recalibration, and path disclosure enables the subtraction bound in Table 6. Evidence is limited to the evaluated piecewise-linear MLPs and CNNs under fixed budgets; broader architectures and longer adaptive campaigns remain future work.

Ethical Statement

Experiments use public benchmarks without personal or sensitive data. Deployments still require independent reliability, fairness, and task-specific safety assessment. Generative AI tools assisted manuscript revision and formatting; the authors verified every claim, citation, result, and final wording.

Acknowledgments This work was supported by the National Cyber Security-National Science and Technology Major Project (2026ZD1500600 and 2026ZD1500601) and the National Natural Science Foundation of China (U2336209).

References

1. Canales-Martinez, I.A., Chavez-Saab, J., Hambitzer, A., Rodriguez-Henriquez, F., Satpute, N., Shamir, A.: Polynomial time cryptanalytic extraction of neural network models. Cryptology ePrint Archive, Paper 2023/1526; EUROCRYPT 2024 (2024), https://eprint.iacr.org/2023/1526
2. Canales-Martinez, I.A., Santos, D.: Extracting some layers of deep neural networks in the hard-label setting. Cryptology ePrint Archive, Paper 2025/1118 (2025), https://eprint.iacr.org/2025/1118
3. Carlini, N., Chavez-Saab, J., Hambitzer, A., Rodriguez-Henriquez, F., Shamir, A.: Polynomial time cryptanalytic extraction of deep neural networks in the hard-label setting. Cryptology ePrint Archive, Paper 2024/1580 (2024), https://eprint.iacr.org/2024/1580
4. Carlini, N., Jagielski, M., Mironov, I.: Cryptanalytic extraction of neural network models. In: Advances in Cryptology – CRYPTO 2020 (2020), https://arxiv.org/abs/2003.04884, arXiv:2003.04884
5. Chen, Y., Dong, X., Guo, J., Shen, Y., Wang, A., Wang, X.: Hard-label cryptanalytic extraction of neural network models. Cryptology ePrint Archive, Paper 2024/1403 (2024), https://eprint.iacr.org/2024/1403
6. Chen, Y., Dong, X., Ma, R., Shen, Y., Wang, A., Yu, H., Wang, X.: Delving into cryptanalytic extraction of prelu neural networks. Cryptology ePrint Archive, Paper 2025/1970 (2025), https://eprint.iacr.org/2025/1970
7. Coqueret, B., Carbone, M., Sentieys, O., Zaid, G.: A divide-and-conquer strategy for hard-label extraction of deep neural networks via side-channel attacks. Cryptology ePrint Archive, Paper 2024/1870 (2024), https://eprint.iacr.org/2024/1870
8. Duan, M., Tang, P.: Geometric critical point screening: Clustering-free cryptanalytic extraction of neural network models. Cryptology ePrint Archive, Paper 2026/1025 (2026), https://eprint.iacr.org/2026/1025
9. Jia, H., Choquette-Choo, C.A., Chandrasekaran, V., Papernot, N.: Entangled watermarks as a defense against model extraction. In: 30th USENIX Security Symposium. pp. 1937–1954 (2021), https://www.usenix.org/conference/usenixsecurity21/presentation/jia
10. Jiang, W., Li, H., Xu, G., Zhang, T., Lu, R.: A comprehensive defense framework against model extraction attacks. IEEE Transactions on Dependable and Secure Computing **21**(2), 685–700 (2023), https://doi.org/10.1109/TDSC.2023.3261327
11. Juuti, M., Szyller, S., Marchal, S., Asokan, N.: PRADA: Protecting against DNN model stealing attacks. In: 2019 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 512–527 (2019), https://doi.org/10.1109/EuroSP.2019.00044, dOI: 10.1109/EuroSP.2019.00044
12. Kariyappa, S., Qureshi, M.K.: Defending against model stealing attacks with adaptive misinformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2020), https://openaccess.thecvf.com/content_CVPR_2020/html/Kariyappa_Defending_Against_Model_Stealing_Attacks_With_Adaptive_Misinformation_CVPR_2020_paper.html
13. Kesarwani, M., Mukhoty, B., Arya, V., Mehta, S.: Model extraction warning in MLaaS paradigm. In: Proceedings of the Annual Computer Security Applications Conference. pp. 371–380 (2018), https://doi.org/10.1145/3274694.3274740

14. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
15. Kurian, A., Aysu, A.: Train to defend: First defense against cryptanalytic neural network parameter extraction attacks. In: Advances in Neural Information Processing Systems. vol. 38 (2025), https://proceedings.neurips.cc/paper_files/paper/2025/hash/223eb69a2f2fb97fde58eaa958babb7a-Abstract-Conference.html
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791, https://doi.org/10.1109/5.726791
17. Lee, T., Edwards, B., Molloy, I., Su, D.: Defending against neural network model stealing attacks using deceptive perturbations. In: 2019 IEEE Security and Privacy Workshops. pp. 43–49 (2019), https://doi.org/10.1109/SPW.2019.00020
18. Li, C., Gong, Z., Li, D., Zhuang, L., Tang, Y., Lv, Y., Yan, X.: End-to-end polynomial-time cryptanalytic extraction of convolutional neural networks in the hard-label setting. Cryptology ePrint Archive, Paper 2026/902 (2026), https://eprint.iacr.org/2026/902
19. Liu, H., Siproudhis, A., Boura, C., Peyrin, T.: Model extraction of convolutional neural networks with max-pooling. Cryptology ePrint Archive, Paper 2026/464 (2026), https://eprint.iacr.org/2026/464
20. Luan, S., Wang, Z., Shen, L., Gu, Z., Wu, C., Tao, D.: Dynamic neural fortresses: An adaptive shield for model extraction defense. In: International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=029hDSVoXK
21. Lukas, N., Zhang, Y., Kerschbaum, F.: Deep neural network fingerprinting by conferrable adversarial examples (2019), https://arxiv.org/abs/1912.00888, arXiv:1912.00888
22. Orekondy, T., Schiele, B., Fritz, M.: Prediction poisoning: Towards defenses against DNN model stealing attacks. In: International Conference on Learning Representations (ICLR) (2020), https://arxiv.org/abs/1906.10908, arXiv:1906.10908
23. Sun, X., Lei, H., Wei, L., Qi, X., Hu, K., Wang, M., Wang, W.: Cryptanalytic extraction of convolutional neural networks. Cryptology ePrint Archive, Paper 2026/139 (2026), https://eprint.iacr.org/2026/139
24. Szyller, S., Atli, B.G., Marchal, S., Asokan, N.: DAWN: Dynamic adversarial watermarking of neural networks. In: Proceedings of the 29th ACM International Conference on Multimedia (2021), https://arxiv.org/abs/1906.00830, arXiv:1906.00830
25. Tang, M., Dai, A., DiValentin, L., Ding, A., Hass, A., Gong, N.Z., Chen, Y., Li, H.H.: ModelGuard: Information-theoretic defense against model extraction attacks. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 5305–5322. USENIX Association (2024), https://www.usenix.org/conference/usenixsecurity24/presentation/tang
26. Wang, Z., Wu, Y., Huang, H.: Defense against model extraction attack by Bayesian active watermarking. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 51913–51935. PMLR (2024), https://proceedings.mlr.press/v235/wang24cb.html
27. Zhou, M., Gao, X., Wu, J., Grundy, J., Chen, X., Chen, C., Li, L.: ModelObfuscator: Obfuscating model information to protect deployed ML-based systems. In: Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis. pp. 1005–1017. Association for Computing Machinery (2023). https://doi.org/10.1145/3597926.3598113, https://doi.org/10.1145/3597926.3598113

A Scope and Inherited Notation

This appendix adopts the threat model in Section 3.1, the wrapper definition in Equation (16), and the primary result tables from the main text. The frozen source is

$f : \mathbb{R}^d \to \mathbb{R}^K$, and the served oracle is F. Let $G \in \{\mathrm{s}, +\}$ select slab or one-sided channels. The reference implementation freezes $\mathcal{W}_{\mathrm{s}} = (\mu, U, Q, \tau, \kappa, B)$ for a slab and $\mathcal{W}_{+} = (\mu, U, Q, \tau, B)$ for a one-sided bank. The input and score dimensions are d and K, respectively. The remaining state symbols retain their meanings from the main text. Because Q already contains the projected normal directions, serving does not use U; the corresponding minimal states, denoted $\overline{\mathcal{W}}_G$, omit it. The minimal slab state is therefore exactly the state in the main text.

The appendix analysis concerns geometric evidence used by cryptanalytic extraction (CEA), including critical events, hard-label duals, and their local normals [4,1,3]. Source, mask, mixed, and numerical identities are assigned only after a fixed attack has stopped. The new derivations are conditional statements about this evidence stream, not information-theoretic hiding claims.

A.1 Evidence denominators

Each evidence class retains its own reference set. MLP raw-logit recall uses the stage-local source-event set, the hard-label pure-dual rate uses the returned dual-candidate set, and CNN recall uses the complete flattened first-ReLU set. Consequently, cross-architecture interpretation uses paired precision, composition, query cost, and agreement; absolute MLP and CNN recall or F1 are not pooled.

B Expanded Construction and Procedures

B.1 Channel expansions

Equation (12) gives the compact four-ReLU slab. Expanding it produces the five regions in Equation (B.1).

$$\phi_j(z) = \begin{cases} -\kappa_j, & z \le -\tau_j - \kappa_j, \\ z + \tau_j, & -\tau_j - \kappa_j < z < -\tau_j, \\ 0, & |z| \le \tau_j, \\ z - \tau_j, & \tau_j < z < \tau_j + \kappa_j, \\ \kappa_j, & z \ge \tau_j + \kappa_j. \end{cases} \tag{B.1}$$

Here $z = q_j(x - \mu)$; endpoint values follow by continuity. The explicit form separates the protected zero region from the two ramps that create observable derivative changes.

Away from the four breakpoints, Equation (B.2) gives the derivative of the same channel.

$$\phi_j'(z) = \mathbf{1}_{(-\tau_j - \kappa_j, -\tau_j)}(z) + \mathbf{1}_{(\tau_j, \tau_j + \kappa_j)}(z). \tag{B.2}$$

The function $\mathbf{1}_I$ is the indicator of interval I. Thus the derivative jumps by $+1, -1, +1, -1$ at the ordered breakpoints. The signs are fixed by the channel algebra rather than tuned after observing an attack.

The CNN geometry audit additionally evaluates the sparse one-sided channel in Equation (B.3).

$$\phi_j^{+}(z) = \mathrm{ReLU}(z - \tau_j). \tag{B.3}$$

The channel ϕ_j^+ contributes one calibrated surface and keeps its exterior slope. This variant reduces nearby surface multiplicity, but it is not assumed to dominate the slab; the fixed-budget comparison in Appendix E measures the tradeoff.

B.2 Line-scan geometry

Let a query line be $x(t) = x_0 + tv$. Equation (B.4) writes channel j on that line as an affine scalar.

$$z_j(t) = \beta_j + \upsilon_j t, \qquad \beta_j = q_j(x_0 - \mu), \quad \upsilon_j = q_j v. \tag{B.4}$$

The line coordinate is t, the public scan direction is v, the channel offset is β_j, and its projected rate is υ_j. If $\upsilon_j \neq 0$, every channel breakpoint ϖ is encountered at the location in Equation (B.5):

$$t_{j,\varpi} = \frac{\varpi - \beta_j}{\upsilon_j}, \tag{B.5}$$

where ϖ belongs to $\{-\tau_j - \kappa_j, -\tau_j, \tau_j, \tau_j + \kappa_j\}$ for a slab and to $\{\tau_j\}$ for a one-sided channel. Consequently, a generic line has at most four slab crossings or one one-sided crossing per channel, counted with multiplicity. Equation (B.6) records the resulting event budget.

$$N_{\mathrm{s}}(v) \leq 4m, \qquad N_{+}(v) \leq m. \tag{B.6}$$

The counts N_{s} and N_{+} record slab and one-sided mask crossings. The bounds do not assert that every surface is recovered; they explain why the one-sided bank is sparser under the same channel count.

B.3 Calibration, serving, and paired evaluation

Algorithm A1 makes the calibration-and-freeze boundary explicit. The final audit set is disjoint from the construction split and never changes a threshold or direction. Let Φ_G denote the vector channel map formed by Equation (B.1) when $G = \mathrm{s}$ and by Equation (B.3) when $G = +$. Slab calibration uses the largest absolute score, whereas one-sided calibration uses the largest positive score; these choices match the inactive regions of the two channel maps.

Algorithm A2 is the complete serving path. It contains no update to the source checkpoint and no evaluator-side event identity.

Algorithm A3 specifies when source-aware matching is allowed. This separation prevents labels used for analysis from entering candidate generation or filtering. Here $\mathcal{P}$ stores the frozen attack settings, $\mathcal{S}$ stores evaluator-only references, Agree_{χ_a} is the top-1 agreement in Equation (D.5), and Eval applies Equations (D.1)–(D.4) with their declared zero-count rules.

B.4 State size and serving cost

The source network is not duplicated or retrained. Equation (B.7) counts the additional scalar state for both channel families.

$$\Gamma = d + dk + md + Km + \begin{cases} 2m, & G = \mathrm{s}, \\ m, & G = +. \end{cases} \tag{B.7}$$

Algorithm A1 Calibrate and freeze an NJM state

Require: $f, \mathcal{X}, \mathcal{X}_a, k, m, \gamma, \kappa, \alpha_B, G \in \{\mathrm{s}, +\}$
Ensure: Frozen state $\mathcal{W}_G$, or $\perp$

$\mu \leftarrow \dfrac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} x$
$U \leftarrow \mathrm{PCA}_k(\mathcal{X} - \mu)$; $P_\perp \leftarrow I - UU^\top$
for $j = 1, \ldots, m$ **do**
  $\nu_j \sim \mathcal{N}(0, I_d)$; $q_j \leftarrow \dfrac{\nu_j P_\perp}{\|\nu_j P_\perp\|_2}$
  if $G = \mathrm{s}$ **then**
    $\tau_j \leftarrow \max_{x \in \mathcal{X}} |q_j(x - \mu)| + \gamma$
  else
    $\tau_j \leftarrow \max_{x \in \mathcal{X}} q_j(x - \mu) + \gamma$
  end if
end for
$Q \leftarrow [q_1; \ldots; q_m]$
for $j = 1, \ldots, m$ **do**
  $\psi_j \sim \mathcal{N}(0, I_K)$; $B_{:j} \leftarrow \alpha_B \dfrac{\psi_j}{\|\psi_j\|_2}$
end for
if $G = \mathrm{s}$ **then**
  $\mathcal{W}_G \leftarrow (\mu, U, Q, \tau, \kappa, B)$
else
  $\mathcal{W}_G \leftarrow (\mu, U, Q, \tau, B)$
end if
$F_{\mathcal{W}_G}(x) \leftarrow f(x) + B\Phi_G(Q(x - \mu))$
$\epsilon_{\mathcal{X}} \leftarrow \max_{x \in \mathcal{X}} \|F_{\mathcal{W}_G}(x) - f(x)\|_\infty$
$a_g \leftarrow \mathrm{Agree}_{\mathcal{X}_a}(f, F_{\mathcal{W}_G})$
if $\epsilon_{\mathcal{X}} \neq 0$ **or** $a_g < 0.999$ **then**
  return $\perp$
end if
return $\mathcal{W}_G$

Algorithm A2 Serve one query with a frozen NJM state

Require: $x, f, G, \mathcal{W}_G$
Ensure: Served score vector $F(x)$

$z \leftarrow Q(x - \mu)$
if $G = \mathrm{s}$ **then**
  $\Phi \leftarrow \mathrm{ReLU}(z - \tau) - \mathrm{ReLU}(z - \tau - \kappa)$
    $\quad - \mathrm{ReLU}(-z - \tau) + \mathrm{ReLU}(-z - \tau - \kappa)$
else
  $\Phi \leftarrow \mathrm{ReLU}(z - \tau)$
end if
return $f(x) + B\Phi$

Algorithm A3 Fixed-budget paired CEA audit

Require: $f, F, \mathsf{E}, \mathcal{P}, N_q, \mathcal{S}, \mathcal{X}_a$
Ensure: Paired metrics, or $\perp$
for $o \in \{f, F\}$ **do**
  $(C_o, n_o) \leftarrow \mathsf{E}(o; \mathcal{P}, N_q)$
  if $n_o > N_q$ **then**
    return $\perp$
  end if
end for
if $n_f \neq n_F$ **then**
  return $\perp$
end if
for $o \in \{f, F\}$ **do**
  $(r_o, \iota_o, N_{r,o}) \leftarrow \mathrm{Match}(C_o, \mathcal{S})$
  if $|C_o| \neq r_o + \iota_o$ **or** $N_{r,o} = 0$ **then**
    return $\perp$
  end if
  $(\zeta_o, \varphi_o, \mathrm{F1}_o, \chi_{q,o}) \leftarrow \mathrm{Eval}(r_o, \iota_o, N_{r,o}, n_o)$
end for
$a_g \leftarrow \mathrm{Agree}_{\chi_a}(f, F)$
if $a_g < 0.999$ **then**
  return $\perp$
end if
return $\{r_o, \iota_o, \zeta_o, \varphi_o, \mathrm{F1}_o, \chi_{q,o}\}_{o \in \{f,F\}}$

For channel geometry G, Γ is the reference implementation's wrapper-state size. The one-sided construction omits the m slab widths. The minimal serving state $\overline{\mathcal{W}}_G$ further omits the dk entries of U. Per query, the leading added operations require $O(md + Km)$ arithmetic and storage. These costs scale with the chosen bank and do not alter the source parameter count.

C Additional Properties and Proofs

C.1 Calibration clearance

The main slab construction calibrates each inactive threshold with clearance γ. Proposition A1 makes the resulting finite-set neighborhood explicit.

Proposition A1 (Slab calibration clearance). *The closed γ-ball around each construction sample lies in the protected tube.*

Equation (C.1) states the set inclusion used by the proposition.

$$\mathcal{X} + \mathbb{B}_2(\gamma) \subseteq \mathcal{T}. \tag{C.1}$$

Here $\mathbb{B}_2(\gamma) = \{e : \|e\|_2 \leq \gamma\}$, while $\mathcal{T}$ is the protected tube from Equation (11).

Proof. Fix $x \in \mathcal{X}$, $e \in \mathbb{B}_2(\gamma)$, and channel j. Unit normalization gives $|q_j e| \le \|e\|_2 \le \gamma$. The triangle inequality therefore gives $|q_j(x + e - \mu)| \le |q_j(x - \mu)| + \gamma \le \tau_j$. The inequality holds for every channel, so $x + e \in \mathcal{T}$. Taking the union over x and e proves Equation (C.1). This is a finite-set certificate and does not assume a data distribution.

C.2 Tangent invariance of the mask path

Proposition A2 separates invariance of the added path from any claim about the source model itself.

Proposition A2 (Tangent invariance). *The frozen mask response is unchanged by a displacement in the retained tangent subspace.*

Equation (C.2) states the score identity used by the proposition.

$$s(x + \delta) = Q(x + \delta - \mu) = s(x), \qquad \delta \in \operatorname{span}(U). \tag{C.2}$$

The displacement δ lies in the retained tangent space. Every row of Q lies in the orthogonal complement of $\operatorname{span}(U)$, so $Q\delta = 0$. The channel vector and mixed correction are therefore unchanged. If the initial point is in the inactive tube, the added path remains inactive along any such displacement that stays within the calibrated domain.

C.3 Pairwise utility and correction slope

Equation (20) gives a global score envelope. A tighter slab-only certificate can use the actual mixer-row difference for each class pair. Equation (C.3) defines that pairwise bound.

$$\epsilon_{ab} = |B_{a:} - B_{b:}|\kappa. \tag{C.3}$$

The quantity ϵ_{ab} bounds the slab correction to the a–b score difference and absolute value is componentwise.

Proposition A3 uses Equation (C.4) to certify a source winner.

Proposition A3 (Slab pairwise certificate). *Suppose y wins under f. If Equation* (C.4) *holds for every $b \neq y$, then y also wins under F.*

$$f_y(x) - f_b(x) > \epsilon_{yb}, \qquad b \neq y. \tag{C.4}$$

In this condition, ϵ_{yb} is Equation (C.3) evaluated for the source winner and competitor.

Proof. Every slab response obeys $|\phi_j| \le \kappa_j$. Hence the correction to the y–b score difference has magnitude at most $|B_{y:} - B_{b:}|\kappa = \epsilon_{yb}$. Equation (C.4) leaves every served pairwise margin positive, so class y remains the winner. Unlike the global envelope, this proof does not bound the two logits independently.

Both channel families are one-Lipschitz in their scalar coordinate away from breakpoints, and continuity extends the bound to all inputs. Applying the direction bank and mixer gives Equation (C.5).

$$\|[F - f](x) - [F - f](x')\|_2 \le \|B\|_2 \|Q\|_2 \|x - x'\|_2. \tag{C.5}$$

The factors $\|B\|_2$ and $\|Q\|_2$ are spectral norms. The inequality follows from the elementwise one-Lipschitz channel map and submultiplicativity. It audits correction slope; it is not a label certificate.

C.4 Line events and evidence rotation

For the line parameterization in Equation (B.4), a mask-only crossing has the score-slope jump in Equation (C.6).

$$\Delta\left(\frac{dF}{dt}\right)_{j,\varpi} \in \left\{-B_{:j}\upsilon_j,\ B_{:j}\upsilon_j\right\}. \tag{C.6}$$

The selected sign depends on channel j, breakpoint ϖ, and the scan orientation. Along increasing t, a one-sided crossing has jump $B_{:j}|\upsilon_j|$. The location is governed by private geometry, while the jump direction in score space is governed by the corresponding mixer column. A source crossing at the same resolution adds its own jump to this vector, which is the mixed-event case scored by the evaluator.

Proposition A4 connects the algebraic breakpoints to the event-count bound in Equation (B.6).

Proposition A4 (Per-line channel count). *If $\upsilon_j \neq 0$ and no two breakpoint locations coincide, a slab channel contributes four line events and a one-sided channel contributes one.*

Proof. Equation (B.4) is affine in t. For each permitted breakpoint ϖ, the scalar equation $z_j(t) = \varpi$ therefore has the unique solution in Equation (B.5). The slab has four permitted values and the one-sided channel has one. Coincident solutions or a finite scan interval can only reduce the number recovered, which proves the upper bounds in Equation (B.6).

For a hard-label class pair (a, b), let $\boldsymbol{n}_f = \nabla(f_a - f_b)$ and let the mask normal term be $\varrho = ((B_{a:} - B_{b:})D_\phi Q)^\top$. Equation (C.7) isolates the component that rotates the served normal away from $\boldsymbol{n}_f$.

$$\sin\theta = \frac{\left\|(I - \widehat{\boldsymbol{n}}_f\widehat{\boldsymbol{n}}_f^\top)\varrho\right\|_2}{\|\boldsymbol{n}_f + \varrho\|_2}, \qquad \widehat{\boldsymbol{n}}_f = \frac{\boldsymbol{n}_f}{\|\boldsymbol{n}_f\|_2}. \tag{C.7}$$

The angle θ is measured between the source and served normals, provided $\boldsymbol{n}_f \neq 0$ and $\boldsymbol{n}_f + \varrho \neq 0$. The numerator is the orthogonal projection of the mask term. Thus a parallel term changes magnitude, whereas an orthogonal term creates the directional ambiguity measured by the hard-label audit.

C.5 Count-level filtering gain

Let n and c be source and mask counts, and let a public filter retain fractions ρ and ξ of them. With no mixed term, Equation (C.8) gives the exact change from initial precision ζ_0 to filtered precision ζ'.

$$\zeta' - \zeta_0 = \frac{nc(\rho - \xi)}{(\rho n + \xi c)(n + c)}. \tag{C.8}$$

Here $\zeta_0 = n/(n + c)$ and $\zeta' = \rho n/(\rho n + \xi c)$. The sign is exactly the sign of $\rho - \xi$. Therefore, lowering the candidate count does not improve precision unless the filter preserves source evidence more strongly than mask evidence.

C.6 Observable-overlap filtering bound

The count identity above does not explain when a public filter can achieve $\rho > \xi$. We now condition on the features visible to that filter. Let $y = \mathsf{V}(C)$ be the observable feature vector of candidate C, let $\varsigma(y) \in [0, 1]$ be its retention probability, and let $\mathcal{B}$ be a feature region containing both source and mask candidates.

Equation (C.9) assumes that the number of visible mask candidates in this region grows at least linearly with the bank size.

$$C_{\mathcal{B}} \geq \Upsilon m, \qquad \Upsilon > 0. \tag{C.9}$$

The pre-filter mask count in $\mathcal{B}$ is $C_{\mathcal{B}}$, and Υ gives its per-channel visibility lower bound under the fixed scan.

Let μ_r and μ_c be the normalized empirical feature measures of source and mask candidates in $\mathcal{B}$. Equation (C.10) states a conditional overlap assumption.

$$\mu_c(\mathcal{V}) \geq \Lambda \mu_r(\mathcal{V}), \qquad \mathcal{V} \subseteq \mathcal{B}, \quad 0 < \Lambda \leq 1. \tag{C.10}$$

The parameter Λ measures distributional overlap over all measurable subsets of $\mathcal{B}$. It is a condition to audit, not a universal property of every mask bank.

Equation (C.11) defines the source retention of the public filter inside the overlap region.

$$\rho_{\mathcal{B}} = \int_{\mathcal{B}} \varsigma(y)\, d\mu_r(y) > 0. \tag{C.11}$$

For any filter that preserves some source recall in $\mathcal{B}$, $\rho_{\mathcal{B}}$ is positive.

Proposition A5 (Conditional overlap bound). *Under Equations* (C.9)–(C.11)*, any filter depending only on y has the source-precision upper bound in Equation* (C.12)*.*

Equation (C.12) displays the bound.

$$\Pi_{\mathcal{B}} \leq \frac{R_{\mathcal{B}}}{R_{\mathcal{B}} + \Lambda \Upsilon m}. \tag{C.12}$$

The pre-filter source count is $R_{\mathcal{B}}$, and $\Pi_{\mathcal{B}}$ is filtered source precision after nonnegative mixed mass is included in the denominator.

Proof. Setwise domination in Equation (C.10) extends from indicator functions to any nonnegative measurable retention function. Equation (C.13) gives the key step.

$$\int_{\mathcal{B}} \varsigma(y)\, d\mu_c(y) \geq \Lambda \int_{\mathcal{B}} \varsigma(y)\, d\mu_r(y) = \Lambda \rho_{\mathcal{B}}. \tag{C.13}$$

The left side is the mask retention rate in $\mathcal{B}$. The expected retained source count is $\rho_{\mathcal{B}} R_{\mathcal{B}}$. Equations (C.9) and (C.13) make the expected retained mask count at least $\Lambda \rho_{\mathcal{B}} \Upsilon m$. Substituting the two counts into source precision, dropping only nonnegative mixed mass, and cancelling the positive $\rho_{\mathcal{B}}$ proves Equation (C.12).

The proposition exposes a precise recall–precision tradeoff. A public filter can escape the bound by using information outside y, violating the overlap condition, or driving $\rho_{\mathcal{B}}$ to zero. The last option removes source recall rather than restoring a clean recovery stream.

C.7 Monotonicity of conditional selection sensitivity

Equation (32) defines the conditional sensitivity $\omega(L; n, c)$. Without repeating that derivation or its numerical table, Equation (C.14) proves its monotonicity in the remaining mask count.

$$\frac{\omega(L; n, c+1)}{\omega(L; n, c)} = \prod_{\ell=1}^{n} \frac{L\ell + c}{L\ell + c + 1} < 1. \tag{C.14}$$

The bound uses the main text's likelihood-ratio sensitivity $L > 0$, a source count $n \geq 1$, and a mask count $c \geq 0$. Every factor is strictly below one, so their product is strictly below one. Thus an additional competing mask decreases the conditional sensitivity without asserting an empirical attack success probability.

D Experimental Protocol and Metric Definitions

D.1 Datasets, models, and training regimes

Experiments use MNIST and CIFAR-10. The MLP family contains seven width/depth structures under Standard training and TTD training [15], as defined in the main text. The CNN transfer audit uses a small MNIST CNN and a small CIFAR-10 CNN. Each wrapper reuses the paired frozen checkpoint. The MLP structures are 8×2, 16×2, 16×3, 16×8, 32×2, 64×2, and 128×2. The MNIST CNN uses conv16/32 with a fully connected head, while the CIFAR-10 CNN uses conv32/64/96 with a fully connected head. CNN results test transfer of the masking principle to first-ReLU geometry; they do not share the MLP reference set.

D.2 Attack execution spine

The evaluation keeps the attack code fixed within each comparison and changes only the queried oracle. This design isolates evidence acquisition from later recovery heuristics. The raw-logit pipeline proceeds as follows.

1. Select fixed anchors and public probe directions. For the CNN audit, embed the lines in a two-dimensional normal plane associated with each anchor.
2. Query equally spaced points on each line and compute finite-difference slopes from returned scores.
3. Retain locations whose slope changes exceed the frozen threshold, then refine each location without consulting source/mask identities.
4. Use the frozen RANSAC and clustering rules to merge compatible locations into recovered lines or hyperplanes.
5. Fit the unchanged downstream signature or surrogate stage with its fixed fit-query budget.
6. After the pipeline stops, match recovered objects to source and mask references for evaluator-side precision, recall, and composition counts.

The hard-label pipeline uses a different observation but the same separation between attack execution and evaluation. It first chooses fixed pairs with different labels, refines their boundary transitions by binary search, and estimates a local class-boundary normal using label-only probes. It then walks the boundary patch and retains normal-change locations as dual candidates. Only after all candidates are frozen does the evaluator determine whether a candidate is pure source, mixed, mask-induced, or numerical.

D.3 Pairing and freeze rules

Each pair shares a source checkpoint, anchor set, attack seed, query budget, thresholds, matching tolerances, and downstream fit. The wrapper is calibrated and frozen before the held-out audit. Parameter screening uses separate local rows; a screened result is promoted to a four-anchor CNN run only after the prediction-agreement gate is checked. The completed one-sided remote run was subsequently requested as a full validation and is reported with all four anchors rather than only the favorable local anchor. Attack budgets and thresholds may differ across evidence classes because their query primitives and reference objects differ, but neither changes within a paired comparison. Source-aware identities never enter an attacker-side selection rule.

D.4 Fixed attacks and budgets

The raw-logit MLP comparison uses the same detector, line budget, clustering, matching rule, and fit budget for each source/NJM pair. The local six-anchor matrix uses a 17×101 scan budget and Seed 1337. The hard-label comparison uses 24 fixed transitions for its primary table and 96 transitions for the larger necessary-input audit. CNN experiments use four fixed anchors, 52 scan lines per anchor, 81 points per line, 1,600 RANSAC iterations, 3,764 scan queries, and 400 fit queries for each unfiltered condition. These budgets are reported at the evidence-acquisition stage; evaluator-side identity matching does not consume attacker queries.

D.5 Metrics and denominators

Let r be the retained source-event count, ι the retained mask/mixed/false count, and N_r the reference source-event count for the current stage. Equation (D.1) defines precision.

$$\zeta = \frac{r}{r+\iota}. \tag{D.1}$$

An empty retained set is reported explicitly rather than assigned a favorable score. Equation (D.2) defines recall.

$$\varphi = \frac{r}{N_r}. \tag{D.2}$$

The reference count N_r remains local to the attack stage. Equation (D.3) defines the harmonic mean when $\zeta + \varphi > 0$.

$$\mathrm{F1} = \frac{2\zeta\varphi}{\zeta+\varphi}. \tag{D.3}$$

The value of F1 is zero when both components are zero.

If N_q queries recover r source events, Equation (D.4) defines the query-normalized cost.

$$\chi_q = \frac{N_q}{r}. \tag{D.4}$$

When $r = 0$, χ_q is undefined and shown as a dash. Equation (D.5) defines prediction agreement on an audit set $\mathcal{X}_a$.

$$a_g = \frac{1}{|\mathcal{X}_a|} \sum_{x \in \mathcal{X}_a} \mathbf{1}\left\{ \arg\max_a F_a(x) = \arg\max_a f_a(x) \right\}. \tag{D.5}$$

Agreement a_g is checked against the declared gate of 0.999 for promoted CNN configurations. Standard errors in the CNN table are computed across the four fixed anchors. Anchors share a checkpoint, so the standard error describes location variability rather than independent model uncertainty.

D.6 Aggregation and zero-count rules

The primary tables aggregate at the unit that is actually varied. The MLP anchor matrix averages six fixed locations for one paired checkpoint. The three-seed mask-count audit first computes one value per seed and then reports their mean and standard error. The CNN audit computes one value per fixed anchor and reports the four-anchor mean and standard error. Candidate-level rows within the same seed or anchor share geometry and are not treated as independent trials.

For J_a seed- or anchor-level measurements ϑ_i, Equation (D.6) defines the reported standard error.

$$\mathrm{SE}(\vartheta) = \frac{1}{\sqrt{J_a}} \left(\frac{1}{J_a - 1} \sum_{i=1}^{J_a} (\vartheta_i - \bar{\vartheta})^2 \right)^{1/2}. \tag{D.6}$$

In this calculation, $\bar{\vartheta}$ is the arithmetic mean and $J_a \geq 2$. This number summarizes observed location or seed variation; it is not used to claim an IID candidate-level confidence interval.

When no source event is recovered, χ_q from Equation (D.4) is undefined and the table displays a dash. When an attack returns no candidate, the underlying counts remain visible and no perfect precision is imputed. Only completed rows enter paper-facing means. Shared-anchor standard errors describe location variation, and no recall or F1 is pooled across evidence classes with different reference sets.

E Utility, Geometry, and Parameter Audits

E.1 No-test calibration audit

The construction split contains training samples and augmentations. The final test split never calibrates thresholds. Table A1 reports the recomputed exact construction-set check and held-out agreement for four dense wrappers. All construction-set maximum differences are zero in the stored audit. Held-out agreement is 1.0000.

Table A1. No-test calibration and held-out audit for dense mixing. The latency ratio is implementation-specific and is not used as a security metric.

Model	Geometry	Train max diff.	Train+aug max diff.	Test max diff.	Test agreement	Latency ratio
MNIST MLP	One-sided	0.0000	0.0000	0.0758	1.0000	1.7817
MNIST MLP	Slab	0.0000	0.0000	0.0000	1.0000	2.8958
MNIST CNN	One-sided	0.0000	0.0000	0.0846	1.0000	1.1352
MNIST CNN	Slab	0.0000	0.0000	0.0222	1.0000	1.2711

Table A2. Mean MLP event counts over 16 diagnostic lines at $m = 128$ with dense mixing. Source/mask identities are used only for this post-hoc geometry audit.

Geometry	Direction	Candidates	Source	Mask	Mixed	Mask/source	Queries
One-sided	Tangent	66.63	36.63	0.00	0.00	0.0000	201
One-sided	Normal	23.44	10.81	1.31	0.13	0.1214	201
One-sided	Random	34.06	17.00	1.25	0.19	0.0735	201
Slab	Tangent	69.31	39.06	0.00	0.00	0.0000	201
Slab	Normal	25.81	11.19	2.19	0.13	0.1955	201
Slab	Random	35.19	15.88	2.19	0.06	0.1378	201

The result distinguishes exact protected-set fidelity from held-out behavioral fidelity. Exact zeros follow from threshold construction on the train and train-augmentation sets. Agreement on test examples is empirical and must be rechecked after distribution, preprocessing, or checkpoint changes.

E.2 Directional event audit

Table A2 checks the design premise before running a full recovery pipeline. Tangent probes largely remain inside the protected geometry, whereas normal and random probes encounter the private bank. At the same $m = 128$ setting, the slab exposes more mask events than the one-sided channel on both normal and random directions. Boundary-directed probes are dominated by source events within the finite scan range and therefore do not serve as the mask-density endpoint.

The diagnostic does not itself establish extraction resistance. It verifies that the calibrated bank is selective with respect to probe direction and explains why mask count, surface multiplicity, and scan resolution must be evaluated together in a complete attacker. Tables A3 and A4 perform that next step.

E.3 Three-seed mask-count audit

Table A3 reports the MLP seed-repeat summary. Increasing m raises mask events and lowers candidate precision while mean clean accuracy is unchanged. The trend saturates near $m = 512$–1024 and slightly reverses at $m = 2048$, so the result supports a tunable geometry effect rather than a claim of monotonic improvement at arbitrary scale.

Table A3. Three-seed MLP mask-count audit. Acc. is mean clean accuracy, Precision is mean source-candidate precision, SE is its standard error across the three seeds, and Mask events is the mean detected mask-event count.

m	Seeds	Acc.	Precision	SE	Mask events
16	3	0.9764	0.9273	0.0059	7.04
32	3	0.9764	0.8753	0.0049	11.67
64	3	0.9764	0.8165	0.0075	18.50
128	3	0.9764	0.7360	0.0130	28.08
256	3	0.9764	0.6908	0.0085	31.63
512	3	0.9764	0.6893	0.0092	33.67
1024	3	0.9764	0.6783	0.0064	34.08
2048	3	0.9764	0.7075	0.0234	34.54

Table A4. One-anchor CIFAR-10 slab mask-count sweep over the five configurations satisfying the 0.999 prediction-agreement gate. Visible mask counts surfaces intersecting the scan plane; Recovered mask counts mask-matched outputs of the fixed 28-line detector. Precision is source matches divided by all recovered lines; every row has zero spurious matches.

m	Visible mask	Recovered mask	Precision	Agreement
16,384	2,521	5	0.8214	1.0000
32,768	5,036	5	0.8214	1.0000
65,536	9,960	6	0.7857	1.0000
131,072	20,018	6	0.7857	1.0000
262,144	40,403	9	0.6786	1.0000

E.4 CNN mask-count and geometry audit

The fixed one-anchor CIFAR-10 screening sweep in Table A4 increases a slab bank without changing the attacker, amplitude, clearance, width, mixing, seed, scan settings, RANSAC, or fit budget. Visible mask surfaces grow sharply, but only a few survive the fixed detector. The completed utility-preserving rows therefore motivate a geometry comparison rather than further blind scaling.

Table A5 then compares slab and one-sided channels at the same local budgets. One-sided channels expose fewer surfaces and recover more masks per visible surface, but their absolute mask count and resulting source precision are worse at these two local points. The later four-anchor run is still valuable because it directly measures location variability and reaches the lower precision reported in Appendix G.

F Expanded Hard-Label Audit

The support audit shows 2 pure-source inputs at $m = 128$ and none at $m = 256$, 512, or 1024 for both Standard+NJM and TTD+NJM. The larger 96-transition audit in Table A6 preserves the conclusion while quadrupling the transition set. It also records source ReLU-unit coverage. The defended rows observe many source units, but only inside mixed candidates.

Table A5. Local one-anchor slab/one-sided comparison under the same 1,412-query, 28-line attack. Source and Mask are evaluator-side matches among recovered lines, and Precision is Source divided by Source plus Mask. All four rows have zero spurious matches and prediction agreement 1.0000.

Geometry	m	Source	Mask	Precision
Slab	65,536	22	6	0.7857
One-sided	65,536	23	5	0.8214
Slab	262,144	19	9	0.6786
One-sided	262,144	21	7	0.7500

Table A6. Ninety-six-transition hard-label necessary-input audit at $m = 512$. Obs. is the number of observed source units, Pure supp. is the number supported by pure-source candidates, and Cov. is their source-only coverage for the candidate-only orientation stage.

Target	Dual	Pure	Mixed	Obs.	Pure supp.	Cov.
Standard	276	276	0	245	245	1.0000
Standard+NJM	284	0	284	246	0	0.0000
TTD	284	284	0	254	254	1.0000
TTD+NJM	275	0	275	256	0	0.0000

This is a necessary-input analysis, not a completed hard-label sign-recovery implementation. It asks whether a downstream candidate-only orientation stage receives pure-source witnesses. Mixed candidates still contain source components, but their normals also include the private term analyzed by Equation (C.7).

G CNN Transfer and Per-Anchor Results

Table A7 exposes the anchor-level outcomes behind the CNN aggregates in the main text. The MNIST slab returns 52 lines at every anchor, with 33–36 source matches and 16–19 mask matches. On CIFAR-10, the one-sided bank recovers 7–18 mask lines, whereas the smaller slab bank recovers 1–4. The result is a stronger fixed-budget evidence mixture for the one-sided geometry without attributing the mean to one favorable anchor.

The adaptive filters also preserve the geometry distinction. The MNIST slab filter retains 13/4, 13/12, 8/4, and 10/4 source/mask lines across the four anchors. The CIFAR one-sided filter retains 12/5, 22/10, 13/10, and 18/8. These per-anchor counts explain why mean filtered precision can decrease even when total retained lines also decrease.

For denominator auditing, the aggregate files contain roughly 11,526 visible first-ReLU reference hyperplanes for MNIST and 28,108 for CIFAR-10. These counts instantiate N_r in Equation (D.2); Appendix A states the cross-evidence comparison rule.

Table A7. Completed CNN results by anchor. Base is the undefended retained source-line count; Src. and Mask are defended source and mask matches. Prec. is unfiltered source precision, and Filt. prec. is source precision after the frozen adaptive filter. Agreement is 1.0000 for MNIST and CIFAR-10 one-sided rows, and 0.9993 for CIFAR-10 slab rows.

Anchor	Base	Src.	Mask	Prec.	Filt. prec.
MNIST / Slab					
0	19	36	16	0.6923	0.7647
1	17	36	16	0.6923	0.5200
2	16	33	19	0.6346	0.6667
3	18	35	17	0.6731	0.7143
CIFAR-10 / One-sided					
0	10	45	7	0.8654	0.7059
1	12	37	15	0.7115	0.6875
2	12	35	17	0.6731	0.5652
3	15	34	18	0.6538	0.6923
CIFAR-10 / Slab					
0	10	50	1	0.9804	1.0000
1	12	43	4	0.9149	0.9091
2	12	44	4	0.9167	1.0000
3	15	50	3	0.9434	0.9000

H Reproducibility Notes

H.1 Companion artifacts

Frozen CSV/JSON files in the companion code and data are the table authority. The provenance manifest maps every retained result table to the corresponding MLP, hard-label, or CNN artifact. The companion README identifies the three package-check scripts and the commands used to run them.

The core recorded environment uses Python 3.12.3, PyTorch 2.7.1 with CUDA 12.6, torchvision 0.22.1, and an NVIDIA RTX 4060 Laptop GPU. The optional official TTD pipeline requires a separate TensorFlow-compatible environment. Seeds and deterministic flags are set by the individual runners. Checkpoints, downloaded datasets, large logs, and per-sample logits are omitted from the companion code and data.

H.2 Result-integrity checks

The verifier recomputes Tables A4, A5, and A7 from frozen CSVs and checks count identities, precision, four anchor rows, and the 0.999 gate. The one-sided audit additionally verifies $m = 262{,}144$, channel type, Seed 1337, and 12 error-free conditions.