Distill and Conquer: Cryptanalytic Extraction of Local Rules in the Hard-Label Setting

Chun Li[1], Di Li[1], Liping Zhuang[1], Siqi Zhu[1], Yufeng Tang[2], Xingfu Yan[1], and Zheng Gong[1⋆]

[1] School of Computer Science, South China Normal University, China
cis.gong@gmail.com
[2] Institute for Network Sciences and Cyberspace, Tsinghua University, China

Abstract. Cryptanalytic extraction can expose proprietary neural network parameters, but unknown architectures and large parameter counts complicate recovery under hard-label access. Local class-pair rules offer a narrower target by making class-switching conditions explicit and predicting nearby labels under a specified sampling law. Distillation methods mainly return substitute models rather than explicit rules whose local agreement has been tested against the victim. We propose Distill–Probe–Extract to produce such rules without first identifying the victim architecture. Verified boundary events guide student training, and student gradients propose directions at boundary anchors refined through victim queries. Fresh hard-label queries then test the proposed rules, retaining validation inputs and victim labels for each accepted rule. On a digits multilayer perceptron and pooling convolutional networks, fixed-candidate comparisons isolate the savings from replacing victim-normal estimation with student gradients. For the multilayer perceptron, median recorded online cost decreases from 246 to 118 queries per attempt after student preparation. With preparation and failures counted, full Fashion-MNIST sessions produce 2.03–2.23 times Direct's accepted-rule count under the original validation gate at 320,000-call caps. At 320,000 calls, a separate student-level test on trained Blended watermarks yields 99.99% trigger agreement and 91.79% clean agreement, with all three seeds passing joint verification.

Keywords: Cryptanalytic extraction · Knowledge distillation · Hard-label access · Local rules · Neural network security

1 Introduction

Cryptanalytic extraction uses chosen inputs to recover hidden neural-network parameters. Extracting proprietary behavior through prediction interfaces challenges model confidentiality, including services that expose only top-1 labels [12, 35]. Early attacks exploit ReLU piecewise linearity [10], with subsequent work demonstrating polynomial-time sign recovery on approximately 1.2 million parameters [6]. Hard-label attacks analyze decision and activation boundaries [9, 12].

⋆ Corresponding author: Zheng Gong.

Recent methods extend extraction to convolution, pooling, and other activations [22, 28, 34].

Recent work also infers ReLU MLP depth and widths from numerical outputs [33], detects omitted neurons with limited architectural knowledge [24], and recovers CNN structure through raw-output or hard-label pipelines [26]. The reported end-to-end unknown-architecture recoveries cover ReLU MLPs and two-convolution CNNs, leaving deeper heterogeneous architectures beyond their demonstrated scope. Hard-label cross-layer attacks also face structural constraints [8, 19].

Knowledge distillation learns a substitute without explicit parameter recovery [18]. DFME, DFMS, DB-DFMS, and BAM train substitutes using victim responses under their respective access assumptions [5, 27, 32, 36]. Natural-input agreement does not directly measure the local transitions needed for rule extraction [20]. Conditional responses used in watermarks and backdoors also require evaluation beyond clean-input accuracy [2, 37].

We target class-pair rules around queried boundary locations, where the victim's label changes from one class to another. A retained rule exposes a local switching condition and predicts nearby labels under its validation sampling law without further victim calls. This target captures partial model behavior without first recovering the architecture or parameters. Distill–Probe–Extract (DPE) combines event-conditioned training, reusable student directions, and hard-label rule validation. DFMS-HL uses student gradients to guide the inputs queried during substitute training [32], while DPE uses them to propose explicit class-pair halfspaces. Victim queries refine their anchors and test their predictions, producing a rule bank with attached validation records. A separate application study tests whether the trained student preserves conditional responses used for behavioral-watermark ownership verification.

We make three contributions.

- **Transition supervision.** Victim-verified boundary events supply class-switching information for student training, supplementing ordinary labels before the student proposes local rules.
- **Reusable direction estimates.** Student gradients replace repeated victim-normal probes. Fixed-candidate comparisons isolate direction costs, while complete hard-label sessions measure accepted-rule yield with preparation and failures counted.
- **Validation of approximate proposals.** Fresh two-sided victim tests score both observed classes and retain accepted rules with their sampled inputs, labels, and transcript positions as validation evidence.

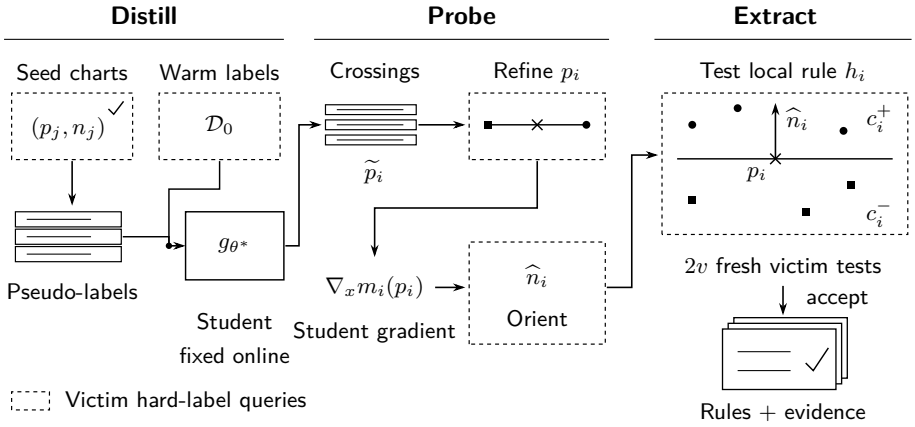


Fig. 1. DPE under hard-label access. Dashed boxes mark victim-query operations, and the tick marks a validated seed chart. Pseudo-labels and warm labels train a student that is fixed during Probe and Extract. Probe refines a victim anchor before differentiating the student margin. Extract tests $2v$ fresh two-sided samples before retaining a local rule and its evidence. Every attempt advances the queue, including rejection. Circles/squares denote the two victim classes, and crosses mark refined anchors. Geometry is schematic.

2 Method

Figure 1 follows *Distill*, *Probe*, and *Extract* from a verified event to an accepted local rule. Algorithm 1 gives the complete procedure, with implementation details and local chart analysis in Appendix A.

The attacker queries a deterministic top-1-label API $f : \mathcal{X} \to [C]$, where $\mathcal{X} = [0,1]^d$ has d input coordinates and $[C] = \{1, \ldots, C\}$ indexes classes. Input preprocessing, numerical precision, and batching follow fixed conventions. The extraction target is a bank of class-pair rules that predict victim labels around refined boundary anchors under a specified local sampling law. Victim scores, gradients, weights, architecture metadata, and hardware side channels are unavailable, and Section 3 specifies experimental assistance, input reuse, and counters. Given setup/online budgets Q_0, Q, DPE trains differentiable students $g_\theta : \mathcal{X} \to \mathbb{R}^C$, returning the selected student, event bank $\mathcal{E}$, rule bank $\mathcal{R}$, and transcript $\mathcal{T}$.

For Algorithm 1, $\mathbf{X}_0$ contains seed inputs, $\mathcal{P}$ supplies candidate event transformations, and $\mathcal{G}$ lists candidate student architectures.

All victim calls use the budgeted query wrapper. Failed or unfunded operations return $\perp$, retaining spent calls. Algorithm 2 expands the gate, and Algorithm 3 gives the query wrapper.

Algorithm 1 Distill–Probe–Extract main procedure

Require: $f, \mathbf{X}_0, \mathcal{P}, \mathcal{G}, Q_0, Q$, chart target $r_{\max}$, samples per side v, scales $a_{\text{syn}}, a_{\text{val}}, \rho > 0$
Ensure: $g_{\theta^*}, \mathcal{E}, \mathcal{R}, \mathcal{T}$

$\mathcal{E}, \mathcal{R}, \mathcal{T} \leftarrow \emptyset$, $q_0, q \leftarrow 0$
Distill: event setup, training, and selection
$(\mathbf{Y}_0, q_0, \mathcal{T}) \leftarrow \textsc{Query}(f, \mathbf{X}_0, \textsc{warm}, q_0, Q_0, \mathcal{T})$
if $\mathbf{Y}_0 = \perp$ **then return** $(\perp, \mathcal{E}, \mathcal{R}, \mathcal{T})$
$\mathcal{D}_0 \leftarrow ((x_{0,l}, y_{0,l}))_{l=1}^{|\mathbf{X}_0|}$
while $|\mathcal{E}| < r_{\max} \wedge q_0 < Q_0$ **do**
  $(B_j, q_0, \mathcal{T}) \leftarrow \textsc{FindBracket}(f, \mathbf{X}_0, \mathcal{P}, \mathcal{D}_0, q_0, Q_0, \mathcal{T})$
  if $B_j = \perp$ **then break** ▷ search exhausted or unfunded
  $(\Gamma_j, q_0, \mathcal{T}) \leftarrow \textsc{BuildChart}(f, B_j, q_0, Q_0, \mathcal{T})$
  if $\Gamma_j \neq \perp$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{\Gamma_j\}$
  end if
end while
$\mathcal{D} \leftarrow \mathcal{D}_0 \parallel \textsc{Synthesize}(\mathcal{E}, a_{\text{syn}})$ ▷ Eq. (3), offline
for $k = 1, \ldots, |\mathcal{G}|$ **do**
  $\theta_k \leftarrow \textsc{Train}(g_k, \mathcal{D}, \mathcal{L}_k)$ ▷ Eq. (4)
  $(A_k, q_0, \mathcal{T}) \leftarrow \textsc{Score}(f, g_{\theta_k}, q_0, Q_0, \mathcal{T})$ ▷ Eq. (7)
  if $A_k = \perp$ **then return** $(\perp, \mathcal{E}, \mathcal{R}, \mathcal{T})$
end for
$k^* \leftarrow \arg\max_k A_k$, $g_{\theta^*} \leftarrow g_{\theta_{k^*}}$
Probe: candidate search and oriented student geometry
$\mathcal{Z} \leftarrow \textsc{RankProposals}(g_{\theta^*})$ ▷ Eq. (8), offline
while $\mathcal{Z} \neq \emptyset \wedge q < Q$ **do**
  $z_i \leftarrow \textsc{PopFirst}(\mathcal{Z})$
  $(p_i, c_i^-, c_i^+, q, \mathcal{T}) \leftarrow \textsc{Refine}(f, z_i, q, Q, \mathcal{T})$ ▷ victim bracket and bisection
  if $p_i = \perp$ **then continue**
  $m_i(x) \leftarrow g_{\theta^*, c_i^+}(x) - g_{\theta^*, c_i^-}(x)$, $d_i \leftarrow \nabla_x m_i(p_i)$
  if $d_i = \perp \vee \|d_i\|_2 = 0$ **then continue**
  $\widehat{n}_i \leftarrow d_i / \|d_i\|_2$, $x_i^{\pm} \leftarrow \Pi_{[0,1]^d}(p_i \pm \rho \widehat{n}_i)$
  $(\mathbf{y}_i, q, \mathcal{T}) \leftarrow \textsc{Query}(f, (x_i^-, x_i^+), \textsc{orient}, q, Q, \mathcal{T})$
  if $\mathbf{y}_i = (c_i^+, c_i^-)$ **then**
    $\widehat{n}_i \leftarrow -\widehat{n}_i$
  else if $\mathbf{y}_i \neq (c_i^-, c_i^+)$ **then**
    continue
  end if
  Extract: local-rule validation
  $(\mathcal{C}_i, q, \mathcal{T}) \leftarrow$
  $\textsc{ValidateRule}(f, p_i, \widehat{n}_i, c_i^-, c_i^+, a_{\text{val}}, v, q, Q, \mathcal{T})$
  if $\mathcal{C}_i \neq \perp$ **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{\mathcal{C}_i\}$
  end if
end while
return $g_{\theta^*}, \mathcal{E}, \mathcal{R}, \mathcal{T}$

2.1 Distill: Event-Conditioned Transfer

Warm labels from natural or benignly perturbed inputs form $\mathcal{D}_0$. DPE searches natural or publicly transformed inputs for opposite-label endpoints, reusing cached labels and charging new labels and failures to setup. Search stops at $r_{\max}$ charts or an exhausted schedule or budget. These events encode where a class transition occurs and a direction across it. Starting from $f(a_j) = c_j^- \neq c_j^+ = f(b_j)$, pair-preserving bisection contracts brackets by 2^{-t} after t exact-arithmetic steps, with midpoint p_j localizing the transition (Eq. (21)). Third classes require rebracketing or rejection, while repeated floating-point midpoints stop refinement when the bracket can no longer advance. Paired victim probes then estimate the seed direction.

Let $\Pi_{[0,1]^d}$ denote clipping to the input domain. For seed construction, draw K_{seed} unit Gaussian directions $d_{j,k}$ at probe scale $\rho_{\text{seed}} > 0$. Query $x_{j,k}^{\pm} = \Pi_{[0,1]^d}(p_j \pm \rho_{\text{seed}} d_{j,k})$ and let $y_{j,k}^{\pm} = f(x_{j,k}^{\pm})$. Retain indices $\mathcal{I}_j$ whose paired labels are the chart's two distinct classes, and estimate

$$\bar{n}_j = \frac{1}{|\mathcal{I}_j|} \sum_{k \in \mathcal{I}_j} \big(\mathbb{1}[y_{j,k}^+ = c_j^+] - \mathbb{1}[y_{j,k}^- = c_j^+]\big)(x_{j,k}^+ - x_{j,k}^-). \tag{1}$$

Where $\bar{n}_j$ is the unnormalized seed direction. Equation (1) averages signed, clipped displacements, rejecting empty $\mathcal{I}_j$ or zero estimates. Normalize $n_j = \bar{n}_j / \|\bar{n}_j\|_2$ and orient it with two clipped victim probes. A chart enters $\mathcal{E}$ only after passing Eq. (15), with third-class responses rejected. Refinement, all $2K_{\text{seed}}$ probes, orientation, and $2v$ validation calls count toward Q_{seed}.

Figure 2 shows how an oriented chart generates training rows by separating signed displacement from tangent variation:

$$x_j(\alpha, z) = p_j + \alpha n_j + \beta H_j z, \qquad H_j = I_d - n_j n_j^\top. \tag{2}$$

Where α is the signed offset, $z \in \mathbb{R}^d$ a perturbation, β its scale, and I_d the identity matrix.

The projector H_j removes components along n_j. Without clipping, Eq. (2) gives $n_j^\top (x_j - p_j) = \alpha$, so tangent variation preserves the modeled side.

DPE synthesizes pseudo-labeled rows as

$$\widetilde{x}_{j,k} = \Pi_{[0,1]^d} x_j(\alpha_{j,k}, z_{j,k}), \qquad \widetilde{y}_{j,k} = \begin{cases} c_j^+, & \alpha_{j,k} \geq 0, \\ c_j^-, & \alpha_{j,k} < 0. \end{cases} \tag{3}$$

Where k indexes synthesized rows. Equation (3) supplies training examples on both modeled sides without further victim calls. These pseudo-labels assume a local halfspace. Gaussian tangent offsets and clipping can violate that assumption, motivating subsequent victim validation. Binary synthesis draws $\alpha \sim U[-a_{\text{syn}}, a_{\text{syn}}]$. It uses a unit projected Gaussian tangent with independent amplitude distributed as $N(0, (0.35 a_{\text{syn}})^2)$. Equal row counts per chart and their retained training multiplicities determine λ_E.

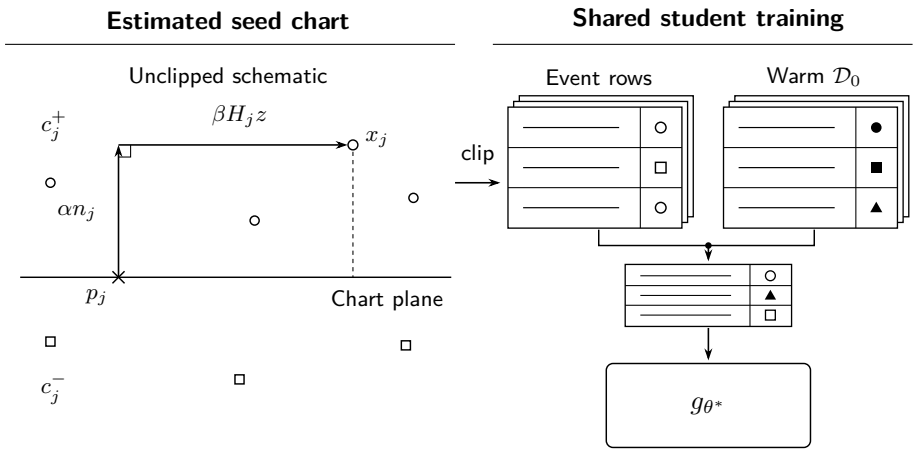


Fig. 2. Event-conditioned transfer (unclipped schematic). Victim probes estimate the unit normal n_j at the refined midpoint p_j. Hollow circles/squares denote chart pseudo-labels c_j^+/c_j^- for $\alpha \geq 0/\alpha < 0$. Rows from all retained charts, clipped to $[0,1]^d$, join victim-labeled warm data $\mathcal{D}_0$ (filled marks). Warm labels may include other classes. Row counts do not specify training proportions. The shared student is fixed online. Synthesis adds no victim calls.

Warm-label and event losses jointly define the training objective

$$\mathcal{L}(\theta) = (1-\lambda_E)\widehat{\mathbb{E}}_0 \ell_{\mathrm{CE}}(g_\theta(x), y) + \frac{\lambda_E}{r}\sum_{j=1}^{r} \widehat{\mathbb{E}}_j \ell_{\mathrm{CE}}(g_\theta(\widetilde{x}), \widetilde{y}). \tag{4}$$

Where $r = |\mathcal{E}| > 0$, $\widehat{\mathbb{E}}_0, \widehat{\mathbb{E}}_j$ average warm/event losses, ℓ_{CE} is cross-entropy, and λ_E is the event fraction.

Equation (4) combines prediction supervision with explicit transitions. Training reverts to warm labels when the event bank is empty.

Consider an unclipped binary chart with $X - p_j = \sigma R n_j + \beta Z$, where σ, R, Z are mutually independent, σ is uniform on $\{-1, +1\}$, $R > 0$, and Z is zero-mean, centrally symmetric, and tangent to n_j. Assume $\mathbb{E}[R] + \mathbb{E}\|Z\|_2 < \infty$. For a local linear margin $m_w(x) = w^\top(x - p_j)$, the regularized population risk is

$$\mathcal{L}_j(w) = \mathbb{E}\big[\log\big(1 + \exp\{-\sigma w^\top (X - p_j)\}\big)\big] + \frac{\lambda}{2}\|w\|_2^2. \tag{5}$$

Where $\lambda > 0$ is the regularization coefficient. Equation (5) isolates the event term under these sampling conditions.

Proposition 1 (Event-normal consistency). *For $\lambda > 0$, the risk in Eq. (5) has a unique minimizer satisfying*

$$w_j^* = \kappa_j n_j \qquad \text{for some } \kappa_j > 0. \tag{6}$$

Where κ_j is a scale determined by the offset distribution and regularization.

Equation (6) aligns the optimum with the pseudo-label normal supplied by seed estimation.

Proof. Write $w = an_j + b$ with $b^\top n_j = 0$. Conditional on R, the signed margin is $aR + \beta b^\top(\sigma Z)$ and $\mathbb{E}[\sigma Z \mid R] = 0$. Jensen's inequality makes $b = 0$ optimal for every a, strictly so with the quadratic term. The regularized risk is coercive and strictly convex. At $a = 0$, its derivative is $-\mathbb{E}[R]/2 < 0$, so the unique optimum has $a = \kappa_j > 0$.

When architecture metadata is hidden, train candidate students on the same setup allocation. Boundary-stress fidelity F_k^{bnd} measures student–victim agreement on perturbed boundary-pair inputs. Verified-event fidelity F_k^{loc} measures agreement on fresh samples around retained seed charts. Combining these with trial acceptance rate S_k^{ver} selects

$$k^* = \arg\max_{k \in [|\mathcal{G}|]} \left(F_k^{\mathrm{bnd}} + F_k^{\mathrm{loc}} + S_k^{\mathrm{ver}} \right). \tag{7}$$

Selection queries in Eq. (7) consume setup calls, and an unfunded score rejects selection. The hard-label specification uses Eq. (15) for trial success. The original selection study instead uses reference-verified localization, with sampling specified in Appendix A.1. The complete-session experiments use fixed student architectures.

2.2 Probe: Student-Guided Geometry

DPE bisects endpoint pairs with differing student predictions offline and ranks crossings by

$$s_i = |\widetilde{m}_i(\widetilde{p}_i)| - \eta \min\{|\widetilde{m}_i(a_i)|, |\widetilde{m}_i(b_i)|\}. \tag{8}$$

Where $\widetilde{p}_i$ is the student crossing, $\widetilde{m}_i$ uses the student endpoint class pair, and $\eta = 0.05$ weights endpoint confidence. Low scores in Eq. (8) prioritize victim search at increasing radii along $b_i - a_i$, followed by opposite-label bracket refinement. This refinement anchors the rule at a victim transition even when the student's boundary is displaced.

At the victim-refined anchor p_i, recompute the margin $m_i(x) = g_{\theta^*, c_i^+}(x) - g_{\theta^*, c_i^-}(x)$ for the observed victim class pair. Differentiability gives the local expansion

$$m_i(p_i + \Delta x) = m_i(p_i) + \nabla_x m_i(p_i)^\top \Delta x + o(\|\Delta x\|_2). \tag{9}$$

Where Δx is an input displacement. Equation (9) supplies a first-order direction even when the victim-refined anchor satisfies $m_i(p_i) \neq 0$.

The class-pair gradient uses two rows of the student's input Jacobian:

$$\mathbf{J}_i = \begin{bmatrix} \partial_1 g_{\theta^*,1} & \cdots & \partial_d g_{\theta^*,1} \\ \vdots & \ddots & \vdots \\ \partial_1 g_{\theta^*,C} & \cdots & \partial_d g_{\theta^*,C} \end{bmatrix}_{x=p_i}. \tag{10}$$

Where $\mathbf{J}_i \in \mathbb{R}^{C\times d}$ contains student-logit derivatives and ∂_ℓ denotes differentiation with respect to input coordinate x_ℓ. The implementation obtains the two-row difference in Eq. (10) by differentiating the margin directly.

Normalizing the difference between the two class rows gives

$$\widehat{n}_i = \frac{\mathbf{J}_i^\top (e_{c_i^+} - e_{c_i^-})}{\|\mathbf{J}_i^\top (e_{c_i^+} - e_{c_i^-})\|_2}. \tag{11}$$

Where $e_c \in \mathbb{R}^C$ is the cth standard basis vector. Since $\mathbf{J}_i^\top (e_{c_i^+} - e_{c_i^-}) = \nabla_x m_i(p_i)$, Eq. (11) requires a nonzero margin gradient. Two clipped symmetric victim probes orient this direction to the refined class pair, flipping its sign or rejecting it as needed.

2.3 Extract: Victim-Validated Local Rules

DPE defines a local decision rule from the oriented direction and victim-refined anchor:

$$h_i(u) = \begin{cases} c_i^+, & (u - p_i)^\top \widehat{n}_i \geq 0, \\ c_i^-, & (u - p_i)^\top \widehat{n}_i < 0. \end{cases} \tag{12}$$

Where u is a query input and h_i the proposed local predictor.

To test Eq. (12) beyond the orientation axis, DPE samples normal and tangent offsets:

$$u_{i,k}^s = \Pi_{[0,1]^d}\left(p_i + s r_{i,k}^s \widehat{n}_i + \widehat{H}_i \xi_{i,k}^s\right), \quad s \in \{-1, +1\}. \tag{13}$$

Where $\widehat{H}_i = I_d - \widehat{n}_i \widehat{n}_i^\top$, $r_{i,k}^s > 0$ is the normal offset, and $\xi_{i,k}^s$ supplies tangent variation. Equation (13) uses v independent draws per proposed side at validation scale $a_{\text{val}} > 0$, with $r_{i,k}^s \sim U[0.05a_{\text{val}}, a_{\text{val}}]$. Tangent offsets use a unit projected Gaussian direction with independent amplitude $N(0, (0.25a_{\text{val}})^2)$. Draws are independent within and across sides. The complete $2v$-call batch must fit the budget before sampling, and all calls count on rejected attempts. Equal proposed-side samples need not balance victim classes.

On validation rows $\mathcal{U}_i$, the gate measures mean recall over the two victim classes:

$$V_i = \frac{1}{2} \sum_{s\in\{-1,+1\}} \frac{1}{N_i^s} \sum_{\substack{u\in\mathcal{U}_i \\ f(u)=c_i^s}} \mathbb{1}[h_i(u) = c_i^s]. \tag{14}$$

Where N_i^s counts rows with victim label c_i^s. Both class counts must be positive to evaluate the score. Equation (14) averages victim-class recalls, whereas ordinary PF, $\widehat{F}_i = (2v)^{-1} \sum_{u\in\mathcal{U}_i} \mathbb{1}[h_i(u) = f(u)]$, weights validation rows equally regardless of their victim class.

With positive-label fraction $\pi_i = N_i^+/(2v)$, acceptance requires

$$V_i \geq 0.8 \quad \wedge \quad 0.2 \leq \pi_i \leq 0.8. \tag{15}$$

Equation (15) excludes one-class neighborhoods, and DPE additionally rejects attempts whose validation rows contain labels outside the refined class pair. Each retained validation record stores the anchor, direction, class pair, validation rows, labels, and transcript positions in $\mathcal{R}$ for subsequent inspection.

Algorithm 2 gives the complete sampling and rejection procedure.

Algorithm 2 Victim-validated local rule

function ValidateRule($f, p_i, \widehat{n}_i, c_i^-, c_i^+, a_{\text{val}}, v, q, Q, \mathcal{T}$)
  if $q + 2v > Q$ **then**
    return $(\perp, q, \mathcal{T})$
  end if
  $\widehat{H}_i \leftarrow I_d - \widehat{n}_i\widehat{n}_i^\top$, $t_0 \leftarrow |\mathcal{T}|$
  for $s \in \{-1, +1\}$ **do**
    for $k = 1, \ldots, v$ **do**
      $r_{i,k}^s \sim U[0.05a_{\text{val}}, a_{\text{val}}]$, $z_{i,k}^s \sim \mathcal{N}(0, I_d)$, $\beta_{i,k}^s \sim \mathcal{N}(0, (0.25a_{\text{val}})^2)$
      $t_{i,k}^s \leftarrow \widehat{H}_i z_{i,k}^s / \max\{\|\widehat{H}_i z_{i,k}^s\|_2, 10^{-12}\}$
      $u_{i,k}^s \leftarrow \Pi_{[0,1]^d}(p_i + sr_{i,k}^s\widehat{n}_i + \beta_{i,k}^s t_{i,k}^s)$
      $\widehat{y}_{i,k}^s \leftarrow h_i(u_{i,k}^s)$ ▷ rule in Eq. (12)
    end for
    $\mathcal{U}_i^s \leftarrow (u_{i,1}^s, \ldots, u_{i,v}^s)$
    $(\mathbf{y}_i^s, q, \mathcal{T}) \leftarrow$ Query($f, \mathcal{U}_i^s$, validate, $q, Q, \mathcal{T}$)
  end for
  if $\exists (s,k) : y_{i,k}^s \notin \{c_i^-, c_i^+\}$ **then**
    return $(\perp, q, \mathcal{T})$
  end if
  for $\sigma \in \{-1, +1\}$ **do**
    $\mathcal{I}_i^\sigma \leftarrow \{(s,k) : y_{i,k}^s = c_i^\sigma\}$, $N_i^\sigma \leftarrow |\mathcal{I}_i^\sigma|$
  end for
  if $N_i^- = 0 \vee N_i^+ = 0$ **then**
    return $(\perp, q, \mathcal{T})$
  end if
  $\pi_i \leftarrow N_i^+/(2v)$, $V_i \leftarrow \frac{1}{2}\sum_{\sigma\in\{-1,+1\}} \frac{1}{N_i^\sigma} \sum_{(s,k)\in\mathcal{I}_i^\sigma} \mathbb{1}[\widehat{y}_{i,k}^s = c_i^\sigma]$
  if $V_i \geq 0.8 \wedge 0.2 \leq \pi_i \leq 0.8$ **then**
    $\mathcal{C}_i \leftarrow (p_i, \widehat{n}_i, c_i^-, c_i^+, \mathcal{U}_i^\pm, \mathbf{y}_i^\pm, \{t_0 + 1, \ldots, |\mathcal{T}|\})$
    return $(\mathcal{C}_i, q, \mathcal{T})$
  end if
  return $(\perp, q, \mathcal{T})$
end function

Draws of r, z, β are mutually independent, within and across sides. Index s denotes the proposed side, and σ indexes the observed victim class. The 10^{-12} floor matches the implementation's tangent normalization. Attached transcript positions identify all $2v$ validation calls.

Let G_i denote agreement under the equal-weight mixture of the two candidate-side distributions in Eq. (13).

Proposition 2 (Finite-sample local interpretation). *For a candidate and sampling law fixed before validation, independent side-conditional draws give the lower bound*

$$\Pr\left[G_i \geq \widehat{F}_i - \sqrt{\frac{\log(1/\delta)}{4v}}\right] \geq 1 - \delta. \tag{16}$$

Where $\delta \in (0,1)$ is the allowed failure probability for this fixed-candidate bound.

Equation (16) bounds ordinary PF for one fixed candidate before conditioning on acceptance. It concerns the stated sampling law, with simultaneous guarantees over selected rules requiring additional analysis.

Proof. Fix the candidate and validation law before sampling. For each side s and draw k, let $Z_k^s = \mathbb{1}[h_i(u_{i,k}^s) = f(u_{i,k}^s)]$. The $2v$ indicators are independent and bounded in $[0,1]$, although the two sides need not have the same distribution. Their average is $\widehat{F}_i$, and its expectation is G_i under the equal-weight side mixture. Hoeffding's inequality gives $\Pr(\widehat{F}_i - G_i > \epsilon) \leq \exp(-4v\epsilon^2)$. Setting $\epsilon = \sqrt{\log(1/\delta)/(4v)}$ proves Eq. (16).

Actual setup calls are $Q_{\text{setup}} = Q_{\text{warm}} + Q_{\text{scan}} + Q_{\text{seed}} + Q_{\text{sel}}$, covering warm labels, event search, chart construction, and student selection. The complete session cost is

$$\begin{aligned} Q_{\text{session}} = Q_{\text{setup}} + Q_{\text{fill}} \\ + \sum_i \big(Q_{\text{pair},i} + Q_{\text{bis},i} + Q_{\text{orient},i} \\ + Q_{\text{val},i} + 2K_i I_i^{\text{VN}}\big). \end{aligned} \tag{17}$$

Where Q_{fill} counts optional filler calls. Per-attempt terms charge pair search, bisection, orientation, and validation, while K_i counts paired victim-normal directions and I_i^{VN} indicates their use. Equation (17) counts every victim-evaluated input, including repeated submissions and failures. DPE sets $I_i^{\text{VN}} = 0$ when the student supplies the normal. Cached labels, training exposures, student inference, differentiation, and synthesis add no victim calls.

Corollary 1 (Matched-attempt query saving). *For the same submitted candidate and identical bracketing, orientation, and validation schedules, replacing K_i paired victim directions by a student normal saves*

$$\Delta Q_i = Q_i^{\text{victim-normal}} - Q_i^{\text{DPE}} = 2K_i. \tag{18}$$

Where $Q_i^{\text{victim-normal}}$ and Q_i^{DPE} are complete online attempt costs, and ΔQ_i is their difference.

Equation (18) isolates matched-attempt direction costs, while Section 3.2 measures yield with preparation and failures included. Appendix D.1 proves the corollary, and Appendix D states global-coverage conditions.

3 Evaluation

3.1 Research Questions and Experimental Setup

RQ1 asks whether reusable students reduce extraction queries when the victim's internal structure is unknown. **RQ2** asks how boundary-event supervision affects local transitions and student retention of conditional responses. Boundary diagnostics measure the former, while ResNet patch transfer and trained-watermark tests measure the latter. Local extraction uses 8×8 sklearn digits for even/odd classification and MNIST 3-vs-8 pooling CNNs. Controlled studies isolate proposal and direction-estimation effects. Complete DPE on Fashion-MNIST CNNs with Patch, Blended, and Content/mask watermarks measures preparation-inclusive rule yield and learned response preservation. The original local studies pair victim/student widths 64/128 for digits MLPs and 128/256 for pooling CNNs. They pool 24 attempts over three seeds for the MLP and 16 attempts over four seeds per CNN, with two-seed pooling selection. Validation uses 96/80 points and $a_{\text{val}} = 0.008/0.012$ for MLP/CNN, respectively. Synthesis uses $a_{\text{syn}} = 0.012$, and orientation uses $\rho = a_{\text{val}}$. Appendix B.1 gives the remaining settings adapted from BAM [5] and Label Diversity [27].

Accuracy measures agreement with ground-truth labels, and PF measures agreement with victim labels on the specified input distribution. Rule PF in the original controlled studies uses candidate-dependent samples from Eq. (13), with medians over available records for costs and rule PF. A 100% median describes sampled local agreement rather than global recovery. Their reusable-rule counts additionally require refinement and reference boundary error below 10^{-3}, alongside the gate in Eq. (15). The preparation-inclusive studies report seed means over accepted, deduplicated online rules without reference filtering.

The original controlled local studies use reference-assisted seed screening and reuse unlabeled evaluation inputs from seeded digits holdouts or the official MNIST test split. Reference checks affect seed retention and reusable-rule counts, but their calls are omitted from the setup counts in Table 7. An actual query evaluates one victim input, whereas cached-label reuse and repeated training exposures incur no calls. Marginal yield includes online failures and excludes setup. The preparation-inclusive studies count setup and all extraction calls, with post-acceptance evaluation charged separately. Equation (17) gives the accounting convention.

3.2 RQ1: Reusable Students for Local-Rule Extraction

The paired comparison isolates the query savings attributable to student gradients by holding event-conditioned candidates fixed across 56 attempts (Fig. 3). Both normal sources accept 17/24, 11/16, and 9/16 rules on MLP/AvgPool/MaxPool, respectively, with median costs falling from 246/180/181 to 118/100/101. Only the victim-normal estimation calls differ, demonstrating lower online cost with unchanged acceptance on these matched attempts.

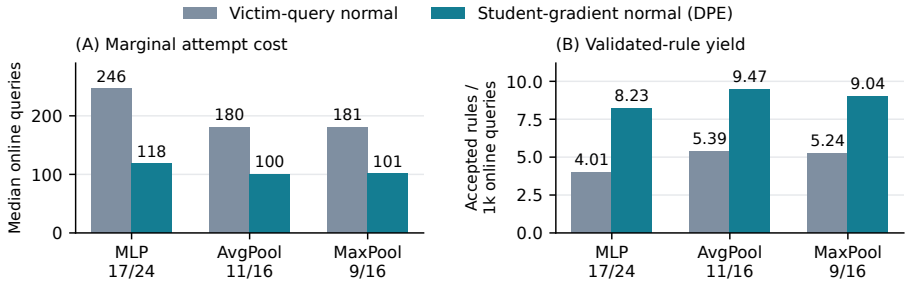


Fig. 3. Fixed-candidate comparison of normal sources, reporting median recorded online attempt cost (A) and accepted rules per 1k online queries (B). Accepted/attempted counts below each victim match across sources, using three seeds for the MLP and four for each pooling CNN. Both costs include orientation and validation, while yield counts queries from failed attempts and excludes setup queries and reference checks.

Table 1 compares DPE against random-candidate victim-normal probing, showing higher recorded marginal rule yields across the tested MLP and pooling CNNs. Its Rule PF column evaluates the extracted halfspaces alongside query costs, linking extraction efficiency to local predictive agreement. MLP median attempt cost falls from 257 to 118 queries and yield rises from 3.88 to 8.23 rules/1k queries. Pooling-CNN yields also rise despite fewer accepted rules, with this comparison reflecting differences in both candidate generation and normal estimation.

Table 1. Local-rule outcomes after reference-assisted seed screening. Accuracy and clean PF are means. Rule PF, costs, and absolute cosine are medians over available records. Counts pool attempts, and yield includes online failures but excludes setup and reference checks. Victim-probe rows use different candidates and victim classifier metrics.

Victim	Normal source	Acc.	Clean PF	Rule PF	Rules	Locate/ Full Q	Rules/ 1k Q	Abs. cosine
Digits MLP	Victim probes	97.0	100.0	100.0	24/24	31/257	3.88	0.651
	Student (DPE)	96.9	98.5	100.0	17/24	**20/118**	**8.23**	**0.939**
AvgPool CNN	Victim probes	96.1	100.0	100.0	16/16	24/186	5.36	0.179
	Student (DPE)	96.3	98.9	100.0	11/16	**18/100**	**9.47**	**0.895**
MaxPool CNN	Victim probes	97.3	100.0	100.0	16/16	24/186	5.36	0.197
	Student (DPE)	97.5	99.1	100.0	9/16	**19/101**	**9.04**	**0.649**

Accuracy and PF values are percentages. Student clone metrics use reused test inputs. Locate/Full Q gives median localization and complete attempt queries, including orientation and validation. Rule PF is agreement on candidate-dependent samples, not a global metric.

The hidden-pooling study asks whether local extraction requires a student with the victim's pooling structure (Fig. 4). Selection uses reference-verified victim-normal trials with a separate query allocation but shares input pools with extraction. Both AvgPool-victim seeds select MaxPool students yet accept 4/8 rules, matching the reference student with the victim's actual pooling type. On MaxPool victims, selection accepts 2/8 rules versus the reference's 3/8. Thus, mismatched pooling can support local extraction, although student selection affects acceptance. Together, these studies establish conditional online savings and extraction without exact pooling identification.

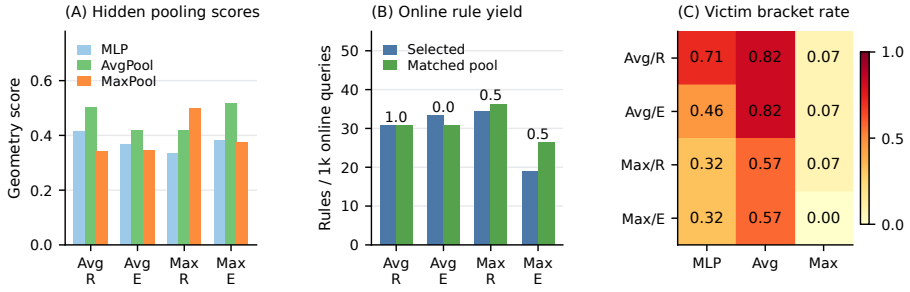


Fig. 4. Hidden-structure diagnostics: student geometry scores (A), selected versus matched-pool student rule yield (B), and victim bracket rate in pooling-switch probes (C). R/E denote Random KD/event-repeat-4 training, and annotations in B report pooling-type selection accuracy, whereas C measures opposite-label bracketing rather than rule acceptance. Panels use related but separate studies, with selection and recovery receiving separate query allocations after reference-assisted seed screening and with shared input pools.

Student preparation consumes victim calls, so we evaluate complete DPE on Fashion-MNIST CNNs with learned Patch, Blended, and Content/mask watermarks. Victims have 16/32 convolutional channels, 3×3 kernels, and a 96-unit hidden layer. Students have 12/24 channels, $5 \times 5/3 \times 3$ kernels, and 64 hidden units. Both apply ReLU and 2×2 max pooling after each convolution. Each family uses three victim seeds and five caps from 20k to 320k, with independent preparation and student training at every cap. The 20k/40k/80k/160k/320k caps target 8/16/32/64/128 charts, with 160k/320k registered after inspecting lower budgets. Setup, extraction, and confirmation use disjoint pools of 9,000, 5,000, and 4,000 bases.

Chart directions and Direct normals each use 128 paired probes, with probe, orientation, validation, and synthesis scales all set to 0.012. Each chart supplies 192 synthesized rows, each included four times in the training set. Students train once for 30 epochs using Adam, batch size 256, and weight decay 0.0001, retaining the final checkpoint. A cosine schedule decreases the learning rate from 0.001 to 0.00001. The attacker receives unlabeled inputs and a public transformation

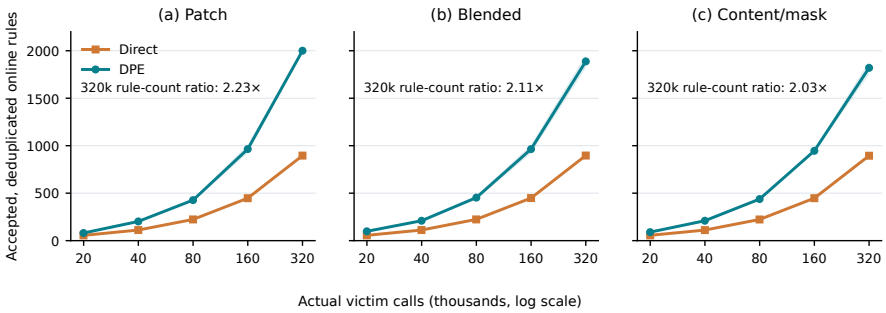


Fig. 5. Complete hard-label extraction on trained behavioral-watermark victims. Actual calls include preparation and failures, and rule counts exclude setup charts. Curves and bands are three-seed means and sample standard deviations. Each cap trains independently, with 160k/320k added after inspecting lower budgets. Annotations give mean rule-count ratios at 320k.

family, using only counted top-1 victim responses. Direct uses its own endpoint stream and paired victim-normal probes. Both methods pay preparation and failed-attempt calls, while student computation and cached-label reuse remain separate.

Figure 5 shows more accepted, deduplicated online rules for DPE across every family and cap under the original directional gate. At 320k, DPE yields 2,000.3/ 1,887.7/1,819.7 rules for Patch/Blended/Content, versus Direct's 895.7/896.3/ 894.3, giving ratios of 2.23/2.11/2.03. Preparation consumes approximately 54k–57k calls. Including those calls, DPE yields 5.69–6.25 accepted rules per thousand total calls, versus approximately 2.80 for Direct. Counts exclude setup charts. Table 2 reports actual costs, directional acceptance PF, and independent ball PF. This compares complete pipelines, while the fixed-candidate study isolates direction-estimation savings.

Table 3 gives the complete-DPE preparation allowances at each independently trained budget. Stable-response discovery uses eight coarse bases and stride-three positions, then radius-one refinement around three shortlisted positions on eight different bases. Each shortlisted transformation is confirmed on 32 separate bases using inferred target responses and their advantage over paired clean predictions. Accepted transformations provide clean-to-transformed endpoint pairs, and natural public pairs remain available when constructing the boundary-event bank. Discovery labels are excluded from direct trigger supervision. Queried warm labels and chart pseudo-labels train each student's fixed final checkpoint.

For the three-watermark study, independent diagnostics freeze each accepted rule's anchor, normal, intercept, and class pair before drawing fresh ball probes. Exact collisions with previous attack inputs are excluded without reading labels, and third-class responses count as disagreement. The diagnostic labels never alter acceptance, the rule bank, or checkpoint selection. PF averages rules within each victim seed, then weights the three seeds equally, with sample standard

Table 2. Complete extraction at the 320k cap. Actual calls are seed means and include setup, repeated submissions, and failures. Counts and PF are means ± sample SD across three victims. Acceptance PF uses the original 80 directional gate samples. Ball PF uses 80 fresh uniform-ball samples at each retained rule's original radius and method-specific frozen anchor. All unique online rules contribute, excluding setup charts. Diagnostic calls are separate from attack cost.

Watermark	Method	Actual calls	Setup calls	Unique online rules	Acceptance PF (%)	Ball PF (%)
Patch	Direct	319,899.7	0.0	895.7 ± 0.6	99.91 ± 0.01	60.75 ± 0.19
	DPE	319,956.0	56,723.7	$2,000.3 \pm 17.6$	99.89 ± 0.06	63.10 ± 2.60
Blended	Direct	319,797.3	0.0	896.3 ± 0.6	99.84 ± 0.01	60.51 ± 0.21
	DPE	319,974.0	54,394.3	$1,887.7 \pm 68.3$	99.92 ± 0.03	66.56 ± 3.49
Content/mask	Direct	319,889.7	0.0	894.3 ± 1.2	99.87 ± 0.03	60.63 ± 0.17
	DPE	319,975.0	55,174.3	$1,819.7 \pm 69.2$	99.91 ± 0.02	63.45 ± 1.07

Table 3. Complete-DPE allocations. Call allowances include all victim-evaluated inputs and are upper bounds without filler traffic. Discovery is incremental, while setup includes warm labels, discovery, and chart construction.

Total cap (calls)	Warm calls	Discovery allowance	Setup cap	Chart target
20,000	2,000	4,000	10,000	8
40,000	4,000	8,000	20,000	16
80,000	8,000	16,000	40,000	32
160,000	8,000	32,000	80,000	64
320,000	8,000	64,000	160,000	128

deviations across seeds. Table 2 reports the original radius. Appendix B.2 specifies the half-radius and double-radius tests.

This study uses nine independently trained victims, 45 students, and 90 DPE/Direct runs across the five budgets. The combined attack ledger records 11,154,985 actual victim calls. The 12,268,800 post-freeze research-evaluation calls are recorded separately from attack budgets. Budgets within a victim seed share its owner and input pools but use separate preparation and training runs. The victim seed, rather than a rule or probe, remains the statistical unit.

Each proposal batch retains up to 600 student crossings after up to 24 offline bisection steps per endpoint pair. Batch limits are 24/48/96 at caps of 20k–80k/160k/320k. Victim search tests radii 0.015, 0.04, 0.09, 0.18, and 0.30, refining brackets to width 10^{-3} or at most 24 steps. Deduplication aligns matching class pairs and requires anchor distance at most $2a_{\text{val}}$ and normal cosine at least 0.995. Offline proposal generation and ranking add no victim calls.

Separate two-seed sessions on digits/AvgPool/MaxPool victims evaluate complete hard-label extraction with disjoint setup, extraction, and clean-evaluation input pools. Each student remains fixed across eight cumulative caps from 1k to

20k actual calls, including preparation and failed attempts. These development sessions differ from the independently trained Fashion-MNIST budgets above. Direct estimates directions through paired victim probes, while Random KD and DPE use student gradients. Random KD matches DPE's actual setup-call cost with natural victim labels, while training rows and offline computation differ.

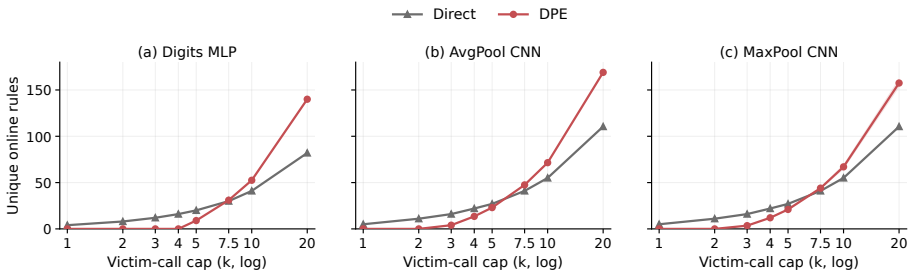


Fig. 6. Counted hard-label extraction on two development victim seeds per family. Curves and bands show seed means and sample standard deviations across eight cumulative attack caps. Counts include accepted, deduplicated online rules, excluding setup charts. Caps charge preparation and failures, without filler calls, and each student trains once per session. The accompanying table includes the matched-call distillation control and independent ball diagnostics.

Figure 6 shows DPE's preparation overhead at small caps and the first observed crossover beyond Direct at 7.5k in every paired victim. At 20k, DPE produces 140.0/169.0/157.5 rules, versus Direct's 82.0/110.5/110.5. Table 4 reports actual costs and fresh, normal-independent ball tests on every retained rule. The methods retain their original anchors, so these diagnostics compare complete rule banks rather than predictions on shared inputs. DPE and Random KD have equal AvgPool mean counts, and their CNN ball-PF differences are small. Appendix B.3 specifies the fixed-student protocol, multiscale tests, and source-record locations.

3.3 RQ2: Event Transfer Beyond Predictive Accuracy

The distillation controls test whether similar natural-input predictions imply similar local transition information (Fig. 7). DPE event assigns roughly one third of the allocation to natural labels and synthesizes 48 rows per retained event. DPE tuned assigns roughly half to natural labels, targets eight events, and synthesizes 96 rows per event. At the nominal 2,400-query setting, all five strategies exceed 97% natural PF, but reference-local PF ranges from 48.44% to 54.43%. DPE event verifies the most boundary events (10/12). DPE tuned verifies 8/12, attains higher reference-local PF than DPE event, and uses 19.1% fewer recorded training queries than Random KD. These differences motivate evaluating transitions separately from natural-input agreement, with additional settings in Appendix B.1.

Table 4. Budget-study outcomes at the configured 20k attack cap. Actual and setup calls are seed means, with setup included in actual calls. Rule counts and normal-independent ball PF/BA are means ± sample SD across two victims. PF/BA use the original validation radius and method-specific original anchors. Diagnostic calls are charged separately.

Victim	Method	Actual calls	Setup calls	Online rules	Ball PF (%)	Ball BA (%)
Digits MLP	Direct	19,939.0	0.0	82.0 ± 0.0	72.44 ± 0.46	72.78 ± 0.41
	Random KD	19,998.5	3,958.0	136.0 ± 2.8	75.42 ± 2.09	75.96 ± 2.21
	DPE	19,934.5	3,958.0	140.0 ± 1.4	89.27 ± 0.10	89.22 ± 0.15
AvgPool CNN	Direct	19,979.0	0.0	110.5 ± 0.7	56.40 ± 0.29	56.70 ± 0.30
	Random KD	19,938.5	2,568.0	169.0 ± 4.2	85.22 ± 2.58	86.65 ± 2.73
	DPE	19,976.0	2,568.0	169.0 ± 0.0	85.90 ± 1.37	87.12 ± 1.68
MaxPool CNN	Direct	19,957.0	0.0	110.5 ± 0.7	56.73 ± 0.16	56.82 ± 0.05
	Random KD	19,946.0	2,563.5	143.5 ± 9.2	69.82 ± 5.08	70.16 ± 5.27
	DPE	19,976.5	2,563.5	157.5 ± 4.9	70.47 ± 4.71	71.04 ± 4.94

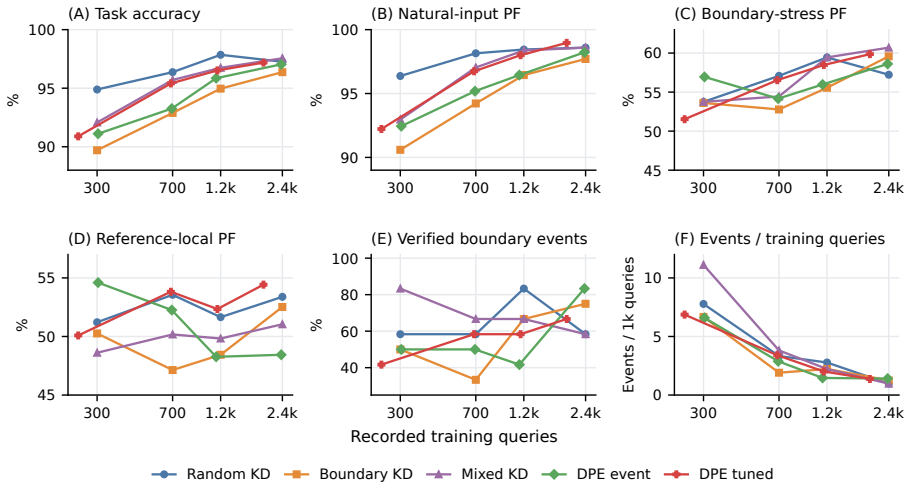


Fig. 7. Comparison of all five digits training strategies, with A–D reporting three-seed mean Accuracy/PF and E pooling 12 boundary-event attempts per setting. F divides verified events by summed recorded training queries, excluding candidate verification calls, while horizontal axes use mean recorded training queries. D evaluates student predictions on a separate reference-rule bank after reference-assisted seed screening and with reused inputs, unlike candidate-dependent rule PF (Table 1).

The historical ResNet study tests discovery and transfer of a controlled conditional response under hard-label access (Table 5). The archived exact-8x setting uses torchvision ResNet-34/18 with stage widths 64/128/256/512 and

a stride-one 3×3 stem without initial max pooling. CelebA-8 labels encode the eight combinations of Male, Mouth_Slightly_Open, and Smiling, while CIFAR-10 uses its ten classes. A deterministic wrapper returns class 0 when every coordinate in a designated bottom-right patch is at least 0.97, retaining base predictions otherwise. The patch-event variant trains students on the selected transformation's observed majority response, rather than boundary-chart pseudo-labels. Trigger PF measures student–wrapped-victim agreement on patched inputs. This tests conditional-response transfer at image scale, while Rule PF in Table 1 measures extracted halfspaces and the subsequent watermark study uses trained victims.

Methods share checkpoint hashes, preprocessing, student architecture, seeds, and evaluation subset protocols within each block. DPE uses natural training-split inputs, while local baseline adaptations use generated inputs. The image-scale budget counts discovery, training exposures, and evaluation, as specified in Eq. (26). With zero-jitter prefetching, cached exposures differ from actual calls and the initial label batch can exceed a small allocation. These historical configured allocations describe transfer, with actual-call cost comparisons provided by the fixed-candidate and complete-session studies.

Across the 14 DPE settings in Table 8, trigger PF ranges from 99.33% to 100.00%. Preserving patch responses can nevertheless trade off against task accuracy. At CelebA-8 1K, DPE achieves 100% trigger PF alongside 12.50% Accuracy, compared with BAM-HL's 21.88% Accuracy (Appendix B.1). Event-only controls show the same trade-off, so both metrics are needed to assess the usefulness of transferred events.

Table 5 includes three-seed results at the configured 1M allocation. DPE achieves 81.71%/77.69% clean PF and 99.64%/99.76% Trigger PF on CelebA-8/CIFAR-10 at 1M. Test subsets and seed counts differ across blocks, so their Trigger PF differences do not isolate the effect of budget.

CelebA-8 shares one victim checkpoint across seeds but regenerates its balanced data partition for each attack seed. An index-only reconstruction places 878/886 of 2,048 seed-1/2 evaluation inputs in the seed-0 victim-training partition. Historical training IDs were not retained, so this overlap is reconstructed from the stored protocol rather than directly observed ID records. The CelebA-8 means characterize this transfer protocol, with strictly held-out generalization requiring a fixed disjoint partition. CIFAR-10 uses separate victim checkpoints and the official test split. Appendix B.1 records the counter convention for both datasets.

Behavioral watermarks support model ownership verification through responses to selected inputs [2], motivating a test of whether those responses survive extraction into a student. We train CNN victims with learned Patch, Blended, and Content/mask watermarks and apply complete DPE, adapting the input transformation families described in [22]. The attacker receives a public candidate-template library. Target labels, registered bases, and localized watermark positions remain private and are used only for final evaluation. Patch registers four responses, whereas Blended and Content/mask each register one, with registered bases and 3,000 independent test bases evaluated under separate

Table 5. Historical ResNet conditional-response transfer at configured budgets. Checkpoints, preprocessing, and evaluation subset protocols match within blocks, while actual-call budgets and auxiliary inputs differ across methods. Base/Behav. V. Acc. measure base/wrapped victim accuracy. Higher clone/PF values are better, and bold ranks displayed values under these configurations.

Method	Base V. Acc.	Behav. V. Acc.	Clone Acc.	Clean PF	Trigger PF
10K configured budget, CelebA-8					
BAM-HL	77.34	77.15	14.06	11.91	0.00
DFMS-SL-HL	77.34	77.15	12.70	11.91	26.95
DB-DFMS-HL	77.34	77.15	12.50	12.30	39.26
DPE	77.34	77.15	**32.62**	**34.96**	**98.83**
10K configured budget, CIFAR-10					
BAM-HL	74.61	73.05	6.25	6.05	2.15
DFMS-SL-HL	74.61	73.05	16.41	17.38	13.48
DB-DFMS-HL	74.61	73.05	10.16	9.77	0.98
DPE	74.61	73.05	**34.38**	**37.11**	**100.00**
100K configured budget, CelebA-8					
DFMS-SL-HL	79.30	78.81	11.72	12.70	98.93
DB-DFMS-HL	79.30	78.81	17.19	17.19	59.08
DPE	79.30	78.81	**73.05**	**79.20**	**100.00**
100K configured budget, CIFAR-10					
DFMS-SL-HL	75.98	74.71	10.55	9.47	0.00
DB-DFMS-HL	75.98	74.71	14.65	14.26	1.17
DPE	75.98	74.71	**64.55**	**69.82**	**100.00**
1M configured budget, CelebA-8†					
DFMS-SL-HL	85.35	85.07	12.48 ± 0.24	12.04 ± 2.13	0.36 ± 0.31
DB-DFMS-HL	85.35	85.07	27.67 ± 9.28	28.52 ± 10.42	47.31 ± 33.32
DPE	85.35	85.07	$\mathbf{78.06} \pm 0.39$	$\mathbf{81.71} \pm 0.49$	$\mathbf{99.64} \pm 0.54$
1M configured budget, CIFAR-10†					
DFMS-SL-HL	79.07	77.85	22.90 ± 5.81	23.52 ± 6.65	2.91 ± 3.40
DB-DFMS-HL	79.09	77.86	28.96 ± 3.83	29.39 ± 5.15	3.76 ± 1.63
DPE	79.09	77.86	$\mathbf{78.87} \pm 2.20$	$\mathbf{77.69} \pm 1.28$	$\mathbf{99.76} \pm 0.27$

Values are percentages. The 10K/100K blocks use seed 0 and 512/1,024 evaluation inputs. The 1M rows (†) report three-seed means ± sample standard deviations, with 2,048 evaluation inputs per seed. CIFAR-10 1M has a one-example difference in seed-0 victim accuracy across source records.

protocols. The independent bases are excluded from this study's training, search, and acceptance stages, while prior use of this historical dataset is disclosed in Appendix B.2.

Each owner retains 36,000 clean rows. Patch augments 700 distinct non-target bases per response with ten marked exposures per base. Blended and Content/mask each use 2,000 marked bases with sixteen exposures. Registered

keys contain 100 bases per Patch response or 300 otherwise, drawn from the marked owner-training subset. Local marked inputs use one-pixel coordinate jitter and amplitudes in $[0.8, 1]$. Blended mixture strengths range from 0.30 to 0.40. Owners train for eight epochs using Adam, learning rate 0.001, weight decay 0.0001, and batch size 256, retaining the final checkpoint.

WSR measures target-label response rate, and WA measures its increase over the paired clean bases. Joint verification requires every registered response to achieve WSR of at least 90% and WA of at least 80 percentage points on the specified verification distribution.

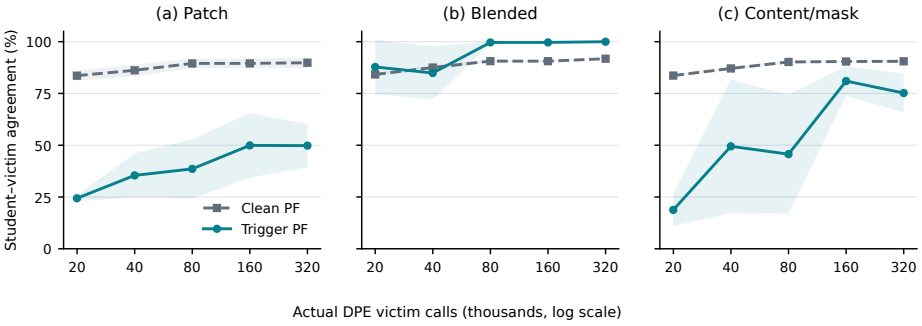


Fig. 8. Complete-DPE agreement on independent test bases. Gray dashed curves show clean PF, teal curves show Trigger PF, and bands show three-seed sample standard deviations. Responses average within seeds for Patch. All panels share the vertical scale. Actual calls include preparation, and each cap trains independently.

Figure 8 reports clean PF alongside Trigger PF, allowing conditional-response preservation to be assessed separately from agreement on untransformed inputs under the same budgets. Blended reaches 99.63% Trigger PF at 80k calls and passes joint verification for all three seeds from that budget onward in both evaluation distributions (Table 9). At 320k, its clean Accuracy, clean PF, and Trigger PF are 85.29%, 91.79%, and 99.99%, with seed variation and verification statistics reported in Table 6. Patch retention remains near 50% at the higher budgets, while Content/mask reaches 80.99% at 160k and 75.21% at 320k on the independent bases. Neither family passes joint verification at the evaluated budgets, so stable watermark retention in this configuration is specific to Blended rather than shared across transformation families.

This application tests whether extraction preserves student responses used in ownership verification. The rule bank has zero measured coverage of the common trigger-region inputs, so its local halfspaces do not explain this observed retention. Appendix B.2 reports trigger-region coverage separately from student behavior preservation and local halfspace acceptance.

Earlier single-patch and two-key studies use patch-event extensions with different training protocols, summarized in Appendix C.1.

Table 6. Student utility and behavioral-watermark verification at the 320k cap. Values are means with sample standard deviations over three victim seeds on independent bases. Accuracy, PF, and WSR are percentages, while WA is measured in percentage points. Joint verification counts seeds satisfying both thresholds for every response. Registered and independent bases give the same pass counts, and all victims pass in 3/3 seeds on both distributions. Patch has four registered responses, whereas the other families each have one.

Watermark	Clean Acc.	Clean PF	Trigger PF	WSR	WA	Joint
Patch	83.59 ±1.69	89.83 ±2.53	49.81 ±10.36	48.46 ±11.34	37.37 ±11.08	0/3
Blended	85.29 ±0.69	91.79 ±0.05	99.99 ±0.02	99.99 ±0.02	90.57 ±0.29	3/3
Content/mask	84.34 ±0.76	90.52 ±0.80	75.21 ±9.40	75.16 ±9.37	63.78 ±9.55	0/3

Published extraction results provide context for clone accuracy beyond these mechanism studies, with source records indexed in Appendix B.1. DPE's Accuracy exceeds BAM's on Fashion-MNIST/CIFAR-10 but remains lower on Monkey. In the six-architecture Label Diversity comparison, DPE leads displayed Accuracy/PF on WRN-32 and VGG-16BN, while other methods lead elsewhere (Appendix B.1). Checkpoints, inputs, interfaces, and accounting differ. Some local victims use a compact three-stage implementation with 32/64/128 channels. These descriptive rankings supplement the transfer evaluation without establishing comparative local-rule recovery.

3.4 Deployment Conditions

Across the loopback tests, 8,934 HTTP requests produce 3,000 model calls. HTTP 429 rejections incur no model calls. Three interleaved fillers per query reduce DPE detection from 3/3 to 0/3 sessions in trace replay under a length-calibrated geometric monitor. This quadruples queries and lowers trace yield from 7.80 to 1.95, with additional controls and interface-specific costs reported in Appendix C.2.

4 Related Work

4.1 Cryptanalytic Extraction

Cryptanalytic attacks use chosen inputs to recover ReLU parameters and signs [6, 10]. Hard-label variants analyze decision and activation boundaries to recover selected deep layers [7, 9, 12, 25]. Extensions address pooling and other activations under model-specific conditions [3, 13, 14, 31].

Shen et al. [33] use numerical-output traces to infer ReLU MLP depth and widths. Li and Wang [24] detect missed neurons during primarily raw-output

extraction with limited architectural knowledge. Liu et al. [26] infer convolution and pooling structure from recovered vectors in raw-output and hard-label pipelines. Their end-to-end tests use two-convolution CNNs, while a varied three-convolution diagnostic uses numerical outputs. Li et al. [22] recover hidden CNN parameters under hard-label access. DPE instead returns a student and local class-pair rules, using victim labels to refine and validate student proposals without reconstructing the network. Appendix B.1 indexes interface and output records without ranking unmatched query counts.

4.2 Distillation-Based Extraction

Learning-based extraction trains substitutes on victim-labeled inputs [20, 30, 35], with Copycat CNN and Knockoff Nets obtaining supervision by querying external images [16,29]. DeGAN uses proxy data and teacher gradients, whereas Black-Box Ripper searches without teacher gradients and DFME estimates zeroth-order gradients from victim–student disagreement [1, 4, 36]. DFMS-HL jointly trains a generator and student using proxy or synthetic priors, student-gradient guidance, and victim hard-label supervision, unlike soft-label DFMS-SL [32]. DB-DFMS simplifies generation around class diversity [27], while BAM samples evolutionarily near low-confidence boundaries [5], both training substitutes to predict victim labels. Original DFME, DFMS-SL, DB-DFMS, and BAM require probability outputs, whereas DPE tests student-gradient proposals through fresh victim hard-label validation of local rules.

5 Conclusion

DPE combines reusable student directions with victim-side validation to extract local class-pair rules without identifying the victim architecture. On the digits MLP, fixed-candidate median online cost falls from 246 to 118 queries per attempt, with both normal sources accepting 17/24 rules. The pooling CNNs show analogous matched-attempt savings. Full Fashion-MNIST sessions count preparation and failures, producing 2.03–2.23 times Direct's accepted-rule count at the 320k caps, with independent ball PF reported separately. The watermark application evaluates ownership-verification responses of the trained student. Blended passes joint verification in every seed from 80k onward, while Patch and Content/mask follow different retention curves. These results establish preparation-inclusive yield gains under the evaluated gate and identify which victim watermark responses survive transfer to a student.

Ethics Statement

Model extraction can enable unauthorized copying and excess service load, so external testing requires authorization from the relevant system owners. We evaluate only local benchmark models and a loopback HTTP service under our control, without querying commercial APIs or private user data. The authors declare that they have no conflicts of interest to disclose.

Reproducibility Statement

Sections 2.1–2.3 and Algorithm 1 specify the method. Appendices A and D give implementation details, the query-saving proof, and recovery conditions. Appendices B and C document protocols and results. The code artifact contains source, configurations, recorded results, and environment requirements. Reproduction also requires the datasets and weights listed in its README.

Acknowledgment

This work was supported by the National Cyber Security–National Science and Technology Major Project under Grants 2026ZD1500600 and 2026ZD1500601, and by the National Natural Science Foundation of China under Grants 62672189 and U2336209.

OpenAI Codex assisted with literature checks, translation, result interpretation, equation and figure checks, LaTeX preparation, drafting, and the implementation and execution of the counted local-rule, trained-backdoor, and behavioral-watermark experiments. The authors are responsible for verifying their code, measurements, and final presentation.

References

1. Addepalli, S., Nayak, G.K., Chakraborty, A., Babu, R.V.: DeGAN: Data-enriching GAN for retrieving representative samples from a trained classifier. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3130–3137 (2020). https://doi.org/10.1609/aaai.v34i04.5709, https://doi.org/10.1609/aaai.v34i04.5709
2. Adi, Y., Baum, C., Cissé, M., Pinkas, B., Keshet, J.: Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: Proceedings of the 27th USENIX Security Symposium. pp. 1615–1631. USENIX Association (2018)
3. Asselineau, R., Derbez, P., Fouque, P.A., Minaud, B.: Cryptanalytic extraction of deep neural networks with non-linear activations. Cryptology ePrint Archive, Paper 2026/253 (2026), https://eprint.iacr.org/2026/253
4. Bărbălău, A., Cosma, A., Ionescu, R.T., Popescu, M.: Black-box ripper: Copying black-box models using generative evolutionary algorithms. In: Advances in Neural Information Processing Systems (2020), https://arxiv.org/abs/2010.11158
5. Biton Dor, M., Mirsky, Y.: Efficient model extraction via boundary sampling. In: Proceedings of the 2024 Workshop on Artificial Intelligence and Security. pp. 1–11. ACM (2024). https://doi.org/10.1145/3689932.3694756, https://doi.org/10.1145/3689932.3694756
6. Canales-Martínez, I.A., Chávez-Saab, J., Hambitzer, A., Rodríguez-Henríquez, F., Satpute, N., Shamir, A.: Polynomial time cryptanalytic extraction of neural network models. In: Advances in Cryptology – EUROCRYPT 2024. pp. 3–33 (2024). https://doi.org/10.1007/978-3-031-58734-4_1, https://doi.org/10.1007/978-3-031-58734-4_1

7. Canales-Martínez, I.A., Santos, D.: Extracting some layers of deep neural networks in the hard-label setting. In: Progress in Cryptology – LATINCRYPT 2025. pp. 399–421 (2025). https://doi.org/10.1007/978-3-032-06754-8_15, https://doi.org/10.1007/978-3-032-06754-8_15
8. Carlini, N., Chávez-Saab, J., Hambitzer, A., Rodríguez-Henríquez, F., Shamir, A.: Polynomial time cryptanalytic extraction of deep neural networks in the hard-label setting (extended version). Cryptology ePrint Archive, Paper 2024/1580 (2024), https://eprint.iacr.org/2024/1580, revised 7 September 2026; version consulted 16 September 2026
9. 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. In: Advances in Cryptology – EUROCRYPT (2025). https://doi.org/10.1007/978-3-031-91107-1_13, https://eprint.iacr.org/2024/1580
10. Carlini, N., Jagielski, M., Mironov, I.: Cryptanalytic extraction of neural network models. In: Advances in Cryptology – CRYPTO (2020). https://doi.org/10.1007/978-3-030-56877-1_7, https://arxiv.org/abs/2003.04884
11. Chen, S., Carlini, N., Wagner, D.: Stateful detection of black-box adversarial attacks. In: Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence. pp. 30–39 (2020). https://doi.org/10.1145/3385003.3410925, https://doi.org/10.1145/3385003.3410925
12. Chen, Y., Dong, X., Guo, J., Shen, Y., Wang, A., Wang, X.: Hard-label cryptanalytic extraction of neural network models. In: Advances in Cryptology – ASIACRYPT (2024), https://eprint.iacr.org/2024/1403
13. Chen, Z., Tang, S., Gao, Z., Su, Y., Qin, L., Dong, X.: Algebraic attack on convolutional neural networks with max pooling. In: Advances in Cryptology – CRYPTO 2026. pp. 3–35 (2026). https://doi.org/10.1007/978-3-032-35415-0_1, https://doi.org/10.1007/978-3-032-35415-0_1
14. Chen, Z., Tang, S., Gao, Z., Su, Y., Qin, L., Dong, X.: Algebraic cryptanalytic extraction on hard-label neural networks. Cryptology ePrint Archive, Paper 2026/1164 (2026), https://eprint.iacr.org/2026/1164
15. 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. IACR Transactions on Cryptographic Hardware and Embedded Systems **2026**(3), 646–678 (2026). https://doi.org/10.46586/tches.v2026.i3.646-678, https://doi.org/10.46586/tches.v2026.i3.646-678
16. Correia-Silva, J., Berriel, R.F., Badue, C., de Souza, A.F., Oliveira-Santos, T.: Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data. In: International Joint Conference on Neural Networks (2018)
17. Feng, R., Hooda, A., Mangaokar, N., Fawaz, K., Jha, S., Prakash, A.: Stateful defenses for machine learning models are not yet secure against black-box attacks. In: ACM SIGSAC Conference on Computer and Communications Security (2023). https://doi.org/10.1145/3576915.3623116
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015), https://arxiv.org/abs/1503.02531
19. Ito, A., Miura, T., Todo, Y.: Is the hard-label cryptanalytic model extraction really polynomial? In: Advances in Cryptology – CRYPTO 2026. pp. 67–98 (2026). https://doi.org/10.1007/978-3-032-35415-0_3, https://doi.org/10.1007/978-3-032-35415-0_3
20. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: USENIX Security Symposium (2020)

21. Juuti, M., Szyller, S., Marchal, S., Asokan, N.: PRADA: Protecting against DNN model stealing attacks. In: IEEE European Symposium on Security and Privacy. pp. 512–527 (2019). https://doi.org/10.1109/EuroSP.2019.00044
22. 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
23. Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., Zhao, B.Y.: Blacklight: Scalable defense for neural networks against query-based black-box attacks. In: USENIX Security Symposium (2022)
24. Li, Y., Wang, G.: Polynomial time cryptanalytic extraction of deep neural networks in the limited architecture knowledge setting. Cryptology ePrint Archive, Paper 2026/2127 (2026)
25. Liu, H., Siproudhis, A., Experton, S., Lorenz, P., Boura, C., Peyrin, T.: Navigating the deep: End-to-end extraction on deep neural networks. In: Advances in Cryptology – EUROCRYPT 2026. pp. 482–512 (2026). https://doi.org/10.1007/978-3-032-25333-0_17, https://doi.org/10.1007/978-3-032-25333-0_17
26. Liu, J., Ma, R., Li, M., Chen, Y., Chen, S.: Extracting CNNs in the unknown-architecture and feedback-agnostic setting. Cryptology ePrint Archive, Paper 2026/2161 (2026)
27. Liu, Y., Wen, R., Backes, M., Zhang, Y.: Efficient data-free model stealing with label diversity (2024), https://arxiv.org/abs/2404.00108
28. Ma, R., Chen, Y., Zhang, J., Yu, H., Wang, X.: Breaking slope and structure restrictions: Broadening hard-label cryptanalytic extraction of prelu neural networks. Cryptology ePrint Archive, Paper 2026/1066 (2026), https://eprint.iacr.org/2026/1066
29. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
30. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: ACM Asia Conference on Computer and Communications Security (2017)
31. Qi, X., Lei, H., Wei, L., Sun, X., Wang, M.: Cryptanalytic extraction of neural networks with various activation functions. IACR Transactions on Symmetric Cryptology **2026**(1), 468–505 (2026). https://doi.org/10.46586/tosc.v2026.i1.468-505, https://doi.org/10.46586/tosc.v2026.i1.468-505
32. Sanyal, S., Addepalli, S., Babu, R.V.: Towards data-free model stealing in a hard label setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15284–15293 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Sanyal_Towards_Data-Free_Model_Stealing_in_a_Hard_Label_Setting_CVPR_2022_paper.html
33. Shen, Y., Chen, Y., Wang, A., Yu, H., Wang, X.: Cryptanalytic extraction of neural networks without known architecture assumption. Cryptology ePrint Archive, Paper 2026/2095 (2026)
34. Sun, X., Lei, H., Wei, L., Qi, X., Hu, K., Wang, M., Wang, W.: Cryptanalytic extraction of convolutional neural networks. In: Information Security and Privacy – ACISP 2026. pp. 123–147 (2026). https://doi.org/10.1007/978-981-92-3012-9_6, https://doi.org/10.1007/978-981-92-3012-9_6
35. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: USENIX Security Symposium (2016)

36. Truong, J.B., Maini, P., Walls, R.J., Papernot, N.: Data-free model extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4771–4780 (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Truong_Data-Free_Model_Extraction_CVPR_2021_paper.html
37. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y.: Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: Proceedings of the 2019 IEEE Symposium on Security and Privacy. pp. 707–723. IEEE (2019). https://doi.org/10.1109/SP.2019.00031
38. Zhang, W., Zhang, B., Lu, T., Ren, K.: Ppml is more vulnerable to cryptanalytic extraction attacks. Cryptology ePrint Archive, Paper 2026/848 (2026)

A Implementation and Local Guarantees

The main text defines the interface, three modules, objectives, and numerical scales. This appendix specifies implementation choices, study-specific protocols, and additional diagnostics. The input–output signature is

$$(g_{\theta^*}, \mathcal{E}, \mathcal{R}, \mathcal{T}) = \mathrm{DPE}(f, \mathbf{X}_0, \mathcal{P}, \mathcal{G}, Q_0, Q). \tag{19}$$

The inputs are defined in the Method section. Equation (19) returns the student, event bank, rules, and query transcript. Algorithms 1–2 give extraction and validation; Algorithm 3 below supplies their query wrapper. Failed or unfunded operations return $\perp$ with spent calls retained. Training-row concatenation preserves multiplicity.

A.1 Event Construction and Student Selection

FindBracket consumes proposals without replacement, reuses cached endpoint labels, and charges new labels to setup. Its fixed trial limit or exhausted budget returns $\perp$. BuildChart refines a distinct-class bracket, estimates and orients a seed direction, and applies the victim validation gate before retaining the chart. An empty bank permits warm-label-only training.

For the ResNet patch scan, candidate c is evaluated on m shared bases at intensities 0.90 and 1.00. Ranking uses

$$S_c = F_c + 1.5J_c + 0.25\Delta_c. \tag{20}$$

Where F_c is high-intensity forced-label consistency, J_c is the fraction switching only between intensities, and Δ_c is clean deviation. Equation (20) selects one patch within the fixed setup budget. Uncached scanning costs $m(1+2|\mathcal{P}|)$ calls, or $2m|\mathcal{P}|$ additional calls when clean labels are cached. This patch-response variant is distinct from full boundary-chart synthesis.

Pair-preserving bisection maintains the exact-arithmetic invariant

$$\begin{aligned} f(a_j^{(t)}) = c_j^- \neq c_j^+ = f(b_j^{(t)}), \\ \|a_j^{(t)} - b_j^{(t)}\|_2 \leq 2^{-t}\|a_j^{(0)} - b_j^{(0)}\|_2. \end{aligned} \tag{21}$$

Equation (21) requires retaining the adjacent class pair. Third classes cause rebracketing or rejection, and repeated floating-point midpoints stop refinement. The final midpoint locates a label transition. The seed estimator in Eq. (1) rejects an empty informative-pair set or zero estimate. All paired probes, two orientation calls, and the subsequent $2v$-point gate count toward Q_{seed}, including rejected chart attempts.

The chart projector expands in input coordinates as

$$H_j = \begin{bmatrix} 1 - n_{j,1}^2 & \cdots & -n_{j,1}n_{j,d} \\ \vdots & \ddots & \vdots \\ -n_{j,d}n_{j,1} & \cdots & 1 - n_{j,d}^2 \end{bmatrix}. \tag{22}$$

Where $n_{j,\ell}$ is coordinate ℓ of the unit seed normal. Equation (22) gives $H_j n_j = 0$ and $H_j^2 = H_j$. Extending the two orientation labels to chart points within radius $r_j > 0$ assumes

$$f(x_j(\alpha, z)) = \begin{cases} c_j^+, & \alpha \geq 0, \\ c_j^-, & \alpha < 0. \end{cases} \tag{23}$$

Equation (23) fixes the side convention at zero. Its neighborhood validity is a modeling condition, followed by victim testing of extracted proposals.

Lemma 1 (Tangent-coordinate invariance). *If clipping is inactive and* $\|x_j(\alpha, z) - p_j\|_2 \leq r_j$*, tangent perturbations do not alter the signed coordinate used in Eq.* (23)*.*

Proof. With clipping inactive, $x_j(\alpha, z) - p_j = \alpha n_j + \beta H_j z$. The unit-length condition gives $n_j^\top n_j = 1$, while symmetry of H_j and $H_j n_j = 0$ give $n_j^\top H_j = 0$. Taking the inner product therefore yields $n_j^\top (x_j(\alpha, z) - p_j) = \alpha$. Thus z changes local coverage but not the modeled side label. The radius condition keeps this statement within the assumed neighborhood. Projection at the input boundary can violate the inner-product identity, which is why DPE retains victim validation.

Synthesis uses the distributions in Eq. (3). Gaussian tangent offsets can leave the modeled neighborhood, and training pseudo-labels are not individually victim-verified. Numerical training minimizes Eq. (4) without a guaranteed global optimum. Row multiplicity sets the event weight, and binary students use one-logit cross-entropy. The main text proves normal consistency relative to the seed direction, whose accuracy depends on its victim probes.

In the hard-label specification, S_k^{ver} in Eq. (7) is the fraction of trial proposals passing the victim gate, with all scoring labels charged to setup. The original

hidden-pooling experiment instead uses reference-verified localizations with error below 10^{-3} and no failed refinement. Its recorded score averages rather than sums the three components, preserving their ranking. Section 3.2 states its reference assistance and input reuse.

The hidden-pooling selector measures boundary PF on 128 perturbed inputs from 64 opposite-label pairs, with reference bisection locating their centers. Local PF uses 48 fresh samples per retained seed chart, with uniform normal offsets and Gaussian tangent offsets at scale 0.012. Both compare student predictions with victim labels. The boundary set is shared across candidates, while local samples use candidate-specific random draws. Seed screening uses reference boundaries and gradients.

A.2 Proposal Refinement and Validation

Student bisection locates crossings without victim calls. Victim refinement searches increasing radii along the endpoint displacement, then bisects an opposite-label bracket. Class-margin gradients use the refined victim class pair. A binary student uses its signed logit gradient directly, with undefined or zero gradients rejected.

The validation wrapper reserves all $2v$ calls before sampling. Repeated sampled inputs retain multiplicity, and ordinary sample agreement is

$$\widehat{F}_i = \frac{1}{2v} \sum_{u \in \mathcal{U}_i} \mathbb{1}[h_i(u) = f(u)]. \tag{24}$$

Equation (24) supplies the empirical statistic in the article's fixed-candidate bound. The implemented rule-bank gate is

$$\mathcal{R} \leftarrow \mathcal{R} \cup \{\mathcal{C}_i\} \quad \text{only if} \quad V_i \geq 0.8 \ \wedge \ 0.2 \leq \pi_i \leq 0.8. \tag{25}$$

Equation (25) combines balanced accuracy and class balance, additionally rejecting third-class responses. The original controlled studies add reference boundary error below 10^{-3} and successful refinement. Counted hard-label sessions apply no reference filter. Validation records retain anchors, normals, class pairs, validation inputs/labels, and transcript positions.

Algorithm 3 counts each submitted victim input, so one batch containing n inputs costs n calls, including repeated submissions. Cached-label lookups bypass this wrapper. Student synthesis, optimization, and differentiation are offline operations. The wrapper checks the entire batch before inference and returns $\perp$ if the remaining budget cannot fund it. Calls spent earlier in a failed attempt remain charged. Each transcript entry records its input, returned label, role, and sequential position, allowing setup, extraction, and evaluation calls to be separated without changing the counting rule.

Equation (16) bounds one candidate fixed before validation under its specified sampling law, before conditioning on acceptance. The gate uses balanced accuracy rather than ordinary PF. The two statistics and their denominators remain separate.

Algorithm 3 Budgeted victim query and transcript append

function QUERY($f, \mathbf{X}, \mathsf{role}, q, B, \mathcal{T}$)
  $n \leftarrow |\mathbf{X}|$
  if $q + n > B$ **then**
    return $(\perp, q, \mathcal{T})$
  end if
  $\mathbf{Y} \leftarrow f(\mathbf{X})$
  for $i \leftarrow 1$ **to** n **do**
    $t_i \leftarrow 1 + |\mathcal{T}|$
    $\mathcal{T} \leftarrow \mathcal{T} \parallel (x_i, y_i, \mathsf{role}, t_i)$
  end for
  $q \leftarrow q + n$
  return $(\mathbf{Y}, q, \mathcal{T})$
end function

B Experimental Protocols and Independent Diagnostics

B.1 Original Controls and Image-Scale Settings

The original digits diagnostics use 8×8 images for even/odd classification, with three seeds at 300/700/1,200/2,400 recorded training queries. Pooling CNNs classify MNIST 3 versus 8. Local rule studies pool 24 digits attempts over three seeds and 16 attempts per pooling CNN over four seeds. Reference neighborhoods use 96/64 labels and balanced-accuracy thresholds 0.75/0.72 for digits/MNIST, with victim-gradient and reference-refinement assistance omitted from the seed counter. Student selection uses two seeds and a separate recovery allocation, with the input reuse stated in the main text.

The original local-study costs and Rule PF are medians over the available attempt records. Accepted-rule counts pool attempts rather than treating each rule as an independent model repeat. Consequently, the observed count and median attempt cost describe different aggregation units. The original selection trials screen boundary localizations before student-normal extraction, using the reference error and refinement criteria stated above. Their selection allocation is separate from the recovery allocation, while their input pools are not independently split.

The original setup counts are 3,590/1,979/1,972 for digits/AvgPool/MaxPool, with 17/11/9 observed rules. Their approximate break-even counts are 40/49/306 rules under observed success rates and median costs. These estimates omit assisted checks and student computation and are not measured total-cost crossovers. The article's counted budget sessions evaluate different victims with preparation included.

Table 7 separates these preparation costs from marginal attempt costs and success-rate-adjusted estimates. Its Direct column uses the original random-candidate control, rather than the fixed-candidate comparison in the abstract.

Table 7. Original controlled-study setup and amortization estimates, using median attempt costs and observed success rates. B/E denotes approximate break-even counts, excluding assisted reference checks and student computation.

Victim	Setup Q	Direct/ DPE Q	Attempt B/E	Rule B/E	Obs.
Digits MLP	3,590	257/118	26	40	17
AvgPool CNN	1,979	186/100	24	49	11
MaxPool CNN	1,972	186/101	24	306	9

Obs. gives the original observed accepted-rule counts. Estimates assume the recorded success rates and omit future changes or duplicate rules. They are not measured total-cost crossovers.

The exact-8x 1K/10K/100K/1M stages evaluate 64/512/1,024/2,048 examples, reserving three passes per example. Seeds are 0 for lower stages and 0–2 at 1M. Configured accounting is

$$Q_{\text{total}}^{\text{rep}} = Q_{\text{extract}}^{\text{rep}} + 3n_{\text{eval}}. \tag{26}$$

Where the two Q^{rep} quantities denote configured total/extraction counters and n_{eval} counts evaluation inputs. Equation (26) includes discovery and base/clean/trigger evaluation, with DPE's cached exposures distinct from actual calls.

At CelebA-8 1K, DPE achieves 12.50% Accuracy and 100% Trigger PF, versus BAM-HL's 21.88% Accuracy and 0% Trigger PF. The 14 historical DPE settings span 99.33–100% Trigger PF. At 2,400 training queries, the five digits strategies in Fig. 7 span 96.37–97.56% Accuracy and 7–10 verified events out of 12. DPE event/tuned accept 10/12 and 8/12, with reference-local PF 48.44%/54.43% versus Random KD's 53.39%. These event checks differ from full local-rule acceptance.

In CIFAR-10 1M, source records differ by one correct victim prediction among 2,048 seed-0 inputs despite matching checkpoint hashes and subset protocols. Individual prediction records are unavailable.

BAM settings configure 0.62M/0.82M/1.41M/1M queries, while Label Diversity uses 2M for SVHN and 20M otherwise. Historical comparisons use one seed and differing inputs, interfaces, and checkpoints. Settings marked effective-width2 use a compact three-stage implementation with 32/64/128 channels. The main text separately reports the historical exact-8x configurations. Historical matrices and source locators remain in the released records indexed by `code/EXPERIMENT_MAP.md`; the literature ledger is `code/experiments/literature_attack_comparison/source_ledger.md`.

Table 8 gives the full historical matrix, identifying the dataset, model combination, configured budget, and clean quality associated with each patch-transfer result.

Table 8. DPE classifier quality and patch-event preservation across 14 configured source-budget settings. Accuracy/PF are percentages. Every row identifies the injected patch exactly within its search family. These are patch-response measurements rather than full-chart extraction.

Setting	Victim → clone	Config. Q	Victim Acc.	DPE Acc.	Clean PF	Trigger PF
BAM						
Fashion-MNIST	11-CNN → ResNet-18	0.62M	86.02	86.40	94.60	99.98
CIFAR-10	AlexNet → ResNet-18	0.82M	78.49	78.62	78.05	99.91
Monkey	ResNet-18 → ResNet-18	1.41M	89.71	67.65	66.91	100.00
CIFAR-10	LeNet → ResNet-18	1M	62.11	62.22	73.70	100.00
CIFAR-10	ResNet-9 → ResNet-18	1M	72.30	71.60	76.17	99.99
CIFAR-10	DenseNet → ResNet-18	1M	72.10	67.15	77.72	99.93
Label Div.*						
CIFAR-10	ResNet-34 → ResNet-18	20M	82.61	81.25	81.70	99.65
SVHN	ResNet-34 → ResNet-18	2M	94.89	94.75	94.61	99.33
CIFAR-10	ResNet-34 → MobileNetV2	20M	82.61	73.42	73.66	99.89
CIFAR-10	ResNet-34 → DenseNet-121	20M	82.61	87.50	85.20	99.99
CIFAR-10	ResNet-34 → WRN-32	20M	82.61	84.15	83.80	100.00
CIFAR-10	ResNet-34 → ResNet-34	20M	82.61	83.44	83.74	99.95
CIFAR-10	ResNet-34 → VGG-16BN	20M	82.61	83.33	82.15	100.00
Label Div. exact						
CelebA-8	ResNet-34-8x → ResNet-18-8x	20M	79.90	79.90	84.25	100.00

*Label Diversity rows use effective-width-2 victims. CelebA-8 uses exact ResNet-34-8x/ResNet-18-8x models. Config. Q gives configured/reported counters subject to cached-label reuse. CelebA-8 at 20M is DPE-only, rather than a same-checkpoint baseline comparison.

B.2 Complete Behavioral-Watermark Extraction

The main text specifies the Fashion-MNIST models, owner training, disjoint 9,000/5,000/4,000 attack pools, discovery procedure, student training, and actual-call ledgers. The runs use seeds 6–8, with Patch/Blended/Content transformations adapted from [22]. Victim responses come from trained argmax predictions, without output replacement. The public library contains four Patch templates, 16 Blended pattern/strength combinations, and two Content masks. Registered

targets, bases, and localized positions remain evaluator-only. Blended is global, so its location is fixed.

The gate uses 80 directional points, class-balanced accuracy at least 0.8, both adjacent classes occupying at least 0.2, and no third class. Each chart contributes 192 synthesized rows, each included four times in the training set. Each cap independently prepares and trains its student once for 30 epochs, using a fixed cosine schedule and keeping the final checkpoint. Initialization, checkpoints, and replacement seeds are not selected using evaluation scores. Discovery labels do not directly supervise trigger responses. Direct uses its own endpoint stream and 128 paired victim-normal directions, with each direction costing two calls, without the DPE candidate bank. All caps use the same discovery and geometric policies. The 160k/320k caps are follow-up runs registered after inspecting lower budgets.

Students and rule banks are frozen before private-key evaluation. Registered owner-training bases and 3,000 official-test bases are evaluated separately. Present-run test bases are excluded from training, search, and acceptance, with earlier historical dataset use disclosed. Patch averages four responses within each seed before equal-weight seed aggregation. All nine victims pass joint verification under both laws, and student pass counts coincide between them at each cap.

Joint verification is applied to individual responses before reporting a family's pass count. A seed passes only when every response satisfies both owner thresholds on the selected verification distribution. For Patch, averaging four response scores summarizes retention but does not replace this all-response decision. Registered-base and independent-base decisions remain separate, even when their aggregate pass counts coincide. The source records retain each response's victim and student labels, clean predictions, and response-level verification outcomes, allowing the joint decision to be reconstructed from the underlying measurements.

Table 9 reports Trigger PF and joint verification for each watermark family across all five budgets. Aggregated utility, verification, and call statistics are in `code/experiments/three_behavioral_watermark_dpe/publication/`. Detailed frozen records reside under that study's `formal/` and `budget_extension/formal/` directories, with larger assets listed in `code/LOCAL_ASSETS.json`.

Table 9. Complete-DPE Trigger PF (percent) on independent bases across all five budgets. Each entry is a three-seed mean ± sample SD, followed by the joint-verification pass count in parentheses. Registered and independent pass counts coincide. Patch averages four responses within each seed, while the other families each have one. All victims pass in 3/3 seeds.

Budget	Patch	Blended	Content/mask
20k	24.38 ± 1.37 (0/3)	87.79 ± 13.13 (2/3)	18.74 ± 7.80 (0/3)
40k	35.43 ± 10.63 (0/3)	84.89 ± 12.84 (1/3)	49.48 ± 32.24 (0/3)
80k	38.58 ± 14.28 (0/3)	99.63 ± 0.55 (3/3)	45.69 ± 28.68 (0/3)
160k	49.91 ± 15.60 (0/3)	99.66 ± 0.60 (3/3)	80.99 ± 7.10 (0/3)
320k	49.81 ± 10.36 (0/3)	99.99 ± 0.02 (3/3)	75.21 ± 9.40 (0/3)

Every accepted, deduplicated online rule receives 80 fresh clipped uniform-ball points at scales $\kappa \in \{0.5, 1, 2\}$ around its frozen anchor. The pre-clipping law ignores its normal. Exact collisions with attack, training, acceptance, or previous diagnostics are excluded without labels. Third classes count as disagreement, and diagnostic labels never change rules. Table 2 shows scale one. Source records contain the remaining budgets and scales.

Common trigger-region tests use 128 separate bases per key, eight ball points at each clean-to-transformed anchor, and interpolation strengths $0, 0.25, 0.5, 0.75, 1$. Methods share test points for each owner. Configured rule balls define routing scope, with nearest-anchor routing and record-ID tie breaking independent of victim labels. Coverage divides routed points by all points, covered PF uses routed points, and correct coverage divides correct predictions by all points. Coverage and correct coverage are zero for every family/method/cap/strength, leaving covered PF undefined. This rule-bank test and student Trigger PF have different targets.

B.3 Counted Hard-Label Budget Protocol

Figure 6 and Table 4 use seeds 91004/91005 with matched victim checkpoints across Direct/Random KD/DPE. Setup/extraction/clean-evaluation pool sizes are 283/284/190 for digits and 600/600/300 for CNNs, with disjoint base IDs per seed. MNIST uses training inputs. Provisioned checkpoints make these development evaluations rather than new confirmatory samples.

Digits victims/students have widths 64/128. CNN victims use 8/16 channels and width-32 classifiers, while fixed AvgPool students use 12/24 channels and width 32 without pooling selection. Victims train for 40/six epochs and students once for 140/55 epochs. DPE warm calls are 900/1,000, chart targets 10/8, seed paired directions 96/48, and Direct online directions 64/40. Charts synthesize 48 rows, repeated once/four times at radius 0.012. Candidate banks draw 1,200/600 pairs and keep 10/6 crossings. Search radii are $0.012, 0.03, 0.07, 0.14$ for digits and $0.015, 0.04, 0.09, 0.18, 0.3$ for CNNs.

The eight cumulative caps are 1k/2k/3k/4k/5k/7.5k/10k/20k actual calls, including preparation, repeated submissions, and failures without fillers. Bisection stops at width 10^{-3} or numerical/trial limits. Gates use 96/80 points at radii 0.008/0.012 and operational deduplication follows the main text. Setup charts are excluded from online counts. Fixed students and saved diagnostics are reused across prefixes.

Random KD obtains natural victim labels at the same actual setup-call cost as DPE for that victim. This matches calls, while training-row composition and offline computation differ.

Each retained rule receives 128 independent ball probes per scale $\kappa \in \{0.5, 1, 2, 4\}$, freezing its original anchor, normal, intercept, and classes. Probes follow $x = \Pi_{[0,1]^d}(p + \kappa a_{\text{val}} U^{1/d} z/\|z\|_2)$ for independent $U \sim U[0,1]$ and $z \sim \mathcal{N}(0, I_d)$. The pre-clipping law is normal-independent, although clipping changes isotropy. All 2,436 unique rules contribute, with zero saved overlaps against acceptance and held-out inputs. PF averages rules within seed, BA uses rules

with both victim classes present, and seeds receive equal weight. Empty banks have zero count and undefined PF/BA. One AvgPool seed favors Random KD in ball PF.

The cumulative caps represent prefixes of a complete session, rather than eight independently trained students. A rule enters a prefix only after its refinement, orientation, and full gate have completed within that cap. Diagnostic labels do not fund extraction or alter previously retained records. The independent tests use each method's original anchors, whereas the separate common-anchor trials hold locations fixed. These designs answer different questions: the former evaluates the retained rule bank, and the latter isolates differences between proposed directions on shared neighborhoods.

The 18-session attack ledger totals 359,290 calls. Separate evaluation uses 421,632 held-out Rule PF calls, 4,740 clean calls, 1,247,232 original-rule ball calls, and 67,475 common-anchor calls. Full curves, per-seed denominators, clipping, and source hashes are in `paper/experiments/query_budget_20261002/`. Shared-anchor trials isolate the normal module rather than the complete pipelines.

B.4 DPE Extension Diagnostics

One Blended seed-6 victim and its frozen 80k student supply eight record-ID-ordered anchors. Adjustments add multidirectional events, average three students' unit gradients, or relocate intercepts with the original normal. Each anchor supplies 64 tangential directions at alternating offsets 0.003/0.006. Three initializations share events and cached rows, training 30 epochs with cached cross-entropy and side/boundary-margin losses.

Tangential event observations use hard-label bisection and labels on both sides of each located transition. The single-student variant uses the first trained member, while consensus averages three unit margin gradients for the same ordered class pair. Intercept correction uses the negative median of boundary displacements projected onto the original normal, preserving that normal and

Table 10. Eight-anchor development diagnostics on one Blended victim. Directional columns are equal-weight anchor means, with final probes separate from acceptance. Ball columns use radius 0.012 and show means ± sample SD across anchors, not model-seed variation. Every variant passes 8/8 original directional gates and 0/8 independent ball gates. BA averages the two adjacent-class recalls.

Variant	Directional acceptance PF (%)	Directional final PF (%)	Independent ball final PF (%)	Independent ball final BA (%)
Original DPE	100.00	99.98	67.97 ± 6.61	65.07 ± 4.40
Multidirectional events	100.00	99.90	63.96 ± 3.12	62.35 ± 2.80
Three-student consensus	100.00	100.00	64.87 ± 4.28	63.87 ± 2.92
Fixed-normal intercept	99.84	100.00	69.85 ± 7.35	63.96 ± 4.61

the support center. These adjustments test different components of DPE, with the same original anchors and independent development probes. Because the new single student changes its initialization and objective alongside its events, its comparison with the original student does not isolate event diversity alone.

Development tests share 256 normal-independent ball points per anchor at radii 0.012/0.006/0.003/0.0015. A 95% Wilson PF lower bound and the average of two Bonferroni-adjusted recall bounds must both reach 0.8. Each class needs 51 responses, with no third class. The largest passing radius would be frozen before independent final testing. Frozen candidates receive 512 fresh shared ball probes per anchor and radius, followed by separate 512-point tests under each candidate's directional law. No variant passes any independent gate. None of the 128 method/anchor/radius combinations achieves both PF and BA point estimates of at least 0.8. No method or radius is selected, and final scores never filter the rule bank.

Table 10 reports anchor variation, not model-repeat uncertainty. Original directional gates pass 8/8 for every variant, while independent gates pass 0/8. Directional banks cover 33.33% of 9,984 probes around 39 shared public boundaries; independently accepted banks are empty. Those public bases overlap search. These diagnostics use 18,849 additional victim calls (8,033 events,8,192 shared ball gates,64 orientation,2,560 directional gates), plus 42,752 post-freeze evaluation calls. Historical preparation/anchor acquisition add 20,066/8,000 calls separately. Shared labels are not independent per-method charges. Full records remain in `code/experiments/three_behavioral_watermark_dpe/reliability_pilot/dpe_extensions/`.

C Earlier Application and Deployment Controls

C.1 Earlier Patch-Event Applications

These studies use supervised patch-event rows rather than complete boundary-chart synthesis. The single-patch victims train eight epochs on 40,000 Fashion-MNIST bases with a white 5×5 patch on 15% of non-target examples relabeled class 0. Students train five epochs. Attack/confirmation pools have 16,000/4,000 bases, and evaluation uses 1,500 fixed official-test images excluded from earlier preliminary evaluations and the 10k-query evaluation. Earlier 4,000-image evaluations had only seed-specific exclusions.

The attack knows patch size/value but not location/target. Its 576-position scan costs 4,616 calls, followed by 408 confirmation calls. All queried labels train the student, with discovery weight 0.2 and one event row per five later calls. Random KD, random-patch KD, DPE, and known-trigger controls receive matched 10k/20k/40k actual calls, with lower-budget prefixes of 40k streams. Evaluation adds 3,000 calls per method/budget. DPE Trigger PF is 99.29/99.20/99.58%, while clean PF is 82.13/88.04/90.89%. At 40k, random-patch KD has 92.07% clean PF and 50.38% Trigger PF. Complete variability and pair/switch/source metrics remain in `code/experiments/trained_fashion_backdoor_budget_curve/summary.csv`.

The two-key victim embeds a horizontal 2×10 bar targeting 7 and a vertical 10×2 bar targeting 3 at secret positions. Each key has 2,000 non-target training bases with four marked exposures, giving 56,000 total rows and a matched unmarked control. Discovery costs 4,516 calls within 20k/40k caps. Five-epoch students share architecture, while budgets use nonnested streams. Verification uses 1,500 held-out bases, with 1,157 shared with the earlier backdoor test. Owner thresholds match the main text. DPE jointly passes 0/3 and 1/3 seeds at 20k/40k and discovers both true positions/targets in 0/3 at both caps. Known-key controls pass 3/3, while natural/random-bar controls pass 0/3. Owners pass 3/3, unmarked controls 0/3, and none of 24 wrong-position pairs or three swapped-target pairs passes. Records are in `code/experiments/fashion_two_key_behavioral_watermark/summary.csv`.

C.2 Query Visibility and Interface Conditions

Loopback profiles use a 20-query/s token bucket with burst four. Paced and four-worker clients each complete 500 requests per seed 0–2, totaling 8,934 HTTP attempts and 3,000 model calls. HTTP 429 responses occur before inference. Mean paced/burst times are $50.098 \pm 0.024/24.819 \pm 0.003$ seconds, with zero/1,978 rejections per seed.

The fixed transcript detector uses a maximum over window scores

$$M(\mathcal{T}) = \max_{t \in \{1, 1+s_w, \ldots\} : t \le T-L+1} \omega^\top \phi(z_t, \ldots, z_{t+L-1}). \tag{27}$$

Where $T = |\mathcal{T}|$ is the transcript length, L the window length, and z_t the tth input–label record. The map ϕ computes geometric trace features, ω is the fixed weight vector, and s_w is stride. Equation (27) uses length 32, stride 4, and weights $(2.4, 2.8, 2.2, 1.2, 0.8, 0.4, 0.25)$ for collinearity, halving, joint shrinking/collinearity, close/very-close steps, label changes, and capped step variation. Three width-128 MLP seeds supply six attempts each. Thresholds use 50 benign sessions per seed/length bucket at 1% nominal false positives.

Replay inserts 0/1/3 fillers after each attack query. Three fillers reduce DPE detection from 3/3 to 0/3 sessions while multiplying queries by four and reducing online yield 7.80 to 1.95. Victim-normal detection changes from 3/3 to 1/3 sessions, with yield 3.88 to 0.97. Without filler, Random KD yields 8.10 under this separate protocol. Appended suffixes cannot lower an already-observed maximum. Full-session throughput is

$$Y_{\text{obs}}(\mathcal{T}) = \frac{1000|\mathcal{R}|}{Q_{\text{session}}}. \tag{28}$$

Where Y_{obs} is accepted rules per 1,000 actual session queries. Equation (28) includes setup/fillers, unlike marginal and trace yields defined in the main text.

Input controls yield 35.71/12.18/2.23 localizations per 1k refinement calls for unprocessed/8-bit/6-bit inputs. Margin randomization/rejection at $\tau = 0.02$ yield 8.04/0.00, with 96.9% clean accuracy. These primitive counts exclude pair

acquisition, reference checks, rule validation, and setup. The account simulation estimates 5,829/24,275 online calls for 20 rules under one account/no filler versus 16 accounts/three fillers, with detection 1.000/0.417; it is not a measured full campaign. Source artifacts and experiment mapping remain in `code/EXPERIMENT_MAP.md`.

D Local Proofs and Scope of Recovery

D.1 Matched-Attempt Saving

For the same candidate and bracketing/orientation/validation schedules, common terms in Eq. (17) cancel. Victim-normal estimation costs $2K_i$ calls, while a trained student's gradient incurs none, proving Eq. (18). Setup, fillers, and success rates are separate. The main text contains the proofs of normal consistency and the fixed-candidate PF bound.

For a continuous region Ω, let $T_q = \{(x_i, f(x_i))\}_{i=1}^{q}$ denote a finite hard-label transcript. A model class admitting label-preserving perturbations may leave unqueried labels ambiguous:

$$\begin{aligned} \exists \tilde{f} \in \mathcal{F} : \quad & \tilde{f}(x_i) = f(x_i) \ \ \forall (x_i, f(x_i)) \in T_q, \\ & \exists x^\star \in \Omega \setminus \{x_i\}_{i=1}^{q} \text{ s.t. } \tilde{f}(x^\star) \neq f(x^\star). \end{aligned} \tag{29}$$

Where $\mathcal{F}$ is the transcript-consistent model class. Equation (29) describes ambiguity outside queried points.

Let $\mathcal{C} = \{(p_i, \widehat{n}_i, U_i)\}_{i=1}^{m}$ collect m local certificates. Boundary certificates support pointwise recovery when their regions cover the boundary and their rules agree with the victim:

$$\begin{aligned} \partial f \cap \Omega \subseteq \bigcup_{i=1}^{m} U_i, \\ h_i(x) = f(x) \ \ \forall x \in U_i. \end{aligned} \tag{30}$$

Where ∂f is the victim boundary and U_i is rule h_i's certified region. Extending sampled evidence to the pointwise condition in Eq. (30) requires a finite domain, grid certificate, or regularity assumptions. Correct labels for every induced cell then give

$$\begin{aligned} \mathsf{GCert}_\Omega(\mathcal{C}) = \mathsf{Cover}_\Omega(\mathcal{C}) \wedge \mathsf{LocalEq}_\Omega(\mathcal{C}) \\ \wedge \mathsf{CellLabel}_\Omega(\mathcal{C}), \\ \mathsf{GCert}_\Omega(\mathcal{C}) = 1 \Rightarrow \mathrm{PF}_\Omega(h_\mathcal{C}, f) = 1. \end{aligned} \tag{31}$$

Where $h_\mathcal{C}$ assigns certified cell labels. The three predicates require coverage, pointwise agreement, and correct cell labels. Equation (31) is a schematic sufficient condition for global PF one. DPE evaluates sampled local rules under its stated laws.