TAILSAFE: DECISION-RELEVANT CERTIFICATION FOR LONG-TAILED ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior correction can improve a long-tailed classifier's overall accuracy while degrading a frequency group. TailSafe determines whether the available evidence justifies executing a correction under baseline-relative overall and group risk budgets. Its key observation is that budget feasibility can be resolved without uniquely recovering the target prior. Frozen logit and prototype witnesses identify compatible environments through mixture-direction concentration. A certification-aware selector anticipates an independent paired-loss audit whose uncertainty adapts to error-change second moments. An intersection–union rule controls erroneous release without a direction-count penalty in the audit threshold. Under label shift, a declared finite prior family, and independent calibration, TailSafe provides finite-sample risk control and protected preservation. An exact ambiguity dual and a nonasymptotic execution bound separate action-relevant prior uncertainty from directional source-evidence requirements. Experiments span ten frozen learners and four image benchmarks. In controlled CIFAR-100-LT deployment, TailSafe achieves a 2.4 percentage-point overall gain with 70% execution and 0.4% test-estimated budget exceedance. Finite-support studies demonstrate useful execution under exactly indistinguishable priors. Code is available at ***Supplement***.

1 INTRODUCTION

Long-tailed recognition involves two class distributions: training frequencies determine learning evidence, whereas deployment frequencies determine how classwise errors contribute to target risk. Rebalanced objectives and logit adjustment address training bias (Ren et al., 2020; Menon et al., 2020); test-agnostic recognition and label-shift correction accommodate changing deployment priors (Zhang et al., 2022; Yang et al., 2024; Wei et al., 2024). For a frozen classifier, however, constructing an accurate correction does not by itself establish deployment safety, leaving a separate question: *does the available evidence justify executing this particular change?*

Overall accuracy alone does not establish admissibility: favoring newly common classes can improve the target average while degrading a frequency group. Even an exactly known target prior and an overall Bayes-optimal correction impose no group-specific deterioration budget. Long-tailed deployment also limits independent examples for verifying rare-class changes. Keeping the baseline avoids modification-induced harm but forgoes useful corrections. The objective is beneficial adaptation within prespecified overall and group budgets.

Two uncertainties govern this decision. Under label shift, unlabeled target observations constrain mixture weights, while source labels identify a correction's class-conditional effects (Lipton et al., 2018; Alexandari et al., 2020). Similar class responses can leave several priors compatible with the observations; uneven source counts make their weighted risk changes difficult to estimate. Additional target inputs can reduce sampling uncertainty but cannot resolve indistinguishable mixtures or increase the independent source evidence available for auditing. Prior recovery and reliable execution consequently demand distinct evidence.

Complete prior recovery is nevertheless stronger than necessary. For a fixed action a, the overall gain is $D_q(a) = q^\top d(a)$, where $d_c(a)$ is its improvement over the baseline on class c. Different priors can imply the same gain, and different gains can satisfy the same budget. Even nearby priors can differ along an unfavorable gain direction. **The inferential target is action-specific budget feasibility, not unique prior recovery.**

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

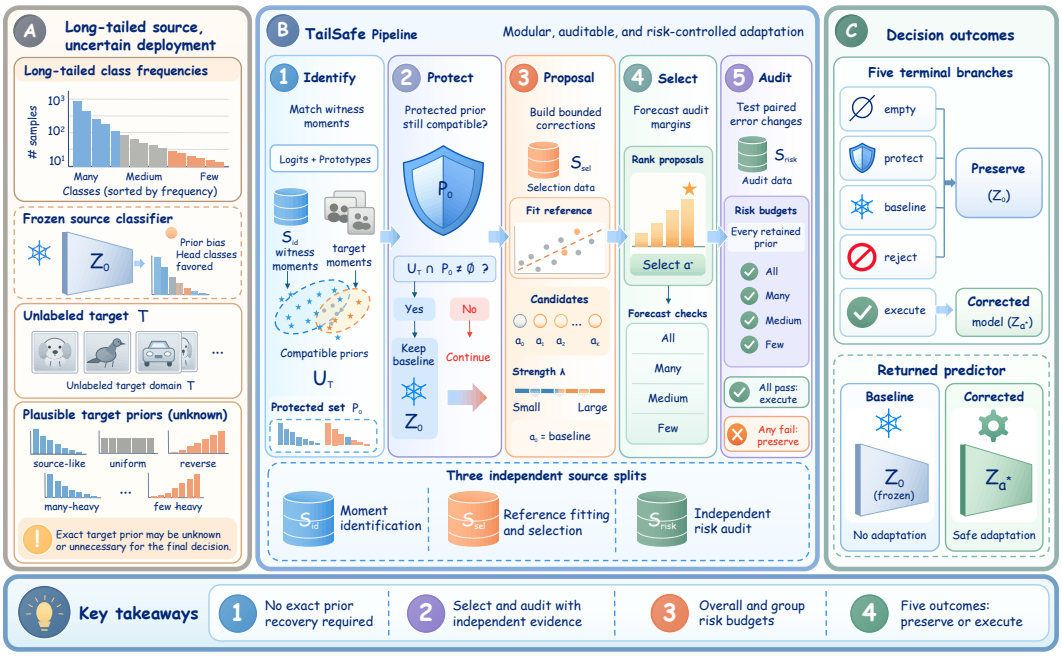


Figure 1: **TailSafe overview.** Frozen responses retain compatible priors, and selection fixes one correction before independent paired auditing. Execution requires every retained overall and group budget check to pass. Empty or protected compatibility, baseline selection, and audit rejection preserve the original scores.

TailSafe implements this principle through Figure 1. Frozen logit and prototype witnesses constrain a declared prior family using mixture-direction concentration. A selector forecasts audit margins from separate selection observations and known audit counts. Once the candidate is fixed, an independent source split measures its paired error changes; a calibrated second-moment allowance exploits sparse changes. Execution requires every retained budget check to pass. This intersection–union rule (Berger & Hsu, 1996) targets erroneous release without requiring simultaneous validity of all numerical lower bounds. Building on independent selection and safety testing (Thomas et al., 2019; Angelopoulos et al., 2025), TailSafe addresses source-to-target certification when deployment labels and exact mixture weights are unavailable.

The contributions connect the deployment decision to its information and evidence requirements:

- ❶ **Decision-relevant prior identification.** A baseline-relative budget formulation and directional mixture concentration retain compatible environments, with an exact ambiguity dual characterizing the witness information needed to resolve action-specific budget margins.
- ❷ **Certification-aware action execution.** Forecast margins guide bounded correction selection, while an independent change-adaptive paired audit supports intersection–union execution with finite-sample risk control and protected-environment preservation.
- ❸ **Evidence requirements for useful adaptation.** A nonasymptotic analysis separates prior ambiguity, missing classes, paired-change frequency, and uneven source counts, yielding sufficient conditions for execution without unique prior recovery.

2 Related Work

Long-tailed recognition under changing priors. Balanced Softmax and logit adjustment address training imbalance (Ren et al., 2020; Menon et al., 2020). SADE and DirMixE aggregate distribution-specialized experts (Zhang et al., 2022; Yang et al., 2024), while LSC estimates deployment priors to correct frozen classifiers (Wei et al., 2024). DART refines predictions using class-

confusion structure (Jang & Chung, 2024); L-TTA combines prototypes, rebalancing shortcuts, and balanced entropy minimization for vision-language adaptation (Wang et al., 2026). TailSafe adds an independently calibrated execution criterion for baseline-relative overall and group risk, applicable to its own or an external correction.

Label-shift inference and uncertainty. BBSE, calibrated maximum likelihood, and RLLS estimate or regularize target mixture weights (Lipton et al., 2018; Alexandari et al., 2020; Azizzadenesheli et al., 2019). Recent semiparametric work studies efficient inference for target-population parameters (Lee et al., 2026); label-shift prediction sets propagate weight uncertainty into coverage (Podkopaev & Ramdas, 2021; Si et al., 2024). TailSafe instead certifies a point classifier's paired risk change across retained priors. Its moment tests require no matrix inversion; agreement on the relevant budgets can permit execution without identifying the mixture.

Safe learning and decision-focused calibration. Seldonian learning and Learn then Test separate selection from calibrated testing (Thomas et al., 2019; Angelopoulos et al., 2025). Shifty extends independent safety testing to demographic shift with specified proportion intervals and anticipates the subsequent test during candidate selection (Giguere et al., 2022). Learn then Test also supplies the multiple-risk intersection–union principle. TailSafe combines these ideas with target-calibrated prior compatibility, source paired-loss evidence, and action-specific execution analysis. Conformal risk control treats expected monotone losses (Angelopoulos et al., 2024); protected adaptation and sequential monitoring address related online risks (Bar et al., 2024; Schirmer et al., 2026). Empirical Bernstein bounds and partial identification provide complementary foundations (Maurer & Pontil, 2009; Berger & Hsu, 1996; Kostin et al., 2024).

3 TailSafe: Setup and Method

TailSafe separates environment identification, candidate selection, and execution certification. The audit construction is presented before the selector because its uncertainty formula also supplies the selection-stage forecast; the actual audit observations remain inaccessible until the candidate is fixed.

3.1 Deployment setting and risk budgets

Let $z_0(x) \in \mathbb{R}^C$ denote frozen scores with fixed class-priority tie-breaking. An action a predicts $h_a(x) = \arg\max_c z_{a,c}(x)$; preservation a_0 returns z_0 exactly. Every source class has positive probability. Under label shift, $P_q(x, c) = q_c P_s(x \mid Y = c)$ and $P_t = P_{\pi_t}$. A prespecified nonempty dictionary $\mathcal{Q} = \{q^{(1)}, \ldots, q^{(L)}\} \subset \Delta^{C-1}$ contains π_t. Training counts determine the nonempty frequency groups $\mathcal{G}$; empty bins are omitted. A protected subset $\mathcal{P}_0 \subseteq \mathcal{Q}$ optionally specifies environments requiring preservation. The template-neighborhood extension is developed in Appendix E.

Define the class risk $r_c(a) = \mathbb{E}_s[\mathbf{1}\{h_a(X) \neq c\} \mid Y = c]$ and paired gain $d_c(a) = r_c(a_0) - r_c(a)$. Label shift preserves these class-conditional quantities, leaving the target prior to determine their aggregates. Overall and group gains use probability directions

$$w_0(q) = q, \qquad w_{g,c}(q) = \frac{q_c \mathbf{1}\{c \in g\}}{q(g)} \quad (q(g) > 0), \qquad D_{q,h}(a) = w_h(q)^\top d(a). \tag{1}$$

Here $q(g) = \sum_{c \in g} q_c$, $\mathcal{H}(q) = \{0\} \cup \{g : q(g) > 0\}$, $D_q = D_{q,0}$, and $R_{q,h}(a) = w_h(q)^\top r(a)$. Conditional group weights need not equal class-macro weights. For nonnegative budgets $\beta_0 = \beta$ and $\{\beta_g\}$, the required complete-decision guarantee is

$$\Pr_{\mathcal{D}}(R_{t,h}(\widehat{a}) - R_{t,h}(a_0) \leq \beta_h \quad \forall h \in \mathcal{H}(\pi_t)) \geq 1 - \delta, \tag{2}$$

where $R_{t,h} = R_{\pi_t,h}$ concerns a fresh target draw. Budgets allow bounded deterioration; δ controls exceedance over the complete decision, including preservation.

Write $\mathcal{D} = (S_{\mathrm{id}}, S_{\mathrm{sel}}, S_{\mathrm{risk}}, T)$. The evidence comprises an independent unlabeled target batch $T = \{X_i^t\}_{i=1}^m$, $m \geq 1$, and three labeled source splits. Model and witness fitting use separate data. Conditional on fitting, source labels, and split assignments, calibration features are mutually independent class-conditional source draws; T is independent and i.i.d. from $P_t(X)$. Assignments

cannot inspect features or losses. All policy and numerical settings are fixed beforehand. Selection can access identification output, S_{sel}, and audit class counts, but no audit features or losses. Allocate $\delta_{\mathrm{id}} + \delta_{\mathrm{r}} = \delta$ with positive components; denote all pre-audit information by $\mathcal{F}_{\mathrm{pre}}$.

3.2 Directional identification of compatible priors

Classifier responses encode the deployed decision structure, while prototype affinities expose its frozen feature geometry. Normalized features $f(x)$ and normalized class means μ_c, fitted independently of calibration, define bounded witnesses

$$\phi_z(x) = \mathrm{softmax}(z_0(x)/\tau_z), \qquad \phi_p(x) = \mathrm{softmax}((\mu_c^\top f(x))_{c=1}^{C}/\tau_p), \tag{3}$$

with positive temperatures and zero vectors retained as zero. The stacked map $\phi = (\phi_z, \phi_p) \in [0,1]^J$, $J = 2C$, need not be a calibrated posterior. Let $M_{jc} = \mathbb{E}_s[\phi_j(X) \mid Y = c]$, so $\mathbb{E}_t\phi(X) = M\pi_t$. For observed class-$c$ identification indices I_c^{id}, estimate

$$\widehat{M}_{jc} = \frac{1}{n_c^{\mathrm{id}}} \sum_{i \in I_c^{\mathrm{id}}} \phi_j(X_i), \qquad \widehat{m}_j = \frac{1}{m} \sum_{i=1}^{m} \phi_j(X_i^t), \tag{4}$$

and use a zero response column when $n_c^{\mathrm{id}} = 0$. With $\mathcal{O}_{\mathrm{id}} = \{c : n_c^{\mathrm{id}} > 0\}$, define

$$\begin{aligned} \eta_{\mathrm{id}}(q) &= \sum_{c \notin \mathcal{O}_{\mathrm{id}}} q_c, & V_{\mathrm{id}}(q) &= \frac{1}{m} + \sum_{c \in \mathcal{O}_{\mathrm{id}}} \frac{q_c^2}{n_c^{\mathrm{id}}}, \\ r_{\mathrm{id}}(q) &= \eta_{\mathrm{id}}(q) + \sqrt{\frac{V_{\mathrm{id}}(q)}{2} \log \frac{2JL}{\delta_{\mathrm{id}}}}, & \mathcal{U}_T &= \{q \in \mathcal{Q} : \|\widehat{m} - \widehat{M}q\|_\infty \le r_{\mathrm{id}}(q)\}. \end{aligned} \tag{5}$$

Weighted Hoeffding concentration (Hoeffding, 1963) yields an event $\mathcal{E}_{\mathrm{id}}$ of probability at least $1-\delta_{\mathrm{id}}$ on which, simultaneously for all $q \in \mathcal{Q}$,

$$\|\widehat{m} - \widehat{M}q - M(\pi_t - q)\|_\infty \le r_{\mathrm{id}}(q). \tag{6}$$

Consequently π_t is retained, and every retained q satisfies $\|M(q - \pi_t)\|_\infty \le 2r_{\mathrm{id}}(q)$. Conversely, separation above $2r_{\mathrm{id}}(q)$ excludes that template on the same event. The squared source weights preserve cross-class averaging: with n observations per class, a uniform mixture contributes $1/(Cn)$ to V_{id}, whereas a single-class mixture contributes $1/n$. Missing mass is charged explicitly; dependent witness coordinates remain admissible. Empty compatibility or $\mathcal{U}_T \cap \mathcal{P}_0 \neq \varnothing$ returns a_0. Protection implements a policy, not a claim that the baseline is optimal.

3.3 Change-adaptive paired auditing

For a candidate a fixed before auditing, let $Z_i(a) = \mathbf{1}\{h_{a_0}(X_i) \neq c\} - \mathbf{1}\{h_a(X_i) \neq c\}$ on class-c examples. Its empirical class means and squared means are $\widehat{d}_c(a)$ and $\widehat{p}_c(a)$. A repaired error contributes $+1$, an introduced error contributes -1, and equal correctness contributes zero. Thus

$$\mathrm{Var}_s(Z(a) \mid Y = c) \le p_c(a) := \mathbb{E}_s[Z(a)^2 \mid Y = c] \le \Pr_s(h_a(X) \neq h_{a_0}(X) \mid Y = c). \tag{7}$$

The audit measures a change, not two unrelated error rates. Predictions can disagree while both are wrong, so error-change second moments can be smaller than prediction disagreement.

For $\mathcal{O} = \{c : n_c^{\mathrm{risk}} > 0\}$ and a probability direction w, define

$$\begin{aligned} \eta_w &= \sum_{c \notin \mathcal{O}} w_c, & V_w &= \sum_{c \in \mathcal{O}} \frac{w_c^2}{n_c^{\mathrm{risk}}}, & b_w &= \max_{c \in \mathcal{O}} \frac{w_c}{n_c^{\mathrm{risk}}}, \\ \widehat{G}_w(a) &= \sum_{c \in \mathcal{O}} w_c \widehat{d}_c(a), & \widehat{v}_w(a) &= \sum_{c \in \mathcal{O}} \frac{w_c^2 \widehat{p}_c(a)}{n_c^{\mathrm{risk}}}. \end{aligned} \tag{8}$$

These terms separate missing mass, range-based uncertainty, and the largest observation weight. An empty maximum is zero. For $0 < \alpha < 1$ and $\ell_\alpha = \log(3/\alpha)$, use

$$\begin{aligned} \overline{v}_w(\alpha; u) &= \min\{V_w, 2u + 2b_w^2 \ell_\alpha\}, \\ r_w(\alpha; u) &= \min\Big\{\sqrt{2V_w\ell_\alpha}, \sqrt{2\overline{v}_w(\alpha;u)\ell_\alpha} + \tfrac{4}{3} b_w \ell_\alpha\Big\}, \\ L_w(a; \alpha) &= \max\{-1, \widehat{G}_w(a) - \eta_w - r_w(\alpha; \widehat{v}_w(a))\}. \end{aligned} \tag{9}$$

The second branch includes second-moment uncertainty (Maurer & Pontil, 2009). Chernoff, Bernstein, and Hoeffding bounds jointly calibrate the minimum, yielding for each fixed direction

$$\Pr\big(L_w(a;\alpha) > w^\top d(a) \mid \mathcal{F}_{\mathrm{pre}}\big) \le \alpha. \tag{10}$$

Appendix C.3 gives the allocation over these three events; no independence between the paired mean and second moment is needed. Missing classes contribute their worst possible gain through $-\eta_w$, without renormalizing observed weights. A wholly unobserved direction has $L_w = -1$.

3.4 Certification-aware selection and execution

Compatible priors determine the environments a correction must withstand, but do not prescribe a unique action. For $q \in \mathcal{U}_T$ and a fixed finite strength grid $\Lambda \subseteq [0, \lambda_{\max}]$, construct

$$z_{q,\lambda,c}(x) = z_{0,c}(x) + \lambda[\log(q_c + \epsilon) - \log(\widehat{\rho}_c + \epsilon)], \qquad \epsilon > 0. \tag{11}$$

The reference $\widehat{\rho} \in \Delta^{C-1}$ uses source-weighted bias calibration on S_{sel} (Appendix B.2); it parameterizes corrections and is not assumed to identify a prior encoded by arbitrary scores. Coordinate changes are bounded by $\lambda_{\max}\log((1+\epsilon)/\epsilon)$. Let $\mathcal{A}_T$ contain these actions and a_0, with zero-strength proposals represented by the baseline.

Selection favors corrections with both estimated benefit and sufficient anticipated audit margin. It computes paired means and second moments on S_{sel}, using $(-1, 1)$ for an absent class of a non-baseline candidate. Audit counts alone determine the planning weights, giving

$$\begin{aligned} \widetilde{v}_w(a) &= \sum_{c\in\mathcal{O}} \frac{w_c^2 \widehat{p}_c^{\mathrm{sel}}(a)}{n_c^{\mathrm{risk}}}, \\ \widetilde{L}_w(a) &= \max\{-1, \sum_{c\in\mathcal{O}} w_c \widehat{d}_c^{\mathrm{sel}}(a) - \eta_w - r_w(\delta_{\mathrm{r}}; \widetilde{v}_w(a))\}. \end{aligned} \tag{12}$$

Set $\widetilde{L}_w(a_0) = 0$ exactly and select

$$\begin{aligned} \mathcal{F}_{\mathrm{sel}} &= \{a \in \mathcal{A}_T : \widetilde{L}_{w_h(q)}(a) + \beta_h \ge 0 \quad \forall q \in \mathcal{U}_T,\ h \in \mathcal{H}(q)\}, \\ a^\star &\in \arg\max_{a\in\mathcal{F}_{\mathrm{sel}}} \min_{q\in\mathcal{U}_T} \widetilde{L}_{w_0(q)}(a). \end{aligned} \tag{13}$$

Ties favor a_0 and then a fixed ordering. The baseline makes $\mathcal{F}_{\mathrm{sel}}$ nonempty, and a nonbaseline selection has positive worst-case overall forecast. A large empirical gain alone can be insufficient when a candidate changes many errors or relies on poorly observed classes. The forecasts guide this search; their statistical calibration is not assumed. Reference fitting and scoring can share S_{sel} because the complete candidate is frozen before S_{risk} is accessed.

Execution. A baseline selection preserves z_0; otherwise, audit the fixed $a^\star$ and execute only if

$$\min_{q\in\mathcal{U}_T,\ h\in\mathcal{H}(q)} \{L_{w_h(q)}(a^\star; \delta_{\mathrm{r}}) + \beta_h\} \ge 0. \tag{14}$$

A failed check returns the original scores and tie rule, without reference recalibration or another search on the audit. The same rule supports external pre-audit proposals. Algorithm 1 specifies the complete decision and its data-access boundary.

4 Decision-Relevant Guarantees

Risk control requires valid testing; useful execution additionally requires sufficient margins. The following analysis distinguishes these two objectives.

4.1 Risk control and protected preservation

Requiring every test to pass is an intersection–union decision (Berger & Hsu, 1996). Its null is that at least one retained budget is violated. The relevant error is releasing such a candidate, not reporting any inaccurate numerical lower bound.

Theorem 1 (Risk control and protected preservation)**.** *Under Section 3.1, $\pi_t \in \mathcal{Q}$, and $\delta_{\mathrm{id}} + \delta_{\mathrm{r}} = \delta < 1$, TailSafe retains the true prior with probability at least $1 - \delta_{\mathrm{id}}$ and satisfies Eq 2. If $\pi_t \in \mathcal{P}_0$, then $\Pr(\widehat{a} \neq a_0) \leq \delta_{\mathrm{id}}$. The result holds for any pre-audit candidate selector.*

Conditional on $\mathcal{F}_{\mathrm{pre}}$, the candidate and retained directions are fixed. If any budget is violated, choose its first violated direction in a fixed order. Release entails falsely passing this particular test, whose probability is at most δ_{r} by Eq 10. Identification failure costs at most δ_{id}; preservation has zero excess risk. A protected true prior triggers preservation whenever retained. The proof therefore uses neither a direction-count nor a candidate-count factor in the audit threshold (Appendix C.4). It does not assert that all reported L_w values form a simultaneous confidence band.

4.2 Budget identification under prior ambiguity

Let $e_0 = \mathbf{1}$, $e_g = \mathbf{1}_g$, and $s_h(q) = q^\top e_h$. Define the linear budget direction

$$u_h(a) = e_h \odot (d(a) + \beta_h \mathbf{1}), \qquad q^\top u_h(a) = s_h(q)[w_h(q)^\top d(a) + \beta_h] \quad (s_h(q) > 0). \tag{15}$$

The functional is nonnegative exactly when its positive-mass budget is feasible. Zero-mass groups have zero margin and require no check. Estimating conditional risk is distinct from deciding its admissibility. The decision-relevant ambiguity modulus is

$$\omega_M(u, \varepsilon) = \sup_{\substack{q, q' \in \Delta^{C-1} \\ \|M(q-q')\|_\infty \leq \varepsilon}} |u^\top (q - q')|. \tag{16}$$

The simplex modulus also bounds ambiguity in the finite dictionary; it measures the effect of unresolved mixtures on a particular margin.

Proposition 1 (Exact ambiguity dual)**.** *For every M, u and $\varepsilon \geq 0$,*

$$\omega_M(u, \varepsilon) = \inf_{v \in \mathbb{R}^J} \{\varepsilon \|v\|_1 + \mathrm{osc}(u - M^\top v)\}, \qquad \mathrm{osc}(b) = \max_c b_c - \min_c b_c. \tag{17}$$

Linear-programming duality establishes the identity, including $\varepsilon = 0$ (Appendix D.1). The terms quantify witness sensitivity and residual representation error. If $u = M^\top v + b\mathbf{1}$, ambiguity is at most $\varepsilon \|v\|_1$, even when the moment fiber $\mathcal{F}_M(\pi_t) = \{q \in \mathcal{Q} : Mq = M\pi_t\}$ contains multiple priors. Margin identification requires a constant functional on that fiber; budget identification requires only agreement about feasibility. Consequently, the relevant uncertainty is adverse margin variation across compatible priors, which their number alone does not determine.

On $\mathcal{E}_{\mathrm{id}}$ every fiber member remains compatible. A protected member forces preservation; a budget-violating member obstructs reliable release of that action. Adding witnesses can reduce the population modulus at a fixed tolerance, but useful finite-sample improvement also requires sufficiently precise estimated moments. These statements concern the selected moment procedure, not universal nonidentifiability from the full input distribution (Appendix D.2).

4.3 Evidence sufficient for execution

For the selected action, let $\nu_w = \sum_{c \in \mathcal{O}} w_c^2 p_c(a^\star)/n_c^{\mathrm{risk}}$, the population paired-second-moment proxy. For $0 < \kappa < 1$, define $K_T = \sum_{q \in \mathcal{U}_T} |\mathcal{H}(q)|$, $\ell = \log(3/\delta_{\mathrm{r}})$, $t_\kappa = \log(2K_T/\kappa)$, and

$$E_w(\kappa) = 2\eta_w + \sqrt{10\nu_w(\ell + t_\kappa)} + \tfrac{10}{3} b_w(\ell + t_\kappa). \tag{18}$$

Conditional on the pre-audit information, weighted concentration bounds $w^\top d(a^\star) - L_w(a^\star; \delta_{\mathrm{r}})$ by $E_w(\kappa)$ simultaneously over retained directions with probability at least $1 - \kappa$. Hence a local sufficient condition for execution is $w_h(q)^\top d(a^\star) + \beta_h \geq E_{w_h(q)}(\kappa)$ for every retained (q, h). Population quantities in this analysis are not algorithm inputs.

Theorem 2 (Execution with unresolved prior uncertainty)**.** *Fix a pre-audit realization on $\mathcal{E}_{\mathrm{id}}$ reaching the audit with $a^\star \neq a_0$. Set $r_{\max} = \max_{q \in \mathcal{U}_T} r_{\mathrm{id}}(q)$ and $\mathcal{U}_{T,h} = \{q \in \mathcal{U}_T : s_h(q) > 0\}$. Suppose, for all h with $\mathcal{U}_{T,h} \neq \varnothing$,*

$$\pi_t^\top u_h(a^\star) \geq \underbrace{\omega_M(u_h(a^\star), 2r_{\max})}_{\textit{budget-relevant ambiguity}} + \underbrace{\max_{q \in \mathcal{U}_{T,h}} s_h(q) E_{w_h(q)}(\kappa)}_{\textit{directional audit evidence}}, \tag{19}$$

then $\Pr(\widehat{a} = a^\star \mid \mathcal{F}_{\mathrm{pre}}) \geq 1 - \kappa$.

Compatibility bounds moment distances; the modulus then lower-bounds each retained budget margin. Dividing by positive $s_h(q)$ gives the local sufficient condition, so every check passes on the common audit-gap event (Appendix D.3). K_T enters the all-checks power analysis, not the release threshold. The statement concerns the selected action, not optimal candidate discovery.

Long-tailed evidence and sample requirements. With all positive weights observed, $n_{\mathrm{eff}}(w) = V_w^{-1}$. Uniform weights over C_g classes with n examples each give $N_g = C_g n$, $\nu_w = \overline{p}_g / N_g$, and $b_w = 1/N_g$, where $\overline{p}_g$ is their mean paired-change probability. Writing $s = \ell + t_\kappa$, a positive local margin $\gamma_g = w^\top d(a^\star) + \beta_g$ satisfies that direction's power condition whenever (Appendix D.5)

$$N_g \geq \max\left\{ \frac{40\overline{p}_g s}{\gamma_g^2}, \frac{20s}{3\gamma_g} \right\}. \tag{20}$$

5 Experiments

Experiments evaluate recognition across frozen learners, deployment utility under uncertain priors, and the contributions of identification, selection, and independent auditing.

5.1 Evaluation protocol

Evaluation covers CIFAR-100-LT at IR $100/50/10$, ImageNet-LT, iNaturalist 2018, and Places-LT (Krizhevsky et al., 2009; Cao et al., 2019; Liu et al., 2019; Van Horn et al., 2018). Tables 1–2 compare ten frozen learners. Controlled deployment uses BCL, with ProCo and ICL as complementary controls (Zhu et al., 2022; Du et al., 2024; Chen et al., 2025). Seven priors comprise source-frequency, uniform, reverse-frequency, two interpolations, and two group-heavy distributions; only source-frequency is protected. Appendices A and G specify the data protocols.

The shared settings are $\delta_{\mathrm{id}} = \delta_r = 0.025$, $\beta = 0$, $\beta_g = 0.02$, and $\Lambda = \{0, 0.25, 0.5, 0.75, 1\}$. Identification, selection, and auditing use disjoint source roles and their realized class counts. Paired methods share checkpoints and observations, with results averaged over three seeded runs. Recognition reports class-uniform Top-1; iNaturalist uses label-hidden transductive evaluation. Deployment gains are percentage points (pp), Int. denotes execution, and Exc. denotes joint test-estimated budget exceedance. Controlled deployment averages weight the six unprotected priors equally and retain preservation outcomes in the complete decision denominator.

5.2 Recognition across frozen learners

TailSafe improves overall Top-1 for all reported learner–benchmark pairs in Tables 1–2. On CIFAR-100-LT at IR=100, Few gains are 3.5–7.2 pp and overall gains are 0.6–1.9 pp. At IR=50, these ranges are 6.7–11.4 and 0.8–1.9 pp; overall gains narrow to 0.2–0.5 pp at IR=10, whose Few bin is empty. On ImageNet-LT, iNaturalist 2018, and Places-LT, Few improves by 7.2–16.0, 1.4–3.9, and 6.9–10.7 pp, respectively. The accompanying Many/Medium trade-offs motivate evaluating overall utility and group protection separately.

5.3 Selective deployment and component attribution

Table 3 quantifies the gain–risk trade-off. TailSafe achieves $+2.4$ pp All and $+4.0$ pp Few with 70% execution and 0.4% estimated exceedance. Applying its execution layer to the identical LSC proposal reduces execution from 90% to 60% and Exc. from 6.0% to 0.5%, retaining $+1.9$ pp All gain under substantially tighter risk control. Common-veto and label-budget controls in Appendix G.6 separate protection policy and label access from certification.

Table 4 identifies each component's contribution. Coordinate-wise identification lowers execution to 59%; range-based and joint audits reduce overall gain to 1.6 and 1.8 pp, respectively, reflecting more conservative certification. Removing group checks or auditing raises Exc. to 3.2% and 4.8%. Appendix G.7 specifies the matched end-to-end and fixed-action controls.

Table 1: CIFAR-100-LT uniform-target Top-1 (%) under the reserved-calibration protocol ($N_{\max} = 400$). Parentheses give Few/All gains (pp). IR=10 has no Few classes.

Frozen learner	IR=100				IR=50				IR=10			
	Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
LDAM-DRW (2019)	62.6	43.0	20.8	41.1	61.8	41.4	29.6	45.7	62.2	53.7	–	58.8
+ TailSafe	60.7	42.1	27.1 (+6.3)	42.4 (+1.3)	60.1	40.9	36.8 (+7.2)	46.6 (+0.9)	62.0	54.5	–	59.0 (+0.2)
Balanced Softmax (2020)	65.6	49.6	23.2	45.2	66.3	47.7	29.8	49.9	65.6	53.2	–	60.6
+ TailSafe	64.5	48.8	30.4 (+7.2)	47.1 (+1.9)	64.8	47.0	41.2 (+11.4)	51.8 (+1.9)	65.5	53.8	–	60.8 (+0.2)
RIDE (3E) (2020)	69.4	50.2	24.7	47.0	66.7	50.1	28.8	50.8	68.4	51.3	–	61.6
+ TailSafe	68.1	49.2	31.3 (+6.6)	48.6 (+1.6)	65.2	49.6	35.9 (+7.1)	51.8 (+1.0)	68.3	52.0	–	61.8 (+0.2)
BCL (2022)	67.6	53.0	33.9	50.7	69.6	54.9	34.9	55.2	66.8	61.0	–	64.5
+ TailSafe	66.1	52.3	39.5 (+5.6)	52.0 (+1.3)	68.6	54.2	41.6 (+6.7)	56.2 (+1.0)	66.4	62.3	–	64.8 (+0.3)
SADE (2022)	68.8	54.6	25.9	48.8	67.1	53.0	32.3	53.0	65.6	60.6	–	63.6
+ TailSafe	67.3	54.1	32.8 (+6.9)	50.6 (+1.8)	65.4	52.1	42.2 (+9.9)	54.4 (+1.4)	65.5	61.8	–	64.0 (+0.4)
BalPoE (2023)	68.9	55.4	32.0	51.3	69.2	55.6	34.7	55.3	67.8	61.9	–	65.4
+ TailSafe	66.9	54.4	37.3 (+5.3)	52.2 (+0.9)	67.5	55.0	41.8 (+7.1)	56.2 (+0.9)	67.5	62.8	–	65.6 (+0.2)
PRL (2024)	69.9	56.5	32.1	52.0	69.0	55.8	40.0	56.6	67.9	61.3	–	65.3
+ TailSafe	68.8	55.9	37.9 (+5.8)	53.5 (+1.5)	67.8	55.0	47.5 (+7.5)	57.7 (+1.1)	67.8	62.5	–	65.7 (+0.4)
ProCo (2024)	70.0	52.0	35.1	51.5	69.0	55.6	37.1	55.9	67.8	60.1	–	64.7
+ TailSafe	68.3	51.2	39.6 (+4.5)	52.3 (+0.8)	67.7	54.7	43.9 (+6.8)	56.7 (+0.8)	67.6	61.4	–	65.1 (+0.4)
ICL (2025)	71.6	61.3	36.4	55.7	72.2	60.5	38.8	59.4	69.8	64.3	–	67.6
+ TailSafe	69.6	60.7	41.3 (+4.9)	56.6 (+0.9)	70.6	59.8	45.7 (+6.9)	60.2 (+0.8)	69.4	65.8	–	68.0 (+0.4)
LOS (2025)	74.4	59.3	31.1	54.0	71.0	58.0	38.5	57.9	72.7	64.5	–	69.4
+ TailSafe	73.0	58.8	34.6 (+3.5)	54.6 (+0.6)	69.0	57.5	46.6 (+8.1)	58.9 (+1.0)	72.5	66.0	–	69.9 (+0.5)

Table 2: Uniform-target Top-1 (%) under dataset-specific calibration holdouts. Parentheses give Few/All gains (pp). Groups use actual post-reservation training counts (Appendix A).

Frozen learner	ImageNet-LT				iNaturalist 2018				Places-LT			
	Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
LDAM-DRW (2019)	58.5	44.7	30.0	48.0	63.8	66.4	62.4	64.0	32.3	31.6	25.6	30.7
+ TailSafe	56.5	43.8	42.3 (+12.3)	48.5 (+0.5)	61.8	65.6	64.0 (+1.6)	64.4 (+0.4)	30.3	30.7	33.8 (+8.2)	31.2 (+0.5)
Balanced Softmax (2020)	64.6	47.7	33.4	52.3	71.2	71.2	71.3	71.3	42.1	39.7	32.8	39.2
+ TailSafe	63.0	47.2	41.5 (+8.1)	52.5 (+0.2)	69.5	70.4	75.2 (+3.9)	73.0 (+1.7)	40.5	39.0	39.8 (+7.0)	39.7 (+0.5)
RIDE (3E) (2020)	66.4	50.6	34.8	54.5	65.2	70.5	69.8	69.7	42.5	40.0	32.9	39.5
+ TailSafe	65.0	50.1	46.3 (+11.5)	55.3 (+0.8)	63.4	69.7	72.3 (+2.5)	70.6 (+0.9)	41.1	39.4	43.6 (+10.7)	40.8 (+1.3)
BCL (2022)	66.1	54.2	35.6	56.3	67.7	72.9	71.7	71.8	42.0	40.4	33.2	39.6
+ TailSafe	65.1	53.4	42.8 (+7.2)	56.5 (+0.2)	66.6	71.9	74.8 (+3.1)	73.1 (+1.3)	40.5	39.8	41.4 (+8.2)	40.4 (+0.8)
SADE (2022)	67.5	55.6	42.5	58.4	75.1	72.3	73.2	73.0	40.7	43.5	36.7	41.2
+ TailSafe	66.0	55.1	56.4 (+13.9)	59.5 (+1.1)	73.1	71.3	75.7 (+2.5)	73.9 (+0.9)	39.0	42.6	45.6 (+8.9)	41.9 (+0.7)
BalPoE (2023)	66.3	56.9	43.3	58.7	74.3	77.1	74.2	75.3	42.9	40.0	33.8	39.8
+ TailSafe	64.8	55.9	54.8 (+11.5)	59.2 (+0.5)	72.5	76.1	76.1 (+1.9)	75.8 (+0.5)	41.2	39.1	42.6 (+8.8)	40.5 (+0.7)
PRL (2024)	71.9	56.7	41.2	60.4	76.8	76.2	72.8	74.4	46.1	41.4	33.9	41.6
+ TailSafe	70.9	55.7	50.5 (+9.3)	60.8 (+0.4)	75.3	75.4	75.2 (+2.4)	75.3 (+0.9)	44.7	40.9	40.8 (+6.9)	42.2 (+0.6)
ProCo (2024)	68.0	54.3	38.5	57.4	74.8	74.8	71.0	72.7	41.7	39.8	33.5	39.3
+ TailSafe	66.8	53.4	54.5 (+16.0)	58.7 (+1.3)	73.5	74.2	72.6 (+1.6)	73.3 (+0.6)	40.4	39.2	42.0 (+8.5)	40.2 (+0.9)
ICL (2025)	67.5	56.6	44.2	59.1	75.3	75.3	75.4	75.4	45.1	40.1	33.8	40.7
+ TailSafe	66.1	55.8	53.1 (+8.9)	59.4 (+0.3)	74.1	74.4	76.8 (+1.4)	75.7 (+0.3)	43.1	39.1	42.6 (+8.8)	41.2 (+0.5)
LOS (2025)	62.3	51.2	42.7	54.3	70.5	70.3	71.1	70.8	43.9	43.7	35.9	42.3
+ TailSafe	61.0	50.2	52.8 (+10.1)	54.7 (+0.4)	69.4	69.5	74.3 (+3.2)	72.2 (+1.4)	42.5	43.0	42.8 (+6.9)	42.8 (+0.5)

Table 3: Controlled CIFAR-100-LT deployment: paired gains (pp) and decision rates (%).

Rule	ΔAll	ΔFew	Int.	Exc.
Frozen	0.0	0.0	0.0	0.0
Point prior	2.9	1.8	91.0	6.8
BBSE	2.0	2.2	85.0	5.5
MLLS	2.6	3.0	88.0	4.7
LSC direct	3.1	2.5	90.0	6.0
LSC + TailSafe	1.9	3.3	60.0	0.5
TailSafe	2.4	4.0	70.0	0.4

Table 4: Controlled CIFAR-100-LT component ablations with matched data and budgets.

Variant	ΔAll	ΔFew	Int.	Exc.
TailSafe	2.4	4.0	70.0	0.4
Coordinate ID	2.0	3.7	59.0	0.4
Range audit	1.6	2.8	49.0	0.3
Joint audit	1.8	3.2	54.0	0.3
Gain selection	1.9	2.6	51.0	0.4
Budget selection	2.1	3.6	62.0	0.4
Logit witnesses	2.0	3.3	58.0	0.4
No group audit	2.6	3.4	76.0	3.2
No risk audit	2.9	3.6	81.0	4.8

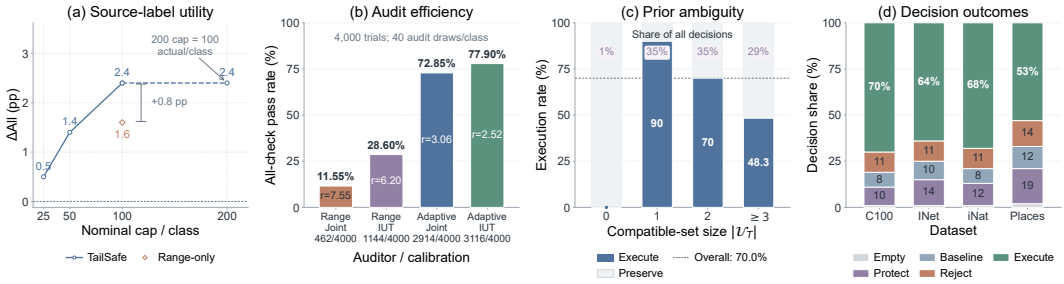


Figure 2: **Evidence and decision diagnostics. (a)** Gain versus nominal source-label cap. **(b)** Fixed-allocation auditor comparison. **(c)** Execution by compatible-set size. **(d)** Five terminal branches across datasets. Panels (a), (c), and (d) use controlled BCL records; panel (b) is a paired-loss experiment. Details are in Appendix I.

Table 5 isolates the auditor using 4,000 paired-loss trials with 40 audit observations per class (Appendix I.2). Adaptive/IUT accepts 77.90% of trials with a 2.52 pp median directional radius, compared with 28.60% and 6.20 pp for Range/IUT. Joint calibration over six distinct checks yields 72.85% and 11.55%, respectively. All four auditors share the same candidate laws and observations.

Table 5: Controlled paired-loss audit ($C = 100$). Median radius (pp); cohort pass rate (%).

Bound	Calibration	Radius	Pass (%)
Range	Joint	7.55	11.55
Range	IUT	6.20	28.60
Adaptive	Joint	3.06	72.85
Adaptive	IUT	2.52	77.90

Table 6: Source-label resource study. Cap and used are per-class image counts; gains are pp.

Cap	Used	ΔAll	Int. (%)	Exc. (%)
25	25	0.5	18.0	0.4
50	50	1.4	44.0	0.4
100	100	2.4	70.0	0.4
200	100	2.4	70.0	0.4

5.4 Evidence requirements and decision diagnostics

Figure 2(a) and Table 6 show that increasing the realized allocation from 25 to 100 images per class raises overall gain from 0.5 to 2.4 pp and execution from 18% to 70%. TailSafe exceeds Range-only by 0.8 pp at the reference, indicating that additional independent source evidence translates into more certifiable corrections rather than merely larger nominal budgets. Panel (b) isolates the bound and calibration effects at a fixed allocation, separating audit efficiency from candidate construction.

In Figure 2(c), non-singleton compatible sets account for approximately 55% of executed corrections: useful execution does not require unique empirical prior resolution when the retained environments agree sufficiently on the relevant budget constraints. Panel (d) separates all five terminal branches, distinguishing unresolved compatibility, baseline selection, audit rejection, and successful execution. Appendix I provides prior-specific, resource, and witness diagnostics.

Exact-risk studies complement these diagnostics (Appendix J). Across 70,000 six-class complete decisions, no violation is observed; execution reaches 66.64% and 97.08% at 1,024 and 4,096 audit draws per class. The four-class study further demonstrates useful execution while two priors with identical full input marginals remain compatible, directly illustrating that certifying an action can require less information than uniquely identifying the deployment prior.

6 Conclusion

TailSafe certifies long-tailed adaptation through evidence about a correction's baseline-relative consequences. Directional identification and a change-adaptive paired audit support execution under uncertain target priors, with an analysis separating environmental ambiguity from source-evidence requirements. Complete prior recovery is unnecessary when the remaining uncertainty permits the prescribed budgets; otherwise, exact preservation retains the deployed classifier.

AI USE STATEMENT

No generative AI tools were used in the development of the research ideas, methodology, experiments, analysis, code, figures, or writing of this work. All content was produced and reviewed by the authors.

REPRODUCIBILITY STATEMENT

Algorithm 1 and Appendices B–E specify the decision rule, fitting details, assumptions, and proofs. Appendix A records dataset-specific holdouts, frequency groups, and the iNaturalist target-access protocol. Appendix G gives sampling and aggregation conventions, and Appendices I–J provide deployment, attribution, sensitivity, and finite-support analyses.

REFERENCES

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19967–19977. IEEE, 2023.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International conference on machine learning*, pp. 222–232. PMLR, 2020.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019.

Yarin Bar, Shalev Shaer, and Yaniv Romano. Protected test-time adaptation via online entropy matching: A betting approach. *Advances in Neural Information Processing Systems*, 37:85467–85499, 2024.

Roger L Berger and Jason C Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Nadine Chang, Jayanth Koushik, Aarti Singh, Martial Hebert, Yu-Xiong Wang, and Michael J Tarr. Alphanet: Improving long-tail classification by combining classifiers. *arXiv preprint arXiv:2008.07073*, 2020.

Yanhao Chen, Zhongquan Jian, Nianxin Ke, Shuhao Hu, Junjie Jiao, Qingqi Hong, and Qingqiang Wu. Enhancing mixture of experts with independent and collaborative learning for long-tail visual recognition. In *IJCAI*, pp. 828–836, 2025.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (9):5890–5904, 2024.

Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip S Thomas, and Scott Niekum. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7607–7616. IEEE, 2020.

Minguk Jang and Hye Won Chung. Label distribution shift-aware prediction refinement for test-time adaptation. *arXiv preprint arXiv:2411.15204*, 2024.

Julia Kostin, Nicola Gnecco, and Fanny Yang. Achievable distributional robustness when the robust risk is only partially identified. *Advances in Neural Information Processing Systems*, 37:83915–83950, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Seong-ho Lee, Yanyuan Ma, and Jiwei Zhao. Efficient inference under label shift in unsupervised domain adaptation. *Journal of Machine Learning Research*, 27(193):1–50, 2026.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2532–2541. IEEE, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pp. 844–853. PMLR, 2021.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Mona Schirmer, Metod Jazbec, Christian Andersson Naesseth, and Eric Nalisnick. Monitoring risks in test-time adaptation. *Advances in Neural Information Processing Systems*, 38:80915–80948, 2026.

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. Pac prediction sets under label shift. In *International Conference on Learning Representations*, volume 2024, pp. 44208–44228, 2024.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, volume 2025, pp. 26681–26697, 2025.

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Xucong Wang, Zhe Zhao, Zekun Wang, Xiaofeng Cao, Xu Wang, Di Wu, Pengkun Wang, and Yang Wang. Long-tailed test-time adaptation for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 46492–46530, 2026.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Tong Wei, Zhen Mao, Zi-Hao Zhou, Yuanyu Wan, and Min-Ling Zhang. Learning label shift correction for test-agnostic long-tailed recognition. 2024.

Zhiyong Yang, Qianqian Xu, Zitai Wang, Sicong Li, Boyu Han, Shilong Bao, Xiaochun Cao, and Qingming Huang. Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In *Forty-first International Conference on Machine Learning*, 2024.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in neural information processing systems*, 35:34077–34090, 2022.

Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. *Advances in Neural Information Processing Systems*, 37:7493–7520, 2024.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6898–6907. IEEE, 2022.

Appendix

TailSafe: Decision-Relevant Certification for Long-Tailed Adaptation

The content of the **Appendix** is summarized as follows:

- **Appendix A: Dataset-Specific Calibration and Data Provenance.**
  - Appendix A.1: Calibration sources and evaluation pools.
  - Appendix A.2: Role allocation and fixed frequency groups.
  - Appendix A.3: Independence, checkpoint eligibility, and data access.
- **Appendix B: Reference Fitting and Selection Details.**
  - Appendix B.1: Frozen witnesses and source class means.
  - Appendix B.2: Source reference fitting.
  - Appendix B.3: Forecasts and terminal decisions.
  - Appendix B.4: Complete deployment algorithm and data access.
- **Appendix C: Finite-Sample Risk Control.**
  - Appendix C.1: Conditioning and probability space.
  - Appendix C.2: Directional moment concentration.
  - Appendix C.3: A marginal change-adaptive audit bound.
  - Appendix C.4: Proof of the complete release guarantee.
- **Appendix D: Decision-Relevant Identification and Execution Power.**
  - Appendix D.1: Exact ambiguity duality.
  - Appendix D.2: Moment fibers and budget decisions.
  - Appendix D.3: Power of the independent audit.
  - Appendix D.4: Effective sample size and unchanged directions.
  - Appendix D.5: Explicit sample requirements for uniformly weighted directions.
- **Appendix E: A Prespecified Template-Neighborhood Extension.**
- **Appendix F: Optional Tail-Macro Protection and Witness Aggregation.**
- **Appendix G: Detailed Experimental Protocol.**
  - Appendix G.1: Dataset, checkpoint, and group inventories.
  - Appendix G.2: Independent source resources and sampling.
  - Appendix G.3: Deployment priors and empty-bin handling.
  - Appendix G.4: Fixed numerical configuration.
  - Appendix G.5: Metrics, identities, and statistical uncertainty.
  - Appendix G.6: Baselines and identical-proposal controls.
  - Appendix G.7: Controlled ablation definitions.
- **Appendix H: Evaluation Consistency and Statistical Scope.**
  - Appendix H.1: Common-weight recognition aggregation.
  - Appendix H.2: Aggregation identities and statistical scope.
- **Appendix I: Extended Deployment, Attribution, and Sensitivity Studies.**
  - Appendix I.1: Complete primary comparison records.
  - Appendix I.2: Fixed paired-loss cohort construction.
  - Appendix I.3: Cross-dataset and cross-learner coverage.
  - Appendix I.4: Prior-specific group gains and terminal decisions.
  - Appendix I.5: Useful decisions under residual ambiguity.
  - Appendix I.6: Selection, forecasts, and execution outcomes.
  - Appendix I.7: Resource allocation and certificate power.
  - Appendix I.8: Identification resolution and the protection bottleneck.

  - Appendix I.9: Risk budgets, confidence allocation, and dictionary size.
  - Appendix I.10: Template mismatch, conditional change, and optional policies.
  - Appendix I.11: Computation and memory.
- **Appendix J: Exact-Risk Calibration and Mechanism Experiments.**
  - Appendix J.1: Complete finite-support deployment experiment.
  - Appendix J.2: A population moment fiber with a useful correction.
  - Appendix J.3: Independent auditing within a nonidentifiable moment fiber.
  - Appendix J.4: Paired-change sparsity at a fixed gain.
  - Appendix J.5: Testing the conditional execution-power result.
- **Appendix K: Limitations.**

A Dataset-Specific Calibration and Data Provenance

A.1 Calibration sources and evaluation pools

The ten learners in Tables 1–2 are evaluated under the matched protocol in Table S1. Calibration membership and roles are fixed before model or witness fitting. The baseline and adapted predictor in each pair share their fitted checkpoint; the reported adaptation is evaluated against that baseline. No published baseline result from another training split is substituted into a paired comparison.

Table S1: Calibration provenance. Counts are per class. N_c^{raw} denotes the class count before calibration reservation; all final evaluation pools are excluded from labeled calibration.

Dataset	Calibration source	Count; roles	Evaluation pool
CIFAR-100-LT	Original training pool, reserved before LT construction	100; $40/20/40$	Complete official test set
ImageNet-LT	Official validation split	20; $8/4/8$	Official test split
iNaturalist 2018	Images indexed by the official training JSON file, reserved before fitting	Minimum-support allocation; Eqs A.2–A.3	Complete official validation JSON split; label-hidden transductive
Places-LT	Official validation split	20; $8/4/8$	Official test split

CIFAR-100-LT. CIFAR-100 has 500 original training images and 100 test images per class (Krizhevsky et al., 2009). Reserving 100 training images per class leaves a 400-per-class parent pool. The training count at frequency rank j is

$$N_{\sigma(j)} = \left\lfloor 400\,\mathrm{IR}^{-(j-1)/99} \right\rfloor, \qquad j = 1, \ldots, 100, \tag{A.1}$$

where σ is a fixed class permutation. This reserved-calibration protocol has $N_{\max} = 400$ and is distinct from the unreserved 500-maximum construction. Each matched baseline–TailSafe pair uses the same fitted checkpoint and actual training-count vector.

ImageNet-LT and Places-LT. The benchmark validation lists contain 20 images per class (Liu et al., 2019; Jamal et al., 2020). Their roles are changed from development data to independent calibration: they are not used for early stopping, checkpoint selection, or hyperparameter selection. ImageNet-LT's validation list is drawn from the original ImageNet training collection; its final test list uses the original ImageNet validation collection (Jamal et al., 2020). Consequently, a checkpoint pretrained on the full original ImageNet training set is not eligible for this calibration protocol when that pretraining includes calibration images. The same exclusion applies to any Places checkpoint that has seen its calibration images during fitting or model selection.

iNaturalist 2018. Source calibration is reserved exclusively from images indexed by the official training JSON file before fitting. The original training collection contains 437,513 images in 8,142 classes (Van Horn et al., 2018). Writing N_c^{raw} for the original class count, reserve

$$h_c = \min\{N_c^{\mathrm{raw}} - 1,\ 100,\ \max\{3, \lfloor 0.2 N_c^{\mathrm{raw}} \rfloor\}\}, \qquad N_c = N_c^{\mathrm{raw}} - h_c. \tag{A.2}$$

This count-based minimum-support rule ordinarily reserves 20%, caps the reservation at 100, and retains at least one fitting image per class. It is not a strict 20% holdout for the smallest classes. For $h_c \geq 3$, the disjoint roles are

$$n_c^{\mathrm{id}} = \max\{1, \lfloor 0.4 h_c \rfloor\}, \quad n_c^{\mathrm{sel}} = \max\{1, \lfloor 0.2 h_c \rfloor\}, \quad n_c^{\mathrm{risk}} = h_c - n_c^{\mathrm{id}} - n_c^{\mathrm{sel}}. \tag{A.3}$$

Thus 3, 20, and 100 reserved images yield $1/1/1$, $8/4/8$, and $40/20/40$. The exceptional class with two original training images retains one for fitting and assigns the other to selection, giving $(n_c^{\mathrm{id}}, n_c^{\mathrm{sel}}, n_c^{\mathrm{risk}}) = (0, 1, 0)$. Identification and audit charge that class's missing mass using Eqs 5 and 8; no image is reused across roles. The realized role manifest covers every class in selection and leaves one class unobserved in each of identification and audit.

Table S2 records the realized image budget. Its roles sum to 78,289 calibration images, and calibration plus fitting sums to the original 437,513 images. Calibration images are excluded from backbone fitting, prototype estimation, witness-temperature selection, and checkpoint selection. Both members of every iNaturalist recognition pair in Table 2 use the same checkpoint fitted on the remaining 359,224 images. Complete selection support removes the missing-selection-class cause of the reference-fit fallback in Appendix B.2; nonfinite fitting outputs still invoke its numerical fallback. This allocation improves role coverage, not necessarily every direction's audit power.

Table S2: Realized iNaturalist 2018 fitting and calibration inventory. Counts are distinct images; a missing class has no observation in that role. Target/evaluation images are outside this training-origin inventory.

Role	Images	Unobserved classes
Model and witness fitting	359,224	0
Identification S_{id}	28,872	1
Reference fitting and selection S_{sel}	16,512	0
Independent audit S_{risk}	32,905	1
Calibration total	78,289	–
Training-origin total	437,513	0

Full-validation target access. For the iNaturalist recognition comparison, T comprises all 24,426 images indexed by the official validation JSON file, with $m = 24{,}426$. Adaptation accesses these images but not their labels. After candidate selection and the independent source audit fix the returned action, the labels associated with the same complete validation set are used to compute the reported classwise and aggregate accuracies. No validation image enters labeled source calibration, no validation image is discarded from evaluation, and no additional target pool is introduced. The complete-validation decision is run once per outer recognition run; the 100-draw controlled-prior convention described below corresponds to a separate evaluation protocol.

This is label-hidden transductive evaluation: the target inputs and evaluated inputs coincide, whereas the source audit remains independent of them. The reported accuracy is a descriptive performance measure on the accessed target pool, not a fresh post-adaptation test estimate. Reusing target inputs for evaluation does not expose labels during adaptation, and it does not create an additional independent evaluation sample. The population guarantees in Section 4 retain their stated source/target sampling and label-shift assumptions; they do not automatically become confidence statements about the realized transductive accuracy. In particular, a fixed benchmark pool and an i.i.d. target sample are not identified merely by hiding their labels. The source-only marginal audit remains conditional on the complete pre-audit information, including the target inputs.

A.2 Role allocation and fixed frequency groups

For CIFAR-100-LT, ImageNet-LT, Places-LT, and the ordinary capped-resource controls, a fixed classwise random permutation assigns the realized pool of h_c distinct images as

$$n_c^{\mathrm{id}} = \lfloor 0.4 h_c \rfloor, \qquad n_c^{\mathrm{sel}} = \lfloor 0.2 h_c \rfloor, \qquad n_c^{\mathrm{risk}} = h_c - n_c^{\mathrm{id}} - n_c^{\mathrm{sel}}. \tag{A.4}$$

All allocations are fixed from class counts and seeds without examining predictions or candidate losses. iNaturalist instead uses the minimum-support rule in Eq A.3 and the stated rare-class exception. The audit uses the actual count vector, including zeros. Calibration images left unused by a resource cap remain excluded from fitting; a smaller cap does not retrospectively enlarge the training set.

Many contains classes with actual fitting count $N_c > 100$, Medium those with $20 \le N_c \le 100$, and Few those with $N_c < 20$. Groups are determined after reservation and remain fixed during adaptation and evaluation. Table S3 records their cardinalities. Empty groups have no conditional metric or risk check and are not replaced by arbitrary rank bins.

For example, the CIFAR IR=100 class-uniform masses are $(0.30, 0.35, 0.35)$ and the iNaturalist masses are $(698, 2929, 4515)/8142$. These are evaluation weights, not source training priors or

Table S3: Post-reservation frequency partitions used in the recognition tables. Class-uniform group masses are the displayed cardinalities divided by C. Every partition uses the actual remaining training counts.

Dataset / setting	C	Many	Medium	Few
CIFAR-100-LT, IR=100	100	30	35	35
CIFAR-100-LT, IR=50	100	35	41	24
CIFAR-100-LT, IR=10	100	60	40	0
ImageNet-LT	1000	385	479	136
iNaturalist 2018	8142	698	2929	4515
Places-LT	365	131	163	71

calibration allocation proportions. The source prior uses the full actual training-count vector, $q_{s,c} = N_c / \sum_k N_k$; group cardinalities alone cannot reconstruct it.

A.3 Independence, checkpoint eligibility, and data access

Calibration exclusions cover classifier training, pretraining, prototype estimation, samplewise witness fitting, checkpoint selection, and hyperparameter search. The training and fitting manifests therefore precede calibration access. A pretrained model is eligible only when its fitting history excludes the calibration observations; a new directory partition cannot undo exposure during pretraining. Within each calibration pool, source roles are disjoint and assigned without inspecting features or candidate losses. Selection receives audit class counts but no audit examples, predictions, or paired losses.

The independence and label-shift assumptions of Section 3.1 are population assumptions. Disjoint file lists are necessary provenance checks, but by themselves do not establish equality of class-conditional source and target laws. Dataset-native training/validation differences and unrecorded pretraining overlap can invalidate transport of the certificate. Final evaluation labels remain unavailable to adaptation and parameter choice.

Each image configuration is indexed by its dataset/version, learner, checkpoint, fitting exclusions, original and remaining training counts, calibration-role manifests, target inputs, and evaluation inputs. The target-access rule distinguishes an i.i.d. draw, fixed quotas, and sampling from a finite reservoir. The full iNaturalist recognition pool follows Appendix A.1; the controlled-prior records use their separately declared target sampling and denominators.

Scope of the recognition records. This appendix specifies Tables 1–2, including all twenty iNaturalist baseline/adaptation rows. The controlled-prior deployment and availability diagnostics in Appendices G and I are separate records; their prior weights and decision counts are not inferred from the recognition accuracies. A single full-validation iNaturalist recognition batch is not a six-prior deployment experiment. The CIFAR resource comparison in Table 6 is inventory-limited to 100 images per class even when its nominal cap is 200.

B Reference Fitting and Selection Details

B.1 Frozen witnesses and source class means

Let $\mathcal{N}(v) = v/\|v\|_2$ for nonzero v and $\mathcal{N}(0) = 0$. With frozen feature extractor F_0 and class-c fitting indices I_c^{fit}, set $f(x) = \mathcal{N}(F_0(x))$ and $\mu_c = \mathcal{N}(|I_c^{\mathrm{fit}}|^{-1} \sum_{i \in I_c^{\mathrm{fit}}} f(X_i))$ when the fitting class is present, otherwise $\mu_c = 0$. These fitting observations may overlap model training but not any calibration split or the target batch. Neither prototypes nor temperatures are refitted from target observations. The two softmax maps are concatenated in the fixed order $\phi = (\phi_z, \phi_p) \in [0,1]^{2C}$.

For each observed identification class, $\widehat{M}_{jc} = (n_c^{\mathrm{id}})^{-1} \sum_{i \in I_c^{\mathrm{id}}} \phi_j(X_i)$; for an absent class the whole column is zero. Target means are $\widehat{m}_j = m^{-1} \sum_{i=1}^{m} \phi_j(X_i^t)$. The resulting paired differences and squared differences are always evaluated with the same fixed class-priority rule. Calibration counts include their zeros and are not changed in response to feature values or candidate losses.

B.2 SOURCE REFERENCE FITTING

Let p^{ref} be a strictly positive simplex vector fixed from source information; the default is the normalized training-count vector. When every selection class is observed, fit a centered bias by minimizing

$$\mathcal{L}_{\mathrm{ref}}(b) = \sum_c \frac{p_c^{\mathrm{ref}}}{n_c^{\mathrm{sel}}} \sum_{i \in I_c^{\mathrm{sel}}} -\log \operatorname{softmax}(z_0(X_i) + b)_c + \frac{\lambda_b}{C}\|b\|_2^2, \qquad \mathbf{1}^\top b = 0, \quad \lambda_b > 0. \tag{B.1}$$

The positive quadratic term gives a unique exact minimizer on the centered subspace. A finite numerical iterate $\widehat{b}$ is sufficient for the independent risk guarantee. Set

$$\rho^{\mathrm{raw}} = \operatorname{softmax}(\log p^{\mathrm{ref}} - \widehat{b}), \qquad \widehat{\rho}_c = \frac{\max\{\rho_c^{\mathrm{raw}}, \epsilon_\rho\}}{\sum_k \max\{\rho_k^{\mathrm{raw}}, \epsilon_\rho\}}, \quad \epsilon_\rho \geq 0. \tag{B.2}$$

An absent selection class or a nonfinite fitting output invokes the fallback $\widehat{b} = 0$, $\widehat{\rho} = p^{\mathrm{ref}}$. The reference convention is part of the fixed proposal construction. Before smoothing and flooring, the relation

$$z_0 + \log q - \log \rho^{\mathrm{raw}} = z_0 + \widehat{b} + \log q - \log p^{\mathrm{ref}} + \text{constant} \cdot \mathbf{1}$$

explains its sign. The deployed baseline is not recalibrated on a preservation branch.

Reference fitting and candidate scoring may share S_{sel}. The forecast scores are not confidence bounds, and finiteness of the fitted candidate list is not used to assert a selection-split concentration result. The actual audit is independent of this entire search.

B.3 FORECASTS AND TERMINAL DECISIONS

Selection computes $Z_i(a)$ and $Z_i(a)^2$ on its own observations. Missing selection classes receive forecast values $\widehat{d}_c^{\mathrm{sel}} = -1$ and $\widehat{p}_c^{\mathrm{sel}} = 1$ for nonbaseline candidates. These are pessimistic planning values; they do not impute observations. The baseline forecast is exactly zero because comparing the baseline with itself has zero paired loss for every input.

The statistics η_w, V_w, b_w used for forecasts are computed from the actual planned audit class counts. Selection does not access audit features, predictions, second moments, or errors. Every forecast is computed from Eq 12 with the same numerical radius function used in the later audit. This favors candidates with both useful estimated gain and plausible audit margins, without claiming an optimal probability of passing.

The baseline belongs to $\mathcal{F}_{\mathrm{sel}}$ because budgets are nonnegative. A selected nonbaseline candidate has strictly positive worst-case overall forecast by the baseline-first tie rule. An empty proposal grid or empty selection data returns the baseline. All candidate parameters, including the reference, prior index, strength, and tie-breaking rule, are frozen before the audit.

The terminal outcomes are empty compatibility, protected compatibility, baseline selection, audit rejection, and execution. The reference-fitting fallback is applied before selection. Other nonfinite computations cause preservation: an identification failure receives the empty-compatibility code, a forecast failure the baseline-selection code, and an audit failure the rejection code, with a separate numerical-failure flag. Early returns leave downstream statistics undefined; a rejected audit is never reused to select a replacement. The proofs concern exact arithmetic. Their finite-precision counterpart excludes a template only when a conservative lower enclosure of its residual exceeds an upper enclosure of its radius, and releases a correction only when a lower enclosure of every audit margin is nonnegative. These outward-compatible and inward-release comparisons preserve the risk argument. Merely detecting nonfinite values is insufficient to certify all rounding effects. An enlarged numerical compatibility tolerance also enlarges the moment-distance allowance in an execution-power analysis.

B.4 COMPLETE DEPLOYMENT ALGORITHM AND DATA ACCESS

Algorithm 1 implements the procedure for one deployment decision. Its fixed configuration $\mathcal{C}$ comprises finite score and feature functions, fitted prototypes, positive training counts, nonempty groups, an ordered simplex-valued prior dictionary, the protected subset, positive confidence allocations,

nonnegative budgets, the reference-fitting rule, a finite nonnegative strength grid, and deterministic tie rules. Only the source-reference parameters and selected candidate are fitted during the pre-audit stages. Neither population gains d, change probabilities p_c, nor the true target prior are implementation inputs.

Algorithm 1 TailSafe for one deployment decision

Input: Frozen configuration $\mathcal{C}$; independent $S_{\mathrm{id}}, S_{\mathrm{sel}}, S_{\mathrm{risk}}$; unlabeled target batch T.
Input: Audit class counts $\mathbf{n}^{\mathrm{risk}}$; audit features remain sealed until line 26.
Output: Returned score function $z_{\widehat{a}}$ and reason $\in \{\texttt{empty}, \texttt{protect}, \texttt{baseline}, \texttt{reject}, \texttt{execute}\}$.

1: Compute $\widehat{M}, \widehat{m}$ from S_{id}, T using Eq 4.
2: Compute $r_{\mathrm{id}}(q)$ for all $q \in \mathcal{Q}$ and construct $\mathcal{U}_T$ by Eq 5.
3: **if** $\mathcal{U}_T = \varnothing$ **then**
4: **return** $(z_0, \texttt{empty})$
5: **end if**
6: **if** $\mathcal{U}_T \cap \mathcal{P}_0 \neq \varnothing$ **then**
7: **return** $(z_0, \texttt{protect})$
8: **end if**
9: **if** $S_{\mathrm{sel}} = \varnothing$ or $\Lambda \cap (0, \lambda_{\max}] = \varnothing$ **then**
10: **return** $(z_0, \texttt{baseline})$
11: **end if**
12: Fit $\widehat{\rho}$ on S_{sel} with the fallback of Appendix B.2.
13: Construct ordered $\mathcal{A}_T$ using Eq 11; represent zero-strength actions by a_0.
14: $\mathcal{F}_{\mathrm{sel}} \leftarrow \{a_0\}$; $\mathrm{score}(a_0) \leftarrow 0$.
15: **for all** $a \in \mathcal{A}_T \setminus \{a_0\}$ in the fixed order **do**
16: Estimate $\widehat{d}_c^{\mathrm{sel}}(a), \widehat{p}_c^{\mathrm{sel}}(a)$; use $(-1, 1)$ when $n_c^{\mathrm{sel}} = 0$.
17: Compute $\widetilde{L}_{w_h(q)}(a)$ by Eq 12, using $\mathbf{n}^{\mathrm{risk}}$ only.
18: **if** $\widetilde{L}_{w_h(q)}(a) + \beta_h \geq 0$ for all $q \in \mathcal{U}_T$, $h \in \mathcal{H}(q)$ **then**
19: $\mathcal{F}_{\mathrm{sel}} \leftarrow \mathcal{F}_{\mathrm{sel}} \cup \{a\}$; $\mathrm{score}(a) \leftarrow \min_{q \in \mathcal{U}_T} \widetilde{L}_{w_0(q)}(a)$.
20: **end if**
21: **end for**
22: $a^\star \leftarrow \arg\max_{a \in \mathcal{F}_{\mathrm{sel}}} \mathrm{score}(a)$, with baseline-first deterministic ties.
23: **if** $a^\star = a_0$ **then**
24: **return** $(z_0, \texttt{baseline})$
25: **end if**
26: Freeze $a^\star$, $\widehat{\rho}$, strength, and tie rule; then unseal audit features.
27: On S_{risk}, compute class means $\widehat{d}_c(a^\star)$ of Z_i and $\widehat{p}_c(a^\star)$ of Z_i^2.
28: **for all** $q \in \mathcal{U}_T$ **do**
29: **for all** $h \in \mathcal{H}(q)$ **do**
30: $w \leftarrow w_h(q)$; compute $\eta_w, V_w, b_w, \widehat{G}_w, \widehat{v}_w$ by Eq 8.
31: Compute $L_w(a^\star; \delta_{\mathrm{r}})$ by Eq 9.
32: **if** an audit quantity is nonfinite or $L_w(a^\star; \delta_{\mathrm{r}}) + \beta_h < 0$ **then**
33: **return** $(z_0, \texttt{reject})$
34: **end if**
35: **end for**
36: **end for**
37: **return** $(z_{a^\star}, \texttt{execute})$

Algorithm 1 abbreviates the pre-audit numerical guards using the failure conventions above. All quantities used for a decision are logged before returning; later audit directions may be evaluated for diagnostics only after the output is irrevocably fixed, with no reselection. The algorithm accepts any fixed nonempty collection of class groups. Empty bins are removed before forming this collection; zero probability mass under an individual prior simply omits that conditional group direction. An empty candidate grid, empty selection data, or a selected canonical zero-strength correction returns the baseline. No score calibration, probability smoothing, interpolated logits, or batch-dependent tie breaking is applied to a preservation output.

Table S4: Allowed data access for one deployment decision. Evaluation labels remain sealed until the returned action is fixed. In the transductive protocol, evaluation inputs have already been accessed as unlabeled target data.

Stage	Accessible observations	Output
Model/witness fitting	Independent fitting data	Frozen scores, features, prototypes
Identification	S_{id} and unlabeled T	Compatible priors
Selection	S_{sel}, compatibility, audit counts only	One frozen candidate
Audit	S_{risk} and that candidate	Execute or preserve
Evaluation	Evaluation labels and fixed predictions; input reuse follows Appendix A.1	Paired metrics and diagnostics

C Finite-Sample Risk Control

C.1 Conditioning and probability space

Fix a generating prior π_t and condition on independent fitting data, the deployment configuration, and all source labels and split assignments. These fix class counts and the dictionary-wide risk directions. Calibration features remain independent with the specified class-conditional laws, and target inputs remain independent draws from the target marginal. The sigma-field $\mathcal{F}_{\mathrm{pre}}$ additionally contains identification and selection features, the unlabeled target batch, and any independent proposal randomness, but no audit features. It fixes the selected action $a^\star$ and $\mathcal{U}_T$. On an early-return branch, define $a^\star = a_0$ for the proof without performing an audit.

C.2 Directional moment concentration

Lemma 1 (Directional compatibility event)**.** *Conditional on the source labels, Eq 6 holds simultaneously for all $q \in \mathcal{Q}$ with probability at least $1 - \delta_{\mathrm{id}}$.*

Proof. For a fixed coordinate j and fixed dictionary vector q, write

$$\widehat{m}_j - (\widehat{Mq})_j - [M(\pi_t - q)]_j = A_{j,q} + B_{j,q},$$

$$A_{j,q} = \frac{1}{m}\sum_{i=1}^{m}\{\phi_j(X_i^t) - (M\pi_t)_j\} - \sum_{c\in\mathcal{O}_{\mathrm{id}}}\frac{q_c}{n_c^{\mathrm{id}}}\sum_{i\in I_c^{\mathrm{id}}}\{\phi_j(X_i) - M_{jc}\},$$

$$B_{j,q} = \sum_{c\notin\mathcal{O}_{\mathrm{id}}} q_c M_{jc}.$$

The summands in $A_{j,q}$ are independent and centered. Their squared range lengths sum to

$$m(1/m)^2 + \sum_{c\in\mathcal{O}_{\mathrm{id}}} n_c^{\mathrm{id}}(q_c/n_c^{\mathrm{id}})^2 = V_{\mathrm{id}}(q).$$

Hoeffding's inequality yields

$$\Pr\left(|A_{j,q}| > \sqrt{\frac{V_{\mathrm{id}}(q)}{2}\log\frac{2JL}{\delta_{\mathrm{id}}}}\right) \le \frac{\delta_{\mathrm{id}}}{JL}.$$

A union bound over the JL coordinate–template pairs gives a common event. On this event $0 \le B_{j,q} \le \eta_{\mathrm{id}}(q)$, so Eq 6 follows. Setting $q = \pi_t$ gives coverage. For a retained q, the triangle inequality between its empirical compatibility residual and the population residual gives

$$\|M(q - \pi_t)\|_\infty \le 2r_{\mathrm{id}}(q).$$

Coordinate independence is not required. Averaging over random source labels preserves the probability statement. □

C.3 A MARGINAL CHANGE-ADAPTIVE AUDIT BOUND

Fix $\mathcal{F}_{\mathrm{pre}}$ and one direction w. Index all positively weighted observed audit examples by i, with coefficient $a_i = w_c/n_c^{\mathrm{risk}}$ for a class-c example. Then

$$\widehat{G} = \sum_i a_i Z_i, \quad G = \mathbb{E}\widehat{G}, \quad V = \sum_i a_i^2, \quad b = \max_i a_i, \quad U = \sum_i a_i^2 Z_i^2, \quad \nu = \mathbb{E}U.$$

These equal $\widehat{G}_w$, the observed population contribution, V_w, b_w, $\widehat{v}_w$, and ν_w, respectively. The total variance of $\widehat{G}$ is at most ν, and $\nu \le V$. The missing contribution B_{miss} obeys

$$w^\top d = G + B_{\mathrm{miss}}, \qquad -\eta_w \le B_{\mathrm{miss}} \le \eta_w.$$

Lemma 2 (Weighted second-moment tails)**.** *For $b > 0$ and $t > 0$,*

$$\Pr(\nu > 2U + 2b^2 t) \le e^{-t}, \qquad \Pr(U > 2\nu + b^2 t) \le e^{-t}. \tag{C.1}$$

Proof. Let $Y_i = (a_i/b)^2 Z_i^2 \in [0,1]$; these variables are independent. Convexity gives $e^{-y} \le 1 - (1-e^{-1})y$ and $e^y \le 1 + (e-1)y$ for $0 \le y \le 1$. Consequently

$$\mathbb{E}e^{-U/b^2} \le e^{-(1-e^{-1})\nu/b^2}, \qquad \mathbb{E}e^{U/b^2} \le e^{(e-1)\nu/b^2}.$$

Since $1 - e^{-1} \ge 1/2$ and $e - 1 \le 2$,

$$\mathbb{E}\exp\{\nu/(2b^2) - U/b^2\} \le 1, \qquad \mathbb{E}\exp\{U/b^2 - 2\nu/b^2\} \le 1.$$

Markov's inequality proves the two claims. If $b = 0$, all quantities are zero and the corresponding events are empty. □

Lemma 3 (Weighted paired-mean tails)**.** *For $t > 0$,*

$$\Pr(\widehat{G} - G > \sqrt{2Vt}) \le e^{-t}, \qquad \Pr\left(\widehat{G} - G > \sqrt{2\nu t} + \frac{4}{3}bt\right) \le e^{-t}. \tag{C.2}$$

The same two bounds hold with $G - \widehat{G}$ in place of $\widehat{G} - G$.

Proof. The range lengths of the independent variables $a_i Z_i$ are $2a_i$, giving the first claim by Hoeffding concentration. Their centered versions are bounded in absolute value by $2b$ and have total variance at most ν. Bernstein's inequality therefore gives

$$\Pr(\widehat{G} - G > x) \le \exp\left\{-\frac{x^2}{2(\nu + 2bx/3)}\right\}.$$

For $x = \sqrt{2\nu t} + 4bt/3$, the exponent is at most $-t$, establishing the displayed bound. Equivalently, Bernstein's inequality follows from the exponential-moment expansion for independent centered bounded summands. Applying these arguments to $-Z_i$ proves the lower-tail statements. The zero-coefficient case is deterministic. □

Lemma 4 (Marginal lower-bound validity)**.** *For $0 < \alpha < 1$, the bound in Eq 9 satisfies Eq 10.*

Proof. Set $t = \ell_\alpha = \log(3/\alpha)$. Outside three events, each of probability at most $e^{-t} = \alpha/3$,

$$\widehat{G} - G \le \sqrt{2Vt}, \qquad \nu \le \min\{V, 2U + 2b^2 t\}, \qquad \widehat{G} - G \le \sqrt{2\nu t} + \frac{4}{3}bt.$$

The last two imply the second data-dependent radius; their intersection with the first implies

$$\widehat{G} - G \le \min\left\{\sqrt{2Vt},\ \sqrt{2\min\{V, 2U + 2b^2 t\}t} + \frac{4}{3}bt\right\} = r_w(\alpha; \widehat{v}_w).$$

Thus the data-dependent minimum is covered by the allocation already included in $\log(3/\alpha)$. No independence between U and $\widehat{G}$ is assumed. Adding the missing-mass allowance gives

$$\widehat{G} - \eta_w - r_w \le G - \eta_w \le w^\top d.$$

Clipping at -1 preserves this inequality because $w^\top d \ge -1$. A union bound over the three events establishes failure probability at most α. This union bound is within one direction; no union over directions has been taken. □

C.4 Proof of the complete release guarantee

Proof of Theorem 1. Identification coverage follows from Appendix C.2. Condition now on $\mathcal{F}_{\mathrm{pre}}$. If there is an early return, the final action is the baseline and all its risk differences are zero. Otherwise the candidate and all compatible directions are fixed.

Define the retained population violation set

$$\mathcal{V} = \{(q, h) : q \in \mathcal{U}_T,\ h \in \mathcal{H}(q),\ w_h(q)^\top d(a^\star) + \beta_h < 0\}.$$

If $\mathcal{V} = \varnothing$, the candidate is feasible for every retained constraint, so either possible output is feasible for those constraints. If $\mathcal{V} \neq \varnothing$, choose its first pair $(q_\dagger, h_\dagger)$ in the predetermined order. This theoretical choice is fixed under the conditioning; it need not be known by the procedure. Acceptance implies

$$L_{w_{h_\dagger}(q_\dagger)}(a^\star; \delta_{\mathrm{r}}) \geq -\beta_{h_\dagger} > w_{h_\dagger}(q_\dagger)^\top d(a^\star).$$

Marginal validity bounds this event by δ_{r}. Therefore, conditional on every pre-audit realization, the probability that the final output violates any retained constraint is at most δ_{r}. Dependence among tests is unrestricted. There is no union bound over the unknown violating pairs: passing all tests implies an error for the one chosen pair.

On $\mathcal{E}_{\mathrm{id}}$ the true prior and every positive-mass true group are retained. Target violation is thus contained in identification failure or erroneous release for a retained constraint. Averaging the conditional bound gives

$$\Pr(\text{any target budget violation}) \leq \delta_{\mathrm{id}} + \delta_{\mathrm{r}}.$$

If $\pi_t \in \mathcal{P}_0$, coverage implies that a protected prior is retained, which triggers preservation before auditing. Hence $\Pr(\widehat{a} \neq a_0) \leq \Pr(\mathcal{E}_{\mathrm{id}}^c) \leq \delta_{\mathrm{id}}$. □

This argument applies to any single candidate measurable with respect to pre-audit information. It does not require candidate optimality, calibrated network probabilities, or inversion of a response matrix. It also does not make the collection $\{L_w\}$ a simultaneous confidence band. The joint object controlled is feasibility of the final output under an all-tests-pass rule.

D Decision-Relevant Identification and Execution Power

D.1 Exact ambiguity duality

Proof of Proposition 1. Exchanging q and q' leaves the feasible set unchanged. The supremum of the absolute objective therefore equals the maximum of the signed objective $u^\top(q - q')$. Consider the linear program

$$\max_{q, q' \geq 0} u^\top q - u^\top q' \quad \text{subject to} \quad \mathbf{1}^\top q = \mathbf{1}^\top q' = 1, \quad M(q - q') \leq \varepsilon \mathbf{1}, \quad -M(q - q') \leq \varepsilon \mathbf{1}.$$

It is feasible and compact. Introduce nonnegative multipliers $\alpha, \zeta \in \mathbb{R}^J$ for the two inequalities and unrestricted s, t for the equalities. Linear-programming duality yields

$$\min_{\alpha, \zeta \geq 0, s, t} \varepsilon \mathbf{1}^\top(\alpha + \zeta) + s + t \quad \text{subject to} \quad u - M^\top(\alpha - \zeta) \leq s\mathbf{1}, \quad -u + M^\top(\alpha - \zeta) \leq t\mathbf{1}.$$

The dual is feasible, for example at $\alpha = \zeta = 0$, $s = \max_c u_c$, $t = -\min_c u_c$. For a fixed $v = \alpha - \zeta$, the smallest $s + t$ is $\operatorname{osc}(u - M^\top v)$, and the smallest $\mathbf{1}^\top(\alpha + \zeta)$ is $\|v\|_1$. Substitution gives the stated identity. Strong duality for this finite linear program includes $\varepsilon = 0$; no strict-feasibility assumption is needed. □

D.2 Moment fibers and budget decisions

For a fixed action and target prior, let $\mathcal{F}_M(\pi_t) = \{q \in \mathcal{Q} : Mq = M\pi_t\}$. Within this finite family, the prior is identified exactly when the fiber is a singleton. A linear margin $q^\top u_h(a)$ is identified exactly when

$$u_h(a)^\top(q - q') = 0 \quad \text{for all } q, q' \in \mathcal{F}_M(\pi_t).$$

The budget indicator for direction h, using zero unnormalized margin for a zero-mass group, is identified exactly when

$$\min_{q \in \mathcal{F}_M(\pi_t)} q^\top u_h(a) \geq 0 \quad \text{or} \quad \max_{q \in \mathcal{F}_M(\pi_t)} q^\top u_h(a) < 0.$$

These statements follow by subtracting linear margins and by checking whether the Boolean budget indicator is constant, respectively. Identification of the conjunction of all budgets only requires that conjunction to be constant; it need not require every individual indicator to be constant on an everywhere-infeasible fiber.

A row-space representation $u_h = M^\top v + b\mathbf{1}$ is sufficient for margin identification. On a particular finite dictionary it is not necessary: orthogonality only to the actual within-fiber differences is the exact criterion. Likewise, budget feasibility can be identified without identifying the margin value.

On $\mathcal{E}_{\mathrm{id}}$, every fiber member is retained by Eq 6. A protected member therefore forces preservation on that event. For the intersection–union audit, an inadmissible retained candidate is released with conditional probability at most δ_{r} by the proof in Appendix C.4. This concerns the all-tests-pass decision and does not require simultaneous validity of all numerical lower bounds. Equality of witness moments is not equality of the complete observable-data laws, so the fiber interpretation concerns the chosen moment procedure.

D.3 Power of the independent audit

Fix a pre-audit realization with a selected nonbaseline action and a nonempty, unprotected compatible set. All quantities in $E_w(\kappa)$ are fixed under this conditioning, including the population change probabilities of the fixed candidate. Let K_T count retained prior–constraint pairs as in the main text; counting duplicates is harmless.

Lemma 5 (Uniform upper bound on the audit gap)**.** *With conditional probability at least $1 - \kappa$, every retained direction satisfies*

$$w^\top d(a^\star) - L_w(a^\star; \delta_{\mathrm{r}}) \leq E_w(\kappa). \tag{D.1}$$

No lower bound of zero on this gap is asserted on the same event.

Proof. Write $\ell = \log(3/\delta_{\mathrm{r}})$, $t = \log(2K_T/\kappa)$, and $s = \ell + t$. For each retained direction, the second upper tail in Eq C.1 and the lower Bernstein tail in Eq C.2 imply

$$U \leq 2\nu + b^2 t, \qquad G - \widehat{G} \leq \sqrt{2\nu t} + \frac{4}{3} b t$$

outside two events each of probability at most e^{-t}. A union bound over at most K_T pairs gives a common event of conditional probability at least $1 - \kappa$.

On this event,

$$\overline{v}_w(\delta_{\mathrm{r}}; U) \leq 2U + 2b^2\ell \leq 4\nu + 2b^2 s.$$

The adaptive branch of the radius, whether or not it is selected by the minimum, therefore gives

$$\begin{aligned} r_w(\delta_{\mathrm{r}}; U) &\leq \sqrt{8\nu\ell + 4b^2\ell s} + \frac{4}{3} b\ell \\ &\leq \sqrt{8\nu\ell} + 2b\sqrt{\ell s} + \frac{4}{3} b\ell. \end{aligned}$$

Since clipping the lower bound can only decrease $w^\top d - L_w$,

$$\begin{aligned} w^\top d - L_w &\leq 2\eta_w + (G - \widehat{G}) + r_w \\ &\leq 2\eta_w + \sqrt{2\nu t} + \sqrt{8\nu\ell} + 2b\sqrt{\ell s} + \frac{4}{3} b s \\ &\leq 2\eta_w + \sqrt{10\nu s} + \frac{10}{3} b s. \end{aligned}$$

The final step uses Cauchy–Schwarz on the two square-root terms and $\sqrt{\ell s} \leq s$. This is Eq D.1. When $b = 0$ the observed contribution vanishes and the deterministic missing-mass argument gives the same result. □

Local sufficient condition. A sharper condition than Eq 19 is

$$w_h(q)^\top d(a^\star) + \beta_h \geq E_{w_h(q)}(\kappa) \quad \text{for all retained } (q,h). \tag{D.2}$$

On the event of the preceding lemma, every numerical audit margin is then nonnegative, so the action is executed. This establishes conditional power at least $1-\kappa$ without any identification assumption once the retained set is fixed.

Proof of Theorem 2. On $\mathcal{E}_{\mathrm{id}}$, every retained prior satisfies $\|M(q-\pi_t)\|_\infty \leq 2r_{\max}$. For each relevant h, the definition of the modulus gives

$$q^\top u_h(a^\star) \geq \pi_t^\top u_h(a^\star) - \omega_M(u_h(a^\star), 2r_{\max}).$$

The theorem's margin condition therefore implies

$$q^\top u_h(a^\star) \geq s_h(q) E_{w_h(q)}(\kappa) \quad \text{for each } q \in \mathcal{U}_{T,h}.$$

Dividing only when $s_h(q) > 0$ yields Eq D.2. Zero-mass group constraints are vacuous. The uniform audit-gap lemma gives execution probability at least $1-\kappa$ conditional on the specified pre-audit realization. □

The value κ indexes a theoretical power guarantee and is not a tuned audit threshold. The event $\mathcal{E}_{\mathrm{id}}$ has probability at least $1-\delta_{\mathrm{id}}$, but the theorem does not assert that its margin condition holds on every realization or that selection finds the best feasible candidate. It is a sufficient execution statement for the independently selected action. In particular, the argument does not substitute the power event for the different intersection–union risk-control proof.

D.4 Effective sample size and unchanged directions

With no missing positive weights, $n_{\mathrm{eff}}(w) = 1/V_w$. Cauchy–Schwarz gives

$$n_{\mathrm{eff}}(w) \leq \sum_{c: w_c > 0} n_c^{\mathrm{risk}},$$

with equality when class weights are proportional to their audit counts. The second-moment proxy is $\nu_w = \sum_{c:w_c>0} w_c^2 p_c / n_c^{\mathrm{risk}}$; it may be substantially smaller than V_w for localized corrections. A concentrated large weight can still dominate the linear term b_w, even if total audit size is large.

If every observed paired change in a direction is zero, then $U = 0$ and $\overline{v}_w(\alpha; 0) \leq 2b_w^2 \ell_\alpha$. Substitution proves

$$r_w(\alpha; 0) \leq \min\{\sqrt{2V_w \ell_\alpha}, (10/3) b_w \ell_\alpha\}.$$

For no missing mass, zero empirical gain, and equal per-example weights $1/N$, a positive budget β_h passes whenever $N \geq 10\log(3/\alpha)/(3\beta_h)$. This is a sufficient algebraic condition for the observed certificate, not an assertion that unobserved population changes are absent. A zero overall budget still needs evidence of positive gain for an ordinary nonbaseline correction; the exactly known baseline comparison is treated separately.

D.5 Explicit sample requirements for uniformly weighted directions

Eq 20 is a consequence of the existing audit-power bound, without changing the algorithm. Consider a direction uniform over C_g classes, each with $n > 0$ audit observations and no missing positive weight. Let $N_g = C_g n$ and $\overline{p}_g = C_g^{-1} \sum_{c \in g} p_c(a^\star)$. Then $V_w = b_w = 1/N_g$, $\nu_w = \overline{p}_g / N_g$, and with $s = \log(3/\delta_{\mathrm{r}}) + \log(2K_T/\kappa)$,

$$E_w(\kappa) = \sqrt{\frac{10\overline{p}_g s}{N_g}} + \frac{10s}{3N_g}.$$

If the population local margin $\gamma_g = w^\top d(a^\star) + \beta_g$ is positive, the two terms are at most $\gamma_g/2$ whenever $N_g \geq 40\overline{p}_g s/\gamma_g^2$ and $N_g \geq 20s/(3\gamma_g)$, respectively. Their sum is then at most γ_g, proving the displayed sufficient condition. To guarantee execution this condition, or the general local condition

of Eq D.2, must hold for every retained direction. It is not a requirement that all dictionary directions be uniform, and this calculation is not used to set counts from unknown population margins. Unequal class weights or counts require the original ν_w, b_w, η_w expressions.

This population-power calculation differs from the certificate evaluated on a realized zero-change audit. The latter uses $\widehat{v}_w = 0$ and still charges a positive finite-sample radius. It does not identify $p_c(a^\star) = 0$ from zero observations of change, nor justify a zero-budget nonbaseline correction without sufficient positive overall evidence. Both statements are analyses of the fixed procedure, not additional empirical findings.

E A Prespecified Template-Neighborhood Extension

The main method uses exact membership in a finite dictionary. This section defines a separate procedure for a bounded template-misspecification model, retaining exact label shift. Fix $\varepsilon_\pi \in [0, 2]$ before adaptation and assume

$$\pi_t \in \bigcup_{q \in \mathcal{Q}} \{p \in \Delta^{C-1} : \|p - q\|_1 \le \varepsilon_\pi\}.$$

For this extension take $0 \le \beta_h \le 1$; budgets greater than one are vacuous and can be omitted. Use

$$\mathcal{U}_T^{\varepsilon_\pi} = \{q \in \mathcal{Q} : \|\widehat{m} - \widehat{M}q\|_\infty \le r_{\mathrm{id}}(q) + \varepsilon_\pi/2\}. \tag{E.1}$$

Protection occurs when this set intersects $\mathcal{P}_0$; the corresponding protected target environments are the union of the neighborhoods of protected templates.

For an independently fixed candidate, define unnormalized audit margins

$$B_h(q, a) = \begin{cases} s_h(q)\{L_{w_h(q)}(a; \delta_{\mathrm{r}}) + \beta_h\}, & s_h(q) > 0, \\ 0, & s_h(q) = 0. \end{cases}$$

The modified release rule requires

$$B_h(q, a) \ge \varepsilon_\pi \quad \text{for all } q \in \mathcal{U}_T^{\varepsilon_\pi} \text{ and } h \in \{0\} \cup \mathcal{G}. \tag{E.2}$$

The candidate is selected before the audit; a compatible-set forecast may use these same inflated margins. Failed identification, protection, or auditing returns the baseline. For $\varepsilon_\pi > 0$, a retained template with zero group mass prevents release under that group's check, because its neighborhood can contain a positive-mass group for which the template supplies no conditional-risk evidence. This behavior follows from the declared robust requirement.

Proposition 2 (Risk control in template neighborhoods)**.** *Under the preceding neighborhood assumption, exact label shift, and the source/target independence protocol, the modified procedure satisfies Eq 2 with failure probability at most $\delta_{\mathrm{id}} + \delta_{\mathrm{r}}$. If the target belongs to a protected template neighborhood, its probability of nonbaseline execution is at most δ_{id}.*

Proof. Choose a template $q_\circ$ with $\|\pi_t - q_\circ\|_1 \le \varepsilon_\pi$. For any vector $f \in [0, 1]^C$ and simplex vectors p, q,

$$|f^\top (p - q)| \le \tfrac{1}{2}\|p - q\|_1.$$

This follows by subtracting the midpoint of the range of f, using $\mathbf{1}^\top (p - q) = 0$. Applying it to every row of M gives $\|M(\pi_t - q_\circ)\|_\infty \le \varepsilon_\pi/2$. The proof of the directional moment event remains valid for an arbitrary fixed π_t, since its centered target moment is $M\pi_t$; only the coverage step in the main method needed dictionary membership. Consequently $q_\circ \in \mathcal{U}_T^{\varepsilon_\pi}$ with probability at least $1 - \delta_{\mathrm{id}}$.

Each budget direction $u_h = e_h \odot (d + \beta_h \mathbf{1})$ has oscillation at most two for $\beta_h \in [0, 1]$. Therefore

$$|(\pi_t - q_\circ)^\top u_h| \le \varepsilon_\pi.$$

If the candidate violates a true positive-mass budget, then $\pi_t^\top u_h < 0$ and $q_\circ^\top u_h < \varepsilon_\pi$. If $s_h(q_\circ) > 0$, acceptance of Eq E.2 forces $L_{w_h(q_\circ)} > w_h(q_\circ)^\top d$. If $s_h(q_\circ) = 0$ and $\varepsilon_\pi > 0$, that check cannot pass. When $\varepsilon_\pi = 0$, $\pi_t = q_\circ$, so a truly positive-mass violated group cannot have zero template mass. Selecting one violated true budget under pre-audit conditioning gives the same marginal-error argument as Appendix C.4. Add the template-retention failure probability to obtain the target risk bound. For a protected target neighborhood, choose a protected $q_\circ$; its retention triggers preservation. □

Setting $\varepsilon_\pi = 0$ recovers the main method's compatibility and release tests. Positive radii define a different configuration, and neither empirical performance nor the exact-dictionary power condition is automatically inherited.

A necessary empirical margin for neighborhood release. Let $\mathcal{G}$ partition the classes, let $G = |\mathcal{G}|$, and suppose that every group has positive mass under a retained template q. Summing its group checks gives

$$\sum_{g\in\mathcal{G}} q(g) L_{w_g(q)}(a;\delta_{\mathrm{r}}) \geq G\varepsilon_\pi - \sum_{g\in\mathcal{G}} q(g)\beta_g. \tag{E.3}$$

Every probability direction satisfies $\widehat{G}_w \geq -1$ and $L_w \leq \widehat{G}_w$, including the clipped and missing-class cases. Since $q(g)w_{g,c}(q) = q_c\mathbf{1}\{c \in g\}$, an executed action therefore obeys

$$\widehat{G}_q(a) = \sum_g q(g)\widehat{G}_{w_g(q)}(a) \geq G\varepsilon_\pi - \sum_g q(g)\beta_g. \tag{E.4}$$

For three positive-mass groups and a common budget 0.02, the lower threshold is 0.04 at $\varepsilon_\pi = 0.02$ and 0.13 at $\varepsilon_\pi = 0.05$. These are necessary bounds on each accepted audit's template-weighted empirical gain. They apply before averaging over deployment decisions; the independently evaluated target gain is a different statistic. Groupwise margins, retained template weights, and the neighborhood radius jointly specify the acceptance event.

F Optional Tail-Macro Protection and Witness Aggregation

When the fixed Few group is nonempty, a deployment policy may additionally include the fixed macro-Few direction

$$w_c^{\mathrm{macro}} = \frac{\mathbf{1}\{c \in G_{\mathrm{Few}}\}}{|G_{\mathrm{Few}}|}.$$

Its paired gain transports under label shift independently of the target prior. Add its prescribed budget to selection and to the all-tests-pass audit. The same intersection–union proof controls the complete enlarged collection without a direction-count factor in the marginal release threshold. The power analysis counts the extra constraint and may require more evidence. This is an optional policy, not a per-class guarantee of the default conditional-group method.

One may also append fixed bounded aggregates $\sum_{c\in g}\phi_{z,c}(x)$ and $\sum_{c\in g}\phi_{p,c}(x)$ to the witness vector, recomputing J and the identification calibration. These statistics are linear combinations of the original coordinates and do not enlarge the population row space. They can nevertheless supply different finite-sample moment constraints because a direct bounded aggregate is concentrated as one statistic. Neither a larger witness vector nor a smaller compatible set implies uniformly better end-to-end performance; the action and its audit margins remain relevant. These aggregates are not part of the default $J = 2C$ construction in the main text.

G Detailed Experimental Protocol

The evaluation distinguishes the dataset-specific recognition records of Appendix A, controlled-prior deployment records, the availability-limited diagnostic, fixed paired-loss experiments, and exact-risk finite-support studies. A comparison is specified by its learner, calibration inventory, target prior, evaluation measure, and decision-rule configuration. The study-specific partitions are explicit: the prior-resolved record in Table S19 and Figure S2, for example, explicitly uses uniform group masses $(0.35, 0.35, 0.30)$, whereas Table 1 at IR=100 uses $(0.30, 0.35, 0.35)$. These records are not pooled or treated as the same run.

G.1 Dataset, checkpoint, and group inventories

The image datasets are CIFAR-10/100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT (Krizhevsky et al., 2009; Cao et al., 2019; Liu et al., 2019; Van Horn et al., 2018). The CIFAR schedule is $N_{\sigma(j)} = \lfloor N_{\max}\,\mathrm{IR}^{-(j-1)/(C-1)}\rfloor$ with fixed class permutation σ. Table 1 uses

$N_{\max} = 400$ after the recognition holdout; the 500-maximum reference remains explicitly identified where used in controlled calculations. Native CIFAR-10 has a 5,000-per-class parent pool and is retained only as an auxiliary reference. Each checkpoint's actual training counts determine its source prior and frequency groups.

BCL is the primary frozen learner in the deployment experiments. ProCo and ICL provide complementary plug-in evaluations, and Balanced Softmax is the architecture-matched rebalancing control. Controlled CIFAR training uses ResNet-32 with 32×32 inputs; the controlled large-scale reference uses ResNet-50 with 224×224 inputs (He et al., 2016). Native checkpoints retain their architecture, expert count, aggregation rule, augmentation, objective, pretraining, and schedule. These checkpoint-specific properties are distinguished from the uniform settings of the controlled comparisons.

Table S5: Controlled model and dataset settings. Native checkpoints retain their original architecture, expert aggregation, and training recipe; the controlled backbones below do not redefine those models. Actual class counts determine the partition of each checkpoint.

Dataset	Controlled backbone / input	Evaluation setting
CIFAR-100-LT	ResNet-32 / 32×32	IR $100/50/10$; $C = 100$
CIFAR-10-LT	ResNet-32 / 32×32	IR 100; rank bins $3/4/3$
ImageNet-LT	ResNet-50 / 224×224	$C = 1000$; fixed frequency bins
iNaturalist 2018	ResNet-50 / 224×224	$C = 8142$; fixed frequency bins
Places-LT	ResNet-50 / 224×224	$C = 365$; fixed frequency bins

Figure S1 gives the separate availability-limited source-evidence diagnostic. Its count vectors and directional summaries are not the complete 100-per-class CIFAR recognition holdout. Stage-isolated CIFAR controls use $40/20/40$ at full allocation. ImageNet-LT and Places-LT recognition use $8/4/8$, and the analytical storage table reports their corresponding actual observation counts.

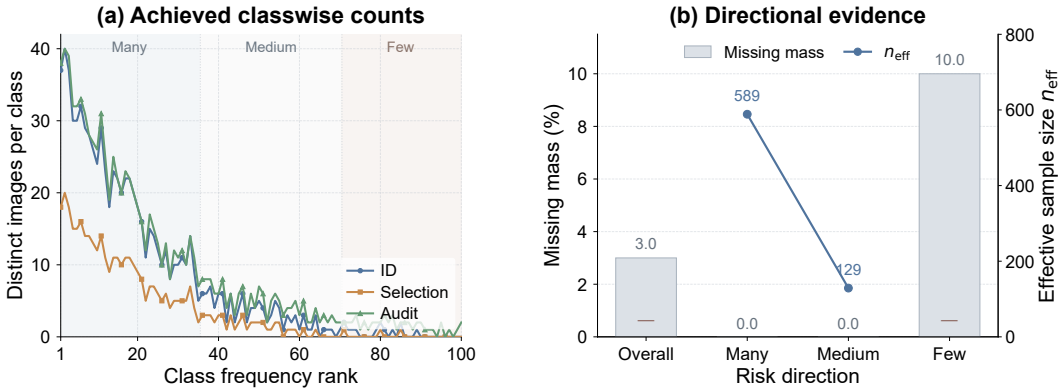


Figure S1: **Availability-limited source-evidence diagnostic.** **(a)** Distinct identification, selection, and audit images per class, ordered by source training frequency; shaded regions denote the fixed Many, Medium, and Few groups. **(b)** Missing mass η_w and directional effective sample size n_{eff} computed from the same audit inventory under the uniform overall and conditional-group directions. Effective sample size is shown only when every positively weighted class has nonzero audit support. The figure illustrates that raw source-label quantity and direction-specific statistical evidence can differ substantially under long-tailed availability.

For controlled CIFAR-100 and large-scale evaluations, Many comprises classes with training count $N_c > 100$, Few those with $N_c < 20$, and Medium those with $20 \leq N_c \leq 100$. Empty bins are omitted from $\mathcal{G}$ and from group-heavy priors; their metrics are inapplicable, not zero. The CIFAR-10 comparison uses a separately declared decreasing-frequency rank partition of $3/4/3$. The risk theorem applies to the fixed collection of nonempty groups and does not require exactly three groups.

Table S3 records the post-reservation groups used in the recognition comparisons. Table S9 separately contrasts these with the 500-maximum reference. The Few bin is empty at IR=10 under both schedules; no rank-based substitute is inserted. The optional macro-Few policy in Appendix F is defined only for a nonempty Few group.

G.2 Independent source resources and sampling

An eligible source-calibration image has not been used to fit the classifier, prototypes, or samplewise witnesses. Independent acquisition and pre-fitting reservation are the two admissible pool constructions; the latter trains both members of every paired comparison on the same remaining examples. Independence is a property of the acquisition and fitting process, not of directory names or an index split performed after fitting. The source-conditional law and any differences between dataset-native training and validation pools determine the population to which calibration evidence applies.

Let H_c be the number of actually eligible class-c images and $H_{\max}$ the design cap. A fixed classwise random permutation selects $h_c = \min\{H_c, H_{\max}\}$ images, then assigns

$$n_c^{\mathrm{id}} = \lfloor 0.4h_c \rfloor, \qquad n_c^{\mathrm{sel}} = \lfloor 0.2h_c \rfloor, \qquad n_c^{\mathrm{risk}} = h_c - n_c^{\mathrm{id}} - n_c^{\mathrm{sel}}. \tag{G.1}$$

The CIFAR source-resource comparison uses nominal caps $25, 50, 100, 200$ on a fixed 100-image-per-class calibration inventory. Its realized class counts are $25, 50, 100, 100$, with role counts $10/5/10$, $20/10/20$, $40/20/40$, and $40/20/40$. The last two settings use the same image/role lists and paired deployment draws, so the 200-cap row is an inventory-truncation control, not a new 200-image experiment. Unused reserved images never return to training. ImageNet-LT and Places-LT recognition have 20 eligible validation images per class, hence $8/4/8$ roles; increasing only a nominal cap cannot expand these pools. iNaturalist uses its count-based minimum-support inventory in Appendix A. Repeated views and forward passes do not increase independent-image counts.

For controlled-prior sampling, labels are drawn from the declared generating prior and hidden from adaptation, and inputs follow the corresponding class-conditional sampling law. Source calibration and target/evaluation roles are kept distinct. Sampling with replacement from a fixed reservoir defines inference for that empirical distribution; it does not create new evidence about an unknown image population. Fixed quotas and sampling without replacement have different probability models. The iNaturalist recognition protocol uses the entire validation input pool once per outer run, as specified in Appendix A.1, and is not assigned this controlled resampling convention.

A target prior is evaluated as the generating distribution, not the random class histogram of its sampled batch. Source-count groups remain fixed throughout the experiment. Neither target labels nor final evaluation outcomes can change groups, dictionary entries, the protection set, confidence allocation, or a hyperparameter.

G.3 Deployment priors and empty-bin handling

Let σ order the actual positive training counts and $N = \sum_c N_c$. Define

$$q_{\mathrm{s},c} = N_c/N, \qquad q_{\mathrm{u},c} = 1/C, \qquad q_{\mathrm{rev},\sigma(j)} = N_{\sigma(C+1-j)}/N, \tag{G.2}$$

$$q_{\mathrm{su}} = (q_{\mathrm{s}} + q_{\mathrm{u}})/2, \qquad q_{\mathrm{ur}} = (q_{\mathrm{u}} + q_{\mathrm{rev}})/2. \tag{G.3}$$

For nonempty Many/Medium/Few bins, group-heavy priors place uniform probability within a bin and nominal group masses $(0.6, 0.2, 0.2)$ or $(0.2, 0.2, 0.6)$. When a bin is empty, drop its nominal mass and renormalize over the nonempty bins; record the resulting full vectors. This is a configuration rule fixed before observing target data, not an inference about an absent target group.

Use $\mathcal{Q} = \{q_s, q_u, q_{\mathrm{rev}}, q_{\mathrm{su}}, q_{\mathrm{ur}}, q_H, q_F\}$ and $\mathcal{P}_0 = \{q_s\}$ before deduplication. Deduplicate exact prior vectors in the declared order and combine their protection flags by logical OR. If a supposedly unprotected template duplicates a protected vector, it is protected. The full vectors, order, deduplication map, and actual L are serialized. Six-prior averaging applies only when these six distinct unprotected templates remain; otherwise report the actual set and denominator. Class-uniform recognition is reported separately from this deployment average; its target prior is q_u.

Table S6: Numerical configuration. Calibration caps bound the eligible inventory; realized counts follow Eq G.1.

Quantity	Setting
Target batch	Controlled-prior reference: $m = 2048$; iNaturalist recognition: all 24,426 validation inputs
Failure allocation	$\delta = 0.05$; $\delta_{\mathrm{id}} = \delta_r = 0.025$
Overall/group budgets	$\beta = 0$; $\beta_g = 0.02$
Calibration inventory	Recognition: Table S1; controlled reference: cap 100, subject to its own inventory
ID/selection/audit allocation	Eq G.1; iNaturalist minimum-support rule in Appendix A
Proposal grid	$\{0, 0.25, 0.5, 0.75, 1\}$
Proposal/reference smoothing	$\epsilon = 10^{-6}$; $\epsilon_\rho = 10^{-8}$
Witness temperatures	$\tau_z = 1$; $\tau_p = 0.1$
Source reference / bias penalty	$p^{\mathrm{ref}} = q_s$; $\lambda_b = 10^{-3}$
Reference optimization	L-BFGS, at most 100 iterations; finite fallback
Outer runs	3; seeds 17, 29, 43
Image deployment repetitions	Controlled-prior study: 100 draws per prior/run; full-validation recognition: one target batch per run
Exact finite-support repetitions	2,000 complete draws per prior/configuration
Power-analysis parameter	$\kappa = 0.05$; analysis only, not an audit tuning knob
Arithmetic and tie rules	Float64 statistics; fixed class order; baseline-first ties

G.4 Fixed numerical configuration

The reference optimizer starts from centered zero bias and returns a finite iterate. An absent selection class or nonfinite fitting result invokes the fallback in Appendix B.2. Missing-selection forecasts are $(-1, 1)$ for nonbaseline candidates and zero for the baseline. The audit uses $\log(3/\alpha)$ and both calibrated branches of Eq 9. Numerical allowances are conservative: they can tighten a margin but cannot relax the zero acceptance threshold.

The outer seeds index the three run-level evaluations. Repeated deployments sharing a checkpoint or image pool are first averaged within their run, then the three run-level values receive equal weight. Training variation and calibration/target-resampling variation are distinct sources of uncertainty. Reusing a finite pool produces clustered deployment diagnostics, not independent replications of the unknown-population guarantee.

G.5 Metrics, identities, and statistical uncertainty

For a fixed returned action and the classwise labeled evaluation sets E_c, define

$$\widehat{A}_c(a) = \frac{1}{|E_c|} \sum_{x \in E_c} \mathbf{1}\{h_a(x) = c\}, \qquad \widehat{d}_c^{E} = \widehat{A}_c(\widehat{a}) - \widehat{A}_c(a_0). \tag{G.4}$$

A class with positive evaluation weight requires support; missing evaluation classes cannot be removed and the remaining weights renormalized. The reported estimands are

$$\widehat{\mathrm{All}}_q(a) = 100 \sum_c q_c \widehat{A}_c(a), \tag{G.5}$$

$$\widehat{\mathrm{Cond}}_{q,g}(a) = 100 \sum_{c \in g} \frac{q_c}{q(g)} \widehat{A}_c(a), \quad q(g) > 0, \tag{G.6}$$

$$\widehat{\mathrm{Macro}}_g(a) = \frac{100}{|g|} \sum_{c \in g} \widehat{A}_c(a). \tag{G.7}$$

The native sample-weighted benchmark metric is separately named when different. The iNaturalist recognition values use all validation images with the class-uniform weighting of Table S3; their transductive input access is specified in Appendix A.1. These descriptive formulas do not assume that evaluated inputs were unseen during adaptation. Under a uniform prior, macro and conditional group metrics coincide. For any fixed nonempty partition and the same evaluated actions, the identity

$$\Delta\mathrm{All}(q) = \sum_g q(g)\Delta\mathrm{Cond}_{q,g} \tag{G.8}$$

must hold before rounding. An average across different priors generally has no single fixed group-weight vector.

Int. is $100\,\Pr(\widehat{a} \neq a_0)$ estimated over all declared decisions. Pres. is the proportion of decisions under a protected generating prior returning the exact score function and tie rule. Mere agreement on finitely many labels after another action executes is a different diagnostic. Empty, Protect, Baseline, Reject, and Execute are mutually exclusive first-terminal-branch outcomes and sum to 100%. Exact baseline-relative gains and exceedances on preservation branches are zero by definition, not inferred from a statistical test.

For one deployment, $\widehat{v} = \mathbf{1}\{\exists h : -w_h(q)^\top \widehat{d}^E > \beta_h\}$. Exc. averages this evaluation-estimated event over all decisions. Finite evaluation noise, and any transductive input reuse, distinguish this descriptive event from the population failure probability in Theorem 1. Rounded test accuracies, including a displayed group reduction no larger than 2 pp, do not establish that every audit margin was nonnegative. The exact-risk study instead enumerates population gains. An execution-conditioned violation frequency uses a different denominator and does not inherit the unconditional δ bound.

For an image-benchmark metric $m_{b,r,q}$, let $b \in \{1,2,3\}$ index outer runs, $r \in \{1,\dots,R\}$ deployment repetitions, and $q \in \mathcal{Q}_u$ the distinct unprotected generating priors. The reported mean is

$$\overline{m} = \frac{1}{3}\sum_{b=1}^{3} m_b, \qquad m_b = \frac{1}{|\mathcal{Q}_u|}\sum_{q\in\mathcal{Q}_u}\frac{1}{R}\sum_{r=1}^{R} m_{b,r,q}. \tag{G.9}$$

For the controlled-prior study, $R = 100$ and six unprotected priors give 1,800 evaluated decisions per learner–dataset configuration, plus 300 protected-prior decisions. These totals count decisions, not distinct images or independent outer runs. They do not describe the full-validation iNaturalist recognition experiment, which fixes one target input pool per outer run. Paired differences are formed within runs before averaging. The image tables report descriptive means; a positive mean alone does not establish statistical significance. Repeated deployments sharing a checkpoint or pool are not independent binomial trials.

For a paired run-level gain Δ_b, the descriptive mean and sample standard deviation are

$$\overline{\Delta} = \frac{1}{3}\sum_{b=1}^{3}\Delta_b, \qquad s_\Delta = \sqrt{\frac{1}{2}\sum_{b=1}^{3}(\Delta_b - \overline{\Delta})^2}. \tag{G.10}$$

Run means do not determine s_Δ, and a resampling count is not an independent-training count. No standard deviation is inferred from rounded table means.

Exact binomial intervals apply to the independently regenerated finite-support experiments and the explicitly randomized paired-loss cohort. Interquartile bands describe the spread of repeated observations, not confidence intervals for a population mean. Conditional audit-pass frequencies and complete-system intervention rates retain their respective denominators.

G.6 Baselines and identical-proposal controls

Frozen and point-prior controls. Frozen always returns a_0. The point-prior control chooses the first minimizer of $\|\widehat{m} - \widehat{M}q\|_\infty$ over the full declared dictionary, then selects a correction strength by empirical weighted selection gain, with the baseline included. It has no independent risk audit. Zero-strength choices may preserve naturally, so its execution rate is measured and is not fixed to 100%.

BBSE and calibrated MLLS. BBSE uses a simplex-constrained response fit with regularization 10^{-4} and a deterministic solver cap of 1,000 iterations (Lipton et al., 2018). Calibrated MLLS uses source bias-fit probabilities and the label-shift likelihood (Alexandari et al., 2020). Its EM initialization is p^{ref}, probability floor is 10^{-8}, and termination occurs when the ℓ_1 iterate change falls below 10^{-8} or after 1,000 iterations. Both controls select correction strength on S_{sel} and execute without a risk audit. Native and common-proposal variants remain distinct. Nonfinite outputs preserve the baseline and remain in the decision denominator.

LSC and external-proposal certification. LSC constructs a corrected classifier using its prior-estimation procedure (Wei et al., 2024). Its estimator output, reference scores, strength, and tie rule are fixed before independent auditing. The execution-layer comparisons share the exact same candidate predictions within each checkpoint and deployment draw. Direct execution and complete TailSafe execution form the primary pair; the common-veto and overall-only controls in Table S7 separate policy preservation from group protection. The default TailSafe candidate bank is evaluated separately as an end-to-end method.

Table S7: Same-candidate LSC execution controls. The candidate is constructed before any veto for this diagnostic; every rule receives its identical stored predictions. Pres. uses the protected source prior; all other columns average unprotected priors. A common veto separates policy preservation from risk certification.

	CIFAR-100-LT					ImageNet-LT				
Rule	ΔAll	ΔFew	Int.	Pres.	Exc.	ΔAll	ΔFew	Int.	Pres.	Exc.
Direct LSC	3.1	2.5	90.0	4.0	6.0	2.0	1.8	89.0	5.0	6.6
Common veto only	2.8	2.6	82.0	99.0	4.8	1.8	1.9	80.0	98.7	5.3
Veto + overall audit	2.3	2.7	70.0	99.0	2.9	1.5	2.2	64.0	98.7	3.3
Veto + full audit	1.9	3.3	60.0	99.0	0.5	1.2	2.5	56.0	98.7	0.6

A method that has been trained with extra labels or supplied with extra information is reported under its actual resource access. Source-native, role-matched, and all-label-budget implementations are separate configurations. In the all-label MLLS control, a prespecified split of the entire eligible labeled pool is available for calibration and response/likelihood fitting; it is not advertised as retaining an independent safety holdout. The same total distinct source-label budget applies to TailSafe. Protection outcomes are also compared with a common explicit veto, because an unprotected baseline is not implementing the same policy by construction.

Table S8: Total-label-budget control. The all-label MLLS variant may use all eligible source labels for its native fitting, but has no independent safety claim. Role-matched MLLS uses the primary split roles. Each comparison holds total distinct labeled source images and evaluation data fixed.

	MLLS, role-matched		MLLS, all-label		TailSafe	
Dataset	ΔAll	Exc.	ΔAll	Exc.	ΔAll	Exc.
CIFAR-100-LT	2.6	4.7	2.9	4.4	2.4	0.4
ImageNet-LT	1.5	5.0	1.8	4.6	1.6	0.5
iNaturalist 2018	2.3	4.9	2.5	4.6	2.0	0.6
Places-LT	1.1	4.6	1.3	4.2	1.1	0.5

G.7 Controlled ablation definitions

Identification ablation. Coordinate ID uses $r_{\mathrm{coord}}(q) = t + \sum_c q_c s_c$, with $\delta_s = \delta_t = \delta_{\mathrm{id}}/2$, $t = \sqrt{\log(2J/\delta_t)/(2m)}$, and $s_c = \sqrt{\log(2JC/\delta_s)/(2n_c^{\mathrm{id}})}$ for observed classes; missing columns use zero and $s_c = 1$. The total identification error budget equals that of the directional method. The action is reselected for the end-to-end row. The two compatible sets are not assumed nested for all counts and dictionaries.

Audit factorial. Range/IUT uses $r_w^{\mathrm{H}}(\alpha) = \sqrt{2V_w \log(1/\alpha)}$ with $\alpha = \delta_r$ and the same missing-mass and clipping conventions. Adaptive/IUT is the main audit. Joint-band controls use the same

marginal formula evaluated at $\alpha = \delta_r/K$, where K counts the fixed or pre-audit-known family actually covered. The primary fixed-cohort comparison uses the same K for the same cohort; the full-dictionary benchmark uses $K_0 = \sum_{q\in\mathcal{Q}} |\mathcal{H}(q)|$ declared in advance. No family is counted twice, and duplicates are handled by the declared rule. A range branch with $\log(3/\alpha)$ is separately identified when testing exact radius-minimum nesting; it is not the strongest range-only baseline. The four variants are compared on identical paired observations and a cohort selected using no audit outcomes.

End-to-end versus fixed-action changes. End-to-end radius ablations replace the radius in both the forecast and the final audit, allowing reselection. Fixed-action ablations hold the candidate, reference, compatible set, observations, and tie rule fixed and change only the audit construction. A tighter radius improves the test for a fixed candidate but does not imply monotonically better accuracy after reselection. Each comparator uses the calibration appropriate to its own risk statement.

Selector ablations. Gain-only maximizes empirical worst-case overall gain over all candidates and the baseline. Empirical-budget selection maximizes that same objective after filtering empirical group gains at $-\beta_g$. The full selector uses forecast lower bounds and planned audit counts. These controls share one already fitted reference, one ordered candidate bank, one compatible set, and identical final audit. Primary comparisons use the same missing-class planning convention; an optional legacy zero-imputation selector is separately named. Forecast scores are not interpreted as confidence bounds. In the fixed-cohort diagnostic, exact toy-population feasibility or test-estimated group infeasibility is recorded only after selection.

Witness and gate ablations. Logit-only drops prototype coordinates. The nested-set fixed-action diagnostic calibrates a single prespecified master family of $J_\star = 2C + 2|\mathcal{G}|$ coordinates, comprising both views and all prespecified group sums. Every subset uses that same master-family radius and confidence allocation. The ordinary dimension-recalibrated $J = C$ and $J = 2C$ algorithms are evaluated separately. The optional bounded group aggregates change J and receive a new identification calibration. A single-prior-audit diagnostic freezes the full candidate but checks only one pre-audit point-prior estimate; it has no robust unknown-prior guarantee. No final group checks removes only final group checks, retaining the full selector. No protection veto removes only the policy veto. No risk audit executes the full selector's nonbaseline choice after the preceding compatibility/protection branches, so its execution share equals full execution plus audit rejection under exactly matched upstream decisions.

All five branches and all-decision gains are reported to distinguish useful execution from rejection shifted upstream. Candidate-conditioned pass rates describe the audit cohort and are not substituted for the complete-system execution rate. With the same candidate and data, removing checks enlarges the acceptance set; this comparison holds upstream selection fixed.

H Evaluation Consistency and Statistical Scope

This section specifies the aggregation and feasibility checks for within-checkpoint comparisons. A method citation identifies the learning algorithm; it does not identify the observations or averaging measure of a particular run. Every paired TailSafe result uses the same frozen checkpoint, class partition, and evaluation observations as its baseline. Scores obtained under another checkpoint, grouping convention, or target distribution are not substituted into the paired difference.

Paired aggregation. Baseline and adapted classwise predictions determine aggregate accuracies, gains, and paired second moments. Deployment branch codes determine execution and exceedance denominators. On any class-c sample, $\widehat{p}_c = (n_c^+ + n_c^-)/n_c$ and $\widehat{d}_c = (n_c^+ - n_c^-)/n_c$, with $n_c^+ + n_c^- \le n_c$. Accuracy means alone do not determine the error-change second moment used by the audit.

H.1 Common-weight recognition aggregation

Recognition uses class-uniform target weighting. The post-reservation CIFAR-100 group cardinalities are $(30, 35, 35)$, $(35, 41, 24)$, and $(60, 40, 0)$ at IR $100, 50, 10$. The groups are $(385, 479, 136)$ for ImageNet-LT, $(698, 2929, 4515)$ for iNaturalist 2018 after its specified reservation, and

$(131, 163, 71)$ for Places-LT. The ImageNet/Places reference partitions are consistent with the dataset-statistics record of Chang et al. (2020); the revised CIFAR/iNaturalist counts belong to the author-specified holdout protocol, not that external reference. CIFAR-10 retains its auxiliary $(3, 4, 3)$ rank partition. Recognition weights are $\lambda_g = |g|/C$ and are common to both members of each pair. The original controlled prior record retains its own group masses.

Table S18 gives the primary CIFAR-100-LT IR=100 group accuracies. For baseline and adapted group accuracies a_g^0, a_g^1, the overall values are $A_0 = \sum_g \lambda_g a_g^0$ and $A_1 = \sum_g \lambda_g a_g^1$, and the overall gain is $A_1 - A_0 = \sum_g \lambda_g (a_g^1 - a_g^0)$. The same weights apply to both predictors. Aggregation precedes display rounding, and empty groups have neither a score nor a risk constraint.

Paired correctness identifies repair and regression events directly. For the probability-scale accuracies a_g^0, a_g^1, gain $d_g = a_g^1 - a_g^0$, and paired second moment p_g, their probabilities are $(p_g + d_g)/2$ and $(p_g - d_g)/2$. Joint error has probability $1 - a_g^0 - (p_g + d_g)/2$, and joint correctness has probability $a_g^0 - (p_g - d_g)/2$. Their nonnegativity gives the feasibility condition in Eq H.1. These identities link the reported accuracies to the paired-loss quantities without replacing prediction-level evaluation.

H.2 Aggregation identities and statistical scope

For a fixed checkpoint, evaluation distribution, and partition, a source-reference number and a local paired measurement are different records. Let λ_g denote the actual evaluation mass of group g. If group scores and All use the same averaging measure, $A_{\text{All}} = \sum_g \lambda_g A_g$ and $\Delta A_{\text{All}} = \sum_g \lambda_g \Delta A_g$, with $\lambda_g \geq 0$ and $\sum_g \lambda_g = 1$. Both All and its paired change must lie in the corresponding convex hull. A sample-weighted All and class-macro group scores do not satisfy this identity without additional within-group information. Estimating weights from the reported accuracies merely to fit the identity does not identify the evaluation protocol.

Table S9 distinguishes the two CIFAR schedules. Each set of weights applies only to its actual count vector and class-uniform evaluation. A changed count vector or source-native grouping requires the corresponding class indices; weights are not fitted to rounded accuracies. Both Table S3 and the controlled prior record keep their own declared partitions.

Table S9: CIFAR group inventories under the two explicitly distinguished schedules. Recognition uses the reserved 400-maximum schedule. The 500-maximum reference describes the original controlled-count calculations; it is not substituted into Table 1.

Schedule	IR	Many	Medium	Few	Uniform-target group masses
Reserved, $N_{\max} = 400$	100	30	35	35	$(0.30, 0.35, 0.35)$
	50	35	41	24	$(0.35, 0.41, 0.24)$
	10	60	40	0	$(0.60, 0.40, 0)$
Original reference, $N_{\max} = 500$	100	35	35	30	$(0.35, 0.35, 0.30)$
	50	41	41	18	$(0.41, 0.41, 0.18)$
	10	69	31	0	$(0.69, 0.31, 0)$

Rounding follows aggregation of the unrounded paired records. A difference between two displayed one-decimal means can differ by up to 0.10 pp from the difference of their unrounded values. Group-weighted identities are therefore compared within their propagated rounding intervals. Image-benchmark observations, controlled numerical experiments, exact calculations, unreported reference entries, and empty groups retain their distinct meanings.

Paired-loss feasibility. For error probabilities r_c^0 and r_c^a, a realizable paired second moment satisfies

$$|d_c| \leq p_c \leq \min\{r_c^0 + r_c^a,\, 2 - r_c^0 - r_c^a\}, \qquad d_c = r_c^0 - r_c^a. \tag{H.1}$$

Indeed, p_c is the probability that the two binary error indicators differ. Their joint correct and joint incorrect probabilities must both be nonnegative. On a fixed sample, $\widehat{d}_c = (n_c^+ - n_c^-)/n_c$ and $\widehat{p}_c = (n_c^+ + n_c^-)/n_c$, where the nonnegative integer repair and regression counts satisfy $n_c^+ + n_c^- \leq n_c$. A claimed audit variance cannot be reconstructed from accuracy means alone.

Decision-level closure. All five branch shares sum to one. With identical upstream choices, the no-audit execution share equals full execution plus audit rejection, and removing final checks yields a superset of the original release decisions. Preservation contributes exactly zero paired gain and no budget violation. Consequently, in percentage units $|\Delta\mathrm{All}| \leq \mathrm{Int.}$ and $\mathrm{Exc.} \leq \mathrm{Int.}$. Stratum summaries are aggregated with their empirical decision shares, while prior summaries use the declared prior weights. These identities use matched denominators and do not assert monotonicity after reselection or after changing the violation budget itself.

Audit-radius feasibility. If every positively weighted class is observed and $N = \sum_{c:w_c>0} n_c^{\mathrm{risk}}$, Cauchy–Schwarz gives $V_w \geq 1/N$. The range-only IUT radius therefore obeys $r_w^{\mathrm{H}}(\delta_r) \geq \sqrt{2\log(1/\delta_r)/N}$. If all 100 classes have exactly 100 eligible calibration images, the stated split gives 40 audit images per class and $N \leq 4000$; at $\delta_r = 0.025$, the lower bound is 4.295 pp. With unequal availability below that cap, the floor-and-remainder allocation can assign 41 audit images to a class with 99 eligible images, so the cap alone yields only $N \leq 4100$ and a 4.242 pp lower bound under complete positive-weight support. A smaller reported range radius must identify different resources or nonzero missing mass; it is not validated merely by a plausible acceptance rate. Missing mass, V_w, b_w, and paired second moments are retained in each radius record. A radius summary is accompanied by its weights, observed counts, and paired second moments.

Audit scope. These identities are necessary consistency conditions, not a characterization of all realizable deterministic multiclass corrections. Verifying the complete procedure additionally requires its frozen score function, coupled predictions, classwise observations, and exact deployment weights. A single aggregate accuracy does not identify those objects.

I Extended Deployment, Attribution, and Sensitivity Studies

I.1 Complete primary comparison records

Tables S11–S14 retain the complete controlled-comparison columns corresponding to Tables 3–6. Table S10 separately summarizes the uniform-target recognition records. The two study families are not treated as identical checkpoints, class partitions, or calibration allocations merely because they use the same learner name.

Table S10: Primary matched All Top-1 (%). C100 and the three large-scale columns reproduce Tables 1–2. C10 is an auxiliary reference, not part of the four-dataset holdout comparison.

Learner / record	C100/100	C100/50	C100/10	C10/100	ImageNet	iNat.	Places
BCL matched	50.7	55.2	64.5	84.0	56.3	71.8	39.6
+ TailSafe	52.0	56.2	64.8	84.6	56.5	73.1	40.4
ProCo matched	51.5	55.9	64.7	85.6	57.4	72.7	39.3
+ TailSafe	52.3	56.7	65.1	86.1	58.7	73.3	40.2
ICL matched	55.7	59.4	67.6	86.2	59.1	75.4	40.7
+ TailSafe	56.6	60.2	68.0	86.5	59.4	75.7	41.2

The 100- and 200-cap rows use the same calibration images, role assignments, and paired deployment records. They represent the same realized resource condition, not independent experimental replicates. Increasing the nominal cap beyond the available inventory adds no source evidence. Calibration images unused at lower caps remain excluded from model and witness fitting.

I.2 Fixed paired-loss cohort construction

The auditor comparison in Tables 5 and S13 fixes 4,000 paired-loss laws before sampling their audit observations. Group sizes are $(35, 35, 30)$ or $(385, 479, 136)$, with 40 audit draws per class. A scalar $t \sim \mathrm{Unif}[-0.012, b]$, where $b = 0.10$ or 0.06, determines group gains $d = (0.8t, 0.95t, 1.6t + 0.01)$ and change probabilities $p = |d| + 0.008$. Baseline risks are $(0.32, 0.46, 0.65)$, and the paired correctness laws satisfy Eq H.1. Group-g repair/regression counts have multinomial sample size $40|g|$ and probabilities $(p_g + d_g)/2$, $(p_g - d_g)/2$, and $1 - p_g$. Group-homogeneous class laws make

Table S11: Deployment utility and budget diagnostics on frozen BCL checkpoints. Gains are pp; rates are percentages. All six unprotected priors and all preservation outcomes enter the averages. LSC direct and LSC+TailSafe use identical pre-audit proposals. Frozen zeros are definitional.

	CIFAR-100-LT				ImageNet-LT			
Rule	ΔAll	ΔFew	Int.	Exc.	ΔAll	ΔFew	Int.	Exc.
Frozen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Point prior	2.9	1.8	91.0	6.8	1.7	0.9	88.0	7.8
BBSE	2.0	2.2	85.0	5.5	1.0	1.6	80.0	5.9
Calibrated MLLS	2.6	3.0	88.0	4.7	1.5	2.1	86.0	5.0
LSC direct	3.1	2.5	90.0	6.0	2.0	1.8	89.0	6.6
LSC + TailSafe	1.9	3.3	60.0	0.5	1.2	2.5	56.0	0.6
TailSafe	2.4	4.0	70.0	0.4	1.6	3.1	64.0	0.5

Table S12: End-to-end attribution with matched observations and confidence budgets. Every variant is fully specified in Appendix G.7. No-group and no-audit controls retain the full selector. Other radius changes update both forecasting and final auditing; fixed-action comparisons isolate the audit separately.

	CIFAR-100-LT, IR=100				ImageNet-LT			
Variant	ΔAll	ΔFew	Int.	Exc.	ΔAll	ΔFew	Int.	Exc.
TailSafe	2.4	4.0	70.0	0.4	1.6	3.1	64.0	0.5
Coordinate ID	2.0	3.7	59.0	0.4	1.3	2.8	55.0	0.5
Range-only audit	1.6	2.8	49.0	0.3	1.0	2.0	43.0	0.3
Simultaneous audit	1.8	3.2	54.0	0.3	1.2	2.4	48.0	0.3
Gain-only selector	1.9	2.6	51.0	0.4	1.3	2.1	47.0	0.4
Empirical-budget selector	2.1	3.6	62.0	0.4	1.4	2.8	57.0	0.4
Logit-only witnesses	2.0	3.3	58.0	0.4	1.3	2.5	52.0	0.5
No final group checks	2.6	3.4	76.0	3.2	1.8	2.6	70.0	3.6
No risk audit	2.9	3.6	81.0	4.8	2.0	2.8	75.0	5.1

Table S13: Executed fixed paired-loss cohort audit. Each scale uses 4,000 independently generated profiles and 40 observations per class. Radius is the median across all profile–direction records; Pass is the all-checks acceptance percentage. The same sampled changes are shared by the four auditors.

		100-class model		1,000-class model	
Bound	Calibration	Radius (pp)	Pass (%)	Radius (pp)	Pass (%)
Range	Joint	7.55	11.55	2.53	59.03
Range	IUT	6.20	28.60	2.08	63.75
Adaptive	Joint	3.06	72.85	0.72	80.65
Adaptive	IUT	2.52	77.90	0.61	81.48

Table S14: CIFAR-100-LT source-resource comparison with a fixed inventory of 100 distinct calibration images per class. Nominal caps are truncated to the available inventory before the disjoint identification, selection, and audit roles are assigned. Gains are pp; decision rates are percentages.

CIFAR-100-LT: fixed 100-image-per-class inventory					
Nominal cap	Images/class	ID / Sel. / Audit	ΔAll	Int.	Exc.
25	25	10/5/10	0.5	18.0	0.4
50	50	20/10/20	1.4	44.0	0.4
100	100	40/20/40	2.4	70.0	0.4
200	100	40/20/40	2.4	70.0	0.4

these sufficient statistics equivalent to the classwise weighted audit. This controlled experiment isolates the auditor independently of neural-network training.

The retained group-mass profiles are $|g|/C$, $(0.6, 0.2, 0.2)$, and $(0.2, 0.2, 0.6)$. Each supplies an overall check and three group checks with budgets $(0, 0.02, 0.02, 0.02)$. Conditional-group directions coincide across profiles, leaving six distinct checks after deduplication. Joint calibration uses $K = 6$ and level $0.025/6$; IUT uses 0.025. Range uses $\log(1/\alpha)$ and Adaptive uses $\log(3/\alpha)$. The 100-class result uses NumPy PCG64 with SeedSequence $[20260918, 73, 0]$: all 4,000 scalar profile parameters are generated first, followed by multinomial observations in trial order and group order (Many, Medium, Few). The four auditors share each trial's observations. The supplied 1,000-class reference uses scale seed $[20260918, 73, 1]$ and retains its separately reported results. All laws, including harmful candidates, enter the denominator. The 100-class model's $(35, 35, 30)$ partition is part of this probability model and is not the $(30, 35, 35)$ partition of the reserved-data recognition checkpoint.

Sampling uncertainty of the controlled auditor comparison. Table S15 gives acceptance counts and their marginal intervals. For k successes among N independent law-and-audit trials, the two-sided $1 - \gamma$ Clopper–Pearson interval (Clopper & Pearson, 1934) is

$$\left[B^{-1}_{\gamma/2}(k, N-k+1),\ B^{-1}_{1-\gamma/2}(k+1, N-k)\right], \tag{I.1}$$

where $B^{-1}_u(a, b)$ is a beta quantile, with lower endpoint zero at $k = 0$ and upper endpoint one at $k = N$. The table uses $\gamma = 0.05$. Coverage concerns the explicitly randomized paired-law population, including candidate-law sampling, not a particular neural checkpoint or the default deployment selector. These are marginal intervals for individual pass rates, not simultaneous intervals or an unpaired significance test between auditors; all four auditors share the observations within a trial.

Table S15: Acceptance counts and marginal two-sided 95% intervals for the controlled paired-loss study. Each row uses 4,000 law-and-audit trials. The 100-class counts correspond to Table 5; the 1,000-class reference is reported separately. Intervals are calculated from integer counts.

Classes	Bound	Calibration	Accepted / total	95% interval (%)
100	Range	Joint	462/4000	[10.58, 12.58]
100	Range	IUT	1144/4000	[27.20, 30.03]
100	Adaptive	Joint	2914/4000	[71.44, 74.22]
100	Adaptive	IUT	3116/4000	[76.58, 79.18]
1000	Range	Joint	2361/4000	[57.48, 60.55]
1000	Range	IUT	2550/4000	[62.24, 65.24]
1000	Adaptive	Joint	3226/4000	[79.39, 81.86]
1000	Adaptive	IUT	3259/4000	[80.24, 82.67]

I.3 Cross-dataset and cross-learner coverage

Table S16 retains the controlled deployment comparison on iNaturalist 2018 and Places-LT. Table S17 summarizes the ten-learner recognition gains from the main tables, while Table S18 reproduces the primary CIFAR-100-LT recognition pairs. The recognition summaries and controlled deployment records have different evaluation scopes; their absolute values are not subtracted across studies.

I.4 Prior-Specific Group Gains and Terminal Decisions

Figure S2 retains the prior-resolved controlled CIFAR-100-LT record. Its uniform group masses are $(0.35, 0.35, 0.30)$, as reported in Table S19; this is not the reserved-calibration recognition record in Table 1. The seven priors comprise the protected source prior and six unprotected uniform, reverse-frequency, interpolated, and group-heavy distributions. Within this record, overall gains combine prior-specific group effects, and terminal branches distinguish unresolved protection, baseline selection, failed certification, and execution.

Table S16: Deployment utility and budget diagnostics on frozen BCL checkpoints. Gains are pp; rates are percentages. All six unprotected priors and all preservation outcomes enter the averages. LSC direct and LSC+TailSafe use identical pre-audit proposals. Frozen zeros are definitional.

	iNaturalist 2018				Places-LT			
Rule	ΔAll	ΔFew	Int.	Exc.	ΔAll	ΔFew	Int.	Exc.
Frozen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Point prior	2.8	1.8	90.0	8.0	1.6	0.5	86.0	7.2
BBSE	1.3	2.0	83.0	6.0	0.6	1.0	75.0	5.6
Calibrated MLLS	2.3	3.3	88.0	4.9	1.1	1.6	82.0	4.6
LSC direct	2.6	3.0	90.0	6.5	1.5	1.3	85.0	6.0
LSC + TailSafe	1.5	3.7	59.0	0.6	0.8	1.9	47.0	0.7
TailSafe	2.0	4.3	68.0	0.6	1.1	2.3	53.0	0.5

Table S17: Summary of uniform-target All gains (pp) for the ten learners in Tables 1–2. This table contains no additional learner experiments.

Learner	C100/100	ImageNet	iNat.	Places
LDAM-DRW	1.3	0.5	0.4	0.5
Balanced Softmax	1.9	0.2	1.7	0.5
RIDE (3E)	1.6	0.8	0.9	1.3
BCL	1.3	0.2	1.3	0.8
SADE	1.8	1.1	0.9	0.7
BalPoE	0.9	0.5	0.5	0.7
PRL	1.5	0.4	0.9	0.6
ProCo	0.8	1.3	0.6	0.9
ICL	0.9	0.3	0.3	0.5
LOS	0.6	0.4	1.4	0.5

Table S18: Primary CIFAR-100-LT IR=100 recognition accuracies (%), reproduced at the main table's precision. The group masses are $(0.30, 0.35, 0.35)$. Rounded displayed values are not unrounded run records.

Learner / rule	Many	Medium	Few	All
BCL / Base	67.6	53.0	33.9	50.7
BCL / TailSafe	66.1	52.3	39.5	52.0
ProCo / Base	70.0	52.0	35.1	51.5
ProCo / TailSafe	68.3	51.2	39.6	52.3
ICL / Base	71.6	61.3	36.4	55.7
ICL / TailSafe	69.6	60.7	41.3	56.6

Table S19 records the group masses and paired gains in Figure S2(a). Write $G_{\text{Many}}, G_{\text{Medium}}, G_{\text{Few}} \in \mathcal{G}$ for the fixed class subsets. The prior-specific overall gain satisfies

$$\Delta\text{All}(q) = q(G_{\text{Many}})\,\Delta_{\text{Many}}(q) + q(G_{\text{Medium}})\,\Delta_{\text{Medium}}(q) + q(G_{\text{Few}})\,\Delta_{\text{Few}}(q), \tag{I.2}$$

where the calculation uses the full-precision group masses before display rounding. The relation is descriptive of the returned action under that generating prior; it does not imply that the same groupwise correction is selected for every prior.

Table S19: Controlled-reference CIFAR-100-LT prior record, whose uniform group masses are $(0.35, 0.35, 0.30)$. The first three columns are group masses, and gains are paired percentage-point changes. ΔAll is computed using the full-precision prior masses before rounding.

Prior	$q(M)$	$q(D)$	$q(F)$	ΔM	ΔD	ΔF	ΔAll
Source	0.8135	0.1584	0.0281	0.000	0.000	0.000	0.000
Uniform	0.3500	0.3500	0.3000	-0.200	0.600	3.533	1.200
Reverse	0.0383	0.2001	0.7617	-0.600	-0.500	5.800	4.295
Source–Uniform	0.5817	0.2542	0.1641	0.400	0.700	2.000	0.739
Uniform–Reverse	0.1941	0.2750	0.5308	-0.300	0.200	5.800	3.076
Many-heavy	0.6000	0.2000	0.2000	1.185	0.400	-0.500	0.691
Few-heavy	0.2000	0.2000	0.6000	-0.500	0.400	7.367	4.400

Under the uniform prior, the reported Many decrease of 0.2 pp and Few improvement of 3.533 pp accompany an overall gain of 1.200 pp. Reverse-frequency and Few-heavy priors yield overall gains of 4.295 and 4.400 pp, respectively. Many-heavy instead improves Many by 1.185 pp while Few decreases by 0.500 pp. These differences combine changes in evaluation weights with changes in the selected action; they do not isolate a weighting effect alone. The source-prior decisions shown here preserve the baseline and have zero gain.

Prior tailness does not specify a unique correction strength. Interpolated and group-heavy priors impose different overall and conditional-group directions, as defined in Section 3.1. Figure S2(a) reports the resulting action-specific risk profiles. Positive overall gains therefore do not imply non-negative gains for every group, and no monotone gain relation is asserted across different generating priors.

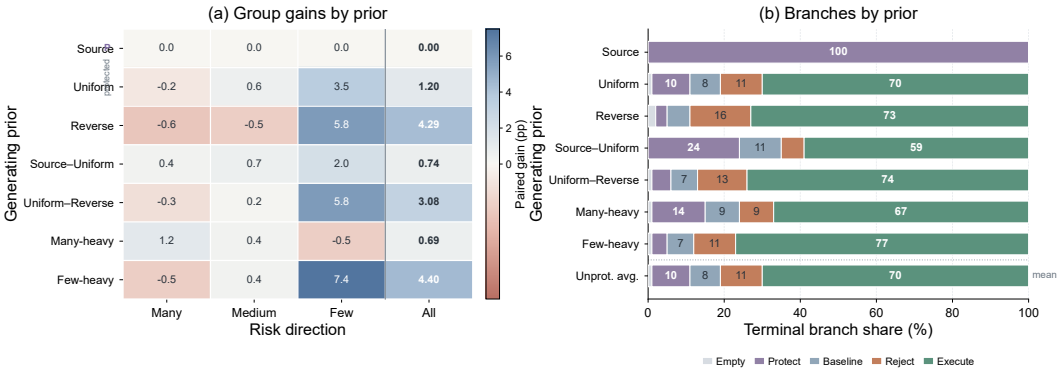


Figure S2: **Prior-resolved outcomes in the controlled reference.** **(a)** Target-conditional paired gains (pp) across the seven declared generating priors and the Many, Medium, Few, and All directions; the All column uses the exact prior masses before display rounding. All displayed source-prior decisions preserve the baseline and have zero gain. **(b)** Five mutually exclusive first-terminal branches for the same prior family. The source prior is shown separately as a protected environment, while the final row is the equal-weight mean over the six unprotected priors. Every decision contributes to exactly one prior-specific branch denominator.

Figure S2(b) decomposes the same records by their first terminal branch. In the displayed source-prior evaluation, every decision terminates at Protect. This observation is consistent with, but distinct

from, the probabilistic preservation guarantee of Theorem 1. An unprotected target can also produce Protect when the source template remains compatible. Empty, Protect, Baseline, Reject, and Execute are mutually exclusive and sum to 100% within each prior.

Source–Uniform has the largest protection-veto share among the displayed unprotected priors. Lower Protect shares create more opportunities to reach selection and auditing, but do not imply execution: a nonbaseline candidate must still pass every retained risk check. The branch decomposition therefore separates environmental resolution from the quality and certifiability of the selected action.

For the CIFAR-100-LT evaluation, the equal-weight mean over the six unprotected priors is

$$(\mathrm{Empty}, \mathrm{Protect}, \mathrm{Baseline}, \mathrm{Reject}, \mathrm{Execute}) = (1, 10, 8, 11, 70)\%. \tag{I.3}$$

This exactly matches the pooled CIFAR branch decomposition used by the main-text diagnostic. The identity is important because a pooled execution rate alone cannot reveal whether preserved decisions arise from unresolved compatibility, the explicit protection policy, baseline selection, or failure of the independent audit.

Table S20: Five exclusive terminal branches over all unprotected deployment decisions. Row totals equal 100% by definition. Early returns do not impute later audit quantities, and every decision remains in the corresponding dataset denominator.

Dataset	Empty	Protect	Baseline	Reject	Execute
CIFAR-100-LT	1.0	10.0	8.0	11.0	70.0
ImageNet-LT	1.0	14.0	10.0	11.0	64.0
iNaturalist 2018	1.0	12.0	8.0	11.0	68.0
Places-LT	2.0	19.0	12.0	14.0	53.0

Table S20 reports dataset-pooled unprotected branch shares, while Figure S2(b) resolves the CIFAR average by prior. Uniform-prior recognition, six-prior deployment means, and dataset-pooled branch statistics are distinct summaries. Their underlying decisions use the same gain identity in Eq I.2 and branch identity in Eq I.3.

I.5 Useful decisions under residual ambiguity

Table S21 stratifies the CIFAR deployment decisions by compatible-set size. Empty sets form a separate stratum; gain and execution are computed within each stratum and then weighted by its decision share. Non-singleton sets account for approximately 55% of executed corrections. This establishes useful execution under residual empirical compatibility, not rank deficiency of the population response matrix.

Table S21: Finite-sample compatibility strata, CIFAR-100-LT. Weight uses all decisions; other columns condition on the stratum. Stratified summaries do not receive the unconditional theorem at the same confidence level. Non-singleton empirical compatibility is not itself proof of a nonidentifiable population prior.

$\lvert\mathcal{U}_T\rvert$	Weight (%)	Mean size	Int. (%)	ΔAll	ΔFew
0	1.0	0.0	0.0	0.0	0.0
1	35.0	1.0	90.0	3.3	5.0
2	35.0	2.0	70.0	2.2	4.0
≥ 3	29.0	3.5	48.3	1.6	2.9
All	100	2.1	70.0	2.4	4.0

Figure S3 distinguishes better environmental resolution from a tighter audit of the same correction. The normalized residuals are computed before the risk audit; the margin distributions are retrospective diagnostics on that prespecified cohort.

For a target batch T, the plotted residual ratio is $\xi_q(T) = \|\widehat{m}(T) - \widehat{M}q\|_\infty / r_{\mathrm{id}}(q)$. Each curve indexes a tested template q, not a separate target-generating distribution. Compatibility is decided

Table S22: Fixed-action witness diagnostics on one pre-audit-defined cohort. One prespecified master witness family calibrates every nested subset; the marginal audit level is fixed. Blocker removal requires removing all old minimizing pairs. Group sums form a separate augmented-witness configuration. Dimension-recalibrated end-to-end controls are evaluated independently.

Witness	Mean $\lvert\mathcal{U}\rvert$	Blockers removed (%)	Pass (%)
Logit only	3.6	0.0	38.0
Joint views	2.1	44.0	61.0
Joint + group sums	1.9	48.0	63.0

batchwise by $\xi_q(T) \leq 1$; a median crossing the boundary is not the probability that a template is excluded. Likewise, the displayed interquartile band describes repeated-batch dispersion and is not a confidence band for identification coverage. Population separation and actual template exclusion are distinct quantities.

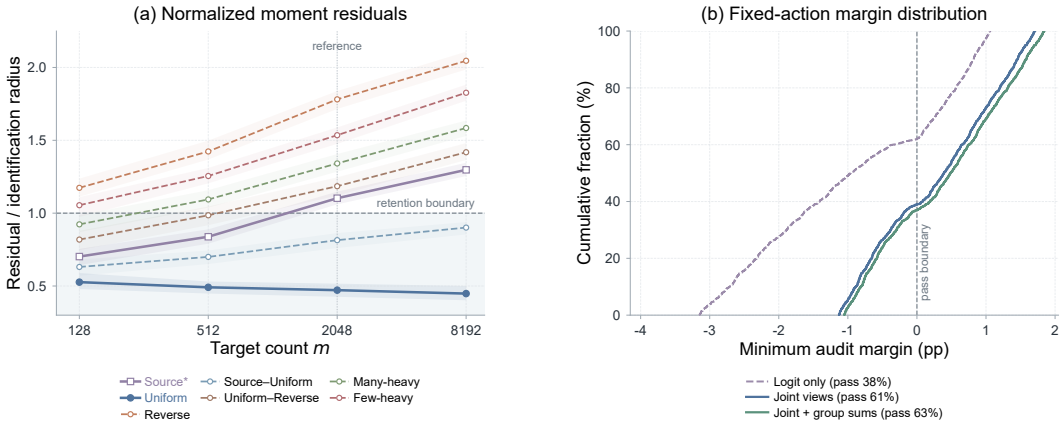


Figure S3: **Identification and action-specific witness diagnostics.** **(a)** Normalized identification residuals across target-batch sizes. Each curve reports the median empirical moment residual divided by its calibrated identification radius for one prior template, and the shaded band denotes the interquartile range across repeated target batches. The horizontal boundary at one separates retained from excluded templates; the dotted vertical line marks the reference target count $m = 2048$. The protected source template is marked by Source*. **(b)** Empirical cumulative distributions of the minimum independently audited margin on the same pre-audit-defined fixed-action cohort. The candidate, audit observations, and master-family calibration are held fixed across Logit-only, Joint-view, and Joint+group-sum witnesses. The vertical line at zero is the execution boundary; the corresponding pass rates are 38%, 61%, and 63%.

The fixed-action witness comparison holds the candidate and all audit statistics fixed and records every retained (prior, budget) margin. Blocker removal means that all previously minimizing pairs are excluded. One prespecified master witness family calibrates every nested subset, preventing an unaccounted increase in identification error. The common cohort is fixed before auditing and contains only cases with nonempty, unprotected compatibility under every compared subset. Empty sets and protection vetoes are not counted as successful certification. Ordinary dimension-recalibrated end-to-end variants are separate controls. Group aggregates are functions of the original witnesses and do not increase their population row-space rank.

Population ambiguity is tested in the exactly specified moment-fiber instance of Appendix J.2. Natural-image calculations of $\widehat{M}$ or empirical classwise gains are only plug-in diagnostics of the modulus; they are not substituted for the population M, d in a theorem verification. Exact primal/dual equality is evaluated on known finite matrices with recorded solver tolerances, not established by favorable prediction accuracy.

Figure 2: data and denominators. Panel (a) uses nominal caps $(25, 50, 100, 200)$, realized counts $(25, 50, 100, 100)$, and all-decision gains $(0.5, 1.4, 2.4, 2.4)$ pp. Only the measured 100-cap Range-only comparison, 1.6 pp, is displayed; no unmeasured comparator trajectory is interpolated. Panel (b) uses the four integer acceptance counts $(462, 1144, 2914, 3116)$ out of 4,000 and median directional radii $(7.55, 6.20, 3.06, 2.52)$ pp at 40 audit observations per class. It does not extrapolate 25% or 50% audit-resource points from the full-allocation trial. Panel (c) uses stratum shares $(1, 35, 35, 29)\%$ and within-stratum execution $(0, 90, 70, 48.3)\%$, whose displayed summaries aggregate to approximately 70.0% execution. The four dataset branch rows in panel (d) are the controlled-deployment records in Table S20, not branch probabilities inferred from recognition accuracies. In particular, its iNaturalist row is not a branch estimate obtained from three full-validation recognition decisions. No confidence intervals or run-level error bars are reconstructed from aggregate means.

I.6 Selection, Forecasts, and Execution Outcomes

The selector determines which candidate reaches the independent audit, so a lower final rejection rate is meaningful only when it is distinguished from earlier baseline selection. The comparison therefore fixes the candidate bank, source-reference fit, compatible set, planned audit counts, and final auditor, and changes only the pre-audit selection rule. Gain-only selection maximizes empirical worst-case overall gain, empirical-budget selection additionally filters candidates by empirical group gains, and the full audit-forecast selector uses the anticipated certification margins of Eq 12. Audit outcomes are never used to filter the diagnostic cohort or to reselect a replacement candidate.

Table S23: Selector attribution on the common candidate bank. Baseline, execution, and gain use all deployment decisions; Infeasible is the test-estimated group-infeasibility percentage among selected nonbaseline candidates.

Selector	Baseline (%)	Infeasible (%)	Execute (%)	ΔAll (pp)
Gain only	2.0	25.0	51.0	1.9
Empirical budgets	7.0	15.0	62.0	2.1
Audit forecast	8.0	4.0	70.0	2.4

For a selected nonbaseline candidate a, define its forecast and independently audited minimum margins as

$$\widetilde{M}(a) = \min_{q \in \mathcal{U}_T,\, h \in \mathcal{H}(q)} \left\{ \widetilde{L}_{w_h(q)}(a) + \beta_h \right\}, \qquad M_{\mathrm{audit}}(a) = \min_{q \in \mathcal{U}_T,\, h \in \mathcal{H}(q)} \left\{ L_{w_h(q)}(a; \delta_r) + \beta_h \right\}. \tag{I.4}$$

The audit-forecast selector admits a nonbaseline candidate only when $\widetilde{M}(a) \geq 0$; Gain-only and empirical-budget selection impose no such forecast-bound requirement. Final execution instead requires $M_{\mathrm{audit}}(a) \geq 0$ on the independent audit. The two margins consequently answer different questions: the first guides candidate choice using pre-audit evidence, whereas the second determines whether the frozen choice can actually be released.

Figure S4(a) plots these two quantities for every selected nonbaseline candidate in the common selector diagnostic. Marker shape identifies the selector and color records the subsequent audit outcome. Gain-only selection produces the broadest set of certification forecasts, including many candidates with negative predicted minimum margins. Empirical-budget filtering shifts the selected population toward more favorable regions but can still choose candidates whose certification forecast is negative. In contrast, every nonbaseline candidate chosen by the audit-forecast selector lies on the nonnegative side of the forecast boundary by construction. Positive forecast margin nevertheless does not guarantee release: some forecast-feasible candidates fall below the independent audit boundary and are rejected, directly illustrating why the selection forecast cannot replace the independent safety test.

The forecast and audit margins are positively associated but dispersed. They use separate observations, and the audit also reflects the selected candidate's realized paired-change statistics. Correlation describes forecast quality, not calibrated confidence: a favorable forecast can still fail the independent audit.

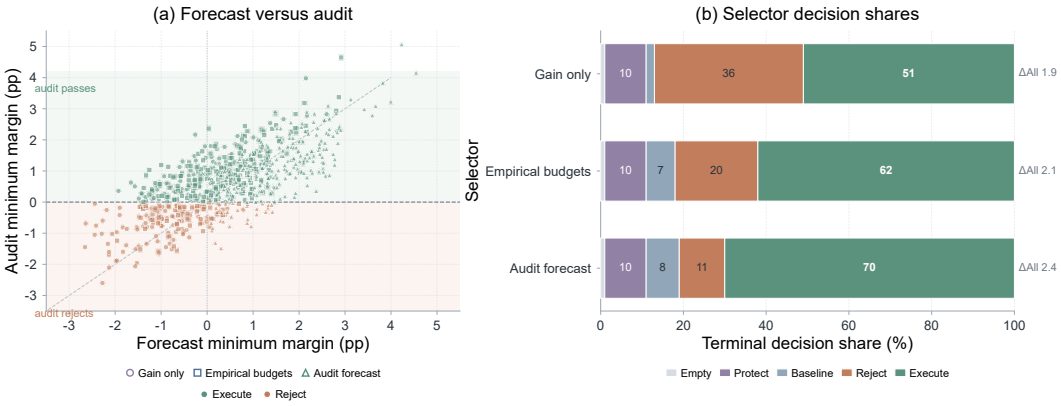


Figure S4: **Selection and independent-audit diagnostics on the same candidate bank. (a)** Forecast minimum margin $\widetilde{M}(a)$ versus independently audited minimum margin $M_{\text{audit}}(a)$ for every selected nonbaseline candidate. Marker shape denotes the selector, while color denotes the final audit outcome. The horizontal zero line is the execution boundary and the vertical zero line is the forecast-feasibility reference; Audit-forecast selections are forecast-feasible by construction, but can still fail the independent audit. **(b)** Complete first-terminal-branch shares for Gain-only, Empirical-budget, and Audit-forecast selection. Empty-compatibility and protection branches are common upstream outcomes, whereas Baseline, Reject, and Execute expose how each selector redistributes decisions after reaching selection. The displayed ΔAll values use all deployment decisions, including preservation and rejection.

Figure S4(b) places the scatter diagnostic back into the complete-decision denominator. The upstream Empty and Protect shares are identical across selectors because identification and protection are fixed before the selector is changed. Gain-only selection returns the baseline on only 2% of all decisions but subsequently sends 36% to audit rejection, leaving 51% execution and $+1.9$ pp all-decision gain. Empirical-budget selection raises the baseline share to 7%, reduces audit rejection to 20%, and increases execution to 62% with $+2.1$ pp gain. Audit-forecast selection changes the baseline branch only modestly further, to 8%, while reducing final rejection to 11% and raising execution to 70% and ΔAll to $+2.4$ pp. The resulting improvement therefore cannot be explained solely by moving difficult cases into an earlier preservation branch.

Conditioned on having selected a nonbaseline candidate, the corresponding independent-audit pass fractions are

$$\frac{51}{51+36} = 58.6\%, \qquad \frac{62}{62+20} = 75.6\%, \qquad \frac{70}{70+11} = 86.4\%, \tag{I.5}$$

for Gain-only, Empirical-budget, and Audit-forecast selection, respectively. These conditional audit-pass fractions and the Infeasible percentages in Table S23 share the selected-nonbaseline denominator but quantify different events. Both differ from intervention rates, which use all deployment decisions. Infeasible checks group budgets only, whereas Exc. checks the union of overall and group violations.

Test-estimated group infeasibility among selected nonbaseline candidates decreases from 25% with gain-only selection to 15% with empirical-budget filtering and 4% with audit forecasting (Table S23). Together with the execution and gain results, this favors the forecast selector in the matched comparison. The forecast remains a selection criterion, while the independent audit supplies the statistical guarantee.

I.7 Resource allocation and certificate power

Table S24 varies target size and source-stage resources separately. Nested classwise allocations are fixed independently of feature values and outcomes. Identification controls hold selection and audit observations fixed. Audit-size controls include both a frozen-action comparison and end-to-end

reselection, because audit counts enter the selection forecast. Increasing target observations leaves the audit statistics unchanged only when the action and audit data are held fixed.

Table S24: Stage-isolated controls at the CIFAR-100-LT reference allocation. ID and Audit give split fractions; m is target count. The displayed $n_{\mathrm{eff}}(q_u)$ is calculated for the uniform overall direction with 40 audit observations per class at 100%, not the smallest effective size over all constraints. Fixed-action and reselected protocols are separated.

Changed resource	m	ID (%)	Audit (%)	Int. (%)	ΔAll	$n_{\mathrm{eff}}(q_u)$
Target	128	100	100	22.0	0.7	4000
Target	512	100	100	49.0	1.6	4000
Target	2048	100	100	70.0	2.4	4000
Target	8192	100	100	73.0	2.5	4000
ID	2048	25	100	43.0	1.4	4000
ID	2048	50	100	59.0	2.0	4000
Audit: fixed action	2048	100	25	38.0	1.2	1000
Audit: fixed action	2048	100	50	58.0	2.0	2000
Audit: reselect	2048	100	25	31.0	1.0	1000
Audit: reselect	2048	100	50	55.0	1.8	2000

Directional audit records contain η_w, V_w, b_w, $\widehat{v}_w$, both radius branches, the selected radius, and the budget margin $L_w + \beta_h$. Effective sample size is reported only for fully observed positive-weight support; missing mass is shown separately. Zero observed changes do not establish zero population change. The execution-power condition remains sufficient: a failed sufficient inequality does not imply a failed realized audit.

I.8 Identification resolution and the protection bottleneck

The target contribution imposes the deterministic radius floor $r_{\mathrm{id}}(q) \geq \sqrt{\log(2JL/\delta_{\mathrm{id}})/(2m)}$, even with unlimited source identification data. Protection may therefore dominate the execution rate before the risk audit is reached. On the common event $\mathcal{E}_{\mathrm{id}}$, population separation larger than $2r_{\mathrm{id}}(q)$ excludes a template, but absence of that sufficient separation is not by itself a necessary rejection condition. The operational diagnostic is the empirical ratio $\|\widehat{m} - \widehat{M}q\|_\infty / r_{\mathrm{id}}(q)$, especially for protected q.

For the specified unreserved CIFAR-100 IR=100 counts, $J = 200$, $L = 7$, $m = 2048$, and $n_c^{\mathrm{id}} = 40$, the target-only radius floor is 0.05328 and $r_{\mathrm{id}}(q_s) = 0.07938$. In the class-indicator limiting case $M = [I; I]$, the source–uniform population discrepancy is $\|q_s - q_u\|_\infty = 0.03610$. Thus the protected source template remains within that radius at the population moments of this limiting example. The calculation isolates finite-sample resolution; it does not assume that neural softmax witnesses equal indicator functions. Figure S3 evaluates the actual witness residuals relative to their radii.

Figure S3 therefore reports normalized identification residuals alongside audit margins. Increasing audit data cannot overcome an upstream protection veto. Conversely, refining an already adequate compatible set cannot replace source evidence for a high-weight, sparsely observed audit class. Both effects must be separated before attributing a low or high execution rate to the audit design.

I.9 Risk budgets, confidence allocation, and dictionary size

Table S25 evaluates the prespecified group-budget grid $0, 0.01, 0.02, 0.05$ at overall budget zero, and total-failure grid $0.01, 0.05, 0.10$ with equal identification/audit allocation. An additional allocation sensitivity at total 0.05 uses $(\delta_{\mathrm{id}}, \delta_r) = (0.01, 0.04), (0.025, 0.025), (0.04, 0.01)$. These are separate procedures fixed before evaluating outcomes. Changing a budget changes the violation event itself; changing confidence changes its guarantee. Neither curve is a license to choose the most favorable point using the audit labels.

Table S25: Prespecified budget and confidence sensitivity. Budgets change tolerated deterioration; failure allocations change confidence. The end-to-end comparison reruns selection; a distinct fixed-action threshold-only diagnostic tests nested acceptance. Exc. uses each row's own budget.

Parameter	Value	Int. (%)	ΔAll	Exc. (%)
Group budget	0.00	41.0	1.5	0.2
Group budget	0.01	58.0	2.0	0.3
Group budget	0.02	70.0	2.4	0.4
Group budget	0.05	77.0	2.7	0.5
Total failure	0.01	53.0	1.9	0.3
Total failure	0.05	70.0	2.4	0.4
Total failure	0.10	78.0	2.6	0.6

Table S26: Nested declared-dictionary sensitivity on the same seven generating priors. Additional distinct templates are specified before deployment; L is updated in identification. A common candidate bank is used for the isolation control, with default reselection reported separately. No monotone utility claim follows from the theory.

L	Int. (%)	ΔAll	Exc. (%)
7	70.0	2.4	0.4
14	65.0	2.2	0.4
28	62.0	2.1	0.5
56	57.0	1.9	0.4

Figure S5 presents the budget, allocation, dictionary, and neighborhood controls as separate estimands. Native exact-dictionary guarantees are not assigned to off-dictionary targets, and a larger prior neighborhood does not cover changes in class-conditional input distributions.

Dictionary sensitivity retains the same seven generating priors and adds distinct convex combinations of template pairs in lexicographic order, first with coefficients $1/4$ and $3/4$ and then finer dyadic coefficients. The evaluated dictionary sizes are $7, 14, 28, 56$. Entries and order are fixed before target sampling. Identification uses the actual L; IUT auditing has no direction-count factor in its marginal threshold, although additional constraints can reduce execution power. Common-bank isolation and default dictionary-dependent reselection remain separate comparisons.

I.10 Template mismatch, conditional change, and optional policies

Prior mismatch and conditional shift are separate experiments in Table S27. The second column denotes the template-neighborhood radius for prior mismatch and the mixing coefficient ζ for conditional change. The exact-method off-dictionary point has zero allowed radius but nonzero actual mismatch. Bounded-prior controls use a zero-sum simplex-preserving perturbation of a nonprotected template and compare the exact method with the prespecified neighborhood method. Neighborhood membership is checked against the entire declared family, including protected templates. The conditional-mixture control instead changes class-conditional distributions and is outside the label-shift guarantee.

The conditional-change control uses the uniform prior, a fixed cyclic class permutation ν, and $P_t^{\zeta}(X \mid Y = c) = (1-\zeta)P_s(X \mid Y = c) + \zeta P_s(X \mid Y = \nu(c))$. Target labels retain their semantic value c. Summation over the uniform prior leaves the unlabeled input marginal unchanged, so coupling the same source and unlabeled evidence fixes the deployment decision across ζ. The analytical paired-prediction model has execution fraction 0.72. Conditional on execution, a profile index has probabilities $(0.05, 0.10, 0.85)$ and crossing points $\tau = (0.04, 0.12, 193/425)$. With endpoint gains $d_0 = \tau/24$ and $d_1 = d_0 - 1/24$, the mixture gain is $d_\zeta = (\tau - \zeta)/24$, giving all-decision means $D(0) = 0.012$, $D(1) = -0.018$, and $D(\zeta) = 0.012 - 0.030\zeta$. Group gains equal the overall gain, so the displayed group budget adds no violation; the zero overall budget is exceeded with probability $0.72\,\Pr(\tau < \zeta)$. The endpoints admit three-label confusion laws that move $1/24 - d_0$ mass from the cyclic alternative to the true label and $1/24 - 2d_0$ mass from the true label to a third label, start-

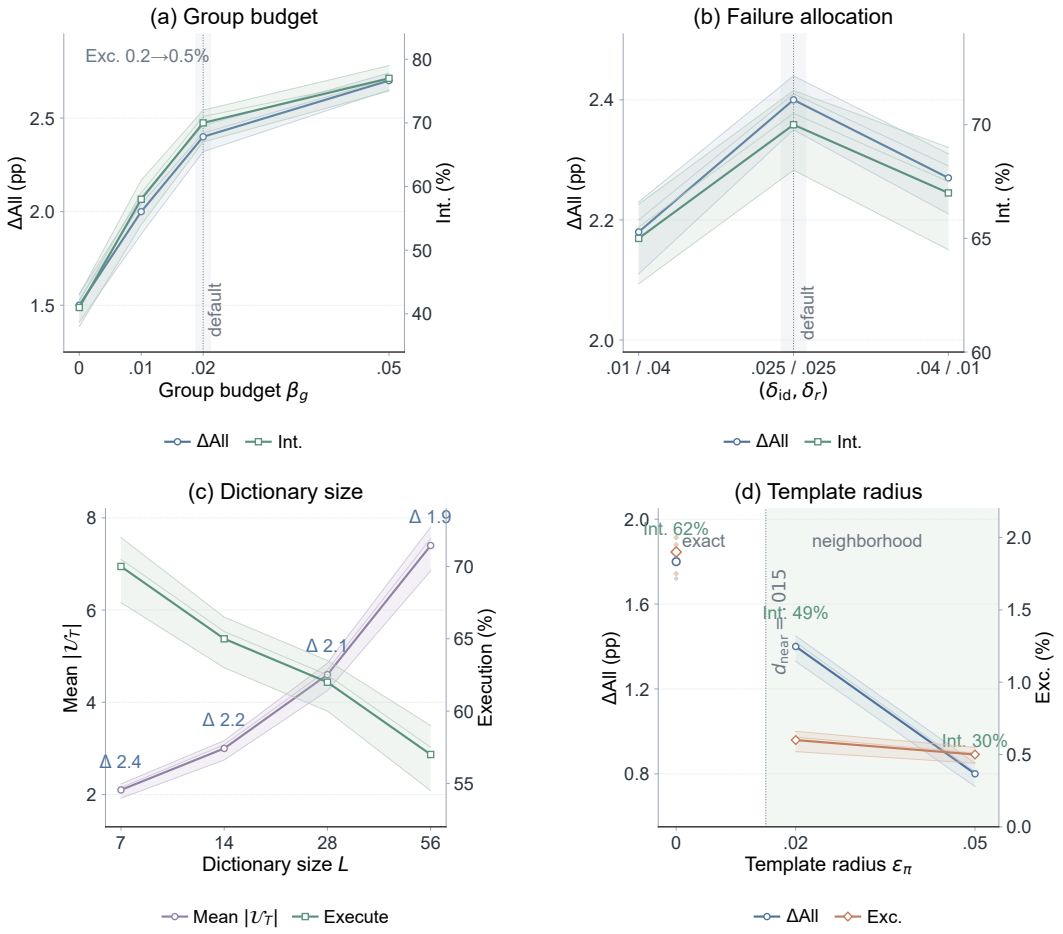


Figure S5: **Prespecified sensitivity diagnostics. (a)** End-to-end sensitivity to the group deterioration budget β_g, showing all-decision gain and intervention. Relaxing the budget changes the tolerated error event and therefore does not preserve the same risk target. **(b)** Identification–audit failure allocation at fixed total $\delta = 0.05$; the central point is the default split $(\delta_{\mathrm{id}}, \delta_r) = (0.025, 0.025)$. **(c)** Sensitivity to the declared prior-dictionary size L, with the generating-prior collection held fixed; mean retained compatible-set size and execution are shown, while the annotations report ΔAll. **(d)** Prior-template mismatch sensitivity. The $\varepsilon_\pi = 0$ point is the off-dictionary exact-method reference, whereas positive radii use the separately defined template-neighborhood procedure on the same perturbed target. The dotted vertical line marks the nearest-template distance. Thin trajectories denote the three outer runs and thick curves their mean.

Table S27: Prior-neighborhood and conditional-mixture controls. The latter share the same uniform-prior decision and execution fraction; their gains and exceedances are computed from the declared coupled endpoint construction. Conditional changes fall outside the label-shift guarantee.

Control	Allowance / ζ	Int. (%)	ΔAll	Exc. (%)
Prior perturbation				
Off-dictionary / exact method	0.00	62.0	1.8	1.9
Neighborhood method	0.02	49.0	1.4	0.6
Neighborhood method	0.05	30.0	0.8	0.5
Exact conditional-mixture model				
Uniform control	0.00	72.0	1.20	0.00
Conditional mixture	0.05	72.0	1.05	3.60
Conditional mixture	0.15	72.0	0.75	10.80

ing with diagonal and cyclic masses 0.5 and 0.25. All masses are nonnegative. The resulting rows are exact evaluations of this defined confusion model, demonstrating an undetectable conditional change in a specified model, not a neural corruption benchmark.

Optional macro-Few protection and fixed aggregate witnesses are tested only as separately named configurations. Macro-Few protects a different statistic from target-weighted conditional Few risk; neither is a guarantee for each individual tail class. An extra constraint may reduce execution. A broader prior neighborhood or a larger dictionary may retain more protected explanations. These costs are reported together with any gain, not suppressed by conditioning on accepted actions.

I.11 Computation and memory

With J witnesses, L priors, A candidates, feature dimension P, and $N_{\mathrm{id}}, N_{\mathrm{sel}}, N_{\mathrm{risk}}$ observations, witness aggregation costs $O((N_{\mathrm{id}} + m)J)$ after feature extraction and prior compatibility costs $O(LJC)$. Candidate predictions cost $O(AN_{\mathrm{sel}}C)$; classwise forecast statistics and disjoint-group aggregates cost $O(ALC)$ after caching. Auditing one selected correction costs $O(N_{\mathrm{risk}}C + LC)$ including prediction and all direction summaries. Storing paired sums and squared sums adds $O(C)$ streaming state to the audit; it does not require retaining all risk images. Prototype affinities cost $O(NCP)$ and reference optimization depends on its logged objective/gradient evaluations. These arithmetic counts do not treat feature extraction, transfers, or cached logits as free.

Table S28: Calibration observation counts and analytical storage. The four primary image benchmarks use their realized recognition inventories; the auxiliary CIFAR-10 row is an explicit 100-per-class reference. Matrix is float64 storage for $\widehat{M} \in \mathbb{R}^{2C\times C}$ (MiB); Paired is two float64 accumulators per class (KiB). Model weights, cached features, and solver workspace are excluded.

Dataset	N_{id}	N_{sel}	N_{risk}	Matrix (MiB)	Paired (KiB)
CIFAR-100-LT	4,000	2,000	4,000	0.153	1.562
CIFAR-10 reference	400	200	400	0.002	0.156
ImageNet-LT	8,000	4,000	8,000	15.259	15.625
iNaturalist 2018	28,872	16,512	32,905	1011.538	127.219
Places-LT	2,920	1,460	2,920	2.033	5.703

Table S28 uses the actual recognition-role counts alongside analytical storage: the dense response matrix requires $8(2C)C$ bytes and two audit accumulators require $16C$ bytes. These storage terms depend on the class count, not the number of images. Streaming avoids retaining every witness value. Model parameters, feature extraction, transfers, cached features, and solver workspace are excluded; the figures are not wall-clock latency measurements.

J Exact-Risk Calibration and Mechanism Experiments

J.1 Complete finite-support deployment experiment

The complete-decision study uses six classes and 12 feature atoms with $f(x) = e_x$. Class c assigns mass 0.70 to $X = c$, 0.27 to $X = 6 + c$, and 0.006 to each $X = 6 + k$ for $k \neq c$; the remaining probabilities are zero. Every conditional distribution sums to one. The source prior is $q_s = (0.72, 0.20, 0.05, 0.02, 0.007, 0.003)$, and consecutive class pairs define the three groups. The deployment dictionary follows Appendix G.3. Frozen logits are $z_{0,c}(x) = \log \max\{q_{s,c}P(x \mid c), 10^{-15}\}$; prototypes are normalized conditional feature means fixed before calibration.

Each repetition samples 256 identification observations and 256 selection observations per class, the displayed audit count per class, and an independent unlabeled target batch. Reference fitting, candidate construction, forecast selection, and independent auditing follow the method. The finite-model reference objective is solved by damped Newton iterations with at most 50 iterations, gradient tolerance 10^{-10}, Armijo coefficient 10^{-4}, and at most 25 step halvings; the finite returned iterate defines the reference. This solver optimizes the same objective as the image-experiment L-BFGS implementation. Each configuration–prior pair has 2,000 independent complete repetitions using NumPy SeedSequence $[1729, i, j]$, indexed by configuration and prior. Every repetition regenerates

source and target observations. Population class gains and budget margins are enumerated over the 12 atoms after candidate selection.

Each repetition evaluates true-prior retention, protected modification, the returned action, and the exact joint population budget-violation event. The three targets are $\Pr(\pi_t \notin \mathcal{U}_T) \le \delta_{\mathrm{id}}$, protected modification probability at most δ_{id} under each fixed protected prior, and complete-decision violation probability at most δ. Independent repetitions concern this defined finite population; repeated feature atoms are legitimate draws from that population, not additional natural-image evidence.

Table S29: Executed complete-procedure finite-support study. Each of the 35 prior–configuration cells contains 2,000 independent complete repetitions. Cov. is the minimum true-template coverage and Viol. the maximum population budget-violation frequency over the seven generating priors; protected modification is evaluated under the source prior. Int. and ΔAll average the six unprotected priors. All rates and confidence bounds are percentages. The final column is the one-sided familywise 95% upper bound for population budget-violation probability in a zero-event cell under the $0.05/35$ allocation.

m	n_r	Cov. (%)	Prot. mod.	Viol.	Int.	ΔAll	Viol. upper
512	64	100.00	0.00	0.00	0.00	0.000	0.327
512	256	100.00	0.00	0.00	0.00	0.000	0.327
2048	256	100.00	0.00	0.00	0.00	0.000	0.327
2048	1024	100.00	0.00	0.00	66.64	8.650	0.327
2048	4096	100.00	0.00	0.00	97.08	12.323	0.327

Table S29 separates calibration from useful execution. Across all 70,000 complete decisions, no true generating prior is excluded, the protected source classifier is never modified, and no exact population budget violation is observed. At the three lower-resource configurations the procedure consequently preserves throughout, whereas increasing the independent audit resource to $n_r = 1024$ and 4096 enables mean intervention rates of 66.64% and 97.08% over the six unprotected priors, together with positive complete-decision population gains. Thus the zero-violation record is not produced by an always-preserving rule: sufficiently informative source and target evidence leads to substantial nonbaseline execution while the exact population checks remain satisfied.

Zero observed events do not establish zero population failure probability. For a fixed prior–configuration cell with $R = 2000$ independent repetitions and zero events, the ordinary one-sided 95% Clopper–Pearson upper bound (Clopper & Pearson, 1934) is $1 - 0.05^{1/2000} \approx 0.0014967$. To cover all 35 prior–configuration cells simultaneously with familywise confidence at least 95%, allocate $0.05/35$ to each cell, giving

$$1 - \left(\frac{0.05}{35}\right)^{1/2000} = 0.00327018, \tag{J.1}$$

or 0.327018%. The final column of Table S29 reports this simultaneous upper bound. It is a finite-experiment confidence statement for the specified collection of independent repetitions, not an empirical estimate of a nonzero violation rate and not a replacement for Theorem 1. The 35-cell family refers to budget violations only; simultaneous confidence statements about identification exclusion and protected modification require their own stated event families.

The aggregate summaries in Table S29 can conceal substantial prior-specific differences in execution power, so Table S30 retains the complete per-prior records. All seven generating priors remain covered in every configuration and every population violation count remains zero. No source-prior execution is observed; its preservation policy remains governed by the probabilistic guarantee. Among the unprotected priors, execution is still zero when audit evidence is insufficient, but becomes strongly prior dependent once n_r reaches 1024. At $(m, n_r) = (2048, 1024)$, for example, execution ranges from 330/2000 repetitions for source–uniform to 1825/2000 for reverse; at $n_r = 4096$, all six unprotected priors execute in at least 1739/2000 repetitions. These differences reflect execution power under distinct population margins instead of changes in the validity target.

For nonzero event count k, the exact binomial interval depends on both k and the number of repetitions. Simultaneous prior–configuration summaries use the allocated familywise confidence level.

Table S30: Per-prior finite-model records. Every row uses 2,000 independent complete repetitions and every true prior is retained. Execute and Viol. are event counts, while ΔAll is the mean complete-decision population gain (pp), including zero gain on preservation and rejection branches.

m	n_r	Prior	Execute	ΔAll	Viol.
512	64	source	0	0.0000	0
512	64	uniform	0	0.0000	0
512	64	reverse	0	0.0000	0
512	64	source-uniform	0	0.0000	0
512	64	uniform-reverse	0	0.0000	0
512	64	many-heavy	0	0.0000	0
512	64	few-heavy	0	0.0000	0
512	256	source	0	0.0000	0
512	256	uniform	0	0.0000	0
512	256	reverse	0	0.0000	0
512	256	source-uniform	0	0.0000	0
512	256	uniform-reverse	0	0.0000	0
512	256	many-heavy	0	0.0000	0
512	256	few-heavy	0	0.0000	0
2048	256	source	0	0.0000	0
2048	256	uniform	0	0.0000	0
2048	256	reverse	0	0.0000	0
2048	256	source-uniform	0	0.0000	0
2048	256	uniform-reverse	0	0.0000	0
2048	256	many-heavy	0	0.0000	0
2048	256	few-heavy	0	0.0000	0
2048	1024	source	0	0.0000	0
2048	1024	uniform	1629	5.8960	0
2048	1024	reverse	1825	22.6632	0
2048	1024	source-uniform	330	0.3764	0
2048	1024	uniform-reverse	1239	9.3342	0
2048	1024	many-heavy	1383	3.0605	0
2048	1024	few-heavy	1591	10.5686	0
2048	4096	source	0	0.0000	0
2048	4096	uniform	1999	8.6548	0
2048	4096	reverse	2000	24.8364	0
2048	4096	source-uniform	1739	3.4360	0
2048	4096	uniform-reverse	1970	16.2799	0
2048	4096	many-heavy	1944	4.8850	0
2048	4096	few-heavy	1998	15.8468	0

Here every error count is zero, so the finite-support evidence consists of resource-dependent non-baseline execution, exact risk evaluation, and an upper confidence bound for each unobserved failure probability.

J.2 A POPULATION MOMENT FIBER WITH A USEFUL CORRECTION

The following separate four-class model makes nonidentifiability exact. Let the feature atoms be x_A, x_B, x_C, x_D with one-hot features. Classes 1 and 2 both put all mass at x_A. Class 3 puts probabilities 0.92 and 0.08 at x_B, x_C; class 4 puts all mass at x_D. The frozen score rows are

	1	2	3	4
x_A	4	0	-8	-8
x_B	-8	-8	4	0
x_C	-8	-8	0	0.6
x_D	-8	-8	0	4

. (J.2)

Use the same fixed logit/prototype witness construction as the main method, with prototypes given by the specified conditional means. Since classes 1 and 2 have identical conditional laws, their response columns agree in both views. Define

$$q_A = (0.2, 0.4, 0.2, 0.2), \qquad q_B = (0.4, 0.2, 0.2, 0.2).$$

These priors have identical witness moments and identical full input marginals. Use groups $\{1, 2\}$, $\{3\}$, and $\{4\}$, with an empty protection set for the isolated budget experiment.

An external fixed additive correction $b_{\text{safe}} = (0, 0, 1, 0)$ changes only the erroneous x_C prediction to class 3, giving the exactly enumerated gain vector $d_{\text{safe}} = (0, 0, 0.08, 0)$. Its overall gain is 0.016 under either prior and its group gains are $0, 0.08, 0$. Thus unresolved prior ambiguity has no effect on its budget feasibility. A second correction $b_{\text{trade}} = (0, 6, 1, 0)$ also changes x_A from class 1 to class 2, giving $d_{\text{trade}} = (-1, 1, 0.08, 0)$. Its overall gains are 0.216 and -0.184, and the first-group gains are $1/3$ and $-1/3$ under q_A and q_B, respectively. These are deterministic calculations from the specified model, not image-benchmark measurements.

The two fixed actions separate budget feasibility from prior identification. The safe action has positive overall and nonnegative group gains under both indistinguishable priors. The trade-off action violates the overall and first-group budgets under q_B, so an auditor that retains both priors must control its release. If q_A is protected, its retention forces preservation for either generating prior in the shared moment fiber. These population facts follow from the specified laws and the execution rule; they do not assert a finite-sample execution frequency for the default selector.

Table S31 summarizes the resulting population gains; the paired actions and identical target marginals make the distinction between environmental identification and budget feasibility explicit.

Table S31: Exactly enumerated four-class moment-fiber gains (pp). Classes 1 and 2 have identical input laws. These values characterize the specified population actions and do not report a finite-sample execution frequency.

Action	Prior	All	Group $\{1, 2\}$	Group $\{3\}$	Group $\{4\}$
Safe	q_A	1.6	0.0	8.0	0.0
Safe	q_B	1.6	0.0	8.0	0.0
Trade-off	q_A	21.6	$100/3$	8.0	0.0
Trade-off	q_B	-18.4	$-100/3$	8.0	0.0

J.3 INDEPENDENT AUDITING WITHIN A NONIDENTIFIABLE MOMENT FIBER

The experiment runs the identification and independent-audit stages on the four-class model above. The dictionary is $\mathcal{Q} = \{q_A, q_B\}$, the protection set is empty, and the budgets are $\beta = 0$ and $\beta_g = 0.02$ for groups $\{1, 2\}, \{3\}, \{4\}$. The two additive corrections b_{safe} and b_{trade} are fixed before sampling. They define two external-proposal procedures covered by Theorem 1; this experiment isolates certification after a proposal has been supplied.

Each repetition draws 256 independent identification observations per class, an independent target batch of size 2048, and $n_r \in \{64, 256, 512, 800, 1024, 4096\}$ audit observations per class. The feature atoms, frozen scores, and prototypes are those of Appendix J.2; the known class-conditional mean features are normalized before constructing prototype affinities. Temperatures are $(\tau_z, \tau_p) = (1, 0.1)$ and $(\delta_{\mathrm{id}}, \delta_{\mathrm{r}}) = (0.025, 0.025)$. Source-class and target atom counts are drawn independently from their multinomial laws, which are sufficient statistics for the specified i.i.d. observations. Both external procedures share the evidence within each repetition and are evaluated separately. Compatibility and audit bounds use Eqs 5 and 9, including empty-compatibility preservation.

Each prior–resource cell contains 2,000 independent repetitions. Sampling uses NumPy PCG64 with `SeedSequence([20260920, i, j, r])`, where i indexes the displayed resource grid, j indexes (q_A, q_B), and r indexes repetitions, all from zero. The order within a repetition is identification classes 1–4, target, then audit classes 1–4. This gives 24,000 independent paired evidence draws and 48,000 decisions across the two procedures, separate from the six-class study's 70,000 decisions. Population gains and budget violations are enumerated from the four-class laws.

Table S32: Finite-sample auditing under exactly indistinguishable priors. Each cell uses 2,000 independent evidence draws. Both templates remain compatible in every repetition. Gains are exact population gains averaged over all returned decisions, including preservation; the two fixed proposals share the evidence.

n_r/class	True prior	Safe executes	Safe gain (pp)	Trade-off executes	Violations[†]
64	q_A	0/2000	0.0000	0/2000	0 / 0
64	q_B	0/2000	0.0000	0/2000	0 / 0
256	q_A	0/2000	0.0000	0/2000	0 / 0
256	q_B	0/2000	0.0000	0/2000	0 / 0
512	q_A	0/2000	0.0000	0/2000	0 / 0
512	q_B	0/2000	0.0000	0/2000	0 / 0
800	q_A	1970/2000	1.5760	0/2000	0 / 0
800	q_B	1968/2000	1.5744	0/2000	0 / 0
1024	q_A	2000/2000	1.6000	0/2000	0 / 0
1024	q_B	2000/2000	1.6000	0/2000	0 / 0
4096	q_A	2000/2000	1.6000	0/2000	0 / 0
4096	q_B	2000/2000	1.6000	0/2000	0 / 0

[†]Exact population violations, Safe / Trade-off, each out of 2,000 decisions.

Both priors remain compatible in all 24,000 evidence draws, consistent with their identical full input marginals. The safe correction executes in 1,970/2,000 and 1,968/2,000 repetitions at $n_r = 800$ under q_A and q_B, respectively, and in all repetitions at $n_r \in \{1024, 4096\}$. Its all-decision population gain therefore approaches its 1.6 pp fixed-action gain while both priors remain unresolved. The trade-off correction is rejected throughout: retaining q_B exposes its negative overall and first-group margins even when the generating prior is q_A.

The lower-resource preservation outcomes expose a finite-sample cost of the audit. Group $\{4\}$ has identically zero paired changes, so on the displayed grid its radius is $(10/3)\log(3/\delta_{\mathrm{r}})/n_r$. Satisfying its 0.02 budget requires $n_r \geq 798$; consequently every draw preserves at $n_r \in \{64, 256, 512\}$, even though the safe action is population-feasible. The calculation identifies an audit-evidence requirement for this fixed proposal, independently of prior recovery.

No population budget violation is observed in either procedure. Applying a one-sided Clopper–Pearson bound at level $0.05/24$ to each of the 24 prior–resource–procedure cells gives the simultaneous 95% upper bound

$$1 - (0.05/24)^{1/2000} = 0.00308213.$$

The bound is 0.3082% per cell. Pairing the two procedures does not require independence between their intervals for this union-bound statement. These finite-experiment intervals concern the displayed fixed-proposal procedures.

Interpretation of the zero-violation record. The safe correction is feasible under either generating prior regardless of the decision. For the trade-off correction, classes 1 and 2 have deterministic audit changes -1 and $+1$. Under q_B, their conditional-group empirical gain is therefore exactly

$-1/3$ at every audit size, and the lower bound cannot pass a 0.02 budget. Equal identification sample counts and identical response columns also cause q_A and q_B to be retained or excluded together. Consequently, the zero-violation record in this construction has a structural explanation. Its main evidence is finite-sample useful execution under nonidentifiability and the resource cost of certifying an unchanged group; it is not a challenging near-boundary test of false release.

J.4 Paired-change sparsity at a fixed gain

The marginal audit experiment uses independent $Z \in \{-1, 0, 1\}$ with $\Pr(Z = 1) = (p + d)/2$, $\Pr(Z = -1) = (p - d)/2$, and $\Pr(Z = 0) = 1 - p$, where $|d| \le p \le 1$. The mean gain and budget are held fixed while p and audit size vary. Figure S6(a) uses $d = 0.02$, budget 0.01, $\delta_r = 0.025$, $p \in \{0.02, 0.05, 0.10, 0.20, 0.40\}$, and $N \in \{1000, 4000, 16000\}$, with 30,000 independent audit redraws per point. The experiment isolates a scalar audit statistic; it is not a claim about a newly trained multiclass predictor.

Table S33 evaluates the radius formulas directly. Equal per-example weights give effective size N, zero missing mass, and $\widehat{v} = \widehat{p}/N$. The table is an algebraic comparison, while Figure S6(a) estimates pass frequencies from sampled observations. Multiple directions in a complete audit share their classwise observations; independent simulation of unrelated directional statistics does not reproduce that joint decision.

Table S33: Calculated radii (pp) for equal per-example weights, no missing mass, specified $\widehat{p}$, $\delta_r = 0.025$, and $K = 28$. These entries are directly calculated from the displayed formulas; they are algebraic evaluations, not natural-image measurements. Joint uses marginal level δ_r/K. The optimally allocated range-only comparator uses $\log(1/\alpha)$, so adaptation is not guaranteed to dominate it at every $\widehat{p}$.

N	$\widehat{p}$	Range/joint	Range/IUT	Adaptive/joint	Adaptive/IUT
1000	0.00	11.850	8.589	2.707	1.596
1000	0.02	11.850	8.589	4.105	2.817
1000	0.20	11.850	8.589	9.304	6.901
4000	0.00	5.925	4.295	0.677	0.399
4000	0.02	5.925	4.295	1.608	1.167
4000	0.20	5.925	4.295	4.321	3.263
16000	0.00	2.962	2.147	0.169	0.100
16000	0.02	2.962	2.147	0.713	0.533
16000	0.20	2.962	2.147	2.085	1.588

J.5 Testing the conditional execution-power result

The execution-power diagnostic fixes the pre-audit information and varies only independent audit observations. Population gains, paired second moments, and directional evidence costs are available from the finite-model laws. Figure S6(b) uses 30,000 independent audit redraws at each displayed configuration. Its horizontal coordinate is $\Gamma = \min_{q,h}\{[w_h(q)^\top d(a^\star) + \beta_h]/E_{w_h(q)}(\kappa)\}$, for positive evidence costs. The reference $1 - \kappa$ applies when every local condition is satisfied, equivalently $\Gamma \ge 1$; configurations below one can also pass because the bound is sufficient, not necessary. The displayed reference fixes $N = 4000$, $p = 0.20$, $K_T = 28$, and $\kappa = 0.05$. Each plotted acceptance indicator concerns all retained checks on a shared audit draw. Thus K_T describes the evidence-cost bound; it is not a substitute for specifying the shared directional observations. The scalar experiment in panel (a) has its own marginal acceptance event.

Conditional pass probabilities are distinct from complete-decision risk coverage, which is evaluated by independently regenerating all stages in Appendix J.1. Compatibility-size strata likewise do not substitute for conditioning on a fixed pre-audit realization. Finite-model primal–dual calculations check numerical agreement and solver feasibility; the exact equality is established by the proof in Appendix D.1.

Figure S6 separates the effects of paired-change frequency and directional margin on audit power. The exact moment-fiber calculations in Appendix J.2 address the distinct question of budget decisions under nonidentifiable priors.

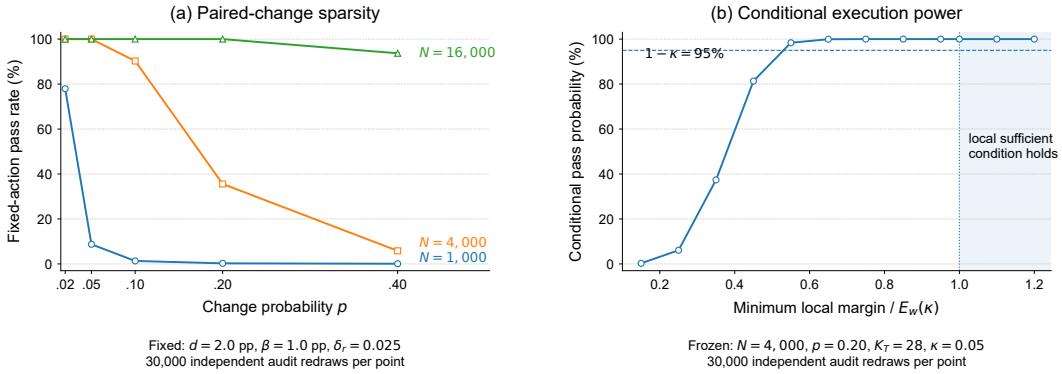


Figure S6: **Mechanism and conditional-power diagnostics. (a)** Fixed-mean paired-change experiments vary the change probability $p \geq |d|$ and audit size while holding the population gain, budget, and marginal audit level fixed; the resulting fixed-action pass rates expose how paired-change sparsity and audit resources jointly determine certification power. **(b)** For frozen pre-audit realizations in the finite model, independently redrawn audits estimate the conditional pass probability as a function of the minimum local population margin normalized by $E_w(\kappa)$. The vertical boundary at one marks the local sufficient-condition threshold, and the $1 - \kappa$ reference applies to realizations satisfying all retained local conditions; realizations below this threshold may still pass because the condition is sufficient, not necessary.

K LIMITATIONS

Distributional assumptions. The principal guarantee assumes exact label shift, target membership in a declared finite family, and independent source calibration. A nonempty compatible set does not validate these assumptions. Template neighborhoods cover bounded prior mismatch, not unrestricted prior or class-conditional shifts. Distinct labeled target laws can share the same input marginal; current unlabeled observations cannot generally resolve them without further assumptions.

Source evidence and selection. Missing or scarce independent source observations can make a valid audit uninformative. Paired-change sparsity reduces uncertainty, but large observation weights and missing mass can still block execution. Repeated images or augmented views do not create new independent population evidence. The forecast selector need not recover the best feasible correction; its reference fit may share selection data because the independent audit, not forecast calibration, supplies risk control.

Statistical target. The guarantee concerns population zero-one risk changes for the final action under one deployment decision, not per-input correctness or each individual tail class. Target-conditional group protection and optional macro-Few protection are different policies. IUT controls erroneous release, not simultaneous coverage of all numerical bounds or coverage at the same level conditional on execution. iNaturalist recognition uses label-hidden transductive evaluation; its observed accuracy is not an independent post-adaptation risk estimate. The execution theorem is a sufficient condition for a selected action, not a universal necessary sample size or an assertion that every image configuration meets it.

Repeated deployment and data reuse. The single-candidate proof does not permit reselection after audit rejection. Adaptive deployments, changing dictionaries, and reused audit feedback require additional simultaneous or sequential control. Related monitoring and protected-adaptation methods (Schirmer et al., 2026; Bar et al., 2024) do not automatically extend the present guarantee.

Paired second-moment bounds can be conservative and are not claimed minimax-optimal. Practical utility depends on the actual independent source-label inventory.