Certified Conflict Policies for Few-Shot Class Registration

Anonymous authors
Paper under double-blind review

Abstract

Deployed vision systems often need to recognize new categories from only a few labeled examples while keeping the existing classifier fixed. However, freezing the classifier does not freeze its predictions. New classes compete with old classes in the same decision rule and can cause old examples to be classified as new. We study this problem as few-shot class registration and introduce Certified Arbitration for Post-training Expansion (CAPE), which provides finite-sample control of old-to-new interference over an old-class calibration frame. CAPE separates risk certification from conflict ranking, using a scorer to prioritize new-class predictions that are less likely to interfere with old classes. We study two realizations of this decomposition: CAPE-BF fits a conflict scorer separately for each registration, while Certified Conflict Policy Learning (CCPL) learns a reusable scorer across registrations and applies it without refitting. Across 900 development episodes under a common certification protocol, CAPE-family scorers retain substantially more new-class utility than two external scoring pipelines, and amortized cross-registration scoring improves over registration-specific fitting. CCPL also transfers without retraining to both held-out populations, including a 3.27 percentage-point gain over CAPE-BF on FGVC Aircraft. These results show that reusable conflict scoring can improve utility under finite-sample risk control for class registration without retraining the old classifier.

1 Introduction

A deployed recognition system rarely operates with a fixed label set: new categories may arrive from a few labeled examples while the existing classifier stays fixed. Prior incremental-learning work already studies competition between old and new classes. We ask a more specific question: can each new registration provide finite-sample control of its old-to-new error rate?

Even with frozen old parameters, adding new classes can change old predictions: all classes compete in the same decision rule, so a new class can win an old-class query. We refer to this as *insertion-induced class-competition interference*. Across the 900 development episodes, direct insertion causes an old-class accuracy loss greater than 0.5 percentage points in 57.3% of episodes. Freezing parameters does not freeze behavior.

Certified Arbitration for Post-training Expansion (CAPE) addresses this interference by using old-class calibration examples to select a per-registration threshold whose 95% upper confidence bound on the old-to-new violation rate is at most 2%. Candidate new-class wins above the threshold are retained, while the others fall back to the old classifier (Figure 1B).

Two factors determine how much new-class utility the certificate retains: more calibration examples make the finite-sample bound less conservative, and the conflict scorer should prioritize new-class predictions that are less likely to interfere with old classes. Stronger frozen features also separate correct new-class predictions from old-to-new errors better (§5). These observations motivate *Certified Conflict Policy Learning* (CCPL), which keeps the certification rule fixed while learning a reusable conflict scorer across registrations.

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.

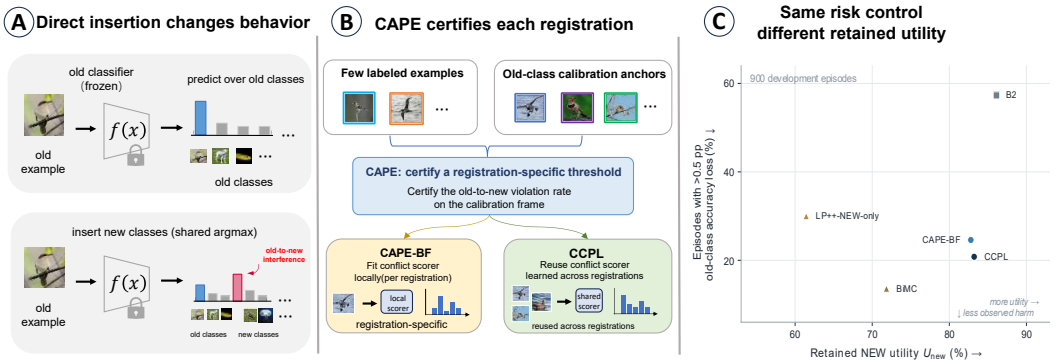


Figure 1: Overview of certified few-shot class registration. (A) Registering new classes can redirect old examples to new classes even when the old classifier is frozen. (B) CAPE uses old calibration anchors to certify a per-registration threshold that either retains a candidate new prediction or falls back to the old prediction. (C) Observed utility–harm operating points on the same 900 development episodes. B2 is the uncertified direct-insertion reference; CAPE-BF and CCPL use the CAPE framework, while BiMC and LP++-NEW-only are protocol-compatible external comparators. The harm axis reports observed old-class degradation and is not the quantity controlled by the certificate.

This decomposition supports two scorer regimes: registration-specific CAPE-BF fits a small conflict scorer separately for each registration, and amortized CCPL learns one scorer across development-training episodes and reuses it on new registrations without refitting. The certification step does not change: both regimes use the same calibration procedure, activity budget, and finite-sample rule (Figure 2). Across 900 development episodes, CAPE-family scorers retain substantially more new-class utility than external pipelines, and the amortized-over-registration-specific advantage grows on held-out populations and as episode-local fitting evidence shrinks (§7.2–§7.4).

Contributions. **(1)** We formalize the old-to-new errors that arise when new classes enter a frozen classifier and introduce CAPE, which provides a 95% finite-sample guarantee on each registration's calibration-frame old-to-new violation rate (§3–§4; Appendix A).

(2) We show that conflict scoring materially affects how much new-class utility can be retained under the same certification rule. Across 900 development episodes, CAPE-family scoring retains substantially more new-class utility than external scoring pipelines (§7.2; Appendix D).

(3) We instantiate CAPE in two regimes—registration-specific CAPE-BF and amortized cross-registration CCPL, which learns a reusable conflict scorer and applies it without refitting. Shared estimation improves the development macro and transfers without retraining to both held-out populations; mechanism and linear-scorer controls associate the advantage with shared estimation rather than network capacity (§6, §7.3–§7.4; Appendices I–J).

2 Related Work

Few-shot and class-incremental learning. Few-shot learning methods use small support sets through metric-based, prototype-based, or rapid-adaptation approaches (Snell et al., 2017; Vinyals et al., 2016; Qi et al., 2018; Finn et al., 2017; Raghu et al., 2020). Our setting differs in that newly registered classes compete directly with an incumbent old classifier, creating old-to-new interference that must be handled at registration time.

Class-incremental learning studies related competition between old and new classes, typically while updating representations or classifiers (Rebuffi et al., 2017; Wu et al., 2019; Hou et al., 2019; Zhao et al., 2020; Zhang et al., 2021; Zhou et al., 2022; Yang et al., 2023;

Tao et al., 2020). Training-free variants instead retain a frozen feature extractor while adjusting the new-class decision rule. These include prototype calibration (Wang et al., 2023) and BiMC's bi-level modality calibration (Chen et al., 2025); further FSCIL methods are surveyed in Zhang et al., 2025; Zhao et al., 2026; Wang et al., 2024. Our focus is complementary: per-registration arbitration with finite-sample control of old-to-new interference.

Risk control, arbitration, and conflict scoring. Our certificate uses exact binomial confidence inversion (Clopper & Pearson, 1934) and is related to Neyman–Pearson classification (Tong et al., 2018) and modern risk-control methods (Bates et al., 2021; Angelopoulos et al., 2024; 2025; Vovk et al., 2005; Angelopoulos et al., 2021; Einbinder et al., 2025). In class registration, the relevant event is induced by the shared OLD∪NEW decision rule: an old example can be reassigned to a newly registered class. CAPE applies finite-sample certification to this registration-specific event on a class-stratified calibration frame while treating conflict ranking as a separate component.

Selective classification rejects uncertain inputs (Geifman & El-Yaniv, 2017); our arbitration instead falls back from a candidate new-class win to the frozen old predictor. Conflict scoring uses frozen CLIP-style (Radford et al., 2021) and DINOv3 (Siméoni et al., 2025) features. LP++ (Huang et al., 2024) and related few-shot CLIP adaptation methods (Gao et al., 2024; Zhang et al., 2022; Silva-Rodríguez et al., 2024) primarily target predictive accuracy. CCPL instead learns a reusable conflict ranking across registrations, while finite-sample risk control is handled separately by the certificate.

3 Problem Setting

An episode e contains n_{old} *old* classes whose classifier is frozen, and N *new* classes, each with a small labeled support set S_{new}. Registration adds the new classes to a shared decision rule over $\mathcal{Y}_{\text{old}} \cup \mathcal{Y}_{\text{new}}$; queries then arrive from both old and new classes. Throughout, utility is U_{new}, the fraction of new-class queries that remain correctly classified after registration; old-side safety concerns queries whose correct label is an old class.

Direct insertion and its failure mode. A simple way to register a new class is to add its prototype or weight to the shared classifier (Qi et al., 2018). We use this direct-insertion rule as the reference method B2; it makes no attempt to protect old behavior. Because old and new classes share the same argmax, a new class can win an old-class query. We call this insertion-induced class-competition interference. It differs from classical forgetting because no old parameter is updated, yet old-class predictions can still change. Arbitration addresses this interference with an explicit decision layer. After direct insertion, a *conflict scorer* assigns a score to each candidate new-class win. A certified threshold determines whether that win is retained or the prediction falls back to the old classifier.

Risk contract. For a registration, we count an *OLD-to-NEW violation* whenever an example from an old class is reassigned to a new class. Our certificate controls the corresponding class-weighted rate on the finite calibration frame at $\epsilon = 2\%$. Held-out old-class queries are used only to evaluate behavior beyond that frame and are not part of the formal guarantee. The contract counts the change regardless of whether the old classifier was originally correct (correctness-damaging flips: Appendix E.1). Throughout, OLD and NEW denote the incumbent and newly registered class sets when used as formal labels.

Each episode also contains a fixed old-class calibration frame. We sample *calibration anchors* from this class-stratified frame independently of any scorer output. Direct insertion then identifies which sampled anchors would be routed to a new class, and the conflict scorer provides the scores used for threshold selection. The certificate's denominator is all sampled calibration anchors.

We develop the method on six development populations drawn from five source datasets: Cars (Krause et al., 2013), Caltech-A/B (Fei-Fei et al., 2007), Food-101 (Bossard et al., 2014), DTD (Cimpoi et al., 2014), and SUN397 (Xiao et al., 2010). tieredImageNet (Ren et al., 2018) is held out from method development. FGVC Aircraft (Maji et al., 2013)

is a second held-out population, used for a prospective held-out evaluation. §7 and the appendices give the full construction and separation rules.

4 Certified Class Registration

We call this certified registration framework **Certified Arbitration for Post-training Expansion (CAPE)**. It separates conflict scoring from finite-sample certification. The scorer orders candidate wins and calibration anchors, while the certificate selects the operating threshold for each registration. **CAPE-BF** is the registration-specific instantiation, fitting a balanced logistic scorer separately in each registration. **CCPL** (Section 6) is the amortized instantiation: it retains the CAPE framework and estimates the scorer across registrations rather than per registration.

Each threshold τ defines an operating point. A *conflict scorer* assigns a real score to each candidate new-class win and calibration anchor. Raising τ retains fewer wins and yields a violation count $k(\tau)$ among the n sampled calibration anchors. For each count we compute

$$\mathrm{CP}_u^{95}\big(k(\tau), n\big) = \sup\big\{p \in [0,1] : \Pr\big[\mathrm{Bin}(n,p) \le k(\tau)\big] > 0.05\big\}, \tag{1}$$

the exact one-sided 95% upper confidence bound on the violation probability, obtained by binomial inversion (Clopper & Pearson, 1934). A threshold is *certified* if $\mathrm{CP}_u^{95}(k(\tau), n) \le \epsilon$, with $\epsilon = 2\%$.

Minimum activity. Let k_{B2} be the number of calibration anchors that direct insertion routes to a new class, and let $k_{\max}(n)$ be the largest violation count satisfying $\mathrm{CP}_u^{95}(k, n) \le \epsilon$. The admitted-violation budget is

$$k^\star = \begin{cases} k_{\max}(n), & k_{B2} = 0, \\ \min\{k_{\max}(n),\, k_{B2} - 1\}, & k_{B2} > 0, \end{cases} \tag{2}$$

We choose the least conservative threshold on the grid constructed from sampled B2-intruded calibration-anchor scores that admits at most $k^\star$ violations. Since $k^\star \le k_{\max}(n)$, this choice preserves the certificate. When $k_{B2} > 0$, the rule rejects at least one sampled B2 intrusion.

This minimum-activity (*anti-vacuity*) clause is not an additional safety guarantee. It encodes the lexicographic operating objective "certificate validity $\succ$ non-vacuous intervention $\succ$ maximum new-class retention," preventing the certified rule from degenerating to direct insertion when the calibration frame contains observed B2 intrusions. On Caltech-A, episodes average 1.4 intruded anchors against a per-episode budget of at most five violations ($k_{\max} \in \{2, \dots, 5\}$ at the realized anchor-pool sizes), and the rule forces at least one intervention on 78 of 103 fitted episodes.

Validity. The certificate is a finite-sample statement about a precisely defined estimand. For a fixed scorer and calibration frame, the certified quantity is the design-weighted finite-pool old-to-new violation rate

$$\bar{p}_F(\tau) = \sum_j \frac{n_j}{n} \frac{M_j(\tau)}{N_j}, \tag{3}$$

where the calibration frame's old-query population is partitioned into classes j with N_j rows, n_j calibration anchors are sampled from class j ($n = \sum_j n_j$), and $M_j(\tau)$ is the number of class-j frame rows that would incur an old-to-new violation at threshold τ. The observed violation count is computed on the sampled anchors. Held-out queries are used only to evaluate behavior beyond the calibration frame.

Theorem 1 (Finite-pool registration guarantee)**.** *Fix an episode and condition on the calibration frames and the fitted conflict scorer. Suppose calibration anchors are drawn independently across old classes by within-class simple random sampling without replacement, as specified above. Let $\hat{\tau}$ be the operating threshold selected by the certified registration rule, including the minimum-activity budget $k^\star \le k_{\max}(n)$ of Eq. (2), and suppose further that $k_{\max}(n) + 1 \le n\epsilon$ (the operating regime of the deployed rule). Then*

$$\Pr\big[\bar{p}_F(\hat{\tau}) \le \epsilon\big] \ge 1 - \alpha, \qquad \epsilon = 0.02, \;\; \alpha = 0.05.$$

Appendix A proves fixed-threshold validity first and then handles data-dependent threshold selection through a finite-grid crossing argument.

To instantiate CAPE, we choose a conflict scorer. CAPE-BF, the registration-specific realization, fits an *episode-local balanced logistic arbiter*. Within each episode, negative (unsafe) training candidates are mined from the episode's legal fit frame under the fixed sampling procedure (Appendix B.3), while positives are correct new-class wins. A class-balanced logistic regression is fitted on the 10-D passive conflict-evidence block (dims 0–9 of the 17-D representation detailed in §6), and its scores feed the certificate above. Across the 900 development episodes, operational CAPE-BF reduces the fraction of episodes with more than a 0.5 percentage-point old-class accuracy loss from 57.3% under direct insertion to 24.6% (Figure 1C). This protection can reduce new-class utility, motivating scorer estimation across registrations rather than per-registration refitting (§6).

5 What Determines Certified Utility?

Certified utility depends jointly on calibration power and conflict-evidence quality. Appendix Figure 4 summarizes both effects.

Calibration power: a measurable finite-sample cost. On Cars we construct three nested old-repair frontiers. F_0 maximizes old-side repair subject only to the utility requirement $\Delta U_{\text{new}} > -5$pp, F_1 additionally imposes the nominal empirical constraint $k/n \leq 2\%$, and F_2 replaces that nominal rate with the finite-sample 95% upper bound in Eq. (1). The nominal 2% constraint costs little relative to F_0, while the exact finite-sample certificate retains 88.5% of the unconstrained old-side repair ($F_0/F_1/F_2$: $+1.420/+1.413/+1.257$pp).

This cost varies with calibration availability across populations. Food-101, whose census anchor pools average 1,109 rows per episode, loses essentially nothing, while Caltech-B falls between Cars and Food-101. For a zero-violation threshold, $1-0.05^{1/n} \leq \epsilon$ requires $n \geq 149$ anchors at $\epsilon = 2\%$. At the 530-row target, the certificate admits $k_{\max} = 5$ violations, compared with a nominal allowance of $n\epsilon = 10.6$.

The second factor is the quality of the conflict representation. Mean real-conflict AUC on Cars improves from 0.619 with incumbent single-level CLIP evidence to 0.725 with passive multi-level CLIP and 0.794 with DINOv3-based evidence. On separate evaluation episodes, DINOv3 evidence reaches 0.827 on Cars and 0.973 on Caltech-B.

Together, these results motivate better conflict ranking under a fixed certificate—and, in §6, learning that ranking across registrations instead of refitting it per registration.

6 Certified Conflict Policy Learning

Representation. The analysis in §5 suggests keeping the certificate fixed while improving the scorer. CCPL instantiates the amortized regime, learning one scorer across development-training registrations without per-registration refitting. It uses a 17-dimensional support-conditioned representation whose exact definitions appear in Appendix B.2. All 17 fields are deterministically available from the same frozen feature sources, but the scorer inputs differ. CAPE-BF uses only the 10-D passive block (dims 0–9), whereas CCPL uses the full 17-D representation.

Ranking objective. Within each episode we form pairs: the positive x^+ is a candidate whose new-class win is correct, the negative x^- is a candidate that would capture an old query or otherwise mislead. The policy q_ϕ is a small multilayer perceptron ($17-64-64-1$; 5,377 parameters) trained once across the development-training episodes with the pairwise logistic ranking loss

$$\mathcal{L}(\phi) \;=\; \mathbb{E}_e\, \mathbb{E}_{x^+,x^-}\; \log\Big(1+\exp\big(-\big(q_\phi(x^+)-q_\phi(x^-)\big)\big)\Big). \tag{4}$$

No risk penalty, certificate term, or utility weight enters the objective. The loss encourages correct new-class wins to receive higher scores than unsafe or misleading candidates. The

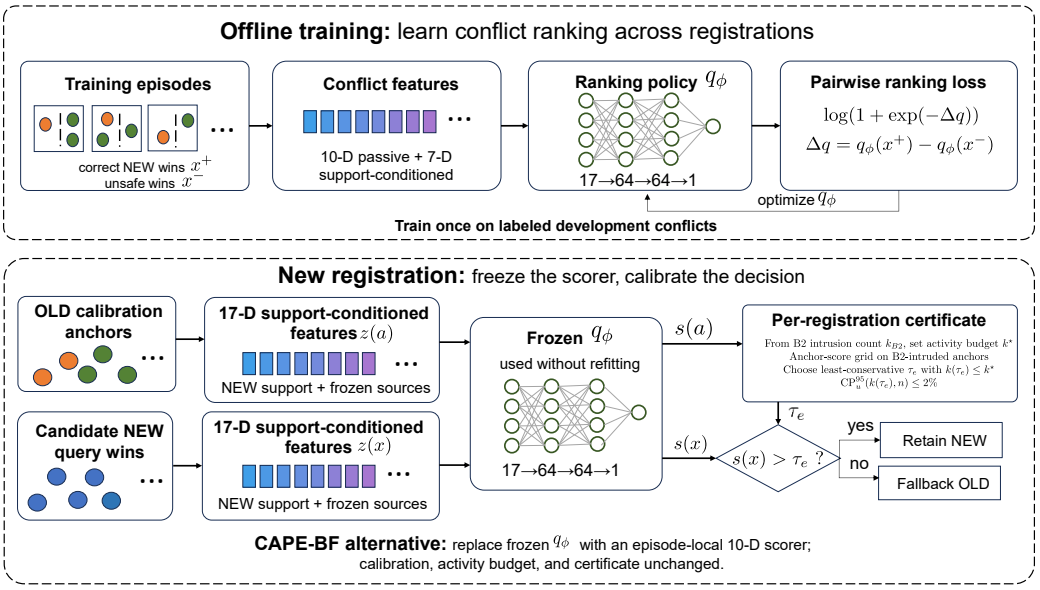


Figure 2: CCPL trains a shared 17-D conflict scorer q_ϕ once across development-training registrations using pairwise logistic loss (top). At registration, OLD calibration anchors and candidate NEW wins share the frozen scorer. Threshold candidates use only the sampled B_2-intruded calibration-anchor scores $s(a)$; the B_2 intrusion count k_{B2} sets the admitted-violation budget $k^\star$, and query scores $s(x)$ do not enter threshold selection. The finite-sample certificate selects τ_e; candidate wins are retained iff $s(x) > \tau_e$, otherwise they fall back to OLD. CAPE-BF uses an episode-local 10-D scorer with calibration and certification unchanged (Algorithm 1).

canonical instantiation evaluated in Tables 1 and 3 is the pairwise MLP (CCPL-MLP when distinguishing architectures); the object of study is cross-registration reuse of conflict structure, not network capacity. CCPL-Linear is a pooled linear instantiation with the same representation and certificate and no episode-local fitting (Appendix D).

Deployment. At test time CCPL scores every candidate and every calibration anchor with q_ϕ and applies the per-registration certificate of §4 unchanged (exact binomial inversion, minimum-activity budget, $\epsilon = 2\%$); a candidate is retained iff $q_\phi(x) > \tau_e$. Because ϵ does not enter scorer training, the same frozen policy spans risk levels by recalibration alone, and its advantage persists across the full certifiable sweep (Appendix I.6). Because the policy is learned across episodes, no scorer parameters are fitted inside the test episode; only the certification threshold is calibrated from the old-class anchors, including episodes where the registration-local fit of CAPE-BF is infeasible. The two regimes differ only in where the scorer is estimated: CAPE-BF fits its balanced logistic arbiter within each registration from the passive evidence block, whereas CCPL amortizes that fit into one ranking policy trained across registrations on the full 17-D representation, under the same calibration pool, activity budget, and certification rule.

7 Experiments

7.1 Setup

All experiments use the episodic registration protocol of §3 with $\epsilon = 2\%$. We evaluate 150 episodes per development population (900 in total) and 300 episodes on each held-out population. The new-class budget is 5-shot per class. Certification uses a class-stratified old-class calibration take (Appendix B.4), so "few-shot" refers to the registered new classes, not the safety system's total data.

Population	ΔU_{new} (CCPL − CAPE-BF), pp	
	Operational (all episodes)	Fitted-only (fit-coverage controlled)
Cars	+0.37	+0.58
Caltech-A	−0.22	−0.12
Caltech-B	+0.94	+1.01
Food-101	−0.17	+0.41
DTD	+2.07	+2.23
SUN397	−0.39	+1.22
Macro	**+0.43** [0.16, 0.71]	**+0.89** (pooled +1.03)

Table 1: **Six development populations.** Operational: all episodes (CAPE-BF deploys direct insertion where it cannot fit its episode-local arbiter). Fitted-only: the 672/900 episodes where CAPE-BF fits, so both arms operate as certified arbitration policies under matched calibration, activity budget, threshold selection, and certification. Macro CIs are 95% paired episode-bootstrap intervals (B=10,000). All certified registrations satisfy $\text{CP}_u^{95} \leq 2\%$.

Algorithm 1: CCPL registration for episode e

1. **Input:** frozen old classifier; support sets S_{new}; calibration anchors $a_1, \ldots, a_n$; policy q_ϕ (trained once, offline); $\epsilon = 2\%$.
2. Build the 17-D representation and score all candidates and anchors: $s_i = q_\phi(x_i)$.
3. Construct threshold candidates from *sampled B2-intruded calibration-anchor scores only* (next-lower values; conservative ties; $-\infty$ if all admissible; the exact maximum anchor score for zero retention); query scores never enter threshold selection. $k(\tau)$ is the number of routed anchors still retained ($s_i > \tau$).
4. Compute the admitted-violation budget $k^\star$ (Eq. (2)) from the B2 intrusion count k_{B2}; select the least conservative candidate on this grid with $k(\tau_e) \leq k^\star$ (exact binomial inversion).
5. Retain candidate new-class wins with $s_i > \tau_e$; all other queries fall back to the old prediction.

The frozen old classifier, support sets, calibration anchors, and certification machinery are shared by all CAPE-family arms. CAPE-BF scores candidates from the 10-D passive evidence block, whereas CCPL uses the full 17-D input. CCPL is trained once on the development-training episodes with the pairwise ranking loss in Eq. (4) and evaluated on held-out data without retraining; training and reported evaluation episodes are disjoint. We report means with 95% paired bootstrap intervals (B=10,000). For descriptive old-side evaluation, harm05 is the fraction of episodes with an old-class accuracy loss above 0.5pp, and ΔOLD the change in old-class accuracy relative to direct insertion.

7.2 Development populations

Using the same evaluation episodes and certification protocol, CAPE-family conflict scoring preserves substantially more new-class utility than the two protocol-compatible external scoring pipelines, LP++-NEW-only (Huang et al., 2024) and BiMC (Chen et al., 2025) (Table 2). CCPL-MLP loses only 2.89pp from the 86.11% direct-insertion reference, compared with 24.67pp for LP++-NEW-only and 14.26pp for BiMC. It meets the −5pp utility criterion in 5/6 populations (Table 2).

Appendix D decomposes the external-pipeline gaps. For LP++-NEW-only, 4.45pp of the gap appears before certification, 18.98pp at the certification stage, and 1.24pp from minimum activity.

BiMC shows a different pattern. Its native insertion is +6.30pp above direct insertion before certification, while the deployed deficit emerges under certified arbitration: −19.10pp at the certificate-only stage and a further −1.46pp from minimum activity (Appendix D).

Registration policy	U_{new} (macro)	ΔU_{new} vs. B2 (pp)	$\Delta U_{\text{new}} > -5$pp
B2 direct insertion (uncertified ref.)	86.11%	– (reference)	–
CAPE-BF (operational)	82.79%	-3.32 $[-3.69, -2.98]$	5/6
CCPL-MLP (certified)	83.22%	-2.89 $[-3.23, -2.55]$	5/6
LP++-NEW-only (matched cert.)	61.44%	-24.67 $[-26.14, -23.21]$	1/6
BiMC (matched cert.)	71.85%	-14.26 $[-16.54, -12.07]$	2/6

Table 2: **Development comparison on the same 900 episodes.** ΔU_{new}: per-episode paired difference against direct insertion (95% paired episode bootstrap); last column: populations with mean $\Delta U_{\text{new}} > -5$pp. All scored arms share the evaluation episodes, calibration anchors, threshold-selection rule, activity budget, and finite-sample certificate; they differ in score construction and, where applicable, the OLD/NEW score matrix and fallback (CAPE-BF and CCPL-MLP: CAPE conflict scorers over frozen B2 old-class evidence; LP++-NEW-only: bridged new-class scores over B2 old-class evidence; BiMC: a native OLD∪NEW score matrix whose rejected new-class wins fall back to its own old-class argmax). B2 is the uncertified reference; CAPE-BF is operational, deploying direct insertion where its episode-local scorer cannot be fit; CCPL-MLP is certified on all 900 episodes; LP++-NEW-only and BiMC are protocol-compatible external comparators (BiMC added post hoc). Appendix D (Table 10) gives per-population results and implementation details.

Population	BF fitted	CAPE-BF U_{new}	CCPL-MLP U_{new}	ΔU_{new} (95% CI), pp
tieredImageNet	291/300	79.51%	80.39%	**+0.88** [0.44, 1.33]
FGVC Aircraft	300/300	31.77%	35.04%	**+3.27** [2.56, 3.99]

Table 3: **Held-out transfer** (300 episodes per population; CCPL-MLP without retraining). CAPE-BF fits all Aircraft episodes, so operational and fitted-only coincide; all certified registrations satisfy $\text{CP}_u^{95} \leq 2\%$.

Operationally, CCPL improves over deployed CAPE-BF by +0.43pp in macro ΔU_{new} (95% CI $[0.16, 0.71]$; the amortized-versus-registration-specific comparison). Population-level effects are heterogeneous, with three negative point estimates whose intervals include zero (Table 1). Restricting to the 672 episodes where CAPE-BF fits yields a +0.89pp macro gain (+1.03pp pooled), showing that the advantage is not explained solely by CAPE-BF fallback episodes.

LP++-NEW-only improves old-class accuracy relative to direct insertion (non-negative ΔOLD on all six populations; Table 7) at a much larger utility cost (61.44% U_{new}; 30.0% harm05). BiMC, included as a post-hoc external comparator, scores the full class set natively (no bridge) and routes almost no old queries (0.42% OLD→NEW; +6.59pp ΔOLD).

On the same 900 episodes, CCPL-MLP attains 83.22% U_{new} with 20.8% harm05, compared with 82.79% / 24.6% for CAPE-BF and 71.85% / 13.6% for BiMC (Figure 1C). BiMC therefore occupies a more conservative operating point, with lower observed old-side harm but substantially lower retained new-class utility than CCPL; neither method dominates the other on these two axes. A prespecified variant with fixed β=0.85 attains 70.80% U_{new} and 17.9% harm05 (Appendix D).

A frozen margin-based alternative trails CAPE-BF by 6.46pp (95% CI $[5.89, 7.06]$; Table 6). The certificate alone does not explain the retained utility. CCPL-Linear, the pooled cross-episode linear instantiation trained on the identical cross-episode items, matches the canonical CCPL-MLP (+0.44pp operationally, within 0.01pp of CCPL; Appendix D), indicating that the effective ingredient is shared cross-registration estimation rather than nonlinear scorer capacity. Representation-matched episode-local controls still favor cross-episode scoring under both the full 17-D and shared 10-D inputs, indicating a representation-by-policy-pipeline interaction (Appendix D).

7.3 Held-out transfer

Both evaluations apply the frozen scorer to unseen populations, testing cross-registration reuse without retraining. On tieredImageNet, CCPL improves U_{new} by +0.88pp (95% CI $[0.44, 1.33]$; Table 3) and by +0.97pp on the 291 episodes where CAPE-BF fits (Appendix E.2).

On FGVC Aircraft, where CAPE-BF fits on all 300 episodes, CCPL-MLP improves U_{new} over CAPE-BF by +3.27pp (95% CI $[2.56, 3.99]$; Appendix B.4). CCPL-Linear improves over CAPE-BF by +3.56pp (95% CI $[2.88, 4.26]$) and is non-inferior to the canonical MLP under the prespecified -0.5pp margin (+0.29pp, 95% CI $[-0.21, 0.78]$).

On Aircraft, direct insertion attains 52.09% U_{new}, routes 7.54% of old queries to new classes, and has harm05 of 96.3%. CCPL-MLP attains 35.04% U_{new} while reducing OLD→NEW routing to 0.92% and harm05 to 14.7%. Appendix I reports additional transfer and representation controls.

7.4 Why does cross-registration scoring help?

We test whether reusable scoring helps by regularizing unstable episode-local estimation: reducing local fitting evidence monotonically increases ranking instability and the CCPL advantage (Appendix J).

Across repeated local refits, reducing the fitting fraction from 1/2 to 1/8 increases within-episode ranking instability by 0.079 on average over the 672 fitted episodes (95% CI $[0.074, 0.083]$). At matched 1/4 evidence, less stable local rankings are associated with larger gains (Spearman = 0.195, 95% CI $[0.126, 0.263]$; Table 20). On the 514 episodes fitted at every fraction, the advantage rises monotonically from +1.26pp to +3.97pp.

A matched linear 10-D shrinkage intervention recovers certified utility as the episode-local estimator approaches the shared scorer, with the fully global endpoint essentially matching CCPL. The prespecified prediction that the preferred shrinkage level would shift with local evidence was not supported (Table 21).

Finally, label shuffling tests whether the cross-episode gain depends on meaningful supervised feature–label correspondence. It removes the global scorer's advantage: the real scorer exceeds the shuffled control by 9.05pp (95% CI $[8.49, 9.62]$; Table 23). The advantage also persists when the evaluated population is excluded from training.

8 Limitations

The aggregate advantage of amortized over registration-specific scoring remains positive across the tested settings, but its magnitude and preferred scorer depend on the representation, and population-level effects vary across calibration draws. Mechanism results are consistent with estimation regularization but do not identify it uniquely; several associations remain exploratory (Appendix J).

The certificate applies per registration under the §4 assumptions. Longer trajectories, larger class spaces, and non-vision modalities remain open. Across four cumulative registration waves, the frozen policy retains a +3.02pp advantage over CAPE-BF, but certification remains per-wave rather than trajectory-level (Appendix I.4).

9 Conclusion

Registration can change old-class behavior even when old parameters are frozen: new and old classes compete in the same decision rule. CAPE provides finite-sample control of the calibration-frame old-to-new violation rate and supports two scorer regimes: registration-specific fitting (CAPE-BF) and amortized cross-registration estimation (CCPL), which retains new-class utility without deployment-time fitting. Across development and both held-out evaluations, the results support separating learned conflict ranking from per-registration risk certification.

AI Use Statement Generative AI tools were used under author direction to assist with manuscript writing and editing and with writing and debugging evaluation code. The authors designed the study, developed the methods, performed all experiments, interpreted the results, verified all reported numbers, and take full responsibility for the final paper.

Reproducibility Statement All evaluation populations are public (Stanford Cars, Caltech-A/B, Food-101, DTD, SUN397, tieredImageNet, and FGVC Aircraft). Episode construction, role splits, and calibration-anchor pools are deterministic functions of documented seeds (Appendix C); scorer training uses fixed seeds with no early stopping or per-run model selection, and the frozen scorer weights and standardizers are included in the supplementary package. Confidence intervals use the paired-bootstrap procedure with the fixed generator and seed described in Appendix F. The pipeline from raw data to every reported table is scripted and deterministic, using the public feature extractors named in the paper (DINOv3 ViT-B/16 canonically; CLIP ViT-B/16 and ImageNet ResNet-50 for the representation controls) under Python 3.10+ with NumPy, SciPy, PyTorch, and scikit-learn. The prospective evaluations use specifications fixed before the corresponding outcomes were evaluated, with no post-outcome model or hyperparameter retuning; external comparators are included post hoc as disclosed in §7.2. Calibration-take sensitivity and correction notes are documented in Appendices B.4 and J. Supplementary materials include the evaluation records and protocol documentation needed to reproduce the reported tables and figures.

References

Anastasios N. Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 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. *Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6): 1–34, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014.

Yiyang Chen, Tianyu Ding, Lei Wang, Jing Huo, Yang Gao, and Wenbin Li. Enhancing few-shot class-incremental learning via training-free bi-level modality calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9881–9890, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

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.

Bat-Sheva Einbinder, Liran Ringel, and Yaniv Romano. Semi-supervised risk control via prediction-powered inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11080–11090, 2025. doi: 10.1109/TPAMI.2025.3594263.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories.

Computer Vision and Image Understanding, 106(1):59–70, 2007. doi: 10.1016/j.cviu.2005.09.012.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23773–23782, 2024.

Stefen Hui and C. J. Park. The representation of hypergeometric random variables using independent bernoulli random variables. *Communications in Statistics – Theory and Methods*, 43(19):4103–4108, 2014. doi: 10.1080/03610926.2012.705941.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. *ICCV Workshops*, 2013.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5822–5830, 2018. doi: 10.1109/CVPR.2018.00610. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Qi_Low-Shot_Learning_With_CVPR_2018_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgMkCEtPB.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018.

Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23681–23690, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman–Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018. doi: 10.1126/sciadv.aao1659.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuan Wang, Zhong Ji, Xiyao Liu, Yanwei Pang, and Jungong Han. On the approximation risk of few-shot class-incremental learning. In *European Conference on Computer Vision*, pp. 162–178, 2024.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip H.S. Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *ICLR*, 2023.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *CVPR*, 2021.

Jinghua Zhang, Li Liu, Olli Silvén, Matti Pietikäinen, and Dewen Hu. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2924–2945, 2025. doi: 10.1109/TPAMI.2025.3529038.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, 2020.

Li-Jun Zhao, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. Hierarchical filtering and refinement classification for few-shot class-incremental learning. *Transactions on Machine Learning Research*, 2026.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shijie Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *CVPR*, 2022.

A Certificate and Finite-Pool Guarantee

This appendix gives a self-contained proof of the certificate stated in §4:

$$\Pr\left[\bar{p}_F(\hat{\tau}) \le \epsilon\right] \ge 0.95$$

for the deployed, data-dependent threshold $\hat{\tau}$.

A.1 Estimand and assumptions

Fix an episode. The *calibration frame* is the fixed, unordered set of candidate rows formed from each old class's support pool and disjoint from all query parts. It is partitioned into per-class pools j of size N_j. For a threshold τ, let $M_j(\tau)$ denote the number of rows in pool j that would be routed to a new class, constituting an *old-to-new violation*. The certified quantity is the design-weighted finite-pool violation rate

$$\bar{p}_F(\tau) = \sum_j \frac{n_j}{n} \frac{M_j(\tau)}{N_j},$$

where n_j calibration anchors are drawn from pool j and $n = \sum_j n_j$. We assume:

1. **Unordered pools.** Each class pool is a fixed multiset of item/score pairs with no sampling order. All probability statements are conditional on these pools.
2. **Within-class SRS-WOR / exchangeability.** The n_j anchors from pool j are drawn by simple random sampling without replacement, independently across pools. In our implementation, the calibration take is generated by a dedicated classwise pseudorandom SRS-WOR draw, with the seed construction specified in §B.4. The draw is used only for evaluation and is fixed before any scorer is applied to the episode. The quota sizes n_j are coupled across pools by the water-filling rule; conditional on these quotas, the pool orderings are treated as independent.
3. **Fit/calibration separation.** For the episode being certified, the conflict scorer is fixed before the calibration-anchor draw and does not use that episode's calibration anchors for fitting or training. CAPE-BF is fitted only on the current episode's new-support positives and fit-frame negatives (§B, §B.4). The CCPL policy is trained on disjoint development-training episodes using the records specified in §C, and is then held fixed on evaluation episodes.

   Validity is design-based conditional on the evaluation episode's fixed calibration frame and scorer. Probability is taken over the classwise SRS-WOR calibration-anchor draws. No assumption on scorer quality is required; scorer quality affects utility, not the guarantee.
4. **Outside the certification regime.** If $k_{\max}(n) < 0$, then no violation count satisfies $\mathrm{CP}_u^{95}(k, n) \le \epsilon$ at that sample size, and no certificate is asserted. The certified system therefore refuses the registration rather than certifying it at a degraded risk level. An operational deployment may instead specify an uncertified fallback, such as B2 pass-through when an episode-local scorer cannot be fitted. Such a fallback is operational and carries no certificate.

A.2 Fixed-threshold validity

Proposition A.1 (Fixed threshold)**.** *Fix τ with $\bar{p}_F(\tau) \ge \epsilon$, and let $K(\tau) = \sum_j K_j(\tau)$ denote the number of sampled anchors that violate at τ, where*

$$K_j(\tau) \sim \mathrm{Hypergeo}(N_j, M_j(\tau), n_j)$$

independently across classes. If $k_{\max}(n) + 1 \le n\epsilon$, then

$$\Pr\left[K(\tau) \le k_{\max}(n)\right] \le 0.05.$$

Proof. Each K_j admits a Bernoulli-sum representation for sampling without replacement (Hui & Park, 2014). Because the classwise draws are independent, $K(\tau)$ is Poisson-binomial with mean $n\,\bar{p}_F(\tau)$ and heterogeneous per-trial success probabilities.

A false certification at the fixed threshold τ can occur only when

$$K(\tau) \le k_{\max}(n).$$

The operating condition $k_{\max}(n) + 1 \le n\epsilon$, together with $\bar{p}_F(\tau) \ge \epsilon$, gives

$$k_{\max}(n) + 1 \le n\bar{p}_F(\tau).$$

Thus the interval $[k_{\max}(n) + 1, n]$ begins no higher than the mean of the Poisson-binomial sum. By the interval comparison of Hoeffding (1956), the equal-probability binomial $\mathrm{Bin}(n, \bar{p}_F(\tau))$ minimizes the probability of this interval. Taking complements gives

$$\Pr\left[K(\tau) \le k_{\max}(n)\right] \le \Pr\left[\mathrm{Bin}(n, \bar{p}_F(\tau)) \le k_{\max}(n)\right].$$

Exact binomial inversion (Clopper & Pearson, 1934) gives

$$\Pr\left[\mathrm{Bin}(n, p) \le k_{\max}(n)\right] \le 0.05$$

for every $p \ge \epsilon$ whenever $\mathrm{CP}_u^{95}(k_{\max}(n), n) \le \epsilon$. Setting $p = \bar{p}_F(\tau) \ge \epsilon$ therefore yields

$$\Pr\left[K(\tau) \le k_{\max}(n)\right] \le 0.05.$$

□

A.3 Proof of Theorem 1

Theorem 1 follows from Proposition A.1 by a crossing argument on the finite threshold grid. Both $\bar{p}_F(t)$ and $K(t)$ are non-increasing in t, and $k^\star \le k_{\max}(n)$ by construction, so the activity budget can only tighten the certificate.

The deployed threshold $\hat{\tau}$ depends on the sampled anchors, and its runtime threshold grid is therefore random. We instead carry out the crossing argument on a deterministic master grid. Let $\mathcal{T}_F$ be the finite set containing (i) every scorer value over the rows of the fixed calibration frame, (ii) for each such value v, the candidate threshold $\mathrm{nextafter}(v, -\infty)$ used by the selection rule, and (iii) the sentinel $-\infty$ together with one value above the largest scorer value.

Every threshold that the selection rule can deploy belongs to $\mathcal{T}_F$. Moreover, $\mathcal{T}_F$ depends only on the fixed calibration frame and fitted scorer; all relevant quantities are constant on the intervals between adjacent scores in $\mathcal{T}_F$.

The runtime grid induced by the sampled anchors is contained in $\mathcal{T}_F$. Its thresholds consist of the next-lower representable values of sampled-anchor scores, the $-\infty$ sentinel, and, when zero retention is required by dense ties or the activity budget, the exact maximum sampled-anchor score.

Define

$$t^\dagger = \max\{t \in \mathcal{T}_F : \bar{p}_F(t) > \epsilon\}.$$

If this set is empty, then $\bar{p}_F(t) \le \epsilon$ for all $t \in \mathcal{T}_F$, and the claim is immediate. Otherwise, $t^\dagger$ is fixed conditional on the calibration frame and satisfies $\bar{p}_F(t^\dagger) > \epsilon$.

Suppose $\bar{p}_F(\hat{\tau}) > \epsilon$. Then $\hat{\tau} \in \mathcal{T}_F$ and, by the definition of $t^\dagger$,

$$\hat{\tau} \le t^\dagger.$$

Since $K(t)$ is non-increasing in t,

$$K(t^\dagger) \le K(\hat{\tau}) \le k^\star \le k_{\max}(n).$$

Therefore

$$\{\bar{p}_F(\hat{\tau}) > \epsilon\} \subseteq \{K(t^\dagger) \le k_{\max}(n)\}.$$

Because $t^\dagger$ is fixed conditional on the frame and satisfies $\bar{p}_F(t^\dagger) > \epsilon$, Proposition A.1 applied at $\tau = t^\dagger$ bounds the probability of this event by 0.05. Hence

$$\Pr\left[\bar{p}_F(\hat{\tau}) \le \epsilon\right] \ge 0.95.$$

The data dependence of $k^\star$, through the observed B2 intrusion count k_{B2}, can only tighten the admission rule. When $k_{B2} = 0$, the rule reduces to the pure certificate; when $k_{B2} > 0$, the admitted count is smaller. The violation numerator counts sampled OLD anchors that direct insertion routes to NEW and arbitration still retains as NEW ($s > \tau$), while the denominator is all n sampled calibration anchors.

No-fit episodes require no separate argument. Under Assumption 4, no certificate is asserted for such episodes.

B Exact Method Specification

B.1 B2: direct insertion

For each class c, let t_c denote its frozen text anchor and p_c its visual prototype. B2 scores a query x by

$$s_{B2}(x, c) \;=\; 0.75\,\langle f_v(x),\, t_c\rangle \;+\; 0.25\,\langle f_v(x),\, p_c\rangle, \tag{5}$$

and predicts

$$\hat{y} \;=\; \arg\max_{c \in \mathcal{Y}_{\text{old}} \cup \mathcal{Y}_{\text{new}}} s_{B2}(x, c), \tag{6}$$

a single shared argmax over old and new classes. The prototype is the class memory: for an old class, p_c is computed from the stored memories of the frozen classifier; for a new class, from the support-set features S_{new}. B2 uses this argmax directly and has no arbitration, calibration threshold, or episode-local fitting; no component of it is learned or tuned per episode.

B.2 The 17-dimensional conflict representation

All 17 fields are computed deterministically from the same frozen feature sources in a fixed order, but the scorer inputs differ: CAPE-BF's episode-local arbiter consumes only the passive block (dims 0–9), whereas CCPL consumes all 17 dimensions—the passive block plus the seven support-conditioned dims 10–16. Both are evaluated under the same calibration, activity-budget, threshold-selection, and certification machinery:

Dimensions	Content
0–9	Eight passive multi-level DINOv3 conflict margins, one cross-level consistency feature, and one B2 evidence margin (exact definitions below)
10	$E_o^{\max}$: strongest old-class evidence for the candidate
11	$E_n^{\max}$: strongest new-class evidence for the candidate
12	$E_o^{\max} - E_n^{\max}$ (evidence margin)
13	cosine similarity between the candidate and the NEW-class prototype
14	evidence between the NEW prototype and the strongest OLD class
15	mean pairwise support cosine (support dispersion)
16	$\cos(x, \text{nearest support}) - \cos(x, \text{farthest support})$ (ambiguity)

Standardization is a per-dimension affine transform (mean/scale). CCPL-MLP and CCPL-Linear share one pooled standardizer fitted once on the development-training items only and reused unchanged at evaluation; CAPE-BF instead fits its own per-episode standardizer on the current episode's legal fit data, together with its logistic scorer and before calibration.

Passive conflict features (dims 0–9). For a conflict candidate x, let $q_\ell(x)$ be its token stack at level ℓ of a frozen DINOv3 ViT-B/16 encoder, levels $\ell \in \{5, 7, 9, 11\}$; each level contributes one CLS token and 196 patch tokens (the encoder's four register tokens are discarded), and all features are L_2-normalized in float64. Each class contributes a per-image bank of five images at every level. Dims 0–3: the patch-Chamfer margins $C_\ell(x, \text{new}) -$

$C_\ell(x, \text{old})$, where $C_\ell(x, c)$ averages over the 196 query patches the maximum cosine to class c's bank patches at level ℓ. Dims 4–7: the CLS-to-prototype margins $\cos\big(h_\ell(x), \mu_\ell(\text{new})\big) - \cos\big(h_\ell(x), \mu_\ell(\text{old})\big)$, where h_ℓ is the level-ℓ CLS token and $\mu_\ell(c)$ the L_2-normalized mean of class c's 197 bank tokens at that level. Dim 8: the B2/CLIP evidence margin $E_n^{\max} - E_o^{\max}$, the only CLIP-derived dimension among dims 0–9; it repeats the active evidence margin of dim 12 with opposite sign, a legacy signed pair retained unchanged for compatibility. Dim 9: the cross-level consistency, the fraction of dims 0–3 whose margin is positive. Under strict leave-one-out, a pseudoquery's own image is excluded from its own class's memory and bank. Dims 0–7 and the consistency dim 9 are DINOv3-derived; dim 8 and the active dims 10–16 use CLIP ViT-B/16 features. Throughout, the class evidence is $E(x, c) = 0.75\,\langle f_v(x), t_c\rangle + 0.25\,\langle f_v(x), p_c\rangle$ (frozen weights $W = (0.75, 0.25, 0)$: the normalized class text anchor t_c, the normalized mean p_c of the class's stored exemplar features, and a zero-weighted memory-consistency term), so $E_o^{\max} = \max_{c\in\mathcal{Y}_{\text{old}}} E(x, c)$, $E_n^{\max} = \max_{c\in\mathcal{Y}_{\text{new}}} E(x, c)$, and dim 14 is $\max_{c\in\mathcal{Y}_{\text{old}}} E\big(p_{\text{new},j},\, c\big)$ for the candidate's predicted new class j.

B.3 Data roles and information conditions

For each evaluation episode, five disjoint item pools have fixed roles:

Pool	Uses
NEW support (5 per NEW class)	prototypes; scorer positives (strict LOO)
OLD memory (5 per OLD class)	prototypes / frozen OLD class memory
Fit frames (remainder of each OLD class's support pool)	CAPE-BF negatives (frozen sequential sampler; up to 16 per NEW class with shared cross-class de-dup/rejection)
Calibration anchors (n_e rows; water-fill target 530, except for the Food-101 census case, subject to per-class frame capacity)	certificate only: violation counts and the threshold
Old and new queries	evaluation only

Within an episode, CAPE-BF fits a class-balanced L_2 logistic scorer (C=1) on standardized margins: positives are the strict-LOO new-support pseudoqueries that B2 routes to a new class; negatives are mined from the legal fit frame by the fixed sampling procedure. The sampler draws candidates sequentially under a single episode-level RNG stream with shared cross-class de-duplication, and admits at most 16 negatives per new class, which defines the registration-time fitting budget used throughout. An episode whose pseudoqueries form a single label class cannot fit; it deploys B2 pass-through and asserts no certificate (9 of 300 tieredImageNet held-out episodes; 228 of 900 development episodes). Queries never enter fitting or calibration; the certificate is computed only on calibration anchors. Development-training episodes use their own query and anchor records for CCPL policy training as specified in §C; no evaluation-episode query or calibration anchor enters policy training.

B.4 Calibration-frame construction and sampling

The calibration anchors are produced by a single fixed construction, run before any scorer is applied to the episode and before any query is evaluated. For each old class j of the episode, let P_j be the class's support-part rows minus the episode's five old-memory rows. Each P_j is split once, by the hash order sha256(`role-split-v1` | ds | j | id). For all populations except Food-101, the first $\min(27, |P_j|)$ rows form the *calibration frame* G_j (the finite pool of §A, $N_j = \min(27, |P_j|)$), and the remainder forms the *fit frame* F_j; the split is disjoint and F_j is never enlarged when G_j is short. Food-101 is the census exception: its disjoint 40-row census part forms G_j, and all rows are used as calibration anchors ($n_j = N_j = 40$). Within each G_j outside the census case, per-class quotas are set by round-robin water-filling to a target of 530 rows (class counts differ by at most one; capped classes are skipped; if the total falls short, all rows are used), and the anchors $C_j \subseteq G_j$ are a classwise simple random sample without replacement of G_j: with a deterministic dedicated namespace string

(exact string and seed derivation in the released code), each class j contributes a generator $\mathrm{PCG64}(\mathrm{int}(\mathrm{sha256}(\texttt{namespace}|\,ds|\,ep|\,j)[:16], 16))$ whose uniform permutation of G_j supplies the first quota_j anchors. The realized episode size is $n_e = \min(530, \sum_j |G_j|)$, and the certificate's $k_{\max}$ is computed from the realized n_e. The namespace is dedicated to this stage (used by no other stage, in particular not during scorer training), and the take is fixed before any scorer is applied to the episode; the remaining seeded draws in an episode are its manifest (classes, memories, supports, queries) and the CAPE-BF negative-mining seed.

Canonical calibration sampling. The reported certificate results use an independent classwise SRS-WOR draw for each episode. An earlier implementation used a fixed hash ordering shared across episodes and contributed calibration-anchor material to scorer training. We therefore adopted the episode-specific sampling procedure above before evaluating the final results. Table 4 summarizes its realized sample sizes.

Population	Support pool per old class	Anchor rule	Realized n_e (fitted episodes)
Cars	29–68 rows	water-fill to 530	530 (250/250)
Caltech-A	6–772 rows	water-fill to 530	mean 506, 359–530 (103)
Caltech-B	5–774 rows	water-fill to 530	mean 491, 342–530 (278)
Food-101	census part, 40 rows	full census (n_j=N_j)	mean 1,109, 800–1,400 (236)
DTD	72 rows	water-fill to 530	530 (297/297)
SUN397	74 rows	water-fill to 530	530 (182/182)
tieredImageNet	72 rows	water-fill to 530	530 (291/291)
FGVC Aircraft	40–41 rows	water-fill to 530	530 (300/300)

Table 4: **Anchor construction by population.** Seven populations share the water-filled 27-row-per-class calibration frame with a 530-row take target; realized n_e falls short only where native pools are small (Caltech-A/B), and the certificate uses the realized n_e throughout. Food-101 is a census special case: anchors are the class's full disjoint 40-row census part, so $n_j = N_j$ and the sampling guarantee holds degenerately.

Calibration-take sensitivity. The aggregate CCPL–BF effects remain positive under both calibration draws, although some population-level estimates change (per-population detail in Appendix D); Table 5 reports the headline quantities under both takes. Results from the earlier draw are reported only as sensitivity analyses; all certificate claims use the final SRS-WOR procedure. Per-population old-side descriptives move the same way: under the earlier take, direct insertion loses more than 0.5 percentage points of old-class accuracy in 72.3% of Cars episodes versus 44.0% under operational CAPE-BF (88.3% to 17.0% on DTD).

Quantity	Earlier calibration draw	Final SRS-WOR draw
Development operational macro	$+0.54$ $[0.25, 0.83]$	$+0.43$ $[0.16, 0.71]$
Development fitted-only macro	$+1.05$	$+0.89$
tieredImageNet	$+0.78$	$+0.88$
FGVC Aircraft	$+3.20$ $[2.51, 3.89]$	$+3.27$ $[2.56, 3.99]$
SUN397 operational	$+0.14$	-0.39

Table 5: **Calibration-take sensitivity.** ΔU_{new} (CCPL − CAPE-BF) in percentage points under the earlier calibration draw and the final SRS-WOR draw. Final-draw values correspond to the canonical results reported in §7.2 and §7.3; earlier-draw values are reported here as calibration-take sensitivity. The aggregate CCPL–BF effects remain positive under both draws.

Cross-episode information separation. Training and evaluation are episode-disjoint but not necessarily image-disjoint within a development population. Because the role split is defined by dataset and class while calibration anchors are sampled independently for each episode, the same underlying image may appear as a calibration anchor in different

training and evaluation episodes. Evaluation-episode queries never enter scorer training, and evaluation-episode calibration anchors are not used to fit the local CAPE-BF arbiter.

Only the final calibration construction described above is used for the reported certificate results. Calibration anchors are first routed by the shared frozen insertion rule, independently of the conflict scorer. All anchors may be scored, but only B2-intruded anchors can contribute violations or threshold candidates. Queries are evaluated only after the registration threshold has been selected.

B.5 The CCPL policy

CCPL is a multilayer perceptron $17-64-64-1$ with tanh hidden activations and a linear logit output (5,377 parameters), trained once as described in §C. At deployment it scores every conflict candidate and every calibration anchor.

C Training and Evaluation Protocol

C.1 Episodes and splits

Each development population provides 300 episodes. Policy training uses fitted episodes 150–299, while the 150 evaluation episodes reported in the main text are episodes 0–149. The two sets are disjoint, and no evaluation episode is used during policy training.

tieredImageNet, the original held-out population, is constructed only after the method and evaluation protocol are fixed. We use a 40-class universe drawn class-disjointly from the training split. For each class, the hash-determined item order is partitioned into 8 old-memory items, 72 support items, 6 old-query items, and 12 new-query items. We construct 300 episodes, with the number of old classes cycling over $\{20, 25, 30, 35\}$ (75 episodes each, shuffled). Every episode contains 5 new classes, fixed 5-shot support for each new class, 5 stored memories per old class, 60 new-query items, and a calibration-anchor target of 530 rows. The construction uses a fixed seed.

No tieredImageNet episode is used for policy training, method development, model selection, or hyperparameter tuning. Per-episode operating thresholds are selected only by the frozen certification rule of §4. Query records are stored for the 291 episodes where CAPE-BF fits; the 9 no-fit episodes are recomputed at evaluation time. This is why the query-item totals in Appendix E refer to the 291 fitted episodes.

C.2 Pair construction and loss

Within a development-training episode, the conflict candidate set includes candidate new-class wins on new queries, candidates that capture old queries, and that training episode's calibration-anchor records. A *positive* x^+ is a candidate whose new-class win is correct. A *negative* x^- is either a candidate that captures an old query, regardless of whether direct insertion classified that old query correctly, or an intruded calibration anchor.

The ranking loss averages

$$\operatorname{softplus}\left(q_\phi(x^-) - q_\phi(x^+)\right)$$

over all positive–negative pairs in the episode. Episodes with no positive or no negative candidate contribute zero loss (19 fitted training episodes). The matched BCE control uses the same candidates, with label 1 for correct new-class wins and 0 otherwise.

C.3 Optimization

CCPL is trained with Adam for 9,000 steps at learning rate 10^{-3}, using one episode per step. Episode order is determined by the run seed, and network initialization uses an independent seed. All training is performed on CPU in a single thread. The CCPL standardizer (§B.2) is fitted once on the development-training items and reused unchanged at evaluation. CAPE-BF instead fits its standardizer and logistic scorer separately within each episode using only that episode's legal fit data.

C.4 Evaluation semantics

At evaluation, CAPE-BF deploys its episode-local balanced arbiter whenever the fit succeeds and otherwise falls back to B2 pass-through. CCPL always uses its trained scorer under the same per-registration certificate. All reported evaluations use the episode-specific classwise SRS-WOR calibration procedure of Appendix B.4.

The main text reports two evaluation views. *Operational* results include all episodes. *Fitted-only* results include only episodes where CAPE-BF successfully fits: 672 of 900 development episodes and 291 of 300 tieredImageNet episodes. CAPE-BF fits on all 300 episodes of the prospective Aircraft evaluation (Appendix I).

New-class utility U_{new} is the fraction of new-label queries correctly classified after arbitration. Development macro statistics give equal weight to the six populations. We compute 95% paired bootstrap intervals with $B = 10{,}000$ by resampling episodes within each population, recomputing the six population means, and then averaging them.

For completeness, the bootstrap tail fractions are

$$\frac{1 + \#\{\text{nonpositive draws}\}}{B + 1},$$

equal to 9/10,001 for the operational macro contrast and 1/10,001 for the fitted-only macro contrast. We report these only as descriptive bootstrap tail fractions, not as p-values.

D Additional Development Results

Figure 3 shows the per-population operational and fitted-only improvements over CAPE-BF that underlie Table 1 of the main text and the held-out tieredImageNet comparison of Table 3.

Contract-matched scorer ladder. Table 6 places CAPE-BF and CCPL in a ladder of scorers under the identical calibration, activity-budget, threshold-selection, and certificate machinery (anchors, risk contract). The Fixed-Margin arm scores every candidate by the frozen evidence margin $E^{\max}_{\text{new}} - E^{\max}_{\text{old}}$ with no learning of any kind; equivalently, it applies a single shared scalar penalty to all new-class scores—the shared argmax compares $E^{\max}_{\text{new}} > E^{\max}_{\text{old}}$, so penalizing every new score by b reduces the decision to the margin exceeding b—with that scalar selected by the same per-registration certificate. The fixed-margin arm trails CAPE-BF by 6.46pp operationally, showing that the certificate alone does not account for the retained utility. CAPE-BF recovers much of the utility lost under fixed-margin scoring, while the cross-episode policy provides an additional positive gain. CCPL-Linear, a pooled linear instantiation, uses L2 logistic regression (C=1, lbfgs, 10,000 maximum iterations, no class weighting, float64) on the 53,335 pooled cross-episode training items after the same standardization — the identical cross-episode items and unweighted labels as the BCE control, with no tuning. It performs comparably to the MLP controls: +0.44pp operational (95% CI [0.15, 0.71]) and +1.64pp on the fitted held-out episodes. CCPL-Linear is a secondary linear instantiation, also used as a capacity diagnostic; the canonical MLP remains the primary CCPL instantiation. For completeness, the all-episode held-out value of CCPL-Linear is 81.10% U_{new} (+1.59pp versus deployed CAPE-BF). The matched ladder therefore suggests that the cross-episode scoring pipeline accounts for the remaining gain in this setting, while neither nonlinear capacity nor the particular MLP loss shows a consistent additional advantage. A representation-matched episode-local diagnostic (CAPE-BF-17D: the same arbiter given the full 17-D input) does not account for the full CCPL advantage; a positive residual remains under the matched 17-D input on the development populations (+0.91pp macro operationally, 95% CI [0.57, 1.25]; +1.44pp macro fitted-only) and on both held-out populations (Aircraft +5.03pp, 95% CI [4.02, 6.04]; tieredImageNet +1.98pp, 95% CI [1.46, 2.52] operationally over all 300 episodes, and +2.10pp, 95% CI [1.58, 2.63] on the 291 BF-fitted episodes); these diagnostic contrasts were computed under the earlier take. This is a pipeline-level diagnostic: representation dimensionality changes together with training sample size, normalization, and fitting structure, so the residual reflects the combined pipeline difference.

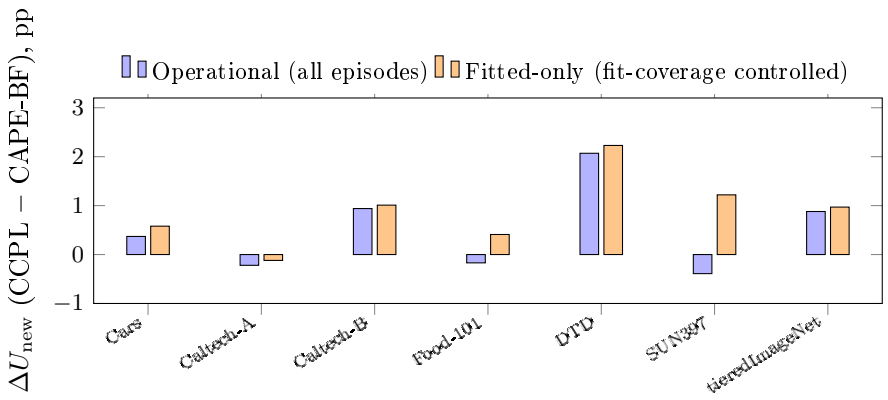


Figure 3: **Improvement over CAPE-BF by population.** Development macros: +0.43pp operational (95% CI [0.16, 0.71]) and +0.89pp fitted-only (+1.03pp pooled). Held-out tieredImageNet: +0.88pp (95% CI [0.44, 1.33]) operational and +0.97pp fitted-only.

Arm	Dev op. U_{new}	ΔU_{new} vs. BF	Fitted-only ΔU_{new}	Held-out fit. ΔU_{new}
Panel A: final SRS-WOR canonical comparisons				
B2 (uncertified ref.)	86.11%	–	–	–
CAPE-BF (canonical; 10-D)	82.79%	–	–	–
CCPL-MLP (canonical)	83.22%	+0.43	+0.89	+0.97
CCPL-Linear (linear inst.)	83.23%	+0.44	+1.00	+1.64
Panel B: earlier-take diagnostics				
CAPE-BF (earlier-take reference)	82.69%	–	–	–
Fixed-Margin (frozen)	76.23%	−6.46	−6.26	−6.84
CAPE-BF-17D (repr.-matched diag.)	82.32%	−0.37	−0.38	−1.23
CCPL-MLP-10D (repr.-matched diag.)	83.05%	+0.36	+0.97	+1.44
CCPL-Linear-10D (repr.-matched diag.)	83.21%	+0.52	+1.14	+1.82
BCE control	83.25%	+0.56	+1.15	+0.51

Table 6: **Contract-matched scorer ladder** (all arms are scored from the same frozen evidence features under identical calibration, activity-budget, threshold-selection, and certificate machinery; every ΔU_{new} is read against the BF reference row of its own panel). *Panel A* reports the final episode-specific classwise SRS-WOR draw: the canonical CAPE-BF (10-D passive block) reference, the uncertified direct-insertion reference B2 (take-independent; no calibration threshold or certificate), and the two cross-episode scorers (full 17-D input). *Panel B* retains the diagnostic rows computed under the earlier calibration take, against that take's CAPE-BF reference: the Fixed-Margin scalar-margin arm, the episode-local full-17-D representation-matched diagnostic (CAPE-BF-17D), the two CCPL-10D representation-matched diagnostics (passive block only), and the BCE-trained control. The panels are not compared row-to-row. Development columns: operational macro over 900 episodes; fitted-only is the six-population macro over the 672 episodes where CAPE-BF fits (pooled fitted-only values: Panel A, final SRS-WOR draw +1.03 (CCPL-MLP), +1.14 (CCPL-Linear); Panel B, earlier take −7.62, −0.49, +1.06, +1.25, +1.25 in row order). Held-out column: the 291 fitted tieredImageNet episodes. Dev op. U_{new} is the operational macro of the arm. CCPL-Linear trains on the identical cross-episode items as the BCE control (deterministic convex logistic fit, no tuning). The two CCPL-10D representation-matched rows retrain the cross-episode arms on dims 0–9 of the identical cross-episode items (fresh 10-D standardizer; single change).

Representation-by-pipeline controls. The final cell of the grid is the cross-episode pipeline given the 10-D passive input only. Its advantage persists there: CCPL-MLP-10D improves over deployed CAPE-BF by +0.36pp operationally (95% CI [0.06, 0.65]; +1.06pp pooled fitted-only) and CCPL-Linear-10D by +0.52pp (95% CI [0.22, 0.80]); the MLP-10D diagnostic also transfers positively to both held-out populations (tieredImageNet +1.39pp operational, +1.44pp fitted-only; Aircraft +2.77pp). Within the cross-episode pipeline the 17-D versus 10-D difference is small (+0.18pp MLP, +0.05pp linear, operational), while the

corresponding episode-local contrast is slightly negative (−0.37pp, BF-17D versus BF-10D). The interaction contrast $(\mathrm{CCPL}_{17}-\mathrm{CCPL}_{10})-(\mathrm{BF}_{17}-\mathrm{BF}_{10})$ is +0.55pp (95% CI $[0.28, 0.84]$; linear-control version +0.42pp, 95% CI $[0.16, 0.68]$), indicating that the effect of representation depends on the scoring pipeline. We treat these post-hoc contrasts as descriptive. The deployed methods remain CAPE-BF-10D and CCPL-MLP-17D, with the same certification rule (both CCPL-10D arms certified in 1500/1500 evaluated episodes). Some post-hoc cross-episode variants attain higher held-out utility than the canonical CCPL-MLP, but all headline results continue to use the CCPL-MLP configuration fixed before evaluation.

LP++-NEW-only comparator identity. The external scores enter the frozen OLD∪NEW argmax through a support-fitted positive affine bridge $\tilde{s}_j(x) = a\, s_j(x) + b$ with $a > 0$, identical for all new classes and fitted per episode by ordinary least squares from new-class support rows only: S collects the method's raw scores for the 25 support images against the 5 new classes, and E contains the corresponding frozen B2 new-class evidence under strict leave-one-out semantics ($a = \sum(s-\bar{s})(e-\bar{e}) / \sum(s-\bar{s})^2$, $b = \bar{e} - a\bar{s}$). The bridge is used only when the fitted parameters are finite, the support-score variance is positive, and $a > 0$; these conditions hold in all 900 episodes. A positive affine map preserves the method's NEW-vs-NEW ranking exactly. The certificate scalar of the comparator is the evidence margin of the joined score matrix: the selection rule of Algorithm 1 is applied unchanged, with the comparator's own conflict scalar in place of a learned scorer. For a row x, let $E(x)$ collect the frozen B2 OLD evidence in the OLD columns and the bridged scores in the NEW columns; the joint winner is $w(x) = \arg\max_k E_k(x)$ and the conflict margin is

$$m(x) \;=\; \max_{j \in \mathrm{NEW}} \tilde{s}_j(x) \;-\; \max_{i \in \mathrm{OLD}} E_i^{\mathrm{B2}}(x). \tag{7}$$

A candidate new-class win is $w(x) \in$ NEW; with the implementation's OLD-first deterministic tie rule (OLD columns precede NEW columns, and ties go to the first maximum), $w(x) \in$ NEW iff $m(x) > 0$. The same margin is evaluated on every old-class calibration anchor; anchors with $w(a) \in$ NEW (candidate intrusions) supply the violation grid $g = \{m(a)\}$. With n anchors and $k_{\max}(n)$ the largest k whose exact Clopper–Pearson 95% upper bound is at most 2%, the certificate-only arm selects the threshold just below the $k_{\max}(n)$-th highest violation margin, so at most $k_{\max}(n)$ intrusions remain retained (ties handled conservatively: the retained count is reduced until it is certifiable; $-\infty$ if every count is certifiable), and the deployed arm uses the minimum-activity budget $k^\star = \min(k_{\max}(n),\, k_{B2} - 1)$ with $k_{B2} = |g|$ (if $k_{B2} = 0$, the certificate-only selection applies). A candidate new-class win is retained iff $m(x) > \tau$; every other query falls back to the OLD argmax of E^{B2}. The same margin is used for query candidates and calibration anchors. The main-text comparator (Table 2) evaluates this fixed comparator specification under the same realized final SRS-WOR calibration anchors as the CAPE-family evaluation on each episode (same derivation as Appendix B.4); the B2 reference is anchor-free and take-independent.

LP++ loss decomposition. Table 7 decomposes the draw-matched LP++ utility gap into three stages: bridged scores inserted without arbitration (pre-certification), the certificate-only rule, and the deployed minimum-activity rule. Of the 24.67pp macro gap, 4.45pp arise before certification (score-and-bridge insertion cost), 18.98pp at the certification stage, and 1.24pp from the minimum-activity rule: under the bridged LP++ scores, satisfying the same certificate requires a substantially more conservative operating point (per-class Q25 −41.67pp pooled). Certification alone reduces pooled harm05 from 64.4% to 36.6% and raises macro ΔOLD from −0.80pp to +0.81pp (Caltech-A and SUN397 remain at −0.10pp); minimum activity further reduces harm05 to 30.0% and leaves ΔOLD non-negative on all six populations (+0.89pp macro). The operating point thus becomes safer and more conservative at the same time; CAPE-family scorers reach a less conservative certified operating point under the same machinery. The stage decomposition localizes where the utility loss occurs but does not causally isolate the bridge from subsequent certification behavior.

BiMC comparator. BiMC is included as a post-hoc external comparator evaluated under the same 900-episode certification protocol (Chen et al., 2025): the scorer replaces the

Stage	Population	U_{new}	ΔU_{new}	Q25	ΔOLD	harm05
B2 (reference)	macro	86.11	–	–	–	0.573
Pre-certification	Cars	83.37	−5.17	−7.69	−0.30	0.740
	Caltech-A	98.50	−0.18	0.00	−0.25	0.327
	Caltech-B	63.60	−9.99	−25.00	−0.24	0.820
	Food-101	93.51	−2.09	−5.00	−0.00	0.393
	DTD	55.73	−7.81	−25.00	−3.58	0.993
	SUN397	95.24	−1.46	0.00	−0.43	0.593
	Macro	**81.66**	**−4.45**	**−8.33**	**−0.80**	**0.644**
Certificate-only	Cars	71.87	−16.68	−30.77	+0.96	0.360
	Caltech-A	59.02	−39.66	−100.00	−0.10	0.233
	Caltech-B	36.34	−37.24	−66.67	+2.52	0.247
	Food-101	93.05	−2.56	−5.00	+0.10	0.307
	DTD	24.96	−38.59	−58.33	+1.51	0.593
	SUN397	90.82	−5.88	−8.33	−0.10	0.453
	Macro	**62.68**	**−23.43**	**−41.67**	**+0.81**	**0.366**
Deployed (min. activity)	Cars	71.11	−17.44	−30.77	+0.98	0.327
	Caltech-A	58.20	−40.48	−100.00	+0.01	0.127
	Caltech-B	35.08	−38.51	−66.67	+2.53	0.240
	Food-101	91.52	−4.09	−5.00	+0.21	0.187
	DTD	24.96	−38.59	−58.33	+1.51	0.593
	SUN397	87.76	−8.94	−8.33	+0.12	0.327
	Macro	**61.44**	**−24.67**	**−41.67**	**+0.89**	**0.300**

Table 7: **LP++-NEW-only loss decomposition under identical final SRS-WOR anchors** (same 900 development episodes as Table 2). $U_{\text{new}}/\Delta U_{\text{new}}/$Q25$/\Delta$OLD in percent or pp (Q25: pooled per-class ΔU_{new} entries; ΔU_{new}: pooled paired difference against B2; macro 95% CI for the deployed ΔU_{new} $[-26.14, -23.21]$); harm05 is the fraction of episodes whose old-class accuracy falls more than 0.5pp below the old-only baseline.

learned CAPE conflict scorer natively and scores the full OLD∪NEW evidence matrix (no bridge, no B2-column substitution; calibration weights λ_t=0.5, λ_i=0.1, temperature 16), and the unchanged certification machinery of §7.2 is applied to those scores. A rejected new-class win falls back to the argmax over BiMC's own old-class scores (its evidence matrix supplies both the OLD and NEW columns); as for every arm, harm05 and ΔOLD are measured against the same frozen B2 old-only baseline. A stage decomposition (Table 8) localizes the deployed −14.26pp utility gap: native BiMC insertion without arbitration attains +6.30pp over direct insertion, the certificate accounts for a −19.10pp swing from that endpoint (per-class Q25 −16.67pp at the deployed point), and minimum activity adds a further −1.46pp. As with the bridged LP++ scores, the external scorer is admitted only at a far more conservative certified operating point than its native scores would require.

Stage	U_{new}	ΔU_{new}	ΔOLD	harm05
B2 (reference)	86.11	–	–	57.3%
BiMC pre-certification (native)	92.41	+6.30	+5.96	23.6%
Certificate-only	73.31	−12.80	+6.49	16.9%
Deployed (min. activity)	71.85	−14.26	+6.59	13.6%

Table 8: **BiMC loss decomposition under identical final SRS-WOR anchors** (same 900 development episodes as Table 2; conventions as Table 7: U_{new} in %, $\Delta U_{\text{new}}/\Delta$OLD in pp versus B2; harm05 is the fraction of episodes whose old-class accuracy falls more than 0.5pp below the old-only baseline).

Its per-population scale parameter β is selected by the method's own query-blind rule (grid 0.05:0.05:0.95, first 20 episodes per population, objective = mean OLD-memory leave-one-exemplar-out accuracy, ties to the smaller β): 0.70 (Cars), 0.95 (Caltech-A), 0.65 (Caltech-B), 0.90 (Food-101), 0.65 (DTD), 0.90 (SUN397). For DTD, a sealed class-description file was unavailable, so the 15 descriptions per class were reconstructed from class names

using the same frozen text encoder. The population-specific β_{DTD} was then selected by the same query-blind OLD-memory LOO rule. The certificate is valid on all 900 episodes. We also report the prespecified fixed-$\beta = 0.85$ sensitivity variant (Table 9): the primary configuration loses 14.26pp pooled with harm05 13.6%, and the fixed-β variant loses 15.31pp with harm05 17.9%. The DTD and Caltech-B rows illustrate the method's operating point: near-zero OLD→NEW routing keeps old-side harm at 1.3% and 0.0% where direct insertion harms 84.0% and 86.0% of episodes, at a large utility cost. On the utility axis the paired CCPL−BiMC difference is +11.37pp (95% CI [+9.10, +13.68]); on the old-side axis, CCPL-MLP's development harm05 is 20.8% against BiMC's 13.6% (Figure 1C), so neither arm dominates the other on the two empirical axes.

Population	U_{new}	ΔU_{new}	ΔOLD	OLD→NEW	harm05
Cars	85.01	−3.54	+2.05	0.65%	0.280
Caltech-A	46.41	−52.27	+0.36	0.03%	0.120
Caltech-B	67.77	−5.82	+27.72	0.03%	0.000
Food-101	94.95	−0.66	+0.49	0.50%	0.193
DTD	51.33	−12.21	+8.08	1.13%	0.013
SUN397	85.62	−11.08	+0.83	0.16%	0.207
Pooled	**71.85**	**−14.26**	**+6.59**	**0.42%**	**0.136**

Table 9: **BiMC post-hoc external comparator on the current six-population protocol** (primary variant; same 900 development episodes as Table 2, identical final SRS-WOR anchor identities). U_{new} in %, ΔU_{new}/ΔOLD in pp versus direct insertion (B2) on the same episodes; harm05 is the fraction of episodes whose old-class accuracy falls more than 0.5pp below the old-only baseline; OLD→NEW is the retained-candidate old-query routing rate. Pooled ΔU_{new} 95% CI [−16.54, −12.07]. Secondary variant ($\beta = 0.85$): pooled U_{new} 70.80, ΔU_{new} −15.31 [−17.55, −13.14], harm05 17.9%.

Minimum-activity ablation. The deployed threshold rule adds a minimum-activity objective to the certificate: the admitted-violation budget is $k^\star = k_{\max}(n)$ when direct insertion routes no anchor, and $\min(k_{\max}(n), k_{B2} - 1)$ otherwise, forcing at least one anchor intervention whenever the certificate alone would permit inaction. A certificate-only ablation variant with k-target $k_{\max}(n)$ (no forced intervention; both rules certified) raises development U_{new} from 83.22% to 83.89% for CCPL-MLP (+0.67pp macro; per-population effects concentrated in Food-101 +1.17 and SUN397 +2.09, with no change on Caltech-B and DTD) and from 82.79% to 83.31% for operational CAPE-BF (+0.52pp); the CCPL − CAPE-BF difference is +0.58pp under the certificate-only rule versus +0.43pp deployed (descriptive post-hoc 95% paired episode bootstrap interval [0.38, 0.79]pp). The minimum-activity objective therefore accounts for 0.67pp of CCPL-MLP's 2.89pp deployed utility cost relative to direct insertion; the remainder is the certified operating point itself, not the forced-intervention objective.

Population	CAPE-BF	CCPL-MLP	CCPL-Linear	LP++-NEW-only	BiMC
Cars	−1.64	−1.27	−1.06	−17.44	−3.54
Caltech-A	−0.27	−0.49	−0.40	−40.48	−52.27
Caltech-B	−1.17	−0.22	−0.37	−38.51	−5.82
Food-101	−1.15	−1.31	−1.29	−4.09	−0.66
DTD	−13.74	−11.68	−11.33	−38.59	−12.21
SUN397	−1.98	−2.37	−2.86	−8.94	−11.08
Macro	−3.32	−2.89	−2.88	−24.67	−14.26

Table 10: **Per-population contemporaneous utility differences against direct insertion** (ΔU_{new} in pp; the same 900 development episodes as Table 2). Macro rows carry 95% paired episode bootstrap intervals in Table 2; per-population 95% paired-bootstrap intervals are reported in Table 11. Population-level retention ($\Delta U_{\mathrm{new}} > -5$pp): five of six for each CAPE-family arm (all but DTD); one of six for LP++-NEW-only (Food-101 only); two of six for BiMC (Cars and Food-101).

Distribution of paired improvements. Beyond the macro means, the pooled empirical distribution of the 900 paired per-episode development differences (CCPL − CAPE-BF) is: minimum −38.00pp, maximum +31.67pp, mean +0.43pp; Q25 = median = 0.00pp (55.4% of episodes are exactly unchanged), Q75 +1.00pp; 26.1% of episodes improve and 18.4% degrade. Held-out tieredImageNet: mean +0.88pp, Q25 = median = 0.00pp (42.7% unchanged), Q75 +1.67pp. Because many episodes are unchanged, the positive mean difference is driven by the subset with nonzero paired changes. Per-population 95% paired-bootstrap intervals appear in Table 11.

Population	ΔU_{new} (CCPL − CAPE-BF)	95% CI
Cars	+0.37	[−0.11, 0.89]
Caltech-A	−0.22	[−0.49, 0.03]
Caltech-B	+0.94	[0.60, 1.34]
Food-101	−0.17	[−0.89, 0.43]
DTD	+2.07	[0.99, 3.14]
SUN397	−0.39	[−1.19, 0.40]

Table 11: **Per-population operational differences** (CCPL − CAPE-BF, development, 150 episodes each). Paired episode bootstrap (B=10,000), same conventions as the main text. Two of six intervals exclude zero (Caltech-B and DTD); the Cars, Caltech-A, Food-101, and SUN397 intervals include zero.

E Held-Out tieredImageNet Diagnostics

E.1 Old-side query endpoints (descriptive validation)

The certificate controls the calibration-frame estimand (§4, Appendix A); this section asks how that control behaves on the held-out old-query population itself. For each of the four evaluated arms (B2, CAPE-BF, CCPL, and the BCE control), we record which old queries the registered rule finally routes to a new class — the query-side event named in the risk contract (§3). The *empirical query-side old-to-new violation rate* of an arm divides its routed-query count by the number of old queries recorded on the 291 fitted episodes (48,090 old queries). The nine no-fit episodes are excluded from this diagnostic because their query records were not retained. The rates below are computed under the earlier calibration draw (Appendix B.4). These query-side rates describe held-out behavior; the formal guarantee remains the calibration-frame certificate of Appendix A.

Arm	Empirical OLD→NEW violation rate	Routed / old queries
B2 (direct insertion)	2.793%	1,343 / 48,090
CAPE-BF (deployed)	0.925%	445 / 48,090
CCPL (ours)	0.913%	439 / 48,090
BCE control	0.869%	418 / 48,090

Three observations. First, direct insertion exceeds the 2% target on the held-out query population, whereas CAPE-BF and CCPL realize rates near 0.93%. Thus, in this tieredImageNet diagnostic, the held-out query-side rates remain below the calibration-frame risk target, although they are not covered by the formal certificate. Second, CCPL closely matches CAPE-BF's realized old-side violation rate on this fitted-291 diagnostic, so its utility gain is not bought with old-side risk. Third, restricting to the subset of retained captures that flip a *correct* old prediction (a secondary diagnostic), the ordering is the same: B2 1.799%, CAPE-BF 0.466%, CCPL 0.405%, BCE 0.391% — the certified policies reduce realized old-side harm as well. These are descriptive held-out measurements and do not extend the formal certificate beyond the calibration frame.

E.2 No-fit decomposition

On the nine of 300 held-out episodes where CAPE-BF cannot fit its episode-local arbiter, it falls back to direct insertion while CCPL still runs its certified scorer. CCPL trails the

B2 fallback by 2.22pp on average on this subset, so the no-fit episodes reduce rather than create the overall +0.88pp held-out advantage. On the remaining 291 episodes, the fitted-only advantage is +0.97pp.

E.3 BCE control

The BCE-trained control arm on the same pairs reaches $U_{\text{new}} = 80.08\%$ on tieredImageNet (CCPL: 80.39%; BF: 79.61%; earlier-take values). On the development populations the control is competitive (operational macro 83.25% vs. 83.23%; fitted-only 79.64% vs. 79.56%). On tieredImageNet, the pairwise objective is 0.31pp higher than the BCE control (paired-bootstrap 95% CI $[0.07, 0.57]$). Across five independent training seeds, however, neither objective shows a consistent advantage over the other (Appendix E.5); the rank-versus-BCE difference is small relative to training-seed variability. This comparison indicates that most of the gain comes from learning a transferable cross-episode conflict scorer rather than from the particular training loss.

E.4 Certificate registrations

CCPL and the BCE control are evaluated as certified registrations on all 300 held-out episodes, yielding 600 learned-arm registrations in total; every registration satisfies $\text{CP}_u^{95} \leq 2\%$. CAPE-BF fits and registers on 291 of 300 episodes; the remaining nine deploy B2 pass-through and assert no certificate. All episodes were certifiable, with $k_{\max}(n) \geq 0$ throughout.

Arm	Episodes	Scorer fit	Certified	Uncertified fallback
CCPL (ours)	300	cross-episode (always)	300	0
BCE control	300	cross-episode (always)	300	0
CAPE-BF	300	episode-local	291	9 (B2 pass-through)

E.5 Training-seed robustness

Both learned objectives were retrained with five independent seeds, including the canonical run. Rank and BCE use matched seeds within each replicate and share the same architecture, optimizer, training budget, and data split. Training is deterministic conditional on the seed. These runs were evaluated under the earlier calibration take and were not repeated under the final SRS-WOR draw, so this experiment isolates training-seed variability rather than calibration-draw sensitivity.

Table 12 summarizes the results. The cross-episode advantage is consistent across seeds: Rank and BCE each outperform CAPE-BF in all five replicates on development (operational and fitted-only) and on held-out tieredImageNet. The held-out advantages are $+0.76 \pm 0.16$pp for Rank and $+0.89 \pm 0.26$pp for BCE.

Rank versus BCE varies across seeds, with no consistent advantage for either objective (development: 3/5 positive; held-out: 1/5). The result shared by both objectives is therefore the advantage of cross-episode scoring over CAPE-BF rather than a consistent benefit of one training loss. All replicate evaluations satisfy the corresponding per-episode certificate under the earlier calibration take.

Under that take, the canonical Rank replicate lies at the lower end of the five-seed development operational range (+0.54pp to +0.77pp). Under the final SRS-WOR draw, its canonical operational macro is +0.43pp (§7.2).

Contrast (pp)	Mean	SD	Range	Positive
Dev operational Rank − BF	+0.69	0.11	[+0.54, +0.77]	5/5
Dev fitted-only Rank − BF	+1.24	0.13	[+1.05, +1.34]	5/5
Dev operational BCE − BF	+0.64	0.07	[+0.56, +0.75]	5/5
Held-out fitted Rank − BF	+0.76	0.16	[+0.57, +0.89]	5/5
Held-out fitted BCE − BF	+0.89	0.26	[+0.51, +1.24]	5/5
Dev Rank − BCE	+0.04	0.08	[−0.04, +0.14]	3/5
Held-out fitted Rank − BCE	−0.14	0.36	[−0.62, +0.36]	1/5

Table 12: **Training-seed robustness** (five independently trained policies per objective; the canonical policy is one of the five; matched randomness within each replicate). Held-out rows are the fitted-only estimand on the 291 cached tieredImageNet episodes; the canonical replicate has an earlier-take fitted-only value of +0.86pp (Rank − BF). Development operational rows use all 900 evaluation episodes; the development fitted-only row is the six-population macro over the 672 episodes where CAPE-BF fits (pooled fitted-only: $+1.36 \pm 0.14$, range $[+1.17, +1.48]$).

F Reproducibility Details

Data. Each development population (Cars, Caltech-A, Caltech-B, Food-101, DTD, and SUN397) contains 300 episodic manifests. The tieredImageNet episodes were constructed once after the method and evaluation protocol were finalized, following Appendix C. The class universe, per-class partitions, and episode construction are deterministic functions of the documented seeds.

Seeds and determinism. Randomness is controlled by documented seeds for initialization, episode order, evaluation, and episode construction, with separate seed namespaces for distinct stages. Training runs on CPU with a single torch thread. Policy training uses exactly 9,000 steps, with no early stopping or per-run model selection.

Statistics. Means are reported with 95% paired bootstrap confidence intervals using $B = 10{,}000$ draws from a fixed PCG64 generator. Episodes are resampled within each population, the corresponding population means are recomputed, and the six development-population means are then averaged with equal weight.

These intervals condition on the realized calibration take and quantify episode-resampling uncertainty; sensitivity to the calibration take is reported separately in Appendix B.4. Bootstrap tail fractions are computed as

$$\frac{1 + \#\{\text{nonpositive draws}\}}{B + 1},$$

giving 9/10,001 for the operational macro contrast and 1/10,001 for the fitted-only macro contrast. We report these as descriptive tail fractions rather than p-values. Confidence intervals for secondary and descriptive analyses are pointwise rather than simultaneous. The middle-quartile statistics in Appendix D use the same paired per-episode differences.

Old-side replay. The descriptive old-side endpoints in Appendix E.1 are obtained by replaying the saved evaluation records and extracting retained-candidate old-query flips from the recorded conflict outputs. The replay requires no retraining. Per-episode utility is recoverable for all 300 tieredImageNet episodes; the query-side diagnostic itself uses the 291 episodes with retained old-query records, as described in Appendix E.1.

Aggregation. Reported tables are generated from the saved per-episode records using the fixed comparator specifications and evaluation code. Both prespecified BiMC β variants and all three LP++ stages in Table 7 are included, and the same CAPE-BF implementation is used throughout the development, held-out, and prospective evaluations. The supplementary materials include the scripts and metadata needed to reproduce these aggregations.

Artifacts. The supplementary package includes per-episode evaluation records, aggregated results, scorer weights, standardizer parameters, and episode-construction manifests. It also contains self-contained Python code for the certificate, episode construction, and evaluation pipeline, including exact binomial inversion, the minimum-activity budget, the documented seed namespaces, the scorer weights, the evaluation replay, the risk-level sweep of Appendix I.6, and a CPU smoke test on a synthetic episode. The full multi-gigabyte feature shards are not redistributed.

G Mechanism Evidence: Full Tables

Table 13 provides the full results summarized in Fig. 4. It reports the frontier decomposition across populations and the conflict-representation comparison, including the Fisher-subspace control and the separate-episode evaluation of DINOv3 evidence.

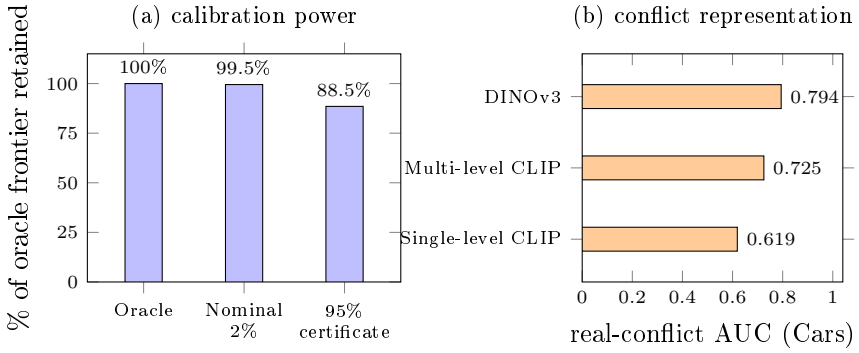


Figure 4: **Two factors shaping certified registration utility.** (a) *Calibration power.* On Cars under the final 530-anchor protocol, the nominal 2% constraint retains 99.5% of the risk-unconstrained old-repair frontier; the 95% exact certificate retains 88.5% (22/250 fitted episodes infeasible at the utility requirement). (b) *Conflict representation.* Real-conflict AUC improves from 0.619 with single-level CLIP evidence to 0.725 with multi-level CLIP and 0.794 with DINOv3-based evidence.

(a) Frontier decomposition

Population	Nominal 2%	95% certificate	Infeasible
Cars	99.5%	88.5%	22/250
Caltech-B	–	repair 2.70 → 2.31pp	32/278
Food-101	–	not binding	1/236

(b) Conflict-evidence representation (Cars real-conflict AUC)

Representation	AUC	Note
Single-level CLIP (incumbent)	0.619	episode-local
Multi-level CLIP	0.725	+10.6pp over single-level CLIP
DINOv3 evidence	0.794	representation-development split
Fisher-subspace projection (active)	−10.2pp	paired control

Table 13: **Mechanism evidence.** (a) Under the final water-filled calibration protocol (530-row target), the finite-sample 95% certificate retains most of the available old-side repair on Cars relative to the nominal 2% constraint. The remaining cost varies with calibration availability, with essentially no loss in the Food-101 census regime. (b) Conflict-evidence quality affects the separation of safe and unsafe candidates. DINOv3 evidence reaches 0.827 on Cars and 0.973 on Caltech-B on evaluation splits disjoint from the representation-development split, while the actively learned Fisher-subspace projection is 10.2pp worse in the paired comparison.

H Edge Deployment Profile

We profile the arbitration layer on a Jetson Nano Developer Kit (4 GB, JetPack R32.7.1, MAXN power mode) to measure the deployment cost of episode-local fitting. All arms are evaluated in the same post-feature harness on the tieredImageNet held-out workload. We use five timing repetitions over the 16,760 conflict-candidate rows, with 100 warm-up calls before timing each arm. Thresholds are computed on device using the same certification logic as in the main evaluation. Because the frozen encoder was unavailable in the Nano software stack, these measurements isolate the post-feature registration layer rather than end-to-end encoder latency.

Removing episode-local fitting reduces median registration latency from 8.02 ms for CAPE-BF to 0.72 ms for CCPL-MLP and 0.17 ms for CCPL-Linear. The serialized CCPL-Linear policy occupies 634 bytes (18 parameters), compared with 46,732 bytes (5,377 parameters) for CCPL-MLP (Table 14). CAPE-BF fitting on the Nano uses a NumPy Newton implementation because scikit-learn is unavailable in the device environment. Device-side keep-or-fallback decisions matched the host implementation for all evaluated rows, so the deployment profile changes only latency and model size, not predictive utility.

Arbiter	Episode-local fit	Held-out U_{new}	Reg. med. (ms)	Reg. p95 (ms)	Scorer p95 (μs)	Scorer params	Artifact bytes
CAPE-BF	yes	79.61%	8.02	18.30	–	–	–
CCPL-Linear	no	81.24%	0.17	0.21	24.5	18	634
CCPL-MLP	no	80.39%	0.72	1.30	122.6	5377	46732

Table 14: **Edge deployment profile on Jetson Nano.** All methods use the same post-feature workload and held-out episodes. The held-out U_{new} values correspond to the earlier calibration draw used for deployment profiling; the CCPL-MLP value rounds to the same value as the final-draw result in Table 3 (calibration-take sensitivity in Table 5). Hardware measurements are collected on device. CAPE-BF latency uses the NumPy Newton fitting implementation; registration latency is reported by its median and 95th percentile.

I Generalization Beyond the Canonical Setting

We test CCPL beyond the canonical development setting along four axes: a new held-out population, larger old-class universes, repeated registration, and alternative feature representations. The cross-episode advantage remains positive in aggregate across these perturbations, while its magnitude and preferred scorer vary with the representation; Table 15 summarizes the results. The first three controls were prospectively specified; the representation-family analyses are descriptive robustness tests. All controls use the features, calibration frame, activity budget, and certificate machinery of the main text—CAPE-BF scoring the 10-D passive block, CCPL the full 17-D representation—with deployed CAPE-BF as the comparator; utility is U_{new} and all intervals are 95% paired bootstrap confidence intervals. The population-transfer control uses the final SRS-WOR calibration draw; the class-universe, repeated-registration, alternative-representation, and risk-level analyses were computed under the earlier calibration take, for which Appendix B.4 reports the take-sensitivity comparison.

I.1 Generalization matrix

Table 15 summarizes these controls, with full results reported in §I.2–§I.5. The population, class-universe, and repeated-registration contrasts are the prospectively specified primaries; the Aircraft Linear contrast is a secondary, diagnostic confirmation, and the representation-family contrasts are descriptive robustness controls.

Axis	Test	CCPL-MLP − BF		CCPL-Linear − BF	
Population	FGVC Aircraft	**+3.27**	[2.56, 3.99]	+3.56	[2.88, 4.26]
OLD-class universe	50–80 OLD classes	**+2.19**	[1.56, 2.86]	+2.48	[1.91, 3.08]
Registration regime	4 sequential waves	**+3.02**	[2.54, 3.53]	+3.13	[2.58, 3.69]
Representation	CLIP-17D (frozen control)	+2.47	[1.77, 3.20]	+3.51	[2.87, 4.16]

Table 15: **External-validity matrix.** ΔU_{new} in percentage points (95% paired bootstrap CI) for the two frozen cross-episode scorers against deployed CAPE-BF along four external-validity axes; the population, class-universe, and registration-regime rows are prospectively specified controls evaluated on FGVC Aircraft episodes, and the representation row is the frozen CLIP-17D control (Aircraft episodes; §I.5). Bold marks the canonical CCPL-MLP contrasts. The ResNet-50 construction is reported separately in §I.5 because its summary estimand is the six-population development macro rather than the Aircraft contrast.

I.2 Population transfer: FGVC Aircraft

On 300 held-out FGVC Aircraft (Maji et al., 2013) episodes, which were excluded from all training and development decisions and drawn with the same episode-construction protocol as the development populations over the 100 aircraft variants as classes—CCPL-MLP improves U_{new} over CAPE-BF by +3.27pp (95% CI $[2.56, 3.99]$) and CCPL-Linear by +3.56pp (95% CI $[2.88, 4.26]$; a secondary, diagnostic contrast); calibration-take sensitivity is reported in Appendix B.4. CAPE-BF fits on all 300 episodes, so the operational and fitted-only comparisons coincide. Under the prespecified non-inferiority margin of −0.5pp, CCPL-Linear is non-inferior to the canonical CCPL-MLP (+0.29pp, 95% CI $[-0.21, 0.78]$). Every registration of every certified arm satisfies the certificate; Table 16 reports all five arms. Direct insertion routes 7.54% of held-out old-class queries to a new prediction and degrades old-class accuracy by more than 0.5pp in 96.3% of episodes; certified arbitration reduces these to 0.68–0.93% and 14.7–26.0% respectively. These results provide prospective evidence that the cross-episode scorer transfers to a second held-out population.

	B2	Fixed-Margin	CAPE-BF	CCPL-MLP	CCPL-Linear
U_{new} (%)	52.09	27.84	31.77	35.04	35.33
OLD→NEW rate (%)	7.54	0.68	0.82	0.92	0.93
OLD-acc. drop >0.5pp (% eps.)	96.3	26.0	15.0	14.7	15.3
Certificates ($\text{CP}_u^{95} \leq 2\%$)	—	300/300	300/300	300/300	300/300

Table 16: **FGVC Aircraft, prospective evaluation** (300 held-out episodes; canonical values use the final SRS-WOR calibration procedure; calibration-take sensitivity is reported in Appendix B.4). B2 is the uncertified direct-insertion reference: it has no calibration threshold and asserts no certificate ("—"). Old-side rows are descriptive.

I.3 Class-universe expansion

The development episodes of the main text cycle the old-class count over $\{20, 25, 30, 35\}$. This control expands the old-class universe to 50, 60, 70, or 80 old classes (75 episodes per stratum, 300 in total). CAPE-BF fits on 299/300 episodes—the single non-fit episode lies in the 50-class stratum and deploys the frozen no-fit fallback (direct insertion), exactly as in the operational scope of the main text. Aggregate U_{new} is 39.03% (B2), 24.14% (CAPE-BF), 26.34% (CCPL-MLP), and 26.63% (CCPL-Linear). The paired advantage persists: CCPL-MLP − CAPE-BF = +2.19pp (95% CI $[1.56, 2.86]$) and CCPL-Linear − CAPE-BF = +2.48pp (95% CI $[1.91, 3.08]$), and it is positive in every class-count stratum (Table 17).

I.4 Repeated cumulative re-registration

The experiments above register one set of new classes per episode. This control repeats registration: 100 frozen sequences each start from 20 old classes and perform four successive five-class insertion waves (40 classes in the final classifier), with per-wave 5-shot support

OLD classes	50	60	70	80
CCPL-MLP − BF	+1.98 [0.60, 3.36]	+2.13 [1.00, 3.24]	+2.29 [1.18, 3.44]	+2.38 [0.93, 3.93]
CCPL-Linear − BF	+2.87 [1.53, 4.16]	+2.02 [1.04, 2.98]	+2.09 [1.00, 3.16]	+2.96 [1.71, 4.36]

Table 17: **Class-universe expansion.** ΔU_{new} (pp, 95% CI) per old-class-count stratum, operational scope, 75 episodes per stratum; every stratum is positive. On the fitted-only scope the 50-class stratum (the one containing CAPE-BF's single non-fit episode) reads +2.03 [0.63, 3.42] (MLP) and +2.95 [1.64, 4.28] (Linear); the other strata are unchanged.

sets, calibration anchors, and queries. The old side at each wave is cumulative—it contains all previously registered classes—and each wave is a complete re-registration: CAPE-BF refits its local scorer, CCPL reuses its frozen cross-episode policy, and the certificate is recalibrated over the cumulative old set with the anchor construction of §B.4; class identity is the only state carried between waves. Each wave-t certificate covers the cumulative old side versus the current wave's five new classes; inference is at the sequence level (paired bootstrap over the 100 sequences, B=10,000).

The advantage persists: CCPL-MLP − CAPE-BF = +3.02pp (95% CI [2.54, 3.53]) and CCPL-Linear − CAPE-BF = +3.13pp (95% CI [2.58, 3.69]). The per-wave advantage is positive at every wave (CCPL-Linear − CAPE-BF wave means: +3.10, +2.37, +3.40, +3.63pp), with no apparent attenuation by the fourth wave (wave 4: 32.43% vs. 28.80%). CAPE-BF fits on 398/400 waves; the two non-fit waves deploy the frozen no-fit fallback. Certified-wave fractions are 99.5% (CAPE-BF) and 100% (CCPL-MLP, CCPL-Linear). Descriptively, final old-class accuracy after four waves is 53.0% for the arbitration arms versus 50.93% under direct insertion, and the cumulative U_{new} loss relative to direct insertion is 18.88pp for CAPE-BF versus 15.86pp (CCPL-MLP) and 15.76pp (CCPL-Linear). The certificate applies separately at each insertion wave; cumulative trajectory-level safety is outside the present analysis.

I.5 Alternative frozen representations

We replace the canonical DINOv3-centered conflict representation with a CLIP-derived 17-D construction—a passive 10-D block built from multi-level CLIP ViT-B/16 features under the same formulas, plus the unchanged seven support-conditioned evidence dims—retraining the cross-episode policies on this representation and re-fitting the episode-local CAPE-BF arm per episode on the same features, while keeping the registration protocol, calibration frame, activity budget, and certificate machinery unchanged. The cross-episode advantage remains under the alternative representation, but the relative ordering of the two scorers changes: on Aircraft the linear policy exceeds the MLP by +1.04pp (95% CI [0.52, 1.56]), so the effect is not tied to nonlinear scorer capacity. On tieredImageNet both alternative constructions show the same qualitative pattern—the MLP interval includes zero, whereas the linear contrast remains positive with an interval excluding zero—and the mechanism behind this repeated representation-dependent ordering is not identified by the present experiments. Caltech-A shows a negative effect under both the canonical and CLIP-derived representations and a near-zero effect under ResNet-50, indicating that this population-level boundary is itself representation-dependent.

On the development populations the cross-episode policies hold a positive macro advantage under the alternative representation (MLP +0.80pp, five of six populations positive; Linear +0.32pp, four of six), with per-population effects heterogeneous exactly as on the canonical representation. On the held-out tieredImageNet axis the MLP advantage crosses zero (+0.31pp, 95% CI [−0.50, 1.13]) while the linear contrast is +3.98pp (95% CI [3.28, 4.70]).

A third construction replaces the passive conflict-evidence extractor with a supervised ResNet-50 (IMAGENET1K_V2), retaining the seven support-conditioned dimensions and the registration, calibration, certificate, and training machinery. Across the six development populations the equal-weight operational macro advantage over CAPE-BF is +4.62pp (95% CI [3.94, 5.33]) for CCPL-MLP and +4.30pp (95% CI [3.62, 5.02]) for CCPL-Linear, with all six MLP population-level point estimates positive. Post-hoc Aircraft contrasts are

Population	CCPL-MLP − BF		CCPL-Linear − BF	
Cars	+1.24	[0.41, 2.11]	+1.08	[0.35, 1.84]
Caltech-A	−0.58	[−1.06, −0.12]	−0.63	[−1.11, −0.18]
Caltech-B	+2.50	[1.73, 3.38]	+1.36	[0.71, 2.07]
Food-101	+0.27	[−0.43, 0.87]	+0.16	[−0.55, 0.78]
DTD	+0.89	[0.02, 1.74]	+0.62	[−0.11, 1.37]
SUN397	+0.47	[−0.63, 1.59]	−0.66	[−1.84, 0.53]
Dev. macro	+0.80	(5/6 pos.)	+0.32	(4/6 pos.)
tieredImageNet (held out)	+0.31	[−0.50, 1.13]	+3.98	[3.28, 4.70]
FGVC Aircraft (held out)	+2.47	[1.77, 3.20]	+3.51	[2.87, 4.16]

Table 18: **Alternative frozen representation (CLIP-17D).** ΔU_{new} in percentage points (95% paired bootstrap CI) for the two cross-episode scorers against the episode-local CAPE-BF arm re-fit on the same features, computed on the same evaluation episodes and certification machinery as the main text, under the earlier calibration take.

also positive (+3.86pp MLP; +3.70pp Linear), whereas on tieredImageNet the MLP contrast is near zero (−0.09pp, 95% CI $[-0.49, 0.30]$) and the linear contrast remains positive (+0.51pp, 95% CI $[0.14, 0.87]$); CAPE-BF is substantially weaker under this construction on some populations. Across DINOv3-centered, CLIP-derived, and supervised ResNet-50 constructions, the aggregate cross-episode advantage persists, but the preferred scorer and effect magnitude vary substantially with the representation.

I.6 Sensitivity to the risk level ϵ

The main evaluation fixes $\epsilon = 2\%$. As a post-hoc descriptive sensitivity analysis, we replay the same evaluation over the grid $\epsilon \in \{0.5, 1, 1.5, 2, 2.5, 3, 4, 5\}\%$ on the two held-out populations, changing only the certificate threshold ($k_{\max}(n_e, \epsilon)$ and the operating point); the full grid is reported. At $\epsilon = 0.5\%$ no certificate exists at the realized anchor-pool sizes ($k_{\max}(n_e, \epsilon) < 0$ on every episode), so all gated arms use the uncertified B2 pass-through fallback; for every $\epsilon \geq 1\%$ the certifiable coverage is 100% of fitted episodes in both populations. The sweep's $\epsilon = 2\%$ cells are the earlier-take values (+3.20/+3.41pp on Aircraft; the deployed-draw values are +3.27/+3.56pp, Table 15). Table 19 reports U_{new} over the grid. Both CCPL scorers improve on CAPE-BF at every certifiable ϵ in both populations. On tieredImageNet the gap is largest at the strictest certified level ($\epsilon = 1\%$: +1.46pp MLP, +3.63pp Linear) and narrows toward the B2 frontier as ϵ loosens; on Aircraft the advantage peaks near the operational level ($\epsilon = 2\%$: +3.20pp MLP, +3.41pp Linear) and is smaller at both ends of the grid (+1.69/+1.46pp at $\epsilon = 1\%$, +1.90/+2.53pp at 5%). The query-side old-to-new rate of the certified arms stays well below ϵ across the grid. The cross-episode advantage therefore persists across the certifiable risk-level sweep rather than being confined to a single operating point. On tieredImageNet the gain contracts as ϵ loosens past the operational level—consistent with the ranking advantage mattering most once the certificate is binding.

I.7 Boundary cases

Places365 (Zhou et al., 2018) remains a boundary case for certified registration. CCPL does not resolve its poor certified-utility regime, and the failure is not explained by calibration availability alone—it does not follow the calibration-availability pattern of §5. Because this population was already examined during method development, we treat it as a documented limitation rather than as additional held-out evidence, and understanding why some populations resist certified arbitration remains open. Our class universes contain at most dozens of old and new classes per episode, and all conflict evidence is computed from vision-encoder features; the controls of this appendix extend the studied range to 80 old classes and four sequential insertion waves, while substantially larger class spaces, longer registration trajectories, and non-vision modalities remain unstudied.

ϵ	tieredImageNet U_{new} (%) BF	MLP	Linear	FGVC Aircraft U_{new} (%) BF	MLP	Linear
0.5%	no certificate admissible ($k_{\max}(n_e, \epsilon) < 0$ everywhere); B2 pass-through					
1%	66.18	67.64	69.81	19.32	21.01	20.78
1.5%	76.19	77.62	78.70	27.37	30.56	30.21
2%	79.61	80.39	81.24	32.33	35.53	35.73
2.5%	81.08	81.64	82.14	35.79	38.77	38.86
3%	81.88	82.26	82.65	38.06	41.03	41.18
4%	82.74	82.97	83.16	41.58	44.11	44.34
5%	83.14	83.21	83.33	44.68	46.58	47.21

Table 19: **Risk-level sensitivity (post-hoc descriptive).** U_{new} over the frozen ϵ grid on the two held-out populations, computed under the earlier calibration take; scorers, anchors, candidate scores, and the anti-vacuity rule held fixed, only $k_{\max}(n_e, \epsilon)$ and the threshold recomputed. Certifiable coverage is 0% at $\epsilon = 0.5\%$ and 100% at every $\epsilon \geq 1\%$ (fitted episodes).

J Mechanism Analysis

We use four complementary interventions, together with a registration-count control, to examine two explanations for CCPL's advantage: regularization of an unstable episode-local estimator and reuse of conflict structure across registrations. We vary the amount of evidence available to the local scorer, interpolate between local and global scorers through shrinkage, disrupt cross-episode feature–label correspondence, leave the target population out of scorer training, and vary the number of source registrations.

Together, these experiments show that the cross-episode gain depends on meaningful feature–label correspondence and transfers beyond the registrations used for scorer training. The local-estimation results are consistent with a regularization effect but do not uniquely identify it.

All controls use the same episode data, threshold-selection rule, and certificate machinery as the corresponding evaluation, with $\text{CP}_u^{95} \leq 2\%$ for every certified decision. Intervals use the population-stratified paired bootstrap of §F. These diagnostic analyses were computed under the earlier calibration take; Appendix B.4 reports sensitivity to the final draw.

The local-evidence intervention uses the canonical CCPL comparison; at $r = 1$ it gives the earlier-take fitted-only contrast of +1.05pp. The remaining controls use a matched linear 10-D global scorer.

J.1 Local evidence and instability

We first reduce the fitting evidence available to the episode-local scorer. Nested subsampling of the deployed arm's fit rows (fractions $r \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, five deterministic subsample seeds) while holding candidates, calibration, certification, and queries fixed makes the paired CCPL advantage rise monotonically, from +1.05pp at full evidence to +3.97pp at 1/8 evidence (+1.75pp at 1/2, +2.69pp at 1/4; Table 20), strictly monotone within every subsample seed. The rise is not an artifact of changing fit success: on the 514 episodes where the reduced-data refit succeeds at every fraction and seed, the same curve reads +1.26 $[0.88, 1.66]$ at full evidence, +1.91 $[1.50, 2.35]$ at 1/2, +2.85 $[2.38, 3.35]$ at 1/4, and +3.97 $[3.43, 4.55]$ at 1/8—strictly monotone, every interval excluding zero.

The advantage grows in step with the local estimator's instability. Across B=20 refits per (episode, fraction), the mean pairwise ranking instability $I_{\text{rank}} = 1 - \overline{\rho_{\text{Spearman}}}$ increases within episode by +0.079 (95% CI $[0.074, 0.083]$; the episode-pooled difference is +0.086) when the fraction falls from 1/2 to 1/8, and at matched r=1/4 episodes with less stable local rankings obtain larger gains (Spearman +0.195, 95% CI $[0.126, 0.263]$, tail fraction $1{\times}10^{-4}$; exploratory, §J.5); the secondary associations at r=1/2 (+0.110, $[0.040, 0.181]$) and r=1/8 (+0.264, $[0.200, 0.329]$) point the same way. Because the resampling diagnostic

is identically zero at r=1 by construction, it measures sensitivity to reduced fitting evidence rather than the full-data estimator's sampling uncertainty; the full-evidence gain under this take remains +1.05pp. The instability is not explained by poor fit on the sampled training rows: train-side ranking AUC is 1.0 at every fraction. The fit rows therefore remain separable even as rankings on deployment candidates become less stable. Instability measured on the certificate-relevant candidate band (the top $k_e^\star$+2 candidates under the full-data refit) is similarly associated with the advantage (Spearman +0.192).

J.2 Shrinkage continuum

Moving the episode-local scorer toward the global scorer rapidly recovers certified utility and forms a broad plateau by $\lambda \approx 0.1$. Let $w_e(\lambda)$ denote the solution of an L-BFGS-B fit to the mean logistic loss on the episode's fit rows plus $\lambda\|w - w_G\|_2^2$ (intercept unpenalized; initialized at w_G with zero intercept, gradient tolerance 10^{-8}, and at most 2000 iterations; evaluated on a fixed grid of λ values). The paired gain over the λ=0 intervention endpoint is +1.86pp at λ=10^{-3}, +2.51pp at λ=0.1, and flat within 0.01pp through λ=100 and the global endpoint w_G (Table 21). The prespecified prediction that the preferred λ would decrease with increasing local evidence was not supported: the response surface is too flat for the grid to identify an optimum, and the largest observed utility occurs at λ=10 for every data fraction. The unregularized objective has no finite minimizer on separable fit rows, so the λ=0 endpoint is the finite numerical solution returned by L-BFGS-B at λ=0 — a pooled-coordinate local fit, not the deployed arm. Absolute levels on the same episodes (pooled; descriptive across families): the global endpoint reaches 79.64% U_{new}, versus 79.56% for the deployed CCPL scorer, 79.71% for the separately trained pooled linear scorer, 76.69% for the λ=0 endpoint, and 78.39% for the deployed episode-local arm. The λ=∞ endpoint is the same w_G scorer evaluated in Table 23: its macro contrast against the deployed arm on all 900 episodes is +0.52pp [0.23, 0.80]; no same-estimand contrast against the deployed arm exists for the 672-fitted subset alone.

J.3 Effect of the number of source registrations

As a secondary diagnostic, we vary the number of source registrations, using the matched linear 10-D instantiation. Training the linear 10-D instantiation with more source registrations D improves the global scorer, saturating by D=12–24 (Table 22). The rising-and-saturating pattern is the same under both optimization-budget conventions—a fixed 9000-step budget for every D, or $1500 \times D$ steps holding per-source exposure constant—so differing per-source exposure does not explain the pattern.

J.4 Structure and transfer

Destroying instance-level feature–label correspondence reverses the global scorer's advantage. Permuting the training labels within each training episode, while preserving episode sizes, within-episode label marginals, the global feature distribution, training volume, and optimization, drops the shuffled scorer 8.53pp below the deployed local arm (paired 95% CI $[-9.16, -7.92]$), whereas the real scorer beats the shuffled control by +9.05pp ($[8.49, 9.62]$, mean of five shuffle seeds, tail fraction 1×10^{-4}; Table 23). Retraining the scorer and standardizer with the evaluated population excluded also preserves a positive aggregate advantage, +0.39pp ($[0.09, 0.68]$, tail fraction 0.006), with per-population heterogeneity (Table 24). Populations with non-positive full-model effects change little under population exclusion, while Cars, Caltech-B, and DTD retain 74–98% of their full-model effects. The shuffle control shows that the gain depends on meaningful feature–label correspondence. Together with the held-out evaluations of §7.3, the LOPO results provide evidence that the learned conflict structure transfers across registrations and populations.

J.5 Protocol note

The r=1/4 instability–gain association is treated as exploratory. A sign error in its prespecified decision rule was discovered after outcome inspection, so the corrected analysis is not

considered confirmatory. The original and corrected bootstrap tails are provided in the supplement. The instability-increase, label-shuffle, and leave-one-population-out analyses were specified before outcome inspection. The prespecified prediction concerning the preferred shrinkage level was not supported.

Local fitting fraction r	CCPL − local (95% CI), pp	ranking instability I_{rank}	episodes (advantage)
1.0	+1.05 [0.73, 1.39]	0.000	672
0.5	+1.75 [1.39, 2.13]	0.047	672
0.25	+2.69 [2.28, 3.11]	0.089	661
0.125	+3.97 [3.43, 4.54]	0.133	514

Table 20: **Local-evidence and instability interventions.** Left/middle: nested subsampling of the deployed arm's fit rows (five deterministic subsample seeds, common fitted intersection); the local scorer is refit with the identical recipe while candidates, calibration, certification, and queries are held fixed. The episodes column counts episodes on which the reduced-data local refit succeeds; a refit failure removes the episode from that row's advantage estimate only, while the CCPL arm and its certification are unchanged. Right: mean refit-to-refit ranking instability $I_{\text{rank}} = 1 - \overline{\rho_{\text{Spearman}}}$ over B=20 refits per episode, all 672 fitted episodes at every fraction (zero at r=1 by construction). Failed instability refits (subsamples containing fewer than two label classes) are excluded from the refit pool; I_{rank} is computed pairwise over the remaining valid refits, with constant-score refits counted as full agreement. The matched common set of episodes succeeding at every fraction and seed (n=514) is strictly monotone: +1.26/+1.91/+2.85/+3.97 pp.

Shrinkage λ	gain over λ=0 (95% CI), pp	tail fraction
0.001	+1.86 [1.57, 2.17]	0.0001
0.01	+2.32 [1.99, 2.66]	0.0001
0.1	+2.51 [2.16, 2.86]	0.0001
1.0	+2.51 [2.16, 2.88]	0.0001
10.0	+2.52 [2.16, 2.88]	0.0001
100.0	+2.52 [2.16, 2.88]	0.0001
∞ (global endpoint)	+2.52 [2.16, 2.88]	0.0001

Table 21: **Shrinkage continuum** (matched 10-D coordinates; equal-weight six-population macro paired ΔU_{new} over the λ=0 intervention endpoint; 672 fitted episodes). Local rows are standardized by the pooled standardizer and shrunk toward the cross-episode scorer w_G; the λ=0 endpoint is the finite solution returned by L-BFGS-B at λ=0 (a pooled-coordinate local fit, not the deployed arm).

Budget	D=6	D=12	D=24
FIXED_TOTAL	+0.13 [-0.15, 0.41]	+0.44 [0.16, 0.71]	+0.49 [0.21, 0.77]
FIXED_PER_SOURCE	+0.13 [-0.15, 0.41]	+0.39 [0.11, 0.66]	+0.49 [0.21, 0.76]

Table 22: **Training-registration diversity under two budget conventions** (macro ΔU_{new} vs. deployed arm, pp; number of source registrations D; linear 10-D instantiation, five subset seeds). FIXED_TOTAL: 9000 steps for every D; FIXED_PER_SOURCE: $1500 \times D$ steps, holding per-source exposure constant.

Control	macro ΔU_{new} vs. deployed arm (95% CI), pp
Real cross-episode scorer	+0.52 [0.23, 0.80]
Shuffled correspondence (5 seeds)	−8.53 [-9.16, -7.92]
Leave-one-population-out	+0.39 [0.09, 0.68]

Table 23: **Structure and transfer controls** (linear 10-D instantiation, all 900 episodes). Shuffling: within-episode permutation of training labels preserving episode sizes, within-episode label marginals, the global feature distribution, training volume, and optimization; the shuffle row is the direct paired contrast of the seed-averaged shuffled scorer against the deployed arm. LOPO: the evaluated population is excluded from scorer and standardizer training.

Population	LOPO (pp)	Full global (pp)
cars	+0.61	+0.62
caltecha	-0.06	-0.06
caltechb	+0.74	+1.00
food	-0.24	-0.24
dtd	+1.73	+2.20
sun397	-0.46	-0.40

Table 24: **Leave-one-population-out heterogeneity.** Per-population macro ΔU_{new} of the LOPO and full global scorers against the deployed arm; populations with non-positive full-model effects change little under exclusion (at most 0.06pp).