PRIOR-AWARE TRANSDUCTIVE CONFORMAL PREDICTION FOR VISION–LANGUAGE MODELS UNDER LABEL SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction gives zero-shot vision–language models prediction sets with a guaranteed probability of containing the true label. Transductive methods shrink these sets by adapting scores to an unlabeled batch of query images, preserving coverage when calibration and query points are exchangeable and treated symmetrically. Conf-OT jointly fits optimal transport to calibration and query scores while fixing the pooled average class probabilities to the calibration label proportions. Compared with split conformal prediction, which calibrates fixed scores, Conf-OT reports smaller sets when both subsets share a distribution. We examine its coverage under label shift, where query class proportions change while the image distribution within each class stays fixed. We show that enforcing the calibration proportions drives the fitted class bias against classes that become more frequent in the query batch. Standard prior correction upweights these same classes. In a class-symmetric population model, we prove that the resulting coverage shortfall grows quadratically with small mismatches in class proportions. Under the strongest tested shift, Conf-OT averages 69.3% coverage versus 89.7% for split conformal prediction at a 90% target. To address this mechanism, we fit transport on half of the calibration set alone and estimate query class proportions from the other half and the query predictions. Holding the map fixed removes most of the shortfall, and prior correction reduces the remaining shortfall when class proportions can be reliably estimated from predictions. Under the same shift, our predictor attains 91.0% mean coverage with 3.30 labels per set versus 4.35 for split conformal prediction. These results link the coverage loss to pooled transport fitting and support prior-aware correction for compact prediction sets under label shift.

1 INTRODUCTION

Conformal prediction constructs prediction sets with a user-chosen coverage target, where coverage is the probability that the set contains the true label (Vovk et al., 2005; Angelopoulos & Bates, 2021). For zero-shot vision–language models such as Contrastive Language–Image Pretraining (CLIP), compact sets let users inspect plausible labels (Radford et al., 2021; Silva-Rodríguez et al., 2025).

To obtain smaller sets, transductive methods adapt scores to an unlabeled query batch (Gazin et al., 2024; Silva-Rodríguez et al., 2025). Coverage is preserved when calibration and query scores remain exchangeable under this adaptation (Gazin et al., 2024). For example, Conf-OT jointly fits optimal transport to calibration and query scores, rescaling pooled class probabilities toward calibration label proportions (Silva-Rodríguez et al., 2025). It reports smaller sets than split conformal prediction (SCP) when calibration and query subsets share a distribution (Silva-Rodríguez et al., 2025).

However, label shift changes query class proportions while preserving the image distribution within each class (Lipton et al., 2018; Podkopaev & Ramdas, 2021). For example, a product-recognition model may serve a store where some products occur more frequently than in its calibration data. Pooled transport then adjusts every image's scores using the changed query batch while still targeting calibration proportions. Label-conditional and prior-weighted prediction account for changing class frequencies in calibration (Vovk et al., 2005; Podkopaev & Ramdas, 2021). With pooled transport, these frequencies also affect the scores being calibrated.

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.

To analyze this dependence, we express pooled transport as a class-specific logit bias shared across images (Sec. 3). For an informative classifier in a class-symmetric population model, small increases in query class proportions reduce the fitted biases of the corresponding classes, whereas standard prior correction upweights those classes (Saerens et al., 2002). Classes that become more frequent consequently lose conditional coverage while receiving greater weight in query coverage. Formally, we prove that the resulting coverage shortfall grows quadratically with small mismatches between calibration and query class proportions.

Guided by this analysis, we separate transport fitting from query-prior estimation (Sec. 4). Specifically, we fit transport on one half of the calibration set and estimate query class proportions from the remaining labeled half and query predictions. The estimated prior then corrects class probabilities and reweights conformal calibration. Our analysis further connects prior-estimation error to coverage error and relates estimation reliability to how well class proportions can be distinguished from model predictions. Accordingly, shift detection and estimation-stability checks select when to apply correction, while shrinkage and a bootstrap criterion control its strength.

In experiments across 41 dataset–backbone combinations and three shift strengths, Conf-OT has lower coverage than SCP in 121 of 123 shifted configurations (Table 12). Under the strongest tested shift, Conf-OT attains 69.3% mean coverage versus 89.7% for SCP at a 90% target. Component comparisons show that calibration-only fitting removes most of this shortfall, and prior correction further improves coverage when query proportions can be reliably estimated. Under the same shift, our predictor achieves 91.0% mean coverage with 3.30 labels per set, compared with 4.35 for SCP (Table 1).

Our contributions are as follows.

- We derive a shared-bias representation of pooled transport and use it to analyze how query class proportions affect the fitted scores (Sec. 3).
- We establish a local quadratic coverage loss in a class-symmetric population model and characterize how prior-estimation error affects coverage (Theorems 1 and 3).
- We develop and evaluate a prior-aware predictor that combines calibration-only transport fitting with estimated-prior correction and weighted calibration to obtain compact prediction sets under label shift (Secs. 4 and 5).

2 Related Work

Conformal prediction under distribution shift. Conformal methods reweight calibration for covariate shift (Tibshirani et al., 2019) and label shift (Podkopaev & Ramdas, 2021; Choi, 2026), including PAC (Si et al., 2024) and federated settings (Plassier et al., 2023). Other approaches use uncertainty sets (Cauchois et al., 2024; Aolaritei et al., 2025), total-variation bounds (Barber et al., 2023) or score-distribution adjustment (Xu et al., 2025; Kasa et al., 2025; Correia & Louizos, 2025). Mixture methods combine calibration distributions (Yang & Jin, 2026; Wang et al., 2025), whereas label-conditional prediction calibrates classes separately (Vovk et al., 2005; Sadinle et al., 2019; Ding et al., 2023; 2026). We show how a constraint on the pooled class marginal moves the fitted map with the query prior.

Unlabeled inputs in conformal scores. Transductive inference incorporates unlabeled inputs while preserving score exchangeability (Vovk et al., 2005; Gazin et al., 2024). Conf-OT fits transport to pooled logits (Silva-Rodríguez et al., 2025), whereas SemiCP adds estimated calibration scores (Zhou et al., 2026). Optimal transport also couples calibration and test distributions for conformal calibration without a class-marginal constraint (Thurin et al., 2025). Reusing labels for score fitting and threshold selection creates tuning bias (Zeng et al., 2025). Temperature scaling changes coverage and set size (Guo et al., 2017; Dabah & Tirer, 2025). Empirical Bayes scores exploit prediction variability (Zeng et al., 2026).

Class proportions in vision–language models. CLIP adaptation and prompt ensembling use query statistics to improve accuracy (Zanella et al., 2024; Martin et al., 2024; Allingham et al., 2023). Balanced queries can conceal class-balance biases in few-shot learning (Veilleux et al., 2021) and vision–language adaptation (Zanella et al., 2025). Zero-shot calibration and label bias (LeVine et al.,

2023; Tu et al., 2024; Zhu et al., 2023) and class-conditional coverage under covariate shift (Sharma & Dutta, 2026) are related problems.

Label-shift estimation. Black-box shift estimation (BBSE) (Lipton et al., 2018) and EM (Saerens et al., 2002) recover query priors from unlabeled predictions. Calibration (Alexandari et al., 2020) and regularization (Azizzadenesheli et al., 2019) refine these estimates, with unified analyses (Garg et al., 2020) and probability matching (Wen et al., 2024) connecting alternatives. Online adaptation (Wu et al., 2024) and realistic benchmarks (Garg et al., 2023) extend this problem beyond a stationary target. We relate prior-estimation error to marginal coverage and use this relation to select correction strength.

3 Label Shift and Pooled Transport

We first isolate how pooled transport responds to query class proportions, then use this analysis to motivate prior estimation and correction.

3.1 Prediction sets and label shift

We observe N labeled calibration pairs and M unlabeled query inputs from K classes, with pool size $n = N + M$ and query fraction $\lambda = M/n$. For image x, logits $\ell(x) \in \mathbb{R}^K$ give probabilities $p(x) = \mathrm{softmax}(\ell(x)/\tau)$ at temperature $\tau > 0$. The least ambiguous classification (LAC) score $s(x, y) = 1 - p_y(x)$ assigns smaller values to more plausible labels (Sadinle et al., 2019). SCP (Vovk et al., 2005; Lei et al., 2018) sets $\hat{q}$ to the $\lceil (N+1)(1-\alpha) \rceil$-th smallest calibration score and returns $\mathcal{C}(x) = \{y : s(x, y) \le \hat{q}\}$, using $\hat{q} = +\infty$ if this index exceeds N. Exchangeable calibration and query scores give coverage at least $1 - \alpha$, for $\alpha \in (0, 1)$.

Under label shift, $P_c(X \mid Y) = P_q(X \mid Y)$ while class priors $\pi_c(y) > 0$ and $\pi_q(y)$ may differ. Their mismatch is $\delta = \pi_q - \pi_c$, with $\mathbf{1}^\top \delta = 0$ and total variation $\mathrm{TV}(\pi_q, \pi_c) = \frac{1}{2}\|\delta\|_1$. For a fixed score map, class-conditional coverages $F_y(t) = P_c(s(X, y) \le t \mid Y = y)$ are preserved, so marginal coverage is $\mathrm{cov}(t, \pi_q) = \sum_y \pi_q(y) F_y(t)$. Write $F(t)$ for the vector of these class coverages.

Conf-OT (Silva-Rodríguez et al., 2025) forms a transport matrix with classes as rows and images as columns. Sinkhorn iterations rescale rows toward the calibration label proportions m and columns toward $u_n = \mathbf{1}_n/n$, assigning equal mass to images. After T iterations, $r^{(T)} \in \mathbb{R}^K$ contains the accumulated row factors. Column normalization gives probabilities $Q^* = (q^*_{ki})$ for class k and image i, replacing $p(x_i)$ in the LAC score (Appendix B.2).

3.2 Pooled fitting and quadratic coverage loss

The transport factors act as a class logit bias b, shared across images, in $p^{(b)}(x) = \mathrm{softmax}(\ell(x)/\tau + b)$.

Lemma 1 (Structure). *For every $T \ge 1$ and positive $r^{(T)}$, $q^*_{ki} = [p^{(b)}(x_i)]_k$ with $b = \log r^{(T)}$. At convergence, $(1-\lambda)\bar{q}^{\mathrm{cal}}_k(b) + \lambda \bar{q}^{\mathrm{query}}_k(b) = m_k$, where the bars average probabilities over the indicated samples.*

The bias representation holds at finite T, and marginal equality at convergence. For the following population analysis, replace sample averages by expectations and m by π_c. For each fixed b, the mixture identity in Sec. 3.1 applies with class coverages $F_y(b, t)$. We study their response to changing query proportions.

Let $b_0, \hat{q}_0$ denote the no-shift population bias and threshold. Three matrices describe this response: $C_{\cdot y} = \mathbb{E}[p^{(b_0)}(X) \mid Y = y]$ averages predictions within classes, $J = \mathbb{E}_{\pi_c}[\mathrm{Diag}(p) - pp^\top]$ maps bias changes to mean probability changes, and $D_{yk} = \partial_{b_k} F_y(b_0, \hat{q}_0)$ maps them to class-coverage changes, with $p = p^{(b_0)}(X)$. Adding a common bias leaves softmax unchanged, so use the subspace $\mathbf{1}^\perp = \{v : \mathbf{1}^\top v = 0\}$ and the pseudoinverse J^+. Class symmetry (H4) requires the joint law of logits and labels to be invariant under consistent relabeling. It gives uniform π_c and reduces C, J, D on this subspace to scalars a_C, a_J, a_D. Assumptions (H1)–(H3) require these responses to be positive: informative predictions, increasing own-class coverage and nondegenerate probabilities (Appendix B.1).

Theorem 1 (Sign law and second-order growth)**.** *Under (H1)–(H4) and a smooth local expansion about $\delta = 0$, choose the population threshold $\hat{q}$ at each fitted bias b so that $\sum_y \pi_c(y) F_y(b, \hat{q}) = 1-\alpha$. The pooled predictor's query coverage satisfies*

$$\mathrm{cov} - (1-\alpha) = -\lambda \frac{a_D a_C}{a_J} \|\delta\|_2^2 + O(\|\delta\|^3), \tag{1}$$

which is negative for sufficiently small nonzero mismatch. In the symmetric two-class model with $\lambda = 1/2$, the sign holds for every nonzero shift yielding a valid query prior.

Classes gaining frequency lose conditional coverage and receive greater weight in query coverage, producing the quadratic loss. Sections 5.2 and 5.5 test these local predictions on real scores. Appendix B gives the proofs and general matrix expansion.

3.3 Pooled adaptation versus prior correction

An oracle knowing π_q adds $\log(\pi_q/\pi_c)$ to b_0, upweighting classes that gain frequency (Saerens et al., 2002). Denote this bias by $b^\star$ and the converged pooled bias by b_{OT}. Let $E(b)$ be excess expected set size over the oracle, with thresholds adjusted to equal query coverage.

Theorem 2 (Reverse adaptation and its price)**.** *Suppose $p^{(b_0)}$ equals the true posterior under π_c. To first order, $b_{\mathrm{OT}} - b_0 = -\lambda J^+ C\delta$ and $b^\star - b_0 = \mathrm{Diag}(\pi_c)^{-1}\delta$. Under (H1), (H3) and (H4), these displacements are anti-parallel on $\mathbf{1}^\perp$. For nonzero δ and $\lambda > 0$, the distance ratio is $\|b_{\mathrm{OT}} - b^\star\|_2 / \|b_0 - b^\star\|_2 = 1 + \lambda a_C/(K a_J) > 1$ at this order. The leading terms obey $E(b_{\mathrm{OT}}) = [1 + \lambda a_C/(K a_J)]^2 E(b_0)$.*

Accordingly, fitting transport on calibration data alone removes its dependence on query proportions. Section 5.3 measures how much coverage this separation recovers.

3.4 Coverage with estimated class proportions

To implement prior correction, we estimate π_q from predictions. For any fixed probability map p, redefine $C_{\cdot y} = \mathbb{E}[p(X) \mid Y = y]$ for that map. The query mean μ_q is the class mixture $C\pi_q$. Black-box shift estimation (BBSE) (Lipton et al., 2018) solves this relation using within-class calibration averages $\hat{C}$ and the unlabeled query mean $\hat{\mu}_q$ (Eq. (9)).

Theorem 3 (Coverage under an estimated prior)**.** *Fix a score map and positive prior vector $\hat{\pi}_q$ under label shift. For a population threshold with $\sum_y \hat{\pi}_q(y) F_y(\hat{q}) = 1 - \alpha$,*

$$\mathrm{cov}(\hat{q}, \pi_q) - (1-\alpha) = -\langle \hat{\pi}_q - \pi_q, F(\hat{q}) \rangle. \tag{2}$$

For the unshrunk linear estimate $\hat{\pi}_q = \hat{C}^{-1}\hat{\mu}_q$ with uniform π_c, assume C is invertible and $\|\hat{C} - C\|_2 \le \sigma_{\min}(C)/2$, where $\sigma_{\min}$ is the smallest singular value. With high probability,

$$\mathrm{TV}(\hat{\pi}_q, \pi_q) = O\left(\frac{\sqrt{K}}{\sigma_{\min}(C)} \left(\frac{1}{\sqrt{M}} + \frac{K}{\sqrt{N}} \right) \right). \tag{3}$$

The identity describes the direction of coverage error, while the rate controls prior-error magnitude: larger $\sigma_{\min}(C)$ makes class proportions easier to distinguish. Section 4 uses a constrained, shrunk estimate and an empirical stability indicator motivated by this rate. Section 5.2 evaluates that selection rule.

4 Method

The pooled-bias analysis motivates fixing a calibration-fitted map before estimating query proportions and selecting correction strength. Algorithm 1 summarizes the construction.

4.1 Separate fitting and prior estimation

Split calibration logits and labels (L_c, y_c) into class-stratified halves A, B of sizes N_A, N_B and priors $\hat{\pi}_A, \hat{\pi}_B$. Half A fits the map, while B supplies class-conditional averages and calibrates thresholds.

Algorithm 1 Split-sample prior-aware conformal prediction

Require: (L_c, y_c), L_q, levels α, α_0, tolerance τ_{rel} and settings in Appendix G.1.3
1: split calibration into class-stratified A, B and fit Sinkhorn scaling r on A
2: fix r, fit temperature τ_A on A, and evaluate $\mathbf{P}_B, \mathbf{P}_q$
3: compute $\hat{C}, \hat{\mu}_q$ and the half-B SCP fallback $\mathcal{C}_A$
4: test shift and compute stability statistics, including bootstrap prior fits
5: if shift is undetected or stability fails, return $\mathcal{C}_A$
6: select $\gamma^\star$ by five-fold query likelihood and refit $\bar{\pi}$ on all queries
7: select $\gamma \le \gamma^\star$ by the bootstrap criterion and set $\hat{\pi}_q = \pi_\gamma$
8: correct probabilities by Eq. (4), form the weighted threshold, and return prediction sets

Specifically, run Sinkhorn on L_A with marginals $\hat{\pi}_A, u_{N_A}$, then fix its scaling r and minimize half-A negative log-likelihood over temperature τ_A. The base map $p_A(x) = \mathrm{softmax}(\ell(x)/\tau_A + \log r)$ is therefore fixed as query proportions change. Its probabilities on B and query logits L_q form matrices $\mathbf{P}_B, \mathbf{P}_q$, with one image per row.

Using these predictions in Eq. (9), estimate $\bar{\pi} = \arg\min_{\pi \in \Delta_K} \|\hat{C}\pi - \hat{\mu}_q\|_2$ on the probability simplex, then shrink toward $\hat{\pi}_B$ via $\pi_\gamma = (1-\gamma)\hat{\pi}_B + \gamma\bar{\pi}$ (Azizzadenesheli et al., 2019; Garg et al., 2020). Thus $\gamma = 0$ gives no prior change and $\gamma = 1$ uses the constrained estimate. Five-fold query likelihood (Alexandari et al., 2020) selects an initial $\gamma^\star$ by evaluating $\log \sum_y p_{A,y}(x)\pi_\gamma^{(-f)}(y)/\hat{\pi}_B(y)$ on fold f, with the prior estimated from the other folds. We then refit $\bar{\pi}$ on all queries (Appendix G.1.1).

4.2 Probability correction and weighted calibration

For a selected prior $\hat{\pi}_q$, probability correction adjusts label plausibility, while calibration weights target the estimated query mixture. With reference prior $\hat{\pi}_B$,

$$p'_y(x) = \frac{p_{A,y}(x)\hat{\pi}_q(y)/\hat{\pi}_B(y)}{\sum_k p_{A,k}(x)\hat{\pi}_q(k)/\hat{\pi}_B(k)}. \tag{4}$$

The fitted base map stays fixed while p' depends on the candidate prior. For half-B scores $s'_i = 1 - p'_{y_i}(x_i)$, assign weights $\hat{w}_i = \hat{\pi}_q(y_i)/\hat{\pi}_B(y_i)$ (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021). Reserve $\max_i \hat{w}_i$ as test-point mass at $+\infty$, normalize all masses, and take the resulting distribution's $(1-\alpha)$ quantile $\hat{q}$. Return $\{y : 1 - p'_y(x) \le \hat{q}\}$. Appendix G.1 specifies class support and positive estimates. When $\hat{\pi}_q = \hat{\pi}_B$, this becomes the fallback $\mathcal{C}_A(x) = \{y : 1 - p_{A,y}(x) \le \hat{q}_B\}$, with $\hat{q}_B$ the ordinary SCP threshold on B. This uncorrected split-sample predictor is Per-split, distinguished from fitting on the full calibration set.

4.3 Selecting the correction

Correction requires evidence of shift and stable estimation. A permutation test compares the histograms of $\arg\max p_A(x)$ on B and queries using Pearson's two-sample χ^2 statistic, rejecting equal predicted-label distributions at nominal level α_0 (Lipton et al., 2018). Stability combines the plug-in scale $\sqrt{K}/\sigma_{\min}(\hat{C})\,(1/\sqrt{M} + K/\sqrt{N_B})$ from Eq. (3) with the mean ℓ_1 dispersion of clipped least-squares bootstrap estimates around their mean. Thresholds and resampling details are in Appendix G.1.3. If shift is undetected or stability fails, return $\mathcal{C}_A$.

For accepted correction, a separate bootstrap of constrained prior estimates limits its strength. Let $\hat{F}_\gamma$ contain half-B class coverages under candidate π_γ, and let $g(\gamma)$ be the 5% bootstrap quantile of $\langle \pi^{\mathrm{boot}} - \pi_\gamma, \hat{F}_\gamma \rangle$. Choose the largest grid value $\gamma \le \gamma^\star$ with $g(t) \ge g(0) - \tau_{\mathrm{rel}}$ at every grid point $t \le \gamma$. Thus $\tau_{\mathrm{rel}} = 0.02$ permits a two-percentage-point decline from no correction, and the final prior is $\hat{\pi}_q = \pi_\gamma$.

Under i.i.d. no-shift sampling and independent splitting, Appendix G.1 proves coverage at least $1-\alpha-\alpha'_0$, where α'_0 bounds null rejection by the implemented pretest at nominal level α_0. Section 5.3 compares this split with the class-stratified implementation. Sinkhorn costs $O(KN_A)$ per iteration and base-map evaluation costs $O(K(N_B + M))$. Selection adds repeated prior fits, permutations and bootstrap evaluations (Appendix G.1.3).

5 EXPERIMENTS

We evaluate the pooled-bias mechanism and prior-aware correction under controlled label shift, then isolate their components. Adapted scores test transfer, the ImageNet ladder tests loss scaling, and natural queries test changes within classes.

5.1 EXPERIMENTAL SETUP

Data and evaluation units. We use EuroSAT (Helber et al., 2019), OxfordPets (Parkhi et al., 2012), DTD (Cimpoi et al., 2014), Caltech101 (Fei-Fei et al., 2004), FGVCAircraft (Maji et al., 2013) and Flowers102 (Nilsback & Zisserman, 2008), spanning 10–102 classes. Four CLIP backbones (Radford et al., 2021), MetaCLIP-B/16 (Xu et al., 2024), EVA02-L/14 (Fang et al., 2024) and SigLIP-SO400M-384 (Zhai et al., 2023) give 41 dataset–backbone configurations, with SigLIP omitted on FGVCAircraft. Each shift strength adds one condition per configuration, evaluated over 20 repetitions with shared logits and splits across methods. Appendix C.1 specifies the models and sample sizes.

Comparisons. SCP (Vovk et al., 2005) and Conf-OT (Silva-Rodríguez et al., 2025) measure the effect of pooled fitting. Label-conditional prediction (Vovk et al., 2005) and Soft-Mondrian (Wang et al., 2025) compare alternative calibration rules for changing class mixtures. Per-split is the uncorrected A-fit, B-calibrated predictor, while full calibration measures the effect of reusing all calibration examples. Always-on applies unshrunk prior-weighted prediction (Podkopaev & Ramdas, 2021) after the pretest, omitting shrinkage, stability selection and the cap. Oracle uses the realized query prior as a known-prior reference.

Sampling and settings. Each repetition first separates calibration and query images within classes. We then divide calibration into A, B for our method. Query priors follow $\mathrm{Dir}(\beta \mathbf{1}_K)$, with smaller β giving stronger imbalance. Class quotas are sampled from held-out images, using replacement when needed, preserving the class-conditional image source. The unresampled stratified split is denoted by $\beta = \infty$. We use LAC at $\alpha = 0.10$ and Sinkhorn settings $\tau = 1, T = 3$. Stability thresholds are selected on ViT-B/16 and transferred to the other settings (Appendix C).

Metrics. We report marginal coverage and mean set size, averaging repetitions first and configurations equally. The paired difference $\Delta\mathrm{cov} = \mathrm{cov}_{Conf-OT} - \mathrm{cov}_{\mathrm{SCP}}$ isolates the empirical comparison with SCP. Equation (1) instead concerns the offset from $1 - \alpha$, so these comparisons assess the predicted pattern on real scores. Appendix C.2 defines standard errors, confidence intervals and configuration-level coverage counts.

5.2 COVERAGE AND EFFICIENCY UNDER CHANGING QUERY PRIORS

We first test whether pooled fitting retains its efficiency advantage when query class proportions change. Table 1 shows that stronger imbalance progressively lowers pooled coverage and enlarges its prediction sets. At $\beta = 0.1$, our predictor reduces SCP's mean set size by 24% while attaining 0.910 coverage. Across all three shifted conditions, it produces the smallest mean sets among methods whose mean coverage falls within the stated comparison band.

The loss extends across the suite: pooled coverage is lower than SCP in 121 of 123 shifted configurations (Table 12), and grows strictly across strengths in 37 of 41 dataset–backbone pairs (Fig. 5). Coverage distributions in Fig. 3 also show pooled prediction moving away from the target while Ours stays concentrated near it.

Alternative calibration rules. The classwise protection of Label-conditional prediction comes with larger sets when class counts are small. At $\alpha = 0.10$, fewer than nine examples give an infinite class threshold (Ding et al., 2023; 2026), which occurs for about half the Caltech101 classes (Table 5). Soft-Mondrian also attains higher coverage with much larger sets (Table 1). These comparisons expose the efficiency cost of the alternative calibration rules at their achieved coverages. Table 9 gives paired intervals and numerical sensitivity, and Table 8 isolates the effect of classwise sample allocation.

	$\beta = \infty$		$\beta = 1.0$		$\beta = 0.3$		$\beta = 0.1$	
Method	Cov ↑	Size ↓	Cov ↑	Size ↓	Cov ↑	Size ↓	Cov ↑	Size ↓
SCP	**.900**	4.42	**.900**	4.41	**.903**	4.41	**.897**	4.35
Conf-OT	**.899**	**3.51**	.855	3.82	.796	4.25	.693	4.97
Label-conditional	.928	20.99	.933	20.91	.934	20.90	.934	20.85
Soft-Mondrian	.935	14.64	.939	21.18	.939	20.07	.940	21.95
Per-split	**.899**	3.53	**.891**	3.53	**.891**	3.53	**.889**	3.47
Always-on	**.899**	3.53	.775	2.98	.791	3.19	.819	2.84
Ours	**.899**	3.53	**.897**	**3.47**	**.902**	**3.39**	**.910**	**3.30**

Table 1: **Coverage and set size under label shift.** Means cover 41 dataset–backbone combinations with 20 repetitions at $\alpha = 0.10$. Bold coverage is within 0.015 of 0.90, an empirical comparison band. Bold size is the smallest within this band. Paired uncertainty at $\beta = 0.1$ is reported in Table 11. Full-calibration and Oracle comparisons appear in Appendix D.1 and are excluded from ranking.

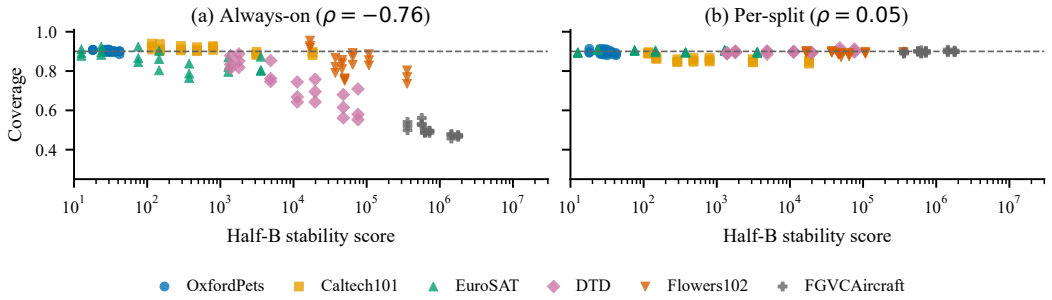

Figure 1: **Prior-estimation stability and coverage.** Each point averages 20 repetitions for one of 123 shifted configurations. Both panels use the same half-B stability score from Eq. (3). Colors and markers identify datasets, ρ is Spearman correlation, and the dashed line marks 0.90.

5.3 Separate fitting and correction components

Isolating query participation. We first remove only query inputs from transport fitting, holding $\tau = 1$ and the full calibration set fixed. Coverage at $\beta = 0.1$ rises from 0.693 to 0.886 (Table 2, top), supporting the query-coupling mechanism of Sec. 3.2. The Per-split and Ours rows in Table 1 then show the additional effect of prior correction after separate fitting: coverage improves while sets become smaller.

When correction is reliable. Theorem 3 motivates checking how accurately class proportions can be estimated before applying correction. Figure 1 shows a strong negative association between the half-B stability score and Always-on coverage, with little association for Per-split. The association also remains negative within every dataset (Table 19), supporting its use to identify settings sensitive to prior error.

Table 2 separates average performance from low-coverage configurations. Removing stability selection improves aggregate efficiency, yet increases the number of configuration means below 0.885 from 9 to 24. Removing the cap also increases this count and lowers mean coverage. Consequently, the two components serve complementary roles: stability selection avoids unreliable corrections, and the cap limits correction strength in the selected settings.

Threshold transfer. Thresholds selected without the evaluated dataset retain similar aggregate coverage and size (Table 6).

Efficiency costs. Separate fitting costs fewer than 0.08 labels relative to full-calibration fitting (Table 7). On stability-passing repetitions, correction recovers 65–73% of Oracle's size reduction from Per-split (Table 10).

To compare the sampling design with that used by the no-shift bound, we randomize both data partitions independently of labels. Independent and stratified splitting both attain mean coverage near 0.90, with smaller sets under stratification (Table 23).

Method	Coverage	Set size	Below 0.885	Below 0.85
Pooled fit, full calibration	0.693	4.97	—	—
Calibration-only fit, full calibration	0.886	3.44	—	—
Per-split	0.890	3.51	24	9
Always-on	0.795	3.00	72	56
No cap	0.897	3.35	14	0
No stability	0.905	2.86	24	0
Ours	0.903	3.39	9	0

Table 2: **Query participation and correction components.** Top: 41 configurations at $\beta = 0.1$, fixing $\tau = 1$ and full calibration. Bottom: means and low-coverage counts over 123 shifted configurations. All use 20 repetitions. Correction variants retain the pretest. Always-on uses an unshrunk prior, and No cap / No stability remove the named component from Ours. Counts use unrounded configuration means.

		$\beta = 1.0$		$\beta = 0.3$		$\beta = 0.1$	
Score source	Method	Cov	Size	Cov	Size	Cov	Size
CoOp	SCP	0.893	2.23	0.894	2.24	0.887	2.22
	Conf-OT	0.840	2.45	0.766	2.65	0.647	3.00
	Per-split	0.890	2.35	0.890	2.36	0.888	2.35
	Ours	0.898	2.30	0.903	2.24	0.914	2.22
Tip-Adapter-F	SCP	0.896	2.40	0.896	2.40	0.890	2.37
	Conf-OT	0.845	2.53	0.775	2.72	0.652	3.08
	Per-split	0.892	2.40	0.891	2.40	0.890	2.38
	Ours	0.898	2.33	0.904	2.27	0.917	2.21

Table 3: **Prior correction with adapted scores.** Means over six datasets and three training seeds per source, with 20 evaluation repetitions per group, ViT-B/16 and $\alpha = 0.10$. Size is the mean number of labels. Paired intervals for correction gains are in Table 25.

5.4 Prior correction with adapted scores

Supervised adaptation changes the score map and can therefore affect both pooled fitting and prior estimation. We evaluate 16-shot CoOp (Zhou et al., 2022) and Tip-Adapter-F (Zhang et al., 2022) on ViT-B/16, with three training seeds per dataset. Pooled loss persists in all 108 shifted configurations (Table 13). Table 3 shows that prior correction improves Per-split coverage and reduces its mean set size for both score sources. At the strongest shift, the corrected sets average about 2.22 labels. Here Per-split alone has larger mean sets than SCP, so correction supplies the net efficiency gain averaged across the two sources.

The full-calibration confusion matrices become more informative after adaptation (Table 24), while the split-sample stability check permits correction on about 70% of CoOp and 75% of Tip-Adapter-F repetitions. To test whether pooled loss depends on improving top-1 accuracy, we additionally evaluate Food101, held out from method design. Its CoOp accuracy stays close to the zero-shot reference, yet pooled coverage falls in every shifted configuration (Table 13). Thus the coverage mechanism also appears when score adaptation leaves classification accuracy nearly unchanged.

5.5 Coverage-loss scaling across class counts

We next test the approximately quadratic dependence on $\|\delta\|_2$ predicted by Theorem 1. We use ImageNet (Deng et al., 2009) with $K \in \{100, 200, 1000\}$, six backbones and five shift strengths. Figure 2(a) plots paired coverage loss relative to SCP. The fitted exponents cluster near two, ranging from 1.70 to 2.12. The logarithmic fit retains 1,737 of 1,800 observations with positive loss. Restricting the fit to the two weakest shifts gives more variable exponents (panel b), with fewer positive-loss observations available for the logarithmic fit. Table 15 reports both fits and their coefficient ratios, with the retained-point counts in Appendix H.3.

We also test whether calibration quantities predict loss magnitude. The theoretical coefficient ranks observed coefficients with correlation 0.765, while overestimating their magnitude (Fig. 6). Top-1

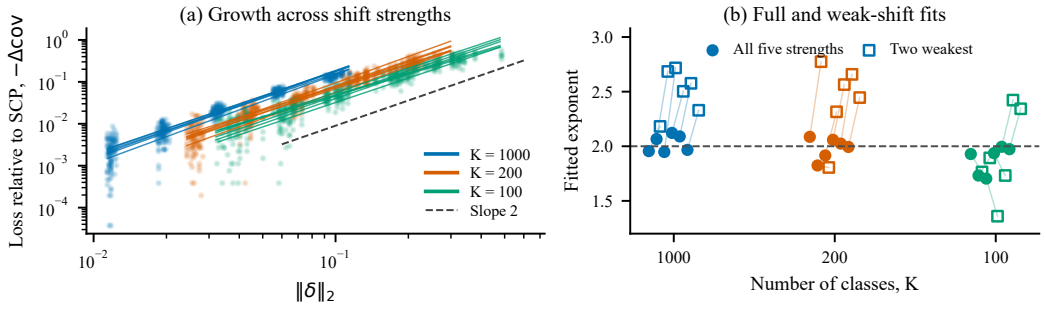

Figure 2: **Coverage-loss growth across class counts.** (a) Positive-loss repetitions and per-backbone log–log fits over five strengths, with a slope-2 reference. (b) Exponents from all five strengths (filled circles) and the two weakest (open squares), paired for each backbone. Dashed horizontal line: exponent 2. Each class count uses six backbones and 20 repetitions per strength.

accuracy also captures the within-class-count ordering. The evidence therefore supports calibration-based ranking, with the class-symmetric coefficient giving an approximate scale on real scores.

5.6 Long-tailed and naturally shifted query sets

To test label shift beyond Dirichlet sampling, we construct long-tailed ImageNet queries at four imbalance levels, using 25 calibration images per class. Pooled loss grows on every backbone, whereas our predictor maintains coverage near 0.90 by selecting Per-split whenever stability fails (Fig. 7, Table 29). This comparison illustrates the role of fallback when query proportions cannot be reliably estimated.

Finally, ImageNet-V2 (Recht et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a) test changes in the within-class image source. Table 30 shows that even Oracle falls below the target, reaching only 0.847 coverage on ImageNet-V2 despite matched class priors. On A/R both selection checks pass in every repetition. Nevertheless, correction lowers mean coverage relative to Per-split by about two and three percentage points, respectively. Thus shift detection and stable prior estimation do not test preservation of $P(X \mid Y)$. The natural-shift comparison identifies the class-conditional stability needed for prior correction to restore coverage.

6 Limitations

Our analysis assumes that calibration and query data share $P(X \mid Y)$. Appearance changes violate this assumption and invalidate the shared class-coverage identity even when the query prior is known. The scalar coverage-loss coefficient assumes class symmetry and overstates measured loss on real scores by a factor of three to four. The no-shift coverage bound uses an independent split and does not establish coverage for the class-stratified splits evaluated here.

7 Conclusion

We identify how pooled transport can reduce conformal coverage under label shift in vision–language models. Matching pooled scores to calibration class proportions induces a shared logit bias. In a class-symmetric population model, this bias produces a quadratic coverage shortfall for small prior mismatches. We also characterize how prior-estimation error affects coverage. Accordingly, our method separates transport fitting from query-prior estimation, then corrects probabilities and reweights calibration. Stability checks guide correction strength.

Experiments with zero-shot and adapted scores show that separate fitting recovers most coverage lost to pooling. Prior correction further improves the coverage–size trade-off. At the strongest Dirichlet shift, zero-shot results reach 91.0% mean coverage at a 90% target with sets 24% smaller than SCP. Natural-shift results show that changes within classes can prevent coverage recovery. These findings motivate evaluating coverage and set size jointly across query class proportions.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, 2020.

James Urquhart Allingham, Jie Ren, Michael W Dusenberry, Xiuye Gu, Yin Cui, Dustin Tran, Jeremiah Zhe Liu, and Balaji Lakshminarayanan. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *International Conference on Machine Learning*, pp. 547–568, 2023. URL https://proceedings.mlr.press/v202/allingham23a.html.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

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

Liviu Aolaritei, Julie Zhu, Oliver Wang, Michael I. Jordan, and Youssef Marzouk. Conformal prediction under Lévy–Prokhorov distribution shifts: Robustness to local and global perturbations. In *Advances in Neural Information Processing Systems*, 2025.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119 (548):3033–3044, 2024.

Seungjin Choi. Conformal bayes under label shift: Post-hoc calibration vs. in-training adaptation. *arXiv preprint arXiv:2606.11865*, 2026. URL https://arxiv.org/abs/2606.11865.

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*, 2014.

Alvaro H. C. Correia and Christos Louizos. Non-exchangeable conformal prediction with optimal transport: Tackling distribution shifts with unlabeled data. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2507.10425.

Lahav Dabah and Tom Tirer. On temperature scaling and conformal prediction of deep classifiers. In *International Conference on Machine Learning*, pp. 11813–11845, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, pp. 64555–64576, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, 2026.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149, 2024.

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. In *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, 2020.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. RLSbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, 2023.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pp. 1504–1512, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b.

Kevin Kasa, Zhiyu Zhang, Heng Yang, and Graham W. Taylor. Adapting prediction sets to distribution shifts without labels. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1990–2010, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Will LeVine, Benjamin Pikus, Pranav Raja, and Fernando Amat Gil. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023. URL https://arxiv.org/abs/2303.12748.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

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

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28816–28826, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pp. 27907–27947, 2023.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, 2021. arXiv:2103.03323.

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*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Jai Kumar Sharma and Amartya Dutta. Does marginal coverage guarantee class-conditional safety for zero-shot VLMs under shift? *arXiv preprint arXiv:2608.19376*, 2026.

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. PAC prediction sets under label shift. In *International Conference on Learning Representations*, 2024.

Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Conformal prediction for zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01856.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *International Conference on Machine Learning*, pp. 48791–48808, 2024. URL `https://proceedings.mlr.press/v235/tu24a.html`.

Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. In *Advances in Neural Information Processing Systems*, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Nien-Shao Wang, Duygu Nur Yaldiz, Yavuz Faruk Bakman, and Sai Praneeth Karimireddy. Conformal prediction adaptive to unknown subpopulation shifts. *arXiv preprint arXiv:2506.05583*, 2025.

Hongwei Wen, Annika Betken, and Hanyuan Hang. Class probability matching with calibrated networks for label shift adaption. In *International Conference on Learning Representations*, 2024.

Ruihan Wu, Siddhartha Datta, Yi Su, Dheeraj Baby, Yu-Xiang Wang, and Kilian Q. Weinberger. Online feature updates improve online (generalized) label shift adaptation. In *Advances in Neural Information Processing Systems*, pp. 106924–106954, 2024.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations*, 2024.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *International Conference on Learning Representations*, 2025.

Yuqi Yang and Ying Jin. Multi-distribution robust conformal prediction. In *International Conference on Machine Learning*, 2026.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.01837.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Hao Zeng, Kangdao Liu, Bingyi Jing, and Hongxin Wei. Parametric scaling law of tuning bias in conformal prediction. In *International Conference on Machine Learning*, pp. 74133–74156, 2025.

Jiapeng Zeng, Yogesh Prabhu, Zhanpeng Zeng, Michael A. Newton, and Vikas Singh. Empirical bayes conformal prediction for vision and language models. *arXiv preprint arXiv:2605.23189*, 2026. URL `https://arxiv.org/abs/2605.23189`.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Xuanning Zhou, Zihao Shi, Hao Zeng, Xiaobo Xia, Bingyi Jing, and Hongxin Wei. Semi-supervised conformal prediction with unlabeled nonconformity score. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17704–17713, 2026.

Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. In *Advances in Neural Information Processing Systems*, pp. 64663–64680, 2023.

A Limitations

This appendix discusses the scope of the theoretical model in Sec. 3 and the coverage and efficiency comparisons in Sec. 5.

The constant orders, it does not measure. Reducing the three matrices of Theorem 1 to scalars needs the class-symmetry assumption (H4), which real data satisfies only approximately, and the closed form overstates the loss threefold to fourfold (Appendix H.3). We use the constant as a ranking device, and even there its content is partly redundant with top-1 accuracy. The top-1 reversal derived in Appendix B.8 needs a monotonicity assumption (H5) we do not prove, so we keep it out of the main text and report it as measured rather than as a predicted crossing.

Only the prior may move. Sec. 3 assumes $P(X \mid Y)$ is shared between the halves. When it is not, Lemma 2 no longer isolates δ, and on the appearance-shifted ImageNet variants even an oracle given the realized query prior does not return to nominal. The identity of Theorem 3 then fails as

well, since it takes the class-conditional coverages to be shared. On ImageNet-A and -R the prior estimate mistakes the appearance shift for a label shift and costs 0.02 and 0.03 against the per-split variant. The halves are equal in size in the Dirichlet evaluation design, so $\lambda = 1/2$ there and the $1/\lambda$ dependence derived in Appendix B.8 is untested.

Coverage is bought with efficiency and with abstention. The bootstrap cap limits downward changes in the estimated coverage offset. Our predictor has coverage 0.910 at $\beta = 0.1$. On the repetitions where the stability check is satisfied it recovers 65 to 73 percent of the set-size reduction that an oracle prior would give over the per-split set. Where the check fails it returns the per-split set, and whether the check is satisfied is mostly a property of the dataset rather than of the shift. Across the seven backbones the check is satisfied on none of the shifted FGVCAircraft repetitions, on 2 and 10 percent of those of Flowers102 and DTD, and on all of those of OxfordPets. With ViT-B/16, $\sigma_{\min}(\hat{C})$ is 0.0001, 0.0008 and 0.0018 on the first three and 0.2233 on OxfordPets. Sample splitting also has a price. Half of the calibration set no longer sets the threshold, which makes the sets larger by 0.04 to 0.08 labels than those of the full-calibration variant.

Potential directions. Alternatives include a weighted quantile on pooled scores under the non-exchangeability penalty of Barber et al. (2023) and reweighting the transport plan itself.

B Assumptions, Proofs and Numerical Checks

This appendix supports the class-bias analysis in Sec. 3 through explicit assumptions and proofs. The numerical comparisons evaluate its identities and scaling predictions.

Notation follows Sec. 3. The pooled logit matrix is $S \in \mathbb{R}^{K \times n}$. The transport plan has class marginal m and uniform sample marginal u_n. The class logit bias b, also called a tilt, gives $p^{(b)}(x) = \mathrm{softmax}(\ell(x)/\tau + b)$. The no-shift tilt b_0 solves $\sum_y \pi_c(y)\mathbb{E}[p^{(b_0)} \mid Y = y] = \pi_c$. All three matrices are evaluated at b_0:

$$
\begin{aligned}
C_{\cdot y} &= \mathbb{E}\big[p^{(b_0)}(X) \mid Y = y\big], \\
J &= \mathbb{E}_{\pi_c}\big[\mathrm{Diag}(p^{(b_0)}) - p^{(b_0)} p^{(b_0)\top}\big] \succeq 0, \\
D_{yk} &= \partial F_y(b, \hat{q})/\partial b_k .
\end{aligned}
$$

$J\mathbf{1} = 0$, so J^+ is taken on $\mathbf{1}^\perp$, where the tilt is identified.

B.1 Assumptions

These assumptions support Theorems 1 and 2. They describe the response of probabilities and coverage to a class bias.

- **(H1)** $a_C > 0$: the soft confusion matrix is positive definite on $\mathbf{1}^\perp$. Writing $C = (1-\rho)I + \rho\,\mathbf{1}v^\top$ gives $a_C = 1 - \rho$, and $\rho < 1$ says the average posterior prefers the true class strictly more than it prefers the others on average. This is the signed version of the identifiability condition $\sigma_{\min}(C) > 0$ used by black-box shift estimation (Lipton et al., 2018).
- **(H2)** $a_D > 0$: raising the tilt of class y raises the class-y conditional coverage. Under LAC this holds pointwise, since $\partial_{b_y} p_y^{(b)} = p_y(1 - p_y) > 0$ and $\partial_{b_k} p_y = -p_y p_k < 0$ for $k \neq y$, so F_y is non-decreasing in b_y and strictly increasing wherever the score density at the threshold is positive.
- **(H3)** $a_J > 0$: the tilted posterior is not a hard one-hot. For $v \perp \mathbf{1}$, $v^\top(\mathrm{Diag}(p) - pp^\top)v = \mathbb{E}_{k\sim p}[v_k^2] - (\mathbb{E}_{k\sim p}[v_k])^2 \geq 0$, with equality only when p is degenerate on the values of v.
- **(H4)** Class symmetry: the joint law of (ℓ, Y) is invariant under relabeling. Then C, J, D are permutation-equivariant, hence of the form $aI + c\,\mathbf{1}\mathbf{1}^\top$, and reduce on $\mathbf{1}^\perp$ to the scalars a_C, a_J, a_D. In addition, $F_y(b_0, \hat{q}_0) = 1 - \alpha$ for every y.
- **(H5)** The top-1 error is monotone in $\|b - b^\star\|$. This unproven assumption is used only for the additional top-1 consequence in Appendix B.8.

Real data satisfies (H4) only approximately, which is why the plug-in constant is used as a ranking device in Sec. 5.5 rather than as a prediction of magnitude.

B.2 TRANSPORT SCALING AND NOTATION

This section gives the matrix construction underlying Lemma 1 and the base map in Sec. 4.

Let $S \in \mathbb{R}^{K\times n}$ contain the pooled logits as columns. The target class marginal is $m = N^{-1}\sum_{i=1}^{N} y_i^{\mathrm{ohe}}$, where y_i^{ohe} is the one-hot vector for label y_i. The uniform sample marginal $u_n = \mathbf{1}_n/n$ assigns equal mass to every image.

To fit both marginals efficiently, Conf-OT uses Sinkhorn iterations (Silva-Rodríguez et al., 2025). Each iteration rescales the rows toward m and the columns toward u_n. We initialize the probabilities and form the scaled matrix after T iterations as follows:

$$Q^{(0)} = \frac{\exp(S/\tau)}{\sum_{k,i}\exp(S_{ki}/\tau)}, \qquad \widetilde{Q} = \mathrm{Diag}(r^{(T)})Q^{(0)}\,\mathrm{Diag}(c^{(T)}), \tag{5}$$

where $Q^{(0)}$ is the initial matrix and $\widetilde{Q}$ is the unnormalized transport plan. The exponential acts elementwise and Diag forms a diagonal matrix from a vector. The vectors $r^{(T)} \in \mathbb{R}^K$ and $c^{(T)} \in \mathbb{R}^n$ contain the accumulated class and sample scaling factors. Column normalization of $\widetilde{Q}$ gives the probability matrix $Q^* = (q^*_{ki})$, whose columns replace $p(x_i)$ in the LAC score. Section 5.1 gives the values of τ and T used for this construction. Calibration-only fitting removes query inputs from this transport step. The default *Per-split* predictor fits transport and then temperature on half A, fixes that map, and calibrates SCP on half B (Sec. 4.2). Its full-calibration counterpart uses the entire calibration set for both fitting and threshold selection.

B.3 LEMMA 1 (STRUCTURE)

This section proves the shared class-bias representation in Lemma 1, supporting Sec. 3.2.

Lemma 1 restated. For every $T \geq 1$, the column-normalized transport probabilities satisfy

$$q^*_{ki} = [p^{(b)}(x_i)]_k, \qquad b = \log r^{(T)} \in \mathbb{R}^K,$$

where the logarithm acts componentwise and the same b applies to every image i. Thus the class scaling alone determines the normalized probabilities. If the iteration converges, the average probability of each class equals its target marginal:

$$\frac{1}{n}\sum_{i=1}^{n}[p^{(b)}(x_i)]_k = (1-\lambda)\bar{q}_k^{\mathrm{cal}}(b) + \lambda\bar{q}_k^{\mathrm{query}}(b) = m_k,$$

where $\bar{q}^{\,\mathrm{cal}}(b)$ and $\bar{q}^{\,\mathrm{query}}(b)$ are the average probability vectors on the calibration and query subsets. The constraint therefore matches a pooled average to a calibration-derived target.

Part (a). Line 11 of the published iteration (Silva-Rodríguez et al., 2025) gives $\tilde{q}_{ki} = r_k^{(T)}Q_{ki}^{(0)}c_i^{(T)}$ with $Q_{ki}^{(0)} = e^{\ell_{ki}/\tau}/Z$ and $Z = \sum_{k',i'} e^{\ell_{k'i'}/\tau}$ a scalar. The column normalization of line 12 then gives

$$q^*_{ki} = \frac{r_k^{(T)}e^{\ell_{ki}/\tau}c_i^{(T)}/Z}{\sum_{k'} r_{k'}^{(T)}e^{\ell_{k'i}/\tau}c_i^{(T)}/Z} = \frac{r_k^{(T)}e^{\ell_{ki}/\tau}}{\sum_{k'} r_{k'}^{(T)}e^{\ell_{k'i}/\tau}},$$

because $c_i^{(T)}$ is constant in k and Z is constant in both indices. This is $[\mathrm{softmax}(\ell_i/\tau + b)]_k$ with $b = \log r^{(T)}$. Since softmax is invariant to adding a constant to b, the effective number of free parameters is $K-1$.

Part (b). At a fixed point, $c_i^{\infty} = (1/n)/\sum_k r_k^{\infty}Q_{ki}^{(0)}$ and $r_k^{\infty}\sum_i Q_{ki}^{(0)}c_i^{\infty} = m_k$. Substituting the first into the second,

$$m_k = \sum_i \frac{r_k^{\infty}Q_{ki}^{(0)}}{n\sum_{k'} r_{k'}^{\infty}Q_{k'i}^{(0)}} = \frac{1}{n}\sum_i q^*_{ki},$$

by part (a). Splitting the pooled average over the two halves with weight $\lambda = M/n$ gives the stated form. □

B.4 Interpretation of the pooled constraint

The set and class-specific coverages used in Sec. 3 can be written as follows.

$$\mathcal{C}(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}\}. \tag{6}$$

$$F_y(t) = P_c\big(s(X, y) \leq t \mid Y = y\big), \qquad R(t) = \max_y F_y(t) - \min_y F_y(t), \tag{7}$$

This section supports Sec. 3.2 by explaining how the shared class bias connects the pooled inputs to coverage.

The proof in Appendix B.3 cancels the sample scaling factor when each column is normalized. Lemma 1 is an equivalent representation of Conf-OT, so it exposes the fitted bias without changing the algorithm. The target m comes from calibration labels while the fitted average also depends on query images. This dependence is the route through which query class proportions affect every image's score.

To translate this score change into coverage, we use the class-specific quantities in Eq. (7). The next identity separates the class proportions from the coverage attained within each class. It also states when a symmetric pooled transformation preserves the usual exchangeability argument.

Equation (5) fits class and sample scaling factors to the combined pool. Its matrix form leaves implicit how a change in the query batch changes an individual image's probabilities. We make this dependence explicit by expressing the same transformation as a shared class bias. This bias adds one offset per class to every image's temperature-scaled logits. We also refer to this additive adjustment as a tilt and write $p^{(b)}(x) = \mathrm{softmax}(\ell(x)/\tau + b)$.

As shown in Fig. 3, the coverage of Conf-OT decreases as query class proportions move away from calibration proportions. Table 1 also shows that its set-size advantage over SCP reverses under stronger shift. Section 3.2 reduces the pooled transport to a shared class bias. We derive its coverage effect in Sec. 3.2 and compare its direction with standard prior correction in Sec. 3.3. Section 3.4 establishes the estimation criterion and Sec. 4 defines the detection criterion used by our predictor. This section gives the proofs of these statements. The pooled constraint couples every score to query class proportions.

B.5 Lemma 2 (Coverage identity and a sharp bound)

This section proves the class-mixture identity and sharp coverage bound in Lemma 2. It supports the coverage analysis in Sec. 3.

Lemma 2 (Coverage identity and a sharp bound)**.** *For a fixed score map and threshold t under label shift, marginal coverage and its change obey*

$$\mathrm{cov}(t, \pi_q) = \sum_y \pi_q(y) F_y(t), \qquad |\mathrm{cov}(t, \pi_q) - \mathrm{cov}(t, \pi_c)| \leq \mathrm{TV}(\pi_q, \pi_c) R(t),$$

where $F_y(t)$ is class-specific coverage and $R(t)$ is its range. The bound is attained when the positive and negative parts of δ concentrate on classes attaining the opposite extrema of $F_y(t)$. For an exchangeable pool, a transport rule with a calibration-label-independent marginal m preserves score exchangeability when it is invariant to permutations of the pooled inputs. The split-conformal coverage guarantee then applies to these scores.

We prove the identity by conditioning on the class and the bound by centering the vector $F(t)$. The result explains why a prior mismatch affects coverage when classes have different coverages. If $R(t) = 0$, the coverage difference is zero for every prior mismatch. The exchangeability statement also identifies the symmetry needed to retain the standard guarantee after pooled adaptation (Gazin et al., 2024).

Identity. By the law of total probability, $P_q(Y \in \mathcal{C}(X)) = \sum_y \pi_q(y) P_q(s(X, y) \leq t \mid Y = y)$. Under label shift the conditional probability equals $F_y(t)$ and does not depend on the prior.

Bound. Since $\mathbf{1}^\top \delta = 0$ we may center F: for any constant c, $\langle \delta, F \rangle = \langle \delta, F - c\mathbf{1} \rangle$. Taking $c = \frac{1}{2}(\max_y F_y + \min_y F_y)$ gives $\|F - c\mathbf{1}\|_\infty \leq \frac{1}{2} R$, and Hölder's inequality gives $|\langle \delta, F \rangle| \leq$

$\|\delta\|_1 \cdot \frac{1}{2}R = \mathrm{TV} \cdot R$. Equality in Hölder requires the mass of δ to sit on the coordinates attaining $|F_y - c| = \frac{1}{2}R$ with matching signs, and $\mathbf{1}^\top \delta = 0$ forces the positive and negative masses each to equal TV.

Validity when m ignores the calibration labels. By Lemma 1(a) the score is $s_i = 1 - q^*_{y_i,i}$ where the code map $\ell \mapsto \mathrm{softmax}(\ell/\tau + \log r)$ is a deterministic, pointwise function of an unlabeled quantity as soon as m does not use the calibration labels. The map is invariant to permutations of the pool index, so it preserves exchangeability of $\{(\ell_i, y_i)\}$, and the split-conformal guarantee applies to the transformed scores unchanged. □

B.6 Theorem 1 (Sign law and second-order growth)

This section proves the local coverage law and the two-class sign statement in Theorem 1. It supports the mechanism described in Sec. 3.2.

Theorem 1 restated. In the population model under (H1)–(H4), use a threshold satisfying $\sum_y \pi_c(y) F_y(b, \hat{q}) = 1 - \alpha$. The pooled predictor has coverage offset $-\lambda \delta^\top A \delta + O(\|\delta\|^3)$, where $A = \mathrm{sym}(DJ^+C)$. On $\mathbf{1}^\perp$, this matrix equals $(a_D a_C / a_J) I$ with I the identity matrix. Consequently, the coverage offset is

$$\mathrm{cov} - (1-\alpha) = -\lambda \frac{a_D a_C}{a_J} \|\delta\|_2^2 + O(\|\delta\|^3),$$

where cov abbreviates the query coverage at the population calibration threshold. The offset is negative for every sufficiently small nonzero mismatch.

In the symmetric two-class model developed at the end of this section, this sign also holds for every admissible nonzero mismatch with equal calibration and query sizes. Here $\pi_c = (\frac{1}{2}, \frac{1}{2})$ and $\pi_q = (\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon)$ with $\varepsilon \in (-\frac{1}{2}, \frac{1}{2}) \setminus \{0\}$. Let G_1 be the class-one cumulative distribution function of the log-odds $z = (\ell_1 - \ell_2)/\tau$. We assume a positive density for G_1 and class-two cumulative distribution function $G_2(t) = 1 - G_1(-t)$. The class-one mean unshifted probability satisfies $\mathbb{E}_{z \sim G_1}[(1 + e^{-z})^{-1}] > 1/2$. Write $b_{\mathrm{bin}} = b_1 - b_2$ for the bias added to these log-odds and $u = \log((1-\hat{q})/\hat{q})$ for their inclusion threshold. The exact sign and coverage offset are

$$\mathrm{sign}\, b_{\mathrm{bin}}(\varepsilon) = -\,\mathrm{sign}\,\varepsilon, \qquad \mathrm{cov} - (1-\alpha) = -\varepsilon[G_1(u - b_{\mathrm{bin}}) - G_1(u + b_{\mathrm{bin}})] < 0,$$

Here b_{bin} measures the relative class bias. Thus the bias opposes the prior change throughout this two-class model.

Step 1. Linear bias response. Write $\mu_y(b) = \mathbb{E}[p^{(b)} \mid Y = y]$. Under label shift the pooled average code is $\bar{P}(b) = \sum_y \pi_c(y)\mu_y(b) + \lambda \sum_y \delta_y \mu_y(b)$, and Lemma 1(b) requires $\bar{P}(b) = \pi_c$. At b_0 we have $\sum_y \pi_c(y)\mu_y(b_0) = \pi_c$. Using $\partial_b \sum_y \pi_c(y)\mu_y = J$ and $\sum_y \delta_y \mu_y(b_0) = C\delta$, linearization gives

$$J(b - b_0) + \lambda C\delta + O(\|b - b_0\|^2 + \|\delta\|\|b - b_0\|) = 0,$$

so $b - b_0 = -\lambda J^+ C\delta + O(\|\delta\|^2)$.

Step 2. Coverage expansion. The threshold is fixed by the calibration side, $\sum_y \pi_c(y) F_y(b, \hat{q}) = 1 - \alpha$, so subtracting from the identity of Lemma 2 gives $\mathrm{cov} - (1-\alpha) = \langle \delta, F(b, \hat{q}) \rangle$ exactly. Expanding F at $(b_0, \hat{q}_0)$,

$$F(b, \hat{q}) = F(b_0, \hat{q}_0) + D(b - b_0) + (\partial_{\hat{q}} F)\, \Delta\hat{q} + O(\cdot^2).$$

Under (H4) the first term is $(1-\alpha)\mathbf{1}$ and $\partial_{\hat{q}} F = \phi\, \mathbf{1}$, and both are orthogonal to δ. Substituting Step 1 and symmetrizing gives $A = \mathrm{sym}(DJ^+C)$ and $\mathrm{cov} - (1-\alpha) = -\lambda\, \delta^\top A\, \delta + O(\|\delta\|^3)$.

Step 3. Scalarization. The algebra of permutation-equivariant $K \times K$ matrices is spanned by $\{I, \mathbf{1}\mathbf{1}^\top\}$, which act on $\mathbf{1}^\perp$ as the scalars a_C, a_J, a_D, so $A|_{\mathbf{1}^\perp} = (a_D a_C / a_J) I$ and $\delta^\top A \delta = (a_D a_C / a_J)\|\delta\|_2^2$. The signs of a_C, a_J, a_D are (H1)–(H3), proved in Appendix B.1. Without (H4) the first-order term $\langle \delta, F(b_0, \hat{q}_0) - (1-\alpha)\mathbf{1} \rangle$ survives, and its expectation vanishes when the direction of δ is drawn symmetrically. The threshold term $\langle \delta, (\partial_{\hat{q}} F)\Delta\hat{q} \rangle$ is then also of second order and unsigned, which is why the theorem is stated under (H4). □

Corollary for the symmetric two-class model. For equal calibration and query sizes, the exact sign statement in Theorem 1 takes the following form. Let $\pi_c = (\frac{1}{2}, \frac{1}{2})$ and $\pi_q = (\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon)$

with $\varepsilon \in (-\frac{1}{2}, \frac{1}{2}) \setminus \{0\}$. Let G_1 be the class-one cumulative distribution function of the log-odds $z = (\ell_1 - \ell_2)/\tau$. We assume a positive density for G_1 and class-two cumulative distribution function $G_2(t) = 1 - G_1(-t)$. The class-one mean unshifted probability satisfies $\mathbb{E}_{z\sim G_1}[(1+e^{-z})^{-1}] > 1/2$. Write $b_{\text{bin}} = b_1 - b_2$ for the bias added to these log-odds and $u = \log((1-\hat{q})/\hat{q})$ for their inclusion threshold. The exact sign and coverage offset are

$$\operatorname{sign} b_{\text{bin}}(\varepsilon) = -\operatorname{sign}\varepsilon, \qquad \operatorname{cov} - (1-\alpha) = -\varepsilon[G_1(u - b_{\text{bin}}) - G_1(u + b_{\text{bin}})] < 0.$$

Here b_{bin} measures the relative class bias. Thus the bias opposes the prior change throughout this two-class model.

The two-class model. Let $K = 2$ with $\pi_c = (\frac{1}{2}, \frac{1}{2})$ and $\pi_q = (\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon)$. The log-odds $z = (\ell_1 - \ell_2)/\tau$ follow $z \mid Y{=}1 \sim G_1$ with positive density. Class symmetry gives $G_2(t) = 1 - G_1(-t)$. Write $h(b_{\text{bin}}) = \mathbb{E}_{G_1}[\sigma(z + b_{\text{bin}})]$ and $\eta = h(0) - \frac{1}{2} > 0$. In the population limit with $\lambda = \frac{1}{2}$, the fixed-point equation reduces to

$$\Phi(b_{\text{bin}}, \varepsilon) = \big(\tfrac{1}{2} + \tfrac{\varepsilon}{2}\big)h(b_{\text{bin}}) + \big(\tfrac{1}{2} - \tfrac{\varepsilon}{2}\big)\big(1 - h(-b_{\text{bin}})\big) - \tfrac{1}{2} = 0.$$

Here $\Phi(0,0) = 0$ and $\partial_\varepsilon \Phi|_{(0,0)} = \eta > 0$. Furthermore $\partial_{b_{\text{bin}}} \Phi|_{(0,0)} = h'(0) > 0$. Hence $b_{\text{bin}}(0) = 0$ and $db_{\text{bin}}/d\varepsilon|_0 = -\eta/h'(0) < 0$. Globally $\partial_{b_{\text{bin}}} \Phi > 0$ for $|\varepsilon| < \frac{1}{2}$, so the solution is unique. The value $\Phi(0, \varepsilon) = \varepsilon\eta$ has the sign of ε, giving $\operatorname{sign} b_{\text{bin}}(\varepsilon) = -\operatorname{sign}\varepsilon$. With $u = \sigma^{-1}(1-\hat{q})$ one has $F_1 = 1 - G_1(u - b_{\text{bin}})$ and $F_2 = 1 - G_1(u + b_{\text{bin}})$. The calibration condition reads $\frac{1}{2}[G_1(u - b_{\text{bin}}) + G_1(u + b_{\text{bin}})] = \alpha$. The identity of Lemma 2 gives $\operatorname{cov} - (1-\alpha) = -\varepsilon[G_1(u - b_{\text{bin}}) - G_1(u + b_{\text{bin}})]$, which is negative for both signs of ε. Expanding $G_1(u - b_{\text{bin}}) - G_1(u + b_{\text{bin}}) = -2b_{\text{bin}} g_1(u) + O(b_{\text{bin}}^3)$ recovers the quadratic law with constant $2g_1(u_0)\eta/h'(0)$, consistent with $-\lambda(a_D a_C / a_J)\|\delta\|_2^2$ at $\lambda = \frac{1}{2}$ and $\|\delta\|_2^2 = 2\varepsilon^2$.

B.7 Coverage expansion and symmetry

This section supports Theorem 1 by explaining its response matrices and the contributions canceled by class symmetry.

Lemma 2 relates a prior mismatch to class-specific coverage differences. However, its absolute bound does not specify the direction of the change induced by the pooled bias. We determine that direction and its local rate by tracking how class proportions change predictions and how predictions change coverage. The calculation uses a population model with class symmetry.

Let b_0 be the converged population bias with no shift and let $\hat{q}_0$ be its population calibration threshold. We write $F_y(b, t)$ when the dependence of class coverage on the bias must be explicit. The vector $F(b, t)$ collects these class coverages. Three matrices describe the response around $(b_0, \hat{q}_0)$. The soft confusion matrix C records the average probability vector for each true class. The matrix J describes how average probabilities respond to the bias. The matrix D describes how class coverage responds at a fixed threshold. Their definitions are

$$C_{\cdot y} = \mathbb{E}[p^{(b_0)}(X) \mid Y = y], \qquad J = \mathbb{E}_{\pi_c}[\operatorname{Diag}(p) - pp^\top], \qquad D_{yk} = \left.\frac{\partial F_y(b,t)}{\partial b_k}\right|_{(b_0, \hat{q}_0)},$$

where p abbreviates $p^{(b_0)}(X)$ inside the expectation and $\mathbb{E}_{\pi_c}$ averages under the calibration law. These matrices connect a change in the prior to a change in the bias and then to class coverage.

Adding the same constant to every coordinate of b leaves softmax unchanged. We therefore compare biases on $\mathbf{1}^\perp = \{v : \mathbf{1}^\top v = 0\}$, the space of relative class offsets. Let J^+ denote the pseudoinverse on this space and let $\operatorname{sym}(H) = (H + H^\top)/2$ denote matrix symmetrization. Assumption (H4) states that relabeling the classes leaves the joint law of logits and labels unchanged. It gives equal calibration priors and equal class coverages at the unshifted threshold. On $\mathbf{1}^\perp$, each matrix acts as a scalar multiple of the identity. We denote these scalars by a_C for C, a_J for J and a_D for D. Assumptions (H1)–(H3) require these scalars to be positive. They express an informative classifier, a non-degenerate posterior and increasing class coverage as that class's bias increases. Appendix B.1 states these conditions in full. The positivity conditions determine the sign of the quadratic response.

Appendix B.6 linearizes the pooled matching equation and substitutes the resulting bias change into Lemma 2. The general expansion includes the linear term $\langle \delta, F(b_0, \hat{q}_0) - (1-\alpha)\mathbf{1}\rangle$. This class-mixture contribution is also present with fixed SCP scores and averages to zero over symmetrically

drawn directions of δ. The threshold response contributes an additional second-order term. Class symmetry makes the unshifted coverage and the threshold-response terms equal across classes, so their contributions cancel against the zero-sum mismatch δ. Equation (1) combines two effects. Classes that become more frequent lose conditional coverage and simultaneously receive more weight in query coverage. Their product gives the quadratic loss. The factor λ measures the query contribution to fitting the bias. The ratio a_C/a_J controls the bias response and a_D converts it into coverage change. The term $O(\|\delta\|^3)$ contains the higher-order remainder. The loss vanishes at $\delta = 0$. Section 5.5 tests the quadratic rate and whether the calibration-derived coefficient ranks the observed losses.

B.8 Theorem 2 (Reverse adaptation and its price)

This section proves Theorem 2, supporting the comparison between pooled adaptation and prior correction in Sec. 3.3.

Appendix B.8 proves the comparison by expanding both biases around b_0. The distance ratio is $(K + \lambda a_C/a_J)/K$, measuring the additional displacement caused by pooled fitting.

Theorem 2 restated. Assume $p^{(b_0)}$ equals the true posterior under π_c. Let b_{OT} denote the bias fitted by the converged pooled transport. To first order in δ, its displacement is $b_{\mathrm{OT}} - b_0 = -\lambda J^+ C\delta$ while the oracle displacement is $b^\star - b_0 = \mathrm{Diag}(\pi_c)^{-1}\delta$. Under (H1), (H3) and (H4) with uniform π_c, these displacements are anti-parallel on $\mathbf{1}^\perp$. At this order, their distances from the oracle obey

$$\|b_{\mathrm{OT}} - b^\star\|_2 = \left(K + \lambda \tfrac{a_C}{a_J}\right)\|\delta\|_2 > K\|\delta\|_2 = \|b_0 - b^\star\|_2,$$

where the comparison is for nonzero δ and $\lambda > 0$. The pooled bias is therefore farther from the oracle than the unadapted bias. At a common query coverage, its leading excess set size over the oracle is the squared distance ratio times the excess for no adaptation. This restatement carries exactly the conclusions that Theorem 2 states in Sec. 3.3.

Full statement. Assume $p^{(b_0)}$ equals the true posterior under π_c. To first order $b_{\mathrm{OT}} - b_0 = -\lambda J^+ C\delta$ while $b^\star - b_0 = \mathrm{Diag}(\pi_c)^{-1}\delta + O(\|\delta\|^2)$. Under (H1), (H3) and (H4) with π_c uniform the two are anti-parallel on $\mathbf{1}^\perp$ and Eq. (8) holds, so the pooled step ends up farther from the oracle than doing nothing. In the tilt family, the expected set size at a given query coverage is minimized at $b^\star$. The excess is quadratic in $b - b^\star$. Therefore the excess size of the pooled step exceeds that of no adaptation by the square of the ratio in (8).

An additional consequence under the unproven (H5). Theorem 2 and the restatement above cover set size alone. If the top-1 error is also monotone in $\|b - b^\star\|$ (H5), there is a threshold $\mathrm{TV}^\dagger \asymp \|b_0\|_2/(\lambda\, a_C/a_J)$ beyond which the top-1 accuracy of the pooled step falls below that of the untransformed model, and $\mathrm{TV}^\dagger \propto 1/\lambda$. We do not prove (H5), so the main text states neither this consequence nor the threshold, and no claim of the paper rests on it. Appendix B.14 locates the crossing in the synthetic model and Appendix A records the assumption it needs.

The published ablation over m. Replacing m by a uniform prior still leaves the method ahead of split conformal prediction under the equal-split evaluation design (Silva-Rodríguez et al., 2025). This is consistent with the statement above, since under an equal split a misspecified m shifts both halves alike. The mechanism isolated here is the asymmetry between the halves, not the choice of m.

Proof. The oracle tilt follows from the prior-correction identity of Saerens et al. (2002): if $p^{(b_0)}$ equals the posterior under π_c, then $b^\star = b_0 + \log(\pi_q/\pi_c)$ componentwise, so $b^\star - b_0 = \log(1 + \delta/\pi_c) = \mathrm{Diag}(\pi_c)^{-1}\delta + O(\|\delta\|^2)$. Step 1 of Appendix B.6 gives $b_{\mathrm{OT}} - b_0 = -\lambda J^+ C\delta$. Under (H1), (H3) and (H4) with π_c uniform these are $+K\delta$ and $-\lambda(a_C/a_J)\delta$ on $\mathbf{1}^\perp$, hence anti-parallel, and $\|b_{\mathrm{OT}} - b^\star\|_2 = (K + \lambda a_C/a_J)\|\delta\|_2 > K\|\delta\|_2 = \|b_0 - b^\star\|_2$.

For the size claim, let $\eta_y(x) = P_q(Y = y \mid X = x)$. At $b = b^\star$ the tilted posterior equals η, so the LAC set is a level set of η, which among all set-valued predictors attaining a given marginal coverage has the smallest expected cardinality (Sadinle et al., 2019). For each b, adjust its threshold to this same query coverage and let $E(b)$ be its excess expected cardinality over the oracle. This excess is non-negative and vanishes to first order at $b^\star$, so $E(b) = \frac{1}{2}(b - b^\star)^\top H(b - b^\star) + O(\|b - b^\star\|^3)$ with $H \succeq 0$. Substituting the two distances gives the ratio-squared statement.

For the top-1 claim, the distance from the pooled tilt to the oracle grows with slope $K + \lambda a_C/a_J$ in $\|\delta\|_2$, while the distance from the untransformed model is bounded by $\|b_0\|_2 + K\|\delta\|_2$. The two lines cross at $\|\delta\|_2 \asymp \|b_0\|_2/(\lambda a_C/a_J)$. Under (H5) the top-1 error is monotone in that distance, so the crossing is a reversal point, and the expression is proportional to $1/\lambda$. Without (H5) the argument gives the crossing of the two distances but not of the two error rates, which is why the reversal appears only as a measured quantity in Appendix B.14. □

B.9 Comparison with prior correction

The distance comparison in Theorem 2 has the explicit form

$$\|b_{\mathrm{OT}} - b^\star\|_2 = \Big(K + \lambda \tfrac{a_C}{a_J}\Big)\|\delta\|_2 > K\|\delta\|_2 = \|b_0 - b^\star\|_2. \tag{8}$$

This section supports Sec. 3.3 by interpreting the displacement and efficiency comparison in Theorem 2.

Appendix B.8 compares the two bias expansions and relates their distances to set size and top-1 error. The distance ratio in Eq. (8) is $(K + \lambda a_C/a_J)/K$. Its excess over unity measures the additional displacement caused by enforcing the calibration proportions on the pool. At $\delta = 0$, both prior-adaptation displacements vanish and $b_{\mathrm{OT}} = b_0$. The comparison motivates fitting the base transport separately and using the query prior to direct subsequent adaptation. Section 5.2 compares the corresponding predictors.

Theorem 1 identifies the coverage loss caused by the pooled constraint in the symmetric model. The remaining question is how its bias compares with a correction directed toward the query prior. We make this comparison using standard prior correction (Saerens et al., 2002). For a posterior under π_c, this correction multiplies each class probability by $\pi_q(y)/\pi_c(y)$ and renormalizes. In logit space, the oracle bias $b^\star$ therefore satisfies $b^\star - b_0 = \log(1 + \delta/\pi_c)$ componentwise. This is the appropriate comparison because it changes the prior while retaining the class-conditional image laws.

B.10 Theorem 3 (Coverage under an estimated prior)

This section proves the coverage identity and estimation rate in Theorem 3. It supports the prior-estimation criteria used in Sec. 4.

Theorem 3 restated. Fix a score map and a positive probability vector $\hat\pi_q$ under label shift. Let $\hat q$ be a population threshold satisfying $\sum_y \hat\pi_q(y) F_y(\hat q) = 1 - \alpha$. Its query coverage offset is

$$\mathrm{cov}(\hat q) - (1 - \alpha) = -\langle \hat\pi_q - \pi_q, F(\hat q)\rangle,$$

where $\mathrm{cov}(\hat q) = \mathrm{cov}(\hat q, \pi_q)$ and $F(\hat q)$ contains the class-specific coverages of the fixed map. This identity weights each prior-estimation error by its class coverage. For the unshrunk linear estimator $\hat\pi_q = \hat C^{-1}\hat\mu_q$ with uniform π_c, assume C is invertible and $\|\hat C - C\|_2 \le \sigma_{\min}(C)/2$. Its prior error satisfies the following rate with high probability:

$$\mathrm{TV}(\hat\pi_q, \pi_q) = O\left(\frac{\sqrt{K}}{\sigma_{\min}(C)}\Big(\frac{1}{\sqrt{M}} + \frac{K}{\sqrt{N}}\Big)\right),$$

where M and N are the query and calibration sample sizes used in Eq. (9). The two sampling terms describe uncertainty in the query mean and in the estimated confusion matrix.

The identity is one line. With the score map and $\hat\pi_q$ fixed, write $G_{\hat\pi}(t) = \sum_y \hat\pi_q(y) F_y(t)$, so that $G_{\hat\pi}(\hat q) = 1 - \alpha$. Lemma 2 gives

$$\mathrm{cov}(\hat q) = \sum_y \pi_q(y) F_y(\hat q) = G_{\hat\pi}(\hat q) - \langle \hat\pi_q - \pi_q, F(\hat q)\rangle.$$

For the estimation error, the linear system perturbs as $\hat\pi_q - \pi_q = \hat C^{-1}[(\hat\mu_q - \mu_q) - (\hat C - C)\pi_q]$, and $\|\hat C^{-1}\|_2 \le 1/(\sigma_{\min}(C) - \|\hat C - C\|_2) \le 2/\sigma_{\min}(C)$ under the stated regularity condition. For failure probability ζ, bounded-vector concentration gives $\|\hat\mu_q - \mu_q\|_2 \le (1 + \sqrt{2\log(2/\zeta)})/\sqrt{M}$. Column-wise averaging gives $\mathbb{E}\|\hat C - C\|_F^2 \le \sum_k 1/(N\pi_c(k)) = K^2/N$ for a uniform calibration prior. Combining with $\|\cdot\|_1 \le \sqrt{K}\|\cdot\|_2$ gives the stated rate. □

B.11 PRIOR-ESTIMATION ERROR AND STABILITY

This section supports Sec. 3.4 by relating BBSE estimation error to coverage and the stability criterion.

The proof in Theorem 2 identifies prior correction as the direction opposite to the pooled response. However, the query prior needed for that correction is unobserved. We characterize the coverage obtained when an estimated prior replaces it. We keep the score map fixed for this calculation and use label-shift weighting to select the population threshold (Podkopaev & Ramdas, 2021). This weighting changes each class's contribution to calibration by its query-to-calibration prior ratio.

To estimate the unobserved prior, we use black-box shift estimation (BBSE) (Lipton et al., 2018). For any fixed probability map, its soft confusion matrix C is defined by the same class-conditional averages used above. Here p denotes that fixed map. Label shift preserves these averages and gives the linear relation

$$\mu_q = C\pi_q, \qquad \hat{C}_{\cdot y} = \frac{1}{N_y} \sum_{i \le N : y_i = y} p(x_i), \qquad \hat{\mu}_q = \frac{1}{M} \sum_{i=N+1}^{N+M} p(x_i), \tag{9}$$

where μ_q is the population mean query probability vector and N_y counts calibration examples of class y. The empirical quantities $\hat{C}$ and $\hat{\mu}_q$ supply the labeled and unlabeled sides of the relation. BBSE recovers the prior by solving this system, which suits our setting because query labels are unavailable. Let $\sigma_{\min}(C)$ denote the smallest singular value of C. A larger value means that different class mixtures produce more distinguishable average predictions.

Appendix B.10 subtracts the estimated-prior calibration equation from the true-prior coverage identity. It then controls the linear-system perturbation to obtain Eq. (3). With an exact prior, Eq. (2) recovers the nominal population coverage of prior-weighted calibration. With an estimated prior, undercoverage occurs when the estimate overweights classes with higher conditional coverage on balance. The identity therefore supplies a signed quantity for choosing how far to adapt. The rate supplies a complementary measure of estimation stability. Its perturbation condition keeps the empirical inverse controlled by the population inverse. In the class-symmetric model, a_C also determines identifiability on $\mathbf{1}^{\perp}$. Our predictor uses the estimated matrix's singular value for stability and an empirical version of the coverage offset to limit adaptation.

B.12 THEOREM 4 (LEVEL AND POWER OF THE PRETEST)

To distinguish prior change from sampling variation, we compare predicted-label histograms with a two-sample Pearson χ^2 permutation pretest. The label map is fitted independently of the tested pool. For an exchangeable pool, a rank-based permutation test has rejection probability at most its level α_0 for any finite number B_{perm} of permutations. Theorem 4 in Appendix B.12 also relates local power to prior mismatch and effective sample size through the hard confusion matrix. This detection criterion complements estimation stability when selecting prior correction.

This section supports the shift-detection step of Algorithm 1 through the level and local power statements of Theorem 4.

Theorem 3 identifies the sampling error introduced when a prior is estimated. It leaves open whether the observed query batch provides evidence of a prior change worth correcting. We address this question with a permutation pretest on predicted-label frequencies. Permuting membership in the calibration and query subsets provides a reference distribution under exchangeability.

Let $h(x) = \arg\max_k p_k(x)$ be a predicted-label map fitted independently of the tested pool. We compute the two-sample Pearson χ^2 statistic G_{obs} from its calibration and query histograms. Let G_j be the same statistic after permutation j, for a total of B_{perm} permutations. The level $\alpha_0 \in (0, 1)$ controls the allowed probability of detecting shift under the null. For the power calculation, define the hard confusion matrix by $C^h_{ky} = P(h(X) = k \mid Y = y)$. It maps true class proportions to predicted-label proportions.

Theorem 4 (Level and power of the pretest)**.** *Suppose the label map is fitted on data independent of the pool. Under exchangeability of the pool, the permutation probability is*

$$\hat{p} = \frac{1 + \#\{j \le B_{\text{perm}} : G_j \ge G_{\text{obs}}\}}{1 + B_{\text{perm}}},$$

where # counts permuted statistics at least as large as the observed statistic. It satisfies $P(\hat{p} \leq \alpha_0) \leq \alpha_0$ for every finite B_{perm}, K, N and M. For the independent multinomial histograms under local alternatives, the statistic has limiting law

$$G \Rightarrow \chi^2_{K-1}(\nu), \qquad \nu \geq \sigma_{\min}(C^h)^2 \|\delta\|_2^2 \frac{NM}{N+M},$$

where ν is the noncentrality parameter and $K-1$ is the number of degrees of freedom. A larger ν separates the alternative distribution further from the central χ^2 null distribution.

The proof below establishes the level statement from permutation ranks and derives the power expression from predicted-label frequencies. The independent label map preserves exchangeability under the null. The power expression connects detectability to the mismatch magnitude and the effective sample size $NM/(N+M)$. The hard confusion matrix describes how much of the true prior change is visible in predicted labels. At $\delta = 0$, the population histograms agree and the local limiting law is central. These results give the stability and detection criteria that Sec. 4 combines with separate transport fitting to construct the prior-aware predictor.

Power in full. Under local alternatives $G \Rightarrow \chi^2_{K-1}(\nu)$ with $\nu = \frac{NM}{N+M} \sum_j (\rho_q(j) - \rho_c(j))^2 / \bar{\rho}(j)$, where ρ_c and ρ_q are the predicted-label frequencies of the two subsets and $\bar{\rho}$ their pooled frequency. Since $\|\rho_q - \rho_c\|_2 \geq \sigma_{\min}(C^h)\|\delta\|_2$ for the hard confusion matrix C^h, the non-centrality is at least $\sigma_{\min}(C^h)^2 \|\delta\|_2^2 \, NM/\big((N+M)\max_j \bar{\rho}(j)\big)$, and since $\max_j \bar{\rho}(j) \leq 1$ this is at least the bound stated in Theorem 4. The power therefore increases with $\|\delta\|_2 \sqrt{NM/(N+M)}$, where $\|\delta\|_2$ lies between $2\text{TV}(\pi_q, \pi_c)/\sqrt{K}$ and $2\text{TV}(\pi_q, \pi_c)$.

The trade-off. It is not symmetric between the two things a conformal predictor is judged by. Adapting shrinks sets from the smallest mismatches onward, while the coverage it recovers outweighs its own estimation noise only at larger ones. A pretest at a fixed level therefore gives up the small efficiency gain below its detection point and leaves the predictor unchanged when no shift is detected.

Proof. Under exchangeability of the pool the vector of predicted labels is distributionally invariant under permutations, so G_{obs} is exchangeable with $G_1, \ldots, G_{B_{\text{perm}}}$ and the permutation p-value stochastically dominates the uniform distribution on $\{1/(B_{\text{perm}}+1), \ldots, 1\}$, giving $P(\hat{p} \leq \alpha_0) \leq \alpha_0$ for every finite B_{perm}. The argument needs the label map to be fixed before the pool is formed, which holds when it is fitted on data outside the pool, as in the split-sample predictor of Sec. 4. If the map is fitted on the calibration subset of the pool itself, the predicted labels are no longer exchangeable and the level statement does not apply.

For power, $h^c \sim \text{Mult}(N, \rho_c)$ and $h^q \sim \text{Mult}(M, \rho_q)$ are independent, and under local alternatives the two-sample Pearson statistic converges to a non-central $\chi^2_{K-1}(\nu)$ with the stated ν. Since $\rho_q - \rho_c = C^h \delta$, we have $\|\rho_q - \rho_c\|_2 \geq \sigma_{\min}(C^h)\|\delta\|_2$, and therefore $\nu \geq \sigma_{\min}(C^h)^2 \|\delta\|_2^2 \, NM/((N+M)\max_j \bar{\rho}(j))$. Since $\|\delta\|_1 = 2\text{TV}$ and $\|\delta\|_2 \leq \|\delta\|_1 \leq \sqrt{K}\|\delta\|_2$, this is within a factor $\sqrt{K}$ of $\text{TV}(\pi_q, \pi_c)^2 \, NM/(N+M)$. □

B.13 The label-leakage remark

This calculation supports the full-calibration comparison in Sec. 4. It describes how using the same labels for score fitting and threshold calibration changes their coverage.

With $m = \hat{\pi}_c$ the same linearization as Step 1 of Appendix B.6, with the right-hand side moved from π_c to $\hat{\pi}_c$, gives $J\Delta b = \Delta m$ where $\Delta m = \hat{\pi}_c - \pi_c$. The change in a calibration score is proportional to $-\Delta b_{y_i}$, whose empirical mean is $\langle \hat{\pi}_c, J^+ \Delta m \rangle$, while the query side sees $\langle \pi_c, J^+ \Delta m \rangle$. The difference is $\langle \Delta m, J^+ \Delta m \rangle \geq 0$ for every realization because $J^+ \succeq 0$. Calibration scores and $\hat{q}$ are systematically smaller. Consequently, the query side under-covers. Taking expectations with $\mathbb{E}\|\Delta m\|_2^2 = (1 - \|\pi_c\|_2^2)/N$ and $a_J^{-1} \asymp K$ for a diffuse posterior gives a deficit of order K/N. The constant that converts a log-odds advantage into a coverage difference is a local density factor which we do not compute in closed form.

Statement	Check
Lemma 1(a)	elementwise $\lvert Q^* - \mathrm{softmax}(\ell/\tau + \log r)\rvert \le 8 \cdot 10^{-16}$ at $T = 3$
Lemma 1(b)	fixed-point residual 3.6–8.0 $\cdot 10^{-3}$ at T=3, K=10. $6.6 \cdot 10^{-5}$–$1.8 \cdot 10^{-4}$ at K=100. $\sim 10^{-16}$ at T=200
Lemma 2, identity	signed difference (closed form minus empirical) ≤ 0.0044
Lemma 2, bound	holds in 100% of repetitions with slack 7.2× to 476× under random shift directions
Lemma 2, sharpness	equality to $5.6 \cdot 10^{-17}$ on 600 random F with δ supported on $\arg\max F$ and $\arg\min F$
Thm. 1, sign	loss negative in 97% (K=10) and 100% (K=100) of repetitions. Split conformal negative in 38%–55%
Thm. 1, rate	log–log slope 2.024 (K=10) and 1.872 (K=100). $\lVert\delta\rVert_2^2$ fits better than $\lVert\delta\rVert_1^2$ (R^2 0.917 vs 0.872, 0.960 vs 0.821)
Thm. 1, λ	slope 0.966, intercept +0.0008 at K=10 in the small-shift regime. It saturates to 0.76 at the largest shift
Thm. 2, size	pooled step larger than the oracle in 100% and larger than the per-split variant in 95–100% of repetitions
Thm. 2, reversal	located at $\mathrm{TV}^\dagger = 0.369$ (K=10) and 0.400 (K=100). moves right when λ falls, by 1.26× and 1.33× against a predicted 2.5×
Thm. 3, identity	residual $\le 1.13 \cdot 10^{-2}$, of the order of the $O(1/\sqrt{M})$ sampling noise
Thm. 3, rate	holds in 100% with slack 9.5×–38×. The regularity condition $\lVert\hat{C} - C\rVert_2 \le \sigma_{\min}(C)/2$ is marginally violated at K=100
Thm. 4, level	4.0% (K=10) and 6.0% (K=100) against a nominal 5% over 100 repetitions
Thm. 4, power	non-central χ^2 formula matches the observed detection rate to 0.8–1.8 points on average
Leakage	deficit negative in 100% of repetitions at K=100. regression of the deficit on the predicted quantity has $R^2 = 0.949$ and intercept -0.002

Table 4: **Numerical checks of the statements of Sec. 3.** Synthetic model, two worlds, twenty repetitions each unless stated.

B.14 Numerical checks

The finite-iteration comparison also qualifies the converged constraint in Lemma 1. At the published $T = 3$, the matching residual is of order 10^{-2} for $K = 10$. The numerical accuracy comparison only partly reproduces the $1/\lambda$ scaling under (H5).

These controlled comparisons support the theoretical identities and scaling relationships in Sec. 3.

We evaluate the statements above with a controlled synthetic model. It directly generates class prototypes, feature noise and shift. Thus π_q, F_y and the three matrices are all available. Two worlds are used, $K = 10$ with $N = M = 1000$ and $K = 100$ with $N = M = 3000$, each over twenty repetitions. Table 4 collects the outcome.

The fixed-point characterization is exact in the limit of many Sinkhorn iterations. At the published $T = 3$, its residual is of order 10^{-2} when $K = 10$. This residual sets the accuracy floor of the first-order analysis at small K. The sharp bound of Lemma 2 is attained in the adversarial direction. Under Dirichlet directions it is loose by one to two orders of magnitude. The quadratic law therefore provides the relevant comparison for these directions. The ordering between the pretest threshold and the onset of adaptation gains is measured empirically. Set size improves below the smallest tested shift. Coverage gains turn positive at $\mathrm{TV} \approx 0.39$ ($K = 10$) and 0.57 ($K = 100$). The pretest detects shift at the intermediate values $\mathrm{TV} \approx 0.10$ and 0.08. The pretest is therefore a compromise between two criteria, and we state it that way.

					Median per class		Classes with < 9	
Dataset	K	N	N_B	M	Full	B	Full	B
EuroSAT	10	4050	2027	4050	412.5	206.5	0	0
OxfordPets	37	1832	918	1837	50	25	0	0
DTD	47	846	423	846	18	9	0	0
Caltech101	100	1213	636	1252	8.5	4.5	50	92
FGVCAircraft	100	1633	833	1700	16	8	0	67
Flowers102	102	1209	629	1254	10	5	42	85

Table 5: **Calibration and query sample sizes.** N is the full calibration size, N_B is the size after the inner stratified split, and M is the query size, including resampled instances. Counts are fixed by the class sizes across the 20 repetitions.

C Experimental Details

C.1 Sampling, sample sizes and predictors

The controlled comparisons in Sec. 5.1 change the class mixture while drawing calibration and query images from the same within-class source. Each repetition first makes a stratified 50/50 split of the test images into calibration and a held-out query pool. For finite β, we draw $\pi_q \sim \mathrm{Dir}(\beta \mathbf{1}_K)$ and query counts from $\mathrm{Mult}(M, \pi_q)$, following Veilleux et al. (2021); Garg et al. (2023). A class quota is filled without replacement when its pool is large enough and with replacement otherwise. The unresampled query half is denoted by $\beta = \infty$. For split-sample predictors, calibration is subsequently divided within classes into fitting half A and calibration half B. These two splits have different purposes: the first defines the evaluation distribution, and the second separates fitting from calibration.

Table 5 gives the resulting sample sizes. The six datasets are EuroSAT (Helber et al., 2019), OxfordPets (Parkhi et al., 2012), DTD (Cimpoi et al., 2014), Caltech101 (Fei-Fei et al., 2004), FGVCAircraft (Maji et al., 2013) and Flowers102 (Nilsback & Zisserman, 2008). We use the released CoOp/CLIP-Conformal test partitions and class orders (Zhou et al., 2022; Silva-Rodríguez et al., 2025), with the official test split for FGVCAircraft and the background category excluded from Caltech101. At $\alpha = 0.10$, class-conditional calibration needs at least nine examples for a finite threshold. The last two columns therefore quantify the data limitation behind its large sets, separately for full calibration and half B.

Models and score sources. The zero-shot grid uses CLIP ViT-B/32, ViT-B/16, ViT-L/14 and ViT-L/14-336 (Radford et al., 2021), MetaCLIP-B/16 (Xu et al., 2024), EVA02-L/14 (Fang et al., 2024) and SigLIP-SO400M-384 (Zhai et al., 2023). SigLIP is unavailable on FGVCAircraft, giving 41 dataset–backbone configurations. All predictors in a configuration share cached logits, prompt templates and evaluation splits. For adapted scores, 16-shot CoOp (Zhou et al., 2022) and Tip-Adapter-F (Zhang et al., 2022) use ViT-B/16 with a fixed visual tower and three training seeds each, giving 36 dataset–method–seed groups. Their support images come from training data and are disjoint from calibration and query images. Food101 (Bossard et al., 2014) is reserved for the additional score-source comparison. ImageNet validation (Deng et al., 2009) supplies the first $K \in \{1000, 200, 100\}$ classes in canonical order for the class-count ladder, with $\beta \in \{10, 3, 1, 0.3, 0.1\}$.

Predictor definitions. SCP uses raw zero-shot probabilities at $\tau = 1$. Conf-OT uses three Sinkhorn iterations on pooled calibration and query logits, targeting the empirical calibration prior. Per-split fits transport on A, then fits its temperature on A, and uses B to select the SCP threshold. The full-calibration variant instead fits and calibrates on all N examples. Label-conditional uses the raw probabilities and the classwise order statistic $\lceil (n_k + 1)(1 - \alpha) \rceil$, with an infinite threshold when this exceeds n_k. Soft-Mondrian uses classes as subpopulations and a full-calibration fitted probability map for both the LAC scores and soft membership weights (Appendix D.1.1). Always-on retains the pretest and otherwise uses the unshrunk least-squares prior, omitting stability selection and the cap. Ours follows Algorithm 1. Oracle replaces the estimated prior by the realized query prior in probability correction and weighted calibration on the same split-sample map. It measures the benefit of known class proportions at its own achieved coverage, distinct from the population equal-coverage optimum in Theorem 2.

		Fixed thresholds		LODO thresholds	
Backbones	β	Cov	Size	Cov	Size
All	1.0	0.897	3.47	0.896	3.47
All	0.3	0.902	3.39	0.901	3.39
All	0.1	0.910	3.30	0.908	3.30
Without ViT-B/16	1.0	0.897	3.18	0.896	3.18
Without ViT-B/16	0.3	0.904	3.10	0.903	3.10
Without ViT-B/16	0.1	0.912	3.02	0.910	3.02

Table 6: **Threshold transfer with a matched reference.** LODO selects thresholds on five datasets and evaluates the omitted sixth. Fixed thresholds are those used by Ours in Table 1. Means cover 41 configurations per strength, or 35 after excluding ViT-B/16, with 20 repetitions each.

C.2 Aggregation and uncertainty

Each configuration and shift condition has 20 sampling repetitions. We average repetitions first, then weight configurations equally. Consequently, the zero-shot aggregate weights 41 dataset–backbone combinations equally, with six combinations for FGVCAircraft and seven for each other dataset. The three finite Dirichlet strengths give 123 shifted configurations. Adapted scores similarly give 36 groups and 108 shifted configurations, each including a specified training seed.

Coverage is the fraction of query sets containing the true label, set size is the mean number of labels, and top-1 accuracy uses each predictor's own decision rule. The empirical loss comparison is $\Delta\text{cov} = \text{cov}(Conf - OT) - \text{cov}(\text{SCP})$, paired on the same queries. It differs from the population offset relative to $1-\alpha$ in Eq. (1). Configuration counts below 0.85 describe mean shortfalls exceeding five percentage points. The boundary 0.885 is instead the lower edge of the 0.90 ± 0.015 comparison band in Table 1. These summaries describe empirical performance at the stated aggregation level.

Standard errors for the paired differences in Sec. 5.2 are the sample standard deviation of 41 paired configuration means divided by $\sqrt{41}$. The reported two-sided 95% bootstrap intervals use 10,000 resamples of the 20 repetition indices, sharing each index draw across configurations and paired methods. They quantify split-seed variation on the fixed configuration grid. The one-sided 99% bounds for pooled loss use the 0.99 quantile of 1,000 such seed-block bootstrap means. For adapted scores these bounds average over the 54 shifted configurations of each method, and for Food101 over its nine configurations per method. They are bounds on aggregate mean loss rather than simultaneous bounds on every configuration. The separate bootstrap procedures used inside the predictor are specified in Appendix G.

C.3 Selection settings and transfer

LAC is evaluated at $\alpha = 0.10$. The pretest uses 200 pool permutations at level $\alpha_0 = 0.05$, and the cap allows an estimated coverage decrease of 0.02 from the uncorrected predictor. Five-fold query likelihood selects shrinkage strength, with bootstrap and grid details in Appendix G. The two stability thresholds are selected on ViT-B/16 zero-shot scores. On this backbone, Always-on coverage averages within 0.03 of nominal on OxfordPets, Caltech101 and EuroSAT, and more than 0.12 below nominal on DTD, Flowers102 and FGVCAircraft. For each stability statistic, the threshold is the geometric mean of the largest value in the first group and the smallest value in the second, over repetitions and shift strengths. This gives $\theta_B = 4493$ and $\theta_D = 0.694$, transferred unchanged to other backbones, adapted scores and ImageNet.

To examine transfer beyond the datasets used for threshold selection, we also select thresholds on five datasets and evaluate the omitted sixth. Table 6 reports this leave-one-dataset-out (LODO) comparison, including results that exclude the threshold-selection backbone ViT-B/16. The matched fixed-threshold and LODO columns show similar aggregate coverage and size at each strength. The lowest LODO configuration mean is 0.8496 on Caltech101 with ViT-B/16 at $\beta = 0.3$, compared with a minimum of 0.853 under the fixed thresholds. This comparison establishes similar aggregate transfer while showing a change in the lowest-coverage configuration.

Computational work. Evaluation uses cached logits and requires no model gradients. Transport performs three iterations on a sample-by-class matrix, followed by one-dimensional temperature

	$\beta = \infty$		$\beta = 1.0$		$\beta = 0.3$		$\beta = 0.1$	
Method	Cov ↑	Size ↓	Cov ↑	Size ↓	Cov ↑	Size ↓	Cov ↑	Size ↓
Ours, full calibration	*.898*	*3.51*	*.896*	*3.43*	*.901*	*3.32*	*.909*	*3.22*
Oracle	*.900*	*3.54*	*.905*	*2.83*	*.911*	*2.03*	*.923*	*1.41*

Table 7: **Full-calibration and known-prior comparisons.** Coverage and mean set size use the same 41 dataset–backbone combinations and 20 repetitions as Table 1 at $\alpha = 0.10$. The full-calibration variant reuses all calibration examples for fitting and threshold selection. Oracle uses the realized query prior. These references are excluded from the ranking in Table 1.

Score map and calibration sample	Coverage	Set size
Raw scores, full calibration	0.934	20.85
Transport, full calibration	0.934	20.71
Transport, half B	0.943	42.61

Table 8: **Classwise calibration and sample allocation.** Label-conditional LAC at $\beta = 0.1$ and $\alpha = 0.10$, averaged over 41 configurations and 20 repetitions. Each arm uses the same classwise finite-sample correction.

fitting. Prior selection adds simplex-constrained solves for the five query folds and full batch, while the stability and cap bootstraps repeatedly resample both B and the query batch and solve their respective prior-estimation problems. The pretest uses 200 histogram comparisons. Thus a single prior solve describes only one subroutine, with the repeated-solve cost determined by the bootstrap counts and selection grids in Appendix G.

D Results by Dataset and Backbone

This appendix supports the aggregate comparisons in Sec. 5 with results for individual datasets and score sources.

Table 12 gives the coverage loss of the pooled predictor relative to split conformal prediction on every zero-shot configuration, at the three shift strengths, together with the split-conformal coverage under the equal split and the zero-shot top-1 accuracy. Table 13 does the same for the 36 adapted groups, and Table 14 lists the identifiability statistic per group.

D.1 Coverage and component comparisons

D.1.1 Class-specific calibration and Soft-Mondrian

The large class-conditional sets in Table 1 are consistent with the limited classwise counts in Table 5. Table 8 compares raw and transported scores at $\beta = 0.1$. Halving the calibration sample increases classwise thresholds and roughly doubles the mean set size. All variants retain the same classwise finite-sample correction, so fewer examples per class can outweigh improved scores.

Soft-Mondrian (Wang et al., 2025) uses the same full-calibration fitted transport and temperature map for LAC scores and soft class memberships. For each query, its threshold is the smallest t satisfying

$$\sum_k p_k(x) \frac{\#\{i : Y_i = k,\ s_i \le t\}}{n_k + 1} \ge 1 - \alpha.$$

If no finite calibration score satisfies the condition, the prediction set contains every label. Table 9 displays the achieved coverage–size trade-off and paired intervals. For numerical sensitivity, we evaluate the same threshold formula in double precision after centering each class CDF at $1 - \alpha$ before the weighted sum. This handles attainable boundary values such as $9/10$ without subtracting nearly equal mixture sums. The resulting set-size difference is small relative to the separation between Soft-Mondrian and Ours, leaving the qualitative comparison unchanged.

Soft-Mondrian variant	Cov	Size	Δ coverage [95% CI]	Δ size [95% CI]
Main implementation	0.940	21.95	-0.030 [-0.036, -0.025]	-18.65 [-20.07, -17.24]
Centered double precision	0.940	22.00	-0.030	-18.70 [-20.10, -17.30]

Table 9: **Soft-Mondrian comparison and numerical sensitivity.** Means over 41 configurations and 20 repetitions at $\beta = 0.1$. Differences are Ours minus the stated Soft-Mondrian variant, on shared splits. Both Soft-Mondrian variants use the same full-calibration fitted map. Intervals are paired seed-block bootstrap intervals.

β	Selected repetitions	Per-split	Ours	Oracle	Recovered
1.0	389/820	1.53	1.40	1.33	65.4%
0.3	397/820	1.53	1.23	1.12	73.1%
0.1	406/820	1.46	1.13	1.00	72.3%

Table 10: **Known-prior efficiency on stability-passing repetitions.** Set sizes use the same selected configuration–repetition pairs for all three predictors. Recovered is the ratio of summed size reductions from Per-split, as defined in Appendix D.1.2. Oracle uses the realized prior and removes shrinkage and the cap.

D.1.2 Efficiency comparisons and known priors

Table 11 quantifies the paired main comparison. Table 7 compares the split-sample predictor with full-calibration fitting and the known-prior reference. Relative to full calibration, split-sample fitting adds 0.040, 0.065 and 0.077 labels at $\beta = 1.0, 0.3, 0.1$, with coverage increases below 0.002. Under matched priors its size cost is 0.021 labels. At $\beta = 0.1$, the total size reduction relative to SCP decomposes into 20.2 percentage points from Per-split and 3.9 further points from correction, using the SCP mean size as the common denominator.

To measure how much known-prior efficiency is recovered, let $\mathcal{I}_\beta$ contain the configuration–repetition pairs passing stability at strength β, and let S_i^m be mean set size for method m on pair i. We report

$$\frac{\sum_{i\in\mathcal{I}_\beta}(S_i^{\text{Per-split}} - S_i^{\text{Ours}})}{\sum_{i\in\mathcal{I}_\beta}(S_i^{\text{Per-split}} - S_i^{\text{Oracle}})}.$$

Table 10 reports the selected sample counts, mean sizes and this ratio at each strength. It compares Ours with an information-advantaged reference on the same selected repetitions.

Oracle has different achieved coverage and also removes shrinkage and the cap, so the remaining size difference includes both prior information and correction-strength selection.

The complete configuration comparisons below supplement the coverage distributions and component results in Secs. 5.2 and 5.3.

E Published Evaluation with Alternative Scores

This appendix supports the baseline comparison in Sec. 5.1 with measurements under the published equal-split evaluation.

Table 16 reports our measurements under the equal-split evaluation design. We use the released implementation of the split, conformal predictors and transductive step. The results use ViT-B/16 at $\alpha = 0.10$ and are medians over 20 repetitions. Against the values reported by Silva-Rodríguez

Metric	Ours minus SCP [95% CI]	Configuration SE
Coverage	0.013 [0.006, 0.020]	0.003
Set size	-1.05 [-1.11, -0.98]	0.18

Table 11: **Paired main-comparison uncertainty at $\beta = 0.1$.** Equal-weight means over 41 configurations and 20 repetitions. Intervals resample shared evaluation seeds, while SE measures variation of the paired configuration means divided by $\sqrt{41}$. These summarize different sources of variation (Appendix C.2).

Cell	SCP cov ($\beta{=}\infty$)	Δcov $\beta{=}1.0$	Δcov $\beta{=}0.3$	Δcov $\beta{=}0.1$	top-1
aircraft / EVA02-L/14	.898	−.0001	−.018	−.066	.359
aircraft / MetaCLIP-B/16	.900	−.005	−.010	−.054	.304
aircraft / ViT-B/16	.899	−.005	−.002	−.043	.250
aircraft / ViT-B/32	.900	+.003	+.004	−.027	.190
aircraft / ViT-L/14	.897	−.005	−.018	−.066	.325
aircraft / ViT-L/14-336	.897	−.006	−.021	−.068	.333
caltech / EVA02-L/14	.904	−.174	−.354	−.567	.975
caltech / MetaCLIP-B/16	.900	−.119	−.278	−.498	.965
caltech / SigLIP-SO400M	.900	−.145	−.322	−.547	.981
caltech / ViT-B/16	.904	−.139	−.267	−.467	.932
caltech / ViT-B/32	.900	−.111	−.223	−.450	.913
caltech / ViT-L/14	.904	−.181	−.344	−.549	.952
caltech / ViT-L/14-336	.903	−.163	−.329	−.532	.945
dtd / EVA02-L/14	.899	−.015	−.039	−.126	.619
dtd / MetaCLIP-B/16	.900	−.015	−.057	−.156	.613
dtd / SigLIP-SO400M	.904	−.029	−.056	−.133	.649
dtd / ViT-B/16	.899	−.007	−.023	−.056	.439
dtd / ViT-B/32	.898	−.012	−.028	−.072	.425
dtd / ViT-L/14	.899	−.012	−.031	−.100	.534
dtd / ViT-L/14-336	.901	−.010	−.035	−.109	.543
eurosat / EVA02-L/14	.899	−.017	−.063	−.072	.674
eurosat / MetaCLIP-B/16	.899	−.019	−.080	−.101	.497
eurosat / SigLIP-SO400M	.899	−.023	−.074	−.093	.569
eurosat / ViT-B/16	.899	−.020	−.081	−.080	.483
eurosat / ViT-B/32	.899	−.008	−.080	−.072	.452
eurosat / ViT-L/14	.899	−.011	−.059	−.075	.603
eurosat / ViT-L/14-336	.898	−.011	−.067	−.083	.615
flowers / EVA02-L/14	.898	−.026	−.064	−.150	.772
flowers / MetaCLIP-B/16	.900	−.022	−.042	−.067	.753
flowers / SigLIP-SO400M	.904	−.084	−.195	−.404	.929
flowers / ViT-B/16	.903	−.020	−.025	−.055	.709
flowers / ViT-B/32	.901	−.018	−.026	−.056	.667
flowers / ViT-L/14	.903	−.034	−.060	−.146	.796
flowers / ViT-L/14-336	.903	−.035	−.064	−.140	.792
pets / EVA02-L/14	.899	−.045	−.136	−.292	.938
pets / MetaCLIP-B/16	.900	−.056	−.157	−.318	.918
pets / SigLIP-SO400M	.901	−.045	−.146	−.355	.965
pets / ViT-B/16	.900	−.059	−.126	−.281	.890
pets / ViT-B/32	.899	−.045	−.106	−.240	.874
pets / ViT-L/14	.898	−.048	−.134	−.277	.934
pets / ViT-L/14-336	.899	−.057	−.148	−.304	.937

Table 12: **Zero-shot configurations.** Mean over 20 repetitions. $\Delta\mathrm{cov} = \mathrm{cov}(\textsc{Conf}\text{-OT}) - \mathrm{cov}(\mathrm{SCP})$ on the same split. The only two non-negative entries are FGVCAircraft with ViT-B/32 at the two weaker strengths.

et al. (2025) for the same six datasets and backbone, our coverage differs by at most 0.007 (mean 0.003). Our set sizes differ by at most 0.32 (mean 0.07). Our top-1 differs by at most 0.51 points (mean 0.14). The table contains our measurements because the reported values use different CLIP implementations and random splits. The rightmost column gives the standard deviation over our 20 repetitions to contextualize the observed differences.

Our CLIP weights use a converted checkpoint and manually matched image preprocessing. These implementation differences affect top-1 accuracy and the resulting conformal scores. The random number stream also differs from the published evaluation, so the 20 splits differ. The last column quantifies variation across these splits. APS and RAPS include uniform score randomization, which introduces further variation between realizations.

F Coverage and Prior Identifiability

This appendix supports the coverage and component comparisons in Secs. 5.2 and 5.3. It relates their behavior to class identifiability and sampling.

Where the loss is largest. Figure 4 separates the coverage distributions by dataset and strength, while Table 17 reports the aggregate paired loss and its one-sided upper bound. The largest losses occur on Caltech101 and OxfordPets, and all dataset-level upper bounds remain below zero. Their ordering matches zero-shot top-1 accuracy in this suite. This is an empirical association with score quality, whereas Theorem 1 specifies the response through a_C, a_D, a_J.

Group	SCP cov ($\beta{=}\infty$)	Δcov $\beta{=}1.0$	Δcov $\beta{=}0.3$	Δcov $\beta{=}0.1$	top-1
aircraft CoOp s0	.898	−.008	−.025	−.068	.400
aircraft CoOp s1	.902	−.010	−.033	−.074	.400
aircraft CoOp s2	.899	−.006	−.030	−.066	.415
aircraft Tip-F s0	.898	−.008	−.028	−.065	.438
aircraft Tip-F s1	.901	−.008	−.030	−.076	.433
aircraft Tip-F s2	.899	−.011	−.030	−.077	.437
caltech CoOp s0	.898	−.074	−.193	−.350	.959
caltech CoOp s1	.900	−.085	−.222	−.392	.955
caltech CoOp s2	.900	−.090	−.194	−.378	.957
caltech Tip-F s0	.899	−.084	−.189	−.371	.959
caltech Tip-F s1	.899	−.080	−.186	−.361	.954
caltech Tip-F s2	.900	−.079	−.190	−.359	.955
dtd CoOp s0	.902	−.021	−.058	−.141	.682
dtd CoOp s1	.899	−.018	−.048	−.127	.690
dtd CoOp s2	.895	−.018	−.055	−.150	.700
dtd Tip-F s0	.899	−.027	−.065	−.174	.701
dtd Tip-F s1	.899	−.025	−.058	−.169	.688
dtd Tip-F s2	.897	−.021	−.062	−.184	.688
eurosat CoOp s0	.899	−.048	−.117	−.147	.798
eurosat CoOp s1	.898	−.046	−.121	−.157	.816
eurosat CoOp s2	.898	−.041	−.118	−.133	.782
eurosat Tip-F s0	.899	−.041	−.113	−.151	.743
eurosat Tip-F s1	.900	−.051	−.111	−.158	.741
eurosat Tip-F s2	.898	−.041	−.114	−.146	.725
flowers CoOp s0	.900	−.107	−.235	−.435	.960
flowers CoOp s1	.902	−.102	−.220	−.437	.964
flowers CoOp s2	.901	−.104	−.236	−.419	.960
flowers Tip-F s0	.901	−.101	−.222	−.405	.972
flowers Tip-F s1	.901	−.102	−.216	−.410	.972
flowers Tip-F s2	.903	−.100	−.229	−.413	.969
pets CoOp s0	.898	−.057	−.132	−.287	.933
pets CoOp s1	.900	−.051	−.139	−.296	.932
pets CoOp s2	.898	−.051	−.132	−.274	.933
pets Tip-F s0	.900	−.056	−.112	−.260	.932
pets Tip-F s1	.899	−.052	−.126	−.270	.932
pets Tip-F s2	.899	−.045	−.108	−.245	.925
Food101 CoOp s0	n/a	−.031	−.097	−.187	.861
Food101 CoOp s1	n/a	−.029	−.087	−.181	.858
Food101 CoOp s2	n/a	−.032	−.098	−.195	.860
Food101 Tip-F s0	n/a	−.033	−.101	−.206	.874
Food101 Tip-F s1	n/a	−.033	−.099	−.210	.873
Food101 Tip-F s2	n/a	−.034	−.099	−.204	.874

Table 13: **Adapted score sources, ViT-B/16.** The six-dataset main batch contains 108 shifted entries. The separate Food101 batch, held out from method design, contains 18 additional shifted entries. All entries in both batches are negative. Food101 has zero-shot top-1 accuracy 0.859.

Group	CoOp	Tip-F	Group	CoOp	Tip-F
aircraft s0	.0002	.0006	eurosat s0	.4954	.3639
aircraft s1	.0001	.0008	eurosat s1	.5311	.3771
aircraft s2	.0002	.0008	eurosat s2	.4406	.3596
caltech s0	.2007	.1536	flowers s0	.6918	.6986
caltech s1	.2299	.1443	flowers s1	.6220	.6926
caltech s2	.2226	.2011	flowers s2	.6505	.7040
dtd s0	.1399	.1536	pets s0	.5113	.3975
dtd s1	.1606	.1630	pets s1	.5203	.3865
dtd s2	.1245	.1350	pets s2	.4782	.3444
Food101 s0 / s1 / s2: CoOp .3410 / .2721 / .2412 and Tip-F .3454 / .3187 / .3110					
Zero-shot reference: aircraft .0001, caltech .0268, dtd .0018, eurosat .0098, flowers .0008, pets .2233, Food101 .3037					

Table 14: **Identifiability statistic $\sigma_{\min}(\hat{C})$ per group**, computed from the full-calibration transport and temperature map, with medians over the three shift strengths. It rises on all 36 groups relative to the corresponding zero-shot map. FGVCAircraft remains of order 10^{-4}, while the other 30 groups show larger increases. On Food101 it rises for all three Tip-Adapter-F seeds but for only one of the three CoOp seeds.

Query coupling against temperature and splitting. Table 18 fits the transport of Conf-OT on the calibration inputs alone and leaves every other choice at its pooled setting. The calibration-only fit returns coverage to within 0.015 of the target at every shifted strength and removes most of the

K	Backbone	Exp.	Lower	Weak exp.	κ_{pred}	κ_{obs}	Ratio	Weak ratio
1000	EVA02-L/14	1.96	1.86	2.18	54.17	16.07	0.297	0.387
1000	MetaCLIP-B/16	2.07	1.95	2.68	37.43	15.79	0.422	0.501
1000	ViT-B/16	1.95	1.85	2.72	31.61	14.32	0.453	0.607
1000	ViT-B/32	2.12	2.00	2.50	25.84	12.78	0.495	0.527
1000	ViT-L/14	2.09	1.99	2.58	46.87	15.20	0.324	0.372
1000	ViT-L/14-336	1.97	1.85	2.33	47.82	15.79	0.330	0.432
200	EVA02-L/14	2.09	1.98	2.77	51.20	8.37	0.164	0.206
200	MetaCLIP-B/16	1.82	1.76	1.81	27.60	6.35	0.230	0.331
200	ViT-B/16	1.92	1.83	2.32	24.02	5.86	0.244	0.352
200	ViT-B/32	2.06	1.92	2.56	16.52	4.88	0.295	0.390
200	ViT-L/14	2.02	1.93	2.66	32.66	6.75	0.207	0.265
200	ViT-L/14-336	1.99	1.89	2.45	38.86	7.13	0.184	0.244
100	EVA02-L/14	1.93	1.81	1.76	41.96	4.29	0.102	0.137
100	MetaCLIP-B/16	1.73	1.62	1.89	19.70	3.20	0.162	0.304
100	ViT-B/16	1.70	1.62	1.36	13.79	3.06	0.222	0.377
100	ViT-B/32	1.94	1.77	1.73	11.99	2.55	0.213	0.318
100	ViT-L/14	2.00	1.90	2.42	25.90	3.28	0.127	0.169
100	ViT-L/14-336	1.97	1.86	2.34	29.72	3.73	0.126	0.176

Table 15: **Growth exponent and loss magnitude.** Exp. fits all five strengths, Lower is its one-sided 95% lower bound, and Weak exp. fits only $\beta = 10, 3$. Logarithmic fits retain positive losses. Ratio and Weak ratio are $\kappa_{\mathrm{obs}}/\kappa_{\mathrm{pred}}$ for the corresponding strengths, fitted using all losses. Each row uses 20 repetitions at each strength.

Dataset	Method	Top-1	Cov	Size	sd (cov / size)
EuroSAT	LAC	48.26	.899	4.13	.0051 / .032
EuroSAT	LAC + OT	58.20	.902	3.03	.0065 / .066
EuroSAT	APS (Romano et al., 2020)	48.26	.898	4.21	.0052 / .040
EuroSAT	APS + OT	58.20	.902	3.48	.0072 / .060
EuroSAT	RAPS (Angelopoulos et al., 2021)	48.26	.899	4.21	.0052 / .040
EuroSAT	RAPS + OT	58.20	.902	3.48	.0073 / .061
DTD	LAC	43.79	.900	10.84	.0113 / .464
DTD	LAC + OT	46.39	.897	8.77	.0122 / .418
DTD	APS	43.79	.902	12.47	.0111 / .631
DTD	APS + OT	46.39	.899	10.06	.0145 / .479
DTD	RAPS	43.79	.901	12.20	.0110 / .509
DTD	RAPS + OT	46.39	.900	9.98	.0138 / .443
Caltech101	LAC	93.33	.899	0.96	.0101 / .012
Caltech101	LAC + OT	92.53	.897	0.93	.0072 / .009
Caltech101	APS	93.33	.902	2.21	.0160 / .143
Caltech101	APS + OT	92.53	.897	1.28	.0150 / .048
Caltech101	RAPS	93.33	.902	2.09	.0160 / .120
Caltech101	RAPS + OT	92.53	.897	1.27	.0150 / .046
Flowers102	LAC	70.89	.900	4.77	.0076 / .276
Flowers102	LAC + OT	75.40	.903	2.95	.0079 / .142
Flowers102	APS	70.89	.896	5.66	.0102 / .298
Flowers102	APS + OT	75.40	.898	3.96	.0128 / .234
Flowers102	RAPS	70.89	.898	5.52	.0089 / .257
Flowers102	RAPS + OT	75.40	.898	3.93	.0121 / .209
OxfordPets	LAC	89.00	.900	1.02	.0086 / .013
OxfordPets	LAC + OT	90.61	.895	0.97	.0098 / .011
OxfordPets	APS	89.00	.901	1.36	.0119 / .038
OxfordPets	APS + OT	90.61	.903	1.34	.0126 / .039
OxfordPets	RAPS	89.00	.902	1.36	.0120 / .038
OxfordPets	RAPS + OT	90.61	.903	1.33	.0125 / .038
FGVCAircraft	LAC	24.91	.899	18.23	.0059 / .495
FGVCAircraft	LAC + OT	27.82	.897	13.95	.0076 / .408
FGVCAircraft	APS	24.91	.898	17.98	.0061 / .533
FGVCAircraft	APS + OT	27.82	.898	14.62	.0077 / .447
FGVCAircraft	RAPS	24.91	.899	17.95	.0061 / .441
FGVCAircraft	RAPS + OT	27.82	.898	14.46	.0075 / .441

Table 16: **The equal-split evaluation design, our measurement.** ViT-B/16, $\alpha = 0.10$, median over 20 stratified 50/50 splits. "+ OT" is the pooled transductive step on the same score. Top-1 is a percentage, coverage and size are on their natural scales.

pooled shortfall. It exceeds the pooled coverage on all but two of the 123 shifted configurations, and mean sets return to their matched-prior size. Per-split in Table 1 differs from this variant by the class-stratified halves and by the temperature refitted on half A. The controlled comparison therefore attributes its coverage recovery to removing query participation. The further Per-split comparison includes both splitting and temperature fitting, while Theorem 2 concerns population bias and set size at equal coverage.

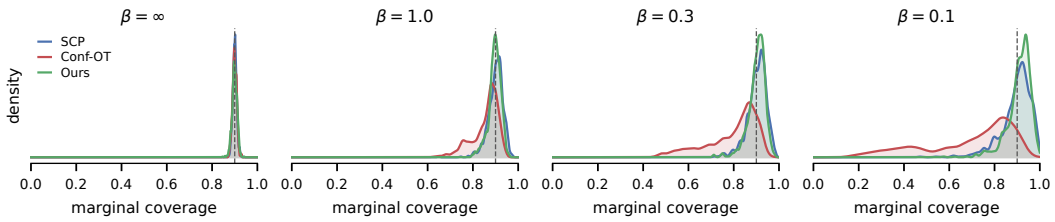

Figure 3: **Coverage as class proportions change.** Each curve pools 820 configuration–repetition coverage values (41 configurations and 20 repetitions) at one shift strength. The dashed line marks the 0.90 target.

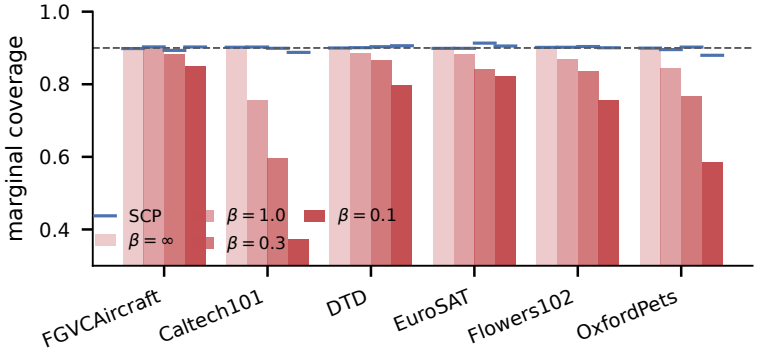

Figure 4: **Per-dataset view of the loss.** Mean coverage of the pooled predictor at the four shift strengths, one group of bars per dataset, with the split-conformal coverage as short horizontal marks and the nominal 0.90 dashed. The dataset averages allow comparison of the loss across score quality and class counts.

Where the correction acts. Table 19 displays pass counts and within-dataset coverage associations for the half-B statistics. The check passes most often on OxfordPets, EuroSAT and Caltech101, and seldom on DTD, Flowers102 and FGVCAircraft. The pretest rejects on every shifted repetition, so this variation comes from the stability check. Passing stability permits correction, while its selected strength can still be zero: one passing repetition on EuroSAT and one on Caltech101 have $\gamma = 0$. Table 20 reports the quantities used for selection, whereas Table 14 describes the separate full-calibration maps.

Low-coverage configurations. Table 2 shows that the corrected predictor avoids the configuration means below 0.85 found for Always-on and Per-split. For Always-on these configurations span FGVCAircraft, DTD, EuroSAT and Flowers102, whereas the Per-split cases are all on Caltech101. The adapted-score comparison has the same ordering of methods (Table 26), with higher minimum coverage for Ours.

Always-on coverage across shift strengths. The pretest detects shift in every shifted repetition, so its rejection frequency does not explain the increase in Always-on coverage across strengths. The Always-on row in Table 1 rises with shift strength, driven most strongly by DTD, EuroSAT and Flowers102. Oracle coverage also rises (Table 7), while its gap from Always-on narrows. Stronger

Dataset	Mean Δcov	One-sided 99% upper bound
Caltech101	−0.322	−0.294
OxfordPets	−0.161	−0.138
Flowers102	−0.083	−0.070
EuroSAT	−0.057	−0.040
DTD	−0.053	−0.042
FGVCAircraft	−0.023	−0.016

Table 17: **Pooled loss by dataset.** Means average over backbones and the three shifted strengths, using 20 evaluation repetitions. Bounds use the seed-block bootstrap of Appendix C.2.

	Pooled fit		Calibration-only fit	
Query prior	Cov	Size	Cov	Size
$\beta = \infty$	.899	3.51	.897	3.50
$\beta = 1.0$	.855	3.82	.888	3.50
$\beta = 0.3$	.796	4.25	.887	3.51
$\beta = 0.1$	.693	4.97	.886	3.44

Table 18: **Removing the query inputs from the transport fit.** Equal-weight means over the 41 dataset–backbone combinations with 20 repetitions at each strength. Both fits use $\tau = 1$ and the whole calibration set for the scaling and for the threshold. They differ only in whether the query inputs enter the transport fit.

Dataset	Configurations	Stability passes	ρ, Always-on	ρ, Per-split
OxfordPets	21	420/420	-0.57	-0.37
Caltech101	21	336/420	-0.70	-0.66
EuroSAT	21	384/420	-0.66	-0.19
DTD	21	43/420	-0.84	0.46
Flowers102	21	9/420	-0.43	-0.27
FGVCAircraft	18	0/360	-0.89	0.32

Table 19: **Stability selection within each dataset.** Pass counts pool 20 repetitions at each of three shifted strengths. Spearman correlations compare configuration means of the half-B stability score and coverage. A passing check permits correction, whose selected strength can still be zero.

imbalance therefore co-occurs with higher coverage for these corrected predictors even as pooled-fit coverage falls.

The statistic inside a dataset. The within-dataset correlations in Table 19 complement the pooled scatter in Fig. 1. The association with Always-on coverage remains negative when the comparison is restricted to backbones and shift strengths within a single dataset.

Correction components. Table 2 compares the correction components with the pretest retained in every variant. Removing the stability check gives higher mean coverage and smaller mean sets than the complete predictor in this aggregate. Removing the bootstrap cap lowers mean coverage. The aggregate hides where the two capped variants differ. Without the stability check, 24 configurations fall below 0.885 and 21 of them lie on FGVCAircraft and DTD, which are the two datasets on which the rate of Eq. (3) is largest. With the check, 9 configurations fall below 0.885 and all of them lie on Caltech101 and Flowers102, where stability selection frequently returns the Per-split sets. This reduction in low-coverage configurations comes with larger aggregate mean sets, as Table 2 shows. Splitting the 123 configurations at the stability threshold θ_B, the variant without the check averages 0.894 coverage on the 57 configurations above the threshold against 0.914 on the 66 below it. Table 20 gives the same comparison by dataset next to the two quantities that enter the rate.

Prior correction without transport. Table 22 keeps the pretest and the shrunk prior and the bootstrap cap and the weighted quantile, and replaces the transported scores by the original probabilities at

				Always-on	Pretest + cap		Ours	
Dataset	K	$\sigma_{\min}(\hat{C})$	Eq. (3)	Cov	Cov	Size	Cov	Size
OxfordPets	37	2.8×10^{-1}	3.0×10^{1}	.902	.910	0.93	.910	0.93
Caltech101	100	2.0×10^{-1}	3.3×10^{3}	.909	.909	0.94	.899	0.92
EuroSAT	10	2.0×10^{-2}	7.9×10^{2}	.854	.921	1.82	.918	1.89
DTD	47	1.0×10^{-2}	2.3×10^{4}	.728	.901	4.14	.899	5.50
Flowers102	102	3.2×10^{-3}	9.7×10^{4}	.835	.907	1.77	.891	2.00
FGVCAircraft	100	1.5×10^{-4}	9.2×10^{5}	.497	.877	8.38	.898	10.01

Table 20: **Correction components by dataset.** Equal-weight means over the shifted zero-shot configurations of each dataset, ordered by $\sigma_{\min}(\hat{C})$. The fourth column is the statistic of Eq. (3) evaluated on half B and the query batch.

Variant	Coverage [95% CI]	Set size [95% CI]
Raw, probability correction and weighting	0.924 [0.918, 0.930]	3.21 [3.09, 3.33]
Raw, candidate-class infinite atom	0.912 [0.906, 0.918]	2.97 [2.85, 3.08]
Raw, weighted quantile only	0.906 [0.900, 0.912]	4.52 [4.20, 4.86]
Transport, probability correction and weighting	0.921 [0.916, 0.926]	2.62 [2.52, 2.70]

Table 21: **Probability correction and calibration weighting.** Matched comparisons at $\beta = 0.1$ over 41 configurations and 20 repetitions, retaining the pretest and cap without stability selection. The candidate variant replaces the maximum-weight infinite atom by the candidate-class weight. Intervals use the paired seed-block procedure.

	Original probabilities		Transported scores	
Query prior	Cov	Size	Cov	Size
$\beta = \infty$	.899	4.38	.899	3.53
$\beta = 1.0$	.895	3.93	.892	3.14
$\beta = 0.3$	.907	3.63	.901	2.83
$\beta = 0.1$	.924	3.21	.921	2.62

Table 22: **Prior correction with and without transport.** Both arms use the pretest and the shrunk BBSE prior and the bootstrap cap and the weighted quantile of Sec. 4.3, over the same 41 combinations with 20 repetitions and shared splits. The right pair reports the *Pretest + cap* arm of Table 2 at each query prior, and its three shifted columns average to that row.

$\tau = 1$. Coverage matches the transported arm within 0.006 at each strength, while mean sets grow by 0.60 to 0.86 labels. The growth holds on 116 of the 123 shifted configurations. Under matched priors this arm uses 4.38 labels against 4.42 for SCP, showing little size gain from prior estimation in this matched-prior comparison. To distinguish probability correction from quantile weighting, Table 21 compares four variants on the same shifted configurations. Using the candidate-class weight for the infinite atom produces smaller sets at lower achieved coverage. Weighting the quantile alone produces larger sets despite lower coverage, showing the efficiency contribution of correcting the class probabilities.

Rank agreement across score sources. Table 26 reports loss-rank agreement with zero-shot scores alongside coverage summaries. Score adaptation preserves part of the ordering of which dataset–shift combinations lose more coverage, despite changing both loss magnitude and prior identifiability.

Duplicate fraction on the ladder. A skewed query prior can request more images of a class than the held-out half contains, so excess quotas use replacement. Table 27 reports image reuse and positive-loss selection for each strength. The weakest shifts have the fewest duplicates and form the restricted fit in Table 15. The strongest shift provides a stress test with more extensive image reuse.

G Algorithm

This appendix supports Sec. 4 by specifying the subroutines of Algorithm 1 and the independent-splitting coverage argument.

Algorithm 2 states the predictor in full, including the subroutines that Algorithm 1 abbreviates. It is the procedure that produces every number reported for our predictor.

G.1 Calibration and selection details

This section specifies the fixed base map, prior estimate and weighted calibration used in Sec. 4, followed by the selection rule and its sampling guarantee.

Fitting the base map. We split calibration logits and labels (L_c, y_c) into fitting half A and calibration half B, of sizes N_A, N_B and empirical label proportions $\hat{\pi}_A, \hat{\pi}_B$. The experiments use one class-stratified split with a fixed seed. Half A fits transport and temperature, while B supplies class-conditional prediction averages and calibrates thresholds. The logit matrices L_A, L_B, L_q contain one image per row, transposing the class-by-image convention for the transport plan in Appendix B.2. We run Sinkhorn on L_A with class marginal $\hat{\pi}_A$ and sample marginal u_{N_A}. After fixing its row scaling

$r = r^{(T)}$, we minimize half-A negative log-likelihood over the scalar temperature τ_A to obtain

$$p_A(x) = \mathrm{softmax}(\ell(x)/\tau_A + \log r). \tag{10}$$

Both fitted quantities depend only on A. We evaluate this map on B and queries to form $\mathbf{P}_B$ and $\mathbf{P}_q$, again with one image per row.

Estimating the query prior. For the fixed map p_A, Eq. (9) uses $N = N_B$. Column y of $\hat{C}$ averages half-B probability vectors with true label y, whereas row k records predicted probability for class k. The query mean $\hat{\mu}_q$ averages rows of $\mathbf{P}_q$. We solve

$$\bar{\pi} = \arg\min_{\pi \in \Delta_K} \|\hat{C}\pi - \hat{\mu}_q\|_2, \qquad \pi_\gamma = (1-\gamma)\hat{\pi}_B + \gamma\bar{\pi}, \tag{11}$$

where Δ_K is the probability simplex. The constraint yields a probability vector, and shrinkage reduces the change from the calibration prior when query evidence is weak. Thus $\gamma = 0$ retains $\hat{\pi}_B$ and $\gamma = 1$ uses the constrained estimate. We reserve $\lambda = M/(N+M)$ for the pooled query fraction and γ for correction strength.

Class support and numerical treatment. The construction requires every modeled class to be represented in A and B, giving positive empirical priors and transport scaling. The stratified experiments supply this support. The independent-splitting comparison below specifies its handling of missing classes. Here $\hat{\pi}_A$ sets the fitting marginal, while $\hat{\pi}_B$ is the reference mixture for estimation and calibration on B. The constrained solver starts at $\hat{\pi}_B$ and uses accelerated projected gradient steps, stopping after at most 2000 iterations or a maximum coordinate change below 10^{-10}. It then clips coordinates below 10^{-12} and renormalizes. These are numerical operations on the estimated query prior, distinct from the positive class-support requirement for the empirical calibration prior. Within bootstrap resamples, an absent class receives the corresponding coordinate unit vector as its confusion-matrix column, as in the implementation.

G.1.1 Selecting shrinkage by query likelihood

The unlabeled likelihood is defined by a density-ratio model. For candidate prior π, the base probabilities define the ratio $d_\pi(x) = \sum_y p_{A,y}(x)\pi(y)/\hat{\pi}_B(y)$. Under label shift this is the query-to-calibration input density ratio when the base probabilities equal the posterior for the reference prior. Accordingly, the query log-likelihood, after dropping the calibration-density term shared by all candidates, is the sum of $\log d_\pi(x)$.

We randomly divide the unlabeled query batch into five folds $\mathcal{Q}_f$. For each fold, keep $\hat{C}$ fixed and solve Eq. (11) using the query mean from the other four folds to obtain $\bar{\pi}^{(-f)}$. Set $\pi_\gamma^{(-f)} = (1-\gamma)\hat{\pi}_B + \gamma\bar{\pi}^{(-f)}$ and select

$$\gamma^\star = \arg\max_{\gamma \in \mathcal{G}_{\mathrm{like}}} \sum_{f=1}^{5} \sum_{x \in \mathcal{Q}_f} \log\left(\sum_y p_{A,y}(x) \frac{\pi_\gamma^{(-f)}(y)}{\hat{\pi}_B(y)} \right), \qquad \mathcal{G}_{\mathrm{like}} = \{0, 0.1, \ldots, 1\}. \tag{12}$$

Ties select the smallest grid value. For numerical evaluation, denominators in this objective are floored at 10^{-12} and the logarithm's argument at 10^{-300}. After selecting $\gamma^\star$, we refit $\bar{\pi}$ using the complete query batch. The selection rule below may further reduce $\gamma^\star$ before forming the final prior $\hat{\pi}_q = \pi_\gamma$.

G.1.2 Corrected scores and weighted calibration

For each candidate prior, Eq. (4) changes label probabilities while retaining the fitted base map p_A. To calibrate the resulting scores for that prior, use label-shift weights (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021). For $i \in B$, let $s'_i = 1 - p'_{y_i}(x_i)$ and $\hat{w}_i = \hat{\pi}_q(y_i)/\hat{\pi}_B(y_i)$. We place mass proportional to $\hat{w}_i$ at each score and additional mass $w_{\max} = \max_{i \in B} \hat{w}_i$ at $+\infty$. Since every class is represented in B, $w_{\max}$ bounds the weight of any candidate query label. The infinity atom accounts for that candidate's contribution without using its unknown label. Normalizing all masses gives

$$\tilde{w}_i = \frac{\hat{w}_i}{\sum_{j \in B} \hat{w}_j + w_{\max}}, \qquad \hat{q} = \inf\left\{ t : \sum_{i \in B} \tilde{w}_i \mathbf{1}\{s'_i \le t\} \ge 1 - \alpha \right\}. \tag{13}$$

We use $\hat{q} = +\infty$ when the finite scores have insufficient total mass, yielding the full label set. The returned set is $\{y : 1 - p'_y(x) \le \hat{q}\}$. At $\hat{\pi}_q = \hat{\pi}_B$, probabilities are unchanged and every weight equals one. The construction therefore equals the fallback $\mathcal{C}_A$, ordinary SCP using p_A and a threshold calibrated on B. This is the uncorrected split-sample Per-split predictor.

G.1.3 Shift detection, stability and correction strength

Evidence of shift. We compare the predicted-label histograms of $\arg\max p_A(x)$ on B and queries with Pearson's two-sample χ^2 statistic. The null is equality of these predicted-label distributions. We permute the pooled predicted labels $B_{\mathrm{perm}} = 200$ times, keeping sample sizes fixed, and reject when the observed statistic exceeds the empirical $(1 - \alpha_0)$ quantile, with $\alpha_0 = 0.05$. Theorem 4 relates this test to prior changes distinguishable by the fixed map.

Stability of prior estimation. The first indicator substitutes empirical quantities into the scale of Eq. (3):

$$\mathrm{score} = \frac{\sqrt{K}}{\sigma_{\min}(\hat{C})} \left(\frac{1}{\sqrt{M}} + \frac{K}{\sqrt{N_B}} \right).$$

It is infinite when the smallest singular value is zero. The second indicator measures sampling dispersion using $B_{\mathrm{stab}} = 50$ bootstrap replicates. Each replicate independently resamples N_B labeled rows from B and M query probability rows, with replacement. From their confusion matrix and query mean, we obtain the unconstrained least-squares solution $v^{(b)}$, clip its entries below 10^{-4} and normalize to a prior $\tilde{\pi}^{(b)}$. Writing $\tilde{\pi}_{\mathrm{mean}} = B_{\mathrm{stab}}^{-1} \sum_b \tilde{\pi}^{(b)}$, define

$$\mathrm{disp} = \frac{1}{B_{\mathrm{stab}}} \sum_b \|\tilde{\pi}^{(b)} - \tilde{\pi}_{\mathrm{mean}}\|_1.$$

We require $\mathrm{score} \le \theta_B$ and $\mathrm{disp} \le \theta_D$, with the thresholds fixed as described in Appendix C. These are empirical indicators motivated by estimation sensitivity, whereas Eq. (3) bounds the unshrunk linear estimator under its stated assumptions. If the shift test does not reject or either stability condition fails, we return $\mathcal{C}_A$.

Limiting correction strength. For candidates that pass these checks, a separate bootstrap estimates the signed coverage offset in Eq. (2). For each $\gamma \in \mathcal{G}_{\mathrm{cap}} = \{0, 0.05, \ldots, 1\}$, compute corrected scores and the weighted threshold on the original half B. Let $\hat{F}_\gamma(y)$ be the fraction of class-y examples covered at this threshold. We draw $B_{\mathrm{cap}} = 50$ new bootstrap resamples of both B and the query batch and recompute the simplex-constrained estimate $\pi^{\mathrm{boot},b}$, including its 10^{-12} floor and normalization. Thus these estimates use the deployed prior solver, while the stability bootstrap uses clipped unconstrained least squares. The original $\hat{\pi}_B$ remains the reference prior, and the same bootstrap priors are used across candidate strengths. Define

$$g(\gamma) = \mathrm{Quantile}_{0.05} \left\{ \left\langle \pi^{\mathrm{boot},b} - \pi_\gamma, \hat{F}_\gamma \right\rangle : b = 1, \ldots, B_{\mathrm{cap}} \right\}, \tag{14}$$

$$\gamma_{\mathrm{sel}} = \max \left\{ \gamma \in \mathcal{G}_{\mathrm{cap}} : \gamma \le \gamma^\star, \quad g(t) \ge g(0) - \tau_{\mathrm{rel}} \text{ for every } t \in \mathcal{G}_{\mathrm{cap}},\ t \le \gamma \right\}. \tag{15}$$

Here the bootstrap prior approximates the unknown query proportions in the population identity, and $\hat{F}_\gamma$ estimates the candidate's class coverages. The lower 5% quantile summarizes downward coverage offsets from prior uncertainty. The tolerance $\tau_{\mathrm{rel}} = 0.02$ allows a decline of two percentage points from the no-correction baseline $g(0)$. Requiring every preceding grid point to pass makes the selected strengths an interval starting at zero. The final estimate is $\hat{\pi}_q = \pi_{\gamma_{\mathrm{sel}}}$.

Computational cost. Each Sinkhorn iteration costs $O(KN_A)$, and evaluating a fixed map on B and queries costs $O(K(N_B + M))$. Temperature optimization repeats half-A likelihood evaluations. Building the confusion matrix and query mean costs $O(K(N_B + M))$. A constrained prior fit with I projected-gradient iterations uses dense $K \times K$ operations, costing $O(K^3 + IK^2)$ including matrix preparation. Five-fold selection fits five priors and evaluates all likelihood-grid candidates, followed by one full-query refit. Stability adds 50 least-squares solves, and the strength cap adds 50 constrained solves, each after rebuilding bootstrap averages. The cap also recomputes corrected half-B scores and weighted quantiles at each grid value. These repeated solves, quantile calculations and the 200 histogram permutations are additional to transport fitting and base-map evaluation.

Coverage under independent splitting. The construction uses calibration labels and query predictions without accessing query labels. Suppose calibration and query pairs are independent and

Splitting design	Coverage [95% CI]	Set size
Independent outer and inner splits	0.900 [0.898, 0.902]	3.73
Stratified outer and inner splits	0.899 [0.898, 0.900]	3.53

Table 23: **No-shift comparison of splitting designs.** Means over all 41 configurations and 20 repetitions. Intervals use the paired seed-block procedure. Independent minus stratified coverage is 0.001 $[-0.001, 0.003]$. The independent design has a set-size interval of $[3.68, 3.77]$ and returns the p_A-based fallback when B lacks a class.

identically distributed, and the split into fitting and calibration halves is independent of them. Conditional on half A, Eq. (10) is fixed and the half-B and query scores are exchangeable. The SCP predictor used for fallback therefore has marginal coverage at least $1 - \alpha$. The pretest also meets the independent-map condition of Theorem 4. With the implemented empirical-quantile test using 200 permutations, its rejection probability is at most $\alpha_0' = 11/201$ under these assumptions. This section derives this bound for the implementation's interpolation rule. The final output differs from the SCP predictor only when the pretest rejects. Subtracting that event's probability gives the no-shift coverage lower bound $1 - \alpha - \alpha_0'$. This bound subtracts the entire rejection event, including cases where the corrected set contains the true label. Table 1 reports coverage 0.899 at $\beta = \infty$. The exchangeability argument concerns the independent split and does not establish coverage for the stratified construction.

The main label-shift comparisons use the class-stratified design in Appendix C, and the additional no-shift comparison below evaluates independent splitting. Under label shift, Theorem 3 supplies the population coverage identity for a fixed corrected map and prior estimate.

Measured coverage under both splitting designs. The independent-splitting experiment randomizes both the outer calibration/query division and the inner A/B division without class stratification, with part sizes differing by at most one. All $41 \times 20 = 820$ configuration–repetition pairs are retained. Half A lacks at least one class in 203 pairs, and half B in 154. Transport scalings are floored at 10^{-300} before taking logarithms during temperature fitting and map evaluation. When B lacks a class, this experimental variant returns the p_A-based SCP fallback, including two pairs where the pretest rejects. This handles undefined empirical-prior ratios while retaining those splits.

Table 23 reports coverage, uncertainty and size under both designs. All 41 independent-split configuration means exceed $1 - \alpha - \alpha_0' = 0.845$. Mean coverage remains close to nominal under both designs, while stratified splitting yields smaller sets. The size comparison includes both partition changes and the support fallback required by the independent design. The bound above applies under its i.i.d. sampling assumptions, while stratified splitting is evaluated empirically.

Full-calibration variant. The separate fitting half addresses the reuse of labels for both adapting scores and selecting their threshold. For comparison, the full-calibration counterpart of Per-split fits transport and temperature on the whole calibration set and uses that same set to calibrate its threshold. Each calibration example then contributes to both fitted quantities. This is an instance of tuning bias (Zeng et al., 2025), whose general bound scales as $\sqrt{(d+1)/N}$ for d tuned parameters. The local calculation in Appendix B.13 yields a leading undercoverage term of order K/N for a diffuse posterior. Separate fitting removes this reuse and allocates half the calibration data to each step. With the predictor and its full-calibration comparator specified, Sec. 5 evaluates their coverage and set size against pooled transport and SCP.

The fixed base map depends only on half A. Under i.i.d. sampling and an independent split, the half-B and query scores are therefore exchangeable conditional on A. For the pretest, NumPy's default linear interpolation places the 0.95 quantile of 200 permuted statistics between the 190th and 191st order statistics. Rejection implies that at most 10 permuted statistics reach the observed value, so the permutation p-value of Theorem 4 is at most $11/201$. Consequently, $\alpha_0' = 11/201$ bounds the null rejection probability used in the coverage argument above. For general B_{perm} and α_0, the count is at most $\lceil (B_{\mathrm{perm}} - 1)\alpha_0 \rceil$. The strength cap then selects a value between zero and the likelihood-selected strength according to Eq. (15).

Algorithm 2 Split-sample prior-aware conformal prediction, full form

Require: calibration logits and labels (L_c, y_c), query logits L_q, level α, tolerance τ_{rel}, pretest level α_0, permutations B_{perm}, thresholds θ_B, θ_D and defaults in Appendix G.1.3

split (L_c, y_c) into halves A and B, stratified by class, with a fixed seed
$\hat{\pi}_A, \hat{\pi}_B \leftarrow$ label marginals of y_A and y_B
$r \leftarrow \textsc{Sinkhorn}(L_A,\, m = \hat{\pi}_A,\, u_{N_A},\, \tau,\, T)$ ▷ half A only
$\tau_A \leftarrow \arg\min_{\tau_A} \text{NLL}(L_A, y_A,\, r, \tau_A)$ ▷ fix r, fit one scalar on A
$\mathbf{P}_B \leftarrow \text{softmax}(L_B/\tau_A + \log r), \quad \mathbf{P}_q \leftarrow \text{softmax}(L_q/\tau_A + \log r)$
Pretest.
$h^B, h^q \leftarrow$ histograms of $\arg\max \mathbf{P}_B$, $\arg\max \mathbf{P}_q$
$G_{\text{obs}} \leftarrow \chi^2(h^B, h^q)$
for $j = 1$ to B_{perm} **do**
    permute the pooled predicted labels, recompute G_j
end for
shift $\leftarrow \left[G_{\text{obs}} > \text{empirical } (1-\alpha_0) \text{ quantile of } G_1, \ldots, G_{B_{\text{perm}}}\right]$
Stability check.
$\hat{C}_{jk} \leftarrow \frac{1}{N_{B,k}} \sum_{i \in B:\, y_i = k} \mathbf{P}_B[i, j], \quad \hat{\mu}_q \leftarrow \frac{1}{M} \sum_i \mathbf{P}_q[i, \cdot]$
$$\text{score} \leftarrow \frac{\sqrt{K}}{\sigma_{\min}(\hat{C})} \Big(\frac{1}{\sqrt{M}} + \frac{K}{\sqrt{N_B}} \Big)$$
independently resample labeled rows of B and query rows for each stability replicate
solve bootstrap least-squares systems, clip entries below 10^{-4}, and normalize
disp $\leftarrow$ mean ℓ_1 distance of these priors from their bootstrap mean
if not shift **or** score $> \theta_B$ **or** disp $> \theta_D$ **then**
    return $\{y : 1 - \mathbf{P}_q[\cdot, y] \le \hat{q}(\mathbf{P}_B, y_B, \alpha)\}$
end if
Shrunk prior estimate.
$\gamma^\star \leftarrow$ maximize Eq. (12), fitting each prior on the other query folds
$\bar{\pi} \leftarrow \arg\min_{\pi \in \Delta_K} \|\hat{C}\pi - \hat{\mu}_q\|_2$ using all queries and the constrained solver
recompute constrained priors on new resamples of both B and queries
compute $\hat{F}_\gamma$ on the original B and the lower-5% offset $g(\gamma)$ in Eq. (14)
$\gamma \leftarrow \gamma_{\text{sel}}$ from Eq. (15), checking every preceding grid point
$\hat{\pi}_q \leftarrow (1-\gamma)\hat{\pi}_B + \gamma\bar{\pi}$
Weighted quantile.
$\mathbf{P}'_B, \mathbf{P}'_q \leftarrow$ rows of $\mathbf{P}_B, \mathbf{P}_q$ multiplied by $\hat{\pi}_q/\hat{\pi}_B$ and renormalized
$s'_i \leftarrow 1 - \mathbf{P}'_B[i, y_i], \quad \hat{w}_i \leftarrow \hat{\pi}_q(y_i)/\hat{\pi}_B(y_i), \quad \tilde{w}_i \leftarrow \hat{w}_i/(\sum_j \hat{w}_j + \max_j \hat{w}_j)$
$\hat{q} \leftarrow$ weighted quantile in Eq. (13), with $+\infty$ for insufficient finite mass
return $\{y : 1 - \mathbf{P}'_q[\cdot, y] \le \hat{q}\}$

H Additional Experiments

The following comparisons examine how pooled loss varies with shift strength and score source, then test its scaling and its behavior under alternative query distributions.

H.1 Ordering across shift strengths

The negative sign and the ordering across strengths address different predictions of the local analysis. Of the 123 shifted configurations in Sec. 5.2, 121 have negative mean Δcov. The two nonnegative values occur on FGVCAircraft with ViT-B/32 at $\beta = 1.0, 0.3$, and both lie within 0.004 of zero.

Figure 5 separately tests whether each dataset–backbone pair obeys $\Delta\text{cov}(1.0) > \Delta\text{cov}(0.3) > \Delta\text{cov}(0.1)$. This strict ordering holds in 37 of 41 pairs. The four exceptions are EuroSAT and FGVCAircraft with ViT-B/16 and ViT-B/32. Their observed departures delimit the empirical ordering claim: the sign holds in more configurations than the monotone-strength pattern. Theorem 1 predicts local behavior under its symmetry and regularity assumptions, while these comparisons span finite shifts on real scores. For the adapted sources below, the ordering is strict in all 36 dataset–method–training-seed groups (Table 13).

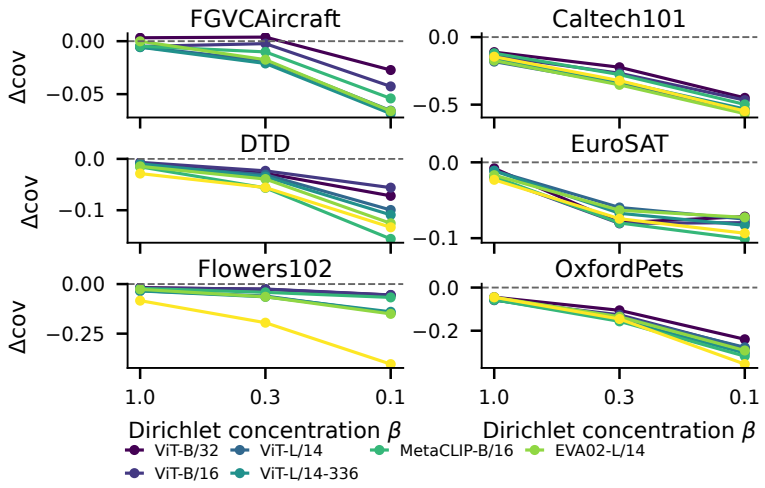

Figure 5: **The loss deepens with the shift.** Mean Δcov against the Dirichlet concentration, one panel per dataset and one line per backbone. Strict ordering across the three strengths holds for 37 of 41 dataset–backbone pairs. The four exceptions lie on EuroSAT and FGVCAircraft.

		$\sigma_{\min}(\hat{C})$			mean Δcov		
Dataset	K	zs	CoOp	Tip-F	zs	CoOp	Tip-F
FGVCAircraft	100	.0001	.0002	.0008	−.017	−.036	−.037
Caltech101	100	.0268	.2226	.1536	−.291	−.220	−.211
DTD	47	.0018	.1399	.1536	−.029	−.071	−.087
EuroSAT	10	.0098	.4954	.3639	−.060	−.103	−.103
Flowers102	102	.0008	.6505	.6986	−.033	−.255	−.244
OxfordPets	37	.2233	.5113	.3865	−.155	−.158	−.142
Food101	101	.3037	.2721	.3187	n/a	−.104	−.113

Table 24: **Pooled loss and class identifiability across score sources.** Left: the smallest singular value of the soft confusion matrix evaluated on the full calibration set using its fitted transport and temperature map, for zero-shot ViT-B/16 and adapted scores (median over three training seeds). Right: mean Δcov over three shift strengths, with adapted columns also averaging training seeds. The selection statistics for Ours instead use the separately fitted p_A map on half B. Food101 is held out from method design.

H.2 Adapted score sources

Score adaptation provides a second test of the pooled-loss mechanism and changes how reliably class proportions can be estimated. All 108 shifted CoOp and Tip-Adapter-F configurations have negative mean Δcov. Table 26 summarizes their low-coverage counts, pooled-loss bounds and agreement with zero-shot loss ranks. The bounds aggregate each method's 54 shifted configurations using the seed-block procedure in Appendix C.2.

To interpret correction on these scores, Table 24 compares confusion matrices for maps fitted on the full calibration set. All 36 groups have a larger smallest singular value than their zero-shot counterpart. FGVCAircraft remains of order 10^{-4}, whereas the other 30 groups show larger changes. The actual split-sample stability check, evaluated on B using p_A, passes on no FGVCAircraft repetitions and all EuroSAT, Flowers102 and OxfordPets repetitions. Across datasets the pass rates are 0.695 for CoOp and 0.749 for Tip-Adapter-F. Thus score adaptation expands the settings where the deployed estimator supports correction, while FGVCAircraft continues to select the fallback.

Table 3 separates the performance of CoOp and Tip-Adapter-F. Table 25 adds paired intervals for the gains from prior correction. These comparisons attribute the improved coverage and smaller sets to correction after separate fitting, while Table 24 explains how adaptation changes the class-identifiability conditions.

Score source	β	Coverage difference	Set-size difference
CoOp	1.0	0.008 [0.006, 0.009]	-0.05 [-0.07, -0.03]
CoOp	0.3	0.013 [0.008, 0.018]	-0.12 [-0.15, -0.10]
CoOp	0.1	0.026 [0.019, 0.033]	-0.13 [-0.17, -0.09]
Tip-Adapter-F	1.0	0.007 [0.004, 0.010]	-0.07 [-0.09, -0.05]
Tip-Adapter-F	0.3	0.013 [0.008, 0.017]	-0.13 [-0.15, -0.11]
Tip-Adapter-F	0.1	0.027 [0.020, 0.034]	-0.17 [-0.20, -0.14]

Table 25: **Correction gains on adapted scores.** Ours minus Per-split, averaged over 18 dataset–training-seed groups per source and strength. Brackets are paired 95% seed-block bootstrap intervals over 20 shared evaluation repetitions (Appendix C.2).

Score source	Configs	Loss upper bound	Rank agreement	Below 0.85	Min Ours
CoOp	54	-0.132	0.699	13 / 3 / 0	0.877
Tip-Adapter-F	54	-0.127	0.695	13 / 0 / 0	0.877
Food101 CoOp	9	-0.096	—	—	—
Food101 Tip-Adapter-F	9	-0.105	—	—	—

Table 26: **Adapted-score coverage and loss summaries.** Upper bounds are one-sided 99% bounds on pooled $\Delta\mathrm{cov}$. Rank agreement compares adapted and zero-shot loss over 18 dataset–strength pairs after averaging training seeds. Counts below 0.85 are Always-on / Per-split / Ours. Main rows use six datasets, three training seeds and three shifted strengths. Food101 is a separate held-out batch, with three training seeds and three strengths, evaluated for pooled loss only.

Food101 and accuracy changes. Food101 tests the same pattern on a dataset held out from method design. The separate Food101 block of Table 13 gives accuracy and paired coverage loss for each training seed. CoOp retains approximately the zero-shot accuracy, whereas Tip-Adapter-F improves it, and both sources lose coverage at every tested shift. The Food101 rows of Table 26 report the corresponding aggregate upper bounds. Loss also deepens strictly across strengths in all six method–seed groups.

Prior identifiability changes differently: the smallest singular value rises for all three Tip-Adapter-F seeds and one of the three CoOp seeds (Table 14).

H.3 Class-count ladder

We use the ImageNet class-count ladder to ask two separate questions: whether coverage loss grows approximately quadratically with the prior mismatch, and whether its magnitude can be ranked from calibration quantities. For each of 18 backbone–class-count pairs, let $L = \mathrm{cov}(\mathrm{SCP}) - \mathrm{cov}(Conf - OT) = -\Delta\mathrm{cov}$ on a paired repetition. We fit $\log L$ against $\log \|\delta\|_2$ across five strengths to estimate the exponent. The one-sided 95% lower bounds in Table 15 use the ordinary least-squares slope standard error and a Student t quantile with $n - 2$ degrees of freedom, where n is the retained point count. This compares the empirical loss relative to SCP with the local population prediction relative to nominal coverage in Eq. (1).

Figure 2 compares full-ladder and weak-shift exponents, and Table 15 gives their values by backbone. The weak-shift fit uses only $\beta \in \{10, 3\}$, moving closer to the local regime while reducing the range of prior mismatches. Both logarithmic fits retain only $L > 0$. Table 27 shows that the same 63 nonpositive losses are excluded from the full ladder's 1800 points and the weak-shift comparison's 720 points. By class count, exclusions are 16 of 600 at $K = 1000$, 19 at $K = 200$ and 28 at $K = 100$. Positive-loss selection therefore has greater influence on the restricted fit.

To compare magnitudes, we estimate κ_{obs} by least squares through the origin in $L = \kappa\|\delta\|_2^2$ using all repetitions, including nonpositive losses, and compare it with the calibration plug-in $\kappa_{\mathrm{pred}} = \frac{1}{2}a_D a_C / a_J$, summarized by its median over calibration splits. Figure 6(a) shows that predicted coefficients preserve part of the observed ordering but systematically overestimate magnitude. The rank correlation is 0.765, with $p = 10^{-4}$ under 10,000 permutations of backbones within class counts. Table 28 compares the full and weak-shift coefficient ratios across two independently drawn batches. Overestimation persists in both batches, with a smaller median discrepancy for the weakest shifts. The per-backbone ratios in Table 15 further show that the discrepancy grows as the class count falls.

β	Repetitions	Nonpositive losses	Duplicate fraction
10.0	360	60	0.148
3.0	360	3	0.238
1.0	360	0	0.374
0.3	360	0	0.577
0.1	360	0	0.756

Table 27: **Sampling and logarithmic-fit selection.** Each strength pools 18 backbone–class-count pairs and 20 repetitions. Duplicate fraction is the mean fraction of repeated query images. Nonpositive losses are excluded only from the logarithmic exponent fits. All losses enter the coefficient fits.

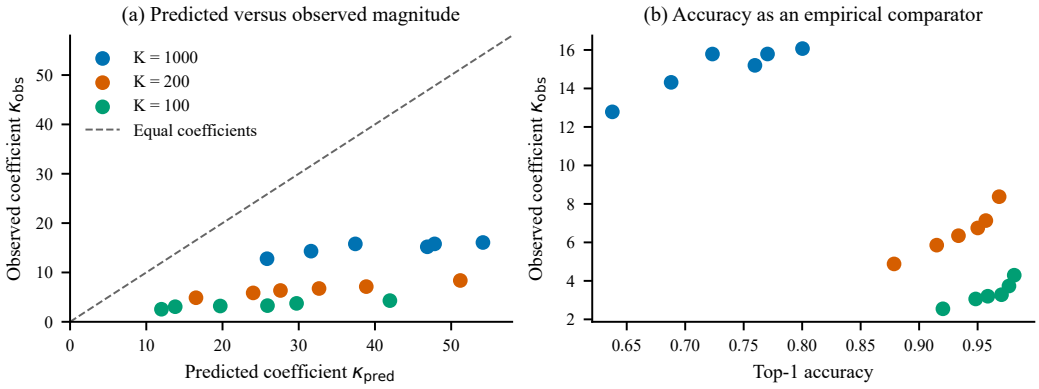

Figure 6: **Predicting the magnitude of pooled loss.** Each point is one backbone–class-count pair. (a) The calibration plug-in coefficient versus the coefficient fitted to all losses, including nonpositive losses. The dashed line denotes equality. (b) Top-1 accuracy versus the same observed coefficient. Colors identify class counts in both panels.

Top-1 accuracy is an empirical comparator to the soft-confusion response a_C. Figure 6(b) shows that accuracy also orders the observed coefficients within each class count, with correlations reported in Table 28. This comparison establishes calibration-based ranking, while leaving the additional explanatory contributions of a_D and a_J unresolved.

H.4 Long-tailed queries and natural image shifts

The long-tailed ladder changes class proportions using a fixed geometric profile, providing a comparison beyond random Dirichlet mixtures. Each ImageNet class contributes 25 calibration images. For imbalance parameter $r \in \{4, 16, 64, 256\}$, the requested query quota for class index $j = 0, \ldots, K-1$ is proportional to $r^{-j/(K-1)}$, with the largest quota 25. The implemented quotas round these values to integers and retain at least one image per class. Thus r describes the largest-to-smallest quota ratio before rounding, and the total variation in Table 29 describes the realized prior mismatch.

Batch	Full-ladder ratio	Weak-shift ratio
Main batch (seeds 100–119)	0.226	0.341
Additional batch (seeds 0–19)	0.228	0.317
Accuracy–κ_{obs} rank correlation in the main batch		
$K = 1000$	0.943	
$K = 200$	1.000	
$K = 100$	1.000	

Table 28: **Coefficient magnitude and accuracy comparison.** Top: median $\kappa_{\text{obs}}/\kappa_{\text{pred}}$ over 18 backbone–class-count pairs, using five strengths or the two weakest. Each batch contains 20 repetitions per strength with distinct random seeds. Bottom: Spearman correlations across six backbones at each class count.

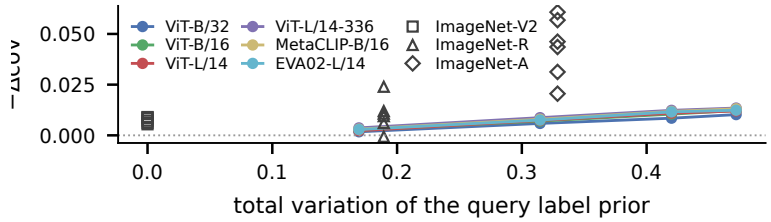

Figure 7: **Loss against the total variation of the query prior.** Lines: the long-tailed ladder, one per backbone. Hollow markers: ImageNet-V2, -R and -A, which change the class-conditional image source. ImageNet-V2 has matched class priors.

Query set	TV	SCP	Conf-OT	Per-split	Ours	Oracle
LT, ratio 4	.169	.900	.898	.901	.901	.901
LT, ratio 16	.315	.900	.893	.900	.900	.900
LT, ratio 64	.420	.900	.889	.900	.900	.900
LT, ratio 256	.472	.900	.888	.900	.900	.901

Table 29: **Long-tailed query coverage.** Marginal coverage at $\alpha = 0.10$ averaged over six backbones and 20 repetitions, with 25 calibration images per class from ImageNet validation. Ours returns the Per-split sets in every repetition because the stability check does not pass.

Pooled loss grows with total variation on every backbone, from 0.002–0.004 at ratio 4 to 0.010–0.014 at ratio 256, with strict ordering across all four ratios. SCP and Per-split remain near nominal coverage. Ours returns Per-split on every repetition because stability fails, so the corrected predictor's coverage in this ladder is supplied entirely by its fallback. This both extends the pooled-loss pattern to another label-shift design and identifies a setting where the method selects uncorrected prediction.

The natural-query comparison tests what happens when the within-class image source changes as well. Table 30 gives the natural-query coverage comparison. ImageNet-V2 has matched priors, yet all methods fall below the target. On ImageNet-A and ImageNet-R, the pretest and stability check both pass in all 120 repetitions per dataset. Their stable prior estimates therefore coexist with a coverage decrease after correction. These results identify changes in $P(X \mid Y)$ as a separate source of error that known class proportions alone do not remove.

I Extended Related Work

This appendix supports the positioning in Sec. 2 by comparing the adaptation targets of conformal prediction and label-shift estimation.

To handle changing class proportions, label-conditional conformal prediction calibrates each class separately (Vovk et al., 2005; Sadinle et al., 2019). This approach retains marginal coverage under label shift when scores remain exchangeable within classes, although scarce examples per class can produce large sets (Ding et al., 2023; 2026). Alternatively, prior-weighted conformal prediction weights calibration examples by query-to-calibration class-proportion ratios when selecting the threshold (Podkopaev & Ramdas, 2021). To estimate these proportions, Black Box Shift Estimation

		Coverage ↑				
Query set	TV	SCP	Conf-OT	Per-split	Ours	Oracle
ImageNet-V2	.000	.853	.846	.847	.847	.847
ImageNet-A	.329	.466	.423	.458	.438	.449
ImageNet-R	.189	.751	.741	.750	.718	.728

Table 30: **Coverage under natural image shifts.** Means over six backbones and 20 repetitions at $\alpha = 0.10$, using 25 ImageNet calibration images per class. V2 uses 1,000 classes, and A/R use their 200-class subsets. TV measures prior mismatch within the evaluated label space. Oracle uses the realized query prior.

(BBSE) inverts a confusion matrix relating true classes to model predictions (Lipton et al., 2018). An alternative estimator uses expectation–maximization (EM) to alternate between correcting class probabilities and updating the query proportions (Saerens et al., 2002; Alexandari et al., 2020). Relatedly, the optimal-transport method of Correia & Louizos (2025) learns calibration weights from unlabeled inputs. For subpopulation shifts, Wang et al. (2025) instead combine calibration distributions using estimated group membership. These approaches adjust calibration to the query distribution, whereas Conf-OT also makes that distribution participate in fitting the scores themselves. This additional dependence raises the question of how enforcing calibration proportions changes the scores and their resulting coverage.

Conformal prediction under distribution shift. Weighted conformal prediction handles covariate shift with likelihood-ratio weights (Tibshirani et al., 2019), and its label-shift counterpart reweights the quantile by the class-prior ratio (Podkopaev & Ramdas, 2021). Label shift has also been treated for PAC prediction sets, where the uncertainty of the estimated ratio enters the guarantee (Si et al., 2024), and for federated calibration (Plassier et al., 2023). Robust variants protect coverage over a ball of distributions (Cauchois et al., 2024; Aolaritei et al., 2025). Wasserstein regularization reduces score discrepancies (Xu et al., 2025). The non-exchangeable framework of Barber et al. (2023) bounds the gap by a total-variation term. Kasa et al. (2025) adapt the sets to an unlabeled test batch through predictive entropy. Label-conditional conformal prediction (Vovk et al., 2005; Sadinle et al., 2019) calibrates each class separately and is therefore unaffected by the class prior of the query batch, at the price of large sets when classes have few calibration examples. Cluster-wise calibration (Ding et al., 2023) and the analysis of Ding et al. (2026) study class-conditional coverage when calibration and test data share one law. Correia & Louizos (2025) bound the coverage gap using an optimal-transport distance between score distributions of a fixed score function. They learn calibration weights from unlabeled test data. These approaches target shifted calibration laws or distribution-wide protection without isolating how a pooled class constraint changes the score map. We analyze this dependence on the query batch. Mixture-based methods instead specify the components of the test distribution. Multi-distribution conformal prediction (Yang & Jin, 2026) treats the test law as a mixture of known components, and Wang et al. (2025) mix per-subpopulation calibration sets through a domain classifier. We use the latter as a baseline with classes in the role of subpopulations.

Unlabeled inputs in the conformal score. Transductive conformal inference allows scores that depend on the covariates of the calibration and test samples, and its guarantees hold as long as the scores stay exchangeable (Vovk et al., 2005; Gazin et al., 2024). Conf-OT (Silva-Rodríguez et al., 2025) applies this principle to zero-shot vision–language models. It fits a transport plan on the pooled calibration and query logits and computes the conformal score from the transported class probabilities. Semi-supervised conformal prediction (Zhou et al., 2026) instead enlarges the calibration set with pseudo-scores of unlabeled samples. When the same calibration data both tune the score and set the threshold, coverage suffers a tuning bias that grows with the number of tuned parameters and shrinks with the calibration size (Zeng et al., 2025), and even a single fitted temperature changes the set size and class-conditional coverage of conformal methods (Dabah & Tirer, 2025). The pooled transport class marginal is set by calibration labels. We study how that target changes scores when query priors differ and separate transport fitting from threshold calibration before prior correction.

Transductive vision–language models and class proportions. Transductive inference for CLIP (Zanella et al., 2024; Martin et al., 2024) uses the statistics of the query batch to improve accuracy. Veilleux et al. (2021) showed that balanced query sets hide a class-balance bias in transductive few-shot methods and drew query proportions from a Dirichlet distribution to expose it, and Zanella et al. (2025) make the same point for test-time adaptation of vision–language models. Zero-shot models also carry a label bias of their own, which logit adjustment can remove (Zhu et al., 2023). These works measure accuracy. Concurrent work studies class-conditional coverage of zero-shot models under covariate shift (Sharma & Dutta, 2026). We study marginal coverage and set size under label shift after separate fitting and prior correction. This target differs from accuracy under imbalance and class-conditional coverage under covariate shift.

Label-shift estimation. Estimating a query class prior from unlabeled data is well understood when the confusion matrix is well conditioned. Black-box shift estimation inverts that matrix (Lipton et al., 2018), the EM procedure of Saerens et al. (2002) and its calibrated refinement (Alexandari et al., 2020) sharpen the estimate, and regularized variants stabilize it when classes are rare (Azizzadenesheli et al., 2019). Garg et al. (2020) unify these estimators, later work matches class probabilities through

calibrated networks (Wen et al., 2024) or adapts online (Wu et al., 2024), and RLSbench (Garg et al., 2023) studies which combinations perform reliably under realistic shifts. We use shrunk soft confusion-matrix estimation. Prior-estimation consistency alone does not specify how its errors affect a conformal threshold. We relate prior error to marginal coverage after separate fitting, guiding the use of prior correction for compact sets.