INJECTIVE RECALIBRATION FOR CANONICAL CALIBRATION UNDER COVARIATE SHIFT WITHOUT LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating canonical calibration error (CE) on a target domain without target labels is a challenge in deployed machine learning systems facing covariate shift. We study this problem on the probability simplex, where a Dirichlet-kernel density estimator yields a consistent surrogate for the source-domain conditional $\mathbb{E}_S[Y \mid g(X)]$ that can be evaluated on unlabeled target predictions. Our central observation is that the resulting target CE admits a triangle-inequality upper bound into two terms: a tractable term $A(h)$ that depends on the recalibration map h and is estimable from labeled source data, and a residual term B that captures the source–target drift of the prediction-space label conditional. We show that injectivity of h is sufficient to make the residual mismatch term independent of the choice of recalibration map within the injective family, providing a principled justification for restricting label-free recalibration to maps such as temperature scaling. We further demonstrate empirically that the injectivity of the base classifier's representation is associated with the source-to-target estimation gap of the Dirichlet-KDE estimator under rotation shift on MNIST. Finally, we introduce a label-free overlap diagnostic that detects regions of the target prediction distribution lying outside the source support, where absolute continuity fails and the surrogate becomes unreliable. Across MNIST and SVHN under strict reweighting-induced covariate shift, CIFAR-10-C, CIFAR-10→CINIC-10, Camelyon17-WILDS, and OfficeHome, the proposed method matches or improves upon SOTA methods on most settings; we additionally characterize the regime—severe shift with collapsed support overlap—where all source-only methods, including ours, fail, and argue that the overlap diagnostic should be reported alongside any such CE estimate.

1 INTRODUCTION

Calibration asks whether predicted probabilities reflect empirical frequencies. In deployment, it is common to face domain shift (Quiñonero-Candela et al., 2008) while lacking target labels; yet practitioners still need to diagnose target-domain miscalibration to make safe decisions. This work focuses on covariate shift (CS), where $\mathbb{P}_S(Y \mid X) = \mathbb{P}_T(Y \mid X)$ but $\mathbb{P}_{S,X} \neq \mathbb{P}_{T,X}$. The setting is practically important because deployed systems routinely encounter input-distribution changes — scanner and protocol differences in medical imaging, sensor and seasonal variation in remote sensing, weather and traffic conditions in autonomous driving, evolving transaction patterns in fraud detection — while target-domain labels are scarce, delayed, or expensive to obtain.

The technical obstruction to label-free recalibration under covariate shift is a target-side conditional expectation $\mathbb{E}_T[Y \mid g(X')]$ that is not identifiable from labeled source data alone. Existing approaches sidestep this either by reweighting source samples by an estimated input-space density ratio (CPCS (Park et al., 2020), TransCal (Wang et al., 2020)), which becomes unstable when source–target overlap collapses (Cortes et al., 2010; D'Amour et al., 2021), or by composing target-accuracy estimators with post-hoc scalers (e.g. Penso & Goldberger, 2025). Both routes leave open the question of when, and to what extent, source-only recalibration can be trusted. Our method uses labeled source data and unlabeled target predictions, and it is subject to the same identifiability limit. We approach the problem differently, by working in prediction space rather than input space and by making explicit which part of the target calibration error source data can and cannot identify.

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.

We address this question via three contributions. First, we give a triangle-inequality upper bound of the target calibration error into a source-estimable term $A(h)$ and a residual B that captures the source-to-target drift of the prediction-space label conditional, and we identify injectivity of the recalibration map as the structural condition that removes h from B — providing a principled justification for restricting label-free recalibration to maps such as temperature scaling. Second, we approximate $A(h)$ by a coverage-weighted Dirichlet-KDE surrogate on the simplex, which down-weights target predictions where the source-conditional estimate is least reliable without using input-space density ratio. Finally, we introduce a label-free overlap diagnostic that converts the absolute-continuity precondition from a hidden assumption into a reportable quantity.

We validate the surrogate on six benchmarks. Under reweighting-induced strict CS (MNIST, SVHN), Ours tracks the labelled-target oracle within seed variance. Under absolute-continuity violations (CIFAR-10-C, CIFAR-10→CINIC-10, OfficeHome) Ours matches or improves upon CPCS and TransCal on most settings. On rotated-MNIST and Camelyon17-WILDS we exhibit and interpret the regime where the residual B dominates.

2 Background

We review related work on calibration under distribution shift (Section 2.1), then define covariate shift and absolute continuity (Section 2.2) and temperature scaling (Section 2.3), which fix the notation used in the rest of the paper.

2.1 Related Work: Calibration under Distribution Shift

Supervised post-hoc calibration. When labeled target validation data are available, post-hoc methods such as temperature scaling (TS, Guo et al., 2017)((Huang et al., 2025)) and multiclass extensions like Dirichlet calibration (Kull et al., 2019) address the recalibration problem directly and serve as supervised upper baselines.

Importance-weighted recalibration under covariate shift. The dominant line of work for our setting is importance weighting. CPCS (Park et al., 2020) estimates source-to-target density ratios and reweights the source-domain calibration objective to approximate target-domain risk; TransCal (Wang et al., 2020) extends this to unsupervised domain adaptation with an explicit bias–variance correction; Pampari & Ermon (2020) use a domain discriminator for the same density-ratio role. Importance weighting inherits the generalisation bounds of Cortes et al. (2010) and the diagnostics of Vehtari et al. (2024), but breaks down when source–target divergence is large — a regime that D'Amour et al. (2021) show is generic in high dimensions. Bayesian variants leveraging unlabeled target data have also been studied (Chan et al., 2020).

Methods that bypass explicit density-ratio estimation. Density-aware calibration (Tomani et al., 2023) modulates a per-sample temperature with a kNN density estimator in feature space; adaptive ensembles of calibrators (Zou et al., 2023) tune across test difficulty; pseudo-calibration (Hu et al., 2024) synthesises a labeled pseudo-target set via inference-stage mixup. Estimated-accuracy approaches (Penso & Goldberger, 2025) compose ATC-style (Garg et al., 2022) target-accuracy estimates with TS to bypass density-ratio estimation entirely. Source-free variants address the case where source data is also unavailable at deployment time (Wang et al., 2023b); training-time alternatives incorporate shift-awareness into uncertainty estimation (Wang et al., 2023a).

Positioning. Across this body of work, we observe: (i) weighting-based methods rely on source–target overlap that is rarely verified, and (ii) target-side conditional quantities involve target labels that are never observed. Our approach differs from importance-weighted recalibration in a structural way: rather than reweighting source samples by an input-space density ratio, we derive a target-domain post-hoc surrogate from an injective transformation of target predictions and reweight on the simplex. Further related work is in Appendix G.

2.2 Covariate Shift and Absolute Continuity

We consider a source distribution $\mathbb{P}_S$ and a target distribution $\mathbb{P}_T$ over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{1, \dots, K\}$, and denote their $\mathcal{X}$-marginals by $\mathbb{P}_{S,X}$ and $\mathbb{P}_{T,X}$. When taking conditional expectations we identify Y with its one-hot encoding $e(Y) \in \{0,1\}^K$. Where unambiguous, we write $X' \sim \mathbb{P}_T$ as shorthand for $X' \sim \mathbb{P}_{T,X}$. Under the assumption of **covariate shift**, the conditional distribution of Y given X remains invariant, $\mathbb{P}_S(Y|X) = \mathbb{P}_T(Y|X)$, while $\mathbb{P}_{S,X} \neq \mathbb{P}_{T,X}$.

Absolute continuity (denoted $\mathbb{P}_{T,X} \ll \mathbb{P}_{S,X}$) holds if for every measurable set $A \subseteq \mathcal{X}$:

$$\mathbb{P}_{S,X}(A) = 0 \implies \mathbb{P}_{T,X}(A) = 0. \tag{1}$$

This condition ensures that the support of the target marginal is contained within the support of the source marginal. When this holds, the target risk can be expressed as an expectation over the source distribution using a density ratio $w(x) = d\mathbb{P}_{T,X}/d\mathbb{P}_{S,X}$. On the probability simplex $\Delta^{K-1} = \{u \in \mathbb{R}^K : u_k \geq 0,\ \sum_k u_k = 1\}$, this implies that any prediction $g(x)$ produced by the model in the target domain must fall within a region where the source density is non-zero.

2.3 Temperature Scaling

Temperature scaling (TS) is a post-hoc recalibration map applied to logits $z(x) \in \mathbb{R}^K$:

$$p_T(x) = \mathrm{softmax}(z(x)/T), \quad T > 0.$$

When $T > 1$, predictions are softened; when $T < 1$, predictions are sharpened. TS is commonly fitted on labeled validation data in IID settings; here we compare multiple label-free strategies for choosing T under covariate shift.

3 Methodology

3.1 Target CE bound via injective recalibration

Let $g : \mathcal{X} \to \Delta^{K-1}$ be the classifier and $h : \Delta^{K-1} \to \Delta^{K-1}$ an injective recalibration map (e.g. temperature scaling). For $p \geq 1$, let $\|u\|_p = \left(\sum_{k=1}^K |u_k|^p\right)^{1/p}$ denote the ℓ_p norm on $\mathbb{R}^K$. The target calibration error of the recalibrated model is

$$\mathrm{CE}_T(h) \ = \ \mathbb{E}_{X' \sim \mathbb{P}_T}\Big[\big\|h(g(X')) - \mathbb{E}_T[Y \mid g(X')]\big\|_p^p\Big]. \tag{2}$$

Adding and subtracting $\mathbb{E}_S[Y \mid g(X) = g(X')]$ inside the norm, applying the triangle inequality, and taking expectations over the distribution of $v = g(X')$ induced by $X' \sim \mathbb{P}_T$ (full line-by-line derivation in Appendix A) yields

$$\begin{aligned} \mathrm{CE}_T(h) \ \leq\ & c_p \underbrace{\mathbb{E}_{X' \sim \mathbb{P}_T}\Big[\big\|h(g(X')) - \mathbb{E}_S[Y \mid g(X) = g(X')]\big\|_p^p\Big]}_{=:\,A(h)} \\ & + c_p \underbrace{\mathbb{E}_{X' \sim \mathbb{P}_T}\Big[\big\|\mathbb{E}_S[Y \mid g(X) = g(X')] - \mathbb{E}_T[Y \mid g(X')]\big\|_p^p\Big]}_{=:\,B}, \qquad c_p := 2^{p-1}. \end{aligned} \tag{3}$$

For $p = 1$, $c_1 = 1$; for $p > 1$ the factor 2^{p-1} arises from the convexity inequality $(\alpha + \beta)^p \leq 2^{p-1}(\alpha^p + \beta^p)$ (Appendix A). Since c_p does not depend on h, it does not affect the minimiser in (4). Term $A(h)$ involves only the source conditional and is therefore estimable from labeled source data; term B measures the source-to-target drift of the prediction-space label conditional and does not depend on h. Injectivity of h is what makes B independent of the choice of h within the injective family: because $h(g(X')) = h(v)$ implies $g(X') = v$, conditioning on the recalibrated output equals conditioning on the original prediction. Consequently, minimising the upper bound over $h \in \mathcal{H}_{\mathrm{inj}}$ reduces to

$$h^* \ = \ \arg\min_{h \in \mathcal{H}_{\mathrm{inj}}} A(h). \tag{4}$$

3.2 SOURCE-CONDITIONAL KDE ESTIMATOR

Given labeled source samples $\{(x_i, y_i)\}_{i=1}^n$ with predictions $p_i = g(x_i) \in \Delta^{K-1}$, we estimate the source conditional via the Dirichlet-kernel Nadaraya–Watson regressor on the simplex (Popordanoska et al., 2022):

$$\hat{\pi}_Y(q) \;=\; \frac{\sum_{i=1}^n K_b(q \mid p_i)\, e(y_i)}{\sum_{i=1}^n K_b(q \mid p_i)}, \tag{5}$$

where K_b is the Dirichlet kernel of bandwidth b and $e(y_i)$ is the one-hot encoding of y_i. Evaluating on unlabeled target predictions $\{p'_j = g(x'_j)\}_{j=1}^m$ gives a uniform empirical surrogate for $A(h)$,

$$\widehat{A}(h) \;=\; \frac{1}{m}\sum_{j=1}^m \big\| h(g(x'_j)) - \hat{\pi}_Y(g(x'_j)) \big\|_p^p, \tag{6}$$

3.3 COVERAGE-WEIGHTED SURROGATE AND OPTIMISATION

The uniform target average in (6) treats every target prediction as equally informative. However, to leading order the pointwise variance of $\hat{\pi}_Y(q)$ carries a factor $1/f_S(q)$, where f_S is the source-prediction density on the simplex (Appendix B), so residuals at poorly covered target predictions are the least reliable. We therefore weight each target residual by the plug-in density $\hat{f}_S(q) = \frac{1}{n}\sum_i K_b(q \mid p_i)$:

$$\widehat{A}^{\mathrm{cov}}(h) \;=\; \frac{\sum_{j=1}^m \hat{f}_S(p'_j)\, \big\| h(g(x'_j)) - \hat{\pi}_Y(g(x'_j)) \big\|_p^p}{\sum_{j=1}^m \hat{f}_S(p'_j)}. \tag{7}$$

Because $\hat{f}_S$ shares the same kernel and bandwidth as $\hat{\pi}_Y$, coverage weighting introduces no additional hyperparameters and no auxiliary estimator on the input space $\mathcal{X}$: both ingredients are evaluated entirely from $\{(p_i, y_i)\}$ on the source side and $\{p'_j\}$ on the target side. We instantiate h as temperature scaling and optimise

$$T^\star \;=\; \arg\min_{T>0}\; \widehat{A}^{\mathrm{cov}}(h_T) + \lambda\,(\log T)^2, \tag{8}$$

with a small regulariser λ that suppresses degenerate temperatures when the coverage mass is small. All experiments below use (8); which is referred to as recalibrator ***Ours***. The full procedure, including bandwidth selection, is given in Algorithm 1 (Appendix C).

Comparison to importance-weighted recalibration. Coverage weighting is structurally distinct from importance-weighted recalibration (CPCS (Park et al., 2020), TransCal (Wang et al., 2020)). Importance weighting reweights source samples by an estimated input-space density ratio $\hat{w}(x) \approx \mathrm{d}\mathbb{P}_T/\mathrm{d}\mathbb{P}_S$, typically via a domain discriminator, and inherits the well-known instability of high-dimensional density ratios under positivity violations (Cortes et al., 2010; D'Amour et al., 2021): weights become very large where the source density is small relative to the target density. Coverage weighting instead reweights target predictions by the source density on the $K-1$-dimensional simplex; the weight $\hat{f}_S(q)$ is bounded, and target points outside the source support receive a weight that decays to zero. The price is that the population functional being estimated is the coverage-weighted $A^{\mathrm{cov}}(h)$ rather than $A(h)$ (Appendix B), i.e. the calibration error weighted towards the part of the simplex that source data cover. The complementary uncovered mass is reported separately as the overlap diagnostic in Section 4.4.

4 EXPERIMENTS AND RESULTS

We evaluate the recalibrator first where its preconditions hold by construction (Section 4.1), then under violations of absolute continuity (Section 4.2), and finally examine how the injectivity of the representation (Section 4.3) and the source coverage of target predictions (Section 4.4) determine when a source-only estimate can be trusted.

4.1 STRICT COVARIATE SHIFT: FEATURE REWEIGHTING

We first validate the proposed surrogate in the regime where its preconditions hold by design. We construct strict covariate shift by reweighting an existing test pool: each test image is assigned a scalar acquisition-level covariate μ_i (mean pixel intensity for MNIST; luminance for SVHN), importance weights $w_i(\lambda) \propto \exp(\lambda \cdot \text{sign}(\mu_i - \text{median}(\mu)))$ are formed for $\lambda \in \{-5, \ldots, 5\}$, and a target set of size 5,000 is sampled with replacement. By construction $P(Y \mid X)$ is preserved exactly and $\mathbb{P}_T^{(\lambda)} \ll \mathbb{P}_S$ at every λ, so the residual term B in the bound (3) is expected to be small. The full construction, the standard density-ratio existence argument, and an E_s vs. E_t diagnostic confirming that the source-KDE tracks the target-KDE oracle without recalibration are in Appendix D.

The samples used to optimise T and the samples used for evaluation are drawn independently. The bandwidth b is selected without labels in every experiment: we maximise the leave-one-out log-likelihood of the Dirichlet KDE over a grid $\mathcal{B}$, separately on a subsample of source predictions (b_S) and of target predictions (b_T), and set $b = \max(b_S, b_T)$ (Algorithm 1).

We compare *Ours* ((8)) against the labelled-target oracle (Oracle TS, KDE-L_1 grid search), CPCS (Park et al., 2020), and TransCal (Wang et al., 2020).

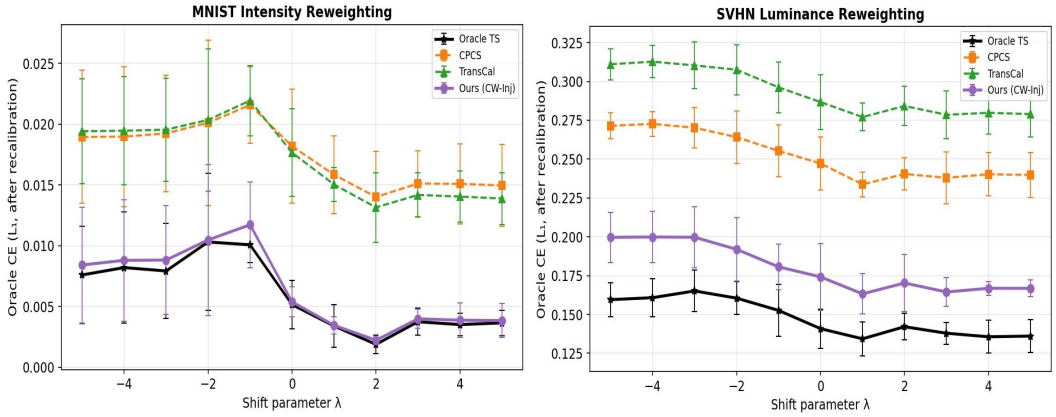


Figure 1: Post-recalibration curves using existing methods. Strict covariate shift via reweighting on MNIST (left, pixel intensity) and SVHN (right, mean luminance). Oracle TS selects T on labeled target data via grid search; CPCS, TransCal, and Ours are label-free.

Results. On MNIST (Figure 1, left), Ours tracks the labelled-target oracle to within seed variance at every λ — the two curves are visually superimposable, and Ours sits lower than CPCS and TransCal throughout the sweep. This suggests that when AC holds and $\hat{\pi}_Y$ is accurate, minimising (7) recovers a temperature close to the oracle's. The gap to CPCS and TransCal may stem from their reweighting of source samples by an input-space density ratio, which adds variance that the simplex coverage weight does not. The slight $\lambda < 0$ vs. $\lambda > 0$ asymmetry on MNIST may reflect the intrinsic difficulty of the dark-image subset, not an asymmetric AC violation (AC holds in both halves by construction).

On SVHN (Figure 1, right) the qualitative ordering is preserved: Oracle TS lowest, then Ours, then CPCS, then TransCal. Ours sits consistently between the oracle and the importance-weighted methods across all λ, recovering most of the available gain from access to source labels alone.

4.2 VIOLATION OF ABSOLUTE CONTINUITY

Real deployments rarely satisfy AC: target inputs are transformations or recaptures of the underlying scene rather than reweightings of source samples, so they fall in regions of input space where $\mathbb{P}_S$ has zero density. We organise our benchmarks below by mechanism and severity of AC violation; full data-construction details for each are in Appendix E.

Rotated-MNIST: source = MNIST test set; target = same images rotated by $\theta \in \{0°, \ldots, 90°\}$.
CIFAR-10-C: target = CIFAR-10 test images under one of eight pixel-space corruptions (Hendrycks

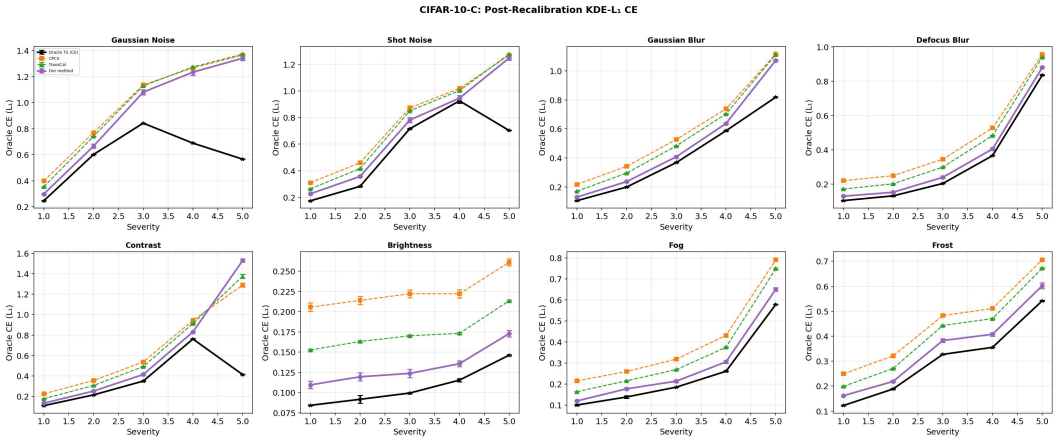


Figure 2: CIFAR-10-C: post-recalibration canonical CE across corruptions and severities.

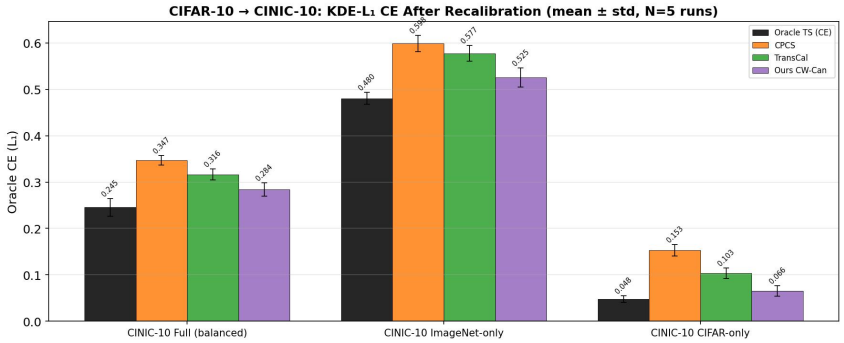


Figure 3: CIFAR-10 to CINIC-10: post-recalibration canonical CE.

& Dietterich, 2019) at five severities; the empirical MMD^2 between source and target prediction distributions (Figure 8) confirms that even severity 1 produces a measurable shift. **CIFAR-10→CINIC-10**: target = three CINIC-10 (Darlow et al., 2018) subsets spanning a controlled spectrum from CIFAR-only (AC holds) through balanced (partial violation) to ImageNet-only (maximal violation). **OfficeHome**: leave-one-domain-out across {Art, Clipart, Product, Real World}; AC is violated for every split. **Camelyon17-WILDS** (Appendix F): WILDS source hospitals {0,3,4} vs. target {1,2}.

Figures 2–5 report post-recalibration KDE-L_1 canonical CE for CIFAR-10-C, CIFAR-10→CINIC-10, OfficeHome, and rotated-MNIST respectively.

Among the unsupervised methods, our method performs favourably across several corruption types, while remaining broadly comparable to CPCS and TransCal in others. It improves over CPCS and TransCal on Gaussian noise, shot noise, Gaussian blur, defocus blur, brightness, fog, and frost in our experiments, and remains competitive on contrast for severities 1–4. These results suggest that matching recalibrated target predictions to a KDE-based source-conditional surrogate can be beneficial when target predictions remain reasonably covered by the source prediction distribution. However, the figure also shows an important limitation: for contrast at severity 5, the canonical CE of our method increases sharply and becomes worse than both CPCS and TransCal. This behaviour is consistent with the possibility that, under sufficiently severe distribution shift, the source-derived prediction-space surrogate may become unreliable.

Rotated-MNIST results. On Rotated-MNIST, Ours gives the lowest KDE-L_1 canonical CE among the three label-free methods across the rotation sweep, although the margin over CPCS and TransCal becomes small at larger angles. For $\theta \geq 60°$, all label-free methods plateau well above the labelled-target oracle, indicating that the main bottleneck is no longer the unsupervised objective but the lack of reliable source support for target prediction regions. Rotated images can produce softmax

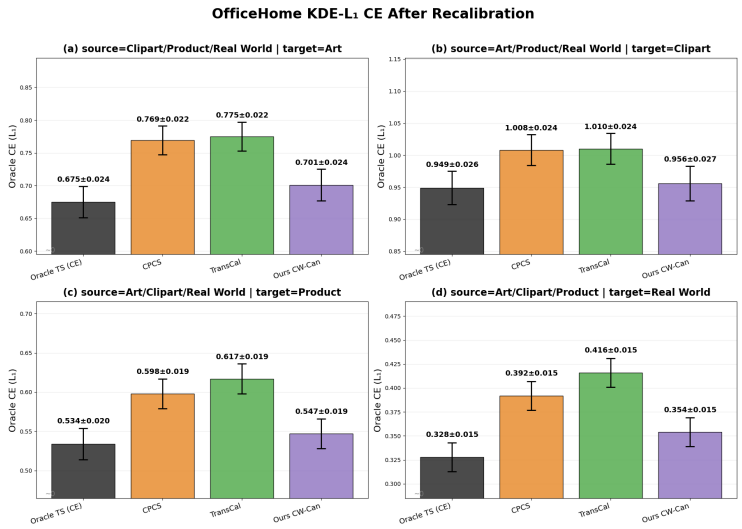


Figure 4: OfficeHome four scenes: post-recalibration canonical CE.

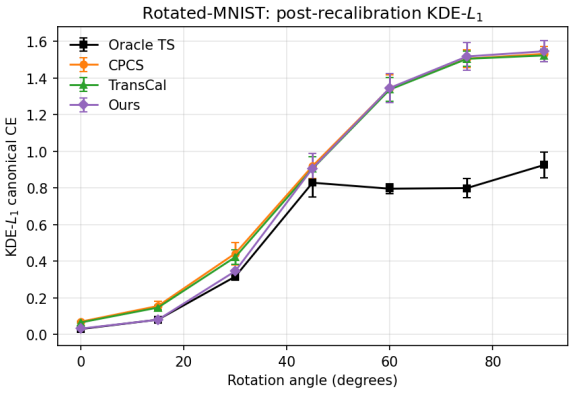


Figure 5: Rotated-MNIST: post-recalibration KDE-L_1 canonical CE vs. rotation angle. Error bars are ± 1 std across three independent training seeds. Oracle TS is grid-search KDE-L_1 on labeled target.

predictions poorly covered by the source prediction distribution, making both importance-weighted and coverage-weighted source losses less informative. The oracle gap is therefore consistent with the residual term B in (3). A per-class visualisation is in Appendix H; Camelyon17 in Appendix F shows a related high-shift regime.

4.3 Injectivity level and KDE estimation

We test the hypothesis that more injective representations reduce the source-to-target conditional estimation gap. Three output representations of increasing injectivity are evaluated as kernel spaces for the source-KDE estimator under rotated-MNIST shift: softmax probabilities ($p \in \Delta^9$, Dirichlet kernel), logits ($z \in \mathbb{R}^{10}$, Gaussian kernel), and penultimate features ($h \in \mathbb{R}^{128}$, Gaussian kernel). The oracle CE is computed via leave-one-out Dirichlet-KDE on the target domain using target labels.

The estimation gap $|\mathrm{Est} - \mathrm{Oracle}|$ orders consistently across all rotation angles: penultimate features (most injective) achieve the smallest gap, logits an intermediate gap, and softmax (least injective) the largest. All three remain small under mild shift ($\leq 30^\circ$) and grow with severity, but the rate of growth is strongly modulated by injectivity. This is a direct empirical consequence of the role of injectivity in (3): non-injective g collapses distinct input regions onto similar prediction vectors, so under covariate shift the mixture weights inside each preimage set drift, and $\hat{\pi}_S$ diverges from π_T.

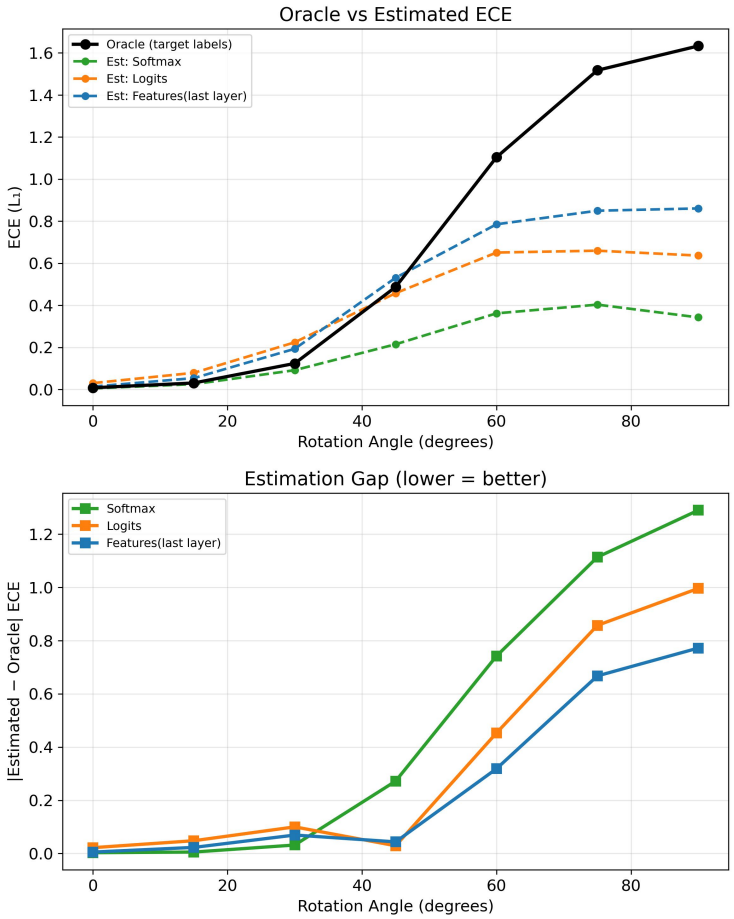


Figure 6: Comparison between representations with different injectivity levels. y-axis: KDE-L_1 Canonical CE; x-axis: rotation angle.

4.4 Overlap diagnostic

We use the uncovered target mass, i.e. the fraction of target predictions whose source-coverage score falls below a threshold, as a label-free indicator of when the source-only estimate should not be trusted. The error it is meant to predict is the estimation gap $|\mathrm{Est} - \mathrm{Oracle}|$ between the source-only CE estimate and the labelled-target oracle CE. A positive association means that lower source coverage goes with a larger estimation error.

Correlation. Pooled over conditions, the Spearman correlation between uncovered mass and $|\mathrm{Est} - \mathrm{Oracle}|$ is 0.94 on Rotated-MNIST (seven angles) and 0.89 on CIFAR-10-C (40 corruption–severity conditions). Because both quantities increase with shift severity, a pooled correlation may reflect severity rather than coverage. We therefore also report partial Spearman correlations controlling for severity: 0.797 on CIFAR-10-C and 0.502 on Rotated-MNIST.

Flagging at a fixed threshold. We flag a condition as unreliable if its uncovered mass is at least κ, and call a condition bad if $|\mathrm{Est} - \mathrm{Oracle}| > 0.2$. Table 1 reports, on CIFAR-10-C, the fraction of bad conditions that are flagged and the false-alarm rate among the remaining conditions. On Camelyon17-WILDS Center 2, where every method is at or above the no-recalibration baseline (Appendix F), the uncovered mass is 0.836, so this condition is flagged at every threshold in Table 1.

Table 1: Overlap diagnostic on CIFAR-10-C (40 conditions; 14 with $|\mathrm{Est} - \mathrm{Oracle}| > 0.2$). A condition is flagged if its uncovered mass is at least κ.

κ	bad conditions flagged	false alarms
0.15	13/14	5/26
0.20	13/14	0/26
0.25	5/14	0/26

5 Conclusions and Limitations

We derived a two-term upper bound into a tractable, source-estimable term $A(h)$ and a non-identifiable residual B, and that injectivity of the recalibration map is the structural condition that isolates the controllable term. Building on this, we developed a coverage-weighted Dirichlet-KDE surrogate that downweights poorly covered target predictions and uses no input-space density ratio. Empirically the surrogate matches the labelled-target oracle on strict-CS reweighting and improves upon importance-weighted recalibration on multiple real-world benchmarks. Beyond the recalibrator itself, we make the source-only identifiability boundary explicit and provide a label-free overlap diagnostic that converts the boundary from a hidden assumption into a reportable quantity.

Limitations. First, B is not identifiable from labeled source alone, so when prediction-space conditional drift is large, minimising $A(h)$ does not guarantee a reduction of target CE dominantly. Second, the Dirichlet-KDE surrogate degrades sharply on target points poorly covered by source — the simplex analogue of a positivity violation, and a failure mode equally affecting other existing methods. Lastly, Our analysis covers the injective recalibration family only; non-injective or input-dependent recalibrators fall outside the bound (3). In deployment, the method requires retaining labelled source predictions, and its estimate is only as reliable as the source coverage of the target predictions; we therefore recommend reporting the uncovered mass (Section 4.4) with every estimate.

AI use statement

In this work, we used generative AI tools for [clean and reformat dataset, assist in the writing of proofs, implement methods]. We have not used generative AI tools for [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results, support qualitative and thematic data analysis] and [assist with translation], are not applicable to this work.

Additionally, we used generative AI tools for [summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The derivation of the bound in (3) is given in full in Appendix A, and the motivation for the coverage weights in Appendix B. The full procedure, including label-free bandwidth selection, is given in Algorithm 1 (Appendix C). The construction of the strict covariate-shift benchmarks is detailed in Appendix D, and the construction of each absolute-continuity-violating benchmark in Appendix E and Appendix F. Code will be made publicly available upon publication.

References

John Aitchison and Ian J Lauder. Kernel density estimation for compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):129–137, 1985.

Amr M. Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International*

Conference on Machine Learning (ICML), volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232, 2020.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Jaroslaw Blasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, 2024.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1727–1740, 2023.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? *Advances in Neural Information Processing Systems*, 36:72071–72095, 2023.

Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International conference on machine learning*, pp. 1392–1402. PMLR, 2020.

Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19978–19987. IEEE, 2023.

Wonjeong Choi, Jungwuk Park, Dong-Jun Han, Younghyun Park, and Jaekyun Moon. Consistency-guided temperature scaling using style and content information for out-of-domain calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11588–11596, 2024.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Weijian Deng and Liang Zheng. Autoeval: Are labels always necessary for classifier accuracy evaluation? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1868–1880, 2021.

Weijian Deng, Yumin Suh, Stephen Gould, and Liang Zheng. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In *International conference on machine learning*, pp. 7658–7674. PMLR, 2023.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=o_HsiMPYh_x.

Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8938–8947. IEEE, 2021.

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.

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. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dapeng Hu, Jian Liang, Xinchao Wang, and Chuan-Sheng Foo. Pseudo-calibration: Improving predictive uncertainty estimation in unsupervised domain adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19304–19326, 2024.

Wenjian Huang, Guiping Cao, Jiahao Xia, Jingkun Chen, Hao Wang, and Jianguo Zhang. H-calibration: Rethinking classifier recalibration with probabilistic error-bounded objective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):9023–9042, 2025.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

Yuzhe Lu, Yilong Qin, Runtian Zhai, Andrew Shen, Ketong Chen, Zhenlin Wang, Soheil Kolouri, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Characterizing out-of-distribution error via optimal transport. *Advances in Neural Information Processing Systems*, 36:17602–17622, 2023.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Frédéric Ouimet and Raimon Tolosana-Delgado. Asymptotic properties of dirichlet kernel density estimators. *Journal of Multivariate Analysis*, 187:104832, 2022.

Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. *arXiv preprint arXiv:2006.16405*, 2020.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3219–3229. PMLR, 2020.

Coby Penso and Jacob Goldberger. Calibration of network confidence for unsupervised domain adaptation using estimated accuracy. In *Computer Vision – ECCV 2024 Workshops*, volume 15639 of *Lecture Notes in Computer Science*, pp. 153–169. Springer, 2025. doi: 10.1007/978-3-031-91585-7_10.

Alexandre Perez-Lebel, Marine Le Morvan, and Gael Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6w1k-IixnL8`.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35: 7933–7946, 2022.

Teodora Popordanoska, Sebastian Gregor Gruber, Aleksei Tiulpin, Florian Buettner, and Matthew B Blaschko. Consistent and asymptotically unbiased estimation of proper calibration errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 3466–3474. PMLR, 2024a.

Teodora Popordanoska, Gorjan Radevski, Tinne Tuytelaars, and Matthew B. Blaschko. LaSCal: Label-shift calibration without target labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024b.

Mingda Qiao and Letian Zheng. On the distance from calibration in sequential prediction. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4307–4357. PMLR, 2024.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence (eds.). *Dataset shift in machine learning*. Mit Press, 2008.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in neural information processing systems*, 33:13456–13467, 2020.

Mélanie Roschewitz, Raghav Mehta, Fabio De Sousa Ribeiro, and Ben Glocker. Where are we with calibration under dataset shift in image classification? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1NYKXlRU2H.

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

Dirk Tasche. Calibrating sufficiently. *Statistics*, 55(6):1356–1386, 2021.

Christian Tomani, Futa Waseda, Yuesong Shen, and Daniel Cremers. Beyond in-domain scenarios: Robust density-aware calibration. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

Haoxuan Wang, Zhiding Yu, Yisong Yue, Anima Anandkumar, Anqi Liu, and Junchi Yan. Learning calibrated uncertainties for domain shift: A distributionally robust learning approach. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1460–1469, 2023a. doi: 10.24963/ijcai.2023/162.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020.

Zixin Wang, Yadan Luo, Zhi Chen, Sen Wang, and Zi Huang. Cal-sfda: Source-free domain-adaptive semantic segmentation with differentiable expected calibration error. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1167–1178, 2023b.

Yaodong Yu, Stephen Bates, Yi Ma, and Michael I. Jordan. Robust calibration with multi-domain temperature scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022a.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, pp. 25721–25746. PMLR, 2022b.

Yuli Zou, Weijian Deng, and Liang Zheng. Adaptive calibrator ensemble: Navigating test set difficulty in out-of-distribution scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19333–19342, 2023.

A BOUND DERIVATION IN FULL

We give the step-by-step derivation of (3). Throughout, $h : \Delta^{K-1} \to \Delta^{K-1}$ is injective and Borel measurable, and $v = g(X')$ with $X' \sim \mathbb{P}_T$. All vectors involved lie in Δ^{K-1}, so every norm below is bounded and all expectations are finite.

(A.1) Injectivity. $h(v') = h(v) \implies v' = v$, so the preimage $\{v' : h(v') = h(v)\} = \{v\}$ is a singleton.

(A.2) Conditioning on $h(g(X'))$ equals conditioning on $g(X')$. Since Δ^{K-1} is a Polish space and h is an injective Borel map, h has a Borel-measurable inverse on its image (Lusin–Souslin theorem). Hence the σ-algebras generated by $g(X')$ and by $h(g(X'))$ coincide, and

$$\mathbb{E}_T\big[Y \mid h(g(X')) = h(v)\big] \;=\; \mathbb{E}_T\big[Y \mid g(X') = v\big].$$

This is why $\mathbb{E}_T[Y \mid g(X')]$ appears in (2). (For temperature scaling, h_T is a homeomorphism of the simplex interior, so this step is immediate.)

(A.3) Pointwise identity. For every v in the support of $g(X')$,

$$h(v) - \mathbb{E}_T[Y \mid g(X') = v] \;=\; \underbrace{\big(h(v) - \mathbb{E}_S[Y \mid g(X) = v]\big)}_{=:\,a(v)} \;+\; \underbrace{\big(\mathbb{E}_S[Y \mid g(X) = v] - \mathbb{E}_T[Y \mid g(X') = v]\big)}_{=:\,b(v)}.$$

(A.4) Norm inequality. For $p = 1$, the triangle inequality gives $\|a+b\|_1 \le \|a\|_1 + \|b\|_1$ pointwise. For $p > 1$, the triangle inequality followed by convexity of $t \mapsto t^p$ on $[0, \infty)$ gives

$$\|a+b\|_p^p \;\le\; \big(\|a\|_p + \|b\|_p\big)^p \;\le\; 2^{p-1}\big(\|a\|_p^p + \|b\|_p^p\big).$$

Both cases are covered by $\|a+b\|_p^p \le c_p\big(\|a\|_p^p + \|b\|_p^p\big)$ with $c_p = 2^{p-1}$.

(A.5) Taking expectations. Substituting $V = g(X')$ into (A.4) gives, pointwise,

$$\big\|a(V) + b(V)\big\|_p^p \;\le\; c_p\big(\|a(V)\|_p^p + \|b(V)\|_p^p\big),$$

where, by (A.3), the left-hand side equals $\big\|h(g(X')) - \mathbb{E}_T[Y \mid g(X')]\big\|_p^p$, and the two terms on the right are the integrands of $A(h)$ and B. The inequality therefore holds $\mathbb{P}_T$-almost surely. Taking expectations over the distribution of $v = g(X')$ induced by $X' \sim \mathbb{P}_T$ preserves the inequality, and linearity of expectation splits the right-hand side into two terms. The left-hand side is $\mathrm{CE}_T(h)$ by (2), and we obtain

$$\mathrm{CE}_T(h) \;\le\; c_p\big(A(h) + B\big),$$

with $A(h)$ and B as defined in (3). For $p = 1$, $c_1 = 1$ and the bound reads $\mathrm{CE}_T(h) \le A(h) + B$.

B MOTIVATION FOR COVERAGE WEIGHTING

Let $\pi_S(q) = \mathbb{E}_S[Y \mid g(X) = q]$ and $\Sigma(q) = \mathrm{Cov}_S\big(e(Y) \mid g(X) = q\big)$. For an interior point q at which f_S and π_S are smooth, standard Nadaraya–Watson asymptotics (bandwidth $b \to 0$, with the effective number of source predictions near q diverging) give, to leading order,

$$\mathrm{Cov}\big[\hat{\pi}_Y(q)\big] \;\approx\; \frac{c_b(q)}{n\, f_S(q)}\, \Sigma(q), \tag{9}$$

where $c_b(q) > 0$ depends on the bandwidth and on the position of q in the simplex, and typically grows towards the boundary (Ouimet & Tolosana-Delgado (2022) for the corresponding result for Dirichlet kernel density estimators). Pointwise estimates are therefore least reliable where f_S is small. Coverage weighting uses only this $1/f_S$ factor: the residual $\xi_j = \|h(g(x'_j)) - \hat{\pi}_Y(g(x'_j))\|_p^p$ receives weight $\hat{f}_S(g(x'_j))$, which gives (7).

This weighting is a heuristic, not an optimal aggregation: (i) it ignores the q-dependence of $c_b(q)$ and $\Sigma(q)$; (ii) the residuals ξ_j are correlated, since they share the same estimator $\hat{\pi}_Y$; and (iii) the residuals have different means, so the weighting changes the population target from $A(h)$ to

$$A^{\mathrm{cov}}(h) = \frac{\mathbb{E}_T\big[f_S(V)\, r_h(V)\big]}{\mathbb{E}_T\big[f_S(V)\big]}, \qquad V = g(X'),\ X' \sim \mathbb{P}_T, \quad r_h(v) = \|h(v) - \pi_S(v)\|_p^p.$$

Since $A^{\mathrm{cov}}(h) - A(h) = \mathrm{Cov}_T\big(f_S(V), r_h(V)\big)/\mathbb{E}_T[f_S(V)]$, the two coincide only when $f_S(V)$ and $r_h(V)$ are uncorrelated under $\mathbb{P}_T$. Minimising (7) is therefore a proxy for minimising $A(h)$ in (3), not an exact minimisation of the bound.

C Algorithm

Algorithm 1 summarises the procedure used in our experiments for $p = 1$. Here $\sigma(\cdot)$ denotes the softmax, $\mathbf{Z}_S, \mathbf{Z}_T$ are matrices of source and target logits, $\mathbf{P}_S = \sigma(\mathbf{Z}_S)$ and $\mathbf{P}_T = \sigma(\mathbf{Z}_T)$ the corresponding predictions (rows $p_{S,i}$ and p'_j), and z'_j the logit of target point j. $\mathrm{SelectBW\text{-}LOO}(\mathbf{P}, \mathcal{B})$ returns the bandwidth in the grid $\mathcal{B}$ that maximises the leave-one-out log-likelihood of the Dirichlet KDE fitted to the rows of $\mathbf{P}$; it uses no labels. $\mathrm{NW}(q, \mathbf{P}_S, b)$ returns the normalised Nadaraya–Watson weights $w_{\cdot i} = K_b(q \mid p_{S,i}) / \sum_l K_b(q \mid p_{S,l})$ of (5).

Algorithm 1 Coverage-weighted injective recalibration

Require: source logits $\mathbf{Z}_S$, labels y_S, predictions $\mathbf{P}_S = \sigma(\mathbf{Z}_S)$; target logits $\mathbf{Z}_T$, predictions $\mathbf{P}_T = \sigma(\mathbf{Z}_T)$; number of classes K; subsample size n; bandwidth grid $\mathcal{B}$; temperature bounds $[\underline{T}, \overline{T}]$

Ensure: temperature $T^\star$ and recalibrated target predictions $\sigma(\mathbf{Z}_T/T^\star)$

Subsample n source and n target indices; clip $\mathbf{P}_S, \mathbf{P}_T$ to the simplex interior
$b_S \leftarrow \mathrm{SelectBW\text{-}LOO}(\mathbf{P}_S, \mathcal{B})$; $b_T \leftarrow \mathrm{SelectBW\text{-}LOO}(\mathbf{P}_T, \mathcal{B})$; $b \leftarrow \max(b_S, b_T)$
$E_S \leftarrow [\, e(y_{S,i}) \,]_{i=1}^n$ ▷ one-hot source labels
for each target prediction p'_j **do**
  $w_{j\cdot} \leftarrow \mathrm{NW}(p'_j, \mathbf{P}_S, b)$
  $\hat{\pi}_Y(p'_j) \leftarrow \sum_i w_{ji}\, E_{S,i}$
  $\hat{f}_S(p'_j) \leftarrow \frac{1}{n} \sum_i K_b(p'_j \mid p_{S,i})$
end for
Define $\widehat{A}_{\mathrm{cov}}(T) = \dfrac{\sum_j \hat{f}_S(p'_j)\, \big\|\sigma(z'_j/T) - \hat{\pi}_Y(p'_j)\big\|_1}{\sum_j \hat{f}_S(p'_j)}$
$T^\star \leftarrow \arg\min_{\log T \in [\log \underline{T},\, \log \overline{T}]} \widehat{A}_{\mathrm{cov}}(T)$
return $T^\star$ and $\sigma(\mathbf{Z}_T/T^\star)$

D Strict-Covariate Shift reweighting: construction and diagnostic

Proposition D.1. *Given covariate shift with $\mathbb{P}_{T,X} \ll \mathbb{P}_{S,X}$,*

$$\mathbb{P}_{T,X}(X) = \exp(f(X))\, \mathbb{P}_{S,X}(X)$$

for some measurable $f : \mathcal{X} \to \mathbb{R}$.

Proof. As absolute continuity holds, $\mathbb{P}_T/\mathbb{P}_S$ is well-defined on the support of $\mathbb{P}_S$, and $f(X) = \log \mathbb{P}_T(X) - \log \mathbb{P}_S(X)$ for all X such that $\mathbb{P}_S(X) > 0$. □

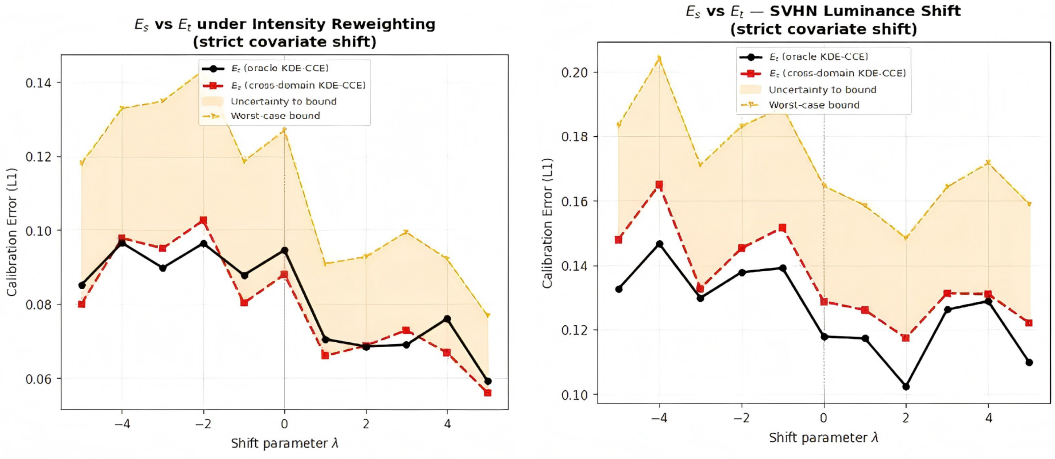
Given an f that defines an instance of covariate shift, $\frac{1}{n} \sum_i \exp(f(x_i))\, \ell(g(x_i), y_i)$ is a consistent estimator of the target risk (Quiñonero-Candela et al., 2008); resampling from $\mathbb{P}_S$ with weights proportional to $\exp(f(x_1)), \ldots, \exp(f(x_n))$ produces a consistent approximation of a sample from $\mathbb{P}_T$.

Construction used in Section 4.1. Each test image x_i is assigned a scalar acquisition-level covariate μ_i. For MNIST, μ_i is the mean pixel intensity in $[0, 1]$, which varies with pen pressure and ink density but does not determine the digit class. For SVHN, μ_i is the mean perceptual luminance computed via the ITU-R BT.601 luma formula, which varies with street lighting, time of day, and camera exposure but does not determine the house-number digit. Importance weights

$$w_i(\lambda) \;\propto\; \exp\big(\lambda \cdot \mathrm{sign}(\mu_i - \mathrm{median}(\mu))\big), \tag{10}$$

normalised to sum to one, define the reweighted target distribution at shift $\lambda \in \{-5, \ldots, 5\}$. We sample a target set with replacement. By construction $\mathbb{P}_T^{(\lambda)} \ll \mathbb{P}_S$ and $P(Y \mid X)$ is preserved exactly.

Pre-recalibration diagnostic (E_s vs. E_t). Before comparing recalibration methods we verify that the source-KDE estimator E_s tracks the labelled-target oracle E_t across the full sweep without recalibration, as Figure 7 shows.



(a) MNIST pixel intensity reweighting.

(b) SVHN mean luminance reweighting.

Figure 7: E_s vs. E_t under controlled input reweighting shifts (no recalibration).

E AC-VIOLATION CONSTRUCTION PER BENCHMARK

Rotated-MNIST. Source = MNIST test set; target = same images rotated by $\theta \in \{0°, 15°, \ldots, 90°\}$. Rotation preserves digit labels by construction, so $\mathbb{P}_S(Y \mid X) = \mathbb{P}_T(Y \mid X)$, but rotated images are physically distinct pixel arrangements not present in the unrotated source pool. The angle θ acts as a continuous knob on the AC-violation severity: $\theta = 0°$ recovers strict IID, $\theta = 90°$ produces maximally displaced predictions. A per-class visualisation of the rotation shift is given in Appendix H.

CIFAR-10-C. Targets are produced by applying the corruptions of Hendrycks & Dietterich (2019) (Gaussian noise, shot noise, Gaussian blur, defocus blur, contrast, brightness, fog, frost) to every CIFAR-10 test image, at five integer severities. Each corruption is a fixed pixel-space mapping $T_c : \mathcal{X} \to \mathcal{X}$ whose image lies outside the support of clean CIFAR-10. The severity index controls the AC-violation budget; Figure 8 reports the empirical MMD^2 between source and target prediction distributions and confirms that even severity 1 produces measurable shift, while severity 5 reaches the regime where the source-only Dirichlet-KDE estimator has essentially no support overlap.

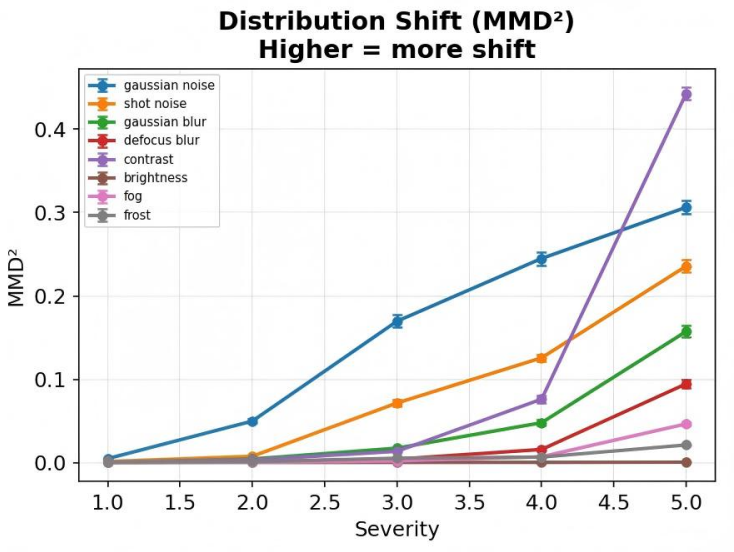


Figure 8: Distribution shift quantification for Corrupted-CIFAR-10. x-axis: corruption severity.

CIFAR-10→CINIC-10. CINIC-10 (Darlow et al., 2018) unifies CIFAR-10 with downsampled ImageNet images of the same ten classes. We evaluate three target subsets: *CINIC-10 CIFAR-only* (AC holds by construction), *CINIC-10 ImageNet-only* (downsampled ImageNet images disjoint from CIFAR-10, maximally violating AC), and *CINIC-10 Full* (balanced union, partial AC violation in which roughly half the target mass lies outside the source support).

OfficeHome. Each of {Art, Clipart, Product, Real World} is captured under different acquisition conditions for the same 65-class label set. We follow the standard leave-one-domain-out protocol: three domains form the labeled source pool, the fourth is the unlabeled target. AC is violated for every leave-one-out split.

F CAMELYON17-WILDS

We additionally evaluated on Camelyon17-WILDS using the official WILDS split with source hospitals {0,3,4} and target hospitals {1,2}. The setup induces a severe acquisition-protocol shift between training and target slides; we report it in the appendix because the distribution shift is large enough that all four methods, including ours, fall close to the no-recalibration baseline on the OOD test centre, illustrating the failure regime characterised in Section 5.

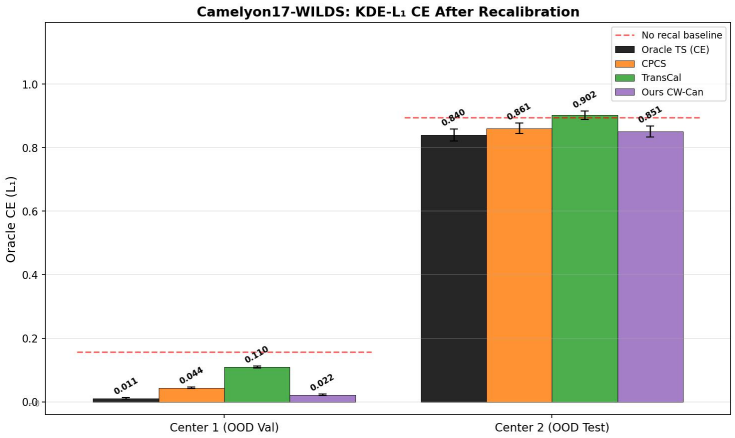


Figure 9: Camelyon17-WILDS: post-recalibration canonical CE. Source: centers 0, 3, 4; target: centers 1, 2.

On the OOD validation centre (Center 1), the differences between methods are within one standard error of zero and well below the no-recalibration baseline. On the OOD test centre (Center 2), every method's calibration error sits at approximately 0.84–0.90, statistically indistinguishable from the no-recalibration baseline of 0.840. The substantial absolute calibration error on Center 2 indicates that source-only recalibration cannot rescue calibration when the target is sufficiently far from the source pool, regardless of the recalibration objective; this is the predicted behaviour when the residual term B in (3) dominates.

G Extended Related Works

Adjacent settings. Several adjacent shift settings have motivated related recalibration methods we do not engage with experimentally but which are methodologically connected: target-free calibration for domain generalization (Gong et al., 2021; Yu et al., 2022a; Choi et al., 2024), label-shift calibration (Alexandari et al., 2020; Popordanoska et al., 2024b), and long-tailed shift (Chen & Su, 2023). A parallel literature on unsupervised accuracy estimation (AutoEval) is complementary to ours, since any consistent target-accuracy estimator can be composed with a post-hoc scaler (Garg et al., 2022; Baek et al., 2022; Lu et al., 2023; Yu et al., 2022b; Deng et al., 2023; Deng & Zheng, 2021).

Empirical context. Work on the consistency of calibration measures, which motivates our choice of metric, is discussed below. Empirical benchmarks suggest that on modern architectures shift-specific calibrators do not always dominate simple post-hoc methods combined with OOD exposure (Roschewitz et al., 2025), that pre-trained transformers are comparatively well-calibrated under shift (Minderer et al., 2021), and that foundation models exhibit qualitatively different miscalibration patterns (Tu et al., 2024); we focus on the ResNet/VGG/DenseNet regime where post-hoc calibration under shift is most widely studied.

KDE Estimator The estimator we use builds on a line of work on consistent estimation of canonical calibration errors: Popordanoska et al. (2022) introduce a consistent and differentiable L_p canonical CE estimator based on Dirichlet-kernel density regression on the simplex, extended by Popordanoska et al. (2024a) to arbitrary Bregman-divergence-induced proper calibration errors. A related foundational formulation is given in Gruber & Buettner (2022), where every proper score is shown to upper-bound a proper calibration error and unbiased upper-bound estimators are provided. Popordanoska et al. (2024a) showed that the model risk R, i.e., the expected loss under the data distribution, decomposes as:

$$R(g) = \underbrace{\mathbb{E}[D_F(\mathbb{E}[Y \mid g(X)],\, g(X))]}_{\text{Calibration}} + \underbrace{\mathbb{E}[-F(\mathbb{E}[Y \mid g(X)])]}_{\text{Refinement}}, \tag{11}$$

where $D_F(\cdot,\cdot)$ is the Bregman divergence induced by a convex function F, $g : \mathcal{X} \to \Delta^{K-1}$ is the classifier mapping inputs to the probability simplex, and $F : \Delta^{K-1} \to \mathbb{R}$ is the convex function associated with the proper scoring rule.

The estimation of the proper calibration error $CE_F(g) = \mathbb{E}[D_F(\mathbb{E}[Y|g(X)], g(X))]$ relies on accurately estimating the conditional expectation $\mathbb{E}[Y|g(X)]$. Violations of absolute continuity break both the density-ratio-based identifiability of target-domain calibration and the statistical regularity of kernel estimators on the simplex. In the probability simplex Δ^{K-1}, a target prediction $g(x)$ lying in a region where no source density exists can still be evaluated under the Dirichlet-KDE conditional-mean estimator, but the estimator's quality degrades sharply because the denominator approaches zero; see Ouimet & Tolosana-Delgado (2022) for the asymptotic theory of Dirichlet kernel density estimators on the simplex (Aitchison & Lauder, 1985), and D'Amour et al. (2021); Petersen et al. (2012) for general positivity-violation diagnostics in high dimension. Injectivity of the base classifier g is a separate condition, which we do not assume: if g is injective, $g(X)$ is a sufficient statistic of X (cf. the sufficiency perspective of Tasche (2021)), so under covariate shift $\mathbb{E}_S[Y \mid g(X) = v] = \mathbb{E}_T[Y \mid g(X') = v]$, whereas non-injective g can introduce a prediction-space label drift that is not recoverable from source alone (Rahimi et al., 2020).

On the consistency of calibration measures. The binned ECE commonly reported in the literature (Guo et al., 2017) is not in general a consistent calibration measure: Błasiok et al. (2023) define the distance-from-calibration d_{CE} and show that binned ECE fails continuity, whereas kernel-smoothed variants are polynomially related to d_{CE} and are consistent. Blasiok & Nakkiran (2024) propose kernel-smoothed reliability diagrams as a principled alternative; our Dirichlet-KDE oracle CE can be viewed as the simplex version of their SmoothECE. Kumar et al. (2019) further show that plug-in ECE is a loose surrogate for scaling-based calibrators such as temperature scaling. Blasiok et al. (2023) characterise when optimising a proper score yields calibration. These results directly motivate our use of a KDE-based canonical CE estimator rather than binned ECE as the primary evaluation metric throughout the paper.

A complementary perspective is given by Perez-Lebel et al. (2023), who separate calibration from grouping loss: if the classifier's representation collapses distinct inputs to the same prediction, the resulting information loss cannot be recovered by any post-hoc calibrator. A general survey of the canonical/class-wise/top-label hierarchy is provided by Silva Filho et al. (2023). For calibration distance in sequential prediction, see Qiao & Zheng (2024).

H ROTATED-MNIST SHIFT VISUALISATION

Figure 10 shows one example digit per class at every rotation angle used in Section 4.2. By $90°$ a rotated '7' is morphologically closer to a '1', a rotated '9' resembles a sideways '6', and rotated '4' and '5' lose their canonical strokes by $60°$. The input-space conditional $\mathbb{P}(Y \mid X)$ is technically preserved (each rotated image's label is fixed), but the prediction-space conditional $\mathbb{P}(Y \mid g(X'))$ drifts substantially because the classifier g was never trained on these orientations. This is exactly the prediction-space conditional drift that makes the residual term B in (3) large and non-identifiable from labeled source data alone.

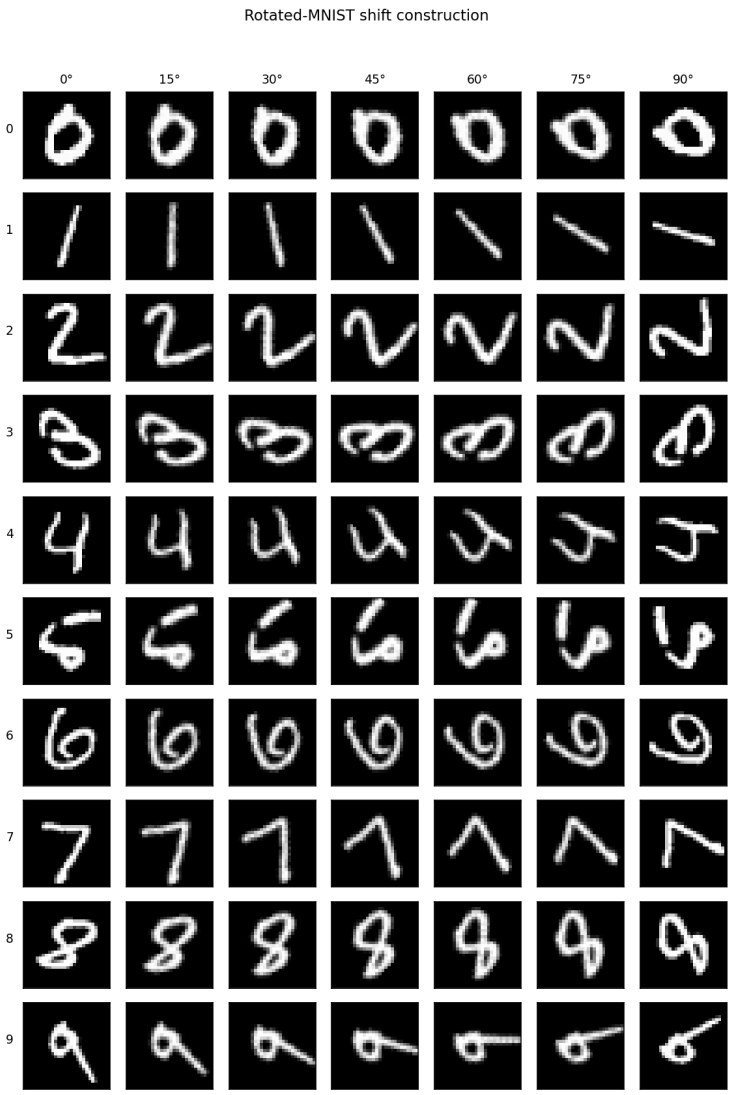


Figure 10: Rotated-MNIST shift construction. One example image per digit class, rotated by $0°, 15°, \ldots, 90°$. Rotation is applied at evaluation time only; the source classifier is trained on $0°$ MNIST.

I RELIABILITY DIAGNOSTICS, FAILURE MODES, AND OVERLAP-AWARE EXTENSIONS

This appendix expands the limitations summary in Section 5.

Non-identifiability under prediction-space conditional drift. Covariate shift assumes $\mathbb{P}_S(Y \mid X) = \mathbb{P}_T(Y \mid X)$ in input space, but this does not automatically imply $\mathbb{E}_S[Y \mid g(X) = v] = \mathbb{E}_T[Y \mid g(X') = v]$ in prediction space. If g is non-injective, multiple input regions with different label behavior may collapse onto the same prediction vector, and under covariate shift the mixture weights of these regions can change, producing a prediction-space conditional drift even when the input-space label mechanism is unchanged. In this regime, minimising the surrogate (7) may reduce $A(h)$ while moving the recalibrated predictions away from the true target conditional. Methods such as CPCS and TransCal implicitly assume the absence of this drift; our injectivity experiment in Section 4.3 makes the assumption empirically visible.

Absolute continuity and uncovered target mass. When target predictions fall in regions poorly covered by source, the Dirichlet-KDE estimator's denominator becomes small and the estimated conditional mean unstable. This is the simplex analogue of a positivity or overlap violation. Neither importance-weighted calibration nor the proposed source-conditional surrogate can reliably recover the target calibration error from source labels alone in this regime. The overlap diagnostic should therefore be viewed as a reliability condition rather than an auxiliary metric: when the uncovered target mass is large, the resulting estimate should be treated as unreliable.

Representation dependence. The simplex is the natural space for canonical CE but is not always the most reliable for estimating the source-to-target relationship. Softmax probabilities can be highly non-injective for confident classifiers; logits and penultimate features (Section 4.3) reduce the source-to-target estimation gap but change the statistical object being estimated.

Safe usage rule. We recommend reporting the proposed calibration estimate together with three reliability indicators: (i) the source–target overlap diagnostic, (ii) uncertainty bars over random seeds or bootstrap resampling, and (iii) sensitivity to the KDE bandwidth and representation space. A point estimate of target-domain CE should only be trusted when target predictions are sufficiently covered by source and the estimate is stable across reasonable bandwidth choices.

Overlap-aware extensions. The continuous coverage-weighted surrogate (7) can also be obtained as the soft-threshold limit of a hard-threshold variant. Let $\hat{s}(p'_j)$ denote a source-coverage score, τ a threshold, and define $c_j(\tau) = \mathbf{1}\{\hat{s}(p'_j) \geq \tau\}$ and $\hat{\rho}_\tau = \frac{1}{m}\sum_j c_j(\tau)$. A conservative covered-region surrogate is

$$\widehat{A}^{\mathrm{cov}}_\tau(h) = \frac{1}{\sum_j c_j(\tau)} \sum_j c_j(\tau) \left\| h(g(x'_j)) - \hat{\pi}_Y(g(x'_j)) \right\|_p^p,$$

reported with the uncovered mass $1 - \hat{\rho}_\tau$. The continuous variant in (7) is recovered when c_j is replaced by $\hat{f}_S(p'_j)$, removing the need to choose τ. Future work should study principled threshold selection for the discrete variant, conservative bounds for the uncovered region, and penalties that discourage extrapolating the KDE conditional into unsupported parts of the simplex.