BUDGETED REGIME CALIBRATION: ALLOCATING CALIBRATION RESOLUTION ACROSS PREDICTIVE REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides marginal coverage guarantees, yet reliability can vary substantially across heterogeneous regions of the input space. Local refinement can reduce this variation, but calibration resolution is finite: assigning more calibration cells to one region leaves fewer cells available elsewhere, while uniform refinement implicitly assumes that all regions benefit equally. We introduce Budgeted Regime Calibration (BRC), which formulates calibration resolution as a fixed-budget allocation problem over preformed regimes. BRC estimates regime-specific responses to candidate refinement levels using out-of-fold calibration on a dedicated STRUCTURE split and allocates the shared resolution budget according to those responses before independent final calibration. BRC selects the allocation that minimizes the largest absolute empirical coverage deviation from the target over fixed STRUCTURE audit groups, thereby controlling an upper bound on worst-group coverage shortfall.

1 INTRODUCTION

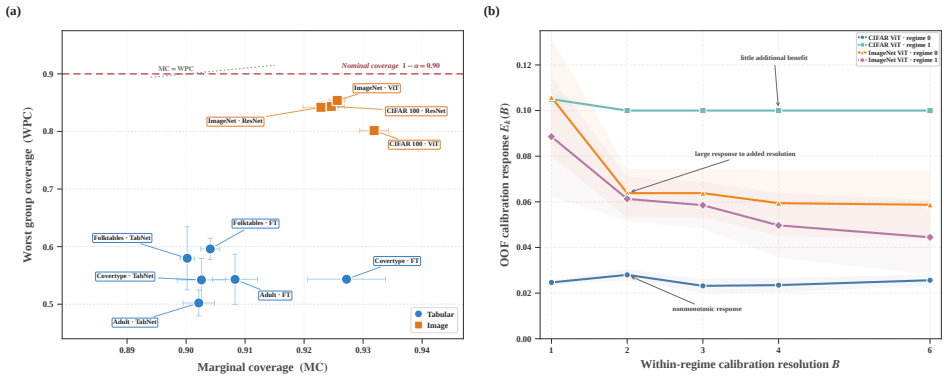

Figure 1: **Motivation for budgeted calibration resolution.** **(a)** Marginal coverage can remain near the nominal target $1 - \alpha = 0.90$ while worst-group coverage is substantially lower under matched Uniform-$2K$ refinement. **(b)** OOF calibration-error responses for representative CIFAR-100 ViT and ImageNet-1K ViT regimes vary with within-regime resolution, exhibiting different improvement rates, saturation, and nonmonotonicity.

Conformal prediction provides distribution-free predictive sets with finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018; Romano et al., 2020; Angelopoulos et al., 2021). Yet marginal validity does not guarantee comparable conditional or subgroup-level reliability across heterogeneous regions of the input space (Barber et al., 2021; Vovk, 2012). This limitation has motivated localized and neighborhood-based calibration (Guan, 2023; Han et al., 2022; Ghosh et al., 2023), feature-space conformal inference (Teng et al., 2023), and learned partitions or groups (Kiyani et al., 2024; Martinez Gil et al., 2024). Figure 1(a) illustrates the practical consequence in

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.

our classification settings: marginal coverage can remain close to the nominal target while worst-group coverage is substantially lower. Thus, satisfactory global coverage need not imply uniformly reliable behavior across heterogeneous regions.

Localized calibration addresses such heterogeneity by determining where calibration should condition or how observations should be grouped. However, a coarse conditioning structure need not eliminate residual heterogeneity within each regime, and different regimes need not benefit equally from further refinement. Additional calibration cells can resolve finer local variation, but they also divide finite calibration support among more local thresholds. Assigning the same number of cells to every regime therefore ignores potentially different responses to additional calibration resolution. Figure 1(b) shows representative response profiles that improve at different rates, saturate, or become nonmonotonic.

This exposes a downstream problem distinct from constructing the conditioning geometry itself: once coarse regimes have been formed, how should a fixed total amount of calibration resolution be distributed across them? We view calibration resolution as a finite resource. If one regime receives additional cells under a fixed total budget, those cells cannot simultaneously be assigned elsewhere. The relevant decision is therefore not simply whether finer localization is useful, but where additional resolution has the greatest estimated value.

We introduce **Budgeted Regime Calibration (BRC)** to address this fixed-budget allocation problem. BRC is intended for settings in which coarse calibration regimes are already available but limited local calibration resolution must be distributed across them. Given preformed regimes and an externally specified calibration-cell budget, BRC estimates how each regime responds to candidate resolutions using out-of-fold calibration on a dedicated STRUCTURE split and jointly allocates the available cells. The resulting structure and allocation are frozen before an independent CAL split is used to estimate final conformal thresholds. Predictive uncertainty determines where refinement occurs within a regime, whereas observed calibration response determines how much resolution that regime receives. BRC therefore does not assume that higher uncertainty should automatically receive more calibration cells.

Our main contributions are:

- **Fixed-budget calibration-resolution allocation.** We formulate local calibration resolution as a finite resource and study how a fixed number of calibration cells should be distributed across already formed regimes.
- **Response-based BRC.** BRC estimates regime-specific resolution responses using STRUCTURE-only out-of-fold calibration, jointly selects an allocation under the shared budget, and freezes the resulting structure before independent final calibration.
- **Matched fixed-budget evidence.** Across tabular and image classification settings, BRC improves worst-group reliability over matched Uniform refinement most strongly at the tightest nontrivial resolution budget, while preserving the same predictor, regime structure, conformity score, and total number of calibration cells.

1.1 Related Work

Prior work addresses heterogeneous conformal reliability along several complementary axes. Localized, group-conditional, and multivalid methods change where calibration conditions or for which groups coverage is controlled (Vovk et al., 2003; Guan, 2023; Jung et al., 2023; Gibbs et al., 2025). Other methods learn representations, partitions, or groups on which calibration operates (Kiyani et al., 2024; Martinez Gil et al., 2024; Zhou & Sesia, 2024; Xu et al., 2026b), or adapt conformity scores and coverage mechanisms (Romano et al., 2019; 2020; Plassier et al., 2025; Gauthier et al., 2026). Recent work additionally treats allocation or discretization as an explicit conformal design choice (Xu et al., 2026a; Das et al., 2026; Toccaceli, 2026; Menadjiev et al., 2026; Li et al., 2026).

BRC studies a complementary downstream decision. Once a coarse regime structure has been formed, it holds the total number of calibration cells fixed and asks how that finite resolution should be distributed across regimes according to their estimated response to further refinement. Appendix A provides a detailed comparison with these literature lines and clarifies the corresponding novelty boundary.

2 PROBLEM FORMULATION

Let $\widehat{f}$ be a trained predictor, $S(x, y)$ a fixed conformity score, and $1 - \alpha$ the target coverage level. Suppose an upstream procedure has already formed K coarse predictive regimes

$$P_1, \ldots, P_K, \qquad r(x) \in \{1, \ldots, K\}.$$

BRC begins from this supplied coarse structure. The methodological question is therefore not how to discover the globally optimal partition, but how much calibration resolution each already formed regime should receive.

Let B_k denote the number of calibration cells assigned to regime P_k, where $B_k = 1$ leaves that regime unrefined. For a finite candidate-resolution set $\mathcal{B}$ and externally specified total resolution budget C, define

$$\mathcal{F}_C = \left\{ \mathbf{B} = (B_1, \ldots, B_K) : B_k \in \mathcal{B},\ \sum_{k=1}^{K} B_k = C \right\}. \tag{1}$$

The shared budget couples the regime-level decisions: assigning additional resolution to one regime consumes resolution that could otherwise be assigned elsewhere.

When $C = qK$ and $q \in \mathcal{B}$, matched Uniform refinement corresponds to

$$\mathbf{B}^{\mathrm{uni}} = (q, \ldots, q) \in \mathcal{F}_C.$$

Thus BRC and Uniform operate under the same total number of calibration cells. The question is whether that fixed resolution can be distributed more effectively using regime-specific calibration response rather than assigned equally. Appendix B documents the development diagnostics that motivated this formulation; those diagnostics are not components of the final algorithm.

3 BUDGETED REGIME CALIBRATION

BRC adds a response-based allocation step to a localized conformal pipeline. Given a frozen predictor, supplied regimes, and a fixed cell budget, BRC evaluates candidate refinement levels using STRUCTURE-only out-of-fold calibration, jointly selects one resolution per regime, freezes the resulting routing structure, and only then performs final conformal calibration on an independent CAL split.

Three quantities have distinct roles. An upstream procedure supplies a frozen coarse regime map; a within-regime uncertainty coordinate determines where candidate refinement occurs; and estimated calibration response determines how much resolution each regime receives. In particular, the BRC decision variable is the regime-level resolution vector rather than the upstream regime constructor. Algorithm 1 summarizes the full decision order, while Appendix C gives the implementation details.

Algorithm 1 Budgeted Regime Calibration

Require: Frozen predictor; STRUCTURE and CAL splits; score S; target α; candidate set $\mathcal{B}$; budget C.

1: Using STRUCTURE, obtain and freeze the upstream regime map, uncertainty references, candidate boundaries, and common audit groups.
2: **for** each regime k and $b \in \mathcal{B}$ **do**
3: Estimate five-fold OOF response costs $E_k(b)$ and $W_k(b)$.
4: **end for**
5: Select

$$\mathbf{B}_C^{\star} \leftarrow \operatorname*{lex\,arg\,min}_{\mathbf{B} \in \mathcal{F}_C} \left(\max_k E_k(B_k), \sum_k W_k(B_k), \mathbf{B} \right).$$

6: Freeze $\mathbf{B}_C^{\star}$, final cells, and routing.
7: Using CAL only, estimate conformal thresholds with the cell $\rightarrow$ regime $\rightarrow$ global support hierarchy.
8: **return** the resulting conformal predictor.

3.1 UPSTREAM REGIMES AND CANDIDATE RESOLUTION

All structural decisions use a dedicated STRUCTURE split and are completed before final calibration. In our classification instantiation, the upstream regime map is formed from a frozen predictor-

derived signature

$$z(x) = [\phi(x); \mathbf{u}_{\mathrm{reg}}(x)], \tag{2}$$

where $\phi(x)$ is a model-internal representation and $\mathbf{u}_{\mathrm{reg}}(x)$ contains predictive-uncertainty information. STRUCTURE-fitted preprocessing and KMeans then produce $r(x)$. The regime count K is selected using a prespecified stability-and-support criterion without CAL or TEST.

This regime constructor is the canonical upstream instantiation rather than the BRC allocation mechanism. BRC requires only a frozen regime map $r(x)$; the downstream response estimation and fixed-budget allocation are unchanged when the regime constructor is replaced. We evaluate this constructor modularity with five-seed non-KMeans experiments in Appendix H.3.

Within each frozen regime, classification examples are ordered using normalized predictive entropy $u_1(x)$ and inverse top-two margin $u_2(x)$. Let $R_{j,k}(x)$ be the percentile rank of $u_j(x)$ relative to STRUCTURE observations assigned to regime k. We define

$$U(x) = \frac{1}{2}R_{1,r(x)}(x) + \frac{1}{2}R_{2,r(x)}(x). \tag{3}$$

For candidate resolution b, STRUCTURE quantiles of U divide regime k into b approximately equal-mass cells. Hence $U(x)$ specifies where refinement occurs, whereas B_k specifies how much refinement regime k receives.

We use the prespecified finite set

$$\mathcal{B} = \{1, 2, 3, 4, 6\}.$$

It is a fixed design grid rather than a theoretically privileged or TEST-selected sequence. The allocation formulation applies to another prespecified finite candidate set whenever the requested exact budget is feasible. Appendix C.3 specifies the percentile, quantile-boundary, tie, and routing conventions.

3.2 Out-of-Fold Resolution Response

BRC does not infer resolution demand directly from uncertainty magnitude. Instead, for every regime k and candidate $b \in \mathcal{B}$, it estimates how calibration reliability responds when that regime receives b cells.

Candidate resolutions are evaluated through five-fold OOF calibration on STRUCTURE, conditional on frozen STRUCTURE geometry. Before foldwise threshold estimation, BRC freezes the regime map, uncertainty references, candidate boundaries, and a common audit partition

$$\mathcal{G}_{\mathrm{audit}} = P \times U_{10}.$$

The audit groups are distinct from the candidate BRC cells and remain identical for every candidate resolution. Thus changing B_k cannot also change the yardstick on which that candidate is evaluated.

Let $A_{k,d}$, $d = 1, \ldots, 10$, denote the supported audit groups within regime k, and let $\widehat{C}^{(b)}_{k,d}$ denote their OOF coverage under candidate resolution b. Define

$$E_k(b) = \max_d \left| \widehat{C}^{(b)}_{k,d} - (1-\alpha) \right|, \tag{4}$$

$$W_k(b) = \sum_d |A_{k,d}| \left| \widehat{C}^{(b)}_{k,d} - (1-\alpha) \right|. \tag{5}$$

The map $b \mapsto E_k(b)$ is the estimated resolution-response profile of regime k. Such profiles may improve, saturate, or vary nonmonotonically because finer localization also divides finite calibration support. The resulting response table may be nonmonotone: uncertainty determines the within-regime refinement order, while the observed calibration response determines the amount of resolution allocated to that regime.

Appendix C.4 specifies the balanced fold construction, foldwise thresholds, fixed audit groups, support requirements, and OOF fallback. We refer to this procedure as five-fold OOF response estimation conditional on frozen STRUCTURE geometry, rather than as nested cross-fitting.

3.3 BUDGET-CONSTRAINED RESPONSE ALLOCATION

For $\mathbf{B} \in \mathcal{F}_C$, define

$$\mathcal{E}_{\max}(\mathbf{B}) = \max_k E_k(B_k), \tag{6}$$

$$\mathcal{E}_{\mathrm{w}}(\mathbf{B}) = \sum_k W_k(B_k). \tag{7}$$

BRC selects

$$\mathbf{B}_C^\star = \operatorname*{lex\,arg\,min}_{\mathbf{B}\in\mathcal{F}_C} \left(\mathcal{E}_{\max}(\mathbf{B}), \mathcal{E}_{\mathrm{w}}(\mathbf{B}), \mathbf{B}\right). \tag{8}$$

The primary criterion minimizes the largest supported local deviation from the nominal target. Among allocations tied on the primary criterion, the secondary term minimizes support-weighted aggregate deviation. Remaining exact ties are resolved deterministically by the allocation vector. This avoids introducing a separately tuned scalar trade-off parameter.

Proposition 1 (Worst-group shortfall control and matched Uniform)**.** *Define*

$$\widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B}) = \min_{k,d} \widehat{C}_{k,d}^{(B_k)}$$

over the supported common audit groups. Then every feasible allocation satisfies

$$\widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B}) \geq (1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}). \tag{9}$$

Equivalently,

$$\left[(1-\alpha) - \widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B})\right]_+ \leq \mathcal{E}_{\max}(\mathbf{B}), \tag{10}$$

where $[z]_+ = \max\{z, 0\}$.

If $C = qK$ *for* $q \in \mathcal{B}$*, then*

$$\mathbf{B}^{\mathrm{uni}} = (q, \ldots, q) \in \mathcal{F}_C,$$

and

$$\mathcal{E}_{\max}(\mathbf{B}_C^\star) \leq \mathcal{E}_{\max}(\mathbf{B}^{\mathrm{uni}}). \tag{11}$$

Consequently,

$$(1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}_C^\star) \geq (1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}^{\mathrm{uni}}), \tag{12}$$

so the BRC allocation has a no-worse STRUCTURE-side lower bound on worst-group coverage than matched Uniform under the primary criterion.

If the primary objectives tie, then

$$\mathcal{E}_{\mathrm{w}}(\mathbf{B}_C^\star) \leq \mathcal{E}_{\mathrm{w}}(\mathbf{B}^{\mathrm{uni}}). \tag{13}$$

Proposition 1 gives the allocation objective a direct local-reliability interpretation. Equation equation 10 shows that $\mathcal{E}_{\max}$ upper-bounds worst-group coverage shortfall from the nominal target on the fixed STRUCTURE audit family. Thus the STRUCTURE objective is not unrelated to WPC: minimizing its primary criterion controls a conservative measure of local undercoverage. Because the criterion is based on absolute deviation from the nominal target, it also penalizes excess coverage rather than rewarding conservatism solely for increasing minimum coverage.

Matched Uniform remains a feasible allocation whenever $C = qK$, so BRC is compared against Uniform within the same total-resolution decision space. After $\mathbf{B}_C^\star$ is selected, the allocation and routing structure are frozen; independent CAL data then estimate the final conformal thresholds, and TEST is reserved for evaluating the resulting predictor out of sample. The proof is given in Appendix C.6.

The finite allocation problem is solved exactly. The implementation first identifies the smallest attainable max-gap threshold permitting an exact-budget allocation and then minimizes the weighted secondary cost under that threshold. Dynamic programming is solely an exact solver for this finite allocation problem rather than a separate methodological contribution. Appendix C.7 gives the recurrence, complexity, and proof of exactness.

3.4 Independent Final Calibration and Validity

After selecting $\mathbf{B}_C^\star$, BRC freezes the regime map, K, uncertainty reference distributions, candidate boundaries, allocation, and final routing. Only then is the independent CAL split used to estimate the final conformal thresholds. CAL and TEST labels therefore do not participate in regime construction, response estimation, or budget allocation.

For a directly calibrated final cell h containing n_h CAL scores $S_{(1)} \le \cdots \le S_{(n_h)}$, define

$$j_h = \lceil (n_h + 1)(1 - \alpha) \rceil, \qquad \widehat{q}_h = \begin{cases} S_{(j_h)}, & j_h \le n_h, \\ +\infty, & j_h > n_h. \end{cases} \tag{14}$$

The resulting predictor is

$$\widehat{\Gamma}_\alpha(x) = \left\{ y : S(x, y) \le \widehat{q}_{h(x)} \right\}. \tag{15}$$

The conformity score determines the form of the prediction set or interval, whereas BRC determines how the fixed calibration-resolution budget is distributed. The formulation is therefore score-modular, not score-invariant: changing the conformity score can alter the response profiles and hence the selected allocation.

Final calibration uses a prespecified support hierarchy,

$$\text{selected cell} \to \text{parent regime} \to \text{global}.$$

Exact support thresholds and score-specific prediction rules are given in Appendix C.8.

Corollary 1 (Inherited groupwise validity for a directly calibrated frozen cell)**.** *Condition on the fitted predictor and every quantity frozen before CAL. Let h be a final cell whose threshold is estimated directly from its own CAL observations. If those calibration scores and the score of a future observation assigned to h are exchangeable, then*

$$\Pr\left\{ Y_{n+1} \in \widehat{\Gamma}_\alpha(X_{n+1}) \mid h(X_{n+1}) = h \right\} \ge 1 - \alpha. \tag{16}$$

Corollary 1 is the standard groupwise split-conformal rank guarantee inherited after the BRC structure is frozen independently of CAL. Validity under fallback is stated at the calibration-group level whose threshold is actually used; full details are given in Appendix C.9.

3.5 Method Scope

BRC exposes the regime-level resolution vector $\mathbf{B} = (B_1, \ldots, B_K)$ as a joint decision under the conserved budget $\sum_k B_k = C$, and estimates that decision from regime-specific calibration response. Predictor representations, KMeans, uncertainty ranks, standard conformal quantiles, support fallback, and dynamic programming instantiate this mechanism but are not separate contributions.

4 Experiments and Results

4.1 Experimental Setup

We evaluate classification on Adult (Kohavi, 1996), Covertype (Blackard & Dean, 1999), and Folktables ACSIncome (Ding et al., 2021) with TabNet (Arik & Pfister, 2021) and FT-Transformer (Gorishniy et al., 2021), and on CIFAR-100 (Krizhevsky, 2009) and ImageNet-1K (Deng et al., 2009) with ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021). Bike Sharing (Fanaee-T & Gama, 2014) and California Housing (Pace & Barry, 1997) provide regression transfer tests. Tabular classification uses LAC, image classification uses deterministic nonrandomized strict-before-candidate APS (Romano et al., 2020), and regression uses absolute residual scores.

For tabular classification and regression, we use disjoint $55/10/15/10/10\%$ TRAIN/VAL/STRUCTURE/CAL/TEST splits. For CIFAR-100, the ImageNet-pretrained feature extractor is frozen and a seed-specific linear classification head is fitted using the official training pool, which is partitioned $60/10/15/15\%$ into TRAIN/VAL/STRUCTURE/CAL; the official CIFAR-100 test set is used as TEST. For ImageNet-1K, the pretrained predictor is held fixed

and the available held-out pool is split $30/30/40\%$ into STRUCTURE/CAL/TEST. In all settings, STRUCTURE determines the frozen regime geometry and allocation before independent CAL calibration, and TEST is reserved exclusively for final evaluation. We study $C \in \{2K, 3K, 4K\}$ with $\mathcal{B} = \{1, 2, 3, 4, 6\}$, and report marginal coverage (MC), worst-group coverage (WPC), and prediction-set size or interval length over five reporting seeds $\{42, 999, 0, 123, 12345\}$, using mean $\pm$ sample standard deviation.

Matched Uniform refinement is the primary comparator: it preserves the predictor, conformity score, regime map, data partitions, audit groups, and total calibration-resolution budget while replacing only the learned allocation by equal allocation. Complete protocol, metric, support, and baseline specifications are given in Appendix D.

4.2 Primary Fixed-Budget Classification

At $C = 2K$, Uniform assigns two cells to every regime, whereas BRC redistributes exactly the same total number of cells using the STRUCTURE-side calibration responses. The predictor, score, upstream regime geometry, data, audit groups, and total resolution are unchanged, so this comparison directly isolates the cross-regime allocation decision.

Table 1 reports all ten classification settings. BRC attains numerically higher WPC in nine settings. The largest matched differences occur on Adult TabNet ($+0.1575$), Adult FT-Transformer ($+0.1505$), Folktables TabNet ($+0.1251$), Folktables FT-Transformer ($+0.0909$), and CIFAR-100 ViT ($+0.0606$). At the primary $2K$ budget, BRC also yields smaller mean prediction sets in $8/10$ settings, indicating that the local-reliability gains are generally not obtained through set inflation.

Table 1: **Primary matched classification comparison at $C = 2K$.** Entries are five-seed mean $\pm$ sample SD. $\Delta\text{WPC} = \text{WPC}_{\text{BRC}} - \text{WPC}_{\text{Uniform}}$. Markers indicate numerical direction only. MC, WPC, and prediction-set size are reported jointly.

Dataset	Backbone	U-MC	BRC-MC	U-WPC	BRC-WPC	ΔWPC	U-Size	BRC-Size
Adult	TabNet	0.9021 ± 0.0027	0.8989 ± 0.0084	0.5021 ± 0.0226●	0.6596 ± 0.0232▲	$+0.1575$	1.1893 ± 0.0192	1.1678 ± 0.0204
Adult	FT-Transformer	0.9083 ± 0.0038	0.9151 ± 0.0135	0.5431 ± 0.0436●	0.6936 ± 0.0509▲	$+0.1505$	1.2004 ± 0.0055	1.1803 ± 0.0164
Covertype	TabNet	0.9026 ± 0.0041	0.9053 ± 0.0043	0.5422 ± 0.0372●	0.6073 ± 0.0945▲	$+0.0651$	1.0814 ± 0.0243	1.0720 ± 0.0441
Covertype	FT-Transformer	0.9272 ± 0.0066	0.9305 ± 0.0156	0.5434 ± 0.0020●	0.5966 ± 0.0934▲	$+0.0532$	1.0101 ± 0.0067	1.0104 ± 0.0154
Folktables	TabNet	0.9002 ± 0.0012	0.9009 ± 0.0024	0.5798 ± 0.0549●	0.7049 ± 0.0122▲	$+0.1251$	1.2418 ± 0.0155	1.2259 ± 0.0210
Folktables	FT-Transformer	0.9041 ± 0.0016	0.9023 ± 0.0014	0.5960 ± 0.0185●	0.6869 ± 0.0270▲	$+0.0909$	1.2552 ± 0.0081	1.2281 ± 0.0055
CIFAR-100	ResNet-50	0.9246 ± 0.0022	0.9263 ± 0.0051	0.8435 ± 0.0077●	0.8515 ± 0.0202▲	$+0.0080$	3.4188 ± 0.0646	3.4021 ± 0.0356
CIFAR-100	ViT-B/16	0.9319 ± 0.0024	0.9361 ± 0.0013	0.8016 ± 0.0060●	0.8621 ± 0.0159▲	$+0.0606$	2.5615 ± 0.1685	2.5352 ± 0.1027
ImageNet-1K	ResNet-50	0.9229 ± 0.0030	0.9224 ± 0.0028	0.8419 ± 0.0091●	0.8471 ± 0.0135▲	$+0.0051$	7.0279 ± 0.5345	6.7731 ± 0.5198
ImageNet-1K	ViT-B/16	0.9256 ± 0.0013	0.9279 ± 0.0007	0.8539 ± 0.0040●	0.8533 ± 0.0093▼	-0.0006	6.1826 ± 0.4920	6.6762 ± 0.2227

Complete matched results at $2K$, $3K$, and $4K$, including all setting-specific values, are reported in Appendix E.

4.3 Budget Relaxation and Mechanism

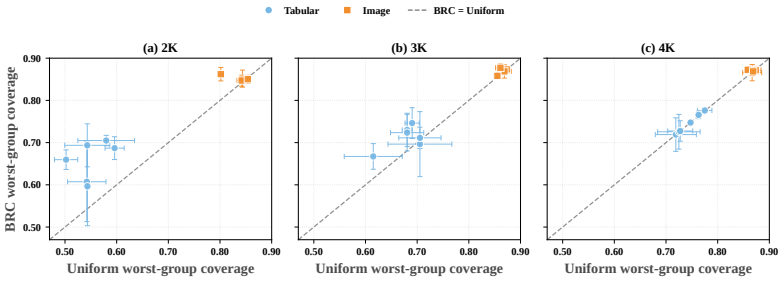

Figure 2: **Budget relaxation under matched resolution.** Each point compares BRC with matched Uniform for one dataset–backbone setting; whiskers show one SD over five seeds and the dashed line denotes equal WPC. Separation is most consistent at $2K$ and becomes smaller or more setting-dependent at larger budgets.

Figure 2 summarizes the effect of relaxing the shared resolution budget across all ten classification settings. BRC has numerically higher WPC in $9/10$, $7/10$, and $4/10$ settings at $2K$, $3K$, and $4K$, respectively, with four exact TEST-WPC ties at $4K$. The strongest and most consistent separation occurs at the tightest nontrivial budget, consistent with the proposed mechanism: when resolution is scarce, differences in regime-specific response make allocation more consequential. As the common budget grows, matched Uniform itself becomes finer, reducing the remaining allocation leverage. Thus the BRC–Uniform advantage generally contracts without implying that BRC's absolute WPC decreases. Points close to the equality line indicate limited allocation leverage rather than necessarily exact ties.

The response behavior introduced in Figure 1(b) provides a mechanism for this variation. For one CIFAR-100 ViT regime, $E_k(B)$ changes from 0.0247 at $B = 1$ to 0.0280 at $B = 2$ and 0.0232 at $B = 3$. Thus, additional refinement need not monotonically improve calibration, motivating direct estimation of regime-specific resolution response rather than allocation from uncertainty magnitude alone. Complete response profiles and downstream mechanism analyses are reported in Appendix G.

Internal controls test whether simpler mechanisms explain the same behavior. To keep this comparison compact, Table 2 shows CIFAR-100 ViT, for which both Random and GlobalU controls are available at the matched $2K$ total-cell budget. Random tests arbitrary nonuniform allocation, while GlobalU is a matched-complexity control that removes the regime partition and uses a single global uncertainty ordering with the same total number of calibration cells.

Table 2: **Internal controls on CIFAR-100 ViT at $C = 2K$.** The setting is shown because both Random and GlobalU are available under the at the matched $2K$ total-cell budget. Complete internal controls across the available datasets are reported in Appendix F.

Method	MC	WPC	Size
Uniform-$q2$	0.9319 ± 0.0024	0.8016 ± 0.0060 ●	2.5615 ± 0.1685
Random-$2K$	0.9334 ± 0.0022	0.8218 ± 0.0041	2.5569 ± 0.1409
GlobalU-$2K$	0.9402 ± 0.0025	0.8685 ± 0.0215	2.6198 ± 0.1496
BRC-$2K$	0.9361 ± 0.0013	0.8621 ± 0.0159 ▲	2.5352 ± 0.1027

Random improves over Uniform in this setting but remains below BRC, showing that arbitrary nonuniformity alone does not reproduce the response-based allocation effect. GlobalU is competitive, indicating that global uncertainty structure can also be informative in some settings. The complete control suite, including GlobalCP, PatternCP, Uniform, Random, GlobalU, and BRC where available, is reported in Appendix F.

4.4 Broader Evaluation

Regime-constructor modularity. The canonical experiments use KMeans to form the upstream regime map, whereas BRC acts on the resulting frozen regimes. Constructor-modularity experiments replace KMeans with non-KMeans alternatives while leaving the downstream allocation procedure unchanged. The tested alternative constructors retain positive matched $2K$ effects, supporting the interpretation that the allocation effect is not specific to KMeans. Complete constructor definitions and results are reported in Appendix H.3, while forced-regime-count sensitivity is reported in Appendix H.

External conformal comparison. We additionally compare BRC with methods representing complementary approaches to local conformal reliability: PLCP (Kiyani et al., 2024) for learned partitioning, NCP (Ghosh et al., 2023) for representation-space neighborhood localization, Clustered CP (Ding et al., 2023) for grouped calibration, and EnergyAPS (Akhavan Attar et al., 2026) for adaptive conformity scoring. These methods adapt different components of the conformal pipeline, whereas BRC keeps the upstream regime structure fixed and reallocates a fixed total calibration-resolution budget across those regimes.

Across the 13 reported external classification comparisons, BRC-$2K$ attains numerically higher WPC in 12. It has higher WPC than PLCP in all four reported Adult/Folktables settings, NCP on CIFAR-100 ResNet-50, all four reported Clustered CP settings, and three of four EnergyAPS settings. The remaining comparison is CIFAR-100 ViT, where BRC-$2K$ and EnergyAPS obtain

WPC 0.8621 and 0.8624, respectively. Complete MC, WPC, prediction-set size, dataset-specific baseline coverage, and implementation provenance are reported in Appendix J.

Regression transfer. The same fixed-budget formulation is evaluated on Bike Sharing and California Housing using absolute-residual conformity scores. Across the six matched BRC–Uniform comparisons over the three budgets, three differences are positive and three are negative, indicating mixed budget-dependent transfer. The appendix reports the complete matched results together with FeatureCP (Teng et al., 2023), RCP (Plassier et al., 2025), PLCP(Kiyani et al., 2024), and LOCART (Cabezas et al., 2025) external regression comparisons; see Appendix J.3.

Robustness and sensitivity. The primary $2K$ effect remains positive under both tested conformity scores on CIFAR-100 ViT, across all tested final-CAL fractions in the representative CIFAR-100 ViT and Folktables settings, and under all tested uncertainty-coordinate constructions on CIFAR-100 ViT. Regime-count and target-coverage sweeps show that the magnitude of the effect varies with the supplied geometry and operating point, while the tested STRUCTURE-support thresholds leave the selected $K = 2$ results unchanged. Complete score, regime-count, target-coverage, support, fallback, CAL-fraction, uncertainty, and representation analyses are reported in Appendix H. Objective-component ablations and runtime results are reported in Appendices I and K, respectively.

5 Discussion and Limitations

BRC is most useful when calibration resolution is constrained and regimes differ meaningfully in their response to refinement. Its advantage over matched Uniform can diminish when equal refinement is already sufficiently fine, when the supplied regime geometry is poorly aligned with the relevant reliability structure, or when a simpler localization already captures that structure. Because allocation is estimated from finite STRUCTURE-side OOF responses, its ordering need not transfer exactly to an independent finite TEST sample. Sensitivity results further show dependence on the conformity score and target coverage, while regression transfer is more mixed than classification. BRC assumes a supplied regime map and externally specified budget C and candidate-resolution grid $\mathcal{B}$; it does not jointly learn these components. Its validity is the standard groupwise split-conformal guarantee for supported frozen calibration cells, not unrestricted pointwise conditional coverage. External methods can use different conditioning structures or score constructions, so those comparisons provide broader empirical context rather than the matched allocation test. Detailed sensitivity and failure-mode results are reported in Appendix H.

6 Conclusion

We introduced Budgeted Regime Calibration (BRC), which allocates a fixed calibration-resolution budget across preformed regimes according to their estimated refinement responses. Matched experiments show that response-based allocation is most useful under constrained resolution, while equal or simpler refinement can suffice in other settings. These results establish calibration-resolution allocation as a distinct design problem for locally reliable conformal prediction.

AI Use Statement

We used large language models to assist with writing and language polishing, literature retrieval and discovery, and research ideation and execution. All methodological decisions, experimental procedures, results, analyses, and claims were reviewed and verified by the authors.

Reproducibility Statement

Appendix C gives the full BRC procedure, including regime construction, candidate resolutions, OOF response estimation, exact allocation, calibration, and validity. Appendix D gives data splits, seeds, support thresholds, scores, evaluation settings, and benchmark configurations. All STRUCTURE decisions are frozen before independent CAL calibration, and TEST is used only for evaluation. Code and configurations are available at `https://anonymous.4open.science/r/budgeted-regime-calibration-iclr2027-anonymous-F710`.

REFERENCES

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2f5e4c7013a6b810c27c347e8ef57d15-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2009.14193.

Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021. doi:10.1609/aaai.v35i8.16826.

Ali Baheri and Marzieh Amiri Shahbazi. Conformal prediction across scales: Finite-sample coverage with hierarchical efficiency. *Results in Applied Mathematics*, 26:100589, 2025. doi:10.1016/j.rinam.2025.100589.

Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2581–2602. PMLR, 2025. URL https://proceedings.mlr.press/v267/bairaktari25a.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi:10.1093/imaiai/iaaa017.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi:10.52202/068431-2129.

Louis Berthier, Ahmed Shokry, Maxime Moreaud, Guillaume Ramelet, and Aymeric Dieuleveut. Self-organized conformal prediction: Reducing regional coverage gaps with unsupervised group discovery, 2026. URL https://arxiv.org/abs/2606.29403.

Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. doi:10.1016/S0168-1699(99)00046-0.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas J-B. Brunel. Automatically adaptive conformal risk control. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 19–27. PMLR, 2025. URL https://proceedings.mlr.press/v258/blot25a.html.

Luben M. C. Cabezas, Mateus P. Otto, Rafael Izbicki, and Rafael B. Stern. Regression trees for fast and adaptive prediction intervals. *Information Sciences*, 686:121369, 2025. doi:10.1016/j.ins.2024.121369.

Luben Miguel Cruz Cabezas, Vagner Silva Santos, Thiago Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 443–470. PMLR, 2025. URL https://proceedings.mlr.press/v286/cruz-cabezas25a.html.

Sayan Das, Bahram Yaghooti, Todd A. Kuffner, and Soumendra N. Lahiri. On optimal data splitting for split conformal prediction, 2026. URL https://arxiv.org/abs/2606.31600.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

Samuel Deng, Navid Ardeshir, and Daniel Hsu. Group conditional validity via multi-group learning, 2023. URL https://arxiv.org/abs/2303.03995.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

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*, volume 36, pp. 64555–64576, 2023. doi:10.52202/075280-2817.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/855b928e3f64498da7d98a3996bd4971-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

John C. Duchi. Sample-conditional coverage in split-conformal prediction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96470–96500, 2025. doi:10.52202/085713-2903.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3):113–127, 2014. doi:10.1007/s13748-013-0040-3.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Adaptive coverage policies in conformal prediction. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3196–3204. PMLR, 2026. URL https://proceedings.mlr.press/v300/gauthier26a.html.

Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7722–7730, 2023. doi:10.1609/aaai.v37i6.25936.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025. doi:10.1093/jrsssb/qkaf008.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.11959.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023. doi:10.1093/biomet/asac040.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction, 2022. URL https://arxiv.org/abs/2206.13092.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi:10.1109/CVPR.2016.90.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025. doi:10.1093/jrsssb/qkae103.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3068–3077. PMLR, 2020. URL https://proceedings.mlr.press/v108/izbicki20a.html.

Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022. URL https://www.jmlr.org/papers/v23/20-797.html.

Sooyong Jang and Insup Lee. Quantifying and improving adaptivity in conformal prediction through input transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22146–22154, 2026. doi:10.1609/aaai.v40i26.39370.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15145.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24749–24769. PMLR, 2024. URL https://proceedings.mlr.press/v235/kiyani24a.html.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996. URL https://cdn.aaai.org/KDD/1996/KDD96-033.pdf.

Jingsen Kong, Wenlu Tang, Dezheng Kong, Linglong Kong, Guangren Yang, and Bei Jiang. Adaptive conformal prediction via mixture-of-experts gating similarity. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/97c8a8eb0e5231d107d0da51b79e09cb-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Tomer Lavi, Bracha Shapira, and Nadav Rappoport. Cluster frequency conformal prediction for local coverage, 2026. URL https://arxiv.org/abs/2605.24872.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014. doi:10.1111/rssb.12021.

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. doi:10.1080/01621459.2017.1307116.

Wangcheng Li, Nan Qiao, Xu Guo, and Wenguang Sun. COINS: Any-stage-valid and utility-oriented sequential conformal prediction, 2026. URL https://arxiv.org/abs/2609.07112.

Natalia Martinez Gil, Dhaval Patel, Chandra Reddy, Giri Ganapavarapu, Roman Vaculin, and Jayant Kalagnanam. Identifying homogeneous and interpretable groups for conformal prediction. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2471–2485. PMLR, 2024. URL https://proceedings.mlr.press/v244/martinez-gil24a.html.

Enver Menadjiev, Jihyeon Seong, Jisu Yeo, and Jaesik Choi. DistMatch: Adaptive binning via distribution matching for robust sequential conformal prediction. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.00690.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi:10.1016/S0167-7152(96)00140-X.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49459–49492. PMLR, 2025. URL https://proceedings.mlr.press/v267/plassier25a.html.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.03222.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.02544.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. URL https://www.jmlr.org/papers/v9/shafer08a.html.

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.00173.

Paolo Toccaceli. CRPS-optimal binning for conformal regression, 2026. URL https://arxiv.org/abs/2603.22000.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490. PMLR, 2012. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine, 2003. URL https://alrw.net/old/04.pdf. On-line Compression Modelling project working paper.

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

Congbin Xu, Yue Yu, Zhaojun Wang, Changliang Zou, and Haojie Ren. Aggregating conformal prediction sets via α-allocation. *Biometrika*, 113(3):asag048, 2026a. doi:10.1093/biomet/asag048.

Senrong Xu, Yanke Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. Fair conformal classification via learning representation-based groups. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0bcfbd99d8a15a5aeac394578ee6a39a-Abstract-Conference.html.

Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively selected groups. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108760–108823, 2024. doi:10.52202/079017-3454.

Appendix

Table of Contents

A Extended Related Work 16

A.1 Marginal, Conditional, and Group Reliability 16

A.2 Localized Conformal Calibration 16

A.3 Learned Representations, Partitions, and Groups 17

A.4 Adaptive Conformity Scores and Coverage Mechanisms 17

A.5 Adaptive Granularity, Binning, and Multi-Scale Calibration 18

A.6 Allocation of Statistical Resources 18

A.7 Positioning of Budgeted Regime Calibration 19

B Development Diagnostics Leading to BRC 19

B.1 Residual Variation Within Coarse Regimes 20

B.2 From Sample-Level Hardness to Regime-Level Diagnostics 20

B.3 Regime-Dependent Uncertainty–Miscoverage Association 20

B.4 Signed Uncertainty Routing 21

B.5 Calibration Resolution as a Decision Variable 21

B.6 Binary Miss-Based Selective Refinement 22

B.7 Cross-Dataset Behavior of Selective Refinement 22

B.8 Representative Development Diagnostics 22

B.9 From Development Diagnostics to Resolution Response 22

B.10 From Resolution Response to a Shared Budget 23

C Method Details 24

C.1 Conformity Scores 24

C.2 Canonical Predictive-Regime Construction 25

C.3 Within-Regime Candidate-Resolution Construction 25

C.4 Common Audit Geometry and OOF Response Estimation 26

C.5 Fixed-Budget Allocation 27

C.6 Proof of the Matched-Uniform STRUCTURE Property 27

C.7 Exact Threshold-First Dynamic Program 28

C.8 Independent Final Calibration and Support 29

C.9 Validity of Frozen Final Calibration 30

C.10 Methodological Scope 30

D Experimental Protocol 30

D.1 Data Partitioning and Data Roles 30

D.2 Datasets and Predictors 31

D.3 Experimental Instantiations 32

D.4 Fixed Structural and Calibration Settings 32

D.5 Resolution Budgets and Matched Comparison 32

D.6 Evaluation Metrics and Common Audit Geometry 33

D.7 Support and Fallback Protocol . . . 33
D.8 Baselines and Comparison Rationale . . . 33
D.9 Robustness Protocol . . . 34
D.10 Seeds and Reporting Convention . . . 34
E Complete Matched Classification Results . . . 35
F Complete Internal Classification Controls . . . 36
G Mechanism Evidence . . . 40
G.1 Regime-Specific Resolution Responses . . . 40
G.2 Downstream Resolution-Demand Check . . . 41
G.3 Resolution Demand and Generic Heterogeneity . . . 42
H Robustness and Failure Modes . . . 42
H.1 Conformity-Score Robustness . . . 43
H.2 Sensitivity to Upstream Regime Count . . . 43
H.3 Regime-Constructor Modularity . . . 44
H.4 Target-Coverage Sensitivity . . . 45
H.5 STRUCTURE-Support Sensitivity . . . 45
H.6 Final-CAL Fallback . . . 45
H.7 Uncertainty Construction . . . 46
H.8 Final-CAL Fraction Sensitivity . . . 46
H.9 Input-Space Regime Construction . . . 47
I Objective Ablation . . . 47
J External Conformal Baselines . . . 48
J.1 Tabular Classification . . . 48
J.2 Image Classification . . . 48
J.3 Regression Baselines . . . 49
K Runtime. . . . 50

A Extended Related Work

This appendix expands the literature positioning summarized in Section 1.1. BRC intersects several established lines of conformal prediction research: conditional and group reliability, localized calibration, learned conditioning structures, adaptive conformity mechanisms, adaptive discretization, and statistical-resource allocation. Prior work establishes important precedents for local calibration, learned groups, adaptive partitions, variable granularity, and allocation. BRC focuses on a distinct downstream decision that arises after a coarse regime structure has been formed: the total number of local calibration cells is treated as a fixed resource, and their distribution across regimes is selected from estimated regime-specific responses to further refinement.

A.1 Marginal, Conditional, and Group Reliability

Conformal prediction provides finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018). Marginal validity, however, averages over the covariate distribution and does not generally imply comparable coverage at every covariate value or within every subpopulation. Early work on conditional conformal validity and distribution-free prediction bands made this distinction explicit (Vovk, 2012; Lei & Wasserman, 2014), while subsequent impossibility results characterized the limitations of obtaining nontrivial distribution-free conditional coverage without additional assumptions (Barber et al., 2021). More recently, Duchi (Duchi, 2025) revisits sample-conditional validity for split conformal prediction and develops guarantees conditional on the validation sample together with results for approximate group-conditional coverage. These results motivate structured relaxations of unrestricted pointwise conditional validity rather than suggesting that marginal coverage alone captures local reliability.

One early structured approach is Mondrian conformal prediction, which performs conformal calibration within a prespecified taxonomy or partition (Vovk et al., 2003; Vovk, 2012). Later work substantially broadens the group-conditional perspective. Practical adversarial multivalid conformal prediction and batch multivalid conformal prediction seek validity over rich, potentially overlapping collections of groups (Bastani et al., 2022; Jung et al., 2023), while multi-group learning provides another route to simultaneous validity over structured group families (Deng et al., 2023). Gibbs et al. (Gibbs et al., 2025) develop conditional guarantees relative to specified classes of weighting functions, and Kandinsky conformal prediction extends group-aware coverage beyond conventional disjoint class or covariate groups (Bairaktari et al., 2025).

Adaptive group selection has also become an explicit object of study. Zhou and Sesia (Zhou & Sesia, 2024) construct classification sets with equalized coverage conditional on adaptively selected features, showing that the groups on which reliability is assessed need not be fixed manually in advance.

This literature establishes the distinction between marginal validity and stronger forms of localized or group-aware reliability, and it provides several mechanisms for obtaining calibration guarantees over structured subpopulations. BRC addresses a different stage of the pipeline. It conditions on an already formed coarse structure and asks how finite calibration resolution should be distributed across that structure under a shared total budget. Thus BRC does not introduce a new definition of conditional validity; it studies how an existing local calibration structure should receive additional resolution.

A.2 Localized Conformal Calibration

A second line of work replaces global calibration with mechanisms that emphasize calibration observations relevant to the particular test point. Localized Conformal Prediction (LCP) (Guan, 2023) introduces test-dependent weighting of conformal scores according to a localizer around the test covariate. It retains a distribution-free finite-sample marginal guarantee and provides additional local guarantees under appropriate assumptions. Split Localized Conformal Prediction develops a related split-conformal approach to local adaptation (Han et al., 2022). Randomized local weighting has subsequently been studied as a mechanism for obtaining robust local-weight guarantees (Hore & Barber, 2025).

Localization can also be performed in learned representation spaces. Neighborhood Conformal Prediction (NCP) (Ghosh et al., 2023) identifies nearby calibration examples through neural representations and uses this neighborhood information to improve uncertainty quantification for deep classifiers. Related density-based constructions use estimated conditional structure to produce flexible predictive bands or high-dimensional conformal regions (Izbicki et al., 2020; 2022).

These approaches primarily determine which calibration observations are locally relevant to a test point or how those observations should be weighted. BRC operates downstream of that localization question. In our instantiation, a test point is routed through a frozen regime and within-regime cell structure, while the allocation decision determines how much candidate calibration resolution each regime receives under a common total budget. The central BRC variable is therefore not a test-point-specific neighborhood or weighting rule, but the amount of resolution assigned across the already formed regimes.

A.3 Learned Representations, Partitions, and Groups

A particularly close literature learns the geometry on which conformal calibration operates. Feature Conformal Prediction uses predictive representations as part of conformal inference (Teng et al., 2023). Partition Learning Conformal Prediction (PLCP) directly learns uncertainty-guided features and a finite partition from calibration information to improve conditional validity (Kiyani et al., 2024). The number and structure of uncertainty regions therefore already constitute adaptive design choices in prior conformal work.

Martinez Gil et al. (Martinez Gil et al., 2024) learn interpretable partitions of the input space so that nonconformity-score behavior is comparatively homogeneous within the resulting groups. The discovered partition can then be combined with group-conditional conformal calibration. LOCART uses regression-tree structure to construct adaptive local prediction intervals (Cabezas et al., 2025). In many-class classification, Clustered CP groups classes with similar conformity-score distributions so that classes with limited calibration support can share information while targeting class-conditional coverage (Ding et al., 2023).

Recent methods extend this direction in several ways. Zhou and Sesia (Zhou & Sesia, 2024) select features adaptively when constructing groups for equalized coverage. FaReG learns representation-based groups for fair conformal classification (Xu et al., 2026b). Cluster Frequency Conformal Prediction (CFCP) clusters learned embeddings, estimates cluster-level label-frequency information, and uses that local information in sample-specific probability construction before conformalization (Lavi et al., 2026). Self-Organized Conformal Prediction (SOCP) discovers input-space groups through self-organizing maps and performs group-local calibration (Berthier et al., 2026). Mixture-of-Experts conformal methods likewise exploit gating-based similarity as information about latent local structure (Kong et al., 2026).

Accordingly, regime construction is an upstream conditioning step in BRC rather than the central allocation contribution. Our implementation provides one concrete way to instantiate this structure from predictor-derived information, but the fixed-budget allocation problem begins once the coarse regimes have been formed. Replacing the upstream regime constructor by another valid partitioning mechanism leaves the downstream allocation problem unchanged. BRC therefore separates the question of how the conditioning geometry is obtained from the question of how much calibration resolution each resulting regime receives.

A.4 Adaptive Conformity Scores and Coverage Mechanisms

Conformal adaptivity can also arise through the score or prediction-set construction rather than through local partitions. In regression, Conformalized Quantile Regression (CQR) (Romano et al., 2019) combines quantile regression with conformal calibration so that interval width can adapt to heteroscedasticity while retaining finite-sample marginal validity. In classification, adaptive prediction-set constructions based on ordered class probabilities provide another influential approach (Romano et al., 2020; Angelopoulos et al., 2021). These methods establish that useful input-dependent behavior can arise from the conformity construction itself even when there is no explicit regime-resolution allocation.

More recent work directly learns or augments conformity mechanisms. Rectified Conformal Prediction (RCP) (Plassier et al., 2025) estimates a conditional quantile of an existing conformity score and learns a transformation intended to make the relevant score quantile more homogeneous across covariates. Standard conformal calibration is then applied to the transformed score, preserving exact marginal coverage while targeting improved approximate conditional behavior. EPICSCORE incorporates epistemic uncertainty into the conformity-score construction (Cruz Cabezas et al., 2025), while energy-based adaptive conformal classification investigates alternatives to relying exclusively on softmax-derived uncertainty (Akhavan Attar et al., 2026).

Other approaches adapt the object being controlled. Jang and Lee (Jang & Lee, 2026) study conformal adaptivity through difficulty-aware input transformations. Automatically Adaptive Conformal Risk Control learns data-dependent conditioning structure for approximately conditional control of statistical risks (Blot et al., 2025). Adaptive Coverage Policies instead allow the desired coverage behavior itself to depend on the input and optimize such policies from calibration information (Gauthier et al., 2026). Related specialized methods address settings such as long-tailed classification (Ding et al., 2026) and latent mixture structure (Kong et al., 2026).

BRC operates on a different design variable. In the matched BRC-versus-Uniform comparison, the predictor and conformity score are held fixed, and BRC does not learn an input-specific target coverage level or a transformation of the conformity score. Predictive uncertainty provides an ordering for candidate refinement within each regime, whereas the amount of refinement assigned to that regime is selected from its estimated calibration response. Thus the role of uncertainty in BRC is to specify where candidate refinement occurs, while the observed calibration response determines how much resolution a regime receives.

A.5 Adaptive Granularity, Binning, and Multi-Scale Calibration

Adaptive granularity is also established in the conformal literature. PLCP learns a finite uncertainty partition whose complexity controls the approximation of conditional score behavior (Kiyani et al., 2024); homogeneous-group and tree-based methods adapt the partition according to observed calibration structure (Martinez Gil et al., 2024; Cabezas et al., 2025). These results establish adaptive partition complexity and calibration granularity as existing conformal design choices. BRC instead couples the candidate resolutions of multiple regimes through a shared total cell budget.

More direct binning methods reinforce this distinction. CRPS-optimal binning selects data-dependent bin boundaries by optimizing a leave-one-out CRPS criterion and solves the resulting fixed-K partitioning problem through dynamic programming (Toccaceli, 2026). DistMatch develops adaptive binning through distribution matching for sequential conformal prediction (Menadjiev et al., 2026). These methods make the partition or binning rule itself an adaptive object.

Multi-scale conformal prediction provides another relevant precedent. Conformal Prediction Across Scales constructs conformal predictors at multiple resolutions and combines them while distributing miscoverage across scales (Baheri & Amiri Shahbazi, 2025). This establishes that scale and resolution can themselves be meaningful conformal design variables.

The BRC decision differs through the coupling imposed across simultaneously available regimes. It begins with a finite set of candidate within-regime resolutions and an externally specified total number of cells. Selecting finer resolution for one regime therefore uses resolution that cannot simultaneously be assigned to another regime. The central problem is consequently not whether adaptive binning or multiple scales are useful, but how a conserved amount of local calibration resolution should be distributed across a fixed set of coarse regimes.

A.6 Allocation of Statistical Resources

The broader concept of allocation also appears in conformal prediction. Several recent methods explicitly optimize how a constrained statistical resource is distributed.

COLA allocates miscoverage levels across multiple conformal prediction sets when constructing an aggregated prediction set (Xu et al., 2026a). Its allocation variable is therefore the coverage or miscoverage assigned to component sets rather than the number of local calibration cells assigned to heterogeneous regions. Work on optimal data splitting studies another resource: how a finite sample

should be divided between model training and conformal calibration to improve the efficiency of split conformal prediction (Das et al., 2026).

Sequential conformal prediction introduces still another allocation problem. COINS coordinates a finite-sample budget across successive stages of a prediction process and develops utility-oriented stagewise allocation while targeting any-stage validity (Li et al., 2026). Multi-scale conformal prediction also distributes miscoverage across scales (Baheri & Amiri Shahbazi, 2025). Together, these methods demonstrate that resource-allocation formulations can be useful in conformal inference, while also illustrating that the allocated statistical object depends on the inferential problem.

BRC studies a different allocated object: *calibration resolution*. Specifically, it distributes a fixed total number of local calibration cells across already formed regimes. The resulting decisions are coupled because additional resolution assigned to one regime is unavailable to the others. This coupling separates fixed-budget calibration-resolution allocation from miscoverage allocation, training-calibration sample allocation, and sequential resource allocation.

A.7 Positioning of Budgeted Regime Calibration

The literature above establishes several closely related design choices in conformal prediction. Localized methods determine which calibration observations should influence a test point (Guan, 2023; Han et al., 2022; Ghosh et al., 2023); learned-partition and group-based methods determine the conditioning structure on which calibration operates (Kiyani et al., 2024; Martinez Gil et al., 2024; Cabezas et al., 2025; Ding et al., 2023; Xu et al., 2026b); and adaptive-score methods modify the conformity construction or its relationship to local difficulty (Romano et al., 2020; Plassier et al., 2025; Cruz Cabezas et al., 2025; Akhavan Attar et al., 2026). Adaptive binning and multi-scale procedures further show that calibration granularity can itself be data dependent (Toccaceli, 2026; Menadjiev et al., 2026; Baheri & Amiri Shahbazi, 2025).

BRC addresses a complementary downstream decision. It assumes that a coarse regime structure has already been formed and that the total amount of local calibration resolution is fixed. The remaining question is how that finite resolution should be distributed across the existing regimes. This distinguishes BRC from methods that primarily learn the conditioning geometry, construct test-dependent neighborhoods, modify conformity scores, or adapt the discretization itself. The distinction is also different from other allocation formulations in conformal prediction: COLA allocates miscoverage across component prediction sets (Xu et al., 2026a), optimal split-conformal procedures allocate observations between training and calibration (Das et al., 2026), and COINS allocates resources across stages of a sequential prediction process (Li et al., 2026). BRC instead treats calibration resolution across already formed regimes as the allocatable resource.

This positioning determines the appropriate empirical comparisons. Matched Uniform refinement is the primary control because it preserves the predictor, conformity score, regime structure, data assignments, and total calibration resolution, isolating the effect of allocation. External localized, partition-based, and score-adaptive conformal methods answer a broader question of empirical relevance rather than the same matched allocation question. Because such methods may differ in conditioning structure, conformity score, coverage target, or efficiency operating point, we interpret their results jointly through marginal coverage, local reliability, and prediction efficiency rather than as a single directly matched ranking.

B Development Diagnostics Leading to BRC

This appendix summarizes the development diagnostics that narrowed the methodological question ultimately addressed by Budgeted Regime Calibration (BRC). These experiments are not additional BRC benchmarks and are not ablations of the final algorithm. Their purpose is to identify which simpler hypotheses were insufficient and to explain why calibration resolution emerged as the relevant decision variable. These experiments use the representative development seed specified below and do not constitute final BRC evaluation, ablation, or model selection on TEST.

The intermediate development stages were examined over multiple runs, but their protocols differed and should not be pooled into a common statistical comparison. For compactness, we report one representative run, seed 42, from each relevant stage. The displayed values illustrate the diagnostic

behaviors that motivated subsequent hypotheses; the development decisions were not based on a single reported run. We therefore use these examples to document the methodological progression, not to estimate final effect sizes or establish general empirical superiority.

The diagnostics below also do not establish that BRC is correct, optimal, or necessarily superior to uniform refinement. They narrow the unresolved question. The resulting BRC hypothesis is evaluated independently under the frozen final protocol in Section 4.

B.1 Residual Variation Within Coarse Regimes

The development process started from regime-specific conformal calibration. Let

$$r : \mathcal{X} \to \{1, \ldots, K\}$$

denote a frozen regime map and

$$P_k = \{x : r(x) = k\}$$

denote regime k.

Separate calibration within coarse regimes can address reliability variation hidden by a single global threshold. However, acceptable aggregate coverage within a regime does not imply homogeneous calibration behavior inside that regime. Development diagnostics revealed residual within-regime coverage variation after coarse regime formation.

This observation shifted the question from whether regimes should be calibrated separately to what information can determine whether additional calibration resolution is useful within an already formed regime.

B.2 From Sample-Level Hardness to Regime-Level Diagnostics

An earlier hypothesis was that a scalar sample-level hardness measure could directly identify observations requiring additional protection. One explored construction modified the nonconformity score as

$$R_\beta(x, y) = R(x, y) + \beta h(x),$$

where $h(x)$ denotes predictive hardness.

Under the nonconformity convention used here, the inclusion condition

$$R(x, y) + \beta h(x) \leq q$$

is equivalent to

$$R(x, y) \leq q - \beta h(x).$$

For $\beta > 0$, increasing hardness therefore decreases the allowable base nonconformity score. This does not provide the intended general interpretation that harder observations automatically receive greater protection.

More broadly, subsequent diagnostics did not support a stable monotone relationship between scalar hardness or uncertainty and residual coverage. Direct sample-level hardness adjustment was therefore not retained as the general mechanism. The subsequent development focused instead on residual calibration behavior conditional on already formed regimes.

B.3 Regime-Dependent Uncertainty–Miscoverage Association

We next examined whether predictive uncertainty could provide a consistent within-regime risk ordering. Using STRUCTURE data, define the out-of-fold miss indicator

$$M_i = \mathbf{1}\left\{Y_i \notin \widehat{\Gamma}_\alpha^{(-f(i))}(X_i)\right\},$$

where $f(i)$ denotes the fold containing observation i.

Within regime P_k, we summarized the association between uncertainty U and the miss indicator by

$$A_k = \mathrm{AUC}(U, M \mid P_k).$$

Values above $1/2$ indicate that larger uncertainty is associated with more misses, whereas values below $1/2$ indicate the opposite orientation.

For the representative Adult development run with seed 42, the three regimes exhibited directions

$$(+, -, +),$$

with stable orientation across the out-of-fold partitions. This example illustrates the broader development observation that uncertainty and residual miscoverage need not follow a common monotone direction across regimes.

The result does not imply that uncertainty is uninformative. Rather, uncertainty magnitude alone did not provide a sufficiently general rule for determining how much calibration resolution a regime should receive.

B.4 Signed Uncertainty Routing

The regime-dependent orientation motivated a signed-routing prototype that estimated the uncertainty direction separately in each regime. Define

$$d_k = \begin{cases} +1, & A_k \geq 1/2, \\ -1, & A_k < 1/2, \end{cases}$$

and let

$$z_k(x) = d_k \left(\mathrm{rank}_{P_k}(U(x)) - \frac{1}{2} \right)$$

denote the corresponding signed within-regime uncertainty position.

This prototype tested whether recovering the direction of the uncertainty–miscoverage relationship was sufficient to identify useful local calibration structure.

In the historical Adult Phase-2 experiment with seed 42 and full calibration support, worst fine-grid coverage was approximately

$$0.7014$$

for signed routing. Under the same development protocol, uniform within-regime refinements gave

Resolution	$q2$	$q3$	$q4$	$q6$
Worst fine-grid coverage	0.7590	0.7943	0.8636	0.8226

.

The representative run illustrates two observations that recurred during development. First, recovering the direction of residual risk was not sufficient to determine the useful amount of refinement. Second, the response to refinement need not be monotone: in this case $q4$ exceeded $q6$.

We therefore did not retain signed routing as the general mechanism. Instead, the next stage treated the amount of calibration resolution as a separate decision variable.

B.5 Calibration Resolution as a Decision Variable

For a common resolution b, uniform within-regime refinement assigns

$$B_1 = \cdots = B_K = b,$$

with each regime subdivided according to a frozen within-regime uncertainty ordering.

This formulation separates two roles. Predictive uncertainty determines where observations lie along the refinement coordinate, whereas B_k determines how much calibration resolution regime k receives. There is no requirement that the useful value of B_k be a monotone function of uncertainty magnitude.

Resolution also interacts with finite calibration support. Increasing the number of cells produces more localized calibration thresholds but reduces the number of calibration observations supporting each threshold. The development evidence therefore did not justify either minimal refinement or maximal refinement as a universal choice.

B.6 Binary Miss-Based Selective Refinement

We next tested whether refinement could be selected directly from out-of-fold calibration failures. The corresponding prototype used the binary variable

$$M_i \in \{0, 1\}$$

and introduced candidate uncertainty splits according to differences in empirical miss rates between the resulting child regions. Structural decisions were made using STRUCTURE and frozen before final calibration.

This experiment tested a stronger adaptive hypothesis: additional resolution would be introduced only where a binary miss diagnostic supported further refinement.

For Adult with seed 42 and full calibration support, the selective prototype obtained worst fine-grid coverage of approximately

$$0.485,$$

whereas the STRUCTURE-selected common resolution obtained approximately

$$0.729.$$

The corresponding structures used approximately 11 and 18 cells, respectively.

This representative case illustrates the failure mode that motivated us to reconsider binary refine-or-not-refine decisions. The broader development branch was evaluated beyond this displayed run; the single example is reported here only to keep the development record compact.

The evidence does not establish a specific causal explanation for the failure. In particular, we do not claim that binary miss indicators necessarily discard the information required for useful refinement. The supported conclusion is narrower: this binary decision rule did not provide a sufficiently reliable general mechanism for determining required calibration resolution.

B.7 Cross-Dataset Behavior of Selective Refinement

The Adult result alone could incorrectly suggest that selective refinement was intrinsically ineffective. The Housing development experiment provides an important counterexample.

For seed 42 at 50% calibration support, selective refinement obtained worst fine-grid coverage of approximately

$$0.838,$$

compared with

$$0.810$$

for the STRUCTURE-selected common resolution. The selective structure used approximately 4 cells, compared with 12 cells for the common-resolution alternative.

This example shows that selective refinement could identify a compact and useful structure in some settings. The development decision was therefore not based on the claim that selective refinement always failed. Rather, its behavior varied substantially across settings, and the binary rule was not sufficiently reliable to serve as the general resolution-selection mechanism.

This motivated replacing the binary refine-or-not-refine decision with direct evaluation of several candidate resolutions.

B.8 Representative Development Diagnostics

Table 3 summarizes the representative evidence used to narrow the methodological question. Results from different development phases should be interpreted within their respective protocols rather than as a single matched benchmark.

B.9 From Development Diagnostics to Resolution Response

Taken together, the development evidence did not support three simple mappings as general decision rules:

$$\text{scalar hardness} \longrightarrow \text{required protection},$$

Table 3: **Representative development diagnostics for seed 42.** The displayed cases illustrate methodological transitions observed during development and are not part of the final BRC evaluation. Phase-2 and Phase-3 used different development protocols.

Development question	Setting	Representative evidence	Consequence
Is the uncertainty–risk direction universal?	Adult direction diagnostic	The three regimes exhibited directions $(+, -, +)$ with stable OOF orientation.	A single monotone uncertainty–risk rule was not supported.
Is learning the uncertainty direction sufficient?	Adult Phase-2, full CAL	Signed routing: 0.7014; uniform $q4$: 0.8636 worst fine-grid coverage.	Direction alone did not determine the useful amount of refinement.
Does binary selective refinement reliably determine resolution?	Adult Phase-3, full CAL	Selective: 0.485 with ~ 11 cells; selected common resolution: 0.729 with 18 cells.	The binary rule was not sufficiently reliable as a general resolution-selection mechanism.
Can selective refinement nevertheless be useful?	Housing Phase-3, 50% CAL	Selective: 0.838 with ~ 4 cells; selected common resolution: 0.810 with 12 cells.	Yes, but its benefit was setting-dependent rather than reliably captured by one binary decision rule.

$$\text{uncertainty direction} \longrightarrow \text{required resolution},$$

or

$$\text{binary refine/no-refine decision} \longrightarrow \text{required resolution}.$$

These observations suggested a different object of interest: the response of calibration quality to the candidate resolution itself.

For each already formed regime P_k and candidate resolution $b \in \mathcal{B}$, let

$$E_k(b)$$

denote an out-of-fold estimate of calibration discrepancy obtained when regime k is assigned resolution b.

The methodological change is therefore from inferring required resolution indirectly from a generic hardness or routing statistic to estimating the resolution response directly. No monotonicity assumption is imposed on $E_k(b)$.

This formulation preserves the role of predictive uncertainty as a within-regime refinement coordinate while separating that role from the decision of how much refinement a regime should receive.

B.10 From Resolution Response to a Shared Budget

Resolution-response estimation alone does not define BRC. If every regime independently selected its preferred candidate,

$$B_k^{\mathrm{ind}} = \arg\min_{b \in \mathcal{B}} E_k(b),$$

there would be no coupling among regimes. Increasing the resolution of one regime would impose no cost on another.

BRC instead specifies a fixed total calibration-cell budget

$$\sum_{k=1}^{K} B_k = C.$$

Fixing C separates the allocation question from the trivial effect of increasing total resolution. Under a common budget, assigning an additional cell to one regime consumes resolution that could otherwise be allocated elsewhere.

The resulting methodological question is therefore: given already formed regimes, estimated responses to candidate calibration resolutions, and a fixed total number of calibration cells, how should that finite resolution be distributed across regimes?

This is the fixed-budget calibration-resolution problem addressed by BRC.

The preceding diagnostics do not imply that the fixed-budget formulation is optimal, nor do they establish that nonuniform allocation must outperform uniform refinement. They motivate the decision variable and the allocation problem. The value of the proposed allocation rule is evaluated separately through matched-budget comparisons that hold the predictor, regime structure, conformity score, data roles, and total number of calibration cells fixed.

Appendix C provides the complete BRC specification, including candidate-response estimation, the fixed audit grid, exact allocation, independent final calibration, fallback behavior, and validity conditions.

C Method Details

This appendix provides the complete specification of Budgeted Regime Calibration (BRC). The main paper presents the allocation problem and its statistical interpretation; here we give the exact construction used to obtain the regimes, candidate cells, out-of-fold response tables, exact-budget allocation, final thresholds, and validity statements.

BRC distinguishes the data roles required for predictor fitting and post-hoc calibration. For the tabular classification and regression experiments, TRAIN fits the predictor and VAL supports ordinary predictive-model selection where required; STRUCTURE determines the regime geometry, refinement references, response estimates, and allocation; CAL estimates final conformal thresholds after routing is frozen; and TEST is reserved exclusively for evaluation.

For CIFAR-100, an ImageNet-pretrained backbone is frozen and used as a feature extractor, while a seed-specific 100-class linear prediction head is fitted on TRAIN with VAL used for predictor selection. STRUCTURE and CAL are drawn from the remaining official CIFAR-100 training set, and the official test set is reserved exclusively for TEST. For ImageNet-1K, the pretrained predictor is fixed before the post-hoc conformal experiment, so the held-out image pool is used only for STRUCTURE, CAL, and TEST. Exact split proportions and dataset-level instantiations are specified in Appendix D.

C.1 Conformity Scores

BRC operates on a fixed conformity score $S(x, y)$. The allocation mechanism does not require a particular score, although the estimated response profiles generally depend on the chosen score.

For tabular classification we use LAC,

$$S_{\mathrm{LAC}}(x, y) = 1 - \widehat{p}_y(x). \tag{17}$$

For image classification we use deterministic nonrandomized APS under the strict-before-candidate convention,

$$S_{\mathrm{APS}}(x, y) = \sum_{j < r_x(y)} \widehat{p}_{(j)}(x), \tag{18}$$

where $\widehat{p}_{(1)}(x) \geq \widehat{p}_{(2)}(x) \geq \cdots$ are the sorted class probabilities and $r_x(y)$ is the rank of candidate class y.

For regression with point predictor $\widehat{\mu}(x)$, we use the absolute residual

$$S_{\mathrm{reg}}(x, y) = |y - \widehat{\mu}(x)|. \tag{19}$$

A threshold q therefore yields

$$\widehat{\Gamma}_\alpha(x) = [\widehat{\mu}(x) - q, \widehat{\mu}(x) + q].$$

The score is held fixed within every matched BRC–Uniform comparison.

C.2 Canonical Predictive-Regime Construction

For each observation, the concrete classification implementation forms a predictor-derived signature

$$z(x) = [\phi(x); \mathbf{u}_{\mathrm{reg}}(x)], \tag{20}$$

where $\phi(x)$ is the frozen representation supplied by the predictor-specific implementation and $\mathbf{u}_{\mathrm{reg}}(x)$ contains frozen predictive-uncertainty statistics.

All preprocessing used to construct the regime geometry is fitted using STRUCTURE only. In particular, standardization parameters are estimated from STRUCTURE and subsequently reused for CAL and TEST. Any backbone-specific representation compression is likewise fitted using STRUCTURE only. Predictor-specific representation extraction is documented with the experimental implementation because it is upstream of BRC rather than part of the allocation objective.

We consider

$$K \in \{2, 3, 5, 7, 10\}. \tag{21}$$

For each K, KMeans uses 20 initializations. Stability is measured by five-fold adjusted Rand index (ARI). A reference KMeans solution is fitted on STRUCTURE; for each fold, another KMeans model is fitted on the remaining STRUCTURE observations and predicts the held-out fold. ARI compares these held-out predictions with the reference assignments.

A candidate K is feasible only if

$$\mathrm{ARI}(K) \geq 0.70$$

and every resulting STRUCTURE regime satisfies the prespecified minimum support: 100 observations for classification and 50 for regression. Among feasible candidates, the implementation selects lexicographically by

$$\text{higher ARI}, \qquad \text{larger minimum regime support}, \qquad \text{smaller } K.$$

The selected KMeans model defines

$$r(x) \in \{1, \ldots, K\}.$$

This construction defines the canonical regime map used in the main experiments. The downstream BRC allocation requires only a frozen regime map and is unchanged when another STRUCTURE-fitted constructor supplies $r(x)$. Five-seed constructor-modularity experiments with non-KMeans regime maps are reported separately in the experimental appendix.

Regression instantiation. For regression, the predictor is a three-member TabNet ensemble and $\widehat{\mu}(x)$ is the ensemble-mean prediction. Let $u_1(x)$ denote the ensemble prediction standard deviation and $u_2(x)$ the ensemble prediction range. The predictive-regime signature concatenates the mask-summary representation $\phi(x)$ from the first TabNet member with $\log(1+u_1(x))$ and $\log(1+u_2(x))$. After STRUCTURE-based standardization, KMeans constructs the predictive regimes using the same ARI threshold of 0.70 and a minimum STRUCTURE-regime support of 50.

Alternative regime constructors. For the constructor-modularity experiments, the canonical KMeans map is replaced by the corresponding non-KMeans STRUCTURE-fitted regime constructor. After $r(x)$ is frozen, the uncertainty coordinate, candidate-resolution set, OOF response estimation, exact-budget allocation, CAL calibration, and TEST evaluation are identical to the canonical BRC pipeline.

C.3 Within-Regime Candidate-Resolution Construction

For classification, let

$$u_1(x) = -\frac{\sum_c \widehat{p}_c(x) \log \widehat{p}_c(x)}{\log |\mathcal{Y}|}$$

denote normalized predictive entropy. Let

$$u_2(x) = 1 - \left(\widehat{p}_{(1)}(x) - \widehat{p}_{(2)}(x)\right)$$

denote inverse top-two margin.

For regime k, define the empirical percentile rank of coordinate u_j relative to the STRUCTURE observations in that regime by

$$R_{j,k}(x) = \frac{\#\{i \in \mathcal{D}_{\text{str}} : r(X_i) = k,\ u_j(X_i) \le u_j(x)\}}{\#\{i \in \mathcal{D}_{\text{str}} : r(X_i) = k\}}. \tag{22}$$

The implementation uses the upper empirical rank for ties. The within-regime coordinate is

$$U(x) = \frac{1}{2}R_{1,r(x)}(x) + \frac{1}{2}R_{2,r(x)}(x). \tag{23}$$

For regression, $u_1(x)$ and $u_2(x)$ are respectively the ensemble prediction standard deviation and prediction range. Their STRUCTURE-referenced percentile ranks are computed within each frozen regime, and the regression uncertainty coordinate uses the same equal-weight form,

$$U(x) = \frac{1}{2}R_{1,r(x)}(x) + \frac{1}{2}R_{2,r(x)}(x).$$

Thus the same equal-mass candidate-resolution construction is used after replacing the classification uncertainty statistics by the two regression ensemble-disagreement statistics.

We use the prespecified candidate-resolution set

$$\mathcal{B} = \{1, 2, 3, 4, 6\}. \tag{24}$$

Here $b = 1$ denotes no within-regime refinement, $b = 2, 3, 4$ correspond to the matched Uniform operating points, and $b = 6$ extends the response grid. The set was frozen before final TEST evaluation. BRC applies to any prespecified finite candidate set for which the requested exact budget is feasible.

For candidate $b > 1$, regime k uses the STRUCTURE quantiles

$$Q_{k,b} = \left\{Q_{U|r=k}(1/b), \ldots, Q_{U|r=k}((b-1)/b)\right\}.$$

For $b = 1$, $Q_{k,1}$ is empty. These boundaries define approximately equal-mass candidate cells.

The same STRUCTURE-derived boundaries are used when routing STRUCTURE, CAL, and TEST observations. In the implementation, observations equal to an interior boundary are assigned deterministically to the higher-index cell. No CAL or TEST labels influence these boundaries.

The role of U is purely to order refinement within a regime. The number B_k of cells assigned to that regime is chosen from the OOF calibration response described next.

C.4 Common Audit Geometry and OOF Response Estimation

Candidate resolutions must be compared on a common evaluation geometry. BRC therefore freezes the STRUCTURE audit partition

$$\mathcal{G}_{\text{audit}} = P \times U_{10}, \tag{25}$$

where each regime is crossed with ten STRUCTURE-defined within-regime uncertainty strata. These audit groups are independent of the candidate resolution b; changing b changes the calibration cells but not the groups on which candidate reliability is measured.

Five OOF folds are assigned within the frozen audit geometry. Fold assignment is balanced inside each audit group so that, when possible, observations from the same audit group are distributed across all five folds.

For held-out fold f, let

$$\mathcal{I}_{-f} = \{i : \text{fold}(i) \neq f\}$$

denote the remaining STRUCTURE observations. A global threshold

$$\widehat{q}_{\text{global},-f} = Q_{\text{conf}}\left(\{S_i : i \in \mathcal{I}_{-f}\}\right)$$

is computed first. For regime k,

$$\widehat{q}_{k,-f} = Q_{\text{conf}}\left(\{S_i : i \in \mathcal{I}_{-f}, r(X_i) = k\}\right)$$

when sufficient training-fold support is available; otherwise the global threshold is used.

For candidate resolution b, the frozen candidate boundaries assign each observation in regime k to one of the b candidate cells. For a held-out observation in candidate cell c, its OOF threshold is estimated from the corresponding cell using only $\mathcal{I}_{-f}$. A local OOF threshold requires at least 80 training-fold observations. If this requirement is not satisfied, the fold-specific parent-regime threshold is used.

Let

$$\widehat{q}^{(b)}_{i,-f(i)}$$

be the resulting OOF threshold for STRUCTURE observation i. For supported audit group $A_{k,d}$, candidate-resolution coverage is

$$\widehat{C}^{(b)}_{k,d} = \frac{1}{|A_{k,d}|} \sum_{i \in A_{k,d}} \mathbf{1}\left\{ S_i \le \widehat{q}^{(b)}_{i,-f(i)} \right\}. \tag{26}$$

The regime-specific primary cost is

$$E_k(b) = \max_d \left| \widehat{C}^{(b)}_{k,d} - (1-\alpha) \right|, \tag{27}$$

and the secondary cost is

$$W_k(b) = \sum_d |A_{k,d}| \left| \widehat{C}^{(b)}_{k,d} - (1-\alpha) \right|. \tag{28}$$

Thus the table

$$\{E_k(b), W_k(b)\}_{k,b}$$

summarizes the observed STRUCTURE response to candidate calibration resolution. No monotonicity constraint is imposed on this table.

C.5 Fixed-Budget Allocation

For total budget C, define

$$\mathcal{F}_C = \left\{ \mathbf{B} \in \mathcal{B}^K : \sum_{k=1}^K B_k = C \right\}. \tag{29}$$

For $\mathbf{B} \in \mathcal{F}_C$,

$$\mathcal{E}_{\max}(\mathbf{B}) = \max_k E_k(B_k), \tag{30}$$

$$\mathcal{E}_{\mathrm{w}}(\mathbf{B}) = \sum_k W_k(B_k). \tag{31}$$

The selected allocation is

$$\mathbf{B}^\star_C = \operatorname*{lex\,arg\,min}_{\mathbf{B} \in \mathcal{F}_C} \left(\mathcal{E}_{\max}(\mathbf{B}), \mathcal{E}_{\mathrm{w}}(\mathbf{B}), \mathbf{B} \right). \tag{32}$$

The third component is not an additional statistical objective. It provides a deterministic ordering only when the first two quantities are exactly tied.

C.6 Proof of the Matched-Uniform STRUCTURE Property

Proof of Proposition 1. For every supported audit group,

$$\left| \widehat{C}^{(B_k)}_{k,d} - (1-\alpha) \right| \le E_k(B_k) \le \mathcal{E}_{\max}(\mathbf{B}).$$

Hence

$$\widehat{C}^{(B_k)}_{k,d} \ge (1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}).$$

Taking the minimum over supported (k,d) gives

$$\widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B}) \ge (1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}),$$

which is Equation equation 9.

Equivalently,

$$(1-\alpha) - \widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B}) \le \mathcal{E}_{\max}(\mathbf{B}).$$

Because the left-hand side may be negative when worst-group coverage exceeds the nominal target, taking its positive part gives

$$\left[(1-\alpha) - \widehat{\mathrm{WPC}}^{\mathrm{OOF}}(\mathbf{B})\right]_+ \le \mathcal{E}_{\max}(\mathbf{B}),$$

which proves Equation equation 10.

Now suppose $C = qK$ with $q \in \mathcal{B}$. Then

$$\mathbf{B}^{\mathrm{uni}} = (q, \dots, q)$$

belongs to $\mathcal{F}_C$. Since $\mathbf{B}^\star_C$ lexicographically minimizes Equation equation 8 over the entire feasible set, its primary objective cannot exceed that of the feasible Uniform allocation:

$$\mathcal{E}_{\max}(\mathbf{B}^\star_C) \le \mathcal{E}_{\max}(\mathbf{B}^{\mathrm{uni}}).$$

Subtracting both sides from $1-\alpha$ yields

$$(1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}^\star_C) \ge (1-\alpha) - \mathcal{E}_{\max}(\mathbf{B}^{\mathrm{uni}}),$$

which proves Equation equation 12.

If the primary objectives are equal, lexicographic optimality further implies

$$\mathcal{E}_{\mathrm{w}}(\mathbf{B}^\star_C) \le \mathcal{E}_{\mathrm{w}}(\mathbf{B}^{\mathrm{uni}}),$$

which proves Equation equation 13. □

The primary STRUCTURE objective therefore provides a conservative upper bound on worst-group coverage shortfall over the fixed audit family. Matched Uniform belongs to the same feasible allocation space, so BRC obtains a no-worse STRUCTURE-side lower bound under the primary criterion.

C.7 Exact Threshold-First Dynamic Program

A dynamic program that retains only one partial state for each used budget is not generally exact for a max-then-sum objective. BRC therefore solves the primary minimax component first over the finite set of attainable local response thresholds.

Define

$$\mathcal{T} = \{E_k(b) : k = 1, \dots, K,\ b \in \mathcal{B},\ E_k(b) < \infty\}. \tag{33}$$

The optimal primary threshold is

$$\tau^\star = \min\{\tau \in \mathcal{T} : \exists \mathbf{B} \in \mathcal{F}_C \text{ such that } E_k(B_k) \le \tau\ \forall k\}. \tag{34}$$

For a fixed candidate threshold τ, let $D_j(c)$ denote the minimum accumulated secondary cost after assigning resolutions to the first j regimes using exactly c cells. Initialize

$$D_0(0) = 0, \qquad D_0(c) = +\infty \quad (c > 0). \tag{35}$$

The recurrence is

$$D_j(c) = \min_{\substack{b \in \mathcal{B},\ b \le c \\ E_j(b) \le \tau}} [D_{j-1}(c-b) + W_j(b)]. \tag{36}$$

Backpointers retain the corresponding allocation prefix, using the lexicographically smaller prefix whenever secondary costs tie.

Thresholds in $\mathcal{T}$ are examined in increasing order. The first threshold for which

$$D_K(C) < \infty$$

is $\tau^\star$.

Proposition 2 (Exactness of the BRC allocation solver)**.** *For fixed response values* $\{E_k(b), W_k(b)\}_{k,b}$*, finite* $\mathcal{B}$*, and feasible exact budget* C*, the threshold-first dynamic program returns*

$$\operatorname*{lex\,arg\,min}_{\mathbf{B}\in\mathcal{F}_C}\left(\max_k E_k(B_k), \sum_k W_k(B_k), \mathbf{B}\right).$$

Proof. Every feasible allocation has primary objective $\max_k E_k(B_k)$ equal to an element of the finite set $\mathcal{T}$. Therefore the smallest threshold in $\mathcal{T}$ that admits an exact-budget allocation is exactly the globally minimal primary value.

Conditioned on this threshold, the constraints $E_k(B_k) \le \tau^\star$ separate across regimes. The recurrence in Equation equation 36 enumerates every admissible local resolution while retaining the exact cumulative budget, and therefore minimizes the additive secondary objective over all allocations attaining the optimal primary value. Deterministic lexicographic backpointer selection resolves any remaining secondary tie. □

The solver is implementation machinery rather than an independent contribution. Its purpose is to ensure that the allocation returned by the code is the exact solution of Equation equation 32 rather than a heuristic approximation.

C.8 Independent Final Calibration and Support

After $\mathbf{B}_C^\star$ has been selected, the following quantities are frozen before CAL is used:

$$r(x), \quad K, \quad \text{uncertainty reference distributions}, \quad \text{candidate boundaries}, \quad \mathbf{B}_C^\star, \quad \text{final routing}.$$

For calibration scores $s_1, \dots, s_n$, define the standard split-conformal quantile operator

$$Q_{\mathrm{conf}}(s_1, \dots, s_n) = \begin{cases} s_{(j)}, & j = \lceil (n+1)(1-\alpha) \rceil \le n, \\ +\infty, & j > n. \end{cases} \tag{37}$$

The global threshold is

$$\widehat{q}_{\mathrm{global}} = Q_{\mathrm{conf}}\left(\{S_i : i \in \mathcal{D}_{\mathrm{cal}}\}\right). \tag{38}$$

For regime k with CAL support n_k^{cal},

$$\widehat{q}_k = \begin{cases} Q_{\mathrm{conf}}\left(\{S_i : r(X_i) = k\}\right), & n_k^{\mathrm{cal}} \ge 30, \\ \widehat{q}_{\mathrm{global}}, & n_k^{\mathrm{cal}} < 30. \end{cases} \tag{39}$$

For selected final cell h with CAL support n_h, let τ_{leaf} denote the prespecified minimum selected-cell CAL support for the corresponding experimental setting, as specified in Table 4. Then

$$\widehat{q}_h = \begin{cases} Q_{\mathrm{conf}}\left(\{S_i : h(X_i) = h\}\right), & n_h \ge \tau_{\mathrm{leaf}}, \\ \widehat{q}_{r(h)}, & n_h < \tau_{\mathrm{leaf}}. \end{cases} \tag{40}$$

The resulting support hierarchy is

$$\text{selected cell} \longrightarrow \text{parent regime} \longrightarrow \text{global}. \tag{41}$$

The minimum STRUCTURE regime support and minimum selected-cell CAL support are dataset-family-specific and are reported in Table 4. The OOF local threshold support of 80 observations and parent-regime CAL support of 30 observations are common to the primary experiments. These support quantities serve different operational roles and are not terms in the BRC allocation objective.

C.9 Validity of Frozen Final Calibration

Condition on the fitted predictor and all quantities fixed before final calibration, including the regime map, selected allocation, uncertainty reference distributions, and final cell boundaries.

Consider a final cell h calibrated directly using its own n_h CAL observations. Let their conformity scores be

$$S_1, \ldots, S_{n_h},$$

and let S_{n_h+1} denote the score of a future observation that is also assigned to h. Under exchangeability of these $n_h + 1$ scores, the rank of S_{n_h+1} among them is exchangeable. With

$$j_h = \lceil (n_h + 1)(1 - \alpha) \rceil,$$

the standard split-conformal rank argument gives

$$\Pr\{S_{n_h+1} \le \widehat{q}_h \mid h(X_{n+1}) = h\} \ge 1 - \alpha. \tag{42}$$

By the definition of the prediction set,

$$\Pr\left\{Y_{n+1} \in \widehat{\Gamma}_\alpha(X_{n+1}) \mid h(X_{n+1}) = h\right\} \ge 1 - \alpha, \tag{43}$$

which proves Corollary 1.

This is ordinary groupwise split-conformal validity for a partition frozen independently of final CAL. This is the standard groupwise split-conformal validity statement for a partition frozen independently of final CAL.

A separate qualification is required for fallback. Whether a selected child cell has sufficient final CAL support is determined from CAL counts. Consequently, when a prediction uses a parent-regime or global threshold, we do not assert Equation equation 43 conditional on membership in the unsupported child cell or conditional on the event that fallback occurred. The threshold used is the ordinary split-conformal threshold of the corresponding parent or global calibration group; the formal statement is interpreted at that calibration-group level rather than as a new guarantee for the fallback-selected subset.

C.10 Methodological Scope

The BRC decision variable is the regime-level resolution vector $\mathbf{B}$, coupled by the shared constraint

$$\sum_{k=1}^{K} B_k = C.$$

Regime construction supplies the frozen geometry, uncertainty orders candidate refinement within each regime, OOF calibration estimates the resolution response, and the allocation objective distributes the fixed budget. The canonical experiments instantiate these components as specified above; constructor-modularity experiments replace the upstream regime map while leaving the downstream BRC allocation and final calibration procedure unchanged.

D Experimental Protocol

This section specifies the protocol used for all final experiments. Methodological definitions, including regime construction, response estimation, the allocation objective, conformity-score formulas, and final conformal calibration, are given in Appendix C. Here we specify only how the method and comparison procedures are instantiated and evaluated.

D.1 Data Partitioning and Data Roles

For tabular classification and regression, the observations are partitioned into five mutually disjoint subsets,

$$\mathcal{D}_{\text{train}}, \quad \mathcal{D}_{\text{val}}, \quad \mathcal{D}_{\text{structure}}, \quad \mathcal{D}_{\text{cal}}, \quad \mathcal{D}_{\text{test}},$$

using the fixed proportions

$$55\%/10\%/15\%/10\%/10\%, \tag{44}$$

respectively. Classification partitions are stratified by the target label. The same five-way proportions are used for the regression experiments.

For CIFAR-100, the ImageNet-pretrained feature extractor is frozen and the official CIFAR-100 training set is partitioned as

$$60\%\ \mathcal{D}_{\mathrm{train}}, \qquad 10\%\ \mathcal{D}_{\mathrm{val}}, \qquad 15\%\ \mathcal{D}_{\mathrm{structure}}, \qquad 15\%\ \mathcal{D}_{\mathrm{cal}}. \tag{45}$$

A seed-specific linear classification head is fitted using TRAIN, with VAL used for predictor selection or early stopping, and the official CIFAR-100 test set is used as $\mathcal{D}_{\mathrm{test}}$.

For ImageNet-1K, the pretrained predictor is held fixed throughout the conformal experiment. Its available held-out pool is partitioned as

$$30\%\ \mathcal{D}_{\mathrm{structure}}, \qquad 30\%\ \mathcal{D}_{\mathrm{cal}}, \qquad 40\%\ \mathcal{D}_{\mathrm{test}}. \tag{46}$$

Within every dataset and reporting seed, BRC and all matched controls use identical data assignments.

The subsets have nonoverlapping roles. For tabular classification, regression, and CIFAR-100, TRAIN is used to fit the task-specific predictor and VAL is used for ordinary predictor selection or early stopping where required. In CIFAR-100, the pretrained feature extractor itself remains frozen while the seed-specific linear head is fitted. For ImageNet-1K, predictor fitting precedes the conformal experiment, which therefore begins with STRUCTURE.

STRUCTURE is used for all structural decisions required by BRC, including regime selection, uncertainty references, candidate refinement boundaries, OOF response estimation, and allocation. These quantities are frozen before independent calibration. CAL is then used only to estimate the final conformal thresholds, and TEST is reserved exclusively for final evaluation.

Consequently, neither CAL nor TEST labels are used to select the regime map, candidate resolution, support configuration, or budget allocation. No TEST metric is used for model selection or hyperparameter selection.

D.2 Datasets and Predictors

Tabular classification. We evaluate Adult (Kohavi, 1996), Covertype (Blackard & Dean, 1999), and ACSIncome from Folktables (Ding et al., 2021). Each dataset is evaluated with TabNet (Arik & Pfister, 2021) and FT-Transformer (Gorishniy et al., 2021), producing six tabular dataset–backbone settings.

Image classification. We evaluate CIFAR-100 (Krizhevsky, 2009) and ImageNet-1K (Deng et al., 2009) with ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021). For CIFAR-100, the ImageNet-pretrained backbone is frozen and used as a feature extractor, and a seed-specific 100-class linear prediction head is fitted on TRAIN with VAL used for predictor selection. STRUCTURE and CAL are drawn from the remaining portion of the official CIFAR-100 training set, while the official test set is reserved exclusively for TEST. For ImageNet-1K, the pretrained predictor is held fixed throughout, and the available held-out image pool is partitioned into $30/30/40\%$ STRUCTURE/CAL/TEST splits as specified in Equation equation 46. Together with the tabular experiments, the classification benchmark comprises ten dataset–backbone settings.

Regression transfer. We additionally evaluate Bike Sharing (Fanaee-T & Gama, 2014) and California Housing (Pace & Barry, 1997) using a three-member TabNet regression ensemble. The ensemble mean provides the point prediction used with absolute-residual conformity scores. These experiments test transfer of the same allocation formulation beyond classification and are analyzed separately from the primary classification study.

BRC is applied after predictor fitting. For every matched comparison, the fitted predictor and its predictions are therefore held fixed while the conformal calibration procedure is changed.

D.3 Experimental Instantiations

Tabular classification uses LAC conformity scores. Image classification uses deterministic nonrandomized APS under the strict-before-candidate convention (Romano et al., 2020). Regression uses absolute residual conformity scores. Their exact definitions are given in Appendix C.

Unless stated otherwise, the target miscoverage level is

$$\alpha = 0.10,$$

corresponding to nominal coverage $1 - \alpha = 0.90$. The target-coverage sensitivity experiment additionally evaluates

$$\alpha \in \{0.05, 0.10, 0.20\}.$$

D.4 Fixed Structural and Calibration Settings

Table 4 summarizes the fixed implementation settings used in the final evaluation. These quantities are specified independently of TEST performance.

Table 4: **Fixed experimental settings.** Structural quantities are selected using STRUCTURE and frozen before independent CAL calibration.

Setting	Value
Tabular/regression TRAIN / VAL / STRUCTURE / CAL / TEST	$55/10/15/10/10\%$
CIFAR-100 train-pool TRAIN / VAL / STRUCTURE / CAL	$60/10/15/15\%$
CIFAR-100 TEST	official test set
ImageNet-1K STRUCTURE / CAL / TEST	$30/30/40\%$
Reporting seeds	$\{42, 999, 0, 123, 12345\}$
Default target miscoverage	$\alpha = 0.10$
Candidate regime counts	$\{2, 3, 5, 7, 10\}$
KMeans initializations	20
Regime-stability threshold	$\mathrm{ARI} \geq 0.70$
Minimum STRUCTURE regime support	100 (classification); 50 (regression)
OOF folds	5
Minimum OOF local support	80
Candidate resolutions	$\{1, 2, 3, 4, 6\}$
Final parent-regime CAL support	30
Final leaf CAL support	30 (tabular cls.); 50 (CIFAR-100); 20 (ImageNet-1K/regression)
Primary resolution budgets	$2K,\ 3K,\ 4K$
Reliability audit grid	$P \times U_{10}$
Minimum TEST audit-group support	30 (tabular cls.); 40 (CIFAR-100); 20 (ImageNet-1K/regression)

The candidate resolutions and allocation mechanism are defined in Appendix C. Candidate resolutions are evaluated by five-fold OOF calibration on STRUCTURE while the structural geometry is held fixed. The resulting allocation is frozen before final CAL calibration.

D.5 Resolution Budgets and Matched Comparison

We evaluate BRC at the three total resolution budgets

$$C \in \{2K, 3K, 4K\}.$$

These budgets represent different amounts of local calibration resolution and do not change the amount of TEST data.

For $C = qK$, matched Uniform refinement assigns exactly q cells to each of the K regimes. Thus the principal matched comparisons are

$$\text{Uniform-}q2 \leftrightarrow \text{BRC-}2K,$$

$$\text{Uniform-}q3 \leftrightarrow \text{BRC-}3K,$$

and

$$\text{Uniform-}q4 \leftrightarrow \text{BRC-}4K.$$

Within each matched pair, we hold fixed the fitted predictor, conformity score, all dataset-specific data assignments, regime map, uncertainty references, evaluation groups, and total number of calibration cells.

The $2K$ setting is the tightest evaluated nontrivial resolution budget. A budget of K would assign one cell to every regime and therefore leave no allocation freedom. Accordingly, "tight budget" refers to limited calibration *resolution*, not to a small number of CAL observations.

D.6 Evaluation Metrics and Common Audit Geometry

All matched methods are evaluated on a common frozen reliability partition

$$\mathcal{G}_{\text{audit}} = P \times U_{10},$$

obtained by crossing the predictive regime identity with ten within-regime uncertainty strata. The evaluation partition is shared across matched methods rather than defined separately for each calibration procedure.

For classification, we report marginal coverage (MC), worst-group coverage (WPC), average prediction-set size where applicable. Worst-group coverage is evaluated as

$$\text{WPC} = \min_{g \in \mathcal{G}_{\text{audit}}} \widehat{\text{Cov}}_g$$

over the supported audit groups. For the primary TEST evaluation, an audit group is included when it contains at least 30 TEST observations for tabular classification, 40 for CIFAR-100, and 20 for ImageNet-1K and regression. Experiment-specific diagnostics that use a different support rule state that rule explicitly. For regression, we report marginal coverage, WPC, and average prediction-interval length.

WPC is a TEST evaluation metric and is distinct from the STRUCTURE-side calibration-response objective used to select the BRC allocation. MC is interpreted relative to the nominal target coverage rather than as a quantity to maximize. Prediction-set size or interval length is reported jointly with coverage so that changes in local reliability can be interpreted together with prediction efficiency.

D.7 Support and Fallback Protocol

Support requirements are fixed before TEST evaluation. In the primary experiments, the minimum STRUCTURE regime support is 100 for classification and 50 for regression. A local threshold used during OOF response estimation requires at least 80 training-fold observations.

During independent final calibration, a parent-regime threshold requires at least 30 CAL observations. The minimum final-cell CAL support is 30 for main tabular classification, 50 for CIFAR-100, and 20 for ImageNet-1K and regression. Diagnostic experiments that intentionally modify these support rules state the corresponding values in their own subsections.

Thus the fixed fallback hierarchy is

$$\text{cell} \rightarrow \text{regime} \rightarrow \text{global}.$$

Support sensitivity is evaluated separately using STRUCTURE thresholds $\{50, 100, 200\}$ and final-CAL leaf thresholds $\{25, 50, 100\}$. These sensitivity settings are diagnostic only and are not used to retune the primary configuration.

D.8 Baselines and Comparison Rationale

The baseline suite separates matched structural controls from external conformal methods.

Matched and internal controls. Uniform refinement is the primary matched control because it preserves the predictor, score, regime geometry, data partitions, total cell budget, and evaluation groups while replacing only the learned allocation by equal allocation.

GlobalCP (Lei et al., 2018) provides the nonlocalized split-conformal reference and tests whether localized calibration is required. PatternCP retains the coarse regimes but removes adaptive within-regime resolution, testing whether regime-level calibration alone explains the effect. Random uses a legal nonuniform allocation without the BRC response criterion, testing whether arbitrary nonuniformity is sufficient. GlobalU uses a single global uncertainty ordering with the same total calibration-cell budget, removing regime-specific allocation and testing whether global uncertainty structure alone captures the relevant heterogeneity.

External classification baselines. PLCP (Kiyani et al., 2024) is included as a learned-partition method, providing a comparison with approaches that adapt the conditioning structure itself. NCP (Ghosh et al., 2023) represents neighborhood-based localization in learned representation space. Clustered CP (Ding et al., 2023) represents grouped calibration for many-class classification. EnergyAPS (Akhavan Attar et al., 2026) represents adaptive conformity-score construction rather than adaptive allocation of calibration resolution.

External regression baselines. FeatureCP (Teng et al., 2023) provides a feature-adaptive conformal comparison. RCP (Plassier et al., 2025) represents conformity-score rectification. PLCP (Kiyani et al., 2024) provides a learned-partition comparison, and LOCART (Cabezas et al., 2025) provides a tree-based local conformal regression comparison.

External methods can differ from BRC in conditioning structure, conformity score, or native operating point. They are therefore reported as contextual comparisons and are not included in the matched BRC–Uniform win/tie/loss analysis. Where source implementations are adapted to the common leakage-free protocol, the corresponding implementation provenance is reported explicitly; paper-aligned reimplementations are not described as official source reproductions.

D.9 Robustness Protocol

Robustness experiments test prespecified dependencies of the final procedure without using TEST to select a preferred variant.

The score analysis compares APS and LAC on CIFAR-100 ViT. The regime-count analysis forces

$$K \in \{2, 3, 5\}$$

to assess dependence on upstream regime geometry. Target-coverage sensitivity evaluates

$$\alpha \in \{0.05, 0.10, 0.20\}.$$

The uncertainty-component analysis compares the default equal entropy/inverse-margin construction with each component individually.

Calibration-availability sensitivity uses

$$25\%, \quad 50\%, \quad 100\%$$

of final CAL while keeping STRUCTURE and TEST fixed. This experiment changes calibration-sample availability but does not change the definition of the resolution budget.

An additional Adult control constructs regimes from input-space information to assess whether the allocation procedure is intrinsically dependent on one model-internal representation.

These analyses are reported as robustness or boundary tests and are not used retrospectively to redefine the primary BRC configuration.

D.10 Seeds and Reporting Convention

All final experiments, controls, ablations, and robustness analyses use the five fitted seeds

$$\{42, 999, 0, 123, 12345\}.$$

Unless stated otherwise, results are reported as mean $\pm$ sample standard deviation across these five runs. The standard deviation summarizes across-run variability and is not treated as a confidence interval or significance test.

For matched comparisons we define

$$\Delta\mathrm{WPC} = \mathrm{WPC}_{\mathrm{BRC}} - \mathrm{WPC}_{\mathrm{Uniform}}.$$

Positive values indicate numerically higher BRC WPC, negative values indicate numerically higher Uniform WPC, and zero denotes a matched tie in WPC rather than zero coverage.

E Complete Matched Classification Results

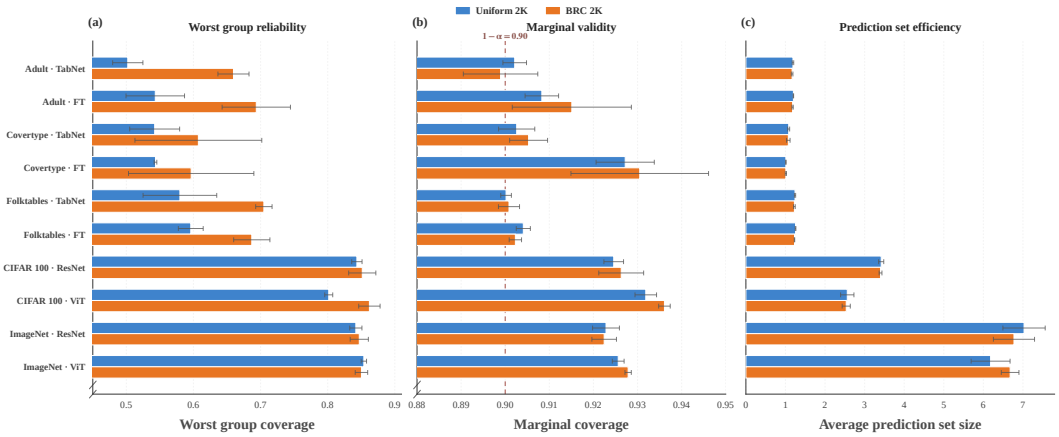

Figure 3: **Visual summary of the primary matched $C = 2K$ comparison.** Bars show five-seed means and whiskers one sample standard deviation. BRC has numerically higher WPC in nine of ten classification settings. MC and prediction-set size are shown jointly to distinguish local reliability changes from changes in marginal coverage or efficiency. Numerical direction does not denote statistical significance.

Tables 5 and 6 report the complete matched BRC–Uniform comparison at $2K$, $3K$, and $4K$. Within each pair, the predictor, conformity score, regime map, audit groups, and total number of calibration cells are fixed; only the allocation of the cell budget changes.

All entries are five-seed mean $\pm$ sample standard deviation over $\{42, 999, 0, 123, 12345\}$. For WPC, a blue bullet marks the matched Uniform reference, a green upward triangle marks numerically higher BRC WPC, a red downward triangle marks numerically lower BRC WPC, and a blue diamond marks a tie. These markers encode numerical direction only.

Table 5: **Complete matched tabular-classification results.** $\Delta\mathrm{WPC} = \mathrm{WPC}_{\mathrm{BRC}} - \mathrm{WPC}_{\mathrm{Uniform}}$. All $2K$, $3K$, and $4K$ operating points are retained.

Setting	Budget	U-MC	BRC-MC	U-WPC	BRC-WPC	ΔWPC	U-Size	BRC-Size
Adult TabNet	$2K$	0.9021 ± 0.0027	0.8989 ± 0.0084	0.5021 ± 0.0226 ●	0.6596 ± 0.0232 ▲	+0.1575	1.1893 ± 0.0192	1.1678 ± 0.0204
	$3K$	0.8979 ± 0.0064	0.9029 ± 0.0031	0.6150 ± 0.0561 ●	0.6672 ± 0.0304 ▲	+0.0522	1.1984 ± 0.0135	1.1974 ± 0.0148
	$4K$	0.9018 ± 0.0052	0.9018 ± 0.0052	0.7190 ± 0.0397 ●	0.7190 ± 0.0397 ◆	0.0000	1.2093 ± 0.0151	1.2093 ± 0.0151
Adult FT-Transformer	$2K$	0.9083 ± 0.0038	0.9151 ± 0.0135	0.5431 ± 0.0436 ●	0.6936 ± 0.0509 ▲	+0.1505	1.2004 ± 0.0055	1.1803 ± 0.0164
	$3K$	0.9194 ± 0.0160	0.9265 ± 0.0042	0.7053 ± 0.0618 ●	0.6964 ± 0.0770 ▼	−0.0089	1.2149 ± 0.0156	1.2216 ± 0.0047
	$4K$	0.9144 ± 0.0126	0.9134 ± 0.0114	0.7249 ± 0.0411 ●	0.7257 ± 0.0411 ▲	+0.0008	1.2104 ± 0.0079	1.2204 ± 0.0049
Covertype TabNet	$2K$	0.9026 ± 0.0041	0.9053 ± 0.0043	0.5422 ± 0.0372 ●	0.6073 ± 0.0945 ▲	+0.0651	1.0814 ± 0.0243	1.0720 ± 0.0441
	$3K$	0.9053 ± 0.0049	0.9103 ± 0.0027	0.6806 ± 0.0096 ●	0.7300 ± 0.0396 ▲	+0.0494	1.0956 ± 0.0196	1.1028 ± 0.0262
	$4K$	0.9119 ± 0.0038	0.9115 ± 0.0044	0.7631 ± 0.0046 ●	0.7657 ± 0.0089 ▲	+0.0026	1.1078 ± 0.0261	1.1084 ± 0.0265
Covertype FT-Transformer	$2K$	0.9272 ± 0.0066	0.9305 ± 0.0156	0.5434 ± 0.0020 ●	0.5966 ± 0.0934 ▲	+0.0532	1.0101 ± 0.0067	1.0104 ± 0.0154
	$3K$	0.9390 ± 0.0096	0.9487 ± 0.0034	0.6903 ± 0.0138 ●	0.7463 ± 0.0364 ▲	+0.0560	1.0259 ± 0.0103	1.0317 ± 0.0027
	$4K$	0.9451 ± 0.0069	0.9263 ± 0.0031	0.7750 ± 0.0139 ●	0.7761 ± 0.0059 ▲	+0.0011	1.0322 ± 0.0081	1.0162 ± 0.0061
Folktables TabNet	$2K$	0.9002 ± 0.0012	0.9009 ± 0.0024	0.5798 ± 0.0549 ●	0.7049 ± 0.0122 ▲	+0.1251	1.2418 ± 0.0155	1.2259 ± 0.0210
	$3K$	0.9006 ± 0.0032	0.9023 ± 0.0016	0.6806 ± 0.0322 ●	0.7236 ± 0.0434 ▲	+0.0430	1.2569 ± 0.0111	1.2533 ± 0.0055
	$4K$	0.9000 ± 0.0033	0.9000 ± 0.0033	0.7474 ± 0.0024 ●	0.7474 ± 0.0024 ◆	0.0000	1.2628 ± 0.0165	1.2628 ± 0.0165
Folktables FT-Transformer	$2K$	0.9041 ± 0.0016	0.9023 ± 0.0014	0.5960 ± 0.0185 ●	0.6869 ± 0.0270 ▲	+0.0909	1.2552 ± 0.0081	1.2281 ± 0.0055
	$3K$	0.9030 ± 0.0025	0.9038 ± 0.0017	0.7053 ± 0.0409 ●	0.7113 ± 0.0252 ▲	+0.0060	1.2641 ± 0.0064	1.2507 ± 0.0080
	$4K$	0.9031 ± 0.0020	0.9031 ± 0.0020	0.7275 ± 0.0242 ●	0.7275 ± 0.0242 ◆	0.0000	1.2757 ± 0.0073	1.2757 ± 0.0073

Table 6: **Complete matched image-classification results.** ΔWPC denotes BRC minus matched Uniform.

Setting	Budget	U-MC	BRC-MC	U-WPC	BRC-WPC	ΔWPC	U-Size	BRC-Size
CIFAR-100 ResNet-50	$2K$	0.9246 ± 0.0022	0.9263 ± 0.0051	0.8435 ± 0.0077•	0.8515 ± 0.0202▲	$+0.0080$	3.4188 ± 0.0646	3.4021 ± 0.0356
	$3K$	0.9301 ± 0.0016	0.9301 ± 0.0016	0.8733 ± 0.0067•	0.8733 ± 0.0067◆	0.0000	3.4566 ± 0.0581	3.4566 ± 0.0581
	$4K$	0.9307 ± 0.0016	0.9310 ± 0.0013	0.8736 ± 0.0103•	0.8717 ± 0.0088▼	-0.0019	3.4804 ± 0.0933	3.4776 ± 0.0483
CIFAR-100 ViT-B/16	$2K$	0.9319 ± 0.0024	0.9361 ± 0.0013	0.8016 ± 0.0060•	0.8621 ± 0.0159▲	$+0.0606$	2.5615 ± 0.1685	2.5352 ± 0.1027
	$3K$	0.9376 ± 0.0028	0.9385 ± 0.0028	0.8689 ± 0.0132•	0.8686 ± 0.0159▼	-0.0002	2.6055 ± 0.1580	2.6183 ± 0.1515
	$4K$	0.9390 ± 0.0016	0.9392 ± 0.0016	0.8667 ± 0.0184•	0.8657 ± 0.0193▼	-0.0010	2.5825 ± 0.1171	2.5820 ± 0.1169
ImageNet-1K ResNet-50	$2K$	0.9229 ± 0.0030	0.9224 ± 0.0028	0.8419 ± 0.0091•	0.8471 ± 0.0135▲	$+0.0051$	7.0279 ± 0.5345	6.7731 ± 0.5198
	$3K$	0.9226 ± 0.0027	0.9230 ± 0.0032	0.8553 ± 0.0010•	0.8582 ± 0.0036▲	$+0.0030$	6.6976 ± 0.5259	6.6617 ± 0.5150
	$4K$	0.9231 ± 0.0031	0.9230 ± 0.0030	0.8582 ± 0.0036•	0.8726 ± 0.0036▲	$+0.0144$	6.6662 ± 0.5094	6.5026 ± 0.4945
ImageNet-1K ViT-B/16	$2K$	0.9256 ± 0.0013	0.9279 ± 0.0007	0.8539 ± 0.0040•	0.8533 ± 0.0093▼	-0.0006	6.1826 ± 0.4920	6.6762 ± 0.2227
	$3K$	0.9302 ± 0.0028	0.9301 ± 0.0029	0.8612 ± 0.0094•	0.8772 ± 0.0092▲	$+0.0161$	6.3148 ± 0.4047	6.2298 ± 0.4446
	$4K$	0.9306 ± 0.0031	0.9306 ± 0.0031	0.8688 ± 0.0103•	0.8688 ± 0.0103◆	0.0000	6.1466 ± 0.3129	6.1466 ± 0.3129

Across the ten classification settings, BRC has numerically higher WPC in nine settings at $2K$, seven at $3K$, and four at $4K$; the $4K$ results additionally contain four exact WPC ties. At the primary $2K$ budget, BRC also produces smaller mean prediction sets in eight of ten settings, so the local-reliability gains are generally not obtained through systematic set inflation. The larger-budget pattern should not be interpreted as a monotone decline in BRC's absolute WPC: rather, the BRC–Uniform difference generally contracts as matched Uniform itself receives finer resolution. Small values near zero therefore represent limited allocation leverage rather than necessarily exact equality.

F Complete Internal Classification Controls

The controls test distinct alternatives to response-based allocation. Uniform is the primary matched comparator because it preserves the regime geometry and total resolution budget while replacing the learned allocation by equal allocation. GlobalCP provides the nonlocalized conformal reference; PatternCP retains the coarse regimes without within-regime adaptive refinement; Random tests whether arbitrary legal nonuniformity is sufficient; and GlobalU tests refinement through a single global uncertainty ordering without regime-specific response allocation. GlobalU was evaluated in the dedicated method-structure ablation, whose prespecified settings were CIFAR-100 ViT-B/16 and Folktables TabNet. It was not run as part of the final CIFAR-100 ResNet-50 or ImageNet-1K control suites; therefore, no GlobalU rows are reported for those settings.

We report each control only for settings completed under the final experimental protocol. An absent row therefore denotes an unavailable comparison rather than a zero-valued result. Directional markers are reserved for the matched BRC–Uniform WPC comparison and indicate numerical direction only.

Table 7: **Complete available internal controls on Adult (tabular classification).** Values are five-seed mean ± sample standard deviation over $\{42, 999, 0, 123, 12345\}$. Uniform and BRC markers refer only to their same-budget matched WPC comparison.

Dataset	Backbone	Method	MC	WPC	Set size
Adult	TabNet	GlobalCP	0.9001 ± 0.0062	0.5964 ± 0.0389	1.1079 ± 0.0224
Adult	TabNet	PatternCP	0.8962 ± 0.0059	0.2315 ± 0.0788	1.1516 ± 0.0206
Adult	TabNet	Uniform-$q2$	0.9021 ± 0.0027	0.5021 ± 0.0226 ●	1.1893 ± 0.0192
Adult	TabNet	BRC-$2K$	0.8989 ± 0.0084	0.6596 ± 0.0232 ▲	1.1678 ± 0.0204
Adult	TabNet	Random-$2K$	0.8988 ± 0.0036	0.4452 ± 0.0463	1.1785 ± 0.0149
Adult	TabNet	Uniform-$q3$	0.8979 ± 0.0064	0.6150 ± 0.0561 ●	1.1984 ± 0.0135
Adult	TabNet	BRC-$3K$	0.9029 ± 0.0031	0.6672 ± 0.0304 ▲	1.1974 ± 0.0148
Adult	TabNet	Random-$3K$	0.9006 ± 0.0023	0.5674 ± 0.0553	1.1959 ± 0.0136
Adult	TabNet	Uniform-$q4$	0.9018 ± 0.0052	0.7190 ± 0.0397 ●	1.2093 ± 0.0151
Adult	TabNet	BRC-$4K$	0.9018 ± 0.0052	0.7190 ± 0.0397 ◆	1.2093 ± 0.0151
Adult	TabNet	Random-$4K$	0.9003 ± 0.0021	0.6222 ± 0.0174	1.2035 ± 0.0129
Adult	FT-Transformer	GlobalCP	0.9043 ± 0.0054	0.6277 ± 0.0258	1.1010 ± 0.0079
Adult	FT-Transformer	PatternCP	0.9014 ± 0.0044	0.1616 ± 0.0832	1.1630 ± 0.0085
Adult	FT-Transformer	Uniform-$q2$	0.9083 ± 0.0038	0.5431 ± 0.0436 ●	1.2004 ± 0.0055
Adult	FT-Transformer	BRC-$2K$	0.9151 ± 0.0135	0.6936 ± 0.0509 ▲	1.1803 ± 0.0164
Adult	FT-Transformer	Random-$2K$	0.9097 ± 0.0078	0.4661 ± 0.0442	1.1928 ± 0.0085
Adult	FT-Transformer	Uniform-$q3$	0.9194 ± 0.0160	0.7053 ± 0.0618 ●	1.2149 ± 0.0156
Adult	FT-Transformer	BRC-$3K$	0.9265 ± 0.0042	0.6964 ± 0.0770 ▼	1.2216 ± 0.0047
Adult	FT-Transformer	Random-$3K$	0.9140 ± 0.0067	0.6461 ± 0.0439	1.2086 ± 0.0041
Adult	FT-Transformer	Uniform-$q4$	0.9144 ± 0.0126	0.7249 ± 0.0411 ●	1.2104 ± 0.0079
Adult	FT-Transformer	BRC-$4K$	0.9134 ± 0.0114	0.7257 ± 0.0411 ▲	1.2204 ± 0.0049
Adult	FT-Transformer	Random-$4K$	0.9164 ± 0.0048	0.6527 ± 0.0370	1.2146 ± 0.0013

Table 8: **Complete available internal controls on Covertype (tabular classification).** Values are five-seed mean ± sample standard deviation over $\{42, 999, 0, 123, 12345\}$. Uniform and BRC markers refer only to their same-budget matched WPC comparison.

Dataset	Backbone	Method	MC	WPC	Set size
Covertype	TabNet	GlobalCP	0.9003 ± 0.0026	0.5064 ± 0.0726	1.0036 ± 0.0164
Covertype	TabNet	PatternCP	0.8993 ± 0.0018	0.1648 ± 0.0162	1.0531 ± 0.0516
Covertype	TabNet	Uniform-$q2$	0.9026 ± 0.0041	0.5422 ± 0.0372 ●	1.0814 ± 0.0243
Covertype	TabNet	BRC-$2K$	0.9053 ± 0.0043	0.6073 ± 0.0945 ▲	1.0720 ± 0.0441
Covertype	TabNet	Random-$2K$	0.9027 ± 0.0032	0.3734 ± 0.0879	1.0770 ± 0.0321
Covertype	TabNet	Uniform-$q3$	0.9053 ± 0.0049	0.6806 ± 0.0096 ●	1.0956 ± 0.0196
Covertype	TabNet	BRC-$3K$	0.9103 ± 0.0027	0.7300 ± 0.0396 ▲	1.1028 ± 0.0262
Covertype	TabNet	Random-$3K$	0.9065 ± 0.0042	0.5193 ± 0.0831	1.0938 ± 0.0222
Covertype	TabNet	Uniform-$q4$	0.9119 ± 0.0038	0.7631 ± 0.0046 ●	1.1078 ± 0.0261
Covertype	TabNet	BRC-$4K$	0.9115 ± 0.0044	0.7657 ± 0.0089 ▲	1.1084 ± 0.0265
Covertype	TabNet	Random-$4K$	0.9094 ± 0.0049	0.6507 ± 0.0206	1.1033 ± 0.0233
Covertype	FT-Transformer	GlobalCP	0.9017 ± 0.0019	0.0000 ± 0.0000	0.9178 ± 0.0007
Covertype	FT-Transformer	PatternCP	0.9014 ± 0.0018	0.0406 ± 0.0186	0.9778 ± 0.0020
Covertype	FT-Transformer	Uniform-$q2$	0.9272 ± 0.0066	0.5434 ± 0.0020 ●	1.0101 ± 0.0067
Covertype	FT-Transformer	BRC-$2K$	0.9305 ± 0.0156	0.5966 ± 0.0934 ▲	1.0104 ± 0.0154
Covertype	FT-Transformer	Random-$2K$	0.9223 ± 0.0048	0.2776 ± 0.1341	1.0040 ± 0.0047
Covertype	FT-Transformer	Uniform-$q3$	0.9390 ± 0.0096	0.6903 ± 0.0138 ●	1.0259 ± 0.0103
Covertype	FT-Transformer	BRC-$3K$	0.9487 ± 0.0034	0.7463 ± 0.0364 ▲	1.0317 ± 0.0027
Covertype	FT-Transformer	Random-$3K$	0.9342 ± 0.0054	0.4878 ± 0.1479	1.0189 ± 0.0053
Covertype	FT-Transformer	Uniform-$q4$	0.9451 ± 0.0069	0.7750 ± 0.0139 ●	1.0322 ± 0.0081
Covertype	FT-Transformer	BRC-$4K$	0.9263 ± 0.0031	0.7761 ± 0.0059 ▲	1.0162 ± 0.0061
Covertype	FT-Transformer	Random-$4K$	0.9398 ± 0.0023	0.6597 ± 0.0241	1.0266 ± 0.0045

Table 9: **Complete available internal controls on Folktables (tabular classification).** Values are five-seed mean ± sample standard deviation over $\{42, 999, 0, 123, 12345\}$. Uniform and BRC markers refer only to their same-budget matched WPC comparison.

Dataset	Backbone	Method	MC	WPC	Set size
Folktables	TabNet	GlobalCP	0.8974 ± 0.0036	0.6918 ± 0.0252	1.1818 ± 0.0051
Folktables	TabNet	PatternCP	0.8975 ± 0.0020	0.5442 ± 0.1397	1.1992 ± 0.0207
Folktables	TabNet	Uniform-$q2$	0.9002 ± 0.0012	0.5798 ± 0.0549●	1.2418 ± 0.0155
Folktables	TabNet	BRC-$2K$	0.9009 ± 0.0024	0.7049 ± 0.0122▲	1.2259 ± 0.0210
Folktables	TabNet	GlobalU-$2K$	0.9013 ± 0.0022	0.6807 ± 0.0730	1.2582 ± 0.0152
Folktables	TabNet	Uniform-$q3$	0.9006 ± 0.0032	0.6806 ± 0.0322●	1.2569 ± 0.0111
Folktables	TabNet	BRC-$3K$	0.9023 ± 0.0016	0.7236 ± 0.0434▲	1.2533 ± 0.0055
Folktables	TabNet	GlobalU-$3K$	0.9017 ± 0.0040	0.7402 ± 0.0719	1.2741 ± 0.0065
Folktables	TabNet	Uniform-$q4$	0.9000 ± 0.0033	0.7474 ± 0.0024●	1.2628 ± 0.0165
Folktables	TabNet	BRC-$4K$	0.9000 ± 0.0033	0.7474 ± 0.0024◆	1.2628 ± 0.0165
Folktables	TabNet	GlobalU-$4K$	0.9023 ± 0.0029	0.8138 ± 0.0491	1.2840 ± 0.0177
Folktables	FT-Transformer	GlobalCP	0.8972 ± 0.0027	0.6717 ± 0.0074	1.1689 ± 0.0029
Folktables	FT-Transformer	PatternCP	0.8986 ± 0.0008	0.3423 ± 0.0969	1.2139 ± 0.0037
Folktables	FT-Transformer	Uniform-$q2$	0.9041 ± 0.0016	0.5960 ± 0.0185●	1.2552 ± 0.0081
Folktables	FT-Transformer	BRC-$2K$	0.9023 ± 0.0014	0.6869 ± 0.0270▲	1.2281 ± 0.0055
Folktables	FT-Transformer	Random-$2K$	0.9016 ± 0.0011	0.5229 ± 0.0423	1.2428 ± 0.0042
Folktables	FT-Transformer	Uniform-$q3$	0.9030 ± 0.0025	0.7053 ± 0.0409●	1.2641 ± 0.0064
Folktables	FT-Transformer	BRC-$3K$	0.9038 ± 0.0017	0.7113 ± 0.0252▲	1.2507 ± 0.0080
Folktables	FT-Transformer	Random-$3K$	0.9032 ± 0.0008	0.6091 ± 0.0324	1.2603 ± 0.0011
Folktables	FT-Transformer	Uniform-$q4$	0.9031 ± 0.0020	0.7275 ± 0.0242●	1.2757 ± 0.0073
Folktables	FT-Transformer	BRC-$4K$	0.9031 ± 0.0020	0.7275 ± 0.0242◆	1.2757 ± 0.0073
Folktables	FT-Transformer	Random-$4K$	0.9051 ± 0.0007	0.6780 ± 0.0113	1.2719 ± 0.0033

Table 10: **Complete available internal controls on CIFAR-100.** Values are five-seed mean ± sample standard deviation. GlobalU is available for the ViT-B/16 setting; no GlobalU row is added to ResNet-50 because no authenticated result is retained for that comparison. Uniform and BRC markers refer only to their same-budget matched WPC comparison.

Dataset	Backbone	Method	MC	WPC	Set size
CIFAR-100	ResNet-50	GlobalCP	0.9015 ± 0.0046	0.8137 ± 0.0123	2.5164 ± 0.0835
CIFAR-100	ResNet-50	PatternCP	0.9201 ± 0.0019	0.7641 ± 0.0170	3.3128 ± 0.0530
CIFAR-100	ResNet-50	Uniform-$q2$	0.9246 ± 0.0022	0.8435 ± 0.0077 ●	3.4188 ± 0.0646
CIFAR-100	ResNet-50	BRC-$2K$	0.9263 ± 0.0051	0.8515 ± 0.0202 ▲	3.4021 ± 0.0356
CIFAR-100	ResNet-50	Random-$2K$	0.9249 ± 0.0018	0.8277 ± 0.0065	3.3961 ± 0.0557
CIFAR-100	ResNet-50	Uniform-$q3$	0.9301 ± 0.0016	0.8733 ± 0.0067 ●	3.4566 ± 0.0581
CIFAR-100	ResNet-50	BRC-$3K$	0.9301 ± 0.0016	0.8733 ± 0.0067 ◆	3.4566 ± 0.0581
CIFAR-100	ResNet-50	Random-$3K$	0.9284 ± 0.0017	0.8640 ± 0.0064	3.4519 ± 0.0702
CIFAR-100	ResNet-50	Uniform-$q4$	0.9307 ± 0.0016	0.8736 ± 0.0103 ●	3.4804 ± 0.0933
CIFAR-100	ResNet-50	BRC-$4K$	0.9310 ± 0.0013	0.8717 ± 0.0088 ▼	3.4776 ± 0.0483
CIFAR-100	ResNet-50	Random-$4K$	0.9288 ± 0.0020	0.8637 ± 0.0075	3.4650 ± 0.0749
CIFAR-100	ViT-B/16	GlobalCP	0.8969 ± 0.0020	0.7616 ± 0.0128	1.7299 ± 0.0088
CIFAR-100	ViT-B/16	PatternCP	0.9307 ± 0.0019	0.8016 ± 0.0060	2.5038 ± 0.1049
CIFAR-100	ViT-B/16	Uniform-$q2$	0.9319 ± 0.0024	0.8016 ± 0.0060 ●	2.5615 ± 0.1685
CIFAR-100	ViT-B/16	BRC-$2K$	0.9361 ± 0.0013	0.8621 ± 0.0159 ▲	2.5352 ± 0.1027
CIFAR-100	ViT-B/16	Random-$2K$	0.9334 ± 0.0022	0.8218 ± 0.0041	2.5569 ± 0.1409
CIFAR-100	ViT-B/16	GlobalU-$2K$	0.9402 ± 0.0025	0.8685 ± 0.0215	2.6198 ± 0.1496
CIFAR-100	ViT-B/16	Uniform-$q3$	0.9376 ± 0.0028	0.8689 ± 0.0132 ●	2.6055 ± 0.1580
CIFAR-100	ViT-B/16	BRC-$3K$	0.9385 ± 0.0028	0.8686 ± 0.0159 ▼	2.6183 ± 0.1515
CIFAR-100	ViT-B/16	Random-$3K$	0.9362 ± 0.0022	0.8451 ± 0.0089	2.5832 ± 0.1454
CIFAR-100	ViT-B/16	GlobalU-$3K$	0.9391 ± 0.0021	0.8658 ± 0.0169	2.6146 ± 0.1524
CIFAR-100	ViT-B/16	Uniform-$q4$	0.9390 ± 0.0016	0.8667 ± 0.0184 ●	2.5825 ± 0.1171
CIFAR-100	ViT-B/16	BRC-$4K$	0.9392 ± 0.0016	0.8657 ± 0.0193 ▼	2.5820 ± 0.1169
CIFAR-100	ViT-B/16	Random-$4K$	0.9370 ± 0.0019	0.8444 ± 0.0111	2.5947 ± 0.1363
CIFAR-100	ViT-B/16	GlobalU-$4K$	0.9391 ± 0.0021	0.8681 ± 0.0122	2.6237 ± 0.1421

Table 11: **Complete available internal controls on ImageNet-1K.** Values are five-seed mean ± sample standard deviation. The authenticated ImageNet internal-control record contains GlobalCP, PatternCP, Uniform, BRC, and Random controls; GlobalU is therefore not added artificially. For ViT-B/16 at $2K$, the corrected final BRC WPC 0.8533 ± 0.0093 is used. Uniform and BRC markers refer only to their same-budget matched WPC comparison.

Dataset	Backbone	Method	MC	WPC	Set size
ImageNet-1K	ResNet-50	GlobalCP	0.8978 ± 0.0017	0.7957 ± 0.0118	3.8197 ± 0.1290
ImageNet-1K	ResNet-50	PatternCP	0.9227 ± 0.0035	0.8213 ± 0.0109	7.6517 ± 0.5083
ImageNet-1K	ResNet-50	Uniform-$q2$	0.9229 ± 0.0030	0.8419 ± 0.0091●	7.0279 ± 0.5345
ImageNet-1K	ResNet-50	BRC-$2K$	0.9224 ± 0.0028	0.8471 ± 0.0135▲	6.7731 ± 0.5198
ImageNet-1K	ResNet-50	Random-$2K$	0.9227 ± 0.0031	0.8395 ± 0.0062	7.1257 ± 0.5213
ImageNet-1K	ResNet-50	Uniform-$q3$	0.9226 ± 0.0027	0.8553 ± 0.0010●	6.6976 ± 0.5259
ImageNet-1K	ResNet-50	BRC-$3K$	0.9230 ± 0.0032	0.8582 ± 0.0036▲	6.6617 ± 0.5150
ImageNet-1K	ResNet-50	Random-$3K$	0.9228 ± 0.0029	0.8518 ± 0.0029	6.7972 ± 0.5220
ImageNet-1K	ResNet-50	Uniform-$q4$	0.9231 ± 0.0031	0.8582 ± 0.0036●	6.6662 ± 0.5094
ImageNet-1K	ResNet-50	BRC-$4K$	0.9230 ± 0.0030	0.8726 ± 0.0036▲	6.5026 ± 0.4945
ImageNet-1K	ResNet-50	Random-$4K$	0.9230 ± 0.0030	0.8576 ± 0.0031	6.7322 ± 0.5123
ImageNet-1K	ViT-B/16	GlobalCP	0.8998 ± 0.0017	0.7407 ± 0.0077	2.7510 ± 0.1164
ImageNet-1K	ViT-B/16	PatternCP	0.9175 ± 0.0011	0.7684 ± 0.0027	6.7904 ± 0.4724
ImageNet-1K	ViT-B/16	Uniform-$q2$	0.9256 ± 0.0013	0.8539 ± 0.0040●	6.1826 ± 0.4920
ImageNet-1K	ViT-B/16	BRC-$2K$	0.9279 ± 0.0007	0.8533 ± 0.0093▼	6.6762 ± 0.2227
ImageNet-1K	ViT-B/16	Random-$2K$	0.9244 ± 0.0017	0.8265 ± 0.0068	6.4293 ± 0.4547
ImageNet-1K	ViT-B/16	Uniform-$q3$	0.9302 ± 0.0028	0.8612 ± 0.0094●	6.3148 ± 0.4047
ImageNet-1K	ViT-B/16	BRC-$3K$	0.9301 ± 0.0029	0.8772 ± 0.0092▲	6.2298 ± 0.4446
ImageNet-1K	ViT-B/16	Random-$3K$	0.9288 ± 0.0024	0.8596 ± 0.0072	6.2147 ± 0.4014
ImageNet-1K	ViT-B/16	Uniform-$q4$	0.9306 ± 0.0031	0.8688 ± 0.0103●	6.1466 ± 0.3129
ImageNet-1K	ViT-B/16	BRC-$4K$	0.9306 ± 0.0031	0.8688 ± 0.0103◆	6.1466 ± 0.3129
ImageNet-1K	ViT-B/16	Random-$4K$	0.9291 ± 0.0023	0.8625 ± 0.0073	6.1579 ± 0.3703

Interpretation. These controls address different empirical questions and should not be read as an interchangeable baseline ranking. Random does not systematically reproduce the BRC behavior, indicating that nonuniform allocation alone is insufficient. GlobalU can exceed BRC in some settings, showing that a single uncertainty ordering may already capture useful reliability structure when regime-specific allocation provides little additional leverage. The controlled claim remains the narrower one tested against matched Uniform: whether the same total calibration resolution is better distributed according to estimated regime response.

G Mechanism Evidence

The mechanism separates two roles. The uncertainty coordinate determines where refinement occurs within a regime, while the estimated OOF calibration response determines how much resolution that regime receives. The analyses below test this response-based interpretation without assuming that generic difficulty or heterogeneity determines the preferred number of cells.

G.1 Regime-Specific Resolution Responses

Table 12: **Regime-specific OOF maximum-gap responses $E_k(B)$.** For candidate resolution B, $E_k(B)$ is the maximum absolute OOF coverage deviation from $1-\alpha$ over the supported common STRUCTURE audit groups within regime k, as defined in Equation equation 4. Lower values indicate smaller STRUCTURE-side maximum calibration deviation. Responses can improve, saturate, or change nonmonotonically with added resolution.

Setting	Regime	$B=1$	$B=2$	$B=3$	$B=4$	$B=6$
CIFAR-100 ViT	0	0.0247 ± 0.0023	0.0280 ± 0.0048	0.0232 ± 0.0061	0.0235 ± 0.0062	0.0256 ± 0.0052
CIFAR-100 ViT	1	0.1050 ± 0.0048	0.1000 ± 0.0000	0.1000 ± 0.0000	0.1000 ± 0.0000	0.1000 ± 0.0000
ImageNet-1K ViT	0	0.1058 ± 0.0517	0.0638 ± 0.0213	0.0638 ± 0.0213	0.0594 ± 0.0289	0.0587 ± 0.0301
ImageNet-1K ViT	1	0.0885 ± 0.0523	0.0613 ± 0.0195	0.0585 ± 0.0205	0.0497 ± 0.0283	0.0445 ± 0.0326

For the downstream diagnostic below, we use the descriptive STRUCTURE-only response quantity

$$D_k^{1\to 3} = E_k(1) - E_k(3).$$

Within each reporting run, the regime with the larger $D_k^{1\to 3}$ is labeled *high-demand* and the other *low-demand* before TEST outcomes are examined. This two-regime label is used only for mechanism analysis. BRC itself does not allocate cells according to $D_k^{1\to 3}$: the actual allocator uses the full response table over $B \in \mathcal{B}$ and solves the shared exact-budget objective.

Regime indices in Table 12 are implementation labels; the table is used to characterize response shape rather than to assign a semantic identity to a cluster across independent fits.

G.2 Downstream Resolution-Demand Check

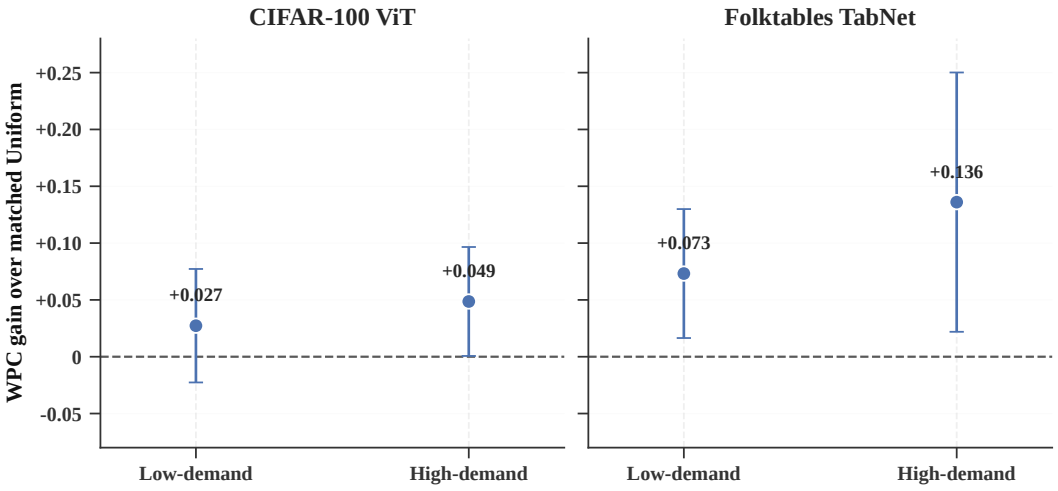

Figure 4: **Downstream effect by estimated resolution demand.** Points show five-seed mean BRC-minus-Uniform regime-specific WPC effects at $2K$, and whiskers denote one sample standard deviation. The comparison is descriptive and does not imply that every high-demand regime must obtain a larger TEST improvement.

The mechanism results are obtained from dedicated diagnostic runs and are distinct from the primary matched benchmark aggregates reported in Table 1. Accordingly, the diagnostic overall ΔWPC values should not be interpreted as replacements for the corresponding primary benchmark effects.

Table 13: **Downstream resolution-demand analysis.** Low-demand and high-demand columns are regime-specific effects and are not additive components of the overall WPC difference.

Setting	Diagnostic ΔWPC	ΔMC	ΔSize	Low-demand ΔWPC	High-demand ΔWPC
CIFAR-100 ViT	$+0.0606 \pm 0.0203$	$+0.0042 \pm 0.0011$	-0.0264 ± 0.0726	$+0.0273 \pm 0.0499$	$+0.0486 \pm 0.0479$
Folktables TabNet	$+0.1251 \pm 0.0560$	$+0.0007 \pm 0.0013$	-0.0159 ± 0.0159	$+0.0732 \pm 0.0567$	$+0.1360 \pm 0.1141$
ImageNet-1K ResNet-50	$+0.0029 \pm 0.0033$	-0.0004 ± 0.0004	-0.2187 ± 0.1934	0.0000 ± 0.0000	$+0.0029 \pm 0.0033$

The high-demand regime has the larger mean downstream WPC effect in the three displayed settings, but the seed variability is substantial. The result is therefore consistent with the response-based mechanism rather than establishing a deterministic ordering of regime behavior.

G.3 Resolution Demand and Generic Heterogeneity

Table 14: **Demand gain and across-regime response heterogeneity.** Increasing average benefit from refinement does not imply increasing heterogeneity in that benefit.

Setting	B	Demand gain	Heterogeneity SD	Heterogeneity range
CIFAR ViT	1	0.0000	0.0000	0.0000
CIFAR ViT	2	0.0008	0.0042	0.0083
CIFAR ViT	3	0.0033	0.0034	0.0069
CIFAR ViT	4	0.0031	0.0037	0.0075
CIFAR ViT	6	0.0020	0.0030	0.0060
ImageNet ViT	1	0.0000	0.0000	0.0000
ImageNet ViT	2	0.0346	0.0266	0.0532
ImageNet ViT	3	0.0360	0.0252	0.0505
ImageNet ViT	4	0.0426	0.0186	0.0372
ImageNet ViT	6	0.0456	0.0157	0.0313

For candidate resolution B, define the regime-specific improvement relative to no refinement as

$$g_k(B) = E_k(1) - E_k(B).$$

The reported *demand gain* is the mean improvement across regimes,

$$G(B) = \frac{1}{K} \sum_{k=1}^{K} g_k(B).$$

The heterogeneity columns summarize the across-regime dispersion of the same response improvements:

$$H_{\mathrm{SD}}(B) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(g_k(B) - G(B)\right)^2}, \qquad H_{\mathrm{range}}(B) = \max_k g_k(B) - \min_k g_k(B).$$

Thus positive $G(B)$ means that resolution B reduces the average STRUCTURE-side maximum-gap response relative to $B = 1$, while H_{SD} and H_{range} measure how unevenly that improvement is distributed across regimes. These quantities are descriptive diagnostics and are not the BRC allocation objective.

For ImageNet ViT, the average cumulative benefit of refinement increases from $G(2) = 0.0346$ to $G(6) = 0.0456$, whereas the across-regime dispersion decreases from $H_{\mathrm{SD}}(2) = 0.0266$ to $H_{\mathrm{SD}}(6) = 0.0157$. Thus, finer resolution can remain increasingly useful on average even as the cumulative response improvements become more homogeneous across regimes. These diagnostics therefore do not support treating generic response heterogeneity as a monotone proxy for useful calibration-resolution demand. Importantly, $G(B)$ measures the overall benefit of refinement rather than the advantage of BRC's nonuniform allocation; the latter is evaluated directly through matched-budget comparisons with uniform refinement.

H Robustness and Failure Modes

Except where stated otherwise, each experiment varies one operating factor while retaining the matched BRC–Uniform comparison and the remaining pipeline. For reporting run s and budget C, define the paired effect

$$\Delta_{s,C} = \mathrm{WPC}_{\mathrm{BRC},s,C} - \mathrm{WPC}_{\mathrm{Uniform},s,C}.$$

Reported means and sample standard deviations are computed from these paired effects over the runs available for the corresponding diagnostic. The sensitivity studies are not used to retune the primary configuration from TEST performance; positive, tied, and negative effects are all retained.

H.1 Conformity-Score Robustness

Table 15: **Score robustness on CIFAR-100 ViT.** Entries are paired BRC-minus-Uniform effects.

Score	Budget	ΔWPC	ΔSize	ΔMC
APS	$2K$	$+0.0606 \pm 0.0203$▲	-0.0264	+0.0042
APS	$3K$	-0.0002 ± 0.0027▼	+0.0128	+0.0009
APS	$4K$	-0.0010 ± 0.0018▼	-0.0004	+0.0002
LAC	$2K$	$+0.1415 \pm 0.0329$▲	-0.1211	-0.0018
LAC	$3K$	$+0.0652 \pm 0.0213$▲	-0.0305	+0.0017
LAC	$4K$	$+0.0727 \pm 0.0227$▲	-0.0418	-0.0016

The $2K$ WPC effect is positive under both conformity scores, while the larger-budget behavior differs between APS and LAC. Under LAC, BRC retains positive WPC effects at $3K$ and $4K$, whereas the corresponding APS effects are approximately neutral. Changes in marginal coverage remain small across all score and budget settings. These results indicate that the allocation mechanism is not specific to a single conformity score, although its effect magnitude depends on the score and budget. The $2K$ WPC effect is positive under both conformity scores, while the larger-budget behavior differs between APS and LAC. Under LAC, BRC retains positive WPC effects at $3K$ and $4K$, whereas the corresponding APS effects are approximately neutral. Changes in marginal coverage remain small across all score and budget settings. These results indicate that the allocation mechanism is not specific to a single conformity score, although its effect magnitude depends on the score and budget.

H.2 Sensitivity to Upstream Regime Count

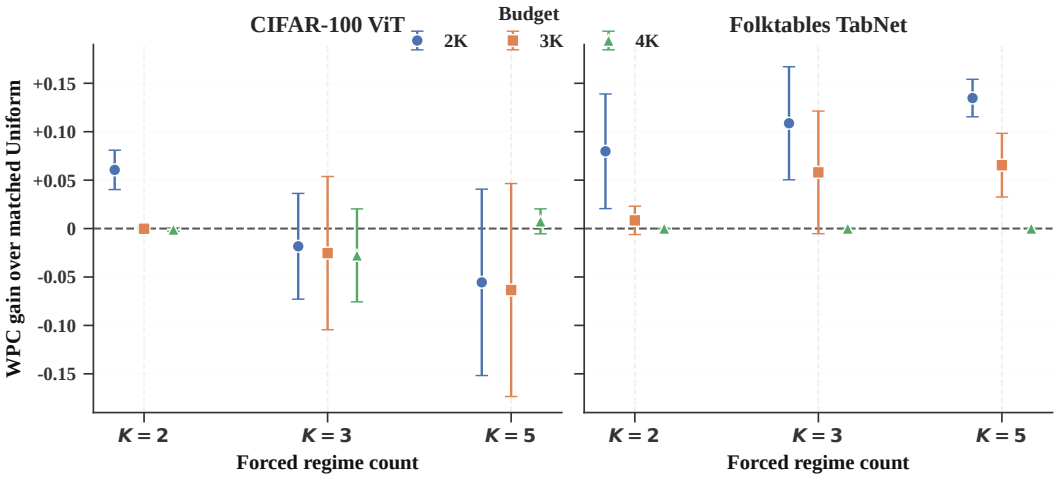

Figure 5: **Sensitivity to upstream regime count.** Points show five-seed mean BRC-minus-Uniform WPC effects under forced $K \in \{2, 3, 5\}$, with one sample standard deviation.

Table 16: **Forced-K sensitivity.** Green triangles denote positive paired effects, red downward triangles negative effects, and blue diamonds exact ties.

Setting	K	$2K$	$3K$	$4K$
CIFAR-100 ViT	2	$+0.0606 \pm 0.0203$▲	-0.0002 ± 0.0027▼	-0.0010 ± 0.0018▼
CIFAR-100 ViT	3	-0.0183 ± 0.0546▼	-0.0253 ± 0.0791▼	-0.0277 ± 0.0480▼
CIFAR-100 ViT	5	-0.0556 ± 0.0962▼	-0.0635 ± 0.1100▼	$+0.0074 \pm 0.0129$▲
Folktables TabNet	2	$+0.0798 \pm 0.0592$▲	$+0.0084 \pm 0.0146$▲	0.0000 ± 0.0000◆
Folktables TabNet	3	$+0.1087 \pm 0.0584$▲	$+0.0580 \pm 0.0633$▲	0.0000 ± 0.0000◆
Folktables TabNet	5	$+0.1348 \pm 0.0193$▲	$+0.0654 \pm 0.0329$▲	0.0000 ± 0.0000◆

Forced K deliberately bypasses the normal STRUCTURE-side stability-and-support selection rule and is used only as a geometry sensitivity intervention. The CIFAR-100 ViT reversals at $K = 3$ and $K = 5$ therefore show that useful allocation depends on the supplied regime geometry rather than implying that these forced geometries should replace the primary constructor. On Folktables, the $2K$ paired effect remains positive for all three tested values of K.

H.3 Regime-Constructor Modularity

This control tests whether the downstream allocation mechanism requires the canonical KMeans constructor. We fix $K = 2$ so that changing the constructor does not simultaneously change the dimension of the allocation problem.

For CIFAR-100 ViT, a diagonal-covariance two-component Gaussian mixture model is fit to the same standardized predictor signature used by the canonical KMeans constructor. For Folktables TabNet, a two-group score-aware neural partition is learned from STRUCTURE features and true STRUCTURE LAC scores; five-fold cross-fitting supplies OOF STRUCTURE regime labels for response estimation, and a final STRUCTURE-fitted model is frozen before CAL/TEST routing. This score-aware constructor is a regime-construction control, not the external PLCP performance baseline reported later.

Within each constructor, Uniform and BRC use the same predictor, frozen regime map, conformity score, CAL/TEST data, total cell budget, support rules, and common audit grid; only the allocation differs. The dedicated constructor-control protocol uses minimum final-CAL-leaf / evaluated-audit support of $50/40$ for CIFAR-100 ViT and $30/30$ for Folktables TabNet. Because changing the constructor also changes the frozen audit-group definition, absolute WPC values should not be ranked across constructors. The diagnostic quantity is the paired BRC-minus-Uniform effect within each constructor.

Table 17: **Non-KMeans regime-constructor modularity.** Entries are mean $\pm$ sample standard deviation over $n = 5$ seeds. $\Delta\mathrm{WPC} = \mathrm{WPC}_{\mathrm{BRC}} - \mathrm{WPC}_{\mathrm{Uniform}}$. Within each row, BRC and Uniform use the same frozen non-KMeans regime map and the same total calibration-resolution budget.

Setting	Constructor	Budget	Uniform WPC	BRC WPC	ΔWPC
CIFAR-100 ViT	Signature GMM	$2K$	0.8501 ± 0.0165	0.8737 ± 0.0033	$+0.0236 \pm 0.0158$
CIFAR-100 ViT	Signature GMM	$3K$	0.8737 ± 0.0033	0.8679 ± 0.0099	-0.0058 ± 0.0116
CIFAR-100 ViT	Signature GMM	$4K$	0.8679 ± 0.0099	0.8692 ± 0.0077	$+0.0013 \pm 0.0023$
Folktables TabNet	Score-aware partition	$2K$	0.6194 ± 0.0348	0.6898 ± 0.0325	$+0.0704 \pm 0.0278$
Folktables TabNet	Score-aware partition	$3K$	0.6898 ± 0.0325	0.7226 ± 0.0137	$+0.0328 \pm 0.0363$
Folktables TabNet	Score-aware partition	$4K$	0.7650 ± 0.0085	0.7650 ± 0.0085	0.0000 ± 0.0000

Both alternative constructors yield positive paired effects at the primary $2K$ budget, showing that BRC can be instantiated with non-KMeans upstream regime maps in these representative settings. The larger-budget effects are smaller or mixed. This supports algorithmic modularity of the downstream allocation step, not invariance to arbitrary regime geometry or a ranking of one constructor against another.

H.4 TARGET-COVERAGE SENSITIVITY

Table 18: **Target-miscoverage sensitivity.** Entries are five-seed paired BRC-minus-Uniform WPC effects.

Setting	α	$2K$	$3K$	$4K$
CIFAR-100 ViT	0.05	0.0000 ± 0.0000 ◆	-0.0033 ± 0.0035 ▼	-0.0022 ± 0.0039 ▼
CIFAR-100 ViT	0.10	$+0.0606 \pm 0.0203$ ▲	-0.0002 ± 0.0027 ▼	-0.0010 ± 0.0018 ▼
CIFAR-100 ViT	0.20	$+0.0027 \pm 0.0095$ ▲	$+0.0104 \pm 0.0097$ ▲	$+0.0020 \pm 0.0035$ ▲
Folktables TabNet	0.05	$+0.0403 \pm 0.0513$ ▲	$+0.0160 \pm 0.0157$ ▲	$+0.0067 \pm 0.0117$ ▲
Folktables TabNet	0.10	$+0.1251 \pm 0.0560$ ▲	$+0.0430 \pm 0.0746$ ▲	0.0000 ± 0.0000 ◆
Folktables TabNet	0.20	$+0.2340 \pm 0.1199$ ▲	$+0.0239 \pm 0.0207$ ▲	$+0.0214 \pm 0.0371$ ▲

The effect varies with both target coverage and resolution budget. This is expected because $E_k(B)$ is defined relative to the target $1 - \alpha$; changing α changes the STRUCTURE-side response landscape and can therefore change the preferred allocation. In particular, the CIFAR-100 ViT $2K$ result changes from an exact tie at $\alpha = 0.05$ to a positive effect at the default $\alpha = 0.10$.

H.5 STRUCTURE-SUPPORT SENSITIVITY

Table 19: **Minimum STRUCTURE-support sensitivity.** The reported runs select $K = 2$ under each tested support threshold. Unlike the forced-K sensitivity experiment, this diagnostic uses the standard regime-selection and local-objective protocol.

Setting	K	Min. support	$2K$	$3K$	$4K$
CIFAR ViT	2	50	$+0.0606 \pm 0.0203$ ▲	-0.0002 ± 0.0027 ▼	-0.0010 ± 0.0018 ▼
CIFAR ViT	2	100	$+0.0606 \pm 0.0203$ ▲	-0.0002 ± 0.0027 ▼	-0.0010 ± 0.0018 ▼
CIFAR ViT	2	200	$+0.0606 \pm 0.0203$ ▲	-0.0002 ± 0.0027 ▼	-0.0010 ± 0.0018 ▼
Folktables TabNet	2	50	$+0.1038 \pm 0.0596$ ▲	0.0000 ± 0.0000 ◆	0.0000 ± 0.0000 ◆
Folktables TabNet	2	100	$+0.1038 \pm 0.0596$ ▲	0.0000 ± 0.0000 ◆	0.0000 ± 0.0000 ◆
Folktables TabNet	2	200	$+0.1038 \pm 0.0596$ ▲	0.0000 ± 0.0000 ◆	0.0000 ± 0.0000 ◆

The identical results over thresholds 50, 100, and 200 indicate that these support changes do not alter the effective selected $K = 2$ geometry or downstream allocation in the reported runs. This is a stability result for these settings, not a claim that the STRUCTURE-support threshold is generally irrelevant.

H.6 FINAL-CAL FALLBACK

Table 20: **Final-CAL fallback diagnostic.** The entry $25/50/100$ denotes three separately evaluated minimum final-CAL leaf-support thresholds. A zero in the fallback columns means that no corresponding fallback event was observed at any of these thresholds; it does not denote missing information.

Setting	Method	Min. CAL leaf	Cell FB	Pred. FB	Mean min. CAL cell
CIFAR ViT	BRC-$2K$	25/50/100	0	0	1553.3333
CIFAR ViT	BRC-$3K$	25/50/100	0	0	1031.0000
CIFAR ViT	BRC-$4K$	25/50/100	0	0	645.0000
Folktables TabNet	BRC-$2K$	25/50/100	0	0	2335.3333
Folktables TabNet	BRC-$3K$	25/50/100	0	0	2277.6667
Folktables TabNet	BRC-$4K$	25/50/100	0	0	1770.6667

No cell-level or prediction-level fallback is observed in these representative sweeps, so fallback does not account for the corresponding matched results.

H.7 Uncertainty Construction

Table 21: **Uncertainty-component sensitivity on CIFAR-100 ViT at $2K$.** The default equal combination is compared with each component alone.

Within-regime coordinate	ΔWPC
Equal $50/50$	$+0.0606 \pm 0.0203$▲
Entropy only	$+0.0156 \pm 0.0138$▲
Inverse margin only	$+0.0700 \pm 0.0208$▲

All three constructions give a positive mean $2K$ effect in this setting. The inverse-margin-only result is numerically larger than the equal combination, so the $50/50$ weighting is not interpreted as an empirically optimal mixture.

H.8 Final-CAL Fraction Sensitivity

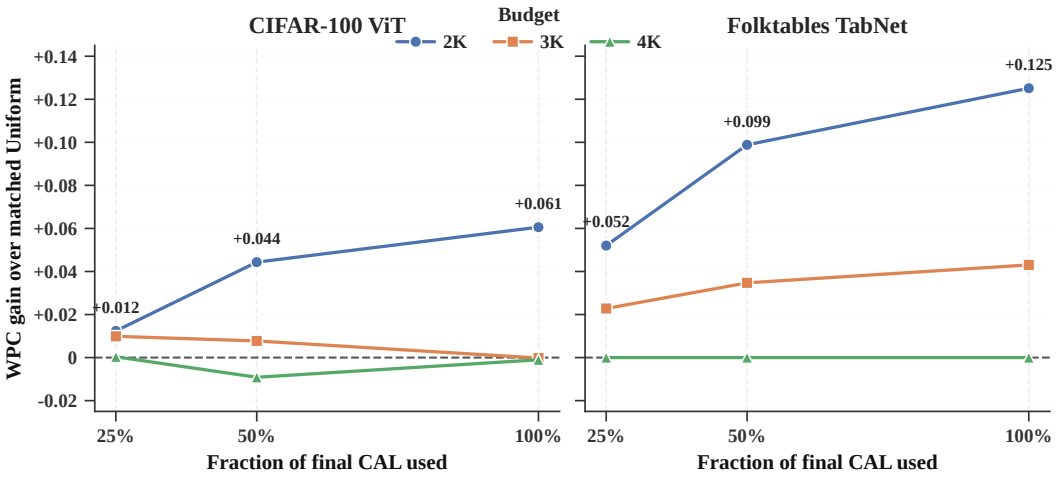

Figure 6: **Sensitivity to final calibration availability.** STRUCTURE and TEST remain fixed while 25%, 50%, or 100% of final CAL is used. Points and lines show five-seed mean paired WPC effects.

Table 22: **Final-CAL fraction sensitivity.** Entries are five-seed paired BRC-minus-Uniform WPC effects.

Setting	CAL used	$2K$	$3K$	$4K$
CIFAR-100 ViT	25%	$+0.0124 \pm 0.0129$▲	$+0.0099 \pm 0.0103$▲	$+0.0004 \pm 0.0006$▲
CIFAR-100 ViT	50%	$+0.0443 \pm 0.0024$▲	$+0.0077 \pm 0.0189$▲	-0.0091 ± 0.0158▼
CIFAR-100 ViT	100%	$+0.0606 \pm 0.0203$▲	-0.0002 ± 0.0027▼	-0.0010 ± 0.0018▼
Folktables Tabnet	25%	$+0.0520 \pm 0.0780$▲	$+0.0228 \pm 0.0395$▲	0.0000 ± 0.0000◆
Folktables Tabnet	50%	$+0.0988 \pm 0.0678$▲	$+0.0347 \pm 0.0601$▲	0.0000 ± 0.0000◆
Folktables Tabnet	100%	$+0.1251 \pm 0.0560$▲	$+0.0430 \pm 0.0746$▲	0.0000 ± 0.0000◆

At $2K$, the mean effect increases rather than disappears as more final-CAL observations are used in both representative settings. Raw final-CAL scarcity is therefore not the sole explanation of the tight-budget effect. At larger budgets, however, the direction can vary with the CAL fraction, so this diagnostic does not imply CAL-fraction invariance.

H.9 Input-Space Regime Construction

This representation-source control uses a minimum STRUCTURE regime support of 50, as specified by its generating implementation; this experiment-specific value does not replace the canonical classification support used in the primary experiments.

Table 23: **Adult input-space regime control.** The same allocation formulation is applied after constructing regimes from input-space information.

Method	MC	WPC	Set size
Input Pattern	0.9017 ± 0.0052	0.5942 ± 0.0422	1.1315 ± 0.0210
Input BRC-$2K$	0.9028 ± 0.0002	0.6679 ± 0.0411	1.1547 ± 0.0151
Input BRC-$3K$	0.9006 ± 0.0089	0.6485 ± 0.0477	1.1683 ± 0.0138
Input BRC-$4K$	0.9012 ± 0.0077	0.7154 ± 0.0467	1.2016 ± 0.0278

This is a representation-source feasibility control: it shows that the downstream BRC procedure can be instantiated after constructing regimes from input-space information rather than only from a model-internal representation. Because this table does not contain matched input-space Uniform-$q2/q3/q4$ rows, it is not used to attribute the displayed differences specifically to budget allocation or to claim equivalence between input-space and representation-space regime constructions.

I Objective Ablation

The objective ablation separates the roles of the maximum-gap criterion, the weighted aggregate criterion, and the full lexicographic objective. Max-only minimizes the largest STRUCTURE audit-group calibration deviation. Weighted-only minimizes the support-weighted aggregate deviation without first enforcing the maximum-gap criterion. The full objective first minimizes the maximum-gap term and then uses the weighted term to distinguish allocations tied under the primary criterion.

Table 24: **Objective-component ablation.** Entries are five-seed mean TEST WPC. The objective itself is selected on STRUCTURE, so TEST WPC is used here for evaluation rather than for choosing among objective variants.

Setting	Objective	$2K$	$3K$	$4K$
CIFAR-100 ViT	Max-only	0.8430	0.8457	0.8457
	Weighted-only	0.8621	0.8686	0.8657
	Full	0.8621	0.8686	0.8657
Folktables TabNet	Max-only	0.6927	0.7032	0.7167
	Weighted-only	0.5201	0.6107	0.7305
	Full	0.6927	0.7339	0.7167

The weighted-only objective performs substantially below the full objective on Folktables at $2K$, while at $4K$ it has the larger TEST WPC. Because these variants are selected according to STRUCTURE criteria rather than TEST WPC, the table is not used to redefine the lexicographic objective according to whichever variant is numerically largest on TEST.

J External Conformal Baselines

External methods can differ from BRC in their conditioning structure, score construction, grouping mechanism, or native operating point. They are therefore reported as external context rather than as part of the controlled BRC–Uniform allocation comparison. Marginal coverage, WPC, and efficiency are reported jointly.

Choice of external classification settings. We evaluate each external classification baseline only on settings where its prediction task and implementation are compatible with the baseline's defining protocol. Accordingly, PLCP (Kiyani et al., 2024) is evaluated on the tabular classification settings, while NCP (Ghosh et al., 2023), Clustered CP (Ding et al., 2023), and EnergyAPS (Akhavan Attar et al., 2026) are evaluated on the image settings for which their corresponding implementations and score constructions are applicable. We do not force every external method onto every dataset, since doing so would require changing native grouping, scoring, or calibration choices and would no longer constitute a faithful baseline.

J.1 Tabular Classification

Partition Learning Conformal Prediction (PLCP) (Kiyani et al., 2024) provides a learned-partition comparison. The reported implementation is a paper-aligned reimplementation rather than an author-run result.

Table 25: **PLCP (Kiyani et al., 2024) on tabular classification.** The reported PLCP implementation is the audited paper-aligned reimplementation used in this study.

Dataset	Backbone	MC	WPC	Set size	Empty rate
Adult	TabNet	0.9007 ± 0.0055	0.5835 ± 0.0391	1.1277 ± 0.0293	0.0160 ± 0.0277
Adult	FT-Transformer	0.8988 ± 0.0039	0.6119 ± 0.0028	1.0927 ± 0.0103	0.0000 ± 0.0000
Folktables	TabNet	0.8972 ± 0.0029	0.6923 ± 0.0307	1.1807 ± 0.0019	0.0000 ± 0.0000
Folktables	FT-Transformer	0.8972 ± 0.0027	0.6722 ± 0.0079	1.1689 ± 0.0029	0.0000 ± 0.0000

J.2 Image Classification

NCP (Ghosh et al., 2023) provides a neighborhood-localization comparison, Clustered CP (Ding et al., 2023) provides a cluster/group calibration comparison, and EnergyAPS (Akhavan Attar et al., 2026) modifies the conformity-score construction. Baselines labeled *source-port* were adapted from the publicly released implementation associated with the original paper, whereas *paper-aligned* baselines were reimplemented from the published method specification; only the changes required for our data splits, seed control, and evaluation interface were introduced, with any score- or protocol-specific modification stated explicitly in the table.

Table 26: **External image-classification baselines.**

Dataset	Backbone	Method	Protocol	MC	WPC	Set size
CIFAR-100	ResNet-50	NCP (Ghosh et al., 2023)	APS source-port	0.9002 ± 0.0127	0.8316 ± 0.0311	2.7703 ± 0.2540
CIFAR-100	ResNet-50	Clustered CP (Ding et al., 2023)	matched nonrandomized APS	0.9017 ± 0.0046	0.8115 ± 0.0080	2.9262 ± 0.1077
CIFAR-100	ViT-B/16	Clustered CP (Ding et al., 2023)	corrected nonrandomized APS	0.9024 ± 0.0052	0.7489 ± 0.0092	2.1106 ± 0.0897
ImageNet-1K	ResNet-50	Clustered CP (Ding et al., 2023)	corrected nonrandomized APS	0.9045 ± 0.0068	0.7792 ± 0.0258	10.5017 ± 1.3270
ImageNet-1K	ViT-B/16	Clustered CP (Ding et al., 2023)	corrected nonrandomized APS	0.9095 ± 0.0024	0.7705 ± 0.0053	7.2715 ± 1.4556
CIFAR-100	ResNet-50	EnergyAPS (Akhavan Attar et al., 2026)	source-port, STRUCTURE-selected	0.9013 ± 0.0041	0.8511 ± 0.0085	3.8345 ± 0.4755
CIFAR-100	ViT-B/16	EnergyAPS (Akhavan Attar et al., 2026)	source-port, STRUCTURE-selected	0.8997 ± 0.0009	0.8624 ± 0.0109	2.8553 ± 0.2301
ImageNet-1K	ResNet-50	EnergyAPS (Akhavan Attar et al., 2026)	source-port, STRUCTURE-selected	0.8972 ± 0.0032	0.8312 ± 0.0136	5.2279 ± 0.1308
ImageNet-1K	ViT-B/16	EnergyAPS (Akhavan Attar et al., 2026)	source-port, STRUCTURE-selected	0.8991 ± 0.0020	0.8290 ± 0.0257	9.9172 ± 1.8964

These rows occupy different marginal-coverage and efficiency operating points. WPC is therefore interpreted jointly with MC and prediction-set size rather than as an isolated external ranking.

Comparison with the primary BRC operating point. We use BRC-$2K$ here because $2K$ is the prespecified primary operating point of the matched study rather than selecting a BRC budget separately for each external method. At $2K$, BRC has numerically higher WPC than PLCP in all

four reported Adult/Folktables comparisons, NCP on CIFAR-100 ResNet-50, and Clustered CP in all four reported image comparisons. BRC also has numerically higher WPC than EnergyAPS in three of four image comparisons; the exception is CIFAR-100 ViT, where BRC-$2K$ and EnergyAPS obtain WPC 0.8621 and 0.8624, respectively. Overall, BRC-$2K$ is numerically higher in 12 of the 13 reported external classification comparisons.

This count is descriptive rather than a matched allocation win rate: the external methods can differ in marginal coverage, prediction-set size, conditioning structure, conformity score, or native operating point. Their WPC values are therefore interpreted jointly with MC and efficiency.

J.3 Regression Baselines

The regression experiments use the same fixed-budget allocation formulation. The matched effects are reported at all three budgets rather than selecting only favorable operating points.

Choice of external regression settings. External regression comparisons are restricted to settings in which the baseline can be implemented without changing its defining regression protocol. We therefore evaluate FeatureCP (Teng et al., 2023), RCP (Plassier et al., 2025), PLCP (Kiyani et al., 2024), and LOCART (Cabezas et al., 2025) on the available compatible regression benchmarks and report their native MC–WPC–interval-length operating points. The purpose is contextual comparison rather than a matched ranking: BRC–Uniform remains the controlled experiment that isolates the effect of allocating a fixed calibration-resolution budget.

Table 28: **Bike Sharing regression results.** No Random-$2K/3K/4K$ rows are included because no corresponding authenticated Bike Random-control results are part of the final result set.

Method	MC	WPC	Interval length
GlobalCP	0.8925 ± 0.0115	0.7196 ± 0.0522	1.8591 ± 0.0829
PatternCP	0.8955 ± 0.0066	0.8058 ± 0.0302	1.7592 ± 0.0270
Uniform-$q2$	0.8961 ± 0.0096	0.7987 ± 0.0246 ●	1.7612 ± 0.0365
BRC-$2K$	0.8950 ± 0.0088	0.7998 ± 0.0219 ▲	1.7648 ± 0.0357
Uniform-$q3$	0.8980 ± 0.0077	0.8164 ± 0.0161 ●	1.7753 ± 0.0310
BRC-$3K$	0.8967 ± 0.0091	0.8108 ± 0.0277 ▼	1.7693 ± 0.0256
Uniform-$q4$	0.8985 ± 0.0087	0.8089 ± 0.0257 ●	1.7947 ± 0.0312
BRC-$4K$	0.8988 ± 0.0081	0.8238 ± 0.0131 ▲	1.7800 ± 0.0389

Table 29: **California Housing regression results.** Random controls are retained because authenticated matched Random-$2K$, Random-$3K$, and Random-$4K$ results are available for this setting.

Method	MC	WPC	Interval length
GlobalCP	0.8908 ± 0.0066	0.8018 ± 0.0228	1.8054 ± 0.0155
PatternCP	0.8934 ± 0.0035	0.8015 ± 0.0286	1.7978 ± 0.0297
Uniform-$q2$	0.8945 ± 0.0065	0.8088 ± 0.0283 ●	1.8185 ± 0.0164
BRC-$2K$	0.8952 ± 0.0054	0.8015 ± 0.0286 ▼	1.8097 ± 0.0164
Random-$2K$	0.8947 ± 0.0053	0.8064 ± 0.0278	1.8136 ± 0.0092
Uniform-$q3$	0.8962 ± 0.0067	0.8088 ± 0.0283 ●	1.8244 ± 0.0105
BRC-$3K$	0.8944 ± 0.0076	0.8069 ± 0.0259 ▼	1.8219 ± 0.0121
Random-$3K$	0.8962 ± 0.0060	0.8106 ± 0.0213	1.8265 ± 0.0058
Uniform-$q4$	0.8978 ± 0.0049	0.8087 ± 0.0172 ●	1.8366 ± 0.0175
BRC-$4K$	0.8983 ± 0.0051	0.8106 ± 0.0203 ▲	1.8372 ± 0.0180
Random-$4K$	0.8967 ± 0.0057	0.8125 ± 0.0171	1.8343 ± 0.0038

The matched BRC-minus-Uniform WPC effects are $+0.0011$, -0.0056, and $+0.0149$ on Bike Sharing and -0.0072, -0.0019, and $+0.0019$ on California Housing at $2K$, $3K$, and $4K$, respectively. The six matched regression comparisons therefore contain three positive and three negative mean effects. The formulation transfers directly to regression, but its empirical effect is mixed and budget-dependent, whereas the classification experiments provide the broader evidence in this study. BRC produces shorter mean intervals in four of the six matched comparisons, so the mixed WPC behavior

is not systematically explained by interval widening. These results support transfer of the allocation formulation to regression, but not a claim of consistent regression improvement.

K RUNTIME

Table 30 separates the method-specific post-hoc cost from the full post-hoc pipeline including regime construction. Predictor fitting and pretrained feature extraction are not included in these timings. Runtime is reported at $3K$ as a representative operating point. This diagnostic is intended to isolate the incremental allocation cost after the shared post-hoc structures exist rather than to characterize runtime scaling with the resolution budget.

Table 30: **Post-hoc runtime in seconds.** Method-specific time measures the corresponding calibration/allocation stage after the required shared structures exist; full post-hoc time includes regime construction.

Setting	Method	Method mean	Method median	Full mean	Full median
CIFAR ViT	GlobalCP	0.2402	0.2201	0.2402	0.2201
CIFAR ViT	PatternCP	0.0186	0.0157	26.7766	26.6782
CIFAR ViT	Uniform-$3K$	0.0202	0.0180	26.7782	26.6810
CIFAR ViT	BRC-$3K$	0.0532	0.0473	26.8112	26.7203
Folktables TabNet	GlobalCP	0.0158	0.0136	0.0158	0.0136
Folktables TabNet	PatternCP	0.0244	0.0213	29.7150	27.7321
Folktables TabNet	Uniform-$3K$	0.0257	0.0220	29.7163	27.7331
Folktables TabNet	BRC-$3K$	0.1047	0.0879	29.7953	27.8164

At $3K$, the BRC-specific stage adds approximately 0.033 s over matched Uniform on CIFAR-100 ViT and 0.079 s on Folktables, whereas the corresponding full post-hoc pipelines require roughly 27–30 s. The allocation overhead is therefore small relative to upstream regime construction in these reported settings.