ENERGY-WEIGHTED CONFORMAL CLASSIFICATION: EFFICIENCY FAILURE MODES AND COST-BOUNDED ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy weighting adapts conformal prediction sets using information beyond class probabilities, yet coverage alone leaves set size uncontrolled. We show that exact joint density can be dominated by label-independent nuisance features, while nearly constant weights can overwhelm narrow temperature-scaled score gaps; both effects survive recalibration. To protect efficiency, we introduce pooled cost projection (PCP), which measures score displacement in labels per input, clips it, and symmetrically recalibrates for each query. Under exchangeability, PCP preserves marginal coverage and bounds expected additions to the same separately calibrated base by a chosen budget B. A half-label budget repairs a nonempty construction from a raw-target lower bound of 47.583 labels to at most $2.5 + \varepsilon$. On an energy-fine-tuned WRN and two ImageNetV2 target architectures, PCP reduces mean set size by 5.36–8.68% relative to adaptive linear shrinkage at the same formal budget. Label-count protection thus provides a principled interface between informative scores and efficient calibrated sets.

1 INTRODUCTION

Under exchangeability, conformal classification turns model scores into prediction sets with marginal coverage (Vovk et al., 2005; Shafer & Vovk, 2008). Probability-based scores such as APS and RAPS adapt these sets to class probabilities (Romano et al., 2020; Angelopoulos et al., 2021), yet a valid set containing most classes offers little resolution. This matters especially when there are many classes: equal coverage can accompany either a short list or a nearly uninformative set. Energy-based scores retain aspects of the logits discarded by softmax normalization (Grathwohl et al., 2020; Liu et al., 2020), and energy-weighted conformal rules use this additional signal (Akhavan Attar et al., 2026). A common shift of all logits leaves softmax probabilities unchanged but changes energy, creating both an opportunity for adaptation and a route for label-independent density variation to affect set size. Can energy weighting improve allocation without adding an uncontrolled number of labels after recalibration?

Why weight control can fail. Recalibration selects a threshold from true-label scores; it does not constrain how many candidate labels lie below it. Changing a score can also move that threshold, so a small numerical perturbation need not imply a small set-size change. We make this gap precise in two settings. Under fixed finite calibration, independent nuisance density leaves the posterior unchanged yet drives raw-energy sets toward an all-label-or-empty limit. Separately, nearly constant weights can, in exact arithmetic and at suitably small scoring temperature, expand APS sets from fewer than two labels to order K at equal population coverage. When many scores lie close together, weight contrast alone cannot bound the number of labels crossing the recalibrated boundary.

Pooled cost projection. These failures suggest measuring a score change by the labels it moves, rather than by its numerical size. PCP maps base and target scores to inclusive counts over all candidate labels on the final-calibration inputs and the query. The common unit is labels per pooled input, even for scores with different scales or temperatures. PCP clips count displacement and uses the base score to resolve ties, retaining moderate target shifts despite a few extreme ones. Pooling the query symmetrically with calibration inputs preserves exchangeability; separate recalibration provides marginal coverage and bounds expected added labels by a chosen budget B relative to the same, separately calibrated base. Adaptive linear shrinkage shares this guarantee, making it a direct comparison for the value of projection.

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

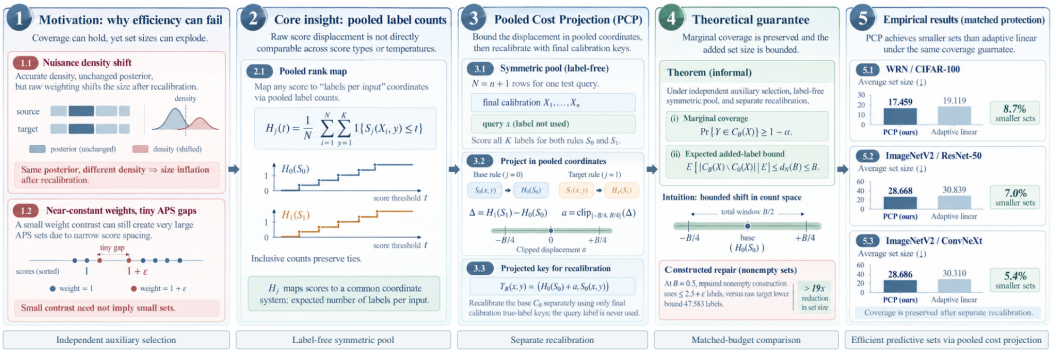

Figure 1: **Protect the label-count cost of adaptation, not the numerical weight.** Density and score-spacing effects motivate a common pooled coordinate; clipping in that coordinate preserves coverage and bounds expected additions after separate recalibration. The construction callout gives theoretical bounds, while the three percentage reductions compare observed mean sizes with adaptive linear at $B = 8$. All empirical comparisons use the same base within each setting.

Results under matched protection. At $B = 0.5$, PCP bounds a nonempty construction's expected size by $2.5+\varepsilon$, against the raw target's 47.583-label lower bound. We also compare PCP with adaptive linear shrinkage under the same formal budget and base rule on real images. At $B = 8$, PCP's mean sets are 5.36–8.68% smaller on an energy-fine-tuned WRN and two ImageNetV2 targets. Paired results and a fixed-strength ablation clarify the contribution of adjustment shape. Together, these comparisons motivate label-count control as an interface between informative scores and efficient calibrated sets.

Our contributions.

- We prove post-recalibration efficiency separations for nuisance density and APS.
- We introduce PCP with coverage and expected-added-label guarantees against a recalibrated base.
- We show constructed repair and matched-budget WRN/ImageNetV2 gains with paired ablations.

2 Related Work

Coverage and score geometry. Exchangeable calibration provides marginal coverage for a broad range of scores (Vovk et al., 2005; Shafer & Vovk, 2008). Their efficiency depends on the information and ordering they retain. Least-ambiguous sets, APS, RAPS, and SAPS exploit probabilities or label ranks (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021; Huang et al., 2024); entropy reweighting, sparse activations, and learned adapters offer complementary designs (Luo & Colombo, 2024; Campos et al., 2025; Liu et al., 2025). Temperature scaling changes this geometry (Guo et al., 2017; Xi et al., 2025; Dabah & Tirer, 2025; Qian et al., 2025). Our APS construction isolates a temperature-nonuniform separation in exact arithmetic, motivating a coordinate that measures the cost of score movement.

Density, energy, and selection. Joint energy models and energy-based detection expose information beyond normalized class probabilities (Grathwohl et al., 2020; Liu et al., 2020), with outlier exposure supplying an additional training signal (Hendrycks et al., 2019). Energy-weighted conformal classification uses this information to allocate set membership (Akhavan Attar et al., 2026). Density can also be driven by background statistics (Nalisnick et al., 2019; Ren et al., 2019; Le Lan & Dinh, 2021); we derive the resulting conformal set-size limit at fixed calibration size. Selection and aggregation can choose among competing scores (Yang & Kuchibhotla, 2025; Liang et al., 2026). PCP complements this choice: independent auxiliary selection fixes a target, and a separate protection mechanism controls its subsequent adaptation.

Protected score adaptation. Expected-size estimation and score-order transformations support efficient conformal design (Dhillon et al., 2024; Hanselle et al., 2025; Yan et al., 2024; Plassier et al., 2025). Symmetric constructions preserve exchangeability (Zargarbashi et al., 2023), while risk-controlling sets address other losses (Bates et al., 2021; Angelopoulos et al., 2024; Fisch et al., 2022). Our distinctive result is a deterministic finite-pool inequality that bounds additions to the *same*, separately recalibrated base. It combines pooled counts, bounded displacement, and inclusive

ordering; adaptive linear shrinkage shares the guarantee and therefore provides a direct test of the value of projection.

A strictly increasing remapping of a single score leaves its recalibrated set unchanged, so useful adaptation must alter the candidate ordering. The challenge is to limit membership changes while allowing the calibration threshold itself to move. PCP handles both movements in pooled label-count coordinates and uses a secondary key to recover the original ordering at zero budget. This makes the protection comparable across score scales and accommodates discrete or tied scores.

3 METHOD

3.1 PROBLEM SETUP

Let $S(x, y, U)$ be a nonconformity score for K labels, with smaller scores preferred and U denoting label-independent randomization. For n final calibration rows, a positive input weight w defines

$$C_w(x, U) = \{y : w(x)S(x, y, U) \le q_w\}, \qquad k = \lceil (n+1)(1-\alpha) \rceil, \tag{1}$$

where q_w is the kth calibration true-label score, or $+\infty$ when $k = n + 1$. We recalibrate every rule separately. Fitting and auxiliary selection precede final calibration, so the coverage question concerns exchangeable final rows conditional on these external choices. The efficiency question concerns $\mathbb{E}|C_w|$.

For logits $f(x) \in \mathbb{R}^K$, write $p_T = \mathrm{softmax}(f/T)$, $m = K^{-1} \sum_j f_j$, and $\ell(p) = \log p - \overline{\log p}\,\mathbf{1}$. Then

$$f = \ell(p_1) + m\mathbf{1}, \qquad e_\tau = \mathrm{LSE}(f/\tau) = h_\tau(p_1) + m/\tau, \quad h_\tau(p) = \mathrm{LSE}(\ell(p)/\tau). \tag{2}$$

The common component m disappears from softmax but enters the energy signal. We study raw weights $w = \phi(e_\tau)$ with $\phi(t) = \log(1 + e^t)$, using the positive-logsumexp convention for e_τ. The score and weight-transform controls are specified in Appendix D.

3.2 WHY A WEIGHT-SCALE CONSTRAINT IS INSUFFICIENT

Nuisance density changes allocation without changing the posterior. Let (A, Y) have positive posterior $\pi_y(A)$ and density $p_A(A)$, and append independent $Z_d \sim N(0, I_d)$. Exact joint logits take the form

$$f_{d,y}(A, Z_d) = \log \pi_y(A) + \log p_A(A) + \log \varphi_d(Z_d). \tag{3}$$

Their posterior is independent of Z_d, whereas e_1 equals the input log density. As nuisance dimension grows, this shared density term can determine the rank of every candidate label on an input.

Theorem 1 (Finite-calibration nuisance limit)**.** *Fix K, n and $1 \le k \le n$. Let $S_y(A, U) > 0$ be independent of Z_d, with almost surely finite log scores and $\log p_A$. Separately recalibrating $\phi(e_1)S_Y$ yields*

$$|\widehat{C}_d| \Rightarrow KV, \quad V \sim \mathrm{Bernoulli}(k/(n+1)), \qquad \mathbb{E}|\widehat{C}_d| \longrightarrow Kk/(n+1). \tag{4}$$

Coverage is exactly $k/(n+1)$ for every positive finite d.

Thus exact density can coexist with an all-label-or-empty limit at fixed calibration size. Setting the top-label score to zero gives a separately calibrated nonempty variant that still has order-K expected size. Appendix A.1 establishes both variants and the relevant scale conditions.

Small contrast must be measured against score gaps. The second mechanism requires no large weight variation. For temperature-scaled APS, probability mass can concentrate on the leading class, compressing the remaining score gaps until a near-constant weight dominates their ordering.

Theorem 2 (Temperature-nonuniform APS separation)**.** *Let a constant positive posterior have leading entries $c > b$ and tail mass $\eta < \alpha = \rho(1 - c)$, with $\rho \in (0, 1)$. For any $R > 1$, choose $T > 0$ such that $a_1 = \pi_1^{1/T} / \sum_j \pi_j^{1/T}$ satisfies $Ra_1 > 1$. Give an independent input group of mass ρ weight R, and all other inputs weight one. Separately population-calibrated APS sets are nonempty, have equal coverage $1 - \alpha$, and satisfy*

$$\mathbb{E}|C_0| = 1 + \frac{1 - \alpha - c}{b} < 2, \qquad \mathbb{E}|C_w| = (1 - \rho)K + \rho. \tag{5}$$

The weights can be realized by raw softplus energy of exact joint logits.

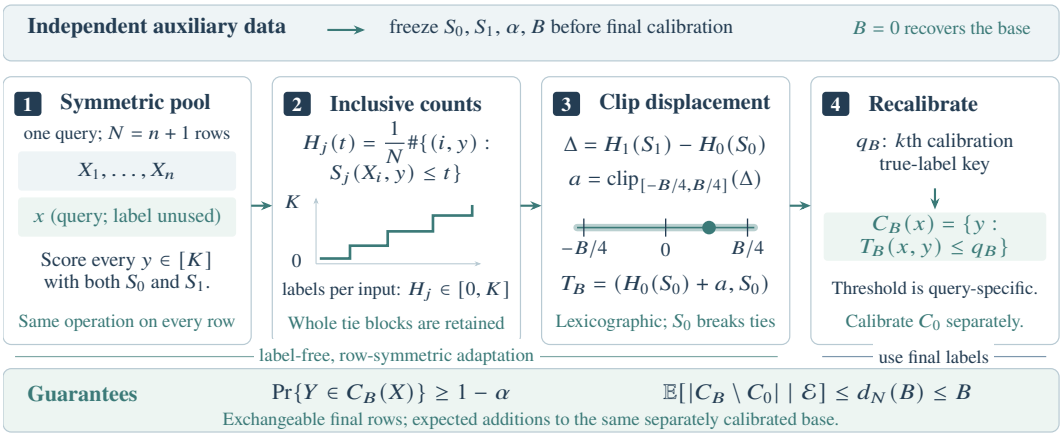

Figure 2: **PCP separates label-free adaptation from labeled recalibration.** Independent data freeze the scoring rules and budget. For each query, all NK candidates enter the inclusive count maps; clipped displacement and the original base score define a lexicographic key. Only then do final calibration true-label keys determine the query-specific threshold. The query label is never used. The same base is calibrated separately at the same quantile rank.

For $c = .9$, $b = .0999$, $\rho = .1$, and $K = 100$, the two expected sizes are 1.9009 and 90.1. The construction allows T to vary as $R \downarrow 1$, isolating a temperature-dependent score-spacing effect. A separate finite-calibration, top-score-zero construction gives a raw-target lower bound of 47.583219 labels and a base bound below $2 + 6.30 \times 10^{-55}$ at $n = 3500$. The proof and a contrasting SAPS bound appear in Appendix A.3.

3.3 Pooled cost projection

A common coordinate for the cost of movement. The preceding results suggest controlling how many labels a score crosses, instead of how far its numerical value moves. Freeze a base S_0, target S_1, and budget $B \ge 0$ using independent auxiliary data. For one query, form a pool of the n final calibration inputs and that query, with $N = n + 1$. Suppressing U for clarity, define

$$H_j(t) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} \mathbf{1}\{S_j(X_i, y) \le t\}, \qquad \Delta(x, y) = H_1(S_1(x, y)) - H_0(S_0(x, y)). \tag{6}$$

Each H_j counts labels per pooled input and ranges from 0 to K. A unit of displacement therefore has the same interpretation across score families and temperatures. Inclusive counting retains whole tie blocks, which is essential when scores are discrete or saturated.

Projection followed by recalibration. PCP moves toward the target by at most $B/4$ in either direction and uses the original base score as a secondary key:

$$T_B(x, y) = (H_0(S_0(x, y)) + \mathrm{clip}(\Delta(x, y), -B/4, B/4),\ S_0(x, y)). \tag{7}$$

Calibrate T_B at rank k using the final true-label keys, compare keys lexicographically, and return all query labels at or below the threshold. The secondary key recovers the base exactly at $B = 0$, including ties. Each query induces its own pool and threshold; the map is constructed without using any labels (Fig. 2).

3.4 Protection through threshold movement

Theorem 3 (Budget for bounded pooled adaptation)**.** *Conditional on independent fitting and selection information $\mathcal{E}$, let the $N = n+1$ final rows, including score randomization, be exchangeable. For any label-free, row-permutation-equivariant adjustment $a_i(y)$ with $|a_i(y)| \le B/4$, separately calibrate the key $T_a = (H_0(S_0) + a, S_0)$. Its set C_a has marginal coverage at least $1 - \alpha$ and satisfies*

$$\mathbb{E}\big[|C_a(X) \setminus C_0(X)| \mid \mathcal{E}\big] \le d_N(B) := \frac{2 \max\{0, \lceil NB/2 \rceil - 1\}}{N} \le B, \tag{8}$$

where C_0 is the same base separately recalibrated on the final true-label scores, with inclusive thresholds and $k = \lceil N(1-\alpha) \rceil$.

Why the bound includes recalibration. A candidate moves by at most $B/4$, and the recalibrated threshold moves by at most another $B/4$. A newly included label must consequently lie within a strict $B/2$ window above the old threshold in base-count coordinates. Across leave-one-out designations, the old threshold takes at most two pooled true-score values. Counting their windows gives $d_N(B)$; symmetry then turns this deterministic finite-pool count into the expected bound. Appendix B proves the result with complete tie blocks.

The expectation averages final calibration and query randomness, conditional on the frozen external choices. In particular, $\mathbb{E}|C_B| \leq \mathbb{E}|C_0| + B$. For the nonempty finite-calibration APS construction, $B = .5$ gives $\mathbb{E}|C_B| \leq 2.5 + \varepsilon$, where $\varepsilon \leq 6.30 \times 10^{-55}$, against the raw lower bound 47.583219. Nonemptiness follows because the projection preserves the shared top-label minimum.

A matched protection control. For $r = B/4$ and $D = \max_{i,y} |\Delta(X_i, y)|$, adaptive linear shrinkage uses $a = \min(1, r/D)\Delta$, with $a = 0$ when $D = 0$. It obeys the same theorem. When $D > r$, one extreme displacement scales down all intermediate movements under linear shrinkage; clipping retains more of each nonextreme movement. This motivates a direct empirical comparison of the two adjustment shapes after recalibration.

Selection and computation. The operator requires no additional backbone training. Independent auxiliary folds select the target settings and budget, keeping $B = 0$ available. Final calibration then applies the frozen policy to each query. Exact direct inference takes $O(NK \log(NK))$ time and $O(NK)$ memory beyond model evaluation; cached calibration outputs and sorted indexes can be reused. Algorithm 1 gives the full procedure, including integer-key and low-temperature ordering rules; Fig. 11 reports the available implementation timings.

4 Experiments

4.1 Experimental setup

We evaluate three questions: whether label-count protection repairs the constructed mechanisms, whether projection retains useful information under the same guarantee as linear shrinkage, and how budget and recalibration shape the resulting gain. All paired methods share their final data roles and receive separate calibration. Real-data means describe overlapping role splits of fixed checkpoints; the corresponding sample SDs appear in Appendix Table 4. The mechanism study instead uses independently generated data and calibration repetitions.

Energy and cross-model adaptation. The native-energy study uses a separately energy-fine-tuned WRN-40-2 on CIFAR-100, a fixed SAPS base, nominal 99% coverage, and five role splits with 2,000 final-calibration and 3,000 test images (Krizhevsky, 2009; Zagoruyko & Komodakis, 2016). The PCP policy selects among four gates at energy strength $\gamma = 1$; the linear policy selects among four strengths. Both select $B \in \{0, .5, 2, 4, 8\}$ using five auxiliary folds and the same empirical preference of at most 0.2 auxiliary added labels per input. This selection preference is separate from the theorem's formal budget.

The transfer study uses ImageNetV2 matched-frequency, a ResNet-18 base, and nominal 95% coverage (Deng et al., 2009; Recht et al., 2019; He et al., 2016). Each of three splits assigns 1,000/2,000/2,000/5,000 images to reference/auxiliary/calibration/test roles. We first evaluate a ResNet-50 target at fixed $B = 8$, then replace it with ConvNeXt-Tiny under a frozen protocol (Liu et al., 2022). The latter independently selects B by two-fold auxiliary size estimation. This second target uses the same image pool and base, isolating target-architecture transfer. Complete protocols and cross-family controls are in Appendices D, E, and G.

4.2 Main results: useful adaptation under matched protection

The key comparison holds the base and formal protection budget fixed while changing the adjustment operator. On the new energy WRN, auxiliary-selected PCP returns 17.459 labels against 19.119 for adaptive linear, a 1.660-label (8.68%) reduction. Both policies choose $B = 8$; their target strengths

Table 1: **Matched-budget real-data comparison ($B = 8$).** Size is mean labels per input (↓); Cov. is mean attained coverage (%). Means use five overlapping WRN or three overlapping ImageNetV2 splits, at nominal coverage 99% and 95%, respectively. Corresponding sample SDs are in Appendix Table 4. On WRN, PCP includes auxiliary gate selection, adaptive linear selects γ, and fixed linear uses $\gamma = 1$. Bold and underline mark the smallest and second-smallest displayed mean sizes.

	WRN / CIFAR-100		**ImageNetV2 / ResNet**		**ImageNetV2 / ConvNeXt**	
Rule	Size ↓	Cov. (%)	Size ↓	Cov. (%)	Size ↓	Cov. (%)
Base	19.327	98.833	31.388	94.827	30.943	94.947
Fixed linear	19.244	98.833	30.840	94.813	30.314	94.933
Adaptive linear	<u>19.119</u>	98.840	<u>30.839</u>	94.813	<u>30.310</u>	94.933
PCP (ours)	**17.459**	98.813	**28.668**	94.753	**28.686**	95.000

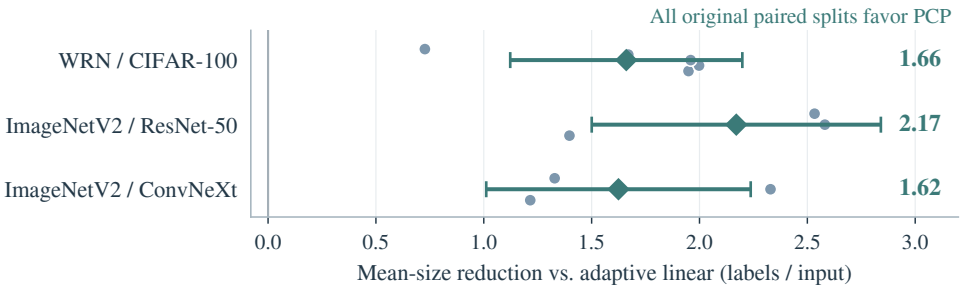

Figure 3: **The matched-budget gain persists across every original paired split.** Points show adaptive-linear minus PCP mean size for each role split; diamonds and bars give the descriptive mean ± sample SD of these differences. Positive values favor PCP. The WRN comparison uses auxiliary-selected policies; the two ImageNetV2 targets use the protocols in Table 1. Image pools overlap across splits, so bars describe partition variation.

are $\gamma = 1$ and 2, respectively. PCP is smaller in all five paired role splits, indicating that the mean gain is distributed across the evaluated partitions rather than concentrated in a single split (Fig. 3).

With the ResNet-50 target, PCP returns 28.668 labels versus 30.839 for adaptive linear, a 2.171-label (7.04%) reduction. Replacing that target by ConvNeXt-Tiny preserves the direction of the comparison: both auxiliary selectors choose $B = 8$ in every split, and PCP returns 28.686 versus 30.310 labels, a 1.624-label (5.36%) reduction. Table 1 reports attained coverage alongside size; the common nominal target and formal budget define these comparisons, while finite-sample coverage proportions may differ.

4.3 Mechanism repair at a half-label budget

To test whether the coordinate change addresses the two theoretical mechanisms, we use $K = 100$, nominal 99% coverage, and the top-score-zero variants. Each of 12 independent repetitions draws 3,500 final-calibration and 12,000 test inputs. This common-minimum construction ensures nonempty sets after recalibration, so set-size repair reflects a genuine reduction in ambiguity.

Raw weighting produces 62.048 and 85.796 labels under 16 and 256 Gaussian nuisance dimensions, whereas PCP remains near the 1.9-label base. Tempering the weight to $w^{.25}$ only partly reduces this inflation. Smooth or hard weight caps handle these Gaussian settings more effectively, yet leave the near-constant APS result unchanged (Fig. 4). There, eight raw-rule repetitions remain near two labels while four rise to about 90 after separate calibration; this two-branch behavior explains why its sample SD exceeds its mean without implying negative set sizes. PCP and fixed linear remain near the base in all three settings. The guarantee explains repair for both bounded operators; the real-data comparison asks which retains more useful target movement.

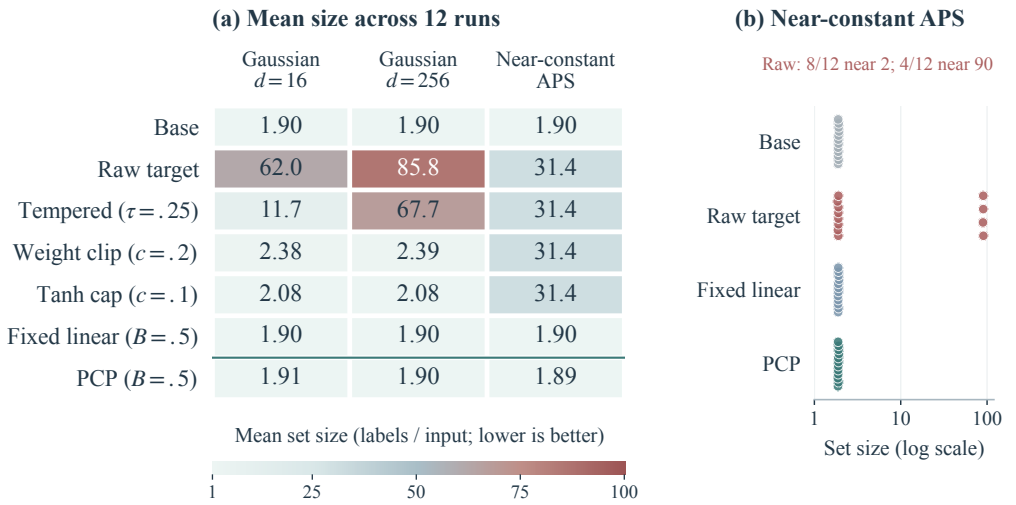

Figure 4: **Constructed repair and calibration branches.** Left: mean size over 12 independent repetitions (labels/input; linear 1–100 color scale). Right: the 12 near-constant APS outcomes (log size axis), with four high-cost raw runs. Gaussian settings use SAPS. All rules are separately calibrated nonempty variants at nominal 99% coverage; fixed linear and PCP share the $B = .5$ expected-addition bound. Each run has 3,500 calibration and 12,000 test inputs. SDs and posterior-expected coverages appear in Table 13.

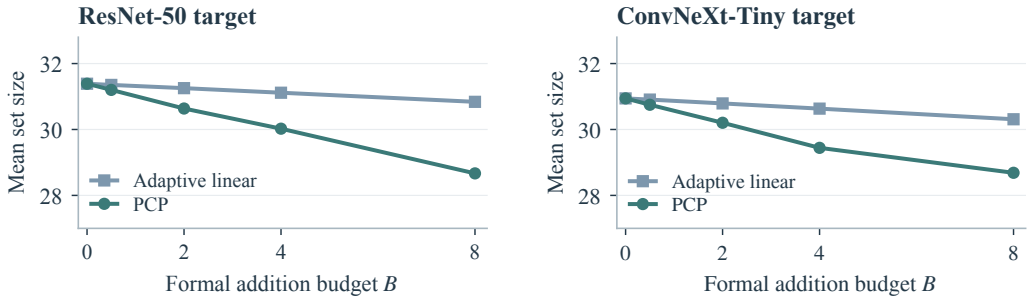

Figure 5: **Increasing the formal budget exposes a larger projection gain.** Clean ImageNetV2 results at nominal 95% coverage, using the five prespecified budgets. Lines give mean set size over three overlapping splits; sample SDs for the main $B = 8$ comparison appear in Appendix Table 4. Both rules recover their base at $B = 0$.

4.4 Budget ablation: how much target movement survives?

A small budget is deliberately conservative. On the clean ResNet transfer, the direct teacher offers a 21.828-label mean reduction, but PCP retains only 0.187 labels of that reduction at $B = .5$. At $B = 8$, the net reduction reaches 2.719 labels (Fig. 5). This behavior follows from the same threshold movement that underlies the protection theorem. For an original base member with score $s \le q_0$, removal requires

$$H_0(q_0) - H_0(s) < B/2. \tag{9}$$

Thus only labels near the calibrated base boundary can respond to a tightly bounded adjustment. In the evaluated ResNet pool, the window excludes 99.02% of teacher-requested deletions at $B = .5$; at $B = 8$, 15.76% are eligible. Figure 6 visualizes this constraint. Eligibility is a necessary condition evaluated using the final base threshold; actual removals also depend on the target direction and the recalibrated key.

The ConvNeXt transfer exhibits the same progression across the frozen grid. PCP's mean size falls from 30.943 at $B = 0$ to 28.686 at $B = 8$, while adaptive linear reaches 30.310. The agreement

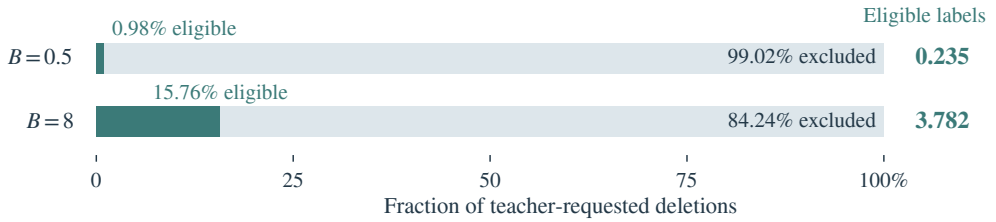

Figure 6: **The removal window explains conservative adaptation at small B.** Clean ResNet-50 transfer at nominal 95% coverage. Bars split teacher-requested deletions into eligible and excluded fractions under Eq. (9); labels report the mean eligible requests per input. Values summarize the three original role splits. Window membership permits a deletion but does not require it.

Table 2: **The projection gain persists when target direction is selected independently.** ResNet-50 transfer at $B = 8$ and nominal 95% coverage. Entries are mean $\pm$ sample SD across three overlapping splits; Size is labels/input (↓), Cov. is attained coverage (%). "Selected" uses the same auxiliary direction library for both operators. Reference projection matches the adjustment amplitude but has a different guarantee scope.

	Clean target		**20% response faults**	
Rule / direction	Size ↓	Cov. (%)	Size ↓	Cov. (%)
Linear / fixed	30.839 *±0.744*	94.813 *±0.336*	30.898 *±0.967*	94.840 *±0.280*
Linear / selected	30.839 *±0.744*	94.813 *±0.336*	30.898 *±0.967*	94.840 *±0.280*
PCP / fixed	28.668 *±1.355*	94.753 *±0.269*	29.292 *±1.315*	94.813 *±0.179*
PCP / selected	**28.611** *±1.408*	94.753 *±0.269*	**28.955** *±0.995*	94.780 *±0.216*
Reference projection	<u>28.648</u> *±1.383*	94.753 *±0.269*	<u>29.201</u> *±1.391*	94.800 *±0.173*

across targets supports the role of the budget as an interpretable control on adaptation, with its operating point selected independently of final calibration.

4.5 Design ablations: adjustment shape, direction, and realized cost

Shape matters beyond satisfying the bound. Fixed and adaptive linear shrinkage share PCP's $B = 8$ expected-addition bound. PCP returns smaller mean sets than both linear rules in the three settings of Table 1. The separately selected WRN policies also differ in target strength and gate choice, so that table alone does not isolate adjustment shape. At fixed $\gamma = 1$ and $B = 8$, ungated PCP returns 17.205 labels against 19.244 for fixed linear on WRN. This comparison holds the target strength and formal budget fixed; the gate attribution study appears in Appendix E.

Direction and construction play distinct roles. Giving both operators the same auxiliary choice among the base, teacher, and four probability-mixture directions yields 28.611 labels for PCP and 30.839 for linear on clean ResNet transfer (Table 2). A matched-amplitude independent-reference projection returns 28.648. These comparisons distinguish the empirical benefit of the adjustment shape from the role of symmetric pooling: the latter supplies the exact same-base certificate, while useful target movement can also be expressed using a reference map.

Track additions and removals separately. A common formal budget defines a shared guarantee, whereas realized additions describe how each method uses it. For ConvNeXt at $B = 8$, PCP removes 2.4365 and adds 0.1794 labels per input relative to the calibrated base; linear removes 0.6328 and adds 0.0001. Their net reductions are therefore about 2.257 and 0.633 labels. Figure 7 makes this accounting explicit across budgets. The resulting size advantage is an improvement at matched formal protection, with observed addition cost reported as a separate axis.

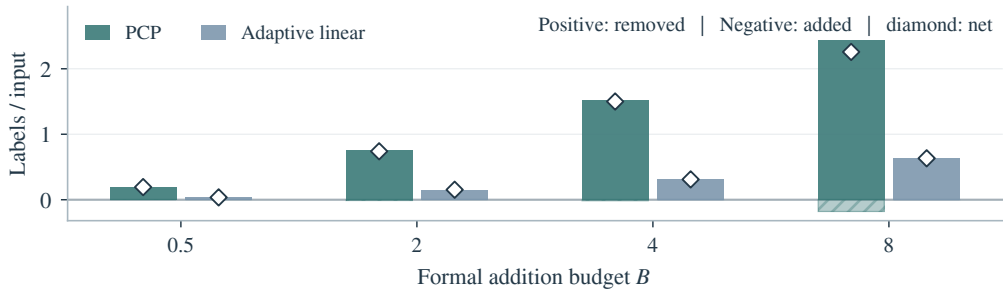

Figure 7: **PCP achieves a larger net reduction with few added labels.** ConvNeXt transfer: positive bars show removed labels, negative bars show added labels, and diamonds mark net size reduction relative to the same separately calibrated base. Values are means over three overlapping splits; all positive budgets in the frozen grid are shown. The equal formal budget does not force equal observed additions.

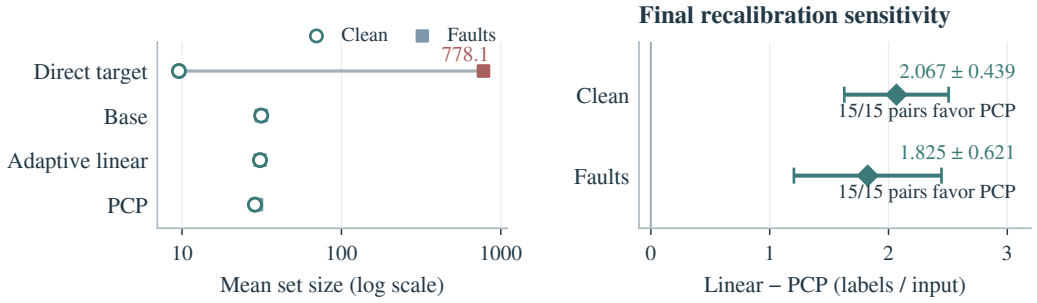

Figure 8: **Target faults and final recalibration.** Left: clean and 20% response-fault mean sizes on ResNet transfer (log scale). Right: linear-minus-PCP size differences across 15 final-pool repartitions per condition at $B = 8$; bars show sample SD, and all pairs favor PCP. Repartitions share models, images, and frozen auxiliary choices.

4.6 Target faults and recalibration sensitivity

Protection when target responses are unreliable. With probability 0.2, the target returns a reference-donor response under the same fault law in auxiliary, calibration, and test roles. On ResNet transfer, the separately calibrated direct target expands from 9.559 to 778.091 labels. At $B = 8$, fixed-direction PCP changes from 28.668 to 29.292 labels, versus 30.898 for linear (Fig. 8); selecting direction on auxiliary data yields 28.955 (Table 2). Appendix F reports condition-specific coverages and unprotected controls.

Sensitivity to the final calibration partition. We repartition each original ResNet split's final 7,000-image pool five times, keeping models, reference data, auxiliary choices, and $B = 8$ fixed. All 15 pairs per condition favor PCP: linear-minus-PCP differences are 2.067 ± 0.439 labels on clean responses and 1.825 ± 0.621 under faults (mean $\pm$ sample SD). Complete comparisons appear in Appendix F.

5 Conclusion

Exact density and nearly constant weights need not yield efficient energy-weighted conformal sets after recalibration. PCP expresses adaptation in pooled label-count coordinates and, under exchangeability, preserves marginal coverage while bounding expected additions to the same separately calibrated base. It repairs large sets in constructed examples and improves mean size over a matched-budget adaptive linear rule on the three evaluated real-data tasks. The resulting tradeoff is how much useful target movement survives a specified label-count allowance.

Reproducibility Statement

We provide implementation details, experimental protocols, evaluation settings, and additional results in the appendices and supplementary materials. Code and supporting materials will be released to facilitate reproduction of the reported results.

AI Use Statement

Generative AI tools were used to assist with literature search, coding, and language polishing. The authors reviewed and verified the resulting content and take full responsibility for the final manuscript.

References

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *ICLR*, 2026.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. arXiv preprint arXiv:2602.20151, 2026.

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

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Margarida M Campos, João Cálem, Sophia Sklaviadis, Mario A. T. Figueiredo, and Andre Martins. Sparse activations as conformal predictors. In *AISTATS*, 2025.

Lahav Dabah and Tom Tirer. On temperature scaling and conformal prediction of deep classifiers. In *ICML*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Guneet S. Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *AISTATS*, 2024.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *ICML*, 2022.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Jonas Hanselle, Alireza Javanmardi, Tobias Florin Oberkofler, Yusuf Sale, and Eyke Hüllermeier. Conformal prediction without nonconformity scores. In *UAI*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *ICML*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021. doi: 10.3390/e23121690.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, pp. 1–13, 2026. doi: 10.1080/01621459.2026.2663588. Advance online publication.

Kangdao Liu, Hao Zeng, Jianguo Huang, Huiping Zhuang, Chi-Man Vong, and Hongxin Wei. C-Adapter: Adapting deep classifiers for efficient conformal prediction sets. In *ECAI*, 2025.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

Ziquan Liu, Yufei Cui, Yan Yan, Yi Xu, Xiangyang Ji, Xue Liu, and Antoni B. Chan. The pitfalls and promise of conformal inference under adversarial attacks. In *ICML*, 2024.

Rui Luo and Nicolo Colombo. Entropy reweighted conformal classification. In *COPA*, 2024.

Thomas Massena, Léo Andéol, Thibaut Boissin, Franck Mamalet, Corentin Friedrich, Mathieu Serrurier, and Sébastien Gerchinovitz. Efficient robust conformal prediction via Lipschitz-bounded networks. In *ICML*, 2025.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

Stephen Obadinma, Xiaodan Zhu, and Hongyu Guo. Calibration attacks: A comprehensive study of adversarial attacks on model confidence. *Transactions on Machine Learning Research*, 2024.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *ICML*, 2025.

Christopher Qian, Feng Liang, and Jason Adams. Extending temperature scaling with homogenizing maps. *Journal of Machine Learning Research*, 26(161):1–46, 2025.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *NeurIPS*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Yan Scholten and Stephan Günnemann. Provably reliable conformal prediction sets in the presence of data poisoning. In *ICLR*, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 1 edition, 2005. doi: 10.1007/b106715.

Huajun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. Does confidence calibration improve conformal prediction? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *ICLR*, 2024.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025. doi: 10.1080/01621459.2024.2344700.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *ICML*, 2023.

Soroush H. Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. In *ICML*, 2024.

Guide to the supplementary material. Appendices A–C give the proofs and implementation. Appendix D specifies shared protocols; Appendices E–G report the real-data comparisons. Appendix H gives a compact score-family check, Appendix I reports mechanism and timing controls, and Appendix J states the scope of the comparisons.

A Proofs of the two efficiency separations

The two constructions hold the label posterior fixed while changing the information that governs score ordering. We first analyze nuisance density at finite calibration size, then study near-constant weights under temperature-dependent APS spacing.

A.1 Gaussian nuisance can determine the entire coverage allocation

Let (A, Y) have a fixed distribution and append $Z_d \sim N(0, I_d)$, independently of (A, Y). Suppose strictly positive posterior probabilities $\pi_j(A)$ and a semantic input density $p_A(A)$ admit finite logits $g_j(A) = \log p_A(A) + \log \pi_j(A)$ almost surely. The normalized exact joint model

$$f_{d,j}(A, Z_d) = g_j(A) + \log \varphi_d(Z_d), \qquad \varphi_d(z) = (2\pi)^{-d/2} e^{-\|z\|^2/2}, \tag{10}$$

has posterior $\pi(A)$ for every d and negative energy $e_d = \log p_A(A) + \log \varphi_d(Z_d)$ at unit temperature. Thus adding these nuisance coordinates changes neither the posterior nor the classification problem. Fix K and any candidate scores $S_j(A, U) > 0$ almost surely, independent of the nuisance, whose distributions do not change with d; informative A is allowed. Use $w_d = \phi(e_d)$ with $\phi(t) = \log(1 + e^t)$.

Proof of Theorem 1. Put $b_d = -d\{\log(2\pi) + 1\}/2$ and $\sigma_d = \sqrt{d/2}$. The normalized nuisance $T_d = (d - \|Z_d\|^2)/\sqrt{2d}$ converges to a standard normal. Since e_d/d tends to a negative constant, $\log \phi(e_d) - e_d \to 0$ in probability. Also $\log p_A(A)$ and every $\log S_j(A, U)$ are $O_p(1)$. Therefore, jointly for all candidate and true-label scores,

$$\{\log(w_d S_j) - b_d\}/\sigma_d = T_d + o_p(1).$$

For the n calibration observations and the test input, the limits are $n + 1$ independent normals. All test candidates share the same test normal. Normal ties have probability zero, so the limiting set contains all labels precisely when the test normal is below the kth calibration normal. Its uniform rank gives probability $k/(n + 1)$. Cardinality is bounded by K, giving expectation convergence. The population version follows from convergence of quantiles to $\Phi^{-1}(v)$. At finite d, conditioning on (A, Y, U) leaves a strictly monotone transformation of a continuous chi-square variable; true scores are therefore continuous and conformal coverage is exact.

A nonempty version. Set the score of the first-ranked label $j_*(A)$ to zero and retain positive scores on the other labels. Let its accuracy be $c = \Pr\{Y = j_*(A)\} < v$ and $\kappa = (v - c)/(1 - c)$. Conditioning positive calibration scores on $Y \neq j_*(A)$ leaves the nuisance independent. Applying the preceding proof to those scores gives

$$\mathbb{E}|C_d| \longrightarrow 1 + (K - 1)\kappa, \qquad \Pr(Y \in C_d) = v, \tag{11}$$

while every set contains its first-ranked label. For finite calibration the size limit is

$$1 + (K - 1)\mathbb{E}\frac{(k - M)_+}{n - M + 1}, \qquad M \sim \mathrm{Bin}(n, c). \tag{12}$$

The forced zeros create ties: finite-sample coverage here equals $\mathbb{E}\max\{k, J\}/(n + 1)$ for $J \sim \mathrm{Bin}(n + 1, c)$, which is at least $k/(n + 1)$. It should not be reported as an exact $k/(n + 1)$ guarantee.

Reference centering does not eliminate the limit. For the nonempty rule, estimate the energy mean from $m \geq 1$ independent reference inputs and use $\phi(e_d - \overline{e}_{d,R} + \log K)$. Here calibration is a population quantile conditional on reference data. If $\kappa > 1/2$, then

$$\lim_{d \to \infty} \mathbb{E}|C_d| \geq 1 + (K - 1)\mathbb{E}\min\{\Phi(V), \kappa\} \geq 1 + (K - 1)\left\{\frac{1}{2} - \frac{(1 - \kappa)^2}{2}\right\}, \quad V \sim N(0, 1/m). \tag{13}$$

For the population reference mean, assuming integrability of $\log p_A(A)$, the stronger bound is $(K+1)/2$. To verify the finite-reference claim, the scaled reference weight tends conditionally to $(N-V)_+$. Write $S_+ = S_Y \mid \{Y \neq j_*(A)\}$. If $\kappa > \Phi(V)$, its positive-scale calibration threshold $q_V > 0$ solves $\mathbb{E}\Phi(V + q_V/S_+) = \kappa$; each non-top candidate is included with probability $\mathbb{E}\Phi(V + q_V/S_j) \geq \Phi(V)$. If $\kappa < \Phi(V)$, the threshold is exponentially small and the same log-score argument as above gives inclusion probability κ for each non-top candidate. The equality case has reference probability zero. Finally $\mathbb{E}\Phi(V) = 1/2$ and $\Pr\{\Phi(V) > u\} \leq 1 - u$ for $u > 1/2$, proving the second bound by integrating over $u \in [\kappa, 1]$.

Temperature conventions lead to different controls. For a constant posterior π and constant semantic density, the reference arguments are exactly

$$\begin{aligned} a + \{\log \varphi_d(Z_d) - \overline{\log \varphi_d}_R\}/\tau \qquad & \text{(code convention)}, \\ a + \log \varphi_d(Z_d) - \overline{\log \varphi_d}_R \qquad & \text{(paper convention)}, \quad a = \log K. \end{aligned} \tag{14}$$

Thus paper-convention energy temperature has no effect at all in this example. In the code convention, the nuisance scale is $\gamma_d = \sqrt{d/2}/\tau_d$: divergent γ_d gives the reference limit above, finite nonzero γ_d retains random nuisance weighting, and $\gamma_d \to 0$ returns constant weighting. For bounded informative logits $|g_j(A)| \leq M$, the semantic energy $H_\tau = \tau\,\mathrm{LSE}(g/\tau)$ differs from $\tau \log K$ by at most M. Consequently, temperature scaling that suppresses the Gaussian nuisance also suppresses this bounded semantic variation. This last statement is a scale comparison, not a new optimality theorem; it does not order arbitrary finite-temperature predictors by efficiency.

Illustration and limitations. With $K = 100$, $\alpha = .01$ and constant posterior leading entries $(.9, .0999)$, leave 10^{-4} positive mass in a strictly decreasing tail. Unweighted SAPS has population size $1 + .09/.0999 = 1.9009009$. The nonempty raw limit is 90.1; the population-reference lower bound is 50.5, and the bound for every finite independent $m \geq 1$ is 50.005. The SAPS penalty may be the empirical default .2. Background-density confounding and the limitations of perfect density models are established (Ren et al., 2019; Le Lan & Dinh, 2021). The results above quantify their effect on conformal coverage allocation for the specified Gaussian nuisance model and softplus weighting rule. It is a claim about this fixed weighting rule. A selected procedure with an unweighted fallback can avoid it; its selection probability and final performance require separate analysis and evidence.

A.2 Coordinate scale and finite-dimensional conditions

For $Z = \sigma G$, $G \sim N(0, I_d)$, put $q(A) = \log p_A(A)$, $a_\sigma = \frac{1}{2}\log(2\pi e\sigma^2)$ and $s_d = \sqrt{d/2}$. The exact input log density is

$$e_d = q(A) - a_\sigma d + s_d N_d, \qquad N_d = (d - \|G\|^2)/\sqrt{2d} \Rightarrow N(0,1). \tag{15}$$

Fix n, K, σ and positive nuisance-independent scores, with almost surely finite q and log scores, as in Theorem 1.

Proposition 4 (Raw weights depend on coordinate scale)**.** *Let $\sigma_* = (2\pi e)^{-1/2} \simeq .241971$ and $1 \leq k \leq n$. If $\sigma > \sigma_*$, the raw-weighted finite-calibration set has the same all-label-or-empty limit as Theorem 1. If $\sigma < \sigma_*$, then $\phi(e_d)/(-a_\sigma d) \to 1$ in probability. The recalibrated set converges in probability to the coupled unweighted set provided each test score equals its unweighted empirical calibration threshold with probability zero. At $\sigma = \sigma_*$, $\phi(e_d)/s_d \Rightarrow N_+$, independently of the semantic variables; this is not generally an all-label-or-empty rule. Ordinary positive-score coverage is exactly $k/(n+1)$ for every finite positive d in all three regimes.*

Proof. In the first regime, $e_d/d \to -a_\sigma < 0$, so $\log \phi(e_d) - e_d \to 0$ and the nuisance-rank proof applies. In the second, $e_d/(-a_\sigma d) \to 1$ and $\phi(e_d) - e_d \to 0$ in probability; normalized scores and their empirical order statistic converge to the unweighted ones. The no-boundary-mass condition transfers this convergence to memberships. In the critical regime, $\sup_t |\phi(s_d t)/s_d - t_+| \leq \log 2/s_d$. Each finite-$d$ true score is continuous by conditioning on its positive base score and using the independent chi-square variable, giving the stated rank coverage.

Weight convergence alone does not remove genuine ties. If every candidate score equals one, the unweighted set is always full, whereas any continuous, nonconstant input weight gives a full set

with probability $k/(n+1)$, even as its relative variation vanishes. Our finite-group synthetic LAC comparisons retain their genuine atoms.

At the critical scale, population calibration at $v > 1/2$ has a positive limiting threshold t_v solving $\mathbb{E}\Phi(t_v/S_Y) = v$. Each candidate is included with limiting probability $\mathbb{E}\Phi(t_v/S_j) \geq 1/2$. For fixed n, the limiting normalized calibration scores have an atom at zero. One valid lower bound is

$$\liminf_d \Pr(C_d = [K]) \geq \frac{\mathbb{E}\min\{k, J\}}{n+1}, \qquad J \sim \mathrm{Bin}(n+1, 1/2). \tag{16}$$

Indeed, a negative test nuisance normal below the kth calibration normal eventually includes every candidate. The normalized log-score comparison resolves the case where both normals are negative. Conditional on J, the test index is among the first $\min(k, J)$ with probability $\min(k, J)/(n+1)$.

Exact reference invariance. Under the same standardized nuisance draw, $f_{d,\sigma}(A, \sigma G) = f_{d,1}(A, G) - d\log\sigma\,\mathbf{1}$. Code energy at any τ changes by $-d\log\sigma/\tau$; paper energy changes by $-d\log\sigma$. Each rule's own reference mean cancels that constant exactly, including a finite reference sample and nonconstant posterior. Consequently reference weights and probability-only rules are identical across coupled scales. A catalogue containing raw candidates need not select identically. This is a consequence of the known representation dependence of density (Le Lan & Dinh, 2021), not a new coordinate-invariant density estimator.

A finite-dimensional sufficient condition. Let $V_i = -\frac{d}{2}\log(2\pi\sigma^2) - \|G_i\|^2/2$, and suppose the realized calibration true scores and test candidate scores have $H = q + \log S$ in an interval of width L. If every calibration/test energy satisfies $q + V \leq -B$, $B > 0$, put $D = L + e^{-B}$. Order-statistic monotonicity gives

$$V_* < V_{(k)} - D \implies C_d = [K], \qquad V_* > V_{(k)} + D \implies C_d = \varnothing. \tag{17}$$

Here $V_{(k)}$ is the calibration nuisance order statistic. The proof uses $-e^{-B} \leq \log\phi(e) - e \leq 0$ for $e \leq -B$: all log scores lie in $V + [H_{\min} - e^{-B}, H_{\max}]$.

For an unconditional version, let $q \leq q_{\max}$ and $H \in [h_-, h_+]$ hold for the batch with probability at least $1 - \eta$, where $L = h_+ - h_-$. For $d \geq 2$ let M_d be the maximum density of V: $M_2 = 1$ and $M_d = (d/2-1)^{d/2-1}e^{-(d/2-1)}/\Gamma(d/2)$ for $d > 2$. If $a_\sigma d \geq q_{\max} + B + \sqrt{dt}$ for $t > 0$, Gaussian concentration (Laurent & Massart, 2000) and independence of test and calibration nuisance imply

$$\Pr(C_d \neq C_d^V) \leq \min\{1, \eta + (n+1)e^{-t} + 2(L + e^{-B})M_d\}, \tag{18}$$

where $C_d^V = [K]$ if $V_* \leq V_{(k)}$ and is empty otherwise. The bound times K also controls the difference in expected sizes. Specifically, the chi-square lower-tail inequality bounds failure of the negative-energy condition; a test nuisance value lies within D of the independent calibration threshold with probability at most $2DM_d$. Since $M_d \sim 1/\sqrt{\pi d}$, this sufficient condition requires semantic and log-score spread small relative to $\sqrt{d}$. Randomized scores near zero need a high-probability envelope with an explicit η. The bound can be vacuous at moderate dimension and does not predict complete-catalogue selection.

A.3 Proof of the near-constant-weight result

Under Theorem 2, write $A_r = \sum_{h \leq r} a_h$, $A_0 = 0$, and $S_r = A_{r-1} + U_r a_r$. The population thresholds are

$$q_0 = a_1 + \frac{1 - \alpha - c}{b} a_2, \qquad q_w = R a_1. \tag{19}$$

Proof. Since c is the unique largest probability, $a_1 = \{1 + \sum_{r \geq 2}(\pi_r/c)^{1/T}\}^{-1} \to 1$ as $T \downarrow 0$, so the condition $Ra_1 > 1$ in Theorem 2 is achievable at finite T. Put $h = (1 - \alpha - c)/b$. We have $h > 0$ because $1 - \alpha - c = (1 - \rho)(1 - c)$, and $h < 1$ because $\eta < \alpha$. For $q \in [a_1, A_2]$, the unweighted true-label score CDF is $c + b(q - a_1)/a_2$. It first reaches $1 - \alpha$ at q_0. Rank one is always included, rank two is included with probability h, and all later ranks are excluded, giving the first size formula.

At $q_w = Ra_1 > 1$, the low-weight group's scores are all at most 1, so its set contains every label. On H, rank one is always included and ranks two and above are excluded almost surely, since their scores are at least Ra_1, with equality only at a uniform endpoint of probability zero. Thus the weighted

coverage is $(1-\rho)+\rho c = 1-\alpha$ and its mean size is $(1-\rho)K+\rho$. For every $0 \le q < Ra_1$, the high-weight group contains no non-top true label, and its top-label contribution to coverage is $\rho c q/(Ra_1)$. The low group's contribution is at most $1-\rho$, so the total CDF is strictly below $1-\alpha$. Hence q_w is the population quantile, not merely a feasible threshold.

For the density realization, take two disjoint intervals I_L, I_H of lengths $(1-\rho)/(e-1)$ and $\rho/(e^R-1)$, and define the input density as $d_L = e-1$ and $d_H = e^R-1$ on them, respectively. Its integral is one. Set $f_r(x) = \log \pi_r + \log d_g$ on I_g. Then $\exp f_r(x)$ is the exact joint density, $\mathrm{softmax}(f) = \pi$, and $\mathrm{softmax}(f/T) = a$ in both groups. Moreover, $e_1(f) = \log \sum_r \exp f_r = \log d_g$, so $\mathrm{softplus}(e_1(f)) = \log(1+d_g)$ is exactly 1 or R. Finally, $0 \le S_r \le 1$, so the weighted and unweighted APS scores differ by at most $R-1$.

Quantifier and scope. For a sequence $R \downarrow 1$, the temperature may also decrease; the result does not assert this separation as $R \downarrow 1$ at one fixed T. The posterior π need not change along that sequence. No numerical truncation, class-order change, or distribution shift between calibration and test is used. With $c = .9$, $b = .0999$, $\eta = 10^{-4}$ and $\rho = .1$, the unweighted mean is $1.9009009\ldots$ and the weighted mean is $.9K + .1$. This is a lack of uniform efficiency stability over temperature, not a claim of a discontinuity for a fixed positive probability vector.

Corollary 5 (Finite calibration for a separately defined nonempty variant)**.** *Under the theorem's assumptions, replace the top-rank score by zero in* both calibration and prediction*, keeping every other APS score unchanged. Denote this score by S^+. Calibrate S^+ and wS^+ separately on n independent observations, with $k = \lceil (n+1)(1-\alpha) \rceil \le n$ and $\ell = n-k+1$. Both resulting rules have nonempty sets and marginal coverage at least $1-\alpha$. Writing $B_n(v) \sim \mathrm{Bin}(n, v)$,*

$$\mathbb{E}|C_w^+| \ge 1 + (1-\rho)(K-1)\mathbb{P}\{B_n(\alpha) \ge \ell\}, \tag{20}$$

$$\mathbb{E}|C_0^+| \le 2 + (K-2)\mathbb{P}\{B_n(\eta) \ge \ell\}. \tag{21}$$

Expectations include calibration data, the independent test input and the uniform randomization. Under population calibration, this variant retains both sizes and exact coverages in Equation (5)*, but its weighted population quantile is* 1*, rather than Ra_1.*

Proof. Every empirical threshold is nonnegative and every top-rank score is zero, so every set contains rank one. Split-conformal validity applies to each fixed score, including the atom at zero; we claim the usual lower coverage bound, not the continuous-score equality $k/(n+1)$.

For weighted calibration, the event $wS^+ \ge Ra_1$ occurs precisely when $X \in H$ and Y is not rank one, up to uniform endpoints of probability zero. Its probability is $\rho(1-c) = \alpha$: low-group scores are at most $1 < Ra_1$ and all top scores are zero. The kth order statistic is at least Ra_1 exactly when at least ℓ calibration scores lie in this event. On that event every low-group test set has size K and every other test set has size at least one. Outside it every test set still has size at least one. Independence of calibration and test gives Equation (20).

For unweighted calibration, a true-label score is at least A_2 precisely when its label has rank at least three, again outside a probability-zero uniform endpoint. This event has probability η. If fewer than ℓ calibration scores lie in it, the calibration threshold is strictly below A_2 and no test label of rank three or higher is included; the size is at most two. Otherwise it is at most K, proving Equation (21).

For population calibration, the unweighted CDF on $[a_1, A_2]$ is unchanged. The weighted CDF first reaches $1-\alpha$ at $q = 1$, and stays at that value through Ra_1: the low group is then fully covered and the high group contains only its zero-score top label. For $q < 1$ the low group has coverage strictly below one, because its last label has positive probability and a score interval ending at one. Its total coverage is therefore below $1-\alpha$. This proves the stated quantile and the unchanged population sets.

Relation to the SAPS contrast bound. On the same constant posterior and at the same T, SAPS with fixed $\lambda > 0$ has unweighted population quantile $q_S = a_1 + \lambda h$ and mean size $1+h$, where $h = (1-\alpha-c)/b$. For any weights in $[1, R]$, its weighted quantile is at most Rq_S, since each weighted true-label score is at most R times the unweighted score. Its effective test threshold is therefore at most Rq_S. The SAPS expected size above a_1 has slope $1/\lambda$ until it is capped at K, yielding Eq. (22). This is a direct specialization of the existing SAPS contrast bound, not a second

Table 3: Deterministic checks of the near-constant-weight construction with $K = 100$, $\rho = .1$, and $\alpha = .01$. Both APS arms use the same score probabilities at the listed temperature and attain exactly 99% population coverage. Their mean sizes are shown; the final column is the SAPS size upper bound at $\lambda = .2$, rounded upwards. These are fixed-rule calculations, not selected-procedure outcomes.

$R-1$	T	$Ra_1 - 1$	APS unweighted	APS weighted	SAPS bound
10^{-2}	0.2	9.982981×10^{-3}	1.900901	90.1	1.959909068
10^{-6}	0.1	9.997161×10^{-7}	1.900901	90.1	1.900906802
10^{-12}	0.05	9.999999×10^{-13}	1.900901	90.1	1.900900901

The exact posterior is fixed across rows: $\pi_1 = .9$, $\pi_2 = .0999$, and $\pi_{j+3} = 10^{-4}2^{97-j}/(2^{98} - 1)$ for $j = 0, \ldots, 97$. All numerical checks use 850-digit arithmetic and true-label probabilities π, rather than their temperature transform a. The limit $R \downarrow 1$ permits T to vary.

new general theorem. Since $q_S \leq 1 + \lambda$, its extra term vanishes uniformly over T as $R \downarrow 1$, whereas Theorem 2 has an order-K APS gap along a joint (R, T) sequence. Neither statement establishes that an efficiency selector will choose the failing rule.

Deterministic numerical verification. Table 3 checks three finite positive temperatures and weight contrasts using 850-digit arithmetic. Every candidate interval remains strictly positive, including the last interval; true-posterior coverage and predicted size are integrated separately. The normalized two-interval input density gives the prescribed raw softplus weights. Direct binomial summation independently checks the finite-calibration bounds at $n = 3500$, $k = 3466$: $\Pr\{\mathrm{Bin}(3500, .01) \geq 35\} = .5228195178$. The weighted nonempty lower bound is 47.583219, and the unweighted upper bound is $2 + 6.29521 \times 10^{-55}$.

A.4 A SCORE-DEPENDENT CONTRAST CONTROL

For the constant-posterior population construction in Theorem 2, the SAPS rank increment provides an explicit contrast scale:

$$\mathbb{E}|C_{S,w}| \leq 1 + h + (R-1)q_S/\lambda, \qquad h = (1 - \alpha - c)/b, \quad q_S \leq 1 + \lambda. \tag{22}$$

This gives the population counterpart to the finite-calibration comparison below.

For a fixed input with leading probability c and SAPS parameter $\lambda > 0$,

$$S_r = \begin{cases} cU_r, & r = 1, \\ c + \lambda(r - 2 + U_r), & r \geq 2, \end{cases} \qquad U_r \sim \mathrm{Unif}(0, 1). \tag{23}$$

Its expected cardinality at a nonnegative threshold t is

$$G_c(t) = \min\{t/c, 1\} + [(t - c)/\lambda]_0^{K-1}. \tag{24}$$

If calibration and test weights lie in $[a, b]$, let $R = b/a$ and let q_S be the finite unweighted calibration order statistic. Monotonicity of order statistics gives $q_w \leq bq_S$, hence

$$C_w(x) \subseteq \{y : S_y(x) \leq Rq_S\}, \qquad \mathbb{E}_U|C_w(x)| \leq G_c(Rq_S) \leq \min\{K, 1 + Rq_S/\lambda\}. \tag{25}$$

More locally, with $b_{\mathrm{cal}} = \max_i w(x_i)$ one can replace R by $b_{\mathrm{cal}}/w(x)$. The proof follows by summing the uniform inclusion ramps in Eq. (23), giving Eq. (24); this uses the existing SAPS score formula (Huang et al., 2024). It is an elementary control, not a new general stability theorem. If q_S is bounded and λ is bounded away from zero as K grows, order-K expected size requires $R = \Omega(K)$. A large ratio is not sufficient for poor efficiency. The bound does not apply to an infinite calibration threshold and its SAPS cardinality expression does not apply to APS or LAC.

B PROOF AND SCOPE OF THE ADDITIONAL-LABEL BUDGET

Conditional setup. Condition on external information $\mathcal{E}$ that fixes the base score S_0, target score S_1, budget $B \geq 0$, and scoring conventions. The $N = n + 1$ future rows (X_i, Y_i, U_i) are exchangeable conditional on $\mathcal{E}$; U_i includes any score randomization. Scores are finite and smaller scores are more

conforming. Write $s_i(y) = S_0(X_i, y, U_i)$ and define the inclusive all-label cost map

$$H(t) = \frac{1}{N}\sum_{j=1}^{N}\sum_{y=1}^{K}\mathbf{1}\{s_j(y) \le t\}, \qquad T_i(y) = \big(H(s_i(y)) + a_i(y),\ s_i(y)\big), \quad |a_i(y)| \le B/4. \tag{26}$$

The modulation array a is label-free and equivariant under row permutations. It may use all pooled candidate scores, but cannot distinguish the query from calibration rows. Thus one shared array T applies to every analytical leave-one-out designation. The target may otherwise be arbitrary: the theorem requires no accuracy or density assumption about it.

Keys are ordered lexicographically, with the original base score as the secondary coordinate. Let $k = \lceil N(1-\alpha)\rceil$ for $0 \le \alpha < 1$. For each held-out row i, separately calibrate s and T using the other $N-1$ true-label scores and their kth order statistics, denoted q_{-i} and Q_{-i}. Both prediction sets use inclusive thresholds. When $k = N$, both thresholds are infinite and both sets contain all labels. Denote these sets by $C_{0,-i}$ and $C_{B,-i}$.

Finite-pool statement. The following deterministic inequality proves the cost part of Theorem 3, including ties:

$$\frac{1}{N}\sum_{i=1}^{N}|C_{B,-i}(X_i) \setminus C_{0,-i}(X_i)| \le \min\{K, d_N(B)\}, \qquad d_N(B) = \frac{2\max\{0, \lceil NB/2\rceil - 1\}}{N} \le B. \tag{27}$$

Proof. It suffices to consider $k < N$. For a finite base threshold $q = q_{-i}$, at least k calibration true scores are at most q. Their transformed keys are at most $(H(q) + B/4, q)$, so Q_{-i} is also at most this key. A newly included candidate has base score $s > q$ and transformed key at least $(H(s) - B/4, s)$. Its inclusion therefore requires

$$s > q, \qquad H(s) - H(q) < B/2. \tag{28}$$

The inequality is strict: equality of the primary coordinates leaves the candidate's secondary score $s > q$, which precludes inclusion.

Let $J(t) = NH(t)$ and $c = \max\{0, \lceil NB/2\rceil - 1\}$. For fixed q, at most c pooled candidate-label pairs satisfy (28). Indeed, if this window is nonempty, let s_* be its largest score. Its entire number of pairs is $J(s_*) - J(q) < NB/2$, an integer at most c. Inclusive counting includes each admitted tie block in full. Deleting one true score from a pool makes the kth remaining true score equal to either the full-pool kth or $(k+1)$th true-score value. Every added pair (i, y) therefore belongs to the union of at most two windows, each containing at most c pairs. This proves (27). Coincident threshold values need only one window. The case $k = N$ has zero additions.

Conditional exchangeability and equivariance make the transformed true keys exchangeable. The inclusive conformal rank argument consequently gives $\mathbb{P}\{Y_N \in C_{B,-N}(X_N) \mid \mathcal{E}\} \ge k/N \ge 1 - \alpha$. Every held-out designation has the same expected number of additions. Taking expectations in (27) proves

$$\mathbb{E}[|C_B(X) \setminus C_0(X)| \mid \mathcal{E}] \le d_N(B) \le B. \tag{29}$$

Since $(|C_B| - |C_0|)_+ \le |C_B \setminus C_0|$, the same upper bound applies to positive net size increase, and hence to signed expected size increase. At $B = 0$, $(H(s), s)$ has exactly the original score preorder, so the base sets are recovered, including ties.

The expectations average over final calibration and query randomness, conditional only on $\mathcal{E}$. They are not conditional on a realized calibration sample, per-input caps, or guarantees under test-only shift. External selection of the base, target, or budget is permitted by this conditioning; selection using final calibration labels is not covered.

Corollary 6 (Nonempty repair of the finite-calibration construction)**.** *Suppose both scores assign the same top label on each row a shared global minimum, with all other scores strictly larger. For the clipped cost projection, every predicted set is nonempty. In the finite-calibration top-score-zero*

Algorithm 1 Pooled cost projection for one query

Inputs: calibration rows $\{(X_i, Y_i)\}_{i=1}^n$, query x, and externally frozen (S_0, S_1, B, α).

1. Form the $N = n + 1$ input pool with $X_N = x$. Evaluate both scores for every candidate label, using the frozen randomization convention.
2. Sort the NK scores of each rule. Construct H_0, H_1 by inclusive pooled counting in Eq. (6), retaining whole tie blocks.
3. Transform every candidate to the key $T_i(y)$ in Eq. (7). Use the original base score as the secondary coordinate; compare keys lexicographically.
4. Read the calibration true-label keys and take their kth order statistic, $k = \lceil N(1-\alpha) \rceil$. Set the threshold to infinity if $k = N$.
5. Return all query labels whose transformed keys are at most this threshold. Repeat with a separate pool and threshold for each new query.

Comparator: recalibrate S_0 independently using the same labeled rows and rank. The pooled map uses no labels; the query label is never accessed.

APS construction of Appendix A.3, $K = 100$, $n = 3500$, and $\alpha = .01$. With $\varepsilon_* = 6.30 \times 10^{-55}$ and $B = .5$,

$$\mathbb{E}|C_{\text{raw}}| \geq 47.583219, \qquad \mathbb{E}|C_0| \leq 2 + \varepsilon_*, \qquad \mathbb{E}|C_B| \leq 2.5 + \varepsilon_*. \tag{30}$$

The repaired rule has marginal coverage at least .99.

Proof. Each score has exactly N top-label entries at its global minimum, so both top-label cost coordinates equal 1; all other coordinates exceed 1. The projected coordinate $H_0 + \text{clip}(H_1 - H_0, -B/4, B/4)$ lies between H_0 and H_1. Every top key thus has primary coordinate 1 and the shared minimum base secondary score, whereas every non-top key has primary coordinate greater than 1. Every calibration true key is at least the top key. Recalibration therefore always includes the top label, without appending a label afterward. The two original size bounds are established in Appendix A.3; adding (29) to the base bound gives the repaired bound. This argument does not assert equal finite-sample coverage.

A protection–adaptation tradeoff. If all NK base candidate scores are distinct almost surely, the same construction also satisfies

$$\mathbb{E}|C_0 \setminus C_B| \leq d_N(B) \leq B, \qquad \mathbb{E}|C_0| - \mathbb{E}|C_B| \leq B, \tag{31}$$

with the same external conditioning as above. To see this, at most $k - 1$ calibration true scores are below $q = q_{-i}$. All others have keys at least $(H(q) - B/4, q)$, so Q_{-i} is at least this key. A removed candidate must satisfy $s \leq q$ and $H(q) - H(s) < B/2$. Under the stated distinctness condition, a held-out candidate cannot equal a calibration threshold, so $s < q$. For each of the same two possible thresholds, only c integer ranks can satisfy $1 \leq J(q) - J(s) < NB/2$. The counting and exchangeability argument then proves (31).

This removal bound needs distinctness of *all candidate scores*, not only true scores. With ties it can fail: if every base score is zero and the true label is always 1, the base set has size K. A fixed target ranking label 1 ahead of tied remaining labels makes the projected set exactly $\{1\}$ for any $0 < B < 4(K-1)$, with coverage one and no additions. In particular, (31) must not be applied to the top-score-zero construction above. Where its premise holds, it explains why a small absolute protection budget also limits attainable mean-size gains; it does not rank projection against other protected adaptations.

C Protected adaptation: algorithm and implementation

External fitting and selection. Before final calibration, fit or select the base score S_0, target score S_1, temperatures, feature transformations, and budget B using independent training and auxiliary data. The method does not require backbone retraining; an already fitted energy rule or learned score may serve as the target. An auxiliary selection procedure can include the base fallback $B = 0$. Its output is frozen before any final calibration labels are used. The additional-label theorem conditions

on these external choices. Auxiliary selection determines which target to use; the theorem controls adaptation of the selected rule on the independent final rows.

Algorithm 1 gives the single-query implementation. All comparisons use inclusive thresholds and exact lexicographic ordering.

The base comparator is separately calibrated at the same rank using its original scores and the same calibration rows. No query label is used. Each new query induces its own pool and threshold; combining all test inputs into one pool is a different procedure. The construction treats calibration and query rows symmetrically before reading calibration labels, which supplies the exchangeability used in Appendix B.

A strong same-budget interpolation control. Let $r = B/4$ and $D = \max_{i,y} |\Delta_i(y)|$ on the same unlabeled pool. The global interpolation

$$a_i^{\mathrm{lin}}(y) = \lambda_D \Delta_i(y), \qquad \lambda_D = \min\{1, r/D\} \quad (D > 0) \tag{32}$$

uses the base when $D = 0$. Its modulation is bounded by r, so it has exactly the same coverage and additional-label guarantee. The more conservative fixed coefficient $\min\{1, r/K\}$ also satisfies the premise. Computing D from the pool is permitted because it is label-free and unchanged by row permutations.

Projection and interpolation instantiate the same bounded cost-coordinate framework. Their distinction is the amount of target movement retained at intermediate displacements. Algebraically,

$$|\operatorname{clip}(\Delta_i(y), -r, r)| \geq |\lambda_D \Delta_i(y)|. \tag{33}$$

The inequality is strict precisely when $D > r > 0$ and $0 < |\Delta_i(y)| < D$. Thus a large pooled maximum can suppress intermediate requested movements under global shrinkage, while projection retains more of them. This coordinate-level distinction motivates the empirical comparison: separate recalibration determines whether the retained movement translates into smaller sets. Adaptive interpolation supplies the matched-guarantee control for that question.

Ordering and numerical representation. Original scores, or codes representing exactly the same score preorder, are the secondary keys. Row indices and arbitrary floating-point epsilons must not replace them. Let $J_j(s) = NH_j(s)$ denote the inclusive integer counts, evaluated at the corresponding candidate score. For quarter-grid budgets, let $b = 4B$ be an integer. An integer primary key proportional to (7) is

$$16J_0 + \operatorname{clip}\big(16(J_1 - J_0), -bN, bN\big). \tag{34}$$

Division by $16N$ recovers the displayed primary coordinate exactly. Use adequate integer range and preserve inclusive score ties. Low-temperature APS requires stable tail/complement ordering before this transform; the projection cannot restore distinctions already lost in rounded input scores.

Computational cost. A naive exact implementation sorts the two NK score arrays separately for each query, taking $O(NK \log(NK))$ time and $O(NK)$ memory, in addition to model evaluation. Calibration order statistics cost no more than this bound. Calibration model outputs can be cached, but the query-specific pooled ranks and transformed threshold still require updating. Sorted calibration indexes and exact insertion/counting operations can reuse work without changing the stated worst-case bound. For M queries, direct inference repeats this calculation M times. Model evaluation and query-specific conformal updates are therefore reported as separate costs.

D EXPERIMENTAL PROTOCOLS

Scope and reporting. The evaluation comprises the nonempty mechanism constructions, a new energy-fine-tuned WRN checkpoint, and a frozen ImageNetV2 base paired with ResNet-50 and ConvNeXt targets. Every candidate is calibrated separately after its fitting and auxiliary choices are fixed. Entries report means and sample standard deviations, with denominator $r - 1$ for r repetitions. Mechanism variability comes from twelve independent data/calibration draws; real-data variability describes overlapping role splits of fixed checkpoints, not independently trained models. The protocol freezes are internal records rather than external preregistrations. Nominal coverage, attained empirical coverage, and posterior-expected coverage are distinguished where applicable.

Scores, weights, and comparison rules. We use $e_\tau = \mathrm{LSE}(f/\tau)$ and the positive raw weight $w(x) = \mathrm{softplus}(e_\tau(x))$. A weighted positive SAPS score is $w^\gamma S$; weighted LAC is equivalently ordered by $-\log p_T(y) + \gamma \log w$. The reference-median weight-clip control acts on centered log weights and is distinct from PCP's label-count projection. Score randomization is independent of labels and shared by paired rules. Protected rules use the same separately recalibrated base and formal budget; an unweighted rule with a different score family or an unprotected direct target is an external comparator. At $B = 0$, PCP and adaptive linear reproduce their common base. Numerical details and tie handling appear in Appendix C.

Real-data roles. The new WRN/CIFAR-100 experiment uses five overlapping partitions of one energy-fine-tuned checkpoint, with 1,000 reference, 2,000 gate-fit, 2,000 auxiliary, 2,000 final-calibration, and 3,000 test images per partition. Its fixed SAPS base, energy-strength library, auxiliary gate policy, and prespecified budgets are detailed in Appendix E. The ImageNetV2 matched-frequency experiment reuses a ResNet-18 base and three overlapping partitions; each assigns 1,000 reference, 2,000 auxiliary, 2,000 final-calibration, and 5,000 test inputs. ResNet-50 and ConvNeXt-Tiny use the same image pool and independently frozen target protocols. Appendices F and G give their budget grids, response-fault construction, coverages, and auxiliary choices. The exploratory archive comparison in Appendix H is a score-family check, not an independent checkpoint replication.

Table 4: **Sample standard deviations for the means in Table 1.** Size SD is in labels per input; coverage SD is in percentage points. Estimates describe five overlapping WRN or three overlapping ImageNetV2 role splits of fixed checkpoints, not independent model seeds.

	WRN / CIFAR-100		**ImageNetV2 / ResNet**		**ImageNetV2 / ConvNeXt**	
Rule	Size SD	Cov. SD	Size SD	Cov. SD	Size SD	Cov. SD
Base	2.274	0.208	0.733	0.332	3.157	0.428
Fixed linear	2.290	0.208	0.745	0.336	3.124	0.437
Adaptive linear	2.259	0.220	0.744	0.336	3.121	0.437
PCP (ours)	2.397	0.222	1.355	0.269	2.605	0.393

Mechanism constructions. These experiments have $K = 100$, nominal coverage 99%, and constant posterior $(0.9, 0.0999, 10^{-4}g)$, where g is the normalized length-98 geometric vector with ratio $1/2$. Gaussian nuisance dimensions are $d \in \{0, 16, 256\}$ at unit scale, with log density $-d\log(2\pi)/2 - \chi^2_d/2$ and its raw softplus weight. SAPS uses temperature one and increment 0.2. The near-constant construction assigns weight $1 + 10^{-6}$ to an independent group of probability 0.1 and weight one otherwise; APS and the same-temperature SAPS control both use $T = 0.1$. Weight clipping centers log weights by their reference median and clips them to $[-0.1, 0.1]$.

Seeds 82101–82112 generate 12 independent data/calibration repetitions, including new calibration and test draws in each repetition; methods within a repetition share the same observations. The saved roles contain 1,000 reference, 2,000 tuning-calibration, 4,000 tuning-validation, 3,500 final-calibration, and 12,000 final-test observations. This fixed-rule repair experiment performs no candidate selection or learning on the tuning roles. Both ordinary scores and the nonempty variant are retained: the latter assigns the top-label score a common strict minimum *before* constructing maps and recalibrating each rule, rather than appending a label after calibration. Table 13 reports the nonempty Gaussian and near-constant APS comparisons with their coverages. The same-temperature SAPS boundary control is summarized in Appendix I. Posterior-expected coverage integrates the known conditional label probabilities on the sampled test inputs; it is not the population expectation over all future calibration samples.

Numerics and computational scope. Order codes and integer primary keys preserve inclusive ties and original-base tie breaking. Compressed APS tails use stable complement representations with high-precision replay; LAC ordering uses the recorded interval/Arb backend with Decimal fallback. These checks address numerical ordering and do not turn a floating-point guard into a global formal certificate. Results save per-input sizes, coverage indicators, packed membership masks, protocol identities, and source/input hashes.

The runtime panel uses saved wall times for the complete five-budget calls on identical 100-class mechanism workloads, with 12 repetitions. It excludes neural inference, initial score construction,

and result serialization. The mechanism freeze records Python 3.12.14 and NumPy 2.3.5; the per-call timing records do not provide sufficient machine metadata for a hardware-normalized speed comparison. Separately, ImageNetV2 export records an NVIDIA GeForce RTX 4090, PyTorch 2.4.1/CUDA 12.4, torchvision 0.19.1, batch size 128, and eight data workers. Both model forwards are required: we do not claim ResNet-18-only inference cost. Timing and size summaries therefore describe different, explicitly identified computational stages.

E NATIVE-ENERGY ADAPTATION WITH AUXILIARY BUDGET SELECTION

Model and roles. Before training this checkpoint, we fixed the scientific plan, role seeds, candidate grid, selection objectives and descriptive success criteria in an internal protocol. This was not an external preregistration; earlier WRN and ImageNetV2 studies informed the design. A WRN-40-2 was energy-fine-tuned for ten epochs from a shared CIFAR-100 cross-entropy checkpoint using CIFAR-100 training images and SVHN training images as auxiliary outliers. Seed 2026092302 and the terminal epoch indexed 9 were fixed without evaluating conformal or test outcomes. The resulting 10,000 CIFAR-100 test logits were exported by a direct FP32 network forward pass, before centering or softmax. The new run is a separate energy fine-tune, not training from an independent initialization or validation on a new image pool.

Each of five seeded partitions assigns 1,000 inputs to an unlabeled reference map, 2,000 to fit the optional gate, 2,000 to auxiliary selection, 2,000 to final calibration and 3,000 to final test. These roles are disjoint within a partition, but the partitions overlap and reuse one checkpoint and image pool. The fixed base is SAPS at temperature one with rank increment .2; the native target multiplies its positive score by $\mathrm{softplus}(\mathrm{LSE}(f))$. The target strength is one for PCP. Adaptive linear may choose strength $\gamma \in \{.25, .5, 1, 2\}$ from the same native logits. Both protected families use $B \in \{0, .5, 2, 4, 8\}$ and the same final-calibration observations. Thus they share the formal same-base addition bound, though their candidate libraries differ.

Selection. The auxiliary role is divided into five held-out folds of 400 observations. Each candidate is calibrated on the other 1,600 and evaluated on the held-out fold; the pooled 2,000 validation predictions determine its objective. The primary selector chooses the smallest auxiliary mean size among candidates with auxiliary mean additions at most .2 labels per image relative to the separately calibrated SAPS base. The $B = 0$ base is always eligible. A second, separately prespecified selector minimizes auxiliary mean size plus mean additions. Neither the .2 preference nor the selected candidate's validation cost is a final-test guarantee. All five selections for both objectives were sealed before any final-calibration or test evaluation. The unconditional expectation bound for the selected formal B follows from Theorem 3 because the final rows are not used to select it.

An optional binary gate is fitted only on the unlabeled reference and labeled fitting roles. The four candidate functions are uniform movement, positive target cost displacement, proximity within two label-count units of a fitting-role base threshold, and the intersection of the latter two. Each gate acts on every candidate row without its label and is frozen before final calibration. The uniform gate reproduces original PCP exactly. Gate search and adaptive-linear strength search each have four candidate types and the same five budgets; ungated PCP is reported separately as an ablation. A strong unweighted LAC comparator selects its temperature from nine predeclared values using the same auxiliary folds, then receives independent final calibration. It is a different score family and is not subject to PCP's same-SAPS-base addition budget.

Interpretation. The auxiliary-selected PCP family returns smaller sets than adaptive linear in all five overlapping role splits, selecting $B = 8$ throughout and the uniform gate in four splits. The methods share the same formal addition budget, while their realized additions differ: in split 4, PCP adds 0.0997 labels per input relative to SAPS, compared with 0.0213 for adaptive linear. A retrospective inspection of the prespecified budget grid additionally shows that, at $B = 2$, ungated PCP with $\gamma = 1$ returns smaller sets than the adaptive-linear $\gamma = 2$ configuration in all five splits, with no observed PCP additions. This observation supplements the primary auxiliary-selected policy comparison.

The auxiliary-selected unweighted LAC comparator is substantially smaller at the same nominal 99% coverage, although its attained coverage differs slightly (Table 6). Consequently, this experiment supports a fixed-SAPS, same-formal-budget local advantage for PCP, not overall score-family superiority. The original experimental pipeline recorded the candidate grid, frozen policy sets, source and

Table 5: All five sealed auxiliary selections on the new WRN checkpoint. Split IDs 1–5 correspond, in order, to archived seeds 2026093701–05. Each split has 2,000 independent final-calibration inputs and 3,000 test inputs; image pools overlap across splits. All entries are observed test means, not independent-run uncertainty intervals. The empirical auxiliary cap is 0.2 added labels/input and is not a final-test guarantee. Gate search selects the uniform/ungated candidate in splits 1–4 and a positive-displacement gate in split 5. In split 4, gate search has 0.0997 observed additions versus 0.0213 for adaptive linear despite a smaller set.

Split	Rule	B	Size ↓	Cov. (%)	Added ↓	Removed ↑	Net ↑
1	SAPS base	0	18.164	98.933	0.0000	0.0000	0.0000
	Adaptive linear ($\gamma = 2$)	8	17.874	98.933	0.0013	0.2920	0.2907
	PCP (ungated)	8	15.925	98.867	0.0000	2.2390	2.2390
	PCP + gate search	8	15.925	98.867	0.0000	2.2390	2.2390
2	SAPS base	0	16.991	98.700	0.0000	0.0000	0.0000
	Adaptive linear ($\gamma = 2$)	8	16.850	98.700	0.0133	0.1543	0.1410
	PCP (ungated)	8	14.852	98.667	0.0000	2.1390	2.1390
	PCP + gate search	8	14.852	98.667	0.0000	2.1390	2.1390
3	SAPS base	0	19.516	98.800	0.0000	0.0000	0.0000
	Adaptive linear ($\gamma = 2$)	8	19.305	98.800	0.0107	0.2217	0.2110
	PCP (ungated)	8	17.346	98.767	0.0103	2.1797	2.1693
	PCP + gate search	8	17.346	98.767	0.0103	2.1797	2.1693
4	SAPS base	0	23.027	99.133	0.0000	0.0000	0.0000
	Adaptive linear ($\gamma = 2$)	8	22.800	99.167	0.0213	0.2490	0.2277
	PCP (ungated)	8	21.130	99.167	0.0997	1.9973	1.8977
	PCP + gate search	8	21.130	99.167	0.0997	1.9973	1.8977
5	SAPS base	0	18.936	98.600	0.0000	0.0000	0.0000
	Adaptive linear ($\gamma = 2$)	8	18.767	98.600	0.0097	0.1790	0.1693
	PCP (ungated)	2	18.387	98.600	0.0000	0.5490	0.5490
	PCP + gate search	8	18.040	98.600	0.0027	0.8987	0.8960

Table 6: Strong unweighted LAC comparator selected using the same auxiliary roles, with its own final conformal calibration at nominal 99% coverage. Each row uses 3,000 test inputs from a role split of the same image pool. Added/removed/net are relative to the separately calibrated SAPS base and are not covered by PCP's formal B budget. Its average size is substantially smaller than the SAPS-family results in Table 1; attained coverage is also lower (98.673% versus 98.813% for selected PCP), so this table is not an equal-attained-coverage efficiency ranking. The reported SD is descriptive across overlapping splits.

Split	T	Size ↓	Cov. (%)	Added	Removed	Net reduction
1	4	9.500	98.733	0.811	9.475	8.664
2	4	8.433	98.767	0.665	9.223	8.558
3	1	9.984	98.900	1.843	11.375	9.532
4	1	8.965	98.500	1.132	15.194	14.062
5	4	10.369	98.467	1.049	9.617	8.567
Mean	–	9.450 ± 0.775	98.673	1.100	10.977	9.877

checkpoint hashes, and integer-count audit. An independent replay of the primary 3,000-test-input selected comparison matched all reported integer totals. The overlap of role partitions precludes treating their sample SD as a standard error across independently trained models.

Gate ablation (exploratory). A retrospective comparison across the full nonuniform-gate grid holds the formal budget and target scale fixed. Relative to ungated PCP, nonuniform gates yield smaller sets in 1 of 60 candidate-by-split comparisons, equal sizes in 4, and larger sets in 55. Thus, nonuniform gating provides no consistent additional size reduction in the evaluated matched settings. In the only split where auxiliary selection chooses a nonuniform gate, the gated policy uses $B = 8$ and the separately selected ungated policy uses $B = 2$; their policy-level difference therefore combines changes in both gating and budget.

F BUDGET–UTILITY STUDY ON IMAGENETV2

F.1 FIXED SCOPE, DATA ROLES AND FAIR CONTROLS

This study extends the analysis of the ImageNetV2 matched-frequency pool and pretrained ResNet-18/ResNet-50 pair under a frozen protocol. Reusing the models and images isolates the effects of budget, direction selection, and final-role partitioning. The complete grid is three inherited split seeds, $\alpha \in \{.05, .01\}$, target-response fault probability $\rho \in \{0, .2\}$ and budgets $B \in \{0, .5, 2, 4, 8\}$. The primary comparison is PCP versus adaptive linear using the same teacher at $\alpha = .05, \rho = 0, B = 8$. It was fixed before computing the new auxiliary objectives and final outcomes. The stopping rule includes all twelve original conditions and thirty additional calibration repartitions, without adding models, budgets or seeds after inspecting outcomes.

Seeds 2026092301, 2026092302 and 2026092303 each assign 1,000 images to reference, 2,000 to auxiliary selection, 2,000 to final calibration and 5,000 to test, using all 10,000 images without cross-role overlap within a split. Base and teacher score configurations are inherited. At 95% coverage, the teacher temperature is one and the base temperature is one or two; at 99%, teacher and base temperatures are 25 and two. Consequently a difference between these operating points is not attributable to the nominal coverage level alone. There is no new classifier, learned gate or router training. Integer order codes preserve all source-score ties; their arbitrary numerical spacing is never treated as a score distance.

For each protected family and each fixed B, the candidate directions are the same frozen teacher and arithmetic probability mixtures with weights $.25, .5, .75, 1$, plus the base fallback. The two 1,000-image auxiliary folds exchange calibration and validation roles. Selection minimizes their summed validation set size, with a shared deterministic priority: base first, then teacher, then increasing mixture weight. This is a two-fold size selector, not a requirement that both folds separately improve. Every auxiliary selection is sealed before any new final evaluation. Separate final calibration is performed for each method and condition.

Faults replace the target response with a reference-donor response with probability .2, leaving the base and true label unchanged. Auxiliary, calibration and test use the same fault law conditional on the external donor bank. The fixed-teacher comparison isolates the protected rule's response to this stress. Selected directions use auxiliary data from each corresponding fault condition, and therefore measure condition-aware reselection. This stress is not evidence about the prevalence of natural failures or a guarantee under test-only distribution shift. The independent-reference comparator matches the displacement amplitude but does not have the same exact pooled addition certificate.

F.2 SENSITIVITY TO FINAL CALIBRATION

For each original 95% condition, five further random permutations repartition its fixed 7,000-image final pool into 2,000 calibration and 5,000 test images; reference, auxiliary roles and selected directions remain fixed. The seed construction uses the original split seed, repetition index and a separate fixed random-number namespace. Clean and faulty conditions use the same repartition. The resulting fifteen comparisons per condition reuse images and models and are not independent deployments. All are evaluated at $B = 8$.

The prerecording used an internal descriptive gate: improvement on all three original clean splits, mean gain at least one label and 3% relative to adaptive shrinkage, and improvement on at least twelve of fifteen clean repartitions. The observed 2.171-label (7.04%) original gain and fifteen successful repartitions pass this gate. These numerical cutoffs are design criteria, not hypothesis-test thresholds or a certificate of population superiority. Table 7 reports all four repeated comparisons. The original clean PCP-minus-adaptive differences are $-1.3972, -2.5812, -2.5334$ labels, with mean -2.1706 and sample SD $.6702$. Mean actual coverage is 94.753% versus 94.813%; independent recalibration does not enforce identical attained coverage on a finite batch.

Table 7: Repeated final calibration at nominal 95% coverage and $B = 8$. Within each of three original splits, the fixed 7,000-image calibration/test pool is repartitioned five times into 2,000 calibration and 5,000 test images; auxiliary choices remain frozen. Size and coverage columns are means over 15 partitions. The paired difference is PCP minus adaptive linear, reported as mean ± sample SD; wins count strictly smaller PCP sets. These are conditional checks on seen images, not 15 independent deployments or a confidence interval.

Target	Direction	PCP size / cov. (%)	Adaptive size / cov. (%)	Paired difference	Wins
Clean	Fixed	30.301 / 95.103	32.368 / 95.127	-2.067 ± 0.439	15/15
Clean	Selected	30.247 / 95.095	32.368 / 95.127	-2.122 ± 0.438	15/15
20% faults	Fixed	30.662 / 95.080	32.487 / 95.120	-1.825 ± 0.621	15/15
20% faults	Selected	30.537 / 95.061	32.487 / 95.120	-1.950 ± 0.600	15/15

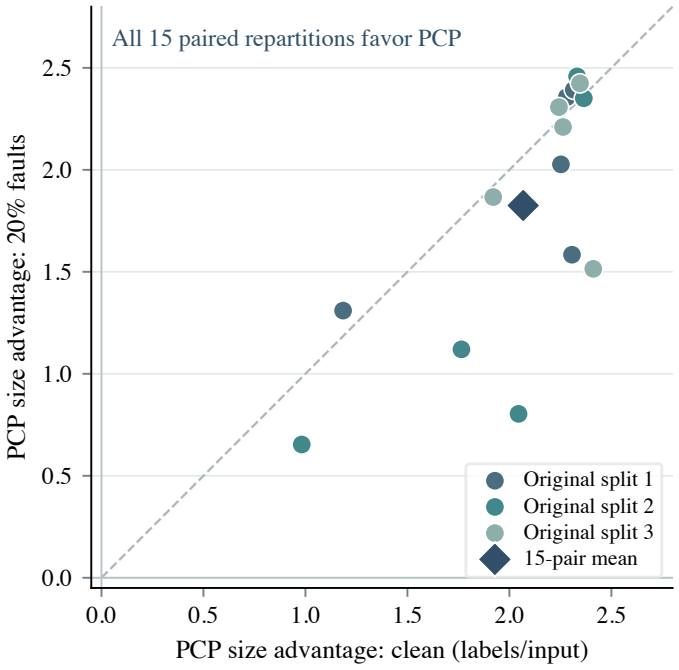

Figure 9: The ResNet-50 ImageNetV2 paired gain across final-calibration repartitions. Each point is a linear-minus-PCP size difference for the same fixed teacher and $B = 8$ at nominal 95% coverage; diamonds mark descriptive means of 2.067 (clean) and 1.825 (shared 20% response faults). All 15 pairs in each condition favor PCP. Models, image pool and auxiliary choices are reused across overlapping repartitions, so the points are not independent deployments.

F.3 A NECESSARY WINDOW FOR USEFUL DELETIONS

Fix one calibration-plus-query pool, a finite original base threshold q_0 with $k \le n$, and any bounded modulation $H_0(S_0) + a$ with $|a| \le B/4$, using the original base score as its secondary key. Recalibration of the transformed score can lower the primary threshold by at most $B/4$ relative to $H_0(q_0)$. For an original member with base score $s \le q_0$, deletion therefore requires Eq. (9). Indeed, when this difference is at least $B/2$, the member's largest possible transformed primary score is at most the transformed threshold's smallest possible primary value. Equality still includes the member: at most $k - 1$ calibration scores have base score strictly below q_0, and the secondary base-score key preserves their order at a tied primary boundary. The inclusive count map preserves source ties. If $k = n + 1$, the threshold is infinite and the finite transformed rule includes all labels.

We apply this necessary condition to members in $C_0 \setminus C_1$, where C_1 is the separately calibrated direct teacher. At the clean 95% operating point the teacher requests 24.008 removals per input on average. With $B = .5$, 23.773 requests are excluded by (9); only .235 remain eligible. At $B = 8$, 20.225 are excluded and 3.782 are eligible. The mean excluded fractions are 99.02% and 84.24%, respectively.

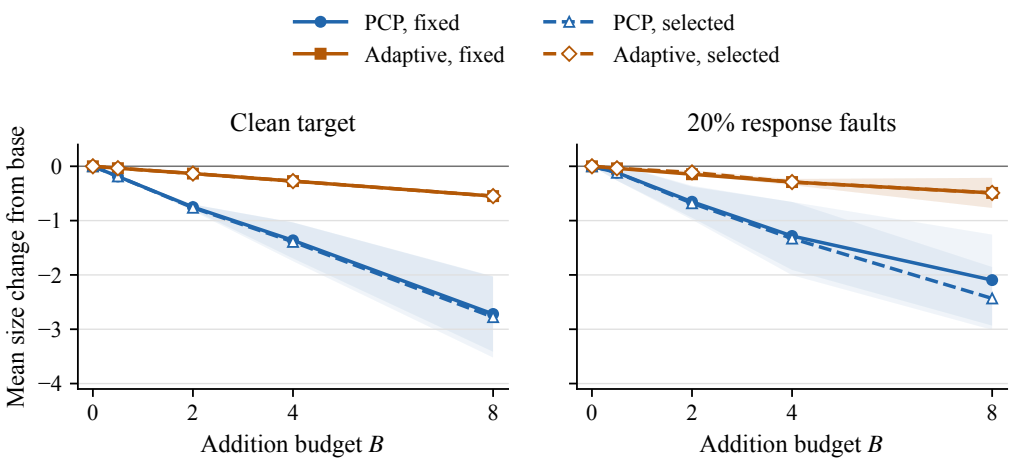

Figure 10: A larger allowed budget exposes a local projection advantage over uniform shrinkage. ImageNetV2 at the frozen 95% nominal operating point: paired mean-size change from the same base, with negative values indicating fewer labels. Curves use the fixed teacher or directions selected from the same auxiliary library. Bands are ± one sample SD across three overlapping data splits, not confidence intervals. Faults replace 20% of target responses in auxiliary, calibration and test roles under the same law; they are not a test-only distribution shift. Budgets were fixed before these new evaluations. Coverage and all 99% results appear in the tables.

PCP removes .187 and 2.794 base members and adds 0 and .075, giving net size reductions .187 and 2.719. Actual removals can include labels outside the teacher-requested subset; they must not be identified with realized eligible requests. The condition is necessary, not sufficient, and uses the final base threshold as a post-hoc diagnostic without test labels. It is neither a new auxiliary-only predictor nor a new risk certificate.

F.4 Complete controls and operating-point dependence

Tables 8 and 9 retain all original conditions. Same-budget clipping improves over adaptive shrinkage in the 95% study, including the shared finite target-direction library. This does not establish dominance over every possible mixture, raw-log direction or temperature. At the 99% operating point the fixed-teacher PCP comparison improves on only one of three clean splits: paired differences are $-2.0778, .1146, .9376$. The fault-condition differences are $-2.1332, 1.2662, 1.0384$. These results preclude a uniform real-data superiority claim.

Unprotected fusion supplies a distinct efficiency reference. In the clean 95% case, direct teacher/selected mixture size is 9.559, substantially below PCP, with observed mean additions 2.179. Under target faults, direct teacher size rises to 778.091, but condition-aware unprotected mixture selection gives 19.744 with observed additions 3.380. Their observed additions describe these finite batches, whereas Theorem 3 supplies an arbitrary-target certificate for the protected constructions. Independent-reference projection has clean/fault means $28.648/29.201$, close to or below PCP's $28.668/29.292$, reinforcing that the numerical advantage is not uniquely due to the pooled map. The guarantee, operator comparison and empirical target quality are separate claims. All $B = 0$ results recover the base, and replay at .5 and 2 matches the original archived size, coverage, addition and removal totals.

Table 8: Complete ImageNetV2 budget panel at nominal 95% coverage. Every entry is mean ± sample SD across the same three overlapping splits; all 32 method/budget pairs are retained for each fault condition. Selected directions use only auxiliary data, with the same base, target and four probability-mixture directions in both protected families. A dash means no specified addition budget. †Reference projection matches the displacement amplitude but lacks the same exact pooled guarantee. Unprotected mixtures are strong efficiency controls, not comparisons at an equal certified budget; all methods independently recalibrate. Coverage is empirical.

Method	B	Clean target		20% response faults	
		Size	Cov. (%)	Size	Cov. (%)
Base	–	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
Direct target	–	9.559 ± 0.872	94.900 ± 0.495	778.091 ± 12.041	95.067 ± 0.295
Prob. mix $\lambda = .25$	–	16.093 ± 3.694	94.807 ± 0.926	20.636 ± 4.346	94.913 ± 0.674
Prob. mix $\lambda = .50$	–	12.838 ± 2.307	94.900 ± 0.828	18.912 ± 2.806	94.953 ± 0.429
Prob. mix $\lambda = .75$	–	10.755 ± 1.306	95.053 ± 0.669	20.041 ± 2.560	94.933 ± 0.584
Prob. mix $\lambda = 1$	–	9.559 ± 0.872	94.900 ± 0.495	778.091 ± 12.041	95.067 ± 0.295
Selected prob. mix	–	9.559 ± 0.872	94.900 ± 0.495	19.744 ± 2.532	94.853 ± 0.602
Adaptive, fixed	0	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
	0.5	31.352 ± 0.734	94.820 ± 0.334	31.350 ± 0.737	94.820 ± 0.334
	2	31.254 ± 0.738	94.813 ± 0.336	31.242 ± 0.751	94.813 ± 0.336
	4	31.115 ± 0.742	94.813 ± 0.336	31.097 ± 0.781	94.820 ± 0.322
	8	30.839 ± 0.744	94.813 ± 0.336	30.898 ± 0.967	94.840 ± 0.280
PCP, fixed	0	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
	0.5	31.201 ± 0.741	94.820 ± 0.334	31.272 ± 0.856	94.840 ± 0.308
	2	30.636 ± 0.780	94.793 ± 0.344	30.735 ± 0.991	94.827 ± 0.306
	4	30.024 ± 1.021	94.827 ± 0.289	30.104 ± 1.293	94.827 ± 0.273
	8	28.668 ± 1.355	94.753 ± 0.269	29.292 ± 1.315	94.813 ± 0.179
Adaptive, selected	0	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
	0.5	31.352 ± 0.734	94.820 ± 0.334	31.350 ± 0.737	94.820 ± 0.334
	2	31.254 ± 0.738	94.813 ± 0.336	31.275 ± 0.699	94.820 ± 0.334
	4	31.115 ± 0.742	94.813 ± 0.336	31.097 ± 0.781	94.820 ± 0.322
	8	30.839 ± 0.744	94.813 ± 0.336	30.898 ± 0.967	94.840 ± 0.280
PCP, selected	0	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
	0.5	31.202 ± 0.746	94.820 ± 0.334	31.269 ± 0.860	94.840 ± 0.308
	2	30.625 ± 0.790	94.793 ± 0.344	30.707 ± 1.002	94.827 ± 0.306
	4	29.997 ± 1.046	94.827 ± 0.289	30.057 ± 1.342	94.827 ± 0.273
	8	28.611 ± 1.408	94.753 ± 0.269	28.955 ± 0.995	94.780 ± 0.216
Reference†	0	31.388 ± 0.733	94.827 ± 0.332	31.388 ± 0.733	94.827 ± 0.332
	0.5	31.208 ± 0.751	94.820 ± 0.334	31.276 ± 0.860	94.840 ± 0.308
	2	30.647 ± 0.804	94.793 ± 0.344	30.745 ± 1.011	94.827 ± 0.306
	4	30.013 ± 1.063	94.813 ± 0.301	30.137 ± 1.262	94.833 ± 0.284
	8	28.648 ± 1.383	94.753 ± 0.269	29.201 ± 1.391	94.800 ± 0.173

Table 9: Complete ImageNetV2 budget panel at nominal 99% coverage. Every entry is mean ± sample SD across the same three overlapping splits; all 32 method/budget pairs are retained for each fault condition. Selected directions use only auxiliary data, with the same base, target and four probability-mixture directions in both protected families. A dash means no specified addition budget. †Reference projection matches the displacement amplitude but lacks the same exact pooled guarantee. Unprotected mixtures are strong efficiency controls, not comparisons at an equal certified budget; all methods independently recalibrate. Coverage is empirical.

Method	B	Clean target		20% response faults	
		Size	Cov. (%)	Size	Cov. (%)
Base	–	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
Direct target	–	134.085 ± 6.270	99.100 ± 0.100	957.117 ± 14.142	99.053 ± 0.300
Prob. mix $\lambda = .25$	–	130.615 ± 11.014	98.940 ± 0.242	130.652 ± 11.167	98.940 ± 0.242
Prob. mix $\lambda = .50$	–	127.273 ± 7.583	98.960 ± 0.171	128.513 ± 8.276	98.947 ± 0.190

Continued on next page.

Continued from previous page.

Method	B	Clean target		20% response faults	
		Size	Cov. (%)	Size	Cov. (%)
Prob. mix $\lambda = .75$	–	119.979 ± 7.935	99.040 ± 0.178	122.192 ± 8.790	99.013 ± 0.234
Prob. mix $\lambda = 1$	–	134.085 ± 6.270	99.100 ± 0.100	957.117 ± 14.142	99.053 ± 0.300
Selected prob. mix	–	128.012 ± 4.556	99.067 ± 0.058	122.192 ± 8.790	99.013 ± 0.234
Adaptive, fixed	0	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
	0.5	131.850 ± 12.911	98.940 ± 0.255	131.852 ± 12.908	98.940 ± 0.255
	2	131.735 ± 12.870	98.940 ± 0.255	131.742 ± 12.864	98.940 ± 0.255
	4	131.583 ± 12.829	98.940 ± 0.255	131.597 ± 12.805	98.940 ± 0.255
	8	131.350 ± 12.671	98.940 ± 0.255	131.301 ± 12.704	98.940 ± 0.255
PCP, fixed	0	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
	0.5	131.777 ± 12.871	98.940 ± 0.255	131.860 ± 12.783	98.940 ± 0.255
	2	131.706 ± 12.469	98.940 ± 0.255	131.784 ± 12.376	98.940 ± 0.255
	4	131.612 ± 11.905	98.940 ± 0.255	131.675 ± 11.816	98.940 ± 0.255
	8	131.008 ± 11.263	98.953 ± 0.252	131.358 ± 10.811	98.973 ± 0.225
Adaptive, selected	0	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
	0.5	131.850 ± 12.911	98.940 ± 0.255	131.852 ± 12.908	98.940 ± 0.255
	2	131.735 ± 12.870	98.940 ± 0.255	131.742 ± 12.864	98.940 ± 0.255
	4	131.583 ± 12.829	98.940 ± 0.255	131.597 ± 12.805	98.940 ± 0.255
	8	131.355 ± 12.668	98.940 ± 0.255	131.301 ± 12.704	98.940 ± 0.255
PCP, selected	0	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
	0.5	131.777 ± 12.871	98.940 ± 0.255	131.791 ± 12.855	98.940 ± 0.255
	2	131.706 ± 12.469	98.940 ± 0.255	131.501 ± 12.678	98.940 ± 0.255
	4	131.459 ± 12.006	98.940 ± 0.255	131.113 ± 12.431	98.940 ± 0.255
	8	130.701 ± 11.441	98.940 ± 0.255	130.358 ± 11.757	98.940 ± 0.262
Reference[†]	0	131.889 ± 12.924	98.940 ± 0.255	131.889 ± 12.924	98.940 ± 0.255
	0.5	131.789 ± 12.859	98.940 ± 0.255	131.861 ± 12.782	98.940 ± 0.255
	2	131.695 ± 12.456	98.940 ± 0.255	131.776 ± 12.358	98.940 ± 0.255
	4	131.583 ± 11.865	98.940 ± 0.255	131.645 ± 11.770	98.940 ± 0.255
	8	130.925 ± 11.153	98.953 ± 0.252	131.269 ± 10.692	98.973 ± 0.225

G CROSS-ARCHITECTURE TRANSFER TO CONVNEXT

Research question and protocol. This study tests whether the protected-utility comparison persists when a ConvNeXt-Tiny checkpoint replaces the ResNet-50 target (Liu et al., 2022). The ImageNetV2 matched-frequency pool and ResNet-18 base remain fixed, isolating target architecture while preserving the paired evaluation setting. Its protocol was frozen internally before target logits were exported or final labels were opened (protocol SHA-256 prefix `af59650a701b`). Native forward logits were converted by binary64 log-sum-exp to LAC-equivalent negative log probabilities at temperature one; the official target resize uses 236 pixels before a 224-pixel crop. The target-logit SHA-256 prefix is `98fed60e3783`; full hashes were recorded by the original experiment.

For each of three prespecified split seeds, 1,000 images serve as a reference, 2,000 as auxiliary data, 2,000 as final calibration and 5,000 as test. Roles are disjoint within a split and paired across methods, but the three partitions overlap. Both PCP and adaptive linear independently select $B \in \{0, 0.5, 2, 4, 8\}$ using the same two-fold auxiliary size objective: each auxiliary half calibrates on one fold and evaluates mean size on the other; summed size is minimized, with smaller B breaking a tie. The $B = 0$ base is always eligible. The selected policies are sealed before final evaluation, and each method is separately calibrated at nominal 95% coverage. The primary comparison is selected PCP versus selected adaptive linear; fixed-budget, direct-target and independent-reference results are controls. Both selectors chose $B = 8$ in every split. A local full-mask replay of all final masks agreed with the recorded outputs; this is an implementation check, not an independently collected test.

Across the three splits, base, selected PCP and selected linear sizes are 30.943 ± 3.157, 28.686 ± 2.605 and 30.310 ± 3.121 (mean $\pm$ sample SD). Their corresponding attained coverages are $94.947 \pm 0.428\%$, $95.000 \pm 0.393\%$ and $94.933 \pm 0.437\%$. PCP reduces size by 1.6244 labels (5.36%) relative

Table 10: Full frozen ConvNeXt formal-budget grid on ImageNetV2. Values are labels/input and mean attained coverage over the same three overlapping role splits after separate final calibration. Added and removed labels compare each rule to the separately recalibrated ResNet-18 base on the same test images. All five budgets were specified before final evaluation. At a matched formal B, observed additions need not match.

	PCP				Adaptive linear			
B	Size ↓	Cov. (%)	Added	Removed	Size ↓	Cov. (%)	Added	Removed
0	30.9429	94.947	0	0	30.9429	94.947	0	0
0.5	30.7499	94.940	0	0.1929	30.9071	94.947	0	0.0357
2	30.2035	94.927	0.0117	0.7511	30.7905	94.947	0	0.1523
4	29.4429	94.913	0.0166	1.5165	30.6326	94.947	0.0001	0.3103
8	**28.6858**	95.000	0.1794	2.4365	30.3102	94.933	0.0001	0.6328

Table 11: Selected $B = 8$ ConvNeXt outcomes by paired role split. Coverage is an attained test proportion, not exact equality between methods. PCP-minus-linear size differences are $-1.2154, -2.3292, -1.3286$; the split overlap precludes treating them as independent training replications.

Split	Base size	PCP size	Linear size	PCP cov. (%)	Linear cov. (%)	PCP added / removed
1	30.5972	28.7292	29.9446	95.06	95.02	0.2848 / 2.1528
2	34.2582	31.2692	33.5984	95.36	95.32	0.0280 / 3.0170
3	27.9732	26.0590	27.3876	94.58	94.46	0.2254 / 2.1396

to linear at the same formal $B = 8$, but uses more of that allowance: additions are 0.1794 ± 0.1338 versus 0.0001 ± 0.0002, while removals are 2.4365 ± 0.5028 versus 0.6328 ± 0.0411. The fixed-budget grid in Table 10 and Fig. 5 therefore support a local advantage at matched *formal* protection, not at matched observed additions.

The direct ConvNeXt target returns 7.046 ± 0.322 labels at $94.747 \pm 0.310\%$ attained coverage, and a fixed-$B = 8$ independent-reference projection returns 28.507 ± 2.597 labels at $94.993 \pm 0.404\%$. The latter uses a matched adjustment amplitude but lacks the exact pooled same-base certificate, while the direct target is not constrained to an arbitrary-target addition budget. They delimit the source and magnitude of the protected gain: PCP recovers 2.2571 of the direct target's 23.8966-label observed net reduction (9.45%). The budget selector chose the largest allowed value in all six method-split decisions; this establishes independent auxiliary selection without demonstrating adaptation to diverse safety preferences. The original experiment recorded per-split masks, scores, protocol metadata, input hashes, and the frozen fixed-budget results.

H Exploratory native-energy score-family check

We retain the earlier native-logit reanalysis as a score-family check, not an independent checkpoint replication. It uses two previously studied archives and three overlapping role splits per model, with auxiliary selection followed by independent final recalibration. Table 12 reports every model/score branch. PCP has a lower descriptive mean size than same-base adaptive linear in all four branches, most clearly for WRN/SAPS (19.295 versus 21.054 labels). The LAC branches do not improve on their unweighted bases (WRN: 11.678 versus 11.485; JEM: 1.626 versus 1.596); independently selected unweighted LAC is also a stronger external score comparator than protected SAPS. Thus the check concerns adaptation within a fixed base, not overall score-family selection. On WRN/SAPS at the selected $B = 8$, PCP and adaptive linear have different realized additions, 0.079 and 0.009 labels per input, despite sharing the formal budget.

Table 12: Exploratory native-energy adaptation on two previously studied native-logit archives. Each method uses its own independently recalibrated threshold at nominal 99% coverage. Size is mean ± sample SD over three overlapping splits (descriptive, not independent-model uncertainty). Coverage and additions are split means; additions are relative to the same-branch unweighted base. Protected methods select energy strength $\gamma \in \{.25, .5, 1, 2\}$ and formal budget $B \in \{0, .5, 2, 4, 8\}$ using only auxiliary folds before final calibration and test. The direct raw rule uses $\gamma = 1$ without the same-base addition protection. The stronger LAC rule is an external score comparator for the SAPS branch, not a same-base protected method. All branches are shown, regardless of effect direction.

Rule	Size ↓	Cov. (%)	Added ↓	Removed ↑
Energy-trained WRN / CIFAR-100: SAPS, $T = 1$, $\lambda = .2$				
Unweighted branch base	21.657 *±3.104*	99.044	0.000	0.000
Direct raw-energy weighting	13.675 *±3.152*	98.767	1.552	9.534
Adaptive linear, selected	21.054 *±3.092*	99.033	0.009	0.612
PCP, selected	19.295 *±3.121*	99.022	0.079	2.440
Strong unweighted LAC	11.485 *±1.050*	99.122	1.364	11.535
Energy-trained WRN / CIFAR-100: Auxiliary-selected LAC				
Unweighted branch base	11.485 *±1.050*	99.122	0.000	0.000
Direct raw-energy weighting	14.172 *±2.276*	98.978	3.913	1.225
Adaptive linear, selected	11.753 *±0.859*	99.100	0.397	0.130
PCP, selected	11.678 *±1.343*	99.089	0.648	0.455
JEM / CIFAR-10: SAPS, $T = 1$, $\lambda = .2$				
Unweighted branch base	2.883 *±0.337*	99.167	0.000	0.000
Direct raw-energy weighting	4.871 *±0.065*	99.289	2.357	0.368
Adaptive linear, selected	2.743 *±0.447*	99.222	0.101	0.241
PCP, selected	2.589 *±0.168*	99.111	0.224	0.518
Strong unweighted LAC	1.596 *±0.084*	99.022	0.062	1.350
JEM / CIFAR-10: Auxiliary-selected LAC				
Unweighted branch base	1.596 *±0.084*	99.022	0.000	0.000
Direct raw-energy weighting	2.843 *±0.913*	99.211	1.347	0.099
Adaptive linear, selected	1.658 *±0.016*	99.067	0.082	0.020
PCP, selected	1.626 *±0.045*	99.033	0.052	0.022

I MECHANISM AND TIMING CONTROLS

Table 13: **Nonempty mechanism controls.** Entries are mean ± sample SD over 12 independent repetitions with 3,500 calibration and 12,000 test inputs each. Size is labels per input; Cov. is posterior-expected coverage (%) at nominal 99% coverage. All rules use the top-score-zero variant and are calibrated separately. Fixed linear and projection satisfy the $B = 0.5$ bound on expected additions relative to the same separately calibrated base; the raw target and weight-transform controls are not subject to this constraint.

Method	**Gaussian** $d = 16$		**Gaussian** $d = 256$		**Near-constant APS**	
	Size ↓	Cov. (%)	Size ↓	Cov. (%)	Size ↓	Cov. (%)
Base	1.8998 *±0.0145*	98.989 *±0.145*	1.9007 *±0.0147*	98.998 *±0.147*	1.8980 *±0.0156*	98.971 *±0.156*
Raw target	62.0475 *±2.9312*	98.954 *±0.197*	85.7955 *±2.4306*	98.980 *±0.198*	31.3519 *±43.5197*	98.948 *±0.164*
Tempered weight ($\tau = .25$)	11.7074 *±0.6751*	98.971 *±0.182*	67.6951 *±3.5348*	98.982 *±0.188*	31.3519 *±43.5197*	98.948 *±0.164*
Weight clip ($c = .2$)	2.3830 *±0.0482*	98.993 *±0.209*	2.3925 *±0.0289*	99.021 *±0.128*	31.3519 *±43.5197*	98.948 *±0.164*
Tanh weight cap ($c = .1$)	2.0777 *±0.0468*	98.990 *±0.211*	2.0830 *±0.0285*	99.020 *±0.131*	31.3519 *±43.5197*	98.948 *±0.164*
Fixed linear	1.9008 *±0.0163*	98.999 *±0.163*	1.9004 *±0.0141*	98.995 *±0.141*	1.8956 *±0.0142*	98.947 *±0.141*
Projection (ours)	1.9061 *±0.0142*	99.007 *±0.141*	1.9026 *±0.0146*	98.997 *±0.145*	1.8925 *±0.0153*	98.916 *±0.153*

Table 13 supplies the dispersion and posterior-expected coverage for the constructed repairs in Fig. 4. The runtime panel below describes the available postprocessing measurements; it is not a neural-network inference benchmark.

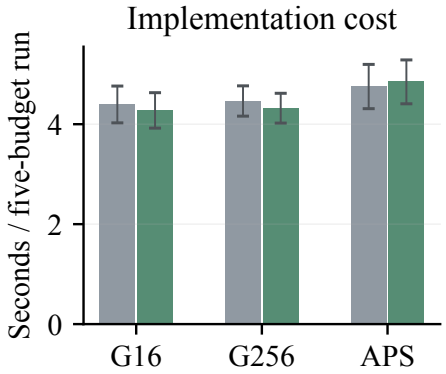

Figure 11: **Postprocessing time on the nonempty mechanism workloads.** Gray bars show fixed linear and green bars show PCP, mean ± sample SD over 12 calibration/data repetitions for the five-budget run (3,500 calibration inputs, 12,000 test inputs, 100 classes). G16 and G256 denote Gaussian nuisance dimensions; APS is the near-constant-weight construction. Timings exclude network inference and initial score construction and are descriptive measurements on one implementation, not a hardware-normalized benchmark.

Score-spacing boundary control. The near-constant-weight construction targets APS rather than every rank-based score. Under the same $T = 0.1$ setting and twelve independent nonempty repetitions, the weighted SAPS control has the same mean size and posterior-expected coverage as its base,

1.8980 ± 0.0156 labels and $98.971 \pm 0.156\%$, respectively. The paired APS raw target instead has mean size 31.3519 ± 43.5197 at $98.948 \pm 0.164\%$ posterior-expected coverage (Table 13). This contrast delimits the role of score spacing in the second construction.

J SCOPE OF THE GUARANTEES AND EMPIRICAL COMPARISONS

What the budget controls. Theorem 3 bounds expected additions to the same separately calibrated base, averaging final calibration and query randomness conditional on independent external choices. It is not a per-image cap or a conditional-on-the-realized-calibration guarantee. Label-free row symmetry and exchangeability are essential; a test-only distribution shift is outside these assumptions. The response-fault experiment instead applies a common law to calibration and test, conditional on its external donor bank.

Formal protection and observed cost. Equal B gives the two bounded operators the same certificate, but their observed additions can differ. For ConvNeXt at $B = 8$, PCP adds 0.1794 labels/input and adaptive linear adds 0.0001. The reported size advantage is therefore a matched-formal-budget comparison, not dominance at matched realized additions. The finite-sample coverage proportions in the tables can also differ even when both methods use the same nominal coverage target.

Dependence on the base and operating point. The real-data claim concerns specified bases, targets, and protection budgets. Unrestricted score selection can yield smaller sets: on the new WRN, auxiliary-selected LAC returns 9.450 ± 0.775 labels at 98.673% attained coverage, compared with selected PCP's 17.459 labels at 98.813% within the SAPS branch. Direct ResNet-50 and ConvNeXt targets return 9.559 and 7.046 labels in their clean 95% settings. These controls use different score families or lack the same arbitrary-target addition constraint; their full results remain in the corresponding experiment sections.

The 99% ResNet operating point also qualifies the target dependence. Its fixed-teacher PCP-minus-linear differences are -2.0778, 0.1146, and 0.9376 across the three clean splits, so the 95% advantage does not extend uniformly across the evaluated configurations. Score temperatures differ between those operating points, and their contrast cannot be assigned to nominal coverage alone. Complete grids are retained in Tables 8–9.

Selection and experimental replication. Auxiliary policies are frozen before final calibration, while protocol design can draw on earlier experiments. The frozen records are internal rather than external preregistrations. Real-data role splits reuse checkpoints and images, including the shared ImageNetV2 pool across target architectures. Their sample SDs measure descriptive partition variability, not uncertainty across independently trained models. In the WRN policy, gate search and linear strength search have different candidate identities; fixed-setting ablations separate those policy choices from the shape comparison. The gate grid gives one smaller, four tied, and 55 larger outcomes among 60 nonuniform-gate comparisons at matched settings, so the primary gain is not attributed to the gate itself.

Relation to other robustness objectives. Nonmonotonic risk control and adversarial or poisoning robustness address distinct losses and assumptions (Angelopoulos, 2026; Zargarbashi et al., 2024; Massena et al., 2025; Liu et al., 2024; Scholten & Günnemann, 2025; Obadinma et al., 2024). These objectives are distinct from the same-law response-fault experiment. Numerical replays validate recorded score orderings without certifying every floating-point library operation. The timing panel reports a specific conformal postprocessing workload and excludes neural inference; its machine metadata is insufficient for a hardware-normalized benchmark.